

GBS 760: BACTERIAL GENETICS & PHYSIOLOGY COURSE PACKET

University of Alabama at Birmingham
Graduate Biomedical Sciences, Microbiology Theme
January / February 2025

Michael J. Gray, M.S., Ph.D.
Associate Professor of Microbiology
email: mjgray@uab.edu
website: graymicrolab.com
GitHub: [lab repository](https://github.com/mjgray)

TABLE OF CONTENTS

3 - 16	<u>Lecture 1: Introduction to Molecular Microbiology</u>
10	Discussion Problem Set #1: Databases and Literature Searches
13	Discussion Problem Set #2: BLAST & Pairwise Alignment
16	Discussion Problem Set #3: BLAST & Multiple Alignment
17 – 28	<u>Lecture 2: Mutants and Mutations</u>
15	Scientific Process 1: Observations and Phenomena
19	Discussion Problem Set #4: Bacterial Phenotypes
23	Scientific Process 2: Models and Hypotheses
24	Discussion Problem Set #5: Proposing Models Based on Data
27	Discussion Problem Set #6: Proposing Hypotheses to test Models
29 – 42	<u>Lecture 3: Mutant Hunts and Experimental Design</u>
27	Scientific Process 3: Experiments, Variables, and Controls
35	Discussion Problem Set #7: Limitations of Transposons
39	Discussion Problem Set #8: Screens and Selections
42	Scientific Process 4: Alternative Approaches and Troubleshooting
43 – 63	<u>Lecture 4: Principles of Regulation</u>
51	Discussion Problem Set #9: Transcriptional Regulation
59	Discussion Problem Set #10: Two-Component Regulators
62	Discussion Problem Set #11: Post-Transcriptional Regulation
64 – 73	<u>Lecture 5: Plasmids</u>
64	Scientific Process 5: Correlation and Causation
65	Discussion Problem Set #12: Correlation and Causation
68	Discussion Problem Set #13: Using Plasmids in Genetic Experiments
74	<u>Lecture 6: Critical Reading (Mutagenesis and Mutant Hunts)</u>
75 – 93	<u>Lecture 7: Principles of Genetic Engineering</u>
86	Discussion Problem Set #14: Designing Constructs for Genetic Experiments
92	Discussion Problem Set #15: Designing Constructs for Genetic Experiments 2
94 – 110	<u>Lecture 8: Gene Transfer and Recombination</u>
95	Discussion Problem Set #16: Gene Transfer
108	Discussion Problem Set #17: Strain Construction with Recombination
110	Scientific Process 6: Scientific Reasoning
111	<u>Lecture 9: Critical Reading (Genetic Engineering)</u>
112 – 126	<u>Lecture 10: Bacterial Cell Envelopes</u>
118	Discussion Problem Set #18: Regulation of Autolysis in <i>Streptococcus gordonii</i>
122	Discussion Problem Set #19: Not All Genetic Systems are Created Equal
126	Discussion Problem Set #20: Antibiotic Resistance in <i>Faecalibacterium prausnitzii</i>
127 – 143	<u>Lecture 11: Bacterial Cytoskeleton and Development</u>
130	Discussion Problem Set #21: Bactofilins from <i>Caulobacter crescentus</i>
138	Discussion Problem Set #22: Cyanobacterial Heterocysts
144	<u>Lecture 12: Critical Reading (Bacterial Cell Structure)</u>
145 – 160	<u>Lecture 13: Protein Secretion</u>
156	Discussion Problem Set #23: Toxin Secretion by <i>Photorhabdus luminescens</i>

160	Discussion Problem Set #24: Protein Complex Formation in <i>Escherichia coli</i>
161 – 180	<u>Lecture 14: Capsule and Biofilms</u>
176	Discussion Problem Set #25: Biofilms in the <i>Vibrio cholerae</i> Life Cycle
179	Discussion Problem Set #26: Cyclic-di-GMP Signaling Specificity
181 – 192	<u>Lecture 15: Motility</u>
190	Discussion Problem Set #27: Intracellular Motility in <i>Mycobacterium marinum</i>
191	Discussion Problem Set #28: Phototaxis by a Bacterial Consortium
193 – 209	<u>Lecture 16: Bacterial Energetics</u>
196	Discussion Problem Set #29: Responding to Stress by Collapsing the PMF
203	Discussion Problem Set #30: Metal Reduction by <i>Shewanella oneidensis</i>
210 – 228	<u>Lecture 17: Central Metabolism</u>
213	Discussion Problem Set #31: PTS and Non-PTS Sugar Phosphorylation
218	Discussion Problem Set #32: Carbon Catabolite Repression
229 – 239	<u>Lecture 18: Secondary Metabolism</u>
237	Discussion Problem Set #33: Activating Silent Biosynthetic Gene Clusters
238	Discussion Problem Set #34: Phage-Inhibiting Secondary Metabolites
240	<u>Lecture 19: Critical Reading (Bacterial Metabolism)</u>
241	<u>Summary of Experimental Design Principles</u>
242	<u>Grading Rubric for Weekly Quiz Problems</u>
243 – 265	<u>Glossary</u>



© 2024 by Michael J. Gray

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

LECTURE I: INTRODUCTION TO MOLECULAR MICROBIOLOGY

INTRODUCTION

The goal of this course packet is to familiarize you with the nomenclature and concepts you will need to participate in each lecture. **Your level of participation in in-class discussions will be a significant determinant of your grade**, and most of the time in class will be dedicated either to small-group problem solving and discussions based on the information and problem sets in each day's reading or to journal club-style discussions of specific scientific papers, so I strongly recommend that you do the reading for each day's class ahead of time. I will begin each lecture with a short question and answer session, so if there's anything in the reading you don't understand or would like clarified, please come prepared to ask. We will then work through the discussion problems during the class period itself. I do not plan to spend very much time actually lecturing at you.

While the digital version of this course packet contains many hyperlinks to useful resources and interesting papers, if you prefer to print out a hard copy, I have formatted it in a way that should be compatible with either spiral or three-ring binding.

In this first lecture, we will discuss some core principles of scientific literacy, including the basics of the scientific method and using the scientific literature. I will also introduce the basics of molecular microbiology, including the fundamentals of genetic nomenclature in bacteria. This will set the foundation for future lectures, in which we will explore the practical and theoretical implementation of the scientific method for experiments in microbial genetics.

EXPECTATIONS AND LEARNING GOALS

In this course, my goal is for you to learn how to think about and apply the tools of bacterial molecular genetics to solve scientific problems, and then to use that knowledge to build a strong foundation of understanding the molecular mechanisms by which bacteria grow, function, and cause disease. To achieve this goal, we will need to build your skills in two fundamental areas:

- **Scientific literacy:** understanding what is and isn't known, how those facts fit into the larger framework of scientific knowledge, and how to search and read the scientific literature
- **Scientific proficiency:** understanding how to design, carry out, and interpret experiments, knowing what tools you have available and creatively applying those tools to answer specific questions

The product of scientific work is knowledge. I want to give you the tools to effectively access and add to that knowledge.

By the end of this course, I want you to:

- be able to define the steps of the scientific method and develop models and hypotheses based on data
- know where to find information about bacterial genes and proteins
- be able to use and understand the nomenclature of bacterial genetics
- understand the principles of mutagenesis and genetic engineering in microbes
- know how to interpret mutant phenotypes in different kinds of genes and with different kinds of mutations
- have a good working knowledge of bacterial physiology and cell biology
- be able to design rigorous experiments to solve biological problems using bacterial genetics

A glossary of important terms, which are indicated in the text in *italics* the first time they appear, can be found starting on page 243. See page 241 for a concise summary of all of the experimental design principles that we will discuss in the course of these chapters.

Class participation will be evaluated using the following scale and will account for slightly less than half of your grade (76 points out of 156 total):

4 points - Student comes to class prepared; contributes readily to the conversation but doesn't dominate it; makes thoughtful contributions that advance the conversation; shows an interest in and respect for others' contributions; participates actively in all groups.

3 points - Comes to class prepared and makes thoughtful comments when called upon; contributes occasionally without prompting; shows interest in and respect for other's views; participates in small groups.

2 points - Student is poorly prepared or participates in discussion, but in a problematic way: e.g. talks too much, rambles, interrupts instructor and others, or does not acknowledge cues of annoyance from others.

1 point - Has not prepared for class or does not contribute to discussion; displays disrespect towards students and/or faculty.

0 points - Absent without explanation.

If you do need to be absent for a day's class, please let me know ahead of time if at all possible. I will then give you an assignment to do to make up the day's points. This will typically be a written answer to one or more of that day's Discussion Problems but may vary depending on the particular day in question.

There will also be weekly quizzes, which will account for the remainder of your grade. These will be made available each Friday morning on the course Canvas page and are due the following **Monday morning at 8 AM**. Each quiz is worth 20 points, accounting for 80 points out of 156 total. You will lose 5 points per day that the quiz is late without explanation. Please do not work with other students on the quizzes; they are meant to assess your personal progress. The rubric I will use to grade the quizzes is on page 242.

If you are concerned about your grade or class status at any point during the class, please contact me immediately. I am very happy to talk to you outside of class to try to clear up any confusing points or help you get the support you need, but I am unlikely to be able to guess what those needs are unless you come to talk to me. Dr.Yother also has tutors available if you feel that you need extra help.

It is worth noting, of course, that this class is structured to emphasize the topics and concepts I think are important, which are not necessarily the same as another instructor might focus on. Those people are not necessarily wrong, and I am certainly not always right, so I will try not to be too dogmatic in this course. A diversity of opinions and approaches is one of the great strengths of science, and you should draw on all of the resources, mentors, and instructional material available to form your own personal scientific knowledge base and approach to research.

Also note that the goal of this class is to teach you to think about scientific problems using your own personal brain. Please do not use generative AI software to attempt to solve any of the problems presented to you in this class.

SCIENTIFIC LITERACY

It has been a very long time since it has been possible for any one person to know everything there is to know about science. What I mean by being scientifically literate has three distinct elements:

- Understanding what scientific knowledge is and is not, and understanding the scientific method.
- Having a good general grasp of the broad state of knowledge across scientific disciplines.
- Having a deep and up-to-date understanding of your own area of specialization.

In this section I will summarize the scientific method, briefly discuss what molecular microbiology is and how it fits into the spectrum and history of science and describe how to read and understand the scientific literature.

THE SCIENTIFIC METHOD

The goal of science is to learn truths about reality. The *scientific method*, more than anything, is a systematic approach we use to uncover those truths in a reliable way. Understanding and appreciating the scientific method is the core of scientific literacy.

Science begins with a **question**. There is something we don't know that we have reason to look at more closely and which we want to understand more fully. This can be very broad (e.g. what affects the spread of influenza?) or very specific (e.g. what is the role of glutamate 245 in the polyphosphate kinase enzyme of *E. coli*?), but the process always begins by identifying something we don't know.

How do we find an answer to the question? The next step in the scientific method is to develop a *hypothesis*. A hypothesis is a **possible** answer to the question and is informed by whatever else the scientist knows about the subject. The most important feature of a hypothesis is that it must be *falsifiable* or *testable*, which leads directly to the next step in the process.

What distinguishes science from other types of inquiry about the nature of reality is that in science we rigorously test our hypotheses. Whether via *observations* or *experiments*, the scientist puts their ideas to the test, **discarding hypotheses that do not match the facts**. This process of testing hypotheses results in the development of a *model* to explain the **mechanism** underlying the observations the scientist has made, and addition of more observations may strengthen or weaken that model. Models to explain natural phenomena start simple and gain complexity and *predictive power* as more facts are discovered and incorrect hypotheses are discarded. If an observation is made that does not fit with the model, the scientist must change the model to incorporate the new data, and test any new predictions made by those changes. Developing a model is a deeply creative process, drawing on all the knowledge of the scientist, with the fundamental constraint that a valid model must explain **all** of the observations of a system. By reiterating this self-correcting process, scientific knowledge converges on truth.

I really like the way that I heard Dr. Nathan Erdmann (from the UAB Department of Medicine) explain his approach to science. To paraphrase, he thinks of the scientist's job as two-fold: 1) to decipher **signal** from **noise** and 2) to ask **meaningful, answerable** questions. I think this is really important. How do you make sure that your observations are not just random variation? Has someone already answered the question? Do the tools exist to allow you to ask the question? What impact will knowing the answer to the question have?

In future lectures we will practice developing hypotheses, models, and experiments and go into more detail about what each of those steps entails.

WHAT IS MOLECULAR GENETICS?

Genetics is an approach to understanding biological systems that involves manipulating the heritable genetic material of an organism (its *genotype*) and observing the changes that result from those manipulations (the *phenotype*). It is often contrasted with *biochemistry*, which focuses on the properties of (usually) purified components of cells like individual proteins, nucleic acids, or lipids. Both approaches are essential to understanding how biological systems work. Very often, genetic experiments will provide the first indication of the role of a protein or other cellular component, which will then guide the detailed biochemical analysis of that component. *Molecular genetics* is simply genetics with an understanding of the biochemical nature of DNA and with tools to directly manipulate that genetic material.

Snyder & Champness's [Molecular Genetics of Bacteria](#), now in its 5th edition, is an excellent textbook on this topic, if you're interested in more in-depth, detailed discussion of specific topics than I'm aiming to achieve here.

Talking about molecular biology and genetics requires understanding quite a bit of technical terminology, and I'll try to define the essential jargon here as simply as possible, but you will inevitably have to learn the vocabulary. You'll also need to have at least a reasonable grasp of how the basic biological processes of *transcription* and *translation* work. If you need to review the basics, these articles may help:

[Transcription \(genetics\) - from Wikipedia](#)

[Translation \(biology\) - from Wikipedia](#)

(Wikipedia is a generally reliable source for information on biochemistry and related topics.)

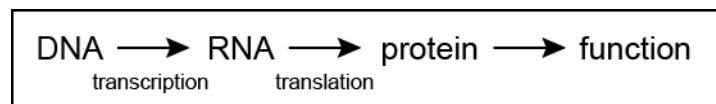
Different organisms are more or less easy to manipulate, and it's important to understand what is technically possible in the species you're studying. An experiment that takes a week in a well-established and easily-manipulated *model organism* like *Escherichia coli* may take months, years, or be impossible in a slow-growing, poorly characterized, or less well-studied species. Of course, new tools and techniques are constantly being developed to try to accelerate difficult procedures, both in academic labs and by commercial companies. We will talk about methods for manipulating DNA molecules in **Lectures 7 - 9**.

Experiments studying the properties of a gene or protein in a living organism, as in genetics, are referred to as *in vivo* studies (Latin for "within the living"). *Ex vivo* ("outside the living") experiments involve the use of cells or tissues removed from a larger organism. *In vitro* ("in glass") experiments, including most biochemistry, involve purified components removed from the cells in which they are normally found. The term *in situ* ("in position") is sometimes used to describe experiments that examine individual cells or organisms in the context of larger systems, without separating them from their natural context. Finally, the term *in silico* (fake Latin for "within silicon") is used to describe experiments performed entirely through computer simulations or calculations. I will note, however, that different fields use these terms differently. See, for an extreme example, [this Wikipedia page](#) on the uses of "*in situ*".

MAJOR CLASSES OF BIOMOLECULES AND THEIR FUNCTIONS

Molecular biology is the study of how organisms function at a biochemical level, and requires an understanding of what kinds of molecules make up living cells and what roles those molecules characteristically play. Since all organisms on Earth are descended from a common ancestor, the types of biomolecules are the same in all cells: bacterial, archaeal, or eukaryotic (and in viruses, too).

To review the so-called Central Dogma of molecular biology, genes are encoded as sequences of nucleotide bases on *chromosomes*, which are long molecules of double-stranded helical DNA (*deoxyribonucleic acid*). These genes are *transcribed* into single-stranded messenger RNA (*ribonucleic acid*) chains. Messenger RNAs (mRNA) are then *translated* into *proteins* (long polymers of amino acids that fold into complex 3-dimensional structures), which carry out enzymatic or regulatory functions within the cell.



This basic picture is, however, a gross oversimplification of the diversity of biomolecular functions, and you should be aware that, for example, there are many forms of *functional RNA* (ribosomal RNA, transfer RNA, small non-coding RNA, *ribozymes*, etc.), that RNA can be *reverse transcribed* into DNA, that some small peptides (short proteins) are synthesized without an mRNA template (we will discuss these in **Lecture 18**), and that *extracellular DNA* (eDNA) can play a structural role (in bacterial biofilms, for example). We've also entirely left out the roles of lipids and carbohydrates. Nothing in biology is simple!

The goal of research in molecular biology is to understand how the complex interactions of these different molecules fit together to form a functioning living cell. Biochemistry and genetics classes will teach you a lot of detailed theory about

what is known so far, and I presume that in order to have gotten this far, you've taken such classes already. In this class, my focus is on giving you the practical and theoretical basis to carry out modern microbial genetics research.

GENES AND GENE PRODUCTS

A gene is a nucleotide sequence that encodes a functional gene product, which is usually a protein, but could also be an RNA molecule. For historical reasons, the terms *gene* and *locus* are often used interchangeably, although loci can also be functional sequences that are not genes themselves (like *operator* sequences involved in controlling expression of certain genes; see [Lecture 4](#)) and sometimes the term *locus* is used to refer to a region containing several related genes or adjacent operons. An *open reading frame* (or *ORF*, sometimes also called a *coding domain sequence* or *CDS*) is a gene sequence that encodes a protein, often predicted based entirely on DNA sequence. Alleles are versions of a particular gene with different sequences, and sometimes with different functional properties. An *operon* is several genes encoded on the same mRNA, so that their transcriptional expression is linked. In bacteria, operons often (but not always) encode several genes that carry out a single biochemical pathway or otherwise related functions. An mRNA encoding more than one gene is still often called a *polycistronic transcript*, although the use of the term *cistron* as a synonym for gene (coined by Seymour Benzer in 1957) has otherwise almost entirely died out. An mRNA encoding only one gene might be referred to, similarly, as being *monocistronic*.

The genotype of an organism is a description of what genes and alleles it contains. The phenotype describes the measurable properties of that organism. The genotype determines the phenotype, but not all changes in the genotype will result in a measurable phenotypic change. More recently, it has also become clear that epigenetic differences in phenotype can exist **without** a corresponding change in the genotype. In bacteria, epigenetics is currently thought to depend mostly on methylation of specific DNA sequences, which changes how genes are expressed (see [Lecture 4](#) for more on regulation).

GENETIC NOMENCLATURE (IN BACTERIA)

For bacteria and archaea, there is a straightforward and fairly consistent system for naming genes and strains that was developed and popularized by Milislav Demerec, a geneticist who was director of the influential Cold Spring Harbor laboratory from 1941 to 1960. The details of this system were published in the journal [Genetics in 1966](#), and spread quickly through the bacterial genetics community. The examples I'll give here are mostly from *Escherichia coli*, the most common laboratory bacterium, but the same rules apply to all prokaryotic organisms.

To illustrate these rules, in the Materials and Methods section of a paper, you might find a table like the following:

Table 1.1. *E. coli* strains used in this study

Strain	Genotype
MG1655	F-, λ-, rph-1 ilvG
MJG238	F-, λ-, rph-1 ilvG Δppk gloA::cat ⁺

At first glance, of course, you may not get a lot out of that, but the information in this table is actually fairly straightforward, once you know the conventions.

Every strain of bacteria created or used in a lab is given a name, usually the initials of the primary investigator followed by a number. For example, strain [MG1655](#) was isolated by Mark Guyer in 1981, and was presumably the 1,655th strain he stored. MG1655 was one of the first bacterial strains to have its complete genome sequenced (in 1997, by Fred Blattner's lab at the University of Wisconsin at Madison), is a very-commonly used lab strain of *E. coli*, and is usually considered to be a *wild-type* strain. (Note that "wild-type" can mean anything from "a strain found in nature" to "any strain which doesn't have the mutation I'm interested in", depending on context. MG1655 is itself derived from an *E. coli* strain called K-12, which was used by Joshua and Esther Lederberg in their foundational studies on bacterial genetic exchange, and is the ancestor of most of the laboratory strains of *E. coli* used today.) MJG238 is a strain I constructed which is derived from MG1655. They are *isogenic strains*, meaning they are identical except for mutations in the genes listed.

The names of some strains of bacteria may include their serotype, which describes what antibodies will react with the surface molecules that strain. MG1655 is, for example, a serotype O:H48:K- strain of *E. coli*. Serotypes affect how the animal immune system responds to a bacterium, and for many species of bacteria strains with certain serotypes are more pathogenic than others. (For example, the Jack-in-the-Box strain of enterohemorrhagic *E. coli* is famously serotype O157:H7.) Each type of antigen is recognized by different sets of antibodies. The O-antigen of *E. coli* refers to lipopolysaccharide in the outer membrane ([Lecture 10](#)), the H-antigen refers to flagella ([Lecture 15](#)), and the K-antigen refers to capsule ([Lecture 14](#)). (The initials come from German and their use in this context is about 100 years old. "Hauch" means "film" and "ohne Hauch" means "without film", referring to the appearance of motile and non-motile strains on agar plates, and "Kapsel" means "capsule".)

A typical *E. coli* genome contains about 4000 genes, which is near the middle of the range for most types of bacteria. Streptomyces species can have more than 8000 protein-coding genes and members of the actinomycete genus *Kibdelosporangium*

can have more than 10,000, while simpler lactic acid bacteria often have fewer than 2000, and obligate intracellular pathogens like *Rickettsia* species have under 1000. Genes are given 3 or 4 letter names that are usually a mnemonic reflecting something about their function. For example, the *cbiA*, *cbiB*, and *cbiC* genes of *Salmonella* are three separate genes involved in cobinamide biosynthesis, and the *ppk* gene mentioned in the genotype of strain MJG238 encodes polyphosphate kinase.

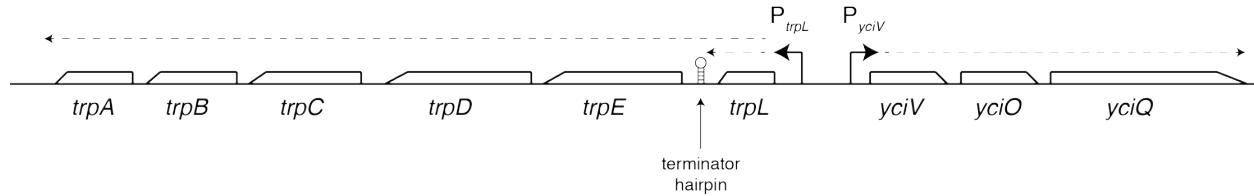


Figure 1.1. The *trp* locus of *E. coli*, to illustrate conventions of drawing genes and operons. The *trpA-E* genes are required for tryptophan synthesis. They and the tryptophan-rich leader peptide encoded by *trpL* are expressed as an operon from the P_{trpL} promoter. Two mRNA transcripts are possible from P_{trpL} : a short *trpL* transcript or, when lack of tryptophan leads to disruption of the terminator hairpin, a full-length 6-gene mRNA (this mechanism of regulation is called transcriptional attenuation, and we will revisit this in [Lecture 4](#)). The *yca* genes are divergently transcribed (that is, transcribed in the opposite direction) from the P_{ycaV} promoter as a 3-gene operon that has no role in tryptophan synthesis. Genes that are close together and transcribed in the same direction are often, but not always, cotranscribed in operons. The only way to tell for sure is to directly test whether they are encoded on the same mRNA. Operons often, but again, not always, contain genes involved in related biochemical functions.

One of the perks of figuring out what a gene does is that you get to choose its name, and there are no hard and fast rules about how to do this, although there are some common conventions. For example, it is very common for genes that encode regulators ([Lecture 4](#)) to be given names ending in the letter R or for genes in an operon that code for proteins involved in the same pathway to be given names with alphabetically-ordered fourth letters. It is generally a good idea to make sure a gene name is not already being used for something else, and to avoid names with alternate meanings. (The [fucose degradation operon](#) of *E. coli* is a somewhat notorious example.)

In *E. coli* and some other species, genes of unknown function (and there are still [many hundreds](#) of these) have been given names starting with the letter y (e.g. *ydjA* or *yeaG*, or the *yca* genes in Figure 1.1). The other letters indicate the location of that gene on the chromosome. These genes are likely to be renamed once their functions are determined. Note that genes of unknown function from different species with the same y-gene symbol may or may not be related in any way. For example, the *yneF* gene of *E. coli* is a putative cytoplasmic diguanylate cyclase, while the gene called *yneF* in *Bacillus subtilis* is an essential membrane protein whose activity is completely unknown. They have no homology to one another. This is another good reason to rename y-genes once their function is known.

Gene names are always written in italics, with the 3 letter first portion in lowercase. The fourth, capitalized letter (not always present, as seen for *ppk*) is used to differentiate genes that are all involved in the same pathway or phenotype. The proteins encoded by these genes would normally be capitalized, but not italicized: e.g. CbiA, CbiB, and CbiC, although in some cases, especially when the gene has only a 3-letter name, all three letters will be capitalized, as is the case for PPK. Systems of gene and protein naming in eukaryotes and viruses are different and vary among model organisms.

When looking at the genotype of a bacterium, the general rule is that any gene **not** mentioned is assumed to have the normal, wild-type sequence and encodes a functional gene product. Unless otherwise noted, it's assumed that any gene that **is** mentioned has **lost** function. In Table 1.1, *ilvG* indicates a *mutation* (or genetic change) destroying the function of the IlvG enzyme (acetolactate synthase, involved in isoleucine and valine synthesis). Some mutations, especially in *E. coli*, may also be given *allele numbers* (as in *rph-1*) to indicate that multiple mutations in those genes exist in different strains. You can look up the function of inactivated genes in genomic databases to determine what effect those mutations might have on the phenotype of the strain. For lab strains of *E. coli*, the [E. coli Genetic Resource Center](#) website allows you to search for specific mutations by gene or allele number and find information about that mutation and a list of publicly available strains that contain it, although it certainly does not contain every *E. coli* mutation that has ever been made. Other useful databases are discussed below. [Google Scholar](#) is particularly useful for tracking down the original references for genes with different allele numbers in the literature.

MJG238 contains two additional mutations, which illustrate additional conventions of genetic nomenclature. The Δppk allele (that's a delta, for those of you not up on your Greek letters) indicates a *deletion* of the *ppk* gene, in which the DNA sequence encoding *ppk* has been completely removed from the genome. In contrast, the *gloA::cat⁺* allele, while still indicating a loss of function of the *gloA* gene (which encodes glyoxalase I), shows (with the double-colon symbol) that it has been disrupted by *insertion* of additional sequence, in this case the *cat⁺* gene. The superscript + associated with the *cat* gene symbol indicates that it encodes a **functional** gene, in this case encoding chloramphenicol acetyltransferase, which makes this strain resistant to the antibiotic chloramphenicol. You may see mutations that combine these properties, e.g. $\Delta glpF786::kan^+$, which indicates a deletion of *glpF* with a kanamycin resistance cassette inserted in its place, and that this specific mutation has been assigned the allele number 786. Sometimes you will run across people using Δ to indicate any mutation destroying gene function, including insertions and point mutations. This is wrong (at least for bacteria). We will discuss types of mutations in much more detail in [Lecture 2](#).

The notations F⁻ and λ⁻ (that's a lambda) are *E. coli*-specific indicators.

The F plasmid (**F**ertility factor) is a large circular DNA element (about 100 kb) found in some natural *E. coli* isolates which is capable of transferring itself to other *E. coli* strains (by conjugation, which we will talk about in [Lecture 8](#)). F⁻ indicates that this strain does **not** contain the F plasmid. F' strains, rarely encountered today, have additional genes incorporated into the F plasmid. λ is a *lysogenic phage*, a virus which can integrate its genome into the chromosome of *E. coli* as a *prophage*. MG1655 has been cured of this viral genetic element, but many *E. coli* strains contain either wild-type λ or, more commonly among laboratory strains, replication-deficient derivatives of λ with important or useful genes inserted into them. Both λ and F and their ability to transfer genes between bacterial strains were discovered by Esther and Joshua Lederberg at the University of Wisconsin around 1950 and were fundamental to the development of bacterial molecular genetics.

Genome sequencing technology has resulted in a tremendous increase in the number of predicted bacterial genes, most of which have no functional information associated with them. To deal with the problem of how to refer consistently to genes from genome sequencing datasets, every predicted gene in a genome is assigned a *locus tag* which is unique to that gene in that specific strain. There are no established rules for how locus tags are formatted, and they are assigned by whatever research group sequenced and annotated the genome in question. For example, the *ppk* gene has the locus tag b2501 in *E. coli* MG1655, but the locus tag ESCCO14588_5033 in the pathogenic *E. coli* strain O157:H7 TW14588, even though these genes differ in DNA sequence by only 15 nucleotides and encode identical proteins. Locus tags are usually assigned in such a way that genes with consecutive numbers are next to one another in the genome, but even this is not always true, especially for small genes or noncoding RNAs that may have been missed in the initial annotation.

While locus tags are not as easy to understand as classical gene names, they are more specific and should be included whenever the identity of a particular gene needs to be established unambiguously. When publishing a paper, I would recommend mentioning the locus tag of a gene once (perhaps the first time it's referred to in the paper or in the Methods section) and then using an easier-to-remember gene symbol to refer to it in the rest of the paper.

For more details on the rules for writing bacterial genotypes, see the instructions on genetic nomenclature from the [Journal of Bacteriology](#). I will note, however, that the nomenclature for promoters is very inconsistent in the literature. I tend to prefer the format P_{gene} (which you can see in Figure 1.1), but you will see many variants, the most technically correct of which is geneP, so that the P_{trpL} promoter in Figure 1.1 should probably actually be annotated trpLp. (I find that aesthetically unpleasant and difficult to read, but c'est la vie.)

You can find the genotypes of many lab strains of *E. coli* on the [Open WetWare wiki](#). You'll note that most of them have many more mutations than MG1655. Bacterial strains are available to researchers from a variety of sources, including large stock centers. The most comprehensive are the [American Type Culture Collection](#) and the German [DSMZ collection](#), but others exist. [BEI Resources](#) is one that I have found useful in the past, for example. These collections maintain stocks of thousands of strains of bacteria and other organisms that have been deposited by researchers around the world. Many common and useful *E. coli* strains are commercially available from biotechnology companies like Novagen, Agilent, and ThermoFisher. Strains generated by individual labs can be requested directly from the labs that made them. Most researchers are happy to share published strains with their fellow academic scientists, and many granting agencies require that they do so. Nevertheless, you may have to do a fair amount of paperwork (called a *material transfer agreement*) to ship bacteria from one university to another, especially when pathogens are concerned.

THE SCIENTIFIC LITERATURE

The *scientific literature* is the summation of all published scientific knowledge. Knowing how to search and interact with it is a critical part of your scientific training. Before embarking on research on a biological system, it's wise to find out what is already known so that you don't waste time repeating experiments someone else has already done. Learning how to find that information is a critical skill. As my Master's thesis advisor once told me, "Six months at the bench can save you an hour at the library."

There are many sources of information about organisms, their genes, and the RNA and protein products of those genes. At the most fundamental level is the *primary literature*: research articles in peer-reviewed scientific journals. These are the basic product of laboratory research, and bacterial genetics papers will often (but not always) be focused on exploring the function of a single gene or protein in a particular organism. Most papers represent a year or more of work from between 2 and 10 scientists. In microbiology papers, the *first author* listed is generally the person who did most of the experimental work, while the last author (or *corresponding author*) is usually the head of the lab where the work was done. This order can vary from field to field. *Reviews* are articles written by experts, summarizing the current state of knowledge in a particular field and collating information from dozens or hundreds of research articles. They are often the best way of learning about a research topic that is new to you and are much more detailed and up to date than any textbook. Generally, the more recent the review, the better; at least to start. *Minireviews* are short reviews (a few pages), which usually either give a very brief introduction to or summarize the most recent developments in a specific topic.

You can find papers and reviews using specialized search engines, the most useful of which for biomedical research is [PubMed](#), provided by the National Center for Biotechnology Information. [Google Scholar](#) is also useful, especially since it allows you to search the full text of articles, and not just the title and *abstract* (which is a brief summary of the paper). Both these

tools let you set up keyword searches that will automatically send you any new references that fit whatever criteria you define. This is a good way to make sure you don't miss any papers on your specific research area.

Video supplement – [Performing a PubMed Search](#)

Video supplement – [Gene Searching on PubMed](#)

[PaperBLAST](#) is a remarkably useful website that allows you to enter a protein sequence and search for papers that contain data on homologs of that protein, although be cautious about whether those homologs are actually functionally related to your protein of interest. See below for a discussion of how BLAST and homology searching works.

It is important to note when searching and reading the literature that not all scientific journals are created equal. Some journals have higher quality standards than others. At one end of the spectrum are the *prestige journals*, which only publish what they consider to be the highest quality, most exciting, cutting edge, and influential results. These journals include Nature, Science, Cell, and the New England Journal of Medicine. At the other end are seemingly endless numbers of *predatory journals*, which have very low or no standards for what they will publish and are mostly just scams for separating naïve scientists from their money. In between are most of the journals in which quality research is published. Most scientific societies (like the American Society for Microbiology or the American Society for Biochemistry and Molecular Biology) publish society *journals*, which are typically not owned by for-profit publishing companies, have rigorous peer review, and are generally reliable, trustworthy publications. Some journals are very specialized (for example, Antimicrobial Agents and Chemotherapy), while others have broader scope (like Molecular Microbiology).

There are a number of metrics that purport to measure journal quality, none of which is perfect. The most common is *impact factor*, which is the number of citations received by articles published in that journal during the two preceding years, divided by the total number of articles published in that journal during that time. A higher impact factor indicates that the papers published in that journal have been cited more frequently, but of course, this only takes into account the last two years and can be skewed by a single highly-cited publication. Journals that publish a lot of reviews tend to have inflated impact factors for this reason. See [EIGENFACTOR.org](#) for an alternative, possibly more robust measure of journal quality, but there is no consensus on a systematic replacement for the impact factor. The consensus among most scientists is probably that these kinds of metrics are actively harmful to the practice of science. See [this paper](#) for an in-depth discussion of the problems they can cause.

While the quality of a paper is not necessarily linked to the quality of the journal in which it is published (however that is measured), higher quality journals will usually have more rigorous peer review and higher standards for publication and may therefore tend to publish more reliable work. Read carefully, and exercise good judgment. Do not automatically assume that something that's been published, even in the most prestigious journals, is necessarily correct.

One useful habit that will help you follow the literature outside of your own narrow research area is to subscribe to the *electronic table of contents* of several journals that publish research relevant to your interests. Those journals will then send you regular emails with the tables of contents for each issue, allowing you to quickly scan through the latest papers and keep up with your research community. As a microbiologist, good broad subject matter journals to follow might include mBio, Cell Host and Microbe, the Proceedings of the National Academy of Sciences, Nature Microbiology, the Journal of Bacteriology, Molecular Microbiology, and Applied and Environmental Microbiology, but you should subscribe to journals that regularly publish papers you are interested in reading. There are also prestigious journals dedicated solely to publishing reviews, which are tremendously useful. These include the Annual Review of Microbiology, Nature Reviews Microbiology, and Current Opinion in Microbiology, and can help you keep current on the most exciting and active research topics.

Preprints are manuscripts that have been made publicly available before they have gone through peer review and are an increasingly popular part of the publication process. [bioRxiv](#) is where most biology preprints are posted. You can read discussions of some of the possible benefits of preprint posting in [these papers](#).

DATABASES

A variety of databases exist which compile data from many individual research papers into a single searchable format, and this is usually the best way to find general information (such as sequence and predicted function) about specific genes or proteins. They also typically contain links to the primary literature, which will contain much more detailed information. A common workflow for obtaining information about a gene or protein would be to consult a database to get a general sense for what is known about that gene, and then drill down more specifically into reviews and individual papers to understand the details and biological context of what we know. The major databases listed below are regularly updated and, in some cases, manually curated to make sure that they contain high-quality data. However, most of the data in them is not exhaustively double-checked, simply due to the impossible amount of work that would be. Be aware of this, and that automated annotations may not necessarily be correct.

The largest of these databases is [GenBank](#), from the National Center for Biotechnology Information (NCBI), which contains all publicly available DNA sequences. A favorite of mine is the [Integrated Microbial Genomes](#) system, which contains all of the information obtained from the full genome sequences of over 140,000 organisms and more than 50,000 metagenomes

from different environments or bacterial communities (as of November 2024, but that number will rapidly become outdated as more become available).

Additional databases and resources that you may find useful include:

[EcoCyc](#): a very well curated repository of information on the model organism *E. coli*, combining large amounts of regularly-updated information from the literature for each gene and pathway in that organism, mostly for the K-12 strain MG1655. Most well-studied model organisms have similar dedicated databases. ([SubtiWiki](#) for *Bacillus subtilis*, for example.) [MetaCyc](#) automatically collates information for all organisms whose genomes have been sequenced, but of course there is generally much less information on genes and pathways in bacteria and archaea that are less well studied than *E. coli*, and much of the information about individual genes is computationally determined based on homology.

[UniProt](#): a comprehensive database of **protein** sequence and functional information. This can be very helpful and is a great database, but is not focused on genes and is not an especially good source for nucleotide information. This is most noticeable in the fact that searching UniProt for locus tags will often not give any hits. In this case, you should use IMG, GenBank, or another database to find the information you need.

[Bacterial and Viral Bioinformatics Resource Center \(BV-BRC\)](#): (combining the former PATRIC and IRD/ViPR databases) a very comprehensive database of bacterial and viral gene information, including genomes, transcriptomes, proteomes, pathways, systems biology, and phenotypic information (including antibiotic resistance), that is intended to be especially useful for those studying pathogenic bacteria. I have not used this much myself, but students in my lab have spoken well of PATRIC, and it has a lot of very powerful tools for genome comparison and analysis.

[RegulonDB](#): compiles the known information on how gene expression is controlled in *E. coli*. Much of this information can be found in EcoCyc, as well, but RegulonDB is organized in a different way that you may find helpful. A *regulon* is the entire set of genes or proteins controlled by a particular regulator (see [Lecture 4](#)).

[BioNumbers](#): a remarkably useful database that collects biological facts that are very difficult to find elsewhere. Do you want to know something like the volume of a *Bacillus subtilis* cell, the number of cells in a bacterial colony, or the concentration of ATP in *E. coli* grown on glucose? BioNumbers will give you the values and (critically) the references you need.

[KEGG](#), the Kyoto Encyclopedia of Genes and Genomes: the database for biochemical pathways in all organisms. KEGG contains a truly vast collection of information on genetics and physiology, with powerful tools for visualizing and comparing pathways in different organisms, although it is less user-friendly than some of the databases listed above, and not all of the data in KEGG is freely available.

[PDB](#): the Protein Data Bank contains three-dimensional structure information for proteins, mostly determined by X-ray crystallography or nuclear magnetic resonance spectroscopy. To visualize and manipulate the data in this database, you will need a specialized structure-viewing program, such as [ChimeraX](#), [PyMOL](#), or [CCP4](#). ChimeraX is particularly nice, since it is both free and cross-platform, so should work on any computer you have access to. The PDB contains only experimentally determined structures, in contrast to [AlphaFold](#), which attempts to computationally predict protein structures from amino acid sequences *in silico*, with decent, but not 100% accuracy.

[BRENDA](#): a comprehensive database of published biochemical information on enzymes. Useful if you want to know things like rate constants for enzymes, cofactor requirements, known inhibitors, and other *in vitro* properties of proteins.

[Microbiology Spectrum](#) and [EcoSal Plus](#): good collections of reviews summarizing all aspects of microbiology and of *E. coli* and *Salmonella*, respectively. They are regularly updated with new material, and if your institution has subscriptions (and UAB does) they are well worth consulting.

DISCUSSION PROBLEM SET #1: DATABASES AND LITERATURE SEARCHES

Use the above databases to answer the following questions and be prepared to discuss your results in class. If you have trouble finding any of the information, that would be a great thing for us to discuss as a group!

That is, in fact, the purpose of all of the discussion problem sets throughout this packet, so don't stress out if you find yourself stuck on something. The bulk of class time will be devoted to talking about and working through these problem sets as a group. You are absolutely welcome to work together or discuss the problems before class, if you want to.

1) What genes are involved in proline synthesis in *E. coli*?

- sketch the pathway of proline synthesis, indicating enzymes and intermediates (no chemical structures necessary)
- draw the operon or operons encoding the genes involved in this pathway
- give a citation for a review article with more information on proline synthesis

2) What is known about the YeaG protein from *E. coli*?

- draw the *yeaG* locus, indicating genes and operons near *yeaG* in the chromosome and their functions (if known)
- summarize briefly what is known about the function or activity of YeaG

- cite two papers from the primary literature that describe research on YeaG

3) What is the function of the gene with locus tag BT2158?

- what species / strain is this gene found in?
 - draw the BT2158 locus, indicating genes and operons near BT2158 in the chromosome and their functions (if known)
 - what is known about the function of this gene? (Hint: you may want to search for this gene in PubMed or Google Scholar, as its functional annotation has not, at time of writing, propagated to all the databases)
-

BLAST SEARCHES

Databases often allow you to search for genes or proteins by name, function, or by *homology*: how similar they are to other sequences (using a search algorithm called **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool)). Searching by homology is often the most useful, since gene names may not be used consistently and automated genome annotation may not necessarily assign the correct function to a gene (searching by locus tag avoids some, but not all of these problems). *Homologs* are genes that share a common ancestor and **may** have similar or related functions. *Orthologs* are homologs found in different species, and *paralogs* are homologs found in the genome of a single species. BLAST is the most common algorithm for identifying regions of similarity between sequences, and therefore for inferring homology. It compares nucleotide or protein sequences, identifies sequences that have significant similarity to each other, and calculates the statistical significance of those matches. BLAST is commonly used to identify members of gene families or to infer evolutionary or functional relationships between sequences.

The statistical significance of the results of a BLAST search are expressed as an *e-value*. Technically speaking, the BLAST e-value is the number of expected hits of similar quality score that could be found by chance in a given database. From a practical standpoint, the smaller the e-value, the more similar the sequences in question are, in much the same way that a smaller P-value in a t test or ANOVA indicates a more significant difference between samples.

The most common place to do BLAST searches is via the [BLAST page](#) at the National Center for Biotechnology Information. This will allow you to search nucleotide or protein sequences against GenBank. GenBank is an extremely large database and includes essentially all published sequence information. This can be problematic, especially if you are BLASTing a gene from an organism (like *E. coli*) for which there are many very similar or identical matches in the database. You can get around this particular problem by clicking the Exclude option in the Organism field and excluding *Escherichia* (or whichever genus you don't want to see results from).

For more focused searches of either single genomes or of specific taxa, it is possible to filter your BLAST search by organism, species, or other taxonomic group. Alternatively, you can use the [Integrated Microbial Genomes database](#), which contains only sequences from complete genomes and can filter searches in a variety of ways. The IMG database also has the advantage of providing (in my opinion) more user-friendly information about genes, gene neighborhoods, and pathways. You will need to create a free account to access the full capacity of this database (particularly BLAST searching against more than 25 genomes at a time). The Top IMG Homolog Hits pulldown menu at the bottom of each gene's page in this database is often exceptionally useful.

It is possible to filter BLAST search results in other useful ways (for example, returning one hit per species or eliminating sequences that are much shorter than your input sequence), but the web-based search platforms do not (at this time) provide for that, and you need to write your own bioinformatics scripts to accomplish these tasks. This is well beyond the scope of this class, and is best addressed by a course in bioinformatics, but I can recommend [BioPython](#) as a very accessible and flexible system for writing bioinformatics programs (see my own [BLASTER.py](#) for an example). Many professional bioinformaticians seem to [prefer R](#), a programming language that provides very powerful tools for statistical analysis. Bioinformatic analysis pipelines may very well contain multiple modules written in different programming languages.

There are alternatives to the BLAST algorithm for identifying homology between sequences. [HMMER](#) is one reasonably popular example that is designed to identify distant homologs with very high sensitivity.

UNDERSTANDING AND ANALYZING BLAST SEARCH OUTPUT

BLAST searches are a key element of almost every project in molecular genetics. The output of a BLAST search will be a list of sequences homologous to your input sequence. The most common format for nucleotide and protein sequences is FASTA format, which looks like this (for the *E. coli* transcription factor RcrR):

```
>646312216 NP_414839 transcriptional regulator, AraC family [Escherichia coli str. K-12 substr. MG1655 chromosome: NC_000913]
MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHALTQGAAKLEMPTEGEIFTLRPGNVVLLPQNSAHRLSHVDNESTCIVCGTLRLQHSARYF
LTSLPETLFLAPVNHSVEYNWLREAIPFLQQESRSAMPGVDALCSQICATFFTЛАVREWIAQVNTEKNILSLLHPRLGAVIQQMLEMPGHAWTVESLASI
AHMSRASFAQLFRDVSGTTPLAVLTKRLQIAAQMF5RETLPVVVIAESVGYASESSFHKAFFVREFGCTPGEYRERVRQLAP
```

The text on the line after the > can be any identifying information for the sequence, from a complex ID like the one above to a simple name or number. The following lines are the amino acid sequence of the protein (as in the above example) or nucleotide sequence of the gene.

A FASTA-formatted sequence file can contain any number of sequences in this format. Here, for example, are the nucleotide sequences for three *E. coli* genes involved in hydrogenase activity (note that a properly constructed file in FASTA format can contain protein, RNA, or DNA sequences, but not a combination thereof):

```
>hyfA
ATGAACCGCTTGTGGTGGCCGAACCACTGTGGTGTACAGGATGTAATACCTGTCGCTGCCTGTCGGACGTGCATAAAACGCAAGGTTACAGCAACA
CCCAGCCTGGCCCTGGCGAACAGCTAACAAATCACTGCCCTGTCGTGTCATCACTGTGAGGAAGGCCCTGCCTGCAGGTCTGCCGGTCAATGCCA
TCTCTCAGAGGGATGATGCGATCCAACCTAACGAAAGCCTGTATTGGCTGCAAGCTTGCGCCGTGGTGTGCCATTGGCGCAATCAGCCTCAGGA
AGCGTCCGGTGAATGCCCATGCGAACATGTTTCAGGCTGAAGGCTCACTCAAAGACGGCAAGAAAACCGCCAACACAAACATGCTTGCGCTG
GGAACCTGGTGTCCAGACCGTGCAGGTGAATGCGACCTGTGTGATTCTTGCCAGAAGGTCCGGCTGCCTGCCTGGCGTACGGC
TGATCACCGGTGATAGCCTGCAACGTCAGATGAAAGAAAACAGCGCCTGCGCAAGCTGGTTGCCAATGGCGGGAGGATCCCCCTTCCCTCACTCAG
GAGCAACGCTAA

>hyfC
ATGAGACAAACTCTTGCAGCGGATATCTGGCATTGGCTAGCACAGGCCGTGATTCTGTCATGCTAACCCCACCTTTACGGTATTCCGGCA
GATACCGCGCGTATGCACTCCCGCCGGCGGGATCGCAGGATTATCGCAGTATCCACAAACTGTTAACGCCAGGAAGTTGCCGACATCTT
CAGGTCTGATGTTCCGCTGATGCCGTGGTATTAAATCAGCAGCATGCTGGCTGGCGATGCCCTTACCACTGTTATTACCGTTCCCTTGGGGC
GGCGCGATCTGATCACCTTATCTATCTTGCCTGTTCTTGTCTTCCGGCTGGTACCGGAAGTCCGGTTGCGGAGTCGGTGC
CAGTCGCGAGTTGACGCTGCCATTCTGGCGAACATGCTTATCTCACTGCTGTATTGGCGCTGATAGCAGGTTCCACGCATATCGAGATGATCA
GCAATACGCTGGCGATGGCGTAACCACCGTACTGGCGTACTGGCGTGTGGCTTTCGCTTACCGGAAGTCCGGTTGCGGAGTCGGTGC
GATGTTGCTGAAGCAGAACAGGAATTACAGGAAGGCCGCTGACCGAACATTCCGGTGCCTGGCGCTAGCGAACATGGGAAACAGGT
CGTATGGCATACTGTTGGCCCTGTTCTGCCCTTGGCGCGCAAGAACATTCTCTGCCCTGCTGACTTCACTTGTGCTACGCTGCTCA
AGGTTTGTGATTTTGACTGGCTCAATCGCAGAAAACACGCTGGCACCGGGCTTTTACTCATTCACCATGTGACCTGGCTTGGCTTACGCTT
GCTGCGCTTGCATGGCTTCTGGTTAACCGGTGTAA

>hyfE
ATGACCGGTTCTATGATGTAATAATCTGGCGGACTGATGATGCTGACATCGCTGTTGTGATTAGCGTCAAAAGCTATCGCTGTCATGCCGATTTA
CGCTGCCAGTCAGTGTGCTGGTGTATTTCGCACTCTCGCTGCTGCTGCCAGAGCAACTGCTGATCTGGTCCGCCAGGCCATTATCACCA
AACTGCTGCTGGTACCGTTAATCATGACTACGCTGACGAAATATTCCCGAAAAGCGTATTGCGTCCGCCAGTGGACTCGCTTACCGATGGCTTACCGGGCTGAAACCGGCGCTGGCGTAGCGTTAGGTCAATT
TCTGCTTGGCCTGCTGTCAGCCAGCGAACATTCTCGCGAACATTGGTACTGCTGATGGAAAACGGCTCCCATCTGGTCTGGCGCTC
TTGCTGGCGAGCACCGAACCTGGTGGAAATAGGTATGCTACCGACGCATCTCGCGTCACTGTGATGGTTACTGGCAAGAAAATGGCTAC
CACGGCACGCTGGCGTGAACAACTGACCGCGTGAAGGGATAA
```

Most of the time, after using a BLAST search to identify homologs of your gene of interest, the next step in your analysis will be to generate an *alignment*, which allows you to visualize the regions of homology between the sequences and identify specific positions that are *conserved* between different sequences. Conserved regions are likely to represent the important functional parts of a gene or protein.

I find that for most purposes, amino acid alignments are the most informative, but in specific cases nucleotide alignments are appropriate. These include identifying the source of an unknown DNA sequence and most *phylogeny* experiments, which examine evolutionary relationships among genes and organisms (since there are three nucleotides per amino acid, DNA sequence contains more potential phylogenetic information). *Phylogenetic trees* can be very valuable for exploring alignments and analyzing the evolutionary relationships among genes, but the details of how they are calculated are beyond the scope of this class.

PAIRWISE ALIGNMENT

In some cases, you may be simply interested in calculating the homology between two sequences. This is a *pairwise alignment*. In this case, I typically use the Align Two or More Sequences option available in the [NCBI BLAST](#) tools. This option will allow you to enter two sequences and BLAST one (the query) against the other (the subject), giving you a sequence alignment and additional information including an e-value, a percent *identity* (how many positions are identical), and a percent *similarity* (for amino acids, how many positions contain residues with similar chemical properties). The output will look something like this, which is an alignment of the *E. coli* RrlR protein sequence above with a homologous sequence from *Klebsiella pneumoniae* (Expect is the e-value, in this case 2×10^{-36} , which is very significant and indicates that these two sequences are closely related to one another):

	Length:	284		
Score	Expect	Identities	Positives	Gaps
121 bits (303)	2e-36	92/301 (31%)	140/301 (46%)	27/301 (8%)
Query 1	MDSLHLLALLAPRCEVNLHCRFGRWQAGHQQMRSVVPWHVLRGEGRNV-GGQTHH			59
	MD+LS LL L AP+ ++ +C G WQ H V+ WH + +G +L + G+			
Sbjct 1	MDALSRLMLNAPQGTIDKNCVLGSDWQLPHGAGELS VIRWHALTQGAAKLEMPTEIFT			60
Query 60	LRAGDVVLPLHGSPHLMESLVEWGQVLPVAHRFNGTVTEMRAGPAEGALEMLCGEFYFGP			119
	LR G+VLLP S AHR + E ++CG			
Sbjct 61	LRPGNVVLLPQNS-----AHRLSHVDNESTC-----IVCGTLRLQH			96

```

Query 120 HVSW-LFSEASTLHLHTDAREDCPELDALLNIVRESLAQRPGGSAIVRSLGDTLLVLL 178
      + L S   TL   +   + L   + L +ES + PG A+ + T L
Sbjct  97 SARYFLTSLPETLFLAPVNHSVEYNWLREAIPFLQQESRSAMPGVDALCSQICATFFTLA 156

Query 179 LRMILLGEQQPPGGLLRLMSDERLMPAVLAVMATPEQPWTLESMAARAFLSRATFARHFAR 238
      +R   +   +L L+ RL   +   ++ P   WT+ES+A+ A +SRA+FA+ F
Sbjct 157 VREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAWTVESLASIAHMSRASFAQLFRD 216

Query 239 VYHLTPQAWLSQLRMALAARLLRLERQTNLEVIAERCGFQSLASF SKRFKMRYGVTPGEW 298
      V   TP A L++LR+ +AA++ E   + VIAE G+ S +SF K F   +G TPGE+
Sbjct 217 VSGTTPLAVLTKLRLQIAAQMFSSRE-TLPVVVIAESVGYASESSHFKAFVREFGCTPGEY 275

Query 299 R 299
      R
Sbjct 276 R 276

```

Notice that in this format amino acid residues identical in both proteins (*conserved residues*, identities, or *percent identity*) are shown with that amino acid letter in between the query and subject sequences and that chemically similar amino acids (*positives* or *percent similarity*; for example, lysine [K] and arginine [R] are both large and positively charged) are indicated with a + sign. Dashes indicate regions of sequence in one of the proteins that do not contain matching sequence in the other; so in this case, there are two regions in the query sequence (from *K. pneumoniae*) that are not found in the subject sequence (from *E. coli*). The more residues which are the same in two aligned sequences, the more closely related those sequences are considered to be. Residues that are more highly conserved (that is, the same in diverse, distantly-related homologs of the protein) are generally more likely to have important functions in the final protein product, since mutants lacking amino acids critical for protein function will be selected against by evolution.

DISCUSSION PROBLEM SET #2: BLAST & PAIRWISE ALIGNMENT

Use the tools linked above to answer the following questions and be prepared to discuss your results in class. (You should probably bring along a laptop so that you can easily share your results with the rest of the class and do additional analysis as necessary.)

For the genes with following locus tags:

- name the species this gene is from
- identify the predicted function of this gene
- align its protein sequence with that of its closest homolog from *E. coli* K-12 MG1655
- report the percent identity and percent similarity between the two proteins

- 1) aq_2095
- 2) SFK218_2554
- 3) USA300HOU_0506

MULTIPLE ALIGNMENT

For alignments of more than 2 sequences (*multiple alignments*), there are a variety of tools and algorithms available, many of the best of which can be found at the [European Bioinformatics Institute site](#). I often use [MUSCLE](#) for protein alignments, but [Clustal Omega](#) is also excellent. Use an alignment program appropriate for your particular sequences. A high-quality alignment is important for future analyses (especially for phylogenetic trees).

Alignment programs accept lists of homologous sequences (commonly in FASTA format) and can present the resulting alignments in a variety of formats. One useful one is the human-readable Clustal format:

CLUSTAL O(1.2.1) multiple sequence alignment

```

Escherichia      -----
Methylobacterium MAGPIRRAGAPETAGADDPLSGIAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
Proteus          -----
Bordetella        -----
Pseudomonas      -----
                  * * . * * . * * : . : * : *
Escherichia      LTQGAAKLEMPTGEIFTLRPGNVVLLPQNNSAHLRSHVDN-----
Methylobacterium VTRGRATLRRPGGAPLQVAAGDILLPRGDAHLFHGAG--PPPSTPLPVAVRHA--HDLRF
Proteus          VLSGQCVCYQIEKSAPIVLSEGTFMLNRRQSHTLWSGERDIEP--PPFLHKNNNGFLPVKY
Bordetella        LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLIDIEATTARRPVPAVVEAGRLPLRR
Pseudomonas      VMAGTCCHCEFLDGSRIDLHPGDILLLPRTGTPHLRSD---SPAPPCEPTVERQGSIPLYQ
                  : * . . . : * . : * : * : * :

```

In Clustal format, the punctuation under each block indicates conserved positions: an asterisk (*) indicates completely conserved residues, a colon (:) indicates very similar residues, and a period (.) indicates a lesser degree of conservation. The similarity is based on the chemical properties of the individual amino acids. This format is an excellent way to present alignments of 3 to perhaps as many as 10 sequences. It is often more visually appealing (for publication, for example) to copy the text into a word processing program and replace the punctuation indicating conservation with colored or shaded backgrounds, as shown here:

Multiple sequence alignment of the C-terminal domain of the RPS14 protein across six bacterial species. The alignment shows highly conserved regions (black bars) and variable regions (grey bars). The species listed vertically are Escherichia, Methylobacterium, Proteus, Bordetella, Pseudomonas, and Escherichia, Methylobacterium, Proteus, Bordetella, Pseudomonas.

You can also record the same alignment in a FASTA formatted alignment file, which is less human-readable, but more convenient for handling larger numbers of sequences:

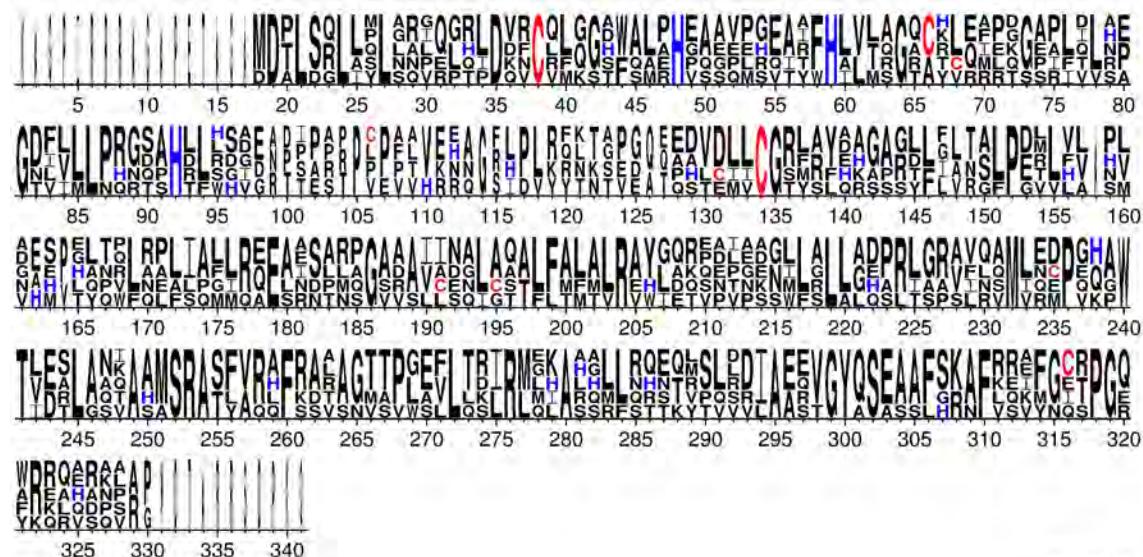
```

>Escherichia
-----MDALSRLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
LTQGAAKLEMPTEIFTLRPGNVVLLPQNSAHLRLSHVDN-----
-----ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKRLQIAAQMFRETLPVVVIAESVGY
ASESSFHKAFFVREFGCTPGYEYRERVRQLAP-----
>Methylobacterium
MAGPIRRAGAPETAGADDPLSGIAPLLRVRPHLDDVCRGGTWAAAHEAEPMRQAYFHL
VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG--PPPSTPLPVAVRHA--HDLRF
KTVGAEPDVELICGRLAFAEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLLAGIREEL
NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRGQPLTARAVLAMVRDPVHPW
TLDALAATAAASRASLVRRAFRAAGVAPLEFLTDLRLGLAHHLRLTETVSLDRLAEVGY
QSAAALSRAFLRKYGIWRPGQARQAEAPPAG-----
>Proteus
-----MDTLSQLLYLSQQLQLQDVFCQMKGHFSLPHVSSVEHETIFHL
VLSGQCYVQIEKSAPIVLQGDFMLNRQSHTLWSGERDIEP--PPFLHKNNGFPLPVKY
TKSEDQTQHVDDLLCGRMAYAKGSGLLLNGFPDMVVA-NLVEMPGILTBLNLFSQLRREA
INANQAAAILNLGAQTLFALRQVYQKPDINSSLWALLAEPRLSRVFNSMLNEPKGW
TLDLANVASMSRATFVRQFKATANTTPGEVLQSIKMLKALSLQQNKYTLSDIAERVGY
QSEAAFSKAFSKVFNCRPGQWKQQSKV-----
>Bordetella
-----MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEEAFHL
LLAGQCRLQARQGPALIINQGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRPLR
NTAPEQQADV DLLCGRFSYDRGAGDLFARS LPGV LH- H-LPQLQPLIAMLRAEA
ASPLPGAAAVINALGQALLALALRAYQREEEV PANMLAADS RIGPSVRAMIQDPGQAW
TIETLGNKAAMS R ATYARHFRS RAGMTVGEFLLRIRMMHAS ALLNHSQRS QRDIAEQVGY
QSEAAFGKAFREIMQTPGQWRRLLHRNARPVDTARRSDPKQ
>Pseudomonas
-----MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI
VMAGTCHECFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ
LNGPG--EALDMICGSYRYHAGSLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA
ESTRSGARSIIDLATALATLFA TL RAYLD RQPLGDGLFGLLDARLGRALQVMLECP EQAW
TVERLAQQAAMS RASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAETGY
QSEAAFSRNFRQAFGE SPGR FRRQADASR-----

```

In FASTA alignment format, each protein sequence is listed separately, with gaps indicated by a dash (-). As you can see, this does not provide an intuitive way to visualize sequence conservation, and you will need to use a separate alignment-drawing program to present the data. This is a good idea when you are aligning large numbers of sequences, where Clustal format becomes unwieldy.

[WebLogo](#) is a convenient online tool for visualizing conservation in large alignments. This program will accept any number of aligned sequences (in FASTA, Clustal, or many other formats), and will generate an image that represents the conserved residues in a very intuitive visual way, called a sequence logo. Take the alignment of RclR homologs from above, enter it into the WebLogo interface, and play with the different options to see what the program can do. Here's an example with 80 stacks per line, units of probability, scaled stack widths, no error bars, no y-axis labels, and a custom color scheme highlighting cysteine residues in red and histidine residues in blue:



WebLogo 3.4

In a sequence logo, the conservation of residues at each position is indicated by the height of the letters. For example, at position 38, all of the proteins in this alignment have a cysteine (C) residue, while at position 39, there are approximately equal chances of finding a glutamine (Q), leucine (L), arginine (R), or valine (V). You should adjust the parameters to give the most useful representation of your own data. WebLogo is a very versatile tool.

An alternative to presenting an alignment or logo is to report the consensus sequence of a set of sequences. This is a single sequence derived from an alignment by reporting only the most common residue at each point. The consensus sequence for the RclR alignment we've been working with is:

```
>RclR_consensus
-----MDLSXLXRXQRLDVRQLXGXWALPHEAAVPGEAXFHLVXAGQCXLXXPGAPLXXXGDFXLLPRGSAHLLXSXE--XPX-
PXPXXVEXAG-LPLRXXTPGX-EDVDLLCGRLXYXAGAXLXLTXILPXXLVXPXXESXXLTXLRPLIALLRXEAXSARPAAAXINALAQALFA
LALRAYXQRXXXXXXLALLXDPRGLGRXVQAMLEDPGXAWTXESLANXAAMSRAFVRAFRAXAGTTPXEXLTRIRMXKAXXLLRQEXXSLXDIA
EEVGYQSEAAFSKAFRRXFGCXPQWRQXRXXXX-----
```

Positions where the most common residue is no amino acid (gaps, or more accurately, positions where one or two sequences in the alignment have a small insertion) are indicated with a dash (-), and positions where no single amino acid is most abundant are indicated with an X. As you can see, this is generally less informative than showing an alignment, but it does take up less space, so may be useful in some situations. For nucleotide sequence alignments, N is used to indicate a position with no conserved or most abundant nucleotide. There are also single letter codes for combinations of nucleotides (e.g. Y = C or T), the complete list of which can be found at [this site](#).

DISCUSSION PROBLEM SET #3: BLAST & MULTIPLE ALIGNMENT

Use the tools linked above to answer the following questions and be prepared to discuss your results in class.

For the genes with following locus tags:

- name the species this gene is from
- identify the predicted function of this gene
- identify homologs of this gene from species belonging to 5 **different genera**

(Note that the more distantly related the homologs you choose, the easier it is likely to be to identify highly conserved regions of the protein. Why is that?)

- generate a multiple alignment with all 6 sequences (in whatever format you find most informative)
- based on your alignment, predict domains or specific amino acids that might be important for function of this protein

(A *domain* is a structural element of a protein, usually between 50 and 250 amino acids long, that folds independently and may carry out a specific kind of function. Many proteins are constructed of several domains, and evolution often builds new proteins by combining domains. From a sequence-gazing standpoint, a region of high homology could potentially be a conserved domain.)

- 1) RCAP_rcc03362
 - 2) USA300HOU_0588
 - 3) PGN_1123
-

LECTURE 2: MUTANTS AND MUTATIONS

INTRODUCTION

In this lecture, we will discuss how bacterial geneticists use mutants and mutations to decipher how biological systems work. We will define different types of mutations and spend considerable time discussing how to interpret mutant phenotypes. We will also begin to explore how observations can lead to models and hypotheses, in the first steps of applying the scientific method to solving biological problems.

SCIENTIFIC PROCESS I: OBSERVATIONS AND PHENOMENA

Every scientific study begins with an *observation*. The scientist looks at the world around them and sees a *phenomenon* that they think might be important or interesting. The key feature of phenomena is that they can be reliably and objectively measured, and therefore represent some real aspect of the physical world. For the purposes of scientific terminology, observation is a noun, not a verb, and refers specifically to the results of a measurement you can make of a phenomenon.

Reproducibility is central to the value of scientific observations. If a phenomenon is representative of something real, then it should be observable by different people in different places whenever the appropriate conditions occur. From a practical standpoint as a scientist, detailed record-keeping and recording of your observations is absolutely central. Only then do your observations rise to the level of being *data*.

The quality of your observations is, in many ways, directly dependent on the tools and instruments you have available. In the history of microbiology, the invention of ever better microscopes (starting with Antonie van Leeuwenhoek and Robert Hooke, followed by many others over the course of the last 400 years) allowed scientists to directly observe the existence of living things too small to be seen by the naked eye. Robert Koch's invention of solid growth media for bacteria and methods for isolating pure cultures in the 1880s made it possible to distinguish and separate different types of microbes from one another, leading directly to observations of specific bacteria and their relationship with particular diseases or environments. Advances in DNA sequencing technology are another example of the same process of technological improvement leading to new kinds of observations.

(As an aside, I will note that it was Angelina Hesse, an assistant in Robert Koch's lab, who introduced the use of agar to solidify growth media. Before that, the Koch lab isolated bacteria on gelatin-solidified media or slices of boiled potato, both of which have distinct disadvantages.)

In this class, our focus is on using genetics, the science of how heritable characteristics are passed from one organism to another, to understand how bacteria function on a molecular level. We will therefore be making observations of how the biochemical and physiological behavior of bacteria is affected by changes in the content and expression of their genes.

When I ask you to describe a set of observations that you plan to make, you should explain:

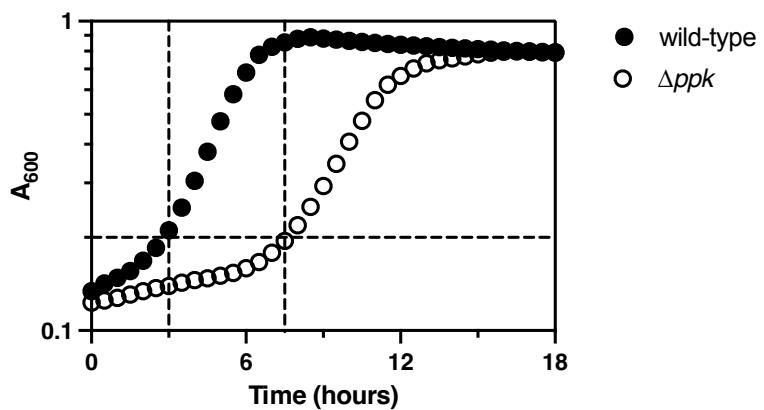
- What will you be measuring, and how will you measure it?
- When and how often will you measure it?
- Is it a *qualitative* or a *quantitative* measurement?

Quantitative measurements result in numerical data, while qualitative measurements result in categorical or descriptive data. "Growth" or "no growth" is qualitative. "Blue" is qualitative, but "absorbance of 0.87 at a wavelength of 595 nm" is quantitative. Beware of assigning numerical values to categorical measurements and then treating them as quantitative. Occasionally, *binning* quantitative measurements to treat them as categories can be useful: colony diameter is quantitative, "greater than 2 mm" versus "less than 2 mm" is qualitative. Both types of measurement are potentially useful, but they need to be treated differently in statistical analyses.

EXAMPLE OF DESCRIBING A SET OF OBSERVATIONS: GROWTH CURVES

Bacterial growth curves have a characteristic shape, with a lag phase in which no growth occurs, a log phase in which cells are growing exponentially, and a stationary phase in which the cells have exhausted one or more resources in the media and have stopped growing again.

Suppose you perform an experiment where you compare the growth of two *E. coli* strains, with the following results:



How can we describe these results?

- 1) Both strains grew (a qualitative measurement).
- 2) The wild-type grew better than the Δppk mutant (also a qualitative measurement).
- 3) The wild-type had a 3 hour lag phase, while the Δppk mutant had a 7.5 hour lag phase (defining lag phase as ending when $A_{600} > 0.2$) (a quantitative measurement, indicated by the dashed lines).
- 4) The wild-type had a faster growth rate (the slope of the curve in exponential phase) than the Δppk mutant (a quantitative measurement).
- 5) Both strains reached the same final cell density after 18 hours (a quantitative measurement, albeit not a very useful one).

Can you think of any others? Each of these descriptions is a potentially valid way to report this result, but some are more rigorous than others. Option 3 defines lag phase in a very arbitrary way, for example. Option 2 does not permit statistical comparison between the two strains.

Since bacterial growth curves can have rather complex shapes, unbiased statistical comparisons between them are not trivial. Several software packages do exist aimed at solving this problem ([AMiGA](#), for example).

THE GENETIC TOOLKIT

At a very simple level, molecular genetics techniques do one of two things:

- 1) move new DNA into a cell or
- 2) change the sequence of the DNA a cell already has.

There are a wide variety of ways to do each of these things, and the methods that allow you to accomplish them in a particular species are referred to as the *genetic toolkit* for that organism. Some species have more fully developed toolkits than others, and this determines what kinds of experiments are possible in each species. In **Lectures 2 and 3**, we'll talk about how and why we can change or remove a cell's genes, and in **Lectures 5 - 9** we'll discuss different ways of moving new DNA into cells, as well as *homologous recombination*, a mechanism which we can use to incorporate new DNA into a cell's chromosomes.

USEFULNESS OF MUTANTS IN BIOLOGICAL EXPERIMENTS

Any change in the genetic material of an organism is a *mutation*, and the resulting organism is a *mutant*. As noted in the last chapter, mutants are interpreted relative to their wild-type *parent strain*, although the definition of wild-type is somewhat arbitrary.

Mutations are the geneticist's best and most fundamental tool for understanding biological systems. We isolate mutants to understand what changes in a cell's genotype affect the phenotype we are studying. This allows us to narrow down the tremendous complexity of cells and focus on only the genes, alleles, and loci that directly influence our particular study system. If a mutation affects our phenotype of interest, it tells us something about how that phenotype works. As several senior microbiologists have expressed it to me, "Let the cells tell you what's important." (I've tracked this phrasing back, anecdotally, to Bruce Ames, a pioneer of *Salmonella* genetics and inventor of the Ames test for measuring how mutagenic chemicals are.) The process of identifying mutations that impact a phenotype of interest is also called *forward genetics*, as opposed to

reverse genetics, which identifies phenotypes resulting from a particular mutation. (See **Lecture 8** for methods we can use to make specific mutations in bacterial genomes).

An analogy I have found useful for explaining the use of mutants in biology is to imagine that you have no idea how automobiles work, and the only resources you have available to figure it out are a hammer and an infinite supply of Volkswagen Beetles. The geneticist's strategy to solve this problem is to break one thing in each car with the hammer and see what happens. If you break the spark plugs, that car won't run, but the headlights will work (at least for a while). If you break the battery, that car won't run **and** the headlights won't work, telling you that the engine depends on both the spark plugs and the battery, but the headlights only require the battery. The hammer is making "mutations", and by interpreting the "phenotype", we are able to piece together how a complex system functions and how the different components are interrelated.

INTERPRETING MUTANT PHENOTYPES

We extract meaning from mutants by examining the phenotypes that result from genetic changes. If we isolate several different mutants that have mutations in different genes, but have similar phenotypes, we can reasonably conclude that those genes are all involved in that phenotype. We might, for example, identify several different mutations in *Vibrio cholerae* that fail to secrete cholera toxin, and are therefore unable to cause disease. Some of these might be genes encoding the toxin protein itself, while others could be important for transport, processing, or regulation. However, since they all have a toxin-minus phenotype, we can conclude that they all must work in concert in the cell to carry out the toxin production process.

DISCUSSION PROBLEM SET #4: BACTERIAL PHENOTYPES

The key feature of a useful mutant is that it has an observably different phenotype than the wild-type. Mutations can change any of the phenotypes we can measure and are our primary tool for interrogating biological functions.

What kinds of phenotypes can we measure for bacteria? List as many as you can think of, indicating whether they are quantitative or qualitative.

There are many technical terms that are used to describe phenotypes. An *auxotroph* is a strain that requires a particular nutrient. This is contrasted with a *prototroph*, which does not require that nutrient. A mutant defective in histidine synthesis would be a histidine auxotroph, for example. (This might typically be written as being His⁻, pronounced "hiss-minus".) A strain that grows more slowly than the wild-type in the absence of a particular nutrient is called a *bradytroph*, although this is a much less commonly used term, and is less likely to be understood by even a specialist audience.

Phenotypes can be "strong", "weak", or "leaky", terms that are not strictly defined, but generally express how easy the phenotype is to observe. If your mutant dies under conditions where the wild-type grows well, that is a strong phenotype. If the difference is a more subtle one in growth rate, for example, that might be referred to as a weak phenotype. A complete lack of histidine synthesis would be a strong phenotype, while a partial lack, with some histidine still being made, would be a leaky phenotype. In this example, you might hypothesize that genes in which mutations result in strong His⁻ phenotypes might be directly involved in the biochemical pathway for histidine synthesis, while those with leakier phenotypes might play roles in regulating the activity of the pathway or reduce the activity of enzymes without eliminating it completely. (The difference between a weak and a leaky phenotype is pretty arbitrary, as you can see.)

Mutations that have several apparently unrelated phenotypic effects are said to have a *pleiotropic phenotype*. This often occurs with mutations in genes for *global regulators* (see **Lecture 4**) or in genes with roles in central cellular functions or stress responses (RNA polymerase or protein folding chaperones, for example).

Does every change in genotype cause a phenotype? I would answer this question with a cautious "no", since many changes in a bacterium's DNA sequence do not cause an obvious change in their appearance or growth. However, this is very much dependent on the growth conditions and on exactly what you are measuring. A mutant defective for uracil synthesis will not appear to have a phenotype until you try to grow it on media containing no uracil. A mutant that cannot make flagella forms colonies perfectly well on plates, and only when you look at it through the microscope in liquid culture will you find that it cannot swim. With a mutation that appears to have no phenotype, you may simply have not yet found the appropriate conditions to see the effect, so be cautious in your interpretations. It is also worth remembering that for bacteria we are often limited to relatively crude measures of gene function like growth rate or survival. Mutations in highly conserved genes which have dramatic effects on multicellular eukaryotes, where developmental problems are very easy to see and can be caused by very subtle biochemical changes, may have no **visible** effect on the growth of a bacterial culture.

See the end of this chapter for more on the use of hypotheses and models in bacterial genetics and how we use mutant phenotypes to develop and test ideas about biological functions.

KINDS OF MUTANTS

There are many kinds of mutations that differ by exactly what sort of change occurs in an organism's genome sequence. *Point mutations* are changes of a single nucleotide in the DNA (sometimes called a *single nucleotide polymorphism* or SNP). *Transitions* are point mutations in which a purine (A or G) is mutated to the other purine or a pyrimidine (C or T) is mutated to the other pyrimidine. *Transversions* are point mutations from a purine to a pyrimidine or vice versa. *Missense mutations* are point mutations in a protein coding sequence that change the amino acid encoded at that point in the gene to a different amino acid. (See [this site](#) for a detailed resource describing the consequences this can have.) *Nonsense mutations* are point mutations that change an amino acid-encoding codon to a stop codon (TAA, TAG, or TGA), terminating translation and resulting in a truncated protein product.

Table 2.1. Types of point mutations

transition	purine (AG) to purine or pyrimidine (CT) to pyrimidine
transversion	purine to pyrimidine or pyrimidine to purine
missense	amino acid-encoding codon to different amino acid-encoding codon
nonsense	amino acid-encoding codon to stop codon
silent	amino acid-encoding codon to a different codon encoding the same amino acid

Silent mutations are point mutations that, due to the degeneracy of the amino acid code (the fact that more than one codon can encode the same amino acid), do not change the amino acid encoded by that codon. However, because some codons are more efficiently translated than others, silent mutations **can** sometimes affect protein expression or even protein folding, since they change the rate of translation. This is because not all tRNAs are present at the same concentration in cells. All species have some *rare codons* which are infrequently used and therefore are translated by a smaller pool of available tRNA. The impact of supposedly "silent" mutations can be remarkably large, as quantified in [this 2024 paper](#).

Insertions and *deletions* are the addition or subtraction of nucleotides into the chromosome. *Frameshift mutations* are small insertions or deletions of a number of nucleotides not divisible by 3, which disrupts translation of the gene downstream of the frameshift. Frameshifts result in scrambled and often truncated proteins. *Duplications* are mutations in which a region of DNA sequence is duplicated (resulting in 2 or more copies of that region). *Inversions* and *rearrangements* are large-scale changes in the structure of the chromosome, in which substantial regions of DNA are either reversed or moved relative to their position in the wild-type.

It is important to note that *null mutations*, in which the product of a mutated gene loses activity (also called *gene knockouts*), are always more common than *gain-of-function mutations*, where a new activity is generated, but all types of mutation can result in the addition of new functions under certain circumstances. Gain-of-function point mutations are often especially informative when trying to understand how a particular gene works. (There are many ways to break something, but usually only a few ways to make something work **better**.)

Large insertions, especially of DNA from a different organism, are the most likely mutations to add new functions, since they may consist of whole new genes. When this happens naturally during evolution it is called *horizontal gene transfer*. Note that, in prokaryotic evolution, new genes are acquired by horizontal gene transfer much more frequently (at least 50-fold) than they arise by the duplication and divergence of existing genes in the genome. This is the opposite of the situation in eukaryotes, where horizontal gene transfer is much less common.

Some mutations are *lethal* and will result in a cell that can no longer grow. Like gain- or loss-of-function, this is a property of the phenotype, not the genotype *per se*. You will not be able to study lethal mutations in the lab without specialized methods. Lethal mutations could include null mutations of *essential genes* or gain-of-function mutations creating toxic effects. Exactly which genes in a bacterium are essential depends heavily on the rest of that organism's genome, and even closely related strains may have different sets of genes that cannot be knocked out.

Many mutations result in *conditional phenotypes*, in which phenotypes (often lethality or failure to grow) are only observed under some conditions. A useful but perhaps non-obvious example of this are *temperature-sensitive mutants*, which contain gene products that are destabilized and therefore do not function at high temperatures.

POLARITY

Because bacterial genes are often found in operons, with more than one protein encoded on a single mRNA molecule, mutations in one gene in an operon can affect expression of *downstream genes* in that operon. This effect is called *polarity*, and can complicate interpretation of mutant phenotypes, since a null mutation in one gene can also prevent expression of several other genes. Large insertions, which can contain entire genes, many stop codons, transcriptional terminators, etc. are especially polar, and commonly completely prevent expression of downstream genes in an operon (Figure 2.1).

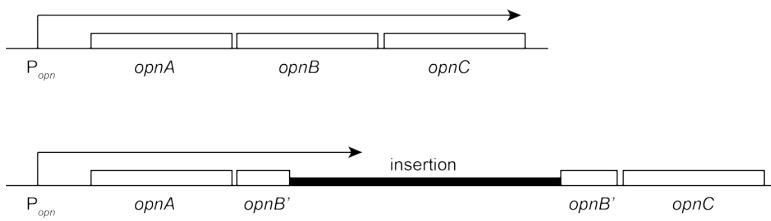


Figure 2.1. Polarity of large insertions. In this example, the large insertion in $opnB$ inactivates that gene but also contains a transcriptional terminator which prevents expression of $opnC$.

Some types of point mutations are also polar, and nonsense mutations and frameshifts are much more likely to have polar effects than other types. To understand this, it helps to understand the mechanism by which these kinds of polar effects occur.

Normally, when bacterial RNA polymerase is transcribing an operon (shown in Figures 2.2 – 2.4 as $orf1$ – $orf2$ – $orf3$), the mRNA produced is coated in ribosomes actively translating that mRNA into protein. (Transcription and translation are linked in many bacteria, although [recent discoveries](#) suggest that this linkage is not as tight in the Gram-positive model organism *Bacillus subtilis* as it is in *E. coli*.)

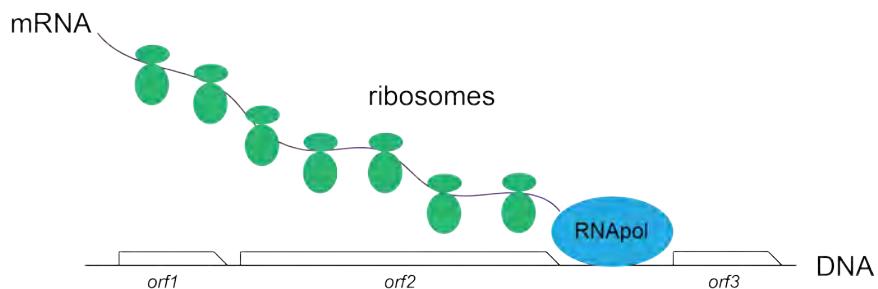


Figure 2.2. Transcription and translation are linked in many bacteria. As an mRNA is being synthesized by RNA polymerase (RNAPol), it is immediately recognized by ribosomes, which begin translating it into protein before the mRNA is finished being transcribed. Those ribosomes protect the mRNA from degradation and from premature transcript termination.

When a mutation prematurely stops translation of a gene in that operon, RNA polymerase continues, producing a stretch of mRNA with no ribosomes on it (until it reaches the ribosome binding site for the next gene; see [Lecture 4](#) for more on ribosome binding sites):

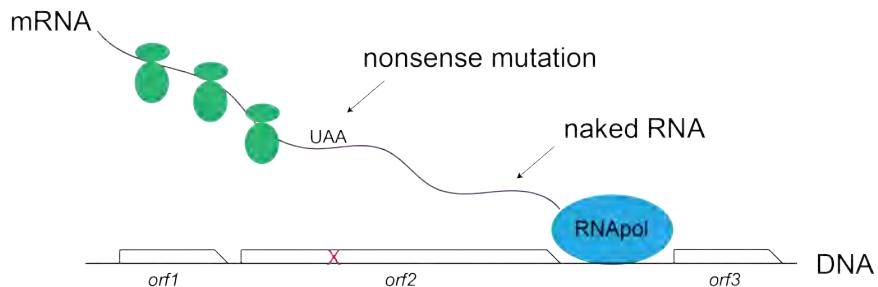


Figure 2.3. Polarity can result from premature stop codons, which result in stretches of RNA unbound by ribosomes which is susceptible to degradation and premature termination.

Many bacteria contain a homolog of a protein called Rho, which forms a complex with actively transcribing RNA polymerase. Rho is normally involved in transcription termination, and recognizes stretches of untranslated RNA with no ribosomes bound. Rho catalyzes changes in the structure of RNA polymerase and causes it to disengage with and fall off the DNA. This is normally a mechanism to ensure that the cell doesn't spend a lot of energy transcribing non-coding RNA after stop codons, it can also cause point mutants to be polar in the same way that insertions can be. Note that $orf3$ will never be transcribed, despite the point mutation being in $orf2$.

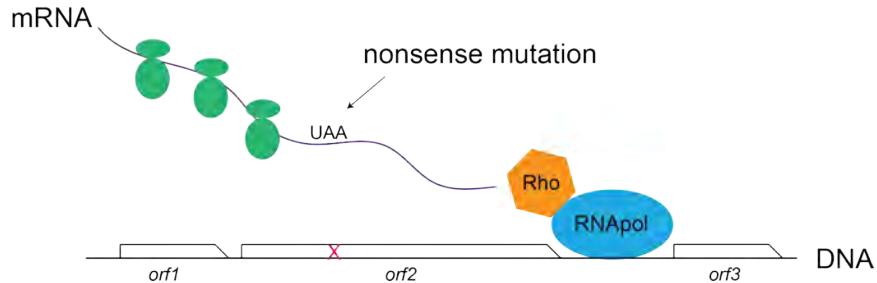


Figure 2.4. Rho recognizes mRNAs that are not covered with ribosomes and interacts with RNAPol to terminate transcription.

A related phenomenon that can lead to mutations in one gene affecting the expression of other genes, even if they're **not** cotranscribed in an operon, is the fact that it is reasonably common for bacterial promoters or regulatory elements ([Lecture 4](#)) to be present inside the coding sequence of adjacent genes, so that mutations in one gene will disrupt the expression of adjacent genes. Since promoter sequences are hard to predict computationally, it can be hard to tell ahead of time when this might be the case for any particular gene locus.

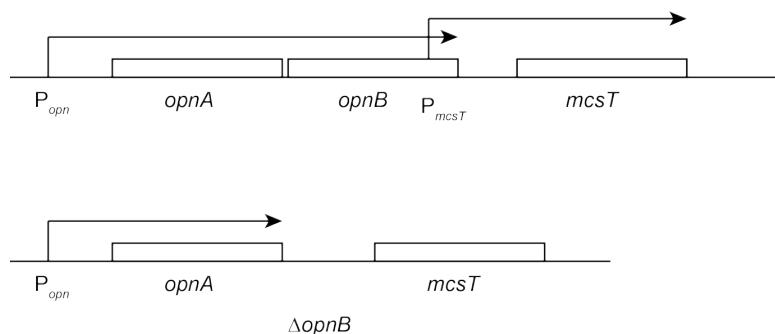


Figure 2.5. Polar effects due to internal promoters. In this example, the promoter driving expression of the monocistronic gene *mcsT* is within the coding sequence of the *opnB* gene. A *opnB* deletion therefore not only eliminates *opnB*, but also prevents expression of *mcsT*.

There are also plenty of examples of overlapping open reading frames in bacteria, where, for example, the start codon of one gene includes one of the same nucleotides as the stop codon of the upstream gene. (e.g. ...**CCCTAA**TGGCT..., where the end of one gene is in bold and the beginning of the second is underlined). Completely deleting the first gene will eliminate the start codon of the second, preventing its expression.

In cases where you suspect that polarity may play a role in the phenotype of a mutant, there are a variety of molecular biology techniques which are used to help clarify which gene in an operon is actually responsible for a given phenotype. The most common is *complementation*, which is a major topic of [Lecture 5](#).

MUTAGENESIS

Mutagenesis is the process of making mutations in an organism. There are many ways to do this, and the technique you use will have important effects on what kinds of mutants you can expect to find. In this chapter we will focus on random mutagenesis. *Random mutagenesis*, as the name implies, creates mutations at random positions within a DNA molecule, and is contrasted with *site-directed* or *targeted mutagenesis*, where you make a specific mutation exactly where you want it. (We will discuss methods for site-directed mutagenesis in [Lectures 7 and 8](#).)

The major advantages of random mutagenesis are that you do not need to know in advance what genes or amino acids are important for a phenomenon of interest, and in the case of random point mutations, there is the potential to find gain-of-function mutations, which are often extremely informative. Depending on what kind of mutagenesis you want to achieve, you may have a variety of tools available, and those form an important part of the genetic toolkit for your model organism.

Because DNA replication is not perfect, *spontaneous mutagenesis* will occur in any population of growing bacteria, and is, of course, one of the underlying processes behind evolution. This is the simplest way to generate mutants in the lab, but since the various different kinds of mutations occur at low frequencies, it may take a very large number of cells or long period of time to identify the mutations you are looking for. As a rough estimate, in *E. coli*, any given single base pair change will occur in about 1 in every 10^8 cells (a frequency of 10^{-8}), while spontaneous null mutation of any given gene (a combination of the rates of all possible types of gene-inactivating mutations) occurs in about 1 in every 10^5 cells, although these numbers can vary depending on the region of the chromosome involved.

Table 2.2 lists the rough frequency to be expected for different kinds of spontaneous mutations, determined mostly from the rates at which they occur in *E. coli* under normal growth conditions. We will discuss recombination, plasmids, and transposons in future lectures.

Table 2.2. Approximate Spontaneous Mutation Frequencies (from Gary Roberts, University of Wisconsin – Madison)

- loss of gene function: 10^{-5}
- any particular single base point mutation: 10^{-8}
- reversion of a frameshift, missense, or nonsense mutation: $10^{-6} – 10^{-8}$
- spontaneous deletions: $10^{-3} – 10^{-10}$ (depends on the region to be deleted)
- duplication of a given region: 10^{-3}
- loss of a duplication: $10^{-1} – 10^{-2}$
- loss of various constructed plasmids: $10^{-2} – 10^{-5}$
- loss of most natural plasmids: $< 10^{-8}$
- precise excision of a transposon: $10^{-6} – 10^{-9}$
- site-specific recombination events: $10^{-1} – 10^{-2}$

The frequency at which mutations yielding a particular phenotype arise can be informative, even if you don't yet know the exact identity of those mutations. For example, if the phenotype you are measuring arises spontaneously in 1 in every 10^5 cells, you can reasonably hypothesize that it could be caused by a loss-of-function mutation in a single gene. If it only occurs once in every 10^8 cells, then your working hypothesis might be that it is caused by a specific point mutation. If obtaining a particular phenotype requires two independent mutations, the frequency of observing that phenotype will be the product of the frequencies of each individual mutation. A phenotype that requires two gene-inactivating knockout mutations would therefore occur spontaneously at a frequency of about $10^{-5} \times 10^{-5} = 10^{-10}$, so once in every 10 billion cells. We will discuss ways of increasing the rate of random mutations in [Lecture 3](#).

The rates listed above are rough estimates. Mutation rate is driven by a combination of the error rate of DNA polymerase and the activity of DNA repair systems. In many organisms, there are also mutational “hot spots”, genes or loci that mutate at much higher rates than the rest of the genome. There are different mechanisms by which this can occur; for example [diversity-generating retroelements](#), which use error-prone reverse transcriptase to rapidly mutate targeted genomic loci. The point is that these are evolvable traits, some organisms mutate more or less rapidly than others, and there is growing evidence that bacteria can regulate their mutation rates in response to environmental conditions. The frequencies in Table 2.2 are only averages.

Spontaneous mutations can result in *genetic drift* as mutations accumulate in bacterial strains. My lab's version of *E. coli* strain MG1655 is not 100% identical to the version whose genome was sequenced (it has two missense mutations in protein-coding genes and three small insertions in intergenic regions). If you have applied some selective pressure to your strains, providing conditions where certain mutants grow better than the parent strain, mutations will be much more likely to become dominant in your populations. For example, many lab strains of *E. coli* are not very motile, but if they are incubated overnight without shaking, this selects for mutants with increased motility (presumably because motility allows bacteria to swim upwards and get better access to oxygen). As many as 25% of the cells in a non-shaken overnight culture of the poorly motile *E. coli* lab strain BW25113 will have mutations that increase their motility, as described in [this paper](#). We will talk more about selections in [Lecture 3](#).

SCIENTIFIC PROCESS 2: MODELS AND HYPOTHESES

There is more to science than simply recording observations. That's just list making, and a list of, for example, 75 mutations that cause a particular phenotype is not useful in and of itself. You must use that information to try to advance your understanding of how the world functions.

Once you have made a set of observations, you can propose a *model* to explain them. A useful model will not only propose a **mechanism** to explain the observations that have already been made, but even more importantly, will make predictions about what might be observed in the future. A model that can make accurate predictions about the world (has *predictive power*) is both more useful and more likely to be correct than a model without such power. Models are always incomplete descriptions of the actual way the world functions (the map is not the territory). Even a model with significant inaccuracies may have some predictive power.

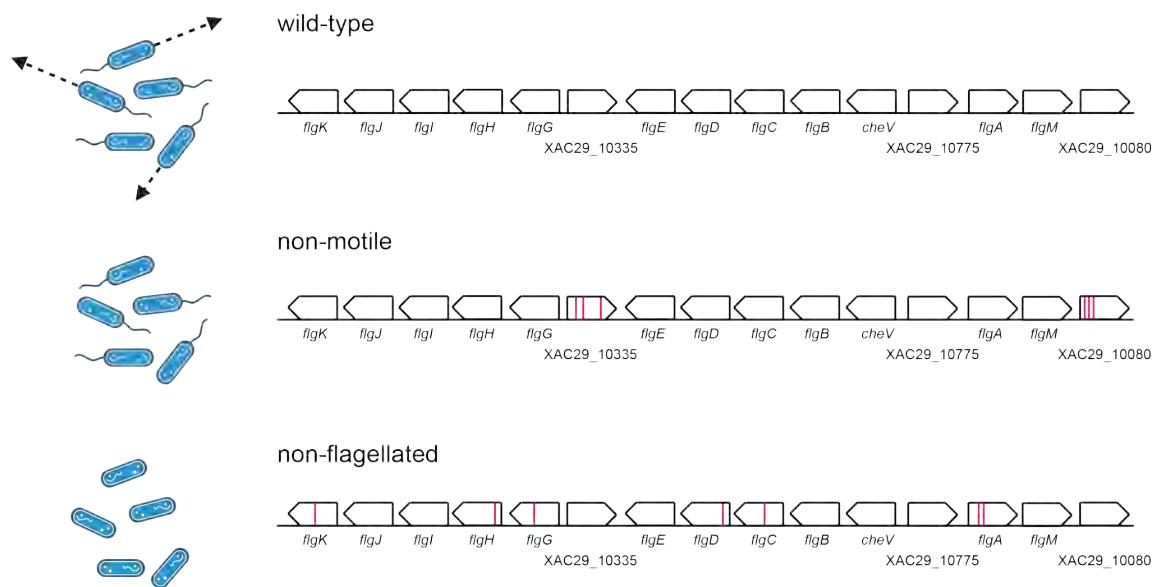
When I ask you to propose a model, it should:

- incorporate all of the available data
- propose a mechanism that explains the behavior of the system
- make testable predictions about the system being studied

EXAMPLE OF PROPOSING A MODEL: MOTILITY

You isolate a set of spontaneous mutants of the plant pathogen *Xanthomonas axonopodis* with defects in motility (which we will discuss in detail in [Lecture 15](#)).

Some of the mutants have flagella, but cannot swim, while others do not have flagella at all. By genome sequencing, you find that many of the mutations in non-motile strains are in one particular genetic locus.



Location of point mutations in the *X. axonopodis* *flg* locus that result in the indicated phenotypes, indicated with red lines. Each individual mutant strain you isolated has only one mutation. Made with Biorender.

Propose a model to explain these data.

Start by gathering data and making sure you know as much as possible about the system.

- 1) Looking up the predicted functions of the *flg* genes (in the Integrated Microbial Genomes database), you find that they are homologous to structural components of the flagellum in other bacteria.
- 2) XAC29_10335 and XAC29_10080 are genes of unknown function, although XAC29_10080 is a member of the histidine kinase family of regulatory proteins.

Based on this information, and on the phenotypes of the mutants you isolated, the following model is reasonable:

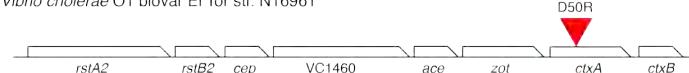
- 1) Mutations in *flgK*, *flgH*, *flgG*, *flgD*, *flgC*, and *flgA* interfere with or delete structural components of the flagellum, resulting in non-flagellated cells.
- 2) XAC29_10335 and XAC29_10080 are regulators of flagellar activity, and mutations in these genes prevent flagellar rotation, but not assembly.

DISCUSSION PROBLEM SET #5: PROPOSING MODELS BASED ON DATA

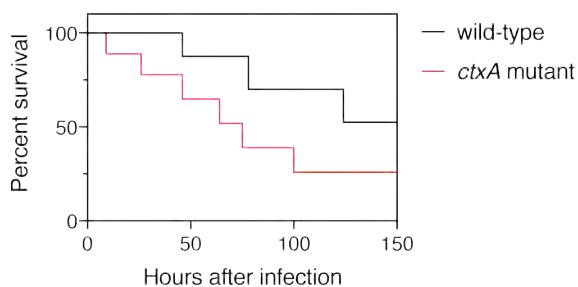
A key skill in science is looking at data and developing models to explain those data. This requires creativity, open-mindedness, and humility (most of your models will end up being wrong, no matter how beautiful or elegant they are), but is the first step in applying the scientific process to solving problems.

Problem #1

The following figure shows an operon from the cholera pathogen *Vibrio cholerae*. You have isolated a strain with the indicated *ctxA* **missense** mutation and compared the survival of mice infected with this mutant and the wild-type strain.



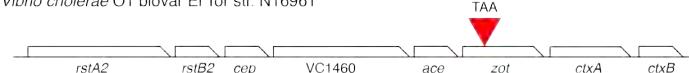
Mice infected with bacteria:



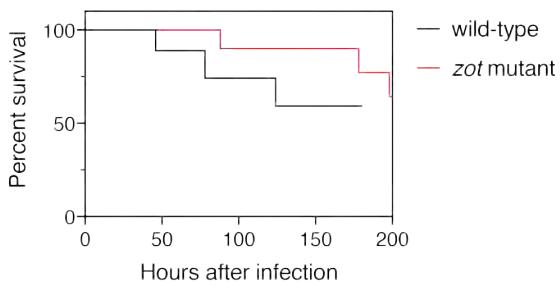
Given these data, propose a model to explain the observed result. (It may be helpful to look up the function of CtxA.) Remember that a model should contain a proposed **mechanism**.

Problem #2

You have isolated a second mutant containing a **nonsense** mutation in the *rstA2* operon, sequenced it, and tested its phenotype. The results are as follows:



Mice infected with bacteria:



Given these data, propose **two** distinct models that could explain the observed result.

The predictions made by a model are *hypotheses*, and testing whether those predictions are accurate or not is a fundamental part of the scientific process. If a model predicts that you will observe X under a particular set of conditions, but you actually observe Y, then the model is wrong and must be changed to include the new information.

To be useful, a hypothesis **must** be *falsifiable*, and therefore testable. "All diseases are caused by bacteria" is a valid hypothesis, since it can be disproved by observing even one case of disease caused by something other than bacteria. "Bacteria are put into sick people by cruel invisible fairies" is **not** a valid hypothesis, because there is no set of observations which could disprove it.

When I ask you to propose a hypothesis:

- it should be falsifiable
- you should be able to propose a set of observations that can be used to test it (generally, using methods we have covered in class)

A final criterion for a good hypothesis is that it should be **interesting**, or at least add something new to the model. A hypothesis that just reiterates what we already know is both boring and useless. Testing such a hypothesis is a waste of time.

In any scientific study, observations lead to models, which lead to hypotheses. Testing hypotheses leads to more observations, the results of which are used to modify the model and improve its predictive power. In this way, science moves ever closer to an understanding of how reality works.

In simple cases, like those we'll be talking about in this class, models and hypotheses are often very similar. Models, however, become more and more complex as they incorporate more data and become able to explain more observations. Hypotheses should always be as simple and straightforward as possible.

Hypotheses do not need to test everything about a model, and in fact, generally only test one aspect of it. A good model will lead to many hypotheses.

EXAMPLE OF PROPOSING A HYPOTHESIS: MOTILITY AGAIN

What is a testable hypothesis based on the model proposed above?



- 1) Mutations in *flgK*, *flgH*, *flgG*, *flgD*, *flgC*, and *flgA* interfere with or delete structural components of the flagellum, resulting in non-flagellated cells.
- 2) XAC29_10335 and XAC29_10080 are regulators of flagellar activity, and mutations in these genes prevent flagellar rotation, but not assembly.

There are several possibilities, but we should choose one which tests an **interesting** part of the model. The idea that mutations in flagellar structural gene homologs prevents flagellar assembly is easily testable, but boring. The genes of unknown function are much more interesting candidates for study.

We know that XAC29_10080 is homologous to a family of transcriptional regulators (more on this in [Lecture 4](#)), so we can use that information to propose a hypothesis:

- XAC29_10080 is required to activate transcription of genes encoding the motor proteins that drive flagellar rotation.

We can test this hypothesis by comparing the transcription of genes encoding motor proteins in the wild-type and XAC29_10080 mutant strains. If they are different, that supports the hypothesis and we can add it to our model. If they are the same, we can reject this hypothesis and come up a new one.

Nothing is known at all about the XAC29_10335 gene, which in some ways makes this more difficult, since there are so many more possibilities for how it could be functioning. However, the following hypotheses are all reasonable and testable:

- XAC29_10335 is required to activate transcription of genes encoding the motor proteins that drive flagellar rotation.
- XAC29_10335 modifies the motor proteins that drive flagellar rotation and affects their activity.
- XAC29_10335 is a component of the flagellum that is essential for rotation, but not assembly.

Testing each of these hypotheses requires a different set of experiments, of course, and this is by no means an exhaustive list of the possible explanations for the observed phenotype. I didn't say what kind of mutations (nonsense, missense, frameshift, etc.) these are, which might be relevant information as well. If all of the mutations are null mutants (premature stop codons and the like), that might lead to a different model / hypotheses than if they were all missense mutations causing more subtle protein changes.

DISCUSSION PROBLEM SET #6: PROPOSING HYPOTHESES TO TEST MODELS

Problem #1

E. coli colonies expressing β -glucuronidase (encoded by the *gusA* gene), are blue on plates containing the indicator compound X-Gluc. Wild-type colonies of *E. coli* MG1655 (whose genome **does** encode *gusA*) are white, suggesting that the *gusA* gene is not usually expressed in this strain.

You spread MG1655 on plates containing X-Gluc, and are able to isolate spontaneous blue colony mutants.



Created with Biorender:

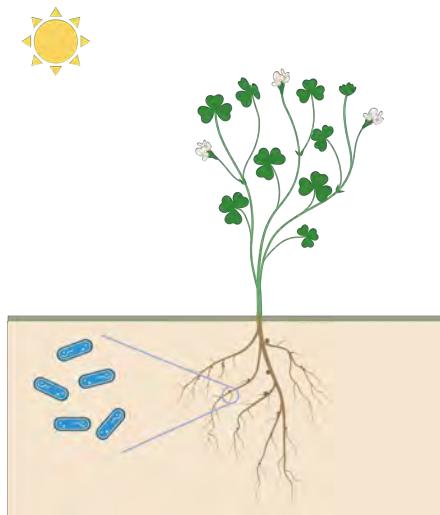
By making serial dilutions and spread plating, you are able to determine the frequency at which blue *gusA*-expressing mutants arise in this strain.

Propose a model and testable hypothesis to explain each of the following possible results:

- blue colonies appear at a frequency of 1 in 10^3 colonies
- blue colonies appear at a frequency of 1 in 10^5 colonies
- blue colonies appear at a frequency of 1 in 10^8 colonies

Problem #2

Sinorhizobium meliloti is a plant symbiont that forms nitrogen-fixing nodules on the roots of alfalfa plants.

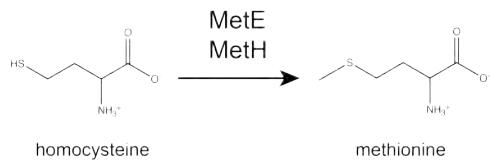


Created with Biorender:

When studying amino acid metabolism in *S. meliloti*, you find the results in the table below.

Strain	Ability to form nodules	Ability to grow on minimal media
wild-type	yes	yes
$\Delta metE$	yes	yes
$\Delta metH$	no	yes
$\Delta metE \Delta metH$	no	no

MetE and MetH are *isozymes*, non-homologous enzymes that catalyze the same reaction, in this case, the last step of methionine synthesis:



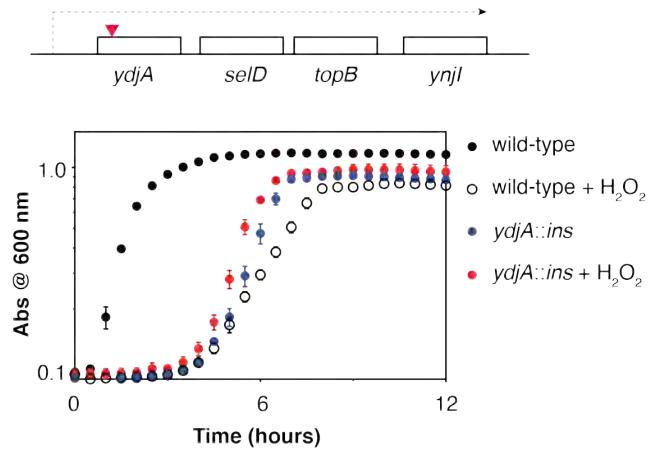
Minimal media are growth media that contain only the compounds that a given organism absolutely requires for growth (as opposed to *rich media*, which contain lots of nutrients). In this case, it indicates media with no amino acids added.

- Propose a model to explain these data.
- State one testable hypothesis derived from that model.
- Propose one observation you could make to test your hypothesis.

Problem #3

While studying hydrogen peroxide (H_2O_2) resistance in *E. coli*, you isolate a H_2O_2 -resistant strain with a mutation in the *ydjA* gene.

Further analysis reveals that *ydjA* is in an operon with several other genes, that the mutation (indicated with a red triangle in the figure below) is a 30 base pair **insertion**, and that the mutant has the following growth phenotype:



- Propose a model to explain these data.
- State one testable hypothesis derived from that model.
- Propose one observation you could make to test your hypothesis.

LECTURE 3: MUTANT HUNTS AND EXPERIMENTAL DESIGN

INTRODUCTION

In this lecture, we will go into more depth about the scientific method, discussing the difference between observations and experiments, and exploring the rules and principles of designing good experiments. We will define screens and selections, the two fundamental genetic techniques for finding interesting mutants, and practice devising mutant hunts for different applications. Finally, we will explore different methods for actively mutating bacteria and discuss more advanced types of mutant analysis. We will also talk about alternative approaches and troubleshooting, emphasizing the importance of creativity and rigor for scientific problem solving.

SCIENTIFIC PROCESS 3: EXPERIMENTS, VARIABLES, AND CONTROLS

Observations are basically passive, making measurements of what occurs naturally in a system. To more aggressively test hypotheses, scientists actively manipulate the systems they are studying to see if the effects of those manipulations fit the predictions made by their models. Such a manipulation is called an *experiment*. A well-crafted experiment is a tremendously powerful way to make discoveries about the physical world, but it is important to understand what makes a **good** experiment.

In any experiment, the experimenter changes one or more *independent variables* (or *treatments*) and observes the effect(s) that these changes have on one or more *dependent variables*. It is usually best to have only **one** independent variable in an experiment, since this makes interpreting the effects on the dependent variable(s) much simpler. Remember: the independent variable is what you **change**, the dependent variables are what you **measure**.

When designing an experiment to test a hypothesis, you must consider the following:

- **Will it answer the question?** Is your experiment capable of falsifying your hypothesis? Will the results of the experiment actually test the predictions of your model? Is it possible to learn anything from a result that is different from what you expect? Are there alternate explanations that could lead to the result you predict?
- **Is it possible?** How difficult will it be to carry out your proposed experiment with the resources you have available? What tools will you use to make your manipulations and measurements?
- **Is it elegant?** Some problems can be solved by *brute-force* approaches that simply test all the possible combinations of factors in a system. This can be effective, but is tedious and often expensive. It is often possible and preferable to test hypotheses with simpler, more creative experiments.

The best experiments are those for which **any** possible outcome gives you new information about the system you are studying and lets you improve your model for how it works. This is not always possible, but is definitely something to strive for.

Pilot experiments are preliminary tests, usually done in a relatively quick and inexpensive way, to see whether a new idea or procedure is worth pursuing further. To use an artistic metaphor, they are like sketches done before a real painting. It's especially important to do pilot experiments before embarking on any really labor-intensive or expensive experiment, so that you don't waste a lot of time and energy on something that will not give you meaningful results.

Experiments always need to have *controls*. Controls are experimental treatments with known outcomes, which allow the experimenter to be certain that their experimental setup is working as intended. *Negative controls* are treatments expected to result in **no change** in the dependent variable, while *positive controls* are treatments that **are** expected to result in such a change. Negative controls are important for ensuring that no contamination or other problems are interfering with measurements to give *false positive results* (called a type 1 error in statistical jargon). Positive controls demonstrate that the measurement system is capable of observing the expected changes in the dependent variable(s), and rule out the possibility of *false negative results* (a type 2 error). If an experiment's controls don't work as expected, then you cannot interpret the results and need to stop and figure out what has gone wrong with your experimental system.

Note that controls are about what you are measuring (the dependent variables), and are not determined by your hypothesis. If you expect nothing to happen or change, it's a **negative** control. If you expect something to happen or change, it's a **positive** control.

When I ask you to design an experiment for this class, you should explicitly:

- define the dependent and independent variables
- explain what you will measure and how
- describe both positive and negative controls
- describe the possible outcomes of the experiment and what they would mean for your hypothesis

One valuable resource you have available to you at UAB is the Center for Clinical and Translational Science [Biostatistics, Epidemiology, and Research Design](#) unit, a team of statisticians whose job is to help UAB scientists design and interpret ex-

periments. Most institutions have similar resources. I encourage you to consult with them as part of your actual research projects, ideally in the experimental **design** phase.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of." - Ronald Fisher (1938)

EXAMPLE OF AN EXPERIMENTAL DESIGN: CARBON SOURCES

Different bacteria are able to use different types of carbohydrates, amino acids, or other chemical compounds as sole carbon sources for growth. The two most abundant carbohydrate polymers on earth are cellulose and chitin, which make up the cell walls of plants and fungi, respectively.



Created with Biorender:

Biotechnology and bioenergy firms are very interested in isolating bacteria that are able to grow on these carbon sources, since genetic engineering of those bacteria might allow us to design bioreactors to turn waste material into valuable chemicals or fuels.

Both cellulose and chitin are abundant in soil, which leads to the following testable hypothesis and experimental design:

Hypothesis: Soil contains bacteria which can use cellulose as a sole carbon source.

Experimental Design:

- 1) Make bacterial growth media in which cellulose is the only available carbon source.
- 2) Spread soil samples on that medium and incubate.
- 3) Isolate bacterial colonies that grow under these conditions.

Independent Variable: (what will you change?)

- The bacteria added to the plate (mainly by changing the source of the soil samples)

Dependent Variable(s): (what will you measure?)

- Ability to form colonies on cellulose growth media (a qualitative measurement)

Negative Control(s): (eliminate false positive results)

- 1) Spread a pure culture of a non-cellulose utilizing soil bacterium on cellulose medium (e.g. *Streptomyces coelicolor*). This should not grow, and will test that the plates contain no other carbon sources.
- 2) Spread autoclaved soil on the plates. Nothing should grow. This will help you distinguish between tiny colonies and bits of dirt.
- 3) Incubate a plate with nothing spread on it. Nothing should grow. This will make sure you don't have any contamination in your media.

Positive Control(s): (eliminate false negative results)

- Spread a pure culture of *Cellvibrio japonicus* (a known cellulose-utilizing bacterium) on your media. This should grow, and will confirm that your plates work as expected.

Potential Outcomes:

- 1) You isolate bacteria that can use cellulose as a sole carbon source. This supports your hypothesis.
- 2) You are unable to isolate bacteria that can use cellulose as a sole carbon source. This result would cause you to reject your hypothesis, especially if your positive control works as expected, but might also mean that any cellulose-utilizing bacteria in your soil samples require other growth factors you haven't included in your media.

In the previous chapter, we discussed using mutants to understand biological phenomena. In this chapter, we will explore this in more depth, beginning to look at the design of experiments in bacterial genetics, and actively manipulating bacterial genomes to test hypotheses.

FINDING INTERESTING MUTANTS

All kinds of mutations occur spontaneously, but not every mutation is interesting. We use forward genetics to ask questions about specific phenomena or phenotypes, which means we need to have methods for identifying mutations that have effects relevant to those phenomena. This kind of experiment is called a *mutant hunt*, and what you're hunting for is mutants that can help answer a specific biological question.

The two broad categories of mutant hunts are *selections* and *screens* (Figure 3.1), and the choice of which to use depends on the phenotype we are interested in measuring.

If there are conditions under which mutants we are interested in will grow but the wild-type will not, then we can select for those mutants. Selections are extremely powerful and allow the isolation of very rare mutations. Since as many as 10^8 or 10^9 cells can be spread on a single agar plate, and only mutant cells will survive to form colonies, it is technically very simple to separate mutants from wild-type with a selection. Whenever possible, you should design mutant hunts as selections, since they will give you better results for much less work. However, it is not always possible to design a selection for your desired mutations, and in those cases, you will need to perform a screen.

Screens are used to isolate mutants that are different from wild-type in a non-selectable way (color, motility, toxin production, etc.) or mutants that die under conditions where the wild-type survives. In either case, the key feature of a screen is that the phenotype of each cell or colony must be examined **individually** to determine if it is an interesting mutant, and even in a best-case scenario no more than about 100-1000 colonies can be screened on a single plate (screens are also commonly now done in liquid media in 96- or 384-well microtiter plates). This means it is rarely practical to screen for mutations that occur at a rate of less than about 1 in 10^5 cells without sophisticated automation. Even with a very expensive robotic setup, screens of more than a few hundred thousand mutants or conditions are usually impractical, although I have seen some [flow cytometry-based screens](#) that can be scaled up very effectively.

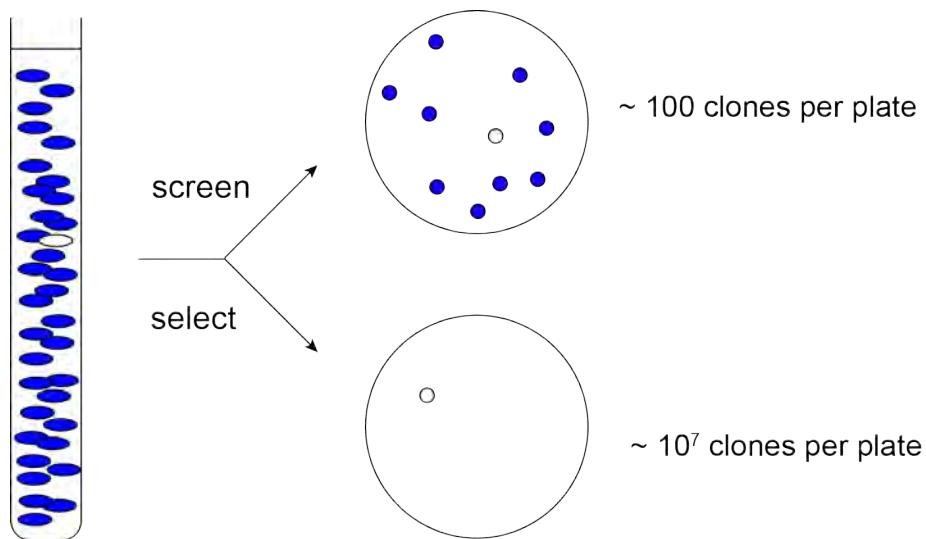


Figure 3.1: Selections allow you to identify much rarer mutants (white cells or colonies in this diagram) than is possible with a screen. A typical bacterial culture contains 10^8 – 10^9 cells per milliliter, and it is possible to spread about 100 microliters of culture on the surface of an agar plate.

An *enrichment* is somewhere between a selection and a screen. For a selection to work, you need conditions where the mutants are alive and the wild-type is dead (or at least does not grow). If you have conditions where the wild-type grows, but the mutants you're interested in grow **faster**, then, over time, the mutants will slowly become a larger and larger proportion of the population, thereby enriching the population with interesting mutants. Similarly, I have seen experiments that enriched for mutants with enhanced motility by allowing a population of bacteria to swim outwards from the point of inoculation in soft-agar plates and harvesting only the outermost cells (presumably those that swam the fastest).

You typically follow up an enrichment (or several cycles of enrichment) with a screen to isolate individual mutant strains. This can greatly reduce the number of colonies that you need to screen to find mutants of interest.

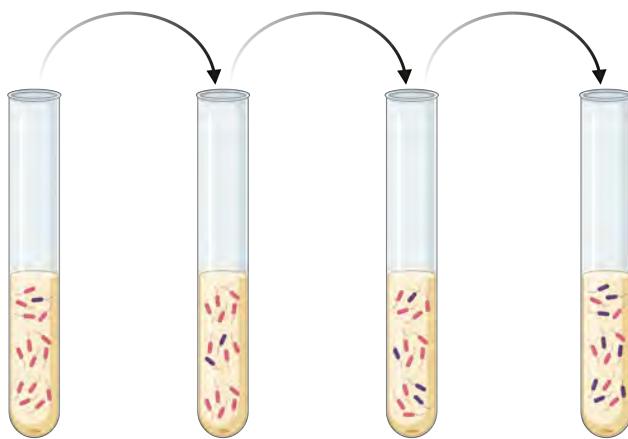


Figure 3.2: Enrichments allow you to increase the proportion of interesting mutants in a population. In this example, purple mutant bacteria grow faster than the red wild-type bacteria under these conditions. Serial passages allow the purple mutants to successively outcompete the wild-type, making them easier to isolate. Created with Biorender.

It can sometimes be challenging to design a mutant hunt that will successfully isolate mutations relevant to a particular biological question. This is where you will need to think creatively about the model you are testing. If your model is correct, what kinds of mutant phenotypes might be possible? Which kind of mutagenesis is most likely to result in interesting and informative changes in the phenotype? We will practice this kind of creative problem solving in class throughout the next several lectures.

I want to end this section by emphasizing one last practical point about looking for mutants: **you get what you select for.** Even if that's not what you think you're selecting for! When you design a mutant hunt to try to identify mutations involved in a particular process, you will have some ideas in mind about what might result in the phenotype you're looking for. Biology is complicated, though, and there may be alternative ways to achieve such a phenotype. Sometimes this is interesting and useful, and leads to discovering unexpected connections between genes, but sometimes it just means you need to think more carefully about your selection conditions.

ARTIFICIAL MUTAGENESIS

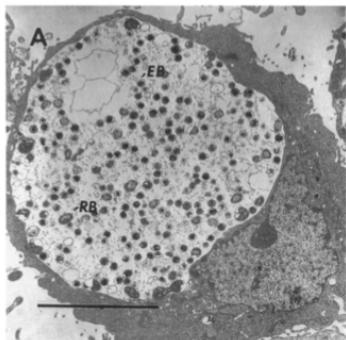
If spontaneous mutagenesis does not give you high enough *mutation rates* to isolate the mutants you are interested in, you can treat bacterial cells with *mutagens* that cause DNA damage and increase the rate at which mutations accumulate. *Chemical mutagens* are toxins that react with DNA and *radiation* (including UV light) delivers energy directly to the DNA. Different mutagens cause different kinds of mutations. For example, UV light primarily causes G:C to A:T transitions, while the chemical mutagen acridine orange preferentially causes frameshifts.

From a practical standpoint, always be extremely careful using mutagens in the lab. You will be adding very toxic chemicals to your bacterial cultures or exposing them to high-energy radiation, either of which will mutate your DNA just as efficiently as they mutate bacterial DNA, and mutations in your cells may increase the chances of those cells becoming cancerous.

Both mutagens and spontaneous mutagenesis have the disadvantage that it can be difficult to locate where exactly in the genome a mutation causing an interesting phenotype actually is, although this has become quite a bit easier with the advent of inexpensive genome sequencing technology. (A complete bacterial genome sequence now costs about \$100. [SeqCenter](#) in Pittsburgh is the company my lab uses for this.) The main limitation now is that mutagens tend to cause many simultaneous mutations across the genome, and it can be difficult to know which one or ones are causing a particular phenotype.

EXAMPLE OF AN EXPERIMENTAL DESIGN: CHEMICAL MUTAGENESIS

Chlamydia trachomatis is an obligate intracellular pathogen that causes a variety of diseases, most notably genital chlamydia, although it is also the leading infectious cause of blindness. It forms inclusions in mammalian cells by modifying host cell responses, retargeting host vesicles, and blocking the induction of apoptosis.



An electron micrograph of *C. trachomatis* cells growing in an inclusion in a mammalian cell. (From [this paper](#).)

C. trachomatis has a very small genome (about 900 genes), but can only replicate inside mammalian cells and is very resistant to transformation with external DNA. This makes the usual methods of bacterial molecular genetics (which we will discuss in much more detail later in this course) much more challenging to implement.

Researchers at Duke University have developed [methods](#) to make *C. trachomatis* more tractable for genetic experiments. Among these is the construction of a library of 923 genome-sequenced strains of chemically-mutated *C. trachomatis*, each of which contains 7 – 25 randomly distributed point mutations in its genome (including a wide range of missense, nonsense, and silent mutations). They have used this library to screen for genes involved in specific elements of pathogenesis in experiments similar to this one:

Hypothesis: Specific genes in *C. trachomatis* encode virulence factors necessary for preventing host cell apoptosis.

Experimental Design:

- 1) Screen each of the 923 mutated *C. trachomatis* strains for the ability to prevent apoptosis in infected mammalian cells. (This seems like a lot, but these cell-based assays can be done in 96-well plates, and it's only 10 of those.)
- 2) Compare the mutations present in mutant strains that are defective in blocking apoptosis.

Independent Variable: (what will you change?)

- The assortment of point mutations present in each *C. trachomatis* strain.

Dependent Variable(s): (what will you measure?)

- How many infected host cells undergo apoptosis (a quantitative measurement).

Negative Control(s): (eliminate false positive results)

- Uninfected mammalian cells, which should not undergo apoptosis.

Positive Control(s): (eliminate false negative results)

- Mammalian cells infected with wild-type *C. trachomatis*, which should undergo apoptosis.

It will be important to include both the positive and negative control conditions on each 96-well plate, to make sure that there aren't problems in any one particular subset of the experiment.

Potential Outcomes:

- 1) One or more of the mutants is defective in blocking apoptosis.

Since each mutant has multiple point mutations, this will indicate that in each strain identified in the screen, **one** of those point mutations is in a gene important for blocking apoptosis. However, each strain has different mutations, so if more than one mutant with a defect is identified, you may be able to identify common genes which are mutated in both strains. The more mutations in a particular gene that you identify in strains with apoptosis-blocking phenotypes, the more likely it is that that gene is linked to that phenotype.

- 2) None of the mutants are defective in blocking apoptosis. This may mean that no mutations are present in the library in genes important for this phenotype, that there are multiple redundant genes involved, or that the gene(s) involved in blocking apoptosis are required for *C. trachomatis* growth and cannot be knocked out.

A common and practical way to make random gene-inactivating null mutations is through the use of *transposons* or *insertion elements*. Transposons are parasitic DNA elements that are able to insert copies of themselves into another DNA molecule,

and many of them have little or no preference for specific target sequences. Barbara McClintock first discovered transposons in corn during the early 1950's, but her results were not widely accepted until nearly 20 years later, though they did ultimately garner her the Nobel Prize in 1983.

Several different kinds of transposons have been engineered to make them useful for random mutagenesis experiments. Common ones include Tn5 and the Mariner transposon, which are able to insert themselves at essentially any point in a DNA sequence (Tn5 has some preference for inserting into G-rich sequences, while Mariner inserts at TA dinucleotides) and have been modified to carry antibiotic resistance genes. This allows you to treat a population of bacteria with the transposon and select for those cells that have successfully integrated it into their chromosome by growing the cells on media containing that antibiotic. Each individual cell will have only one insertion, but since they occur at random positions, pooling together many cells results in a diverse population of mutated cells. This is a *transposon library*, should contain a wide variety of different highly polar insertions, and can be screened or selected for phenotypes as usual (Figure 3.3). The location of the transposon insertion in the desired mutants can then be determined by a variety of methods, the most common of which now is simply genome sequencing.

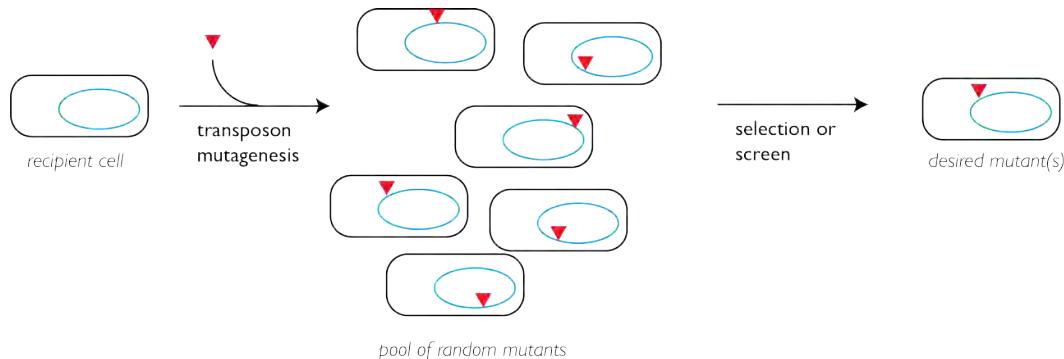
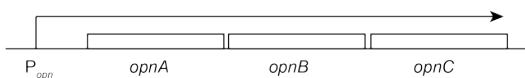


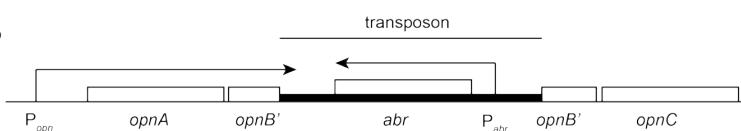
Figure 3.3: Transposon mutagenesis generates a library (or pool) of strains, each containing one randomly located transposon insertion. This library can then be screened to identify insertions that cause phenotypes of interest.

Transposon mutants, like other insertions, nearly always destroy the function of the gene they integrate into, and are highly polar (Figure 3.4), which limits their usefulness for some kinds of mutagenesis experiments. You will, for example, almost never get a gain-of-function mutation or any phenotypes that require more subtle changes in gene function from a transposon mutagenesis experiment.

A



B



C

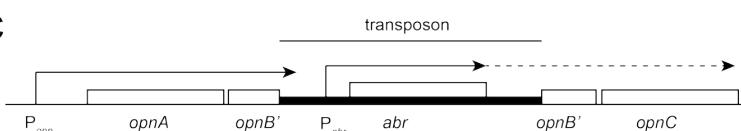


Figure 3.4: Transposon insertions are (almost) always polar null mutations. If the *opnB* in the operon shown in A is interrupted by an antibiotic-resistance (*abr*)-encoding transposon as shown in B, the expression of both *opnB* and *opnC* will be disrupted. Most modern transposons contain multiple transcriptional terminators to prevent the situation shown in C, where expression from a promoter within the transposon extends outward to change the expression of genes outside the bounds of the transposon itself.

A transposon library should contain enough independent mutants to ensure that at least one insertion is present in every non-essential gene. If this is true, the library is said to be saturated. As a general rule of thumb, you need to have at least 4 to 5 times as many independent mutants in your transposon library as there are genes in the genome of the organism you're studying to ensure good saturation. However, the more insertions you have in a transposon library the better, since you will increase the odds of having at least one transposon even in very small genes. I have seen recent papers using libraries of up to a million mutants, which represent transposons in essentially every possible insertion site in an organism's genome.

It is occasionally possible to use naturally-occurring transposons in a bacterial genome as a source of spontaneous mutations, if a species contains a transposon that happens to translocate at high frequency. In [this paper](#), for example, the authors took advantage of a highly-active transposon present in the genome of the lactic acid bacterium *Tetragenococcus halophilus* to isolate spontaneous mutants with phenotypes relevant to soy sauce fermentation. This is only worth doing in organisms where other options are very limited, but is a clever strategy that can be pulled out when necessary.

DISCUSSION PROBLEM SET #7: LIMITATIONS OF TRANSPOSONS

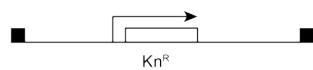
Problem #1

The integration of transposons into a DNA molecule is carried out by enzymes called transposases. Natural transposons encode transposase, but engineered transposons (like the Tn5 and Mariner-based ones mentioned above) do not, and require expression of transposase from a different source (usually a plasmid; see [Lecture 5](#)).

Tn5 (5.8 kb)



mini-Tn5 (2.4 kb)



Genetic structures of the naturally-occurring Tn5 transposon and of the very commonly-used mini-Tn5 transposon derived from it. Black boxes indicate repeat sequences that Tn5 transposase recognizes to catalyze insertion of the sequence between them into other DNA molecules. The *tnp* gene encodes transposase, and *tnp** encodes a null mutant of transposase with a premature stop codon. *Kn^R* = kanamycin resistance, *Bl^R* = bleomycin resistance, *Sm^R* = streptomycin resistance.

Why are transposons intended for generating mutant libraries engineered this way?

Problem #2

Let's suppose you make a transposon library of the cellulose-secreting bacterium *Komagataeibacter xylinus*, with the goal of finding mutants that produce higher than normal amounts of cellulose, which might be useful industrially.

(See [Walker et al. 2023](#) for an example of turning *K. xylinus* cellulose mats into a prototype leather alternative.)



A mat of bacterial cellulose produced by *K. xylinus* (Wikipedia).

However, despite your best efforts you are unable to isolate any transposon mutants that make more cellulose than the wild-type strain.

Why might this have failed? List as many reasons as you can think of.

Tn-seq (transposon sequencing, variations of which are also called INSeq, TraDIS, or HITS) is a very powerful technique that combines transposon mutagenesis with high-throughput DNA sequencing, allowing screening, enriching, or selecting for many transposon mutants in a single experiment without the need to isolate them individually (Figure 3.5).

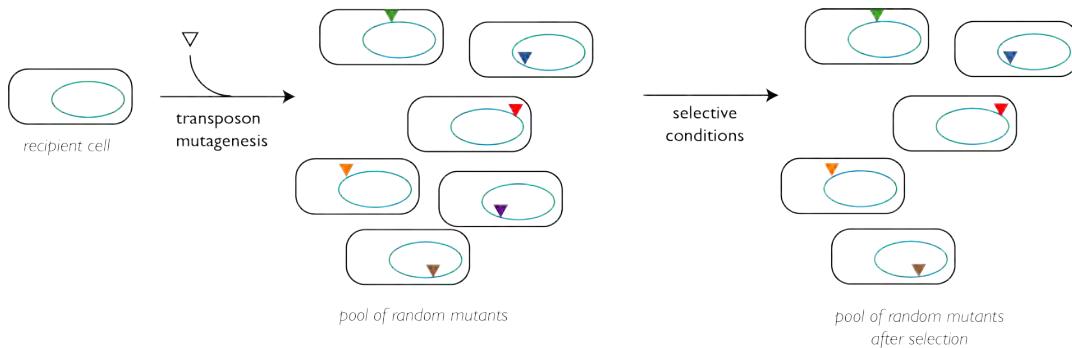


Figure 3.5: Transposon sequencing is a technique in which a pooled transposon library is subjected to selective conditions and high-throughput sequencing techniques are used to determine all of the transposon insertion sites in that library both before and after selection. Transposons whose frequency in the library change after selection (the one indicated in purple, in this case) are assumed to be in genes important for growth under those conditions.

In a Tn-seq experiment, a transposon library is generated and grown under permissive conditions. This pool of bacteria is sequenced to identify all of the sites where transposons have inserted into the genome in that library. The pool is then exposed to the non-permissive conditions, allowed to grow, and re-sequenced.

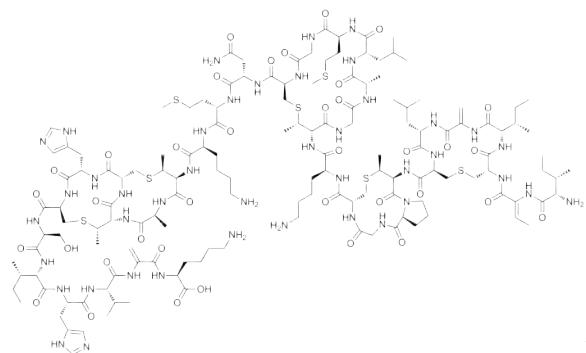
Since the output of Tn-seq experiments is the change in the proportion of a particular mutation in a bacterial population, more nuanced results are possible than from simpler plate-based screens. Any transposons whose presence in the library decreases after treatment are likely to be in genes important for growth under those conditions, even if they are not eliminated entirely. You may also be able to identify mutations that **increase** in frequency after treatment, identifying genes whose products are detrimental under the test condition.

MUTANT HUNTS AS EXPERIMENTS

It may not be immediately obvious how the principles discussed in the section above on experimental design apply to mutant hunts. To illustrate, let's take a look at an experimental problem and see how we could potentially solve it using the genetic techniques we've discussed so far.

EXAMPLE OF AN EXPERIMENTAL DESIGN: SCREENING FOR NISIN RESISTANCE GENES

Nisin is an antimicrobial bacteriocin produced by some, but not all, strains of the lactic acid bacterium *Lactococcus lactis*. Nisin efficiently inhibits the growth of a wide range of Gram-negative and Gram-positive bacteria. It is commonly used to prevent the growth of pathogenic bacteria on the surfaces of food, including hard cheeses.



Wikipedia

L. lactis strains that produce nisin are immune to its effects, as are some strains of the food-borne pathogen *Listeria monocytogenes*. Design a mutant hunt experiment to identify bacterial genes required for nisin resistance.

Hypothesis: Nisin-resistant *L. monocytogenes* strains encode a gene or genes that confer nisin resistance.

Experimental Design:

- 1) Generate a saturated transposon library in a nisin-resistant strain of *L. monocytogenes*. (The genome of *L. monocytogenes* encodes about 3000 genes, so we'll need at least 12-15,000 individual mutants.)
- 2) Screen individual mutants for the ability to grow on media containing nisin.
- 3) Sequence to identify the location of transposon insertions in strains that lose the ability to grow in the presence of nisin. The disrupted genes in these strains will be necessary for nisin resistance.

Independent Variable: (what will you change?)

- The location of the transposon insertion in each mutated strain.

Dependent Variable(s): (what will you measure?)

- The ability of each mutant to grow in the presence of nisin (a qualitative measurement).

Negative Control(s): (eliminate false positive results)

- Confirm that a nisin-sensitive strain of *L. monocytogenes* will not grow on your nisin-containing plates.

Positive Control(s): (eliminate false negative results)

- 1) Confirm that the nisin-resistant wild-type strain of *L. monocytogenes* will grow on your nisin-containing plates.
- 2) Confirm that your transposon library is saturating and contains an adequate diversity of different transposon insertions. There are several ways to do this, either by randomly sampling a subset of the pool or by deep sequencing.

Potential Outcomes:

- 1) One or more transposon insertions in *L. monocytogenes* eliminates the ability to grow in the presence of nisin. The disrupted genes are implicated, either directly or indirectly, in nisin resistance.
- 2) No transposon insertions are identified that make *L. monocytogenes* nisin-sensitive. If the library is saturating, that might mean that there are redundant genes that confer nisin-resistance or that the gene which confers nisin resistance is essential and cannot be knocked out.

The first experiment you may have thought of was to screen for transposon mutants of *L. lactis* that lose the ability to resist nisin. This seems reasonable, but will actually not work, since, as stated in the problem, nisin-resistant strains of *L. lactis* are also nisin producers. This means that any *L. lactis* mutant that cannot resist nisin will kill itself even in media that contains no added nisin. This is, in some ways, a subset of the second Potential Outcome listed above, where the desired mutation might be in a gene that is essential under the tested conditions. If this is the case, then mutagens which make only null mutations (like transposons) will never generate a mutation that satisfies your screen.

BRUTE FORCE AND ELEGANCE IN EXPERIMENTAL DESIGN

Elegance is an elusive, but desirable, property of experiments, and nowhere is this more apparent than in the design of mutant hunts. There are multiple ways to solve any experimental problem. The experiment described above will work, but it's a very labor-intensive, brute-force approach to the problem, requiring individually screening tens of thousands of colonies for nisin sensitivity. Can we redesign our experiment to be more efficient?

Here's another possible approach:

EXAMPLE OF AN EXPERIMENTAL DESIGN: A MORE ELEGANT SCREEN FOR NISIN RESISTANCE GENES

Design a mutant hunt experiment to identify bacterial genes required for nisin resistance.

Hypothesis: Nisin-resistant *L. monocytogenes* strains encode a gene or genes that confer nisin resistance.

Experimental Design:

- 1) Generate a saturated transposon library in a nisin-resistant strain of *L. monocytogenes*.
- 2) Grow the transposon library in regular growth media and sequence to determine the frequency and distribution of transposon insertions.
- 3) Grow the transposon library in growth media containing nisin and sequence to determine the frequency and distribution of transposon insertions.

Independent Variable: (what will you change?)

- The presence or absence of nisin in the growth media.

Dependent Variable(s): (what will you measure?)

- 1) The distribution and frequency of transposon insertions in the absence of nisin (a quantitative measurement).
- 2) The distribution and frequency of transposon insertions in the presence of nisin (a quantitative measurement).

Negative Control(s): (eliminate false positive results)

- Confirm that a nisin-sensitive strain of *L. monocytogenes* will not grow in your nisin-containing medium.

Positive Control(s): (eliminate false negative results)

- 1) Confirm that the nisin-resistant wild-type strain of *L. monocytogenes* will grow in your nisin-containing medium.
- 2) Confirm that your transposon library is saturating and contains an adequate diversity of different transposon insertions. This will be easy to determine in this case, since you will be measuring this as Dependent Variable #1.

Potential Outcomes:

- 1) There are one or more transposon insertions that occur more frequently in the pool of mutants grown without nisin than in the pool of mutants grown with nisin. These insertions are likely to be in genes that are necessary for nisin resistance.
 - 2) There are no differences in the frequency and distribution transposon insertions in media with and without nisin. This is good evidence that no null mutation can confer nisin sensitivity, and means that nisin resistance is likely to be due to a point mutation modifying the function of a gene.
-

This is a Tn-seq experiment, which is still a screen, but it is a much less labor-intensive one, since all of the mutants are tested simultaneously in the same test tube. Note that this does have the disadvantage of not isolating any individual nisin-resistant mutants, so any genes that seem to be involved in nisin resistance will have to be knocked out separately for future study.

Is there an even better way? Here's one more possibility:

EXAMPLE OF AN EXPERIMENTAL DESIGN: A SELECTION FOR NISIN RESISTANCE GENES

Design a mutant hunt experiment to identify bacterial genes required for nisin resistance.

Hypothesis: Nisin resistance is conferred by gain-of-function mutation(s) in nisin-sensitive *L. monocytogenes*.

Experimental Design:

- 1) Expose a culture of nisin-sensitive *L. monocytogenes* to a mutagenic dose of UV light.
- 2) Plate 10^8 CFU from the mutated culture on media containing nisin. If desired, do this on 10 – 100 plates to increase the number of mutants tested.
- 3) Sequence the genomes of colonies that grow in the presence of nisin.

Independent Variable: (what will you change?)

- The genome sequence of nisin-sensitive *L. monocytogenes*.

Dependent Variable(s): (what will you measure?)

- The ability of mutants to grow on nisin (a qualitative measurement).

Negative Control(s): (eliminate false positive results)

- Confirm that the nisin-sensitive wild-type *L. monocytogenes* will not grow on your nisin-containing plates.

Positive Control(s): (eliminate false negative results)

- 1) Confirm that a nisin-resistant wild-type strain of *L. monocytogenes* will grow on your nisin-containing medium.
- 2) Confirm that your mutagenesis was effective. This can probably be best achieved by sequencing the genomes of several randomly chosen colonies (before nisin selection) to get a sense of how many mutations per strain are present and that they are more or less randomly distributed around the genome.

Potential Outcomes:

- 1) You isolate nisin-resistant mutants. Sequencing their genomes will identify candidate mutations responsible for this phenotype, which will very likely cluster in a few genes directly involved in nisin resistance.
- 2) You are unable to isolate nisin resistant mutants. This means that no point mutation caused by UV light is able to make nisin-sensitive *L. monocytogenes* nisin resistant. The most likely explanation for this is that nisin resistance requires a specific gene or genes that is not present in your nisin-sensitive strain, and that nisin resistance is likely to depend on gene acquisition by horizontal gene transfer.

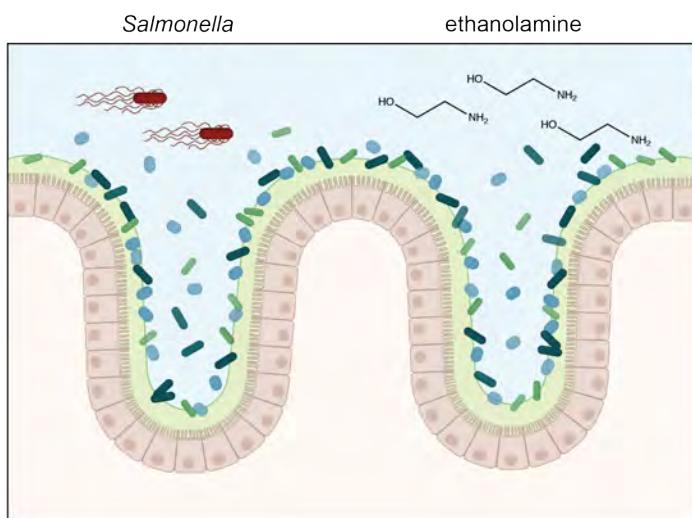
If you are able to isolate nisin-resistant mutants this way, it demonstrates that it is **possible** for *L. monocytogenes* to acquire nisin resistance via point mutations, although not necessarily that this is how clinically-relevant strains have done so. Comparison of the mutations acquired in this experiment to the genomes of nisin-resistant and nisin-sensitive *L. monocytogenes* isolates will help to determine if that has happened in the wild.

This experimental design will, in a single step, test millions or billions of random point mutations at once for their ability to confer nisin resistance, and illustrates why it is pretty much always better and more elegant to use a selection in a mutant hunt, if possible. The problem is that it's not always possible, or might be technically challenging to set up, which is why screens and enrichments are so widely used. As long as the method you choose results in answering the question you are asking, it counts as a successful experiment.

DISCUSSION PROBLEM SET #8: SCREENS AND SELECTIONS

Problem #1

Salmonella enterica can grow on the lipid breakdown product ethanolamine as a sole carbon and nitrogen source. Ethanolamine is abundant in the mammalian gut, especially during inflammation. You hypothesize that *S. enterica* has genes encoding a pathway specifically required for growth on ethanolamine.



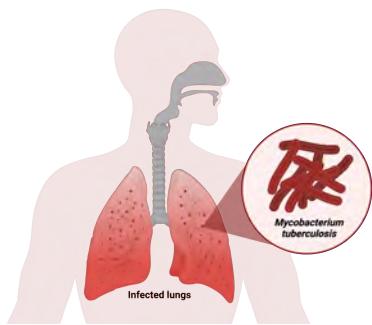
Created with Biorender.

Design a mutant hunt that would allow you to identify *Salmonella* genes necessary for growth on ethanolamine, and state:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, and/or an enrichment to identify relevant mutants?
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #2

Mycobacterium tuberculosis strains often become resistant to antibiotics (like rifampicin) over the course of an infection, but do not typically acquire any new genes by horizontal gene transfer to do so (probably because they don't encounter any other bacteria that they could acquire genes from when sequestered inside a lung granuloma).



Created with Biorender.

Your lab strain of *M. tuberculosis* is **not** resistant to rifampicin.

Design a mutant hunt that would allow you to identify mutations that lead to a Rif^R phenotype, and state:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, and/or an enrichment to identify relevant mutants?
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

REVERTANTS AND SUPPRESSORS

If a mutation causes a slow growth phenotype or prevents growth without actually killing the cells, you will sometimes observe secondary mutations in that strain that restore wild-type-like growth (Figure 3.6).



Figure 3.6: The appearance of spontaneous revertants. Note how faster-growing colonies containing suppressor mutations are arising out of a streak of slower-growing parent cells (which are themselves mutants that do not grow especially well under these conditions).

These arise because in the process of observing the poor growth phenotype, you are also selecting for mutants that **are** able to grow well under those conditions. Such a mutant is called a **revertant**, because the phenotype has reverted back to wild-type. A mutation that causes a reversion in phenotype is called a **suppressor**. In some cases, this may actually be due to a mutation that simply changes the original mutated gene back to the wild-type sequence, which is the least informative type of revertant. This is far more likely with point mutations than with other kinds of mutations.

Revertants caused by mutations in the same gene as the original mutation are referred to as *intragenic suppressors*. However, it is often most interesting and informative to identify *intergenic suppressor* mutations, which are mutations in **other** genes that restore the phenotype of your mutant strain. If mutating one gene causes a growth defect, and you identify suppressor mutations in a second gene that restore growth, you have very strong evidence that those two genes are involved in the same biological process (Figure 3.7).

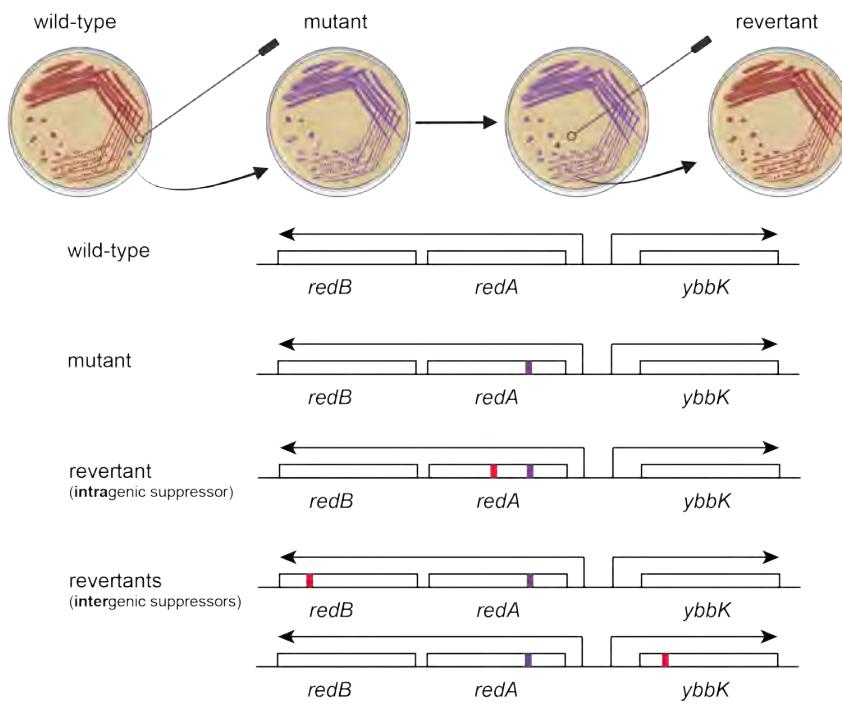


Figure 3.7: Types of suppressor mutations. In this example, a mutation in the *redA* gene converts the phenotype of the wild-type from red to purple. Spontaneous reversion to a red phenotype can be due to either **intragenic** suppressors (additional mutations in the *redA* gene) or **intergenic** suppressors (mutations in *redB* or *ybbK*). This is genetic evidence that all three of these genes contribute to the color phenotype of these bacteria. Made with Biorender.

Suppressor mutations occur at the same rates as any other example of a particular kind of mutation (Table 2.2), but whether you observe them or not depends on the phenotype in question and, in many cases, on whether there is selective pressure that allows revertants to outcompete their parent strain. It is worth noting that revertants may have a **wild-type phenotype**, but nearly always have a **mutant genotype**. It may not be possible to distinguish between the wild-type and a revertant based on the phenotype alone.

Multicopy suppressors are genes that, when present in more copies than in the wild-type (see **Lecture 5** on plasmids), suppress the phenotype of a mutation in a different gene. While this is commonly observed in the lab with overexpression of genes from plasmids, it can result naturally from duplication mutations.

Some mutations are **only** ever found with a suppressor elsewhere in the genome, and it can be hard to know when this is the case without whole-genome sequencing. A mutant of *E. coli* lacking the *rpoH* gene, encoding a heat shock stress response regulator, cannot grow above 18°C, but it is relatively easy to isolate *rpoH* null mutants at 30°C. How does that happen? The strains you isolate turn out to have suppressor mutations that result in an increase in the expression of protein-stabilizing chaperone systems, but if you didn't know that, you might make the wrong conclusions about the function of *rpoH*.

In a related phenomenon, there are also mutations that have no obvious phenotype on their own, but have strong phenotypes when they occur in **combination** with another mutation. When either one of a pair of genes can be knocked out, but you cannot delete both of them simultaneously, they are referred to as being **synthetically lethal**. Synthetic lethality is a strong piece of evidence that two genes are involved in related processes, or may in fact be **functionally redundant** genes that encode the **same** essential function (Figure 3.8).

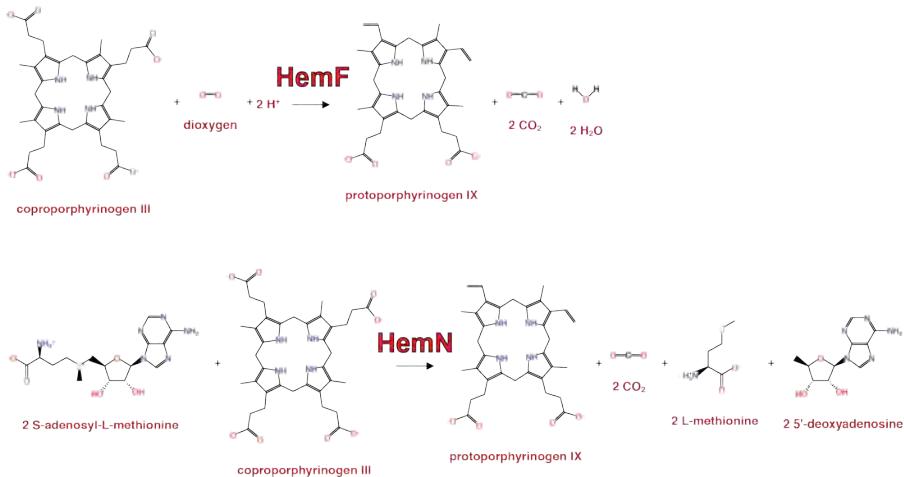


Figure 3.8: Synthetic lethality by functional redundancy. HemF and HemN are both enzymes that catalyze the conversion of coproporphyrinogen III to protoporphyrinogen IX in the *E. coli* heme biosynthesis pathway, albeit by different biochemical mechanisms (reaction drawings from EcoCyc). It is possible to knock out either of the *hemF* or *hemN* genes individually with no ill effect, but a *hemF hemN* double mutant becomes a heme auxotroph, and will die in the absence of heme.

MUTATIONS YOU WILL NEVER ISOLATE

There are some kinds of mutations that are very difficult or impossible to obtain, no matter how clever your mutant hunt design might be. The most common example of this is null mutations in *essential genes*, which encode functions that are absolutely required for the cell to survive. These include genes required for key cellular functions like DNA replication, RNA synthesis, and protein translation.

Of course, which genes are essential depends on what growth conditions you're examining and on what strain you are using. This is also true of synthetic lethality or any other phenotype that results in the lack of growth. The genes described in Figure 3.8 are essential in media without heme, but dispensable if heme is added. A mutant defective in glutamate uptake may not grow with glutamate as a sole carbon source, but grow fine when given glucose. If you can rescue a lethal phenotype by changing the growth conditions, the genes involved are only *conditionally essential*.

Certain kinds of gain-of-function mutations may also be difficult or impossible to obtain if they cause toxic effects or consume all of a critical cellular resource in an uncontrolled way.

SCIENTIFIC PROCESS 4: ALTERNATIVE APPROACHES AND TROUBLESHOOTING

There is never only one way to address a scientific question. Testing a hypothesis in multiple independent ways is, in fact, a great way to ensure that any one experiment is not giving you misleading results. Most scientific papers (the good ones, anyway) will use multiple approaches to test and validate their conclusions.

As we saw in the nisin resistance gene experiments we worked through earlier in the chapter, different approaches to a problem have different advantages and disadvantages and give you different kinds of results, so combining multiple approaches is the most rigorous way to test a hypothesis. When designing experiments for this class, different groups are very likely to come up with different, equally valid approaches to answer each question. This is fine! As we move through the different lectures, the tools you have available will expand and this will make more different kinds of experiments possible.

A related subject is troubleshooting: what do you do when your experiment doesn't work the way you expect it to?

Be very careful when you say that an experiment has "failed". Sometimes equipment breaks or contamination ruins a procedure, and the results of those experiments can be safely ignored while you fix the technical problem. Your controls will help you identify when this has happened.

An experiment that just doesn't give you the results you expected is **not a failed experiment**. It is a **discovery**. This is why the Potential Outcomes section of an experimental design is so important. You need to think about **all** the possible outcomes of your experiment, and be able to adjust your model to account for the result you actually get, not just the one that fits your preferred hypothesis.

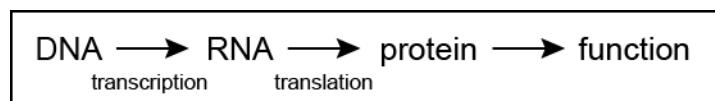
LECTURE 4: PRINCIPLES OF REGULATION

INTRODUCTION

In this lecture, we will discuss regulation in bacteria, with a focus on interpreting the phenotypes of mutations that affect regulation and designing genetic experiments to explore how the expression of bacterial genes are controlled.

GENE EXPRESSION IS NOT CONSTANT

Bacterial genomes typically encode a few thousand different proteins. Not all of these proteins are present at the same concentration or at the same time, and bacteria are able to control the expression and activity of proteins in response to changes in their environments. Recall the basic (and, as we've mentioned, oversimplified) flow of information from DNA to protein function, with DNA transcribed to RNA, which is translated into protein, which then has a biological activity:



The steps in this process that can be regulated include:

1. Transcription initiation
2. Transcription elongation
3. Transcription termination
4. mRNA stability
5. Translation initiation
6. Translation elongation
7. Protein stability
8. Protein activity

Any gene product may be regulated at one or more of these steps in response to either internal or external signals. In this chapter, I will summarize what is known about these processes. As with most fundamental biological mechanisms, the details are understood best in the Gram-negative model bacterium *Escherichia coli* and may differ more or less dramatically in other species, but I will try to focus on broadly-applicable general principles.

MUTATIONS IN REGULATORS

As geneticists, it is important to understand what kinds of phenotypes arise from mutations in regulators and how we can use and interpret those phenotypes to understand biological systems.

At the simplest level, there are two kinds of regulators: *positive* and *negative*. A positive regulator directly **activates** the system being studied in response to a signal. A negative regulator **represses** the system, and that repression is what responds to the signal. When negative regulation is relieved in response to a signal, this is called *derepression*, which, as we will see below, looks phenotypically very similar to activation.

Mutations in positive regulators are often relatively easy to interpret. If a positive regulator is required to activate a particular phenotype, then null mutations in that regulator will have the same phenotype as null mutations in the other genes required for that phenotype. The genes controlled by such a regulator will be *constitutively inactive* (always OFF) in the mutant.

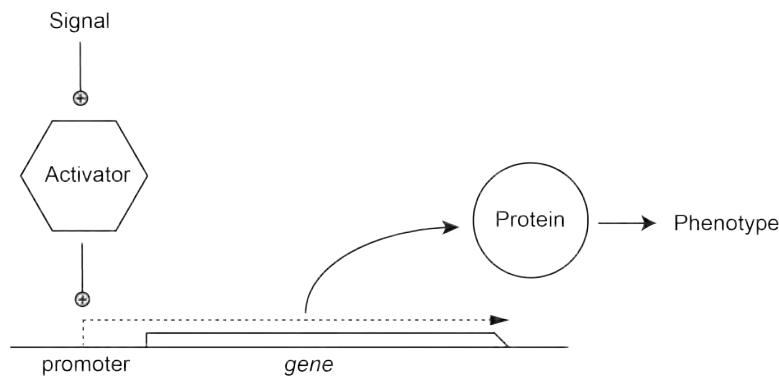


Figure 4.1. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional activator.

The following table illustrates how null mutations in different components of the circuit shown in Figure 4.1 would be expected to change the observable phenotype, which will only occur when the Protein encoded by gene is produced:

Mutation	Signal	Phenotype
wild-type	absent	-
wild-type	present	+
Δ activator	absent	-
Δ activator	present	-
Δ gene	absent	-
Δ gene	present	-

Mutations in negative regulators can have less straightforward phenotypes. A very common regulatory circuit in bacteria involves *transcriptional repressors*, proteins that bind to DNA and prevent transcription of genes until they detect a signaling molecule or metabolite. When that metabolite is present, the repressor loses its ability to bind DNA, and the repressed genes are then expressed. Other transcriptional repressors may respond to signals by becoming **better** at binding DNA, which will **decrease** gene expression or activity. In either case, the result of a null mutation in a negative regulator is likely to be *constitutive expression* or activity of the genes or proteins being regulated (always ON), which can have very different effects depending on the genes in question.

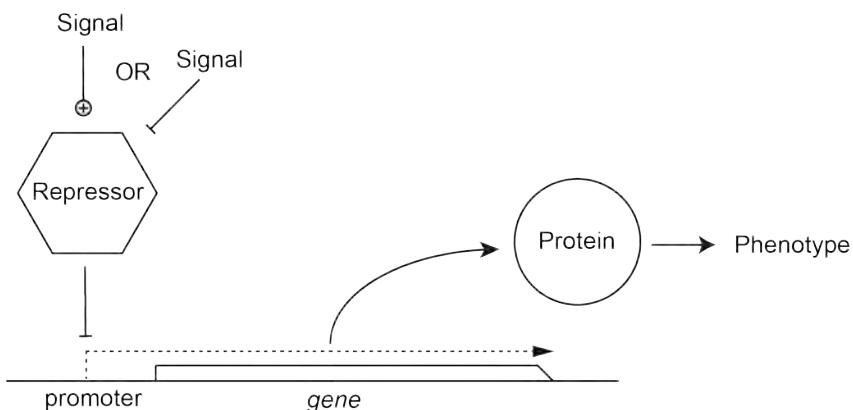


Figure 4.2. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional repressor.

The following table illustrates how null mutations in different components of the circuit shown in Figure 4.2 would be expected to change the observable phenotype if the repressor responds to a signal that **increases** its repressive functions (this is an example of a gain-of-function phenotype resulting from a null mutation):

Mutation	Signal	Phenotype
wild-type	absent	+
wild-type	present	-
Δ repressor	absent	+
Δ repressor	present	+
Δ gene	absent	-
Δ gene	present	-

If the repressor responds to a signal that **decreases** its repressive functions, null mutations in different components of this circuit would be expected to change the observable phenotype as follows:

Mutation	Signal	Phenotype
wild-type	absent	-
wild-type	present	+
Δ repressor	absent	+
Δ repressor	present	+
Δ gene	absent	-
Δ gene	present	-

Operator sequences are sites in DNA (usually within the promoters of genes; see below for more on promoters) where proteins or other regulators bind to control gene expression. In eukaryotes these are called response elements. Mutations in operators can change how well regulators bind to those sequences, leading to a variety of phenotypes depending on the nature of the operator mutation and the regulator.

Global regulators control many genes or gene products throughout a cell's genome. They may be positive regulators of some of those genes and negative regulators of others, and mutations of global regulators often have very complex pleiotropic phenotypes. *Local regulators* control only a small set of genes, often in the same locus or operon as the gene encoding the regulator itself. Many genes are regulated by both global and local regulators, which allows them to respond in sophisticated ways to changes in the cell's environment. The classic example of this is the *lac* operon of *E. coli*, which is repressed by the lactose-specific local regulator LacI and activated by the cyclic AMP-sensing global regulator CAP so that the *lac* operon is only expressed when lactose **is** present, but glucose **isn't**.

It is important to note that not all gene regulation is absolute. In some cases you will have a gene switched entirely on or entirely off, but many regulators only adjust gene expression levels. This is particularly true for genes affected by multiple regulators (which is probably most of them).

SURVEY OF REGULATORY MECHANISMS

Regulation of biological systems has been studied for a long time, with repression of gene expression, covalent modification of enzymes to control their activity, and feedback inhibition of enzyme activity all first being described in the mid-1950s. In the next part of this chapter, I will briefly describe some of the different types of regulation that are known to occur in bacteria without giving many specific examples (which would rapidly become overwhelming). I will also give examples of the different methods available for measuring gene expression in bacteria.

 We will be discussing a lot of different laboratory techniques in this course, some of which are historically important but are now obsolete. This symbol will be used to mark outdated techniques that have been replaced with more effective modern technologies. Not every old technique is obsolete, so some that have remained in use for many decades and are still broadly useful will not be marked this way.

My goal here is to give you a broad sense of how complex regulation can be and what tools are available to study it. The experimental problems below focus on how to decipher and understand phenotypes resulting from mutations in regulatory factors, with an emphasis on being able to narrow down the possible mechanisms leading to particular phenotypes.

REGULATION OF mRNA LEVELS

The first step in production of a protein is transcription of the mRNA encoding that protein by RNA polymerase. This involves three steps that can be regulated: *initiation*, *elongation*, and *termination*. The actual amount of a particular mRNA in a cell is determined by the combination of these factors and the *stability* of that mRNA. The level of an unstable RNA can be changed very rapidly by changes in initiation, elongation, or termination, while it might take several cellular generations to significantly change the levels of a very stable RNA.

Regulating transcription initiation might be the least wasteful method of regulation, from the cell's point of view, since no nucleotides, amino acids, or energy are wasted producing unwanted gene products. However, it is also the slowest to respond to changes in the environment, since the cell must go through the entire process of transcription and translation to produce a final protein product (not **very** slow; this still takes less than a minute in *E. coli*).

I. Transcription initiation. Initiation of transcription takes place at a *promoter* where RNA polymerase binds to the DNA. Many bacterial genes have more than one promoter, each of which can be regulated independently.

Promoters vary in sequence, and the sequence of the promoter has a very strong effect on how efficiently a gene is transcribed. The *sigma subunit* (σ or *sigma factor*) of RNA polymerase is a small protein that determines the DNA sequence to which a particular molecule of RNA polymerase will bind. Typically, bacteria encode a *housekeeping sigma factor*, which is the

most abundant sigma factor in the cell and recognizes the promoters of genes that need to be transcribed under most growth conditions. In *E. coli*, this is σ^{70} (so-called because of its molecular weight of 70 kDa), encoded by the *rpoD* gene, and it recognizes promoters containing consensus sequences of TTGACA and TAATAT centered at positions 35 nucleotides and 10 nucleotides upstream of the *transcriptional start site*, respectively (the -35 and -10 sites). The more similar the sequence of a promoter is to the consensus sequence for a particular sigma factor, the more strongly it will be bound by that sigma factor, which usually increases the amount of mRNA produced from that promoter.

Software exists that attempts to predict promoters (and transcriptional terminators) from DNA sequences (see the tools at [this link](#), for example), but none of the tools I am aware of is especially good at doing so, even when attempting to predict transcriptional regulatory elements in well-characterized species like *E. coli*. It is possible that machine learning / AI approaches will allow this to be done more effectively, but I haven't seen anything systematic in bacterial systems to address the problem.

Alternative sigma factors can replace the housekeeping sigma factor in RNA polymerase, and typically drive the transcription of genes important in responding to particular types of stress (e.g. heat shock, stationary phase growth, or membrane damage), involved in the construction of complex molecular machines (e.g. flagella), or required for processes like virulence or spore formation. They recognize consensus sequences different from those found in promoters transcribed by RNA polymerase containing the housekeeping sigma factor. The concentration and activity of alternative sigma factors are tightly controlled, often using multiple mechanisms of transcriptional and post-transcriptional regulation, but several sigma factors can be present and active at the same time in a cell, all competing for the pool of core RNA polymerase. Different bacterial species may contain anywhere from one to dozens of sigma factors, depending on the complexity of their environment and developmental pathways (see [Lecture 11](#)).

A common form of regulation for alternative sigma factors are *anti-sigma factors*, negative regulator proteins that bind to sigma factors and prevent them from interacting with RNA polymerase. The stability or activity of the anti-sigma factor can then be regulated to control the activity of the sigma factor itself. Anti-sigma factors are usually cytoplasmic, but can be membrane-bound or, in [exceptional cases](#), span both the inner and outer membranes to sense external conditions in Gram-negative bacteria.

Other features of promoters can also influence the efficiency of transcription initiation. UP elements are AT-rich DNA sequences upstream of the -35 site that increase transcription 30 to 70-fold. For some extremely highly active promoters (like those driving transcription of ribosomal RNA), the *initiating nucleotide* (that is, the first nucleotide of the transcribed RNA, most commonly A or G) can influence initiation in response to the levels of ATP or GTP in the cell, directly linking cellular energy state to gene expression.

Transcription initiation

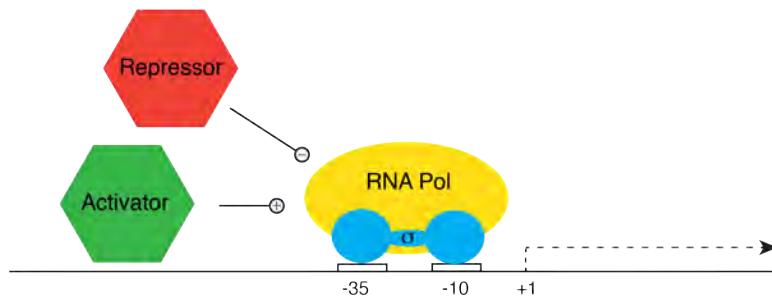


Figure 4.3. Regulators that can affect transcription initiation.

In addition to RNA polymerase itself, there are other proteins that can influence transcription initiation. These are called *transcription factors*. Most of these recognize specific DNA sequences in or near the promoter but in some cases, they may bind to RNA polymerase without interacting directly with the DNA. Repressors are transcription factors that prevent initiation of transcription, often by blocking the -10 or -35 sites or otherwise preventing RNA polymerase from binding to the promoter. Activators increase the rate of initiation when bound to a promoter; either by recruiting RNA polymerase to a promoter or by interacting with an RNA polymerase molecule that is already bound to the promoter and stimulating its activity. The same transcription factor can sometimes act as a repressor or as an activator at different promoters, depending on the nature of the protein and the location of the binding site in the promoter; and multiple transcription factors often regulate a single promoter. The DNA-binding or RNA polymerase-influencing activity of transcription factors is often controlled in response to changes in the metabolism or environment of the cell (see Regulation of Protein Activity section below).

Transcription initiation is probably the best understood stage of gene regulation, and you can find comprehensive reviews [here](#), [here](#), and [here](#).

It is also worth noting that DNA can be used experimentally to purify (or “pull down”, in biochemical jargon) proteins that bind to specific DNA sequences, for example by attaching short DNA molecules to agarose beads or to a chromatography column and then running a cell lysate through that column to see what proteins stick to the DNA. This is a useful way to identify which transcription factors bind to a particular promoter sequence, for example, and is more or less the opposite of **chromatin immunoprecipitation (ChIP)**, which uses antibodies against a specific protein to isolate DNA molecules that bind to that protein.

2. Transcription elongation. Once RNA polymerase has left the promoter and is producing mRNA, it enters the transcriptional elongation phase. The sequence and structure of the transcribed RNA determines the frequency of *transcriptional pause sites*, where RNA polymerase briefly stops producing mRNA. The number and position of pause sites can affect the speed of mRNA production and how it folds, which can affect both elongation and termination. There are proteins that interact with RNA polymerase to influence elongation speed and pausing (e.g. NusA or GreA), thereby regulating the amount of transcript produced.

3. Transcription termination. There is considerably more known about regulation of transcriptional termination than of elongation. In *Rho-dependent termination*, the Rho protein, as part of the RNA polymerase elongation complex, recognizes single-stranded RNA with no ribosomes attached and terminates transcription. (As mentioned in **Lecture 2**, this is why nonsense mutations are polar in Gram-negative bacteria: they result in long stretches of untranslated RNA in mRNAs.) *Rho-independent termination* also occurs (for about half of transcripts in *E. coli* and more in Gram-positive bacteria, generally). In these transcripts, *intrinsic terminators* are encoded in the mRNA itself that lead to the dissociation of RNA polymerase from the transcript. Intrinsic terminators are typically stable, GC-rich stem-loop structures 7 to 20 base pairs long, followed by a stretch of uracil residues.

Transcriptional termination can be regulated by *transcriptional attenuation* or by *anti-terminators*. Anti-terminators are proteins that prevent termination at specific termination sites, allowing RNA polymerase to bypass those sites. The best-studied example of this is mediated by the N and Q anti-terminator proteins of phage λ , but some examples of bacterial genes regulated by anti-termination are known to exist.

Attenuation is a more common mechanism in which an mRNA can take on more than one structural conformation, one of which is an intrinsic terminator. The classic example of this is the tryptophan (Trp) biosynthesis operon (briefly mentioned in **Lecture 1**), the first part of which encodes a small Trp-rich *leader peptide*. When this peptide is translated efficiently, the presence of ribosomes on the mRNA causes it to fold into a structure that includes an intrinsic terminator stem-loop. If translation stalls due to a shortage of Trp-charged tRNA, the mRNA folds differently, eliminating the terminator stem-loop and allowing transcription of the entire operon to continue. UAB Microbiology's own Chuck Turnbough has written [an excellent review](#) on this topic, if you're interested in more details.

Transcription termination

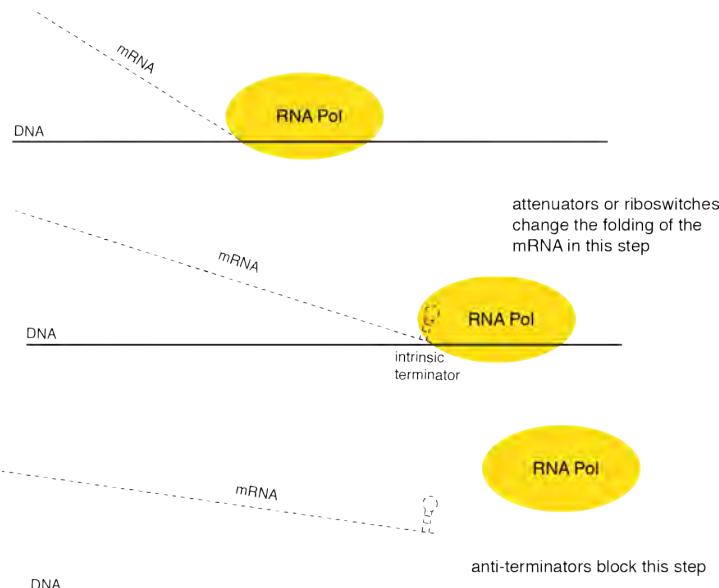


Figure 4.4. Regulators that can affect termination of transcripts with intrinsic terminators.

Riboswitches are widely-distributed regulatory RNA structures, usually found in the 5' untranslated region (UTR) of an mRNA, which fold into structures that bind specific metabolites (e.g. amino acids, metal ions, or vitamins) and can affect transcription termination in the same way as other attenuators.

4. mRNA stability. The final consideration in controlling the amount of a particular mRNA in the cell is *transcript stability*. The half-life of mRNAs varies greatly, ranging in *E. coli* from as little as 40 seconds to longer than 40 minutes, with an average between 2 and 6 minutes. Bacteria contain a variety of ribonucleases, which are enzymes that degrade RNA. The stability of a particular mRNA is determined by several factors. An important one is the presence of endonuclease cleavage sites, which are more common in some sequences than others.

The *translatability* of a particular mRNA (see below) can also affect mRNA stability, since an mRNA that is covered in ribosomes is less susceptible to nucleolytic cleavage.

RNA stability can also be regulated in response to environmental factors. The most common mechanism for this involves transcription of small, non-coding RNAs (sRNAs) that base-pair with the mRNA to be regulated (often overlapping the *ribosome binding site*, see below). The resulting double-stranded RNA then becomes a target for ribonucleases (in Gram-negative bacteria, RNase III, encoded by the *rcn* gene). *Cis*-acting sRNAs are transcribed from the non-coding strand of an open reading frame and are therefore exactly complementary to their target sequences. *Trans*-acting sRNAs are encoded elsewhere in the genome, typically have less exact matches to their target sequences, and can potentially regulate more than one mRNA. *Trans*-acting sRNAs nearly always require the RNA-binding protein Hfq or another RNA chaperone protein for activity. An *hfq* mutant is broadly defective in sRNA-mediated regulation, which can be useful for determining whether sRNAs are involved in a regulatory phenotype, although not every phenotype of an *hfq* mutant is dependent on sRNAs.

sRNA regulation of mRNA degradation

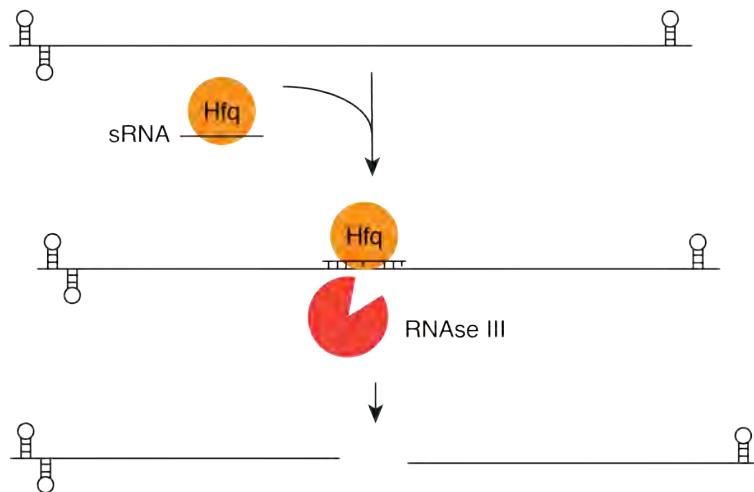


Figure 4.5. sRNA and Hfq-mediated mRNA degradation, by forming a double stranded RNA targeted by RNase III.

MEASURING GENE EXPRESSION: mRNA

There are a wide variety of techniques available for measuring mRNA levels in a cell. They can be divided into **direct** measurements of RNA and **indirect** measurements, and may be able to measure either expression of a single gene or all of the genes in an entire genome.

	Single Gene	Genome-Wide
Direct	qRT-PCR northern blot	RNA-seq DNA microarray
Indirect	transcriptional reporter fusion	--

Techniques to directly measure the amount of an mRNA in a cell include *northern blotting*, *quantitative reverse transcriptase PCR* (*qRT-PCR*), *DNA microarrays*, and *RNA sequencing* (*RNA-seq*). *qRT-PCR*, in which cDNA is produced from mRNA by reverse transcriptase then PCR amplified with gene-specific primers in the presence of fluorescent dsDNA reporters and quantified with a specialized instrument, is useful for measuring the levels of individual mRNAs. *RNA-seq* uses next-generation sequencing technologies to measure the concentrations of all of the mRNAs produced from the entire genome (*transcriptomics*). *Northern blotting* and *microarrays* are largely obsolete methods of accomplishing the same things, respectively.

RNA-seq is probably currently the best technique for assessing transcript abundance, and is growing cheaper all the time, but can still become prohibitively expensive if a lot of different samples need to be analyzed. There are even [multiple methods for single-cell](#) *RNA-seq* in bacteria, if you have a very large amount of money and a problem that requires understand-

ing differences in genome-wide transcription among individual cells, but they are not as well-established as the equivalent methods in eukaryotic cells and tend to detect only a few transcripts per cell. This is a technology in active development, and may well become more standardized and useful in the next few years.

A long-established and common technique to indirectly measure the amount of an mRNA in a cell is by using *transcriptional reporter fusions*. These are plasmids or chromosomal insertions in which the promoters of genes of interest are artificially engineered to place them upstream of genes encoding products that are easy to measure (*reporter genes*). The level of transcription from that promoter is then inferred from the amount of reporter product produced. Commonly used reporters include fluorescent proteins like GFP or mCherry, enzymes with simple colorimetric assays like β -galactosidase (LacZ) or β -glucuronidase (GUS), or one of a variety of luciferase enzymes, which produce light.

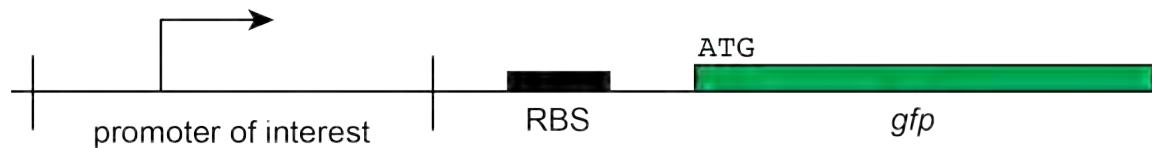


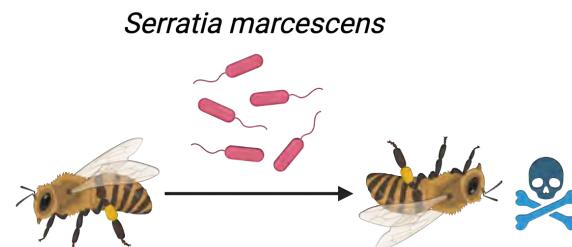
Figure 4.6. Transcriptional fusions link a promoter of interest with an easily-measured reporter gene (in this case, *gfp*).

One advantage of using transcriptional reporters is that they are a simple, cheap way to measure expression from a particular promoter, and with fluorescent reporters and a good microscope, can give information on cell-to-cell variations in gene expression, which RNA-based methods cannot do easily. There are several disadvantages, though. First is that they are intrinsically non-physiological, since they are cloned promoters driving non-physiological products from (usually) multi-copy plasmids (see **Lecture 5** for more on plasmids). Secondly, high production of reporter gene products may be toxic (fluorescent proteins) or be a large drain on the cell's ATP pool (luciferase). Thirdly, cellular growth conditions can affect reporters in ways that they would not affect the actual gene product. Both GFP and luciferase require oxygen, for example, and both LacZ and GUS can be inactivated by oxidative stress (e.g. hydrogen peroxide). Finally, the readout from a reporter fusion is always delayed relative to the actual production of the mRNA due to the time necessary to translate the product and, for fluorescent proteins in particular, the time needed for that product to mature into its active form (from a few minutes to several hours).

Remember that techniques that directly measure the amount of a particular RNA in a population of cells are measuring the **combined** effect of synthesis and stability. This is not true for indirect assays, since the reporters are generally stable transcripts and proteins. Transcriptional fusions therefore typically only measure **synthesis** rates, since the reporter accumulates over time but does not degrade. This also means, of course, that fusions can only reflect the activity of the promoter, which may or may not accurately describe the regulation of the actual mRNA.

EXAMPLE OF AN EXPERIMENTAL DESIGN: HIGH-THROUGHPUT SCREENING FOR REGULATORS

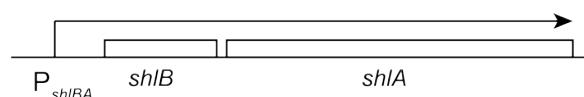
Serratia marcescens is a Gram-negative bacterium that is an opportunistic pathogen of both humans and honeybees. *S. marcescens* infections are one of several factors implicated in honey bee colony collapse, which is a major agricultural and ecological concern.



Created with Biorender.

The virulence factors of *S. marcescens* include a variety of different enzymes capable of breaking down host cells. Important among these is hemolysin, named for its ability to lyse red blood cells, although since insects have neither blood nor hemoglobin, the actual cells being lysed in the bee are more general. As happens with many bacterial pathogens of insects, the result of an *S. marcescens* infection is essentially the liquification of everything inside the bee.

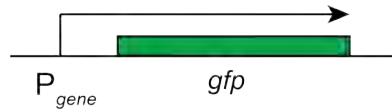
S. marcescens hemolysin is encoded by the *shlBA* genes:



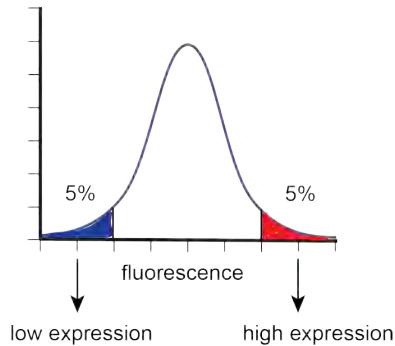
Design an experiment to identify the regulators of hemolysin gene expression in *S. marcescens*.

This is not a trivial problem, but a variety of methods have been developed over the years to make it easier to solve. A new-ish one that is particularly relevant here is [SorTn-seq](#), a method that combines clever reporter construction and transposon mutagenesis with FACS (fluorescence-activated cell sorting; a type of machine which is able to sort thousands of single cells based on their individual levels of fluorescence) to screen for mutations that either increase or decrease expression from a promoter of interest.

The first step in SorTn-seq is to construct a transcriptional fusion reporter strain in which the promoter for the gene of interest (in this case, *shlAB*) is fused to a fluorescent protein, so that expression of the fluorescent protein will reflect expression of the gene of interest.

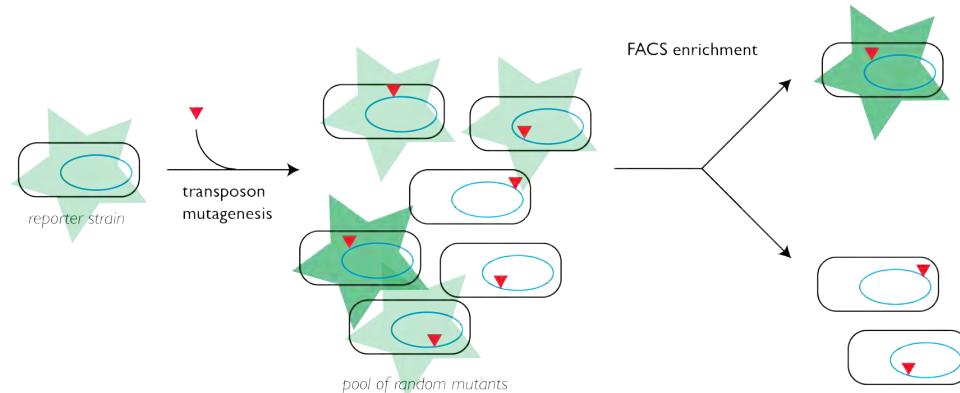


Next, this reporter strain is mutagenized with a transposon, generating a saturated mutant library, which is sequenced to identify the location of all of the transposon insertion sites in the unsorted library (as usual for Tn-seq). Then, the mutant library is separated by FACS, pooling the bacteria that are either much more or much less fluorescent than the average. This is a powerful high-throughput enrichment technique.



The mutated population expressing a fluorescent reporter is passed through a FACS machine, set to collect cells which are either in the top or bottom 5% of fluorescence, enriching for cells with mutations in regulators of the promoter controlling fluorescent protein expression.

Finally, the high- and low-expression mutant pools are sequenced and the population of transposon insertion sites compared to that in the unsorted control, as in a regular Tn-seq procedure, resulting in lists of genes whose disruption either increases or decreases the expression of the gene of interest.



Hypothesis: There are multiple transcriptional regulators of *shlBA* in *S. marcescens*.

Experimental Design:

- I) Construct a P_{shlBA} -*gfp* reporter strain.



2) Use SorTn-seq to identify genes whose disruption significantly changes P_{shlBA} expression.

Independent Variable: (what will you change?)

- The location of the transposon insertion in each strain

Dependent Variable(s): (what will you measure?)

- GFP fluorescence in each cell

Negative Control(s): (eliminate false positive results)

- FACS sorting of an unmutated $P_{shlBA}-gfp$ reporter strain will be important for establishing the natural distribution of expression from the P_{shlBA} promoter, which is not likely to be completely identical from cell to cell.

Positive Control(s): (eliminate false negative results)

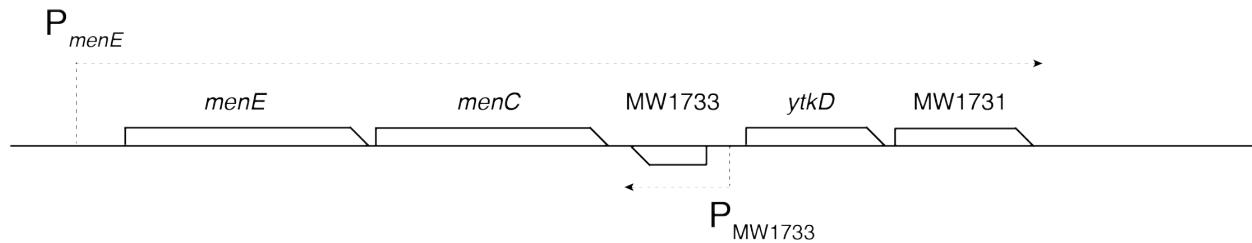
- 1) Sequencing of the unsorted mutant pool will identify genes which cannot be disrupted or are simply not represented in the library.
- 2) If a *shlBA* regulator is already known, then a mutant in that gene would be a useful positive control as well.

Potential Outcomes:

- 1) Mutants are identified which increase or decrease $P_{shlBA}-gfp$ reporter expression beyond the limits of the natural variation in P_{shlBA} expression. This supports the hypothesis, but it will be important to validate these candidates by making individual mutants and directly testing expression of *shlBA* in those strains.
- 2) No mutants are identified which increase or decrease $P_{shlBA}-gfp$ reporter expression beyond the limits of the natural variation in P_{shlBA} expression. This could indicate that *shlBA* are not regulated at the transcriptional level (see below) or that any transcriptional regulators of *shlBA* are either essential or otherwise not represented in the transposon library.

DISCUSSION PROBLEM SET #9: TRANSCRIPTIONAL REGULATION

There is an unusual genetic locus called a “non-contiguous operon” in *Staphylococcus aureus*, involving 5 genes associated with menaquinone biosynthesis:



The *menE*, *menC*, *ytkD*, and *MW1731* genes are all encoded on a single polycistronic mRNA. The *MW1733* gene, located between *menC* and *ytkD* on the opposite strand, has its own promoter and is encoded on its own monocistronic mRNA. Both *MW1731* and *MW1733* encode conserved hypothetical proteins with no known functions.

Under conditions where the *MW1733* mRNA is expressed, the amount of *menE-menC-ytkD-MW1731* mRNA decreases. Consistent with this, replacing P_{MW1733} with a strong constitutive promoter (increasing transcription of *MW1733*) dramatically reduces the amount of *menE-menC-ytkD-MW1731* mRNA.

- 1) Propose a hypothesis to explain the regulation of the *menE-menC-ytkD-MW1731* operon by *MW1733*.
 - 2) Design a genetic experiment that will test your hypothesis. All standard genetic tools are available to manipulate *S. aureus*.
State:
 - the independent and dependent variables
 - both positive and negative controls
 - potential outcomes of your experiment, and how you will interpret them
-

REGULATION OF PROTEIN LEVELS

Similarly to mRNA, protein levels in a cell are controlled at the level of both production and degradation. Translation can be regulated at the *initiation* or *elongation* stages, and protein stability is controlled by the activity of protein-degrading enzymes called *proteases*.

Similar considerations must be taken into account when considering protein regulation as when considering mRNA. Regulation of translation requires the cell to maintain a pool of mRNA that it does not need to transcribe before producing protein, potentially speeding regulatory response. Cellular levels of unstable proteins can be changed much more quickly than stable ones can, and regulated proteolysis is a fast and irreversible way to stop a particular protein from carrying out its function in the cell.

5. Translation initiation. The first step in translation is binding of the 16S ribosomal subunit to the Shine-Dalgarno (S.D.) sequence (also known as a *ribosome binding site* or *RBS*) upstream of the start codon in an mRNA. The sequence of the 3' end of the 16S rRNA (the *anti-Shine-Dalgarno sequence*) of *E. coli* is 5' **ACCUCCUUA** 3', and therefore the consensus sequence for S.D. sites in *E. coli* is 5' AGGAGGU 3', which base pairs with the bolded region of the 16S rRNA. The more similar a gene's RBS is to the consensus, the more efficiently ribosomes will bind to that site, and the more efficiently translation will be initiated. Each gene in a polycistronic mRNA typically has its own RBS, meaning that different proteins encoded on the same RNA can be translated at different rates.

Note that every organism's 16S rRNA sequence is different, and therefore the consensus RBS in each species is also different, although bacterial S.D. sequences are typically G/A rich.

Several types of regulation work by changing the *accessibility* of the RBS. There are proteins that compete with ribosomes for binding to mRNAs, and a variety of factors that can change the structure of the mRNA to make it more or less accessible to ribosome binding. These include riboswitches which fold to expose or hide the RBS when bound to metabolites, sRNAs which base pair with the RBS or change the folding of the 5' UTR, and structural features of the mRNA itself which can conceal or expose the RBS in response to changing conditions. A straightforward example of this are thermosensors in which the RBS forms part of a stem-loop structure at low temperature which unfolds at higher temperatures (found, for example in the virulence-associated *prfA* transcript from *Listeria monocytogenes*).

Translation initiation

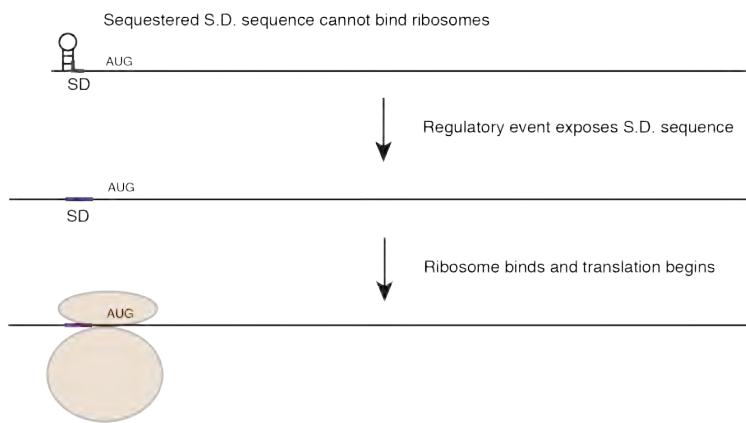


Figure 4.7. An example of how RBS accessibility can regulate translation initiation.

The identity of the start codon also has a strong effect on translation initiation. Most protein-coding gene sequences begin with an AUG codon, but some begin with GUG or UUG and are therefore less efficiently translated. The first amino acid of a protein is always formyl-methionine (fMet), regardless of the nucleotide sequence of the start codon.

6. Translation elongation. The rate of elongation by ribosomes is determined by a number of factors, but the most important one for regulating the relative amounts of protein produced from different transcripts is *codon usage*. While most organisms contain tRNAs capable of translating all of the possible amino acid-encoding codons, different tRNAs are not all present in the same concentrations. A gene with many *rare codons* will not be translated efficiently, since the ribosome will need to pause frequently to wait to encounter an appropriate charged tRNA. Different species have different codon usage patterns, but for example, in *E. coli* the arginine codons AGG and AGA are very rare, and an mRNA with these codons will not be translated efficiently (and is likely to be prone to premature transcriptional termination or degradation, as described above). In many bacteria, there are a few apparent stop codons that are actually translated as the unusual amino acids selenocysteine or pyrrolysine. The sequence properties that make a particular stop codon be translated this way are not completely understood.

[Recent work](#) in *Streptomyces* spp. suggests that some bacteria have the ability to produce alternatively-charged tRNAs under some conditions (e.g. attaching a proline amino acid to an alanine tRNA), meaning that they can regulate how some codons are translated. How common this is and how it affects the proteome as a whole remain to be determined. There is also [evidence](#) that certain combinations of adjacent codons are particularly poorly translated, possibly due to steric clashes between tRNAs in the A and P sites of the ribosome, but the rules determining what combinations those are have not yet been well defined.

7. Protein stability. Protein stability is determined by cytoplasmic proteases, which themselves are tightly regulated to prevent uncontrolled degradation of cellular proteins. They are typically large multi-protein complexes with barrel-like structures. The active sites are inside the barrel, sequestered away from most protein substrates.

In *E. coli*, the primary ATP-dependent proteases are ClpP (in complex with either ClpA or ClpX), Lon, HslUV, and FtsH. These are widely conserved, but some other bacteria have different protease complexes, such as the bacterial proteasome, which is found only in actinobacteria and is homologous to the eukaryotic proteasome. Each protease has different substrate specificities, although they often overlap extensively. Proteases recognize signal sequences (degrons) in their target proteins, and the presence or accessibility of degrons in a protein will determine which proteases degrade it. Lon, for example, recognizes aromatic amino acids that are normally buried in the hydrophobic core of proteins, and is therefore an important protease for degrading unfolded or damaged proteins. The ClpA and ClpX adaptor proteins recognize different degrons and target them for degradation by ClpP.

The N- and C-terminal ends of proteins often contain degron sequences that determine their stability. The N-end rule describes a phenomenon in which the N-terminal amino acid(s) of a protein have a dramatic effect on that protein's degradation. Proteins which still have their N-terminal formyl-methionine (fMet) residue are degraded more quickly (probably by FtsH) than those in which fMet has been removed. Proteins with N-terminal leucine, tyrosine, tryptophan, or phenylalanine residues are recognized by ClpS and then degraded by ClpAP. Proteins with N-terminal arginine, lysine, or methionine residues can have N-terminal phenylalanine residues added by the L/F-tRNA-protein transferase, targeting them to the same system. Endopeptidases that cut within proteins can generate previously unexposed degrons.

Protein degradation

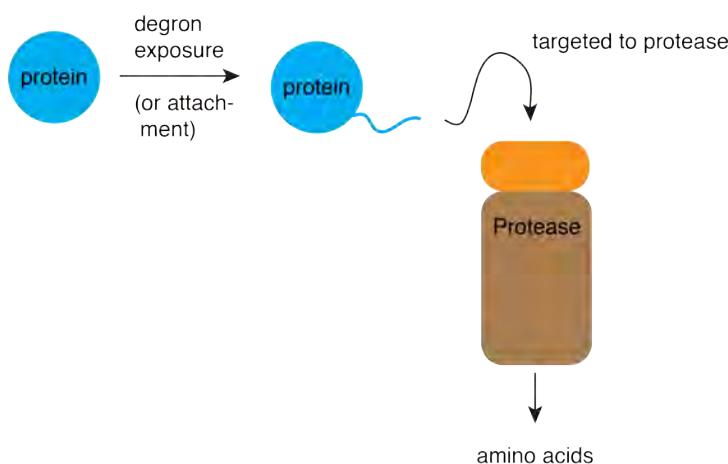


Figure 4.8. Targeting and degradation of proteins by protease complexes.

In eukaryotes, proteins destined for degradation by the proteasome are post-translationally modified by addition of ubiquitin. Bacteria do not contain ubiquitin, but actinobacteria (including *Mycobacterium* spp.) have a similar system in which they conjugate the small, intrinsically disordered Pup protein (**prokaryotic ubiquitin-like protein**) to lysine residues in proteins that are then targeted to a protease complex known as the bacterial proteasome. This *pupylation* system is only found in actinobacteria.

MEASURING GENE EXPRESSION: PROTEIN ABUNDANCE

There are multiple techniques available for measuring protein abundance in a cell. As for measurements of transcripts, they can be divided into **direct** measurements of protein and **indirect** measurements, and may be able to measure either expression of a single protein or a large fraction of the **proteome**.

	Single Protein	Proteome-Wide
Direct	western blotting ELISA	mass spectrometry 2-D gels
Indirect	translational reporter fusion	ribosome profiling

Techniques to directly measure the amount of a protein in a cell include *western blotting*, *mass spectrometry*, and *2-dimensional gel electrophoresis*.

Western blotting (also called *immunoblotting*) relies on antibodies specific to a particular protein to detect and quantify that protein, and is by far the most common method for measuring protein abundance in cells. Whole cell protein extracts can be spotted directly onto membranes or run on polyacrylamide gels and then transferred to membranes before detection by western and quantification by comparison to a standard curve of purified protein. There are numerous variations on using antibodies for protein detection, notably including *ELISA* (enzyme-linked immunosorbent assay), which allows high-throughput quantitation of particular proteins in complex biological samples. ELISA kits are available for many mammalian proteins of interest, but are not as commonly used or available for bacterial systems.

To quantify a protein by immunoblotting or ELISA, you must have a high-quality antibody, which is to say, one that is both sensitive and specific to the protein you want to detect. Generating such antibodies can be challenging and expensive (we recently spent about \$10,000 having a monoclonal antibody made). To get around this issue, there are a number of *epitope tags* that can be engineered into proteins to allow them to be detected with commercially available high-quality antibodies. Common examples include the HA-tag, derived from a fragment of the influenza virus hemagglutinin protein, the myc-tag, derived from human *c-myc* protein, the FLAG-tag, an entirely artificial antigen with the amino acid sequence DYKDDDDK, and the recently-invented C-tag and ALFA tag, which are detected by specific camel-derived *nanobodies*. Adding an epitope tag to a protein makes it far easier to detect, but careful controls must be done to make sure that the tag itself does not have an effect on the abundance or activity of the protein. It is also typically technically simpler to add epitope tags to proteins encoded on plasmids than to chromosomal genes, and effects of plasmid copy number, etc., must be taken into account in such experiments (see **Lectures 5** and **7** for more on plasmids and their engineering).

Proteomics studies attempt to quantify the abundance of all (or a large subset) of the proteins in a cell simultaneously. There are a very wide range of sophisticated methods to do this, but nearly all of them rely on *mass spectrometry* to identify proteins by their molecular weight. The details of how this works are well beyond the scope of this course, but in general, one weakness of this kind of approach is that proteomics is not able to reliably detect low-abundance proteins.

2-D gels are an older proteomics method that you will sometimes run across in the literature which used gel electrophoresis to separate proteins (often from cells fed radioactively-labeled amino acids) on large acrylamide gels in two stages. Proteins were first separated by size, and then by isoelectric point. This method is technically challenging and has been almost entirely supplanted by more modern techniques which give higher-quality data for less work, but did allow separation, visualization, and quantification of hundreds of separate proteins simultaneously.

(mostly) It is possible to indirectly measure the amount of a protein in a cell by using *translational reporter fusions*, which are closely related to the transcriptional fusions discussed earlier in this chapter. The difference is that instead of only including the promoter of the gene of interest, the entire upstream region of that gene, including the RBS and often several codons of the gene itself, is fused to the reporter gene. This makes expression of the reporter dependent on both the transcriptional and translational control signals associated with the gene of interest. All of the same caveats listed for transcriptional fusions apply to translational fusions, and it's important to remember that any translational signals (pause sites, rare codons, etc.) found within the coding sequence of the gene will not be present in the fusion.

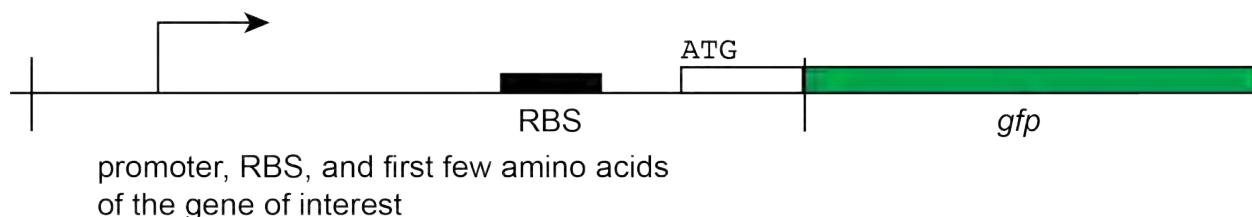


Figure 4.9. Translational fusions link the promoter, RBS, and first few codons of a gene of interest with an easily-measured reporter gene.

A relatively recent development is *ribosome profiling*, an indirect method to infer whole-genome protein translation. In this method, cells are treated with a chemical that reversibly crosslinks ribosomes to mRNA. The ribosome-mRNA complex is purified, treated with RNase to degrade any RNA that is not protected by ribosome binding, and then the crosslinking is reversed to obtain the pool of protected mRNA fragments. Next-generation sequencing is used to compare the ribosome-bound RNA fragments to the total mRNA pool, which quantifies the proportion of any given mRNA that is ribosome-

bound at the moment of measurement. The presence of ribosomes on an mRNA is interpreted as a measure of the amount of translation of that mRNA, and therefore a readout for the amount of that protein being produced.

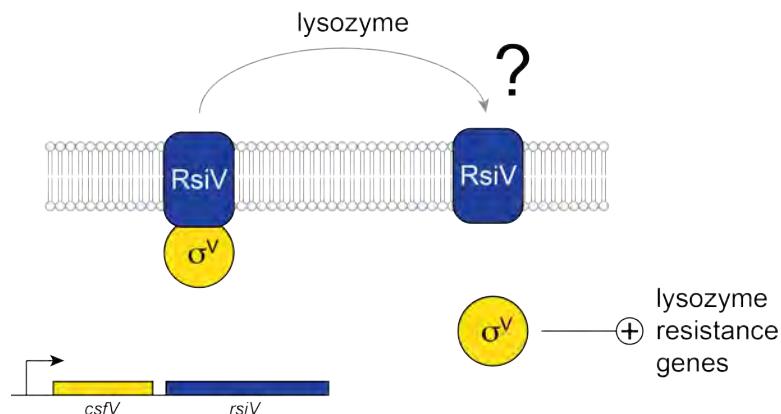
Protein stability can be difficult to measure independently from synthesis. One common approach is to add a translation inhibitor (such as the antibiotic chloramphenicol) to cells and then measure the abundance of a particular protein over time. This has the disadvantage, of course, of having serious effects on cellular physiology in general.

A second, older approach to measuring protein stability is a *pulse-chase* experiment, which, in its original form, involves adding radioactively labeled amino acids to a cell for a short period of time, then replacing them with unlabeled amino acids and tracking how long the radioactively labeled proteins produced during that pulse are maintained in the cell. More sophisticated labeling or immunodetection techniques can be used to focus pulse-chase experiments on a single protein or set of proteins.

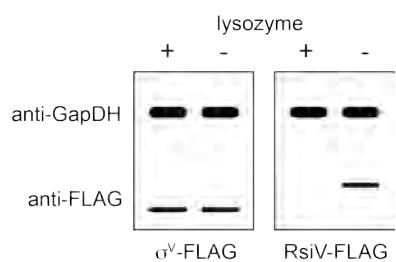
EXAMPLE OF AN EXPERIMENTAL DESIGN: REGULATION OF PROTEIN ABUNDANCE

Clostridioides (formerly Clostridium) difficile is a dangerous pathogen that causes persistent and difficult-to-treat intestinal infections in patients who have received antibiotics. It is a common infection acquired in hospitals, and can be fatal to elderly or immunocompromised people.

Among *C. difficile*'s many virulence factors are a variety of different genes involved in resisting the toxic effects of lysozyme, an enzyme secreted in saliva and other bodily fluids that breaks down bacterial cell walls. Many of the lysozyme resistance genes of *C. difficile* are regulated by an alternative sigma factor called σ^V (encoded by the *csfV* gene), and those genes are only expressed in the presence of lysozyme. A gene for a predicted membrane-bound anti-sigma factor (called *rsiV*) is adjacent to *csfV* in the *C. difficile* genome.

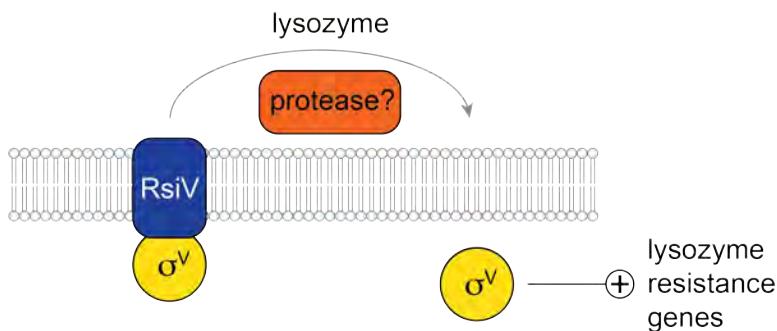


You construct mutants of *C. difficile* in which the chromosomally-encoded σ^V or RsiV proteins are fused to FLAG tags and use those strains to measure protein abundance in the presence and absence of lysozyme by quantitative Western blot with commercially-available anti-FLAG antibodies.



Western blots performed on *C. difficile* expressing either FLAG-tagged σ^V or RsiV in the presence and absence of lysozyme. GapDH is a metabolic enzyme that is commonly used as a loading control for normalization of quantitative Western blots in bacteria.

You observe that the levels of σ^V protein do not change in response to lysozyme, but that RsiV is no longer present after addition of lysozyme. Based on these data, you propose a model in which RsiV is proteolytically cleaved in the presence of lysozyme, releasing σ^V and activating transcription of lysozyme resistance genes.



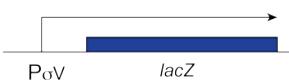
One notable gap in this model is that we do not know what protease(s) might be responsible for RsiV cleavage. Design an experiment to identify the RsiV protease.

Hypothesis: The genome of *C. difficile* encodes a protease that cleaves RsiV in the presence of lysozyme.

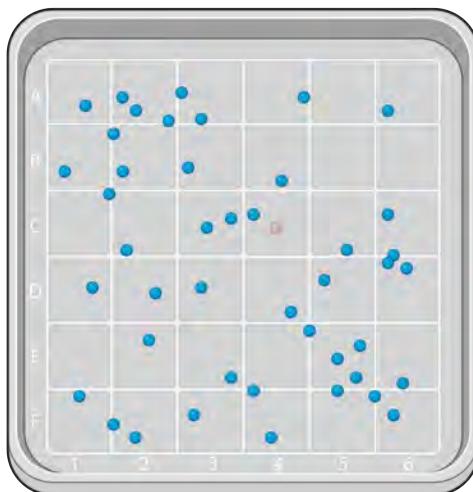
We could just make a list of all of the predicted proteases in the *C. difficile* genome and knock them out one by one, hoping to get lucky (a reverse genetics approach), but an unbiased mutant hunt (i.e. forward genetics) is probably a more elegant way to approach the problem.

Experimental Design:

- 1) Construct a strain of *C. difficile* with *lacZ* under the control of a σ^V -dependent promoter, ideally integrated in single copy in the chromosome. This strain should express *lacZ* in the presence of lysozyme, but not in the absence of lysozyme.



- 2) Construct a saturated transposon library of the *C. difficile* $P\sigma^V-lacZ$ strain.
- 3) Screen the library for mutants that do not express *lacZ* on plates containing lysozyme.



Mutated reporter strain grown on agar medium containing lysozyme and the colorimetric LacZ substrate X-Gal. Note the colony at C4 that is not expressing LacZ and is therefore not blue. Made with Biorender.

Expect to need to screen about 30,000 colonies, which sounds like a lot, but since you can easily distinguish and isolate a white colony from among several hundred blue ones, this needs only about a hundred plates (4 sleeves-full, or about 2 liters of agar). A little tedious, but not unreasonable. Hits will include any mutant which cannot degrade RsiV. Our hypothesis is that this will be a protease, but this design will identify other genes with similar effects on σ^V abundance or activity, if any exist.

Independent Variable: (what will you change?)

- The position of the transposon insertion in each tested strain.

Dependent Variable(s): (what will you measure?)

- $P\sigma^V-lacZ$ expression on lysozyme-containing media (a qualitative measurement, at least when looking at colony color on a plate, although LacZ activity certainly can be measured quantitatively in a pure culture grown in liquid media)

Negative Control(s): (eliminate false positive results)

- 1) Confirm that an un-mutated $P\sigma^V-lacZ$ strain has no LacZ activity on media without lysozyme.
- 2) Use PCR or sequencing to eliminate Lac- strains in which the transposon has disrupted the *lacZ* gene.

Positive Control(s): (eliminate false negative results)

- 1) Confirm that an un-mutated $P\sigma^V-lacZ$ strain has LacZ activity on media containing lysozyme.
- 2) Confirm that a $\Delta csfV$ mutant can grow on media containing a concentration of lysozyme capable of inducing σ^V -dependent gene expression. If it can't, then this experiment will not work, since the desired mutants will be unable to express σ^V -dependent genes.
- 3) Confirm that a $P\sigma^V-lacZ \Delta rsiV$ strain has LacZ activity on media both with and without lysozyme.

Potential Outcomes:

- 1) You identify transposon insertion mutants in one or more protease-encoding genes that inhibit expression of the $P\sigma^V-lacZ$ reporter. It will be important to validate these candidates by measuring RsiV abundance in the mutant strains in the presence and absence of lysozyme (by Western blot).
- 2) You identify transposon insertion mutants in genes that do not encode proteases, but which inhibit expression of the $P\sigma^V-lacZ$ reporter. These are likely to be novel regulators of σ^V , and will need further study. Update your model to include them.
- 3) You are unable to identify transposon mutants that inhibit expression of the $P\sigma^V-lacZ$ reporter. This may mean that RsiV abundance is regulated by redundant or essential proteins, or that RsiV itself is the protease (auto-cleavage is not unheard of in regulatory proteins).

REGULATION OF PROTEIN ACTIVITY

8. Protein activity. The amount of a particular protein in a cell does not necessarily determine the level of activity of that protein. Many proteins' activities vary depending on the concentration of metabolites within the cell or can be regulated by covalent modifications or by physical interactions with other cellular components. These regulatory events can be more difficult to measure *in vivo* than changes in the amount of mRNA or protein.

Regulation of protein activity is the fastest, most agile mode of regulation available to the cell, since all of the components needed are already present. However, producing inactive proteins does require the same resource expenditure as producing active ones does.

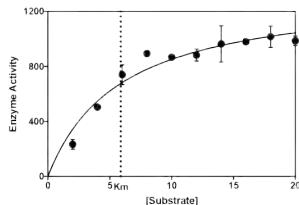
The rest of this section is more biochemistry than genetics, but I think it's important to understand how all of these layers work together in living cells. It is certainly possible to isolate mutations in proteins that effect the regulation of protein activity, and we will discuss the interpretation of such mutations in class.

The nature of enzyme kinetics means that the activity of enzymes varies depending on the concentrations of their substrates and products. The reaction rate of a reversible enzyme operating close to thermodynamic equilibrium can change dramatically or even reverse in response to modest changes in the ratio of substrates and products. Large changes in the activity of an enzyme can result from quite small changes in substrate concentration for enzymes whose K_m (Michaelis constant; the concentration of substrate at which reaction rate V is half of V_{max}) is close to the concentration of substrate found in the cell. Many of the enzymes of central metabolism have these properties, and flux through these pathways therefore rapidly responds to changes in conditions without any changes in gene expression. This mode of regulation is, however, somewhat wasteful, since an enzyme operating near its K_m cannot, by definition, be working at its maximum efficiency, and enzymes operating near thermodynamic equilibrium will spend most of their time catalyzing exchange reactions between substrates and products with no net flux in one direction or the other.

Kinetic Regulation

1. Substrate concentrations near K_m

$$\text{rate (V)} = \frac{V_{\max} \times [\text{Substrate}]}{K_m + [\text{Substrate}]}$$



2. Reaction near thermodynamic equilibrium



3. Product or substrate inhibition



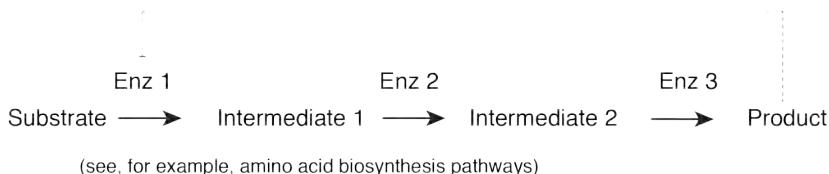
Figure 4.10. A variety of ways enzymes can be regulated by their biochemical properties.

Many enzymes are competitively inhibited by their products and some are inhibited by their substrates, providing additional layers of kinetic control that can affect enzyme activities. This kind of inhibition usually occurs by competition for binding in the active site of the enzyme.

Allostery is a regulatory mechanism by which a molecule controls protein activity by non-covalently binding to a site that is **not** the active site of that protein. **Allosteric effectors** can activate or inhibit protein activity, and are generally thought to function by causing changes in the structure of the protein. Allostery is particularly common in metabolic enzymes. For example, the first enzyme of a complex biosynthetic pathway is often allosterically inhibited by the final product of that pathway, ensuring that the pathway will be inactive when enough of the product is present in the cell.

Allostery

1. Metabolic regulation



2. Second messengers

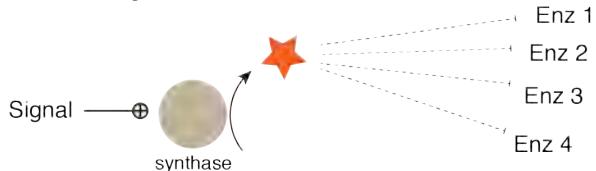


Figure 4.11. Examples of allosteric regulation of protein activity by small molecules.

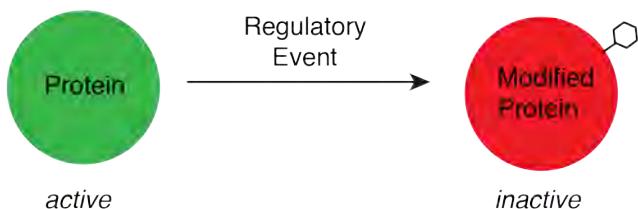
There are many examples of allosteric regulation by second messengers, which are small signaling molecules produced under certain conditions. The second messenger then diffuses through the cytoplasm and regulates the activity of proteins throughout the cell. Many second messengers are derived from nucleotides, and important examples include cyclic AMP (cAMP), cyclic di-GMP (c-di-GMP), and guanosine tetraphosphate (ppGpp). Second messengers regulate complex processes in cellular stress response and development, and typically have multiple enzymes controlling their synthesis and degradation.

Allostery is also important for many transcription factors, whose DNA-binding activity or interactions with RNA polymerase are changed when they bind to the specific small molecules they sense. Riboswitches are an example of allosterically-controlled regulators which are not proteins.

Covalent modifications (often called *post-translational modifications* or PTMs) can also affect protein activity.

A very common mechanism of PTM regulation is phosphorylation. Serine, threonine, tyrosine, histidine, aspartate, arginine, and (very rarely) cysteine residues can be *phosphorylated*, reversibly adding an ATP- or GTP-derived large negatively charged phosphate group that can dramatically affect protein structure and activity. These are controlled by specific *kinases* and *phosphatases*, which are enzymes that add or remove phosphate groups, respectively, and which often function in signaling pathways. See [this recent review](#) for more details on this important signalling mechanism.

Post-Translational Modification

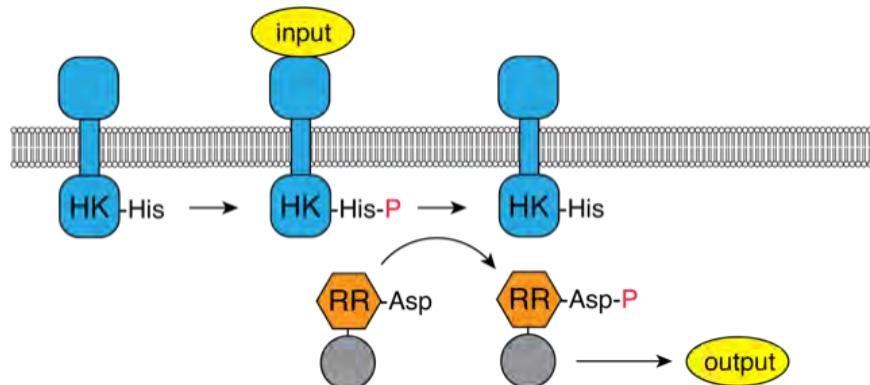


(modifications can activate **or** inactivate proteins, or have other effects on their structure, function, localization, etc.)

Figure 4.12. Protein activity can be changed by covalent modification of the protein. These modifications are usually the result of the activity of other enzymes, which are, in turn, typically regulated by one or more of the mechanisms discussed in this chapter.

DISCUSSION PROBLEM SET #10:TWO COMPONENT REGULATORS

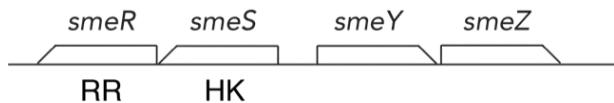
A very common family of regulators in bacteria that use post-translational phosphate modifications to regulate protein activity are the so-called **two-component systems** (or TCS). An archetypical TCS consists, as you might expect, of two proteins, one with a **sensor domain** fused to a **histidine kinase (HK)** domain and a second with an **output domain** fused with a phosphate-accepting **response regulator (RR)** domain. Many HKs, but not all, are integral membrane proteins with their sensor domains on the outside of the cell, allowing them to detect changes in the exterior environment.



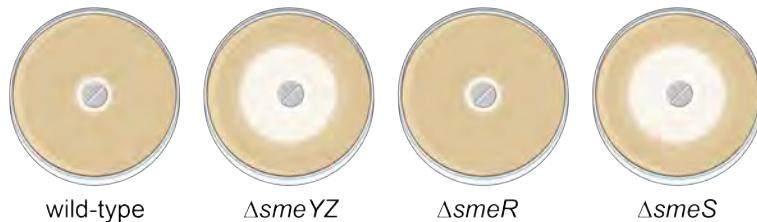
When the sensor domain of the HK detects its cognate input signal (whatever that may be), the HK uses ATP to autophosphorylate itself on a conserved histidine residue, then transfers that phosphate to a conserved aspartate residue in the RR. Phosphorylation affects the activity of the RR, which is often (but not always) a DNA-binding transcription factor. Many so-called TCS actually consist of *phosphorelays* with more than two components, including multiple HK and RR domain containing proteins.

TCS are involved in regulating responses to a wide variety of environmental signals. The number of TCS systems encoded by the genome of any particular species depends on the complexity of that species' environment: obligately intracellular pathogens like *Anaplasma* may have none at all, while free-living bacteria with complex developmental programs (like *Myxococcus* or *Nostoc*; see [Lecture 11](#)) can have close to 200.

Stenotrophomonas maltophilia is a bacterium that is ubiquitous in the environment and occasionally causes opportunistic infections in humans. This is problematic, because *S. maltophilia* is naturally resistant to nearly all antibiotics. While studying antibiotic resistance in this organism, you identify a locus containing homologs of a TCS (encoding a predicted RR called SmeR and a predicted HK called SmeS) divergently transcribed from a predicted small-molecule export pump (SmeYZ).



The exact functions of SmeY and SmeZ are not known, but a Δ smeYZ mutant is more sensitive to aminoglycoside antibiotics than the wild-type, as is a Δ smeS mutant. A Δ smeR mutant has the same antibiotic sensitivity as the wild-type.



A disc diffusion assay for antibiotic sensitivity. Filter discs containing antibiotics are set on top of a petri dish that has been seeded with bacteria. A larger zone of inhibition indicates that the strain of bacteria is more sensitive to the antibiotic. Made with Biorender.

You use qRT-PCR to measure the transcription of *smeY* in media containing antibiotics to determine whether SmeRS is involved in regulating the *smeYZ* operon. You get the following results:

Mutation	<i>smeY</i> mRNA
wild-type	+
Δ smeR	+
Δ smeS	-
Δ smeRS	+

Based on these results and your knowledge of how TCS regulators work, propose a model of how SmeRS regulates the *smeYZ* operon. In your model, is SmeR an activator or a repressor? How do you predict that phosphorylation by SmeS affects the activity of SmeR?

N-acylation is the conjugation of acyl groups (acetyl-, propionyl-, succinyl-, etc.) to lysine residues by acyltransferases, and plays an important role in controlling metabolic enzymes. To give one illustrative example, in *Salmonella enterica*, acetyl-CoA synthase (Acs) is inactivated by acetylation (by the Pat acetyltransferase, which uses acetyl-CoA as a substrate) when acetyl-CoA levels rise in the cell. When acetyl-CoA levels drop, acetylation of Acs is reversed by the activity of the CobB sirtuin deacetylase, reactivating it for acetyl-CoA synthesis. Proteins also can be reversibly methylated, which plays a notable role in controlling the activity of proteins involved in chemotaxis (Lecture 15). Note that, of course, the activity of the modification and demodification enzymes for each of these mechanisms must themselves be regulated.

Oxidative modifications of cysteine or methionine residues are common regulators of protein activity in response to changes in redox conditions. Cysteine is normally found in a reduced thiol state (-SH), and can be reversibly oxidized to sulfenic acid (-SOH) or, if two cysteines are in close proximity to each other, to a disulfide bond (-S-S-), either of which can dramatically affect the structure and activity of a protein. Cysteine residues can also be covalently modified by electrophilic compounds. Reductase systems, including those involving glutathione and thioredoxin, can reduce these oxidized forms of cysteine back to the thiol form. In other proteins, oxidation of methionine to methionine sulfoxide regulates activity, and can be reversed by the activity of methionine sulfoxide reductases.

Most PTMs are reversible, but some regulatory events are irreversible. Cysteine can be oxidized irreversibly to sulfenic ($-\text{SO}_2\text{H}$) or sulfonic ($-\text{SO}_3\text{H}$) acid, and the *Bacillus subtilis* transcription factor PerR responds to peroxide stress via the irreversible oxidation of a histidine residue. Presumably, the resulting permanently inactive proteins are subsequently degraded by proteases. Arguably, any modification that leads to proteolysis is an irreversible PTM.

Finally, a protein's activity can be controlled by physical interactions between the protein and other components of the cell, including proteins, ribosomes, DNA, or the cell membrane. This is a kind of allostery, since no covalent modifications of the proteins are involved, and the interaction surface is often not the active site. Some proteins are only active when they are in complex with other proteins, and the formation of these complexes can be regulated by the mechanisms described above. In other cases, proteins can be sequestered in an inactive state by interactions with other cell components, and become active only when they are released from these interactions. Anti-sigma factors, discussed above, work in this way.

MEASURING GENE EXPRESSION: PROTEIN ACTIVITY

Measuring protein activity is a very direct way to assess the function of a gene product, but can be technically challenging. The techniques required depend on the function of the gene product in question, and differ for every specific protein. There are, however, some general categories of assays which are commonly used, and which I will describe below. One key consideration for protein activity assays is whether they can be performed *in vivo* or if they require the *in vitro* analysis of purified proteins or cell lysates. *In vivo* activity measurements are affected by both how active a given protein is and how abundant it is in the cell, while *in vitro* assays typically allow much simpler normalization for protein abundance.

Enzyme activity assays are the most direct way to assess whether an enzyme (a protein that catalyzes a chemical reaction) is active in a cell or not, but how easy this is to measure depends entirely on the particular enzyme in question. Some enzymes are very simple to assay. Many are not. This is not a biochemistry class, so we won't go into tremendous detail here, but when you're thinking about measuring the activity of an enzyme, consider the following:

1) Is the enzyme cytoplasmic, periplasmic, secreted, or membrane-bound?

Alkaline phosphatase (PhoA) in *E. coli* is a surface-exposed enzyme for which a colorimetric substrate is available, so cells can simply be resuspended in buffer for measurements of PhoA-dependent accumulation of a yellow product. Cytoplasmic enzymes (like LacZ) may need cell permeabilization to allow substrate access. Membrane proteins might or might not retain activity when solubilized with detergents. A secreted protein might need to be concentrated from the spent growth medium of the culture.

2) What are the substrate(s) of the enzyme, and how can you measure them?

How can you measure the conversion of substrate into product? Are they different colors? Do they have different absorbance or fluorescence properties? Can they be separated by chromatography? Are there substrate analogs available that are easier to measure than the physiological substrate? (This is what the commonly used indicator substrate X-Gal is; a colorimetric analog of lactose that turns blue when cleaved by LacZ.)

3) Are there other enzymes in the cell that act on the same substrate(s)?

Many cellular enzymes act on common substrates, like ATP or NADH. Trying to measure the activity of this kind of enzyme *in vivo* or in a complex mixture of proteins is not possible due to interference from other enzymes. You will need to purify the protein and study it *in vitro*.

4) How fast does the enzyme act? Do the products accumulate *in vivo*? Are they stable *in vitro*?

Some enzymes catalyze very slow reactions, others turn over in milliseconds. Both situations make it difficult to measure the activity of the enzyme accurately. Some enzyme products are immediately consumed in cells by the next enzyme in a biosynthetic pathway, making it impossible to measure the synthesis of those products *in vivo* without special interventions. If the product of an enzyme is chemically unstable, it will also be difficult to measure *in vitro* unless it can be trapped or stabilized somehow.

5) How stable is the enzyme?

Some purified enzymes are highly stable. Others lose activity rapidly *in vitro*. It's generally a good idea to keep enzymes cold, but some will lose activity when frozen. Reducing agents and metal chelators are often added to enzyme storage buffers to prevent oxidation and inhibit contaminating proteases, respectively, but can be problematic for enzymes with metal cofactors (roughly a third of all enzymes). *In vivo*, you have much less control over protein stability, although of course, the cell itself has more control, which can be a form of regulation in and of itself, as discussed above.

Not every function of a protein necessarily has the same stability. The catalytic activity of one of the enzymes in which allosteric regulation was first discovered ([aspartate transcarbamylase](#)) is stable at 4°C, but its allosteric regulation by CTP is not. When the protein is refrigerated, it loses the ability to be regulated by CTP. This differential stability, once it was figured out, was part of the evidence that allosteric regulators do not act by binding to the active sites of enzymes.

6) How easy is the enzyme to purify?

When purifying proteins for *in vitro* studies, there are potential problems at each of several steps. Can the protein be over-expressed without toxic effects on the cell? Is the protein soluble? Membrane proteins are never soluble, and there are different detergents and lipid analogs available that can be used to try to keep them in solution. Will the protein tolerate having an affinity chromatography tag (e.g. 6xHis or GST) fused to it? If so, does the tag need to be removed after purification in order for the protein to be active (typically by treatment with a specific protease, if the tag has been designed to allow this)? If not, how can you separate the protein from other cellular proteins without a tag?

There are lots of resources to help you troubleshoot protein expression and purification. [This paper](#) does a reasonable job of introducing some of the key concepts.

Allosteric and kinetic regulation is typically easiest to measure for proteins that can be purified and assayed *in vitro*, but are generally challenging, especially if you don't know what the small molecule regulators might be.

Not all proteins in a cell are enzymes. The activity of DNA- or RNA-binding proteins (like transcription factors) can be measured by a variety of methods, some of which take advantage of modern high-throughput sequencing technology. Purified DNA binding proteins can be mixed with different DNA fragments to see if they interact *in vitro*. The most common method uses gel electrophoresis to separate unbound DNA from protein-bound DNA, which migrates more slowly. This is called an **electrophoretic mobility shift assay** (EMSA). ChIP-seq (**chromatin immunoprecipitation sequencing**) is an *in vivo* technique to identify all of the genomic binding sites of a DNA binding protein. In a ChIP-seq experiment, cells are treated with a chemical to crosslink proteins and DNA, the DNA is fragmented, then an antibody to a particular DNA-binding protein of interest is used to pull down only those fragments of DNA bound to that protein. The resulting pool of DNA fragments is sequenced with next-generation sequencing and compared to the entire genome sequence.

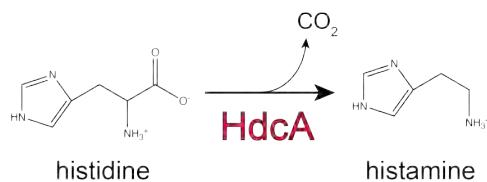
Most PTMs are detectable by mass spectrometry, although some can be detected by other means (antibodies, radioactive tracers, etc.). This is, of course, simplest with purified proteins, but can often be done in a high-throughput way in the course of a mass spectrometric proteomics experiment, if the protein is abundant enough.

Protein-protein interactions can be measured both *in vivo* and *in vitro*, although *in vitro* techniques with purified proteins are much more likely to give quantitative measurements of binding affinity. Two-hybrid assays are clever *in vivo* screens (or sometimes selections) that link protein-protein interactions to easily measured phenotypes. They typically involve generating plasmids with fusions between the proteins of interest and two halves of a protein that has a measurable activity when brought within close proximity to each other. This could be an enzyme or, in the case of the most common yeast two-hybrid system, a transcription factor. Libraries of different proteins fused to these kinds of reporters have been used to generate maps of all of the two-way protein-protein interactions in various kinds of cells. We will revisit these techniques in more detail in **Lecture 13**.

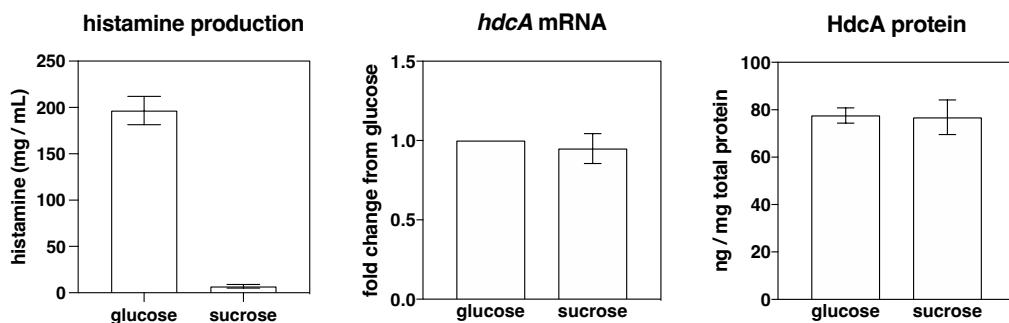
Metabolomics uses mass spectrometry to measure the concentration of molecules in cells that are not proteins or nucleic acids (**metabolites** or small molecules). This can be especially useful to assess how much *metabolic flux* is passing through different pathways, by quantifying the amount of each substrate, intermediate, and product that accumulates under different conditions, although it does require that all of the metabolites of interest be 1) fairly abundant and 2) of a known molecular weight different from other metabolites. As protein levels and activity change, metabolic flux and metabolite concentrations will shift, reflecting changes in cellular metabolism.

DISCUSSION PROBLEM SET #11: POST-TRANSCRIPTIONAL REGULATION

You isolate a *Lactobacillus reuteri* strain from the microbiome of a mouse and discover that it synthesizes the anti-inflammatory compound histamine when grown in media containing glucose, but not in media containing sucrose. This is dependent on the presence of the *hdcA* gene, which encodes histidine decarboxylase, an enzyme that converts the amino acid histidine to histamine.



You would like to understand how histamine synthesis is regulated. You measure histamine production in cultures of *L. reuteri* grown with glucose and sucrose as carbon sources, then do qRT-PCR to measure *hdcA* mRNA levels and quantitative Western blots to measure HdcA protein levels in those cultures, with the following results:



Based on these data and your knowledge of regulation, propose a hypothesis to explain the mechanism of regulation of histamine synthesis in response to sucrose. Design an experiment that would allow you to test your hypothesis. State:

- the independent and dependent variables
 - both positive and negative controls
 - potential outcomes of your experiment, and how you will interpret them
-

CONCLUSIONS

The take-home message from this section is that regulation of bacterial genes can be very complex and that bacteria can regulate multiple steps between gene expression and protein function. Even whole-cell measurements of mRNA levels, protein levels, or enzyme flux only tell part of the story, which is very important to remember when designing and interpreting experiments.

LECTURE 5: PLASMIDS

INTRODUCTION

In this lecture, we will discuss correlation and causation, and how to design experiments that establish causal relationships. Because these kinds of experiments in bacterial systems very often use plasmids, we will spend most of this chapter discussing what plasmids are and how they are used in different experimental applications.

SCIENTIFIC PROCESS 5: CORRELATION AND CAUSATION

It is extremely important to be clear about the difference between phenomena that are *correlated* with each other and phenomena that *cause* other phenomena. How can we distinguish between these experimentally?

In the earliest days of microbiology, there was a very serious debate about whether the microbes found in diseased humans and animals were the **cause** of disease or a **symptom** of disease. A great many observations were made and bitter arguments were had over the course of decades, until Robert Koch was finally able to settle the issue with a series of experiments based on what have come to be known as Koch's Postulates:

1. A specific microbe must be found in abundance in all host organisms suffering from the disease, but should not be found in healthy hosts.
2. The microbe must be isolated from a diseased organism and grown in pure culture.
3. The cultured microbe should cause the same disease symptoms when introduced into a healthy host.
4. The microbe isolated from inoculated host must be identical to the originally isolated microbe.

Koch used these postulates to prove that *Bacillus anthracis* was the causative agent of anthrax in 1884, and he and his coworkers spent much of the next 30 years following essentially this process to identify and isolate the bacterial pathogens that cause various diseases (including cholera, diphtheria, tetanus, typhoid fever, pneumonia, and bubonic plague, among others).

The key aspect of Koch's Postulates that allows the scientist to establish *causality* is the careful addition and subtraction of a single independent variable, in this case a specific microbe. In step 1, a correlation between microbe and disease is established, and then steps 2 - 4 demonstrate that adding **only** that microbe to a healthy host organism leads to development of the same disease. Similar principles can be applied to a wide variety of scientific questions.

In 1988, Stanley Falkow proposed a set of Molecular Koch's Postulates which he applied to the problem of figuring out whether particular genes contribute to the pathogenesis of disease-causing microbes, and which are more directly relevant to this course. Falkow's Postulates (first described in [this paper](#)) are:

1. The phenotype or property under investigation should be associated with pathogenic members of a genus or pathogenic strains of a species.
2. Specific inactivation of the gene(s) associated with the suspected virulence trait should lead to a measurable loss in pathogenicity or virulence.
3. Reversion or allelic replacement of the mutated gene should lead to restoration of pathogenicity.

He also included the alternative steps:

- 2A. The gene(s) associated with the supposed virulence trait should be isolated by molecular methods. Specific inactivation or deletion of the gene(s) should lead to loss of function in the clone.
- 3A. The replacement of the modified gene(s) for its allelic counterpart in the strain of origin should lead to loss of function and loss of pathogenicity or virulence. Restoration of pathogenicity should accompany the reintroduction of the wild-type gene(s).

Falkow's argument was that observing a phenotype that went away when a particular gene was deleted and which came back when that gene was reintroduced is strong evidence that the gene in question **causes** the phenotype. Nearly every molecular genetics experiment follows this logic, and Falkow's postulates are still the gold standard for demonstrating genetic causality. (I would argue, of course, that virulence is not the only interesting bacterial phenotype.)

When designing experiments for this class, think carefully about whether the observations and manipulations you are making test correlation or causation, and interpret the results accordingly. Correlations can be very valuable information. Most of the time, however, an experiment that tests causality is superior to one that tests correlation.

DISCUSSION PROBLEM SET #12: CORRELATION AND CAUSATION

People (and mice) with inflammatory diseases of the gut have different proportions of bacteria in their gut microbiomes than do healthy people. This typically includes higher populations of *E. coli* and lower populations of *Faecalibacterium* species (among other changes).

Propose an experiment to determine whether inflammation causes changes in bacterial populations in the intestine or vice versa. State:

- the independent and dependent variables
 - both positive and negative controls
 - potential outcomes of your experiment, and how you will interpret them
-

The rest of this chapter discusses the use of plasmids, one of the main tools for genetic manipulation of microbes. **Lecture 7** will address the technical aspects of constructing and manipulating plasmids.

PLASMIDS

Plasmids are genetic elements that replicate independently from the chromosome. The term plasmid was first proposed by Joshua Lederberg, and settled on more or less its current meaning around 1968. It replaced François Jacob and Élie Wollman's term *episome*, which is no longer much used in bacterial genetics, but is still used to describe some autonomously replicating DNA molecules in eukaryotes. You will also often hear plasmids referred to as vectors, because they are used to transfer genes from one cell to another. (The analogy is to disease vectors, like ticks or mosquitos. We will discuss gene transfer in more detail in **Lecture 8**.) Plasmids can be used for reverse genetics, by observing the phenotypes that result from addition of a gene or genes encoded on that plasmid to a bacterial strain.

Naturally occurring plasmids vary widely in their size and properties. They may be present in a *copy number* anywhere from one per cell up to hundreds. They may carry a wide variety of genes, some of which are involved in maintaining their own copy number or encode conjugation machinery to transfer themselves to other cells (see **Lecture 8**), and some which may provide evolutionary advantages to their host cell. The classic example of this is antibiotic resistance, but there are many other examples.

Plasmids may be less than 1 kb in size or as large as several Mbp (megabase pairs = 1,000,000 bp), at which point it becomes difficult to distinguish clearly between a plasmid and a chromosome. Generally speaking, in such cases, if an essential gene is encoded on a DNA molecule and it has a copy number of one, it is likely to be considered a chromosome rather than a "megaplasmid". This is kind of arbitrary, of course. The purple photosynthetic bacterium *Rhodobacter sphaeroides*, for example, has two chromosomes: one of 3.1 Mbp and one of 0.9 Mbp. Like bacterial chromosomes, plasmids are usually, but not always, circular. Linear plasmids have been studied in spirochetes and in *Streptomyces* species, but you are unlikely to encounter them in most labs.

The plasmids we use most often in the lab have generally been engineered to make them easy to work with. They are typically quite small (2 – 8 kb), and usually have high copy numbers, which makes them easy to purify and manipulate. They also nearly always encode at least one antibiotic resistance gene, which makes it easy to select for their presence in transformed cells. See below for a reasonably comprehensive list of plasmid features and **Lecture 7** for details on the methods by which plasmids can be engineered and manipulated.

COMPLEMENTATION ANALYSIS

In genetic experiments, plasmids are arguably most important for *complementation analysis*. This is an experimental design that uses plasmid-encoded genes to ensure that the interpretation of mutant phenotypes is correct by fulfilling Falkow's postulates. In a complementation experiment, you replace a mutated gene by expressing the wild-type gene from a plasmid, testing to see if this restores the wild-type phenotype. This is also an excellent way to test whether the phenotype resulting from a particular mutation is due to polar effects on downstream genes.

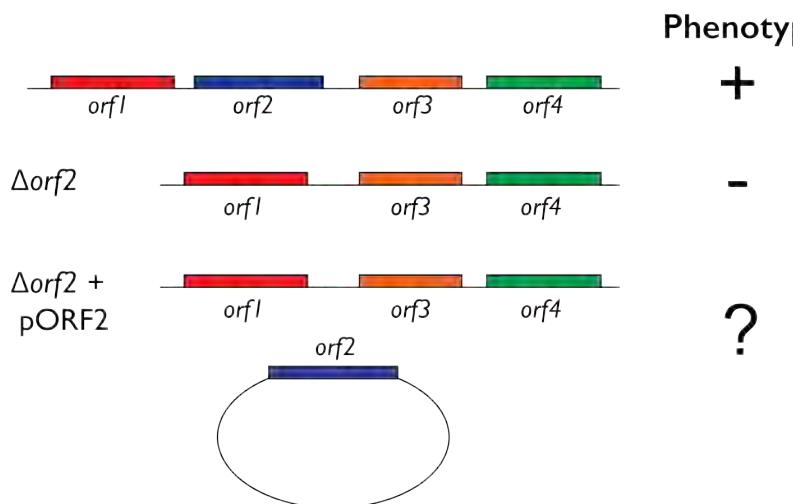
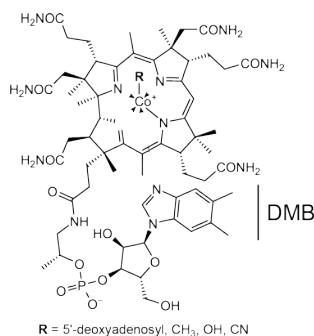


Figure 6.1. Illustration of complementation analysis testing whether *orf2* is responsible for the negative phenotype of the $\Delta\text{orf}2$ mutant.

EXAMPLE OF AN EXPERIMENTAL DESIGN: DMB SYNTHESIS IN *R. RUBRUM*

As a real-world example (drawn from my dissertation research), let's consider the *bluB* gene of the photosynthetic bacterium *Rhodospirillum rubrum*. I constructed a ΔbluB mutant (by allelic exchange, see [Lecture 8](#)) and observed that it grew poorly in the absence of either vitamin B_{12} or the B_{12} lower ligand precursor dimethylbenzimidazole (DMB).



The structure of vitamin B_{12} (cyanocobalamin), with the DMB lower ligand highlighted.

This was intriguing, and suggested that BluB might be necessary for DMB synthesis, but how could I be sure that the phenotype I observed was actually due to the deletion of *bluB* and not to polar effects or to an unrelated mutation elsewhere on the chromosome? It took me most of a year to construct the ΔbluB mutant (site-directed mutagenesis of *R. rubrum* is not trivial), so there was certainly a chance that other mutations could have arisen.

Hypothesis: BluB is responsible for DMB synthesis in *R. rubrum*.

Experimental Design:

- 1) Construct a plasmid expressing the *bluB* gene (*pbluB*⁺) and insert it into the ΔbluB mutant.
- 2) Test whether addition of *pbluB*⁺ restores growth in the absence of DMB.

Independent Variable: (what will you change?)

- The presence or absence of *bluB*.

Dependent Variable(s): (what will you measure?)

- Ability to grow in media with no DMB or vitamin B_{12} added (a qualitative measurement).

Negative Control(s): (eliminate false positive results)

- Include a *vector-only control*, in which the ΔbluB mutant contains a plasmid that does **not** express *bluB*.

This is important in any experiment using plasmids for complementation. Plasmids themselves are not necessarily neutral. Replicating them does have an energetic cost, and plasmids encode genes other than just the ones you

cloned into them. Vector-only controls make sure that any complementation effect you observe is due to the cloned gene and not to the plasmid itself.

Positive Control(s): (eliminate false negative results)

- 1) Include wild-type cells in all growth curves.
- 2) Perform growth curves with DMB or vitamin B₁₂ added.

Does this experiment establish correlation or causation?:

- Because we are fulfilling Falkow's postulates, this design tests **causality**.

Results: (from [this paper](#))

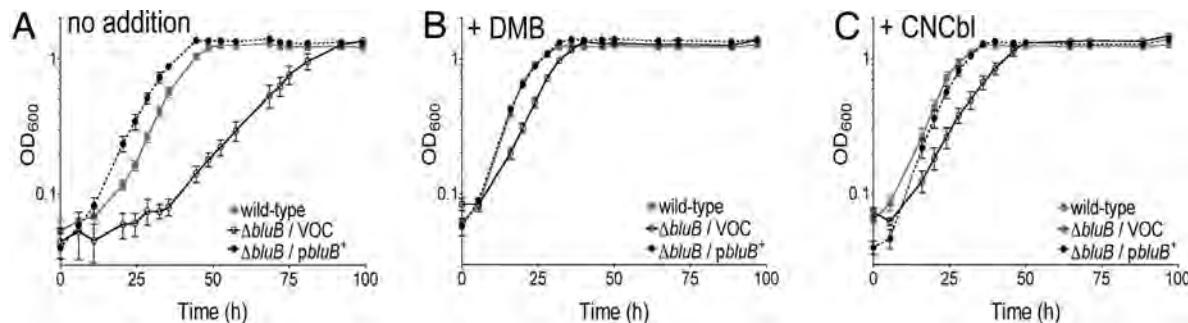


Figure 6.2. BluB is necessary for DMB synthesis. *R. rubrum* wild-type and Δ ablub mutant cultures containing the indicated plasmids were grown photo-synthetically in the presence of DMB or cyanocobalamin (CNCbl), as indicated.

CNCbl is an abbreviation for cyanocobalamin, the chemical name for vitamin B₁₂.

Notice that the wild-type (grey squares) grows well under all three conditions, but the Δ ablub mutant containing an empty plasmid (VOC, open circles) has a significant growth defect in the absence of DMB or B₁₂. Critically, complementing the mutant with the pablub⁺ plasmid (black circles) restored growth in the absence of DMB or B₁₂ (actually allowing it to grow better than wild-type), demonstrating that it was **only** the lack of bluB that was responsible for the observed growth defect phenotype. Panels B and C are important controls to show that the growth defect seen in panel A is due to a problem with DMB and vitamin B₁₂ metabolism, and is not an unrelated problem with growth in the Δ ablub strain.

Whenever possible, you should complement any mutants you make to confirm that the mutation you have made is actually causing the phenotype you observe, and plasmids are almost always the simplest way to do this. This is an especially useful technique when examining mutations that you suspect may have polar effects, since it allows you to distinguish which gene or genes in an operon are responsible for a particular phenotype.

OTHER USES FOR PLASMIDS IN EXPERIMENTS

Far and away the most common use for plasmids in microbiology is as *cloning* vectors. Putting a gene into a plasmid is called *cloning* because it generates many identical copies of the gene in question. The resulting plasmid can then be used for complementation (as with pablub⁺ in the example above) or for a variety of other purposes.

A gene on a plasmid is far easier to manipulate than a gene on the chromosome (see **Lectures 7** and **8** for details). Expression of genes on plasmids can be tightly controlled, depending on the promoter present in the plasmid (**Lecture 4**), so you can tune the amount of RNA or protein product that is produced. You can do this by replacing the promoter entirely or by using an *inducible promoter* whose activity can be modified by addition of a chemical inducer (often a sugar; in practice). Protein products encoded on plasmids can also be fused to GFP or other proteins for detection or purification (discussed in more detail in **Lecture 11**).

Plasmids can easily be mutated, either *in vivo* or *in vitro*, to rapidly test the effect of specific mutations on gene activity. *Site-directed mutagenesis* is much easier on a plasmid than in the chromosome, and allows very precise experimental designs (methods for doing this kind of mutagenesis will be discussed in **Lecture 7**).

Randomly mutating a plasmid allows *localized mutagenesis* of a DNA molecule smaller than the entire genome. One way to do this is by propagating a plasmid containing the gene you want to mutate in a *mutator strain* that lacks DNA repair genes and therefore accumulates random mutations in any DNA molecule it contains (a commercially available example is [XL1-Red](#), which lacks the MutD, MutS, and MutT proteins). The resulting mutated plasmid pool can then be purified and reintroduced into a test strain for experiments. (You would never want to do a genetic experiment in a mutator strain!) Note, of

course, that passaging a plasmid through a mutator strain randomly mutates the entire plasmid, not just a single gene within it.

Another common and very useful technique is the construction of *plasmid libraries*, which are pools of plasmids containing many different cloned inserts. A *genomic library* contains random fragments of the entire genome of an organism, a *metagenomic library* contains random fragments of DNA from a mixture of species, and a *cDNA library* contains DNA reverse transcribed from an mRNA preparation. cDNA libraries are useful not only for enriching for genes which are being actively expressed, but also (for libraries derived from eukaryotic organisms) will lack introns. Libraries can be screened to rapidly identify genes encoding specific functions. A *bacterial artificial chromosome* or *BAC* is a plasmid derived from the F plasmid that can be used to clone very large inserts (up to 350 kb). BACs are commonly used in the construction of genomic libraries from eukaryotic organisms, whose genes are much larger than those in bacteria (due to the presence of introns).

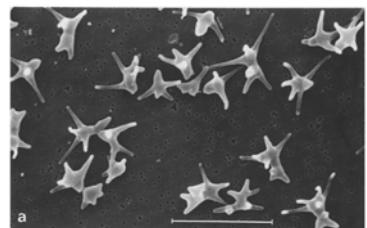
Genomic or cDNA libraries are useful for mutant hunts, since they allow you to select or screen for phenotypes that require the **addition** of a gene or genes from one species into another. In this way, they are kind of the opposite of transposon libraries ([Lecture 3](#)). Many novel antibiotic resistance genes from soil organisms were identified this way by Jo Handelsman's lab in the early 2000's, for example. They prepared metagenomic libraries from DNA purified from soil, transformed them into *E. coli* or other lab bacteria, and selected for antibiotic resistance.

For more information on plasmids, as well as a place you can obtain many useful vectors, the nonprofit plasmid repository [Addgene](#) is an excellent source. Their [Molecular Biology Reference](#) is particularly valuable.

DISCUSSION PROBLEM SET #13: USING PLASMIDS IN GENETIC EXPERIMENTS

Problem #1

You are interested in identifying genes involved in determining the cell shape of the structurally complex bacterium *Anacalymicrobium adetum*.



MicroBestiary

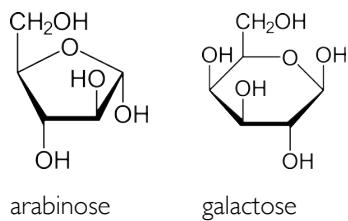
Design a genetic experiment **using plasmids** to identify and confirm which genes are required for cell shape determination in *A. adetum*.

State:

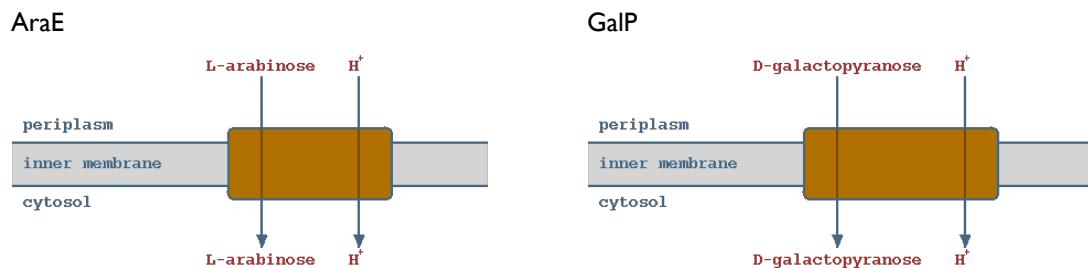
- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?
- what are you using plasmids for in your experiment?
- the independent and dependent variables
- both positive and negative controls
- whether your experiment tests correlation or causation
- potential outcomes of your experiment, and how you will interpret them

Problem #2

Arabinose and galactose are dietary sugars that can affect the levels and proportions of bacteria in the gut. *E. coli* can grow on both of these sugars, although the pathways utilized to break them down are very different (we will discuss metabolic pathways in [Lecture 17](#)).



The transporter proteins for importing arabinose and galactose (AraE and GalP, respectively) in *E. coli* are 65% identical to each other at the amino acid level. They are both > 470 amino acid proteins. AraE cannot transport galactose and GalP cannot transport arabinose (images from EcoCyc).



Design a genetic experiment **using plasmids** to identify amino acids involved in substrate specificity in AraE and/or GalP. (It may be useful to note that neither *araE* nor *galP* are in operons.) State:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?
- the independent and dependent variables
- both positive and negative controls
- whether your experiment tests correlation or causation
- potential outcomes of your experiment, and how you will interpret them

NAMING CONVENTIONS FOR PLASMIDS

There are no rules set in stone for the naming of plasmids, but some guidelines may be helpful. The names of plasmids nearly always start with a lowercase letter p. This is followed by a short name consisting, usually, of capital letters and numbers:

pBR322
pUC18
pET-21b
pBAD18

Like strain identifiers, these letters are often the initials of the researcher(s) who first built or isolated the plasmid. The B and R in pBR322 (one of the original cloning vectors developed for use in *E. coli*) stand for **B**olívar and **R**odríguez, who were postdocs in Herbert Boyer's lab at UCSF in 1977. However, this is not at all a universal rule. The UC in pUC18 stands for **U**niversity of **C**alifornia, the ET in pET-21b (one of a very large family of pET vectors) stands for **e**xpression by **T**7 RNA polymerase, and the BAD in pBAD18 refers to the presence of the arabinose-inducible *P_{araBAD}* promoter in that plasmid. The numbers typically refer to the order in which the plasmid was constructed by the lab in question.

This is all very well, but it becomes somewhat more confusing once a researcher begins to manipulate plasmids for their own work. In a very simple case, a scientist may insert a single gene (say, for example, the metabolic gene *mgsA*) into a common plasmid, such as pUC18. In that case, what should the resulting plasmid be called? There are many possibilities, none of which are really wrong, but I do have my own preferences.

Many people will simply append the name of the gene onto the end of the name of the plasmid and call it a day:

pUC18-mgsA

This is OK in a simple case, but rapidly becomes unwieldy with more complex constructs. Say, for example, that you were constructing a vector with a GFP fusion to a mutant form of the enzyme MgsA. You might end up with something like the following:

pUC18-GFP-mgsA(A745T, G746C, C747T)

While informative, this system is a real nuisance to write and work with, and I personally find it inelegant.

On the other end of the naming spectrum, you might decide that all of this is too complicated and you will just put your initials on every plasmid you build and number them consecutively. In that case, the two plasmids above would just be:

pMjG01
pMjG02

Super simple and concise, and a very common system, but also not very informative, especially since you are likely to be building plasmids for multiple different projects in multiple labs over the course of your research career.

Personally, I like to give plasmids names based on the gene or genes that they encode. I find this strikes a good balance between the two systems above:

pMGS1
pMGS2

Concise, but also reasonably informative, in that it's easy to see that these two plasmids encode alleles of the *mgsA* gene. A table in the resulting publication, along with a detailed description of how each plasmid was constructed in the Methods section, is the appropriate place to describe in detail exactly what alleles and constructs of *mgsA* each plasmid encodes. Your PI will probably have their own preferences, but you will certainly see all three of these methods (and more!) used in the literature.

FEATURES AND TYPES OF PLASMIDS

The following list is not a comprehensive list of everything that might be found on a plasmid, but covers the most common and useful features of plasmids you are likely to encounter, along with some practical considerations for their use. Knowing what these features are will be particularly useful when you are engineering plasmids for your own use ([Lecture 7](#)).

origin of replication (*ori* or *oriC*): Every plasmid has an origin of replication, which controls the ability of the plasmid to replicate within the cell. There are many different types, each of which is associated with a characteristic *copy number* (the number of plasmids per cell; this ranges from 1 to more than 100) and *host range* (the species in which the plasmid will replicate). Origins of replication that work in Gram-negative bacteria will often not work in Gram-positive bacteria, for example. *Shuttle vectors* will replicate in more than one species (allowing you to shuttle genes back and forth between species), and sometimes have separate origins of replication for each species.

If you want to put more than one plasmid into a single strain of bacteria, you need to ensure that they have **different** origins of replication. Because the origin of replication determines the copy number, the replication machinery will not be able to distinguish between multiple plasmids with the same origin and the smaller one (typically) will outcompete the larger one in that population. Two plasmids with the same origin will therefore be *incompatible*, so origins of replication are sometimes called *compatibility groups*.

A single plasmid also **cannot** contain two origins of replication that function in the same organism. The attempts of the two replication machineries to impose different replication cycles and copy numbers will result in double strand DNA breaks, and the plasmid will be unstable, generally losing one or the other origin of replication in unpredictable ways.

A *suicide vector* is a plasmid that can be introduced into a cell but does **not** have a functioning origin of replication for that species (useful, for example, in allelic exchange mutagenesis procedures – see [Lecture 8](#)). This can be accomplished either by using a plasmid with an origin that does not function in the recipient species or by using a vector with a *temperature-sensitive origin of replication*, which will allow replication at a *permissive temperature* (often 30°C) but not at a *restrictive temperature* (often 42°C).

selectable marker: A gene encoding a product which allows you to select for cells containing the plasmid. This is most often an antibiotic resistance gene, in which case only bacteria with the plasmid will survive in media containing that antibiotic. Such a plasmid can only be used in a strain that is otherwise sensitive to that antibiotic, and if you want to have more than one plasmid in a strain, they must have **different** selectable markers. Plasmids may carry more than one selectable marker.

Most plasmids we use in the laboratory are unstable and are lost fairly quickly in the absence of selection (see Table 2.2), and therefore you should always include the appropriate antibiotics in media used to grow strains containing plasmids. Natural plasmids are typically much more stably maintained.

The phenotypes conferred by antibiotic resistance markers are abbreviated in the form Ab^R , as opposed to cells that are sensitive to that antibiotic, which are sometimes indicated as Ab^S . The abbreviations for some common laboratory antibiotics are listed below.

ampicillin = Ap, Amp
chloramphenicol = Cm, Cam
kanamycin = Kn, Kan
tetracycline = Tc, Tet
streptomycin = Sm, Str
spectinomycin = Sp
nalidixic acid = Nx
gentamycin = Gm
rifampicin = Rif
erythromycin = Em, Erm

(A practical note that may save you some headaches: when making antibiotic stock solutions, be sure to look up what concentration and solvent are appropriate. Not all antibiotics are water-soluble. Cm, for example, must be dissolved in 100% ethanol, and Tc will only dissolve in 70% ethanol.)

counter-selectable marker: A gene encoding a product which allows you to select for cells that **don't** contain the plasmid. These typically encode conditionally toxic gene products, and the most common is the *sacB* gene, which confers sucrose sensitivity on many Gram-negative and a few Gram-positive bacteria. This is useful for some allelic exchange procedures, for example, which we will discuss in [Lecture 8](#), or for curing plasmids from cells.

A potentially useful side note is that it is possible (at least in *Salmonella* and *E. coli*) to select **against** tetracycline resistance, allowing *Tc^R* to serve as both a selectable **and** a counter-selectable marker. The method was developed by Barry Bochner, and works because the *tetA* gene makes cells resistant to tetracycline by changing the properties of the cell membrane. Those same changes make the cells **sensitive** to killing by fusaric acid. These [two papers](#) expand on this method and describe media for using it in *E. coli*, should that ever happen to be useful to you.

Multiple Cloning Site (MCS): A small region of the plasmid with several closely spaced restriction enzyme cleavage sites to allow simplified insertion of cloned genes (see [Lecture 7](#)). As we will see, this is drifting ever closer to being an obsolete feature, as methods for restriction enzyme-free cloning become more common.

promoter: A DNA sequence which allows expression of genes on the plasmid. While every gene on the plasmid must have a promoter in order to be expressed (recall [Lecture 4](#)), in most cloning vectors there is a specific promoter directed at the MCS, so that inserted genes will be expressed from that promoter.

Constitutive promoters express genes at a constant level, while inducible promoters can be turned on and off or can have their level of expression regulated by addition of *inducers* to the growth medium. Common inducible promoters used in plasmids include *lac* operon-derived promoters that respond to lactose or unnatural lactose analogs like *IPTG* (isopropyl β-D-1-thio-galactopyranoside) and promoters controlled by other sugars, like arabinose or xylose. Many overexpression vectors used in *E. coli* to produce recombinant proteins for purification contain a very strong promoter from the T7 bacteriophage which, when provided with T7 RNA polymerase (usually on the chromosome of specialized overexpression strains and itself controlled by a *lac* promoter), drives **extremely** high levels of gene expression. It is not uncommon for a protein expressed from a T7 promoter to make up 50% of the total protein within a cell.

Like origins of replication, not all promoters work equally well in all species, and you must use a promoter compatible with the organism you are working with. In lactic acid bacteria, for example, the most common inducible promoter in use is activated by the polycyclic peptide nisin.

ribosome binding site (RBS): also called the *Shine-Dalgarno* sequence after John Shine and Lynn Dalgarno, the Australian scientists who identified it, this is a short AG-rich sequence required for ribosomes to interact with mRNA. In order for a protein to be translated, there must be an RBS between the promoter and the start codon, and different RBS sequences may lead to more or less efficient translation ([Lecture 4](#)). Some plasmids include an RBS, and some do not. In the latter case, you must include an RBS upstream of gene sequences you clone in order for them to be translated.

terminator: a sequence which stops transcription, often included on the opposite side of the MCS from a promoter to prevent *read-through transcription* of other genes on the plasmid from that promoter. These usually consist of stable DNA secondary structures (*hairpins*) that block the progress of RNA polymerase.

fusion proteins / tags: Some plasmids are designed to allow inserted gene sequences to be linked to sequences already encoded on the plasmid. This results in a *chimeric protein* or *protein fusion* with sequence derived from both your inserted gene and another protein. These can be used for purification of the fused protein (as with the 6X-Histidine or GST tags), changing the physical properties of the protein (fusion with the maltose binding protein MBP increases the solubility of a protein), and fusion with a *signal sequence* can target a protein for secretion out of the cell; see [Lecture 13](#), or for easier detection of the expressed protein either *in vivo* or *in vitro* (as with green fluorescent protein, the easily assayed enzyme β-galactosidase, or epitope tags detectable with commercially-available antibodies). To use this kind of plasmid, you must make sure that your gene of interest is *in-frame* with the fusion protein (that is, forms a single continuous open reading frame) and that there is not a stop codon in between your cloned sequence and the coding sequence for the fusion protein.

origin of transfer (oriT): A DNA sequence allowing the plasmid to be *mobilized* from one bacterial cell to another by conjugation. The **transfer genes** (or *tra* functions) necessary for mobilization may be encoded on the plasmid with the *oriT*, on a separate plasmid, or on the chromosome. See [Lecture 8](#) for more on conjugation.

 **fI origin:** Many older plasmids will contain an origin of replication derived from the filamentous phage fI, and are referred to as *phagemids*. This is a site that allows the plasmid to be packaged as long repeating single-stranded DNA molecules when the host bacterium is infected with fI. This was useful when DNA sequencing technologies required large amounts of single-stranded DNA, but is now obsolete.

 **cos sites:** Like the fI origin, cos sites are sequences that allow plasmids to be packaged into phage particles, in this case those of λ phage. Plasmids containing cos sites are called *cosmids*, and can contain much larger DNA sequences than is practical in normal plasmids, limited only by the size of the DNA molecule which will fit in a λ phage capsid (up to about 45 kb).

Like f1 phagemids, cosmids are much less commonly used now than they used to be, having been replaced by more flexible vectors (e.g. **bacterial artificial chromosomes** or BACs).

Typically, when working with a plasmid, you will have a *plasmid map*, which is a drawing showing the location of the various features of that plasmid. You will probably also have a sequence file with the exact DNA sequence of the vector. If you don't, get the plasmid sequenced (for example, by [Plasmidsaurus](#)). If you are unable to find a map for the plasmid you're working with, you should go ahead and draw it yourself, based on the sequence. There is specialized software intended for just this kind of thing. My lab uses [SnapGene](#), but there are many options, including a free web-based tool provided by the company [Genscript](#). I strongly recommend using [Plannotate](#) to annotate plasmid sequence files, since it is able to identify and label many plasmid components that might otherwise go unrecognized.

The figure on the next page illustrates what some plasmid maps might look like:

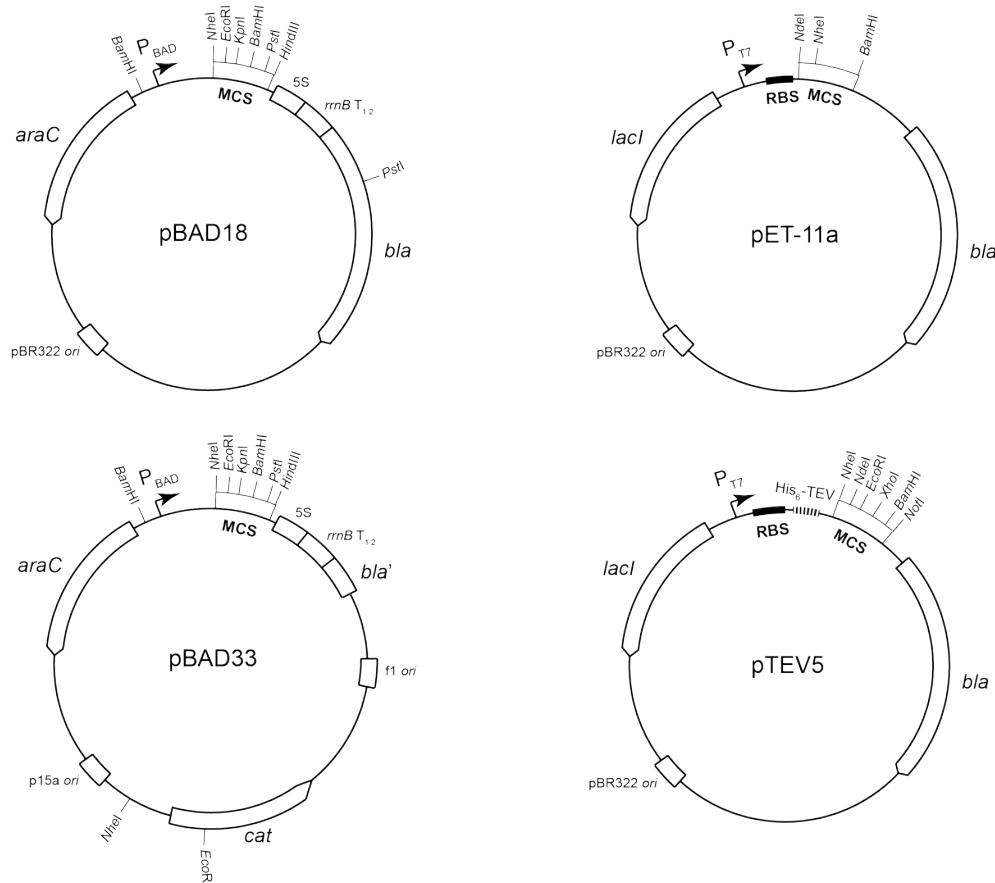


Figure 6.3. Sample plasmid maps.

Note that on these maps I have indicated the direction of transcription for each gene's coding sequence. I have indicated the promoters driving expression of inserts, but not those driving expression of selectable markers and regulators.

- pBAD18 and pBAD33, originally described in [this paper](#), are common vectors used for expressing genes in *E. coli* and other Gram-negative bacteria. They have the P_{BAD} promoter and the *araC* gene encoding the arabinose-sensitive transcriptional activator for that promoter, so that expression of genes inserted into the MCS can be controlled by addition of arabinose to the medium. They also have the 5S $rRNA T_{12}$ transcriptional terminator immediately after the MCS, to prevent read-through transcription from the P_{BAD} promoter.

pBAD18 carries the *bla* gene, encoding β -lactamase, which breaks down the antibiotic ampicillin, and the high copy number pBR322 origin of replication. In contrast, pBAD33 carries the *cat* gene, encoding chloramphenicol acetyltransferase, which confers chloramphenicol resistance, and a lower-copy number p15a origin of replication. It also has an f1 phage origin, making this technically a phagemid. Because pBAD18 and pBAD33 have different origins of replication **and** different antibiotic resistance markers, both of these plasmids can coexist stably in a single bacterial cell.

- pET-11a (available from EMD Millipore) has a much more limited set of restriction sites in its MCS, but includes an RBS, which these particular pBAD vectors do not. It also encodes the IPTG-sensing transcription factor LacI and has a powerful T7 promoter to drive very high-level expression of cloned genes. This only works, of course, in strains containing the T7

RNA polymerase (such as the very common *E. coli* protein overexpression strain BL21 λDE3). pET-11a is intended as an expression vector for recombinant protein purification.

- pTEV5, from [this paper](#), is a similar protein overexpression vector, but has a much improved MCS and incorporates an N-terminal, TEV protease-cleavable 6xHis purification tag into proteins produced from this plasmid. This allows easy purification of the tagged protein, and then removal of the tag from that protein by addition of the sequence-specific TEV protease to the purified protein product.

Note that restriction sites found in the MCS may cut elsewhere in the plasmid (as indicated in the pBAD vectors), so you should take care that any sites you plan to use for cloning (see [Lecture 7](#)) only cut once. Not all possible restriction sites are included in most maps, which is where having the complete sequence becomes helpful. ([RestrictionMapper](#) is a simple online tool that searches DNA sequences for restriction sites.)

LECTURE 6: CRITICAL READING (MUTAGENESIS AND MUTANT HUNTS)

INTRODUCTION & EXPECTATIONS

In today's class, we will discuss a scientific paper from the recent literature in detail, to see how the principles of bacterial genetics we've discussed have been applied to an actual scientific problem. This kind of deep dive into a paper is very valuable for thinking about experimental design and rigor, as well as keeping on top of the current literature. It's also good practice for peer reviewing manuscripts. You will probably participate in *journal clubs* that function more or less this way throughout your career.

To prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
 - What is the central **question** of this paper?
 - Is the experimental design clear and appropriate to address that question?
 - Do you understand the methods used?
 - Are the data clearly presented, with appropriate statistics?
 - Do you agree with the conclusions the authors came to based on their data?
 - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Eickhoff & Bassler (2020) "Vibrio fischeri siderophore production drives competitive exclusion during dual species growth" Mol Microbiol 114:244-261.

As we discussed in **Lecture 1**, you can retrieve this paper from a number of databases. Either PubMed or Google Scholar is probably the simplest option.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including the Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

LECTURE 7: PRINCIPLES OF GENETIC ENGINEERING

INTRODUCTION

Constructing a DNA molecule with sequences from two or more different organisms is called *recombinant DNA technology*. This is the basis of all modern biotechnology, and most frequently involves the use of plasmids. As we saw in the last chapter, plasmids are a critical tool for molecular genetics of bacteria, so knowing how to construct and manipulate them is key to carrying out a lot of experimental designs. This chapter is about how that is done.

PRINCIPLES OF GENETIC ENGINEERING FOR MOLECULAR BIOLOGY

Biotechnology and molecular genetics depend on being able to manipulate the genetic material of cells. From a practical standpoint, this means that your success as a molecular biologist hinges on understanding the technology for constructing and changing the sequence of DNA molecules. At first glance this seems like a very daunting proposition. There are hundreds of different protocols for manipulating DNA, some of which have many steps and seem very complicated or specialized, and new techniques are being invented all the time. However, all of these techniques are built up from a framework of only a few different fundamental procedures. The goal of this chapter is to provide a practical resource that will explain what those building blocks are and how they can be combined to build a DNA molecule of almost any desired sequence.

First, I will describe the six fundamental procedures that make up all molecular genetics protocols and the current technology for carrying out these procedures both on purified DNA *in vitro* and in living cells *in vivo*. Then I will show how these procedures are combined to construct and modify DNA molecules, using common lab techniques as specific examples. I will include notes with links to resources describing specific technologies in detail for readers who want to explore them in more depth. I will also try to highlight common mistakes and points of confusion.

By the end of this section, you should be able to understand any molecular biology protocol by breaking it down to its basic building blocks. I will focus here on molecular genetics in bacterial systems, but the fundamental concepts apply to all molecular biology, and in fact almost all DNA molecular construction is done in *E. coli*, where the most highly developed tools are available. The resulting DNA products are then transferred to other species of interest.

There are lots of software packages that exist to help visualize and manipulate DNA molecules *in silico*. My lab uses [SnapGene](#), mostly, but [Benchling](#) is an alternative that is free to academics.

THE SIX FUNDAMENTAL PROCEDURES OF MOLECULAR BIOLOGY

All of molecular biology is based on carrying out combinations of six different procedures on DNA: **Reading**, **Writing**, **Copying**, **Cutting**, **Pasting**, and **Swapping** sequences, either *in vitro* or *in vivo*. The following table lists these procedures, along with the method(s) we use to accomplish them (some of which we will not discuss until [Lecture 8](#)).

	<i>In vitro</i>	<i>In vivo</i>
Read	DNA sequencing	--
Write	Oligonucleotide synthesis	--
Copy	PCR	Replication
Cut	Nucleases	CRISPR
Paste	Ligase	DNA nick repair
Swap	--	Homologous recombination



In the text and figures below, appropriate icons will be used to indicate each type of procedure.



Read

The technology for determining the nucleotide sequence of DNA molecules continues to advance rapidly, and it is now straightforward and inexpensive to sequence DNA up to and far beyond the length of the genome of a bacterium. Every molecular biology protocol ends with a DNA sequencing Read step to confirm that the correct DNA molecule has been constructed. DNA is always extracted from the organism before sequencing, so the Read step always happens *in vitro*.

For routine sequencing of short sections of DNA molecules (< 1000 bp), Sanger sequencing is a common and fairly inexpensive method. A variety of so-called *next-generation sequencing* (NGS) and emerging third-generation sequencing technologies (e.g. Oxford nanopore sequencing) exist that allow us to sequence whole genomes and complex mixtures of DNA from many organisms (*metagenomes*), often by computationally aligning millions of very short sequence reads (50 –

100 bp), although nanopore and PacBio sequencing technologies are capable of much longer reads. Practically speaking, in most labs you will not do your own DNA sequencing but will outsource it to a company or university core facility.

[SeqCenter](#) provides extremely affordable whole-genome sequencing (about \$100 for a bacterial genome and associated bioinformatic analysis), and my lab has had a very good experience working with them. [Plasmidsaurus](#) offers full-length plasmid sequencing for \$15 per plasmid, and at those price points I'd strongly recommend sequencing every plasmid and strain you work with. It has certainly saved us from some embarrassing mistakes.

Write

It is possible to chemically synthesize DNA molecules with a desired sequence *in vitro*, but current methods typically only allow accurate synthesis of single-stranded DNA chains up to about 100 nucleotides long. These are called *oligonucleotides* ("oligos", in common lab jargon) and are relatively inexpensive (around 20¢ per nucleotide, or cheaper with bulk ordering discounts).

Since even a single gene is usually hundreds or thousands of nucleotides long, it is more complicated to make large and complex DNA molecules from scratch (although see the description of Gene Synthesis below for a common commercially available way to do so, when needed). There is no equivalent technique for generating entirely new DNA sequences *in vivo*.

Almost all of the protocols we'll discuss below begin with an oligonucleotide synthesis Write step. Very few labs have the specialized equipment to synthesize their own oligos, and you will typically order them from a company. There is lots of software available to help you design primers for different applications, often provided by the companies that want to sell you the oligos. Here are a few free web-based options you may find useful:

[Primer-BLAST](#)

[IDT Primer Quest](#)

[NEBuilder](#)

[WebPrimer](#) (this is what I use, mostly)

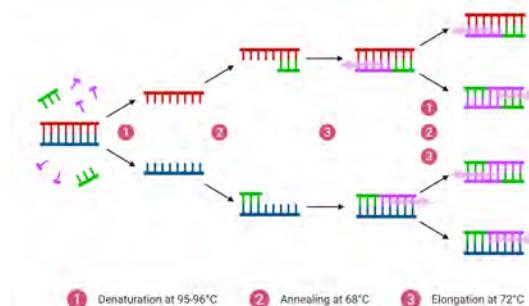
Copy

In order to sequence or manipulate DNA, we typically need to make many copies of the specific DNA molecule of interest. We can do this either *in vivo* or *in vitro*. In either case, DNA polymerase is the essential enzyme for copying DNA sequences and uses an existing DNA molecule as a template to synthesize new DNA.

If the DNA molecule in question has an origin of replication, it is simple to grow large amounts cells containing that DNA and allow normal cellular replication and reproduction to generate copies *in vivo*. This is useful for generating large amounts of chromosomal DNA and plasmids.

One of the key technologies of molecular biology is PCR (the polymerase chain reaction), a technique for copying DNA sequences *in vitro*. (See [this Wikipedia page](#) for a detailed explanation.) For PCR, short oligos (15 – 30 nucleotides) called *primers* are designed which are complementary to the sequence of a double-stranded DNA template molecule. These are annealed to the template and then extended with purified DNA polymerase. If two primers are used which are directed towards each other on the same template (DNA polymerase always extends in the 5' to 3' direction) and multiple cycles of annealing and extension are repeated, the result is exponential copying of the sequence between the two primers.

Polymerase chain reaction - PCR



BioRender (2019). Polymerase Chain Reaction (PCR). <https://app.biorender.com/biorender-templates/figures/all/t-5cf004269cf1ff003377d8cd-polymerase-chain-reaction-pcr>

PCR works best on relatively short sequences of a few hundred to a couple of thousand base pairs but can be used to amplify linear DNA molecules up to about 10,000 base pairs long. Many different thermostable DNA polymerases are

available for use in PCR, some of which are especially good at amplifying long templates or are engineered to make fewer errors during amplification.

Since we can only write short DNA sequences, constructing complex DNA molecules always involves at least one Copy step. In order to get enough DNA to Read the resulting sequence, it is also essentially always necessary to copy the product of your protocol *in vivo*.

Cut

A key step in many genetic engineering protocols is cutting DNA molecules into smaller fragments. Nucleases are enzymes that cleave DNA molecules by breaking the bonds between nucleotides. The most common and useful nucleases for molecular biology are those that cleave DNA only at specific nucleotide sequences, but some protocols use nucleases with less specificity for particular purposes. This is important in Gibson Assembly, for example, a protocol we will examine later in this chapter.

Purified nucleases are used to cut DNA molecules *in vitro*. Restriction enzymes, the most commonly used type, are nucleases that recognize specific short DNA sequences (usually 4 – 8 base pairs long, called *restriction sites*) and introduce a double-strand break in the DNA at or near that recognition sequence. Hundreds of different restriction enzymes with different recognition sequence specificities are commercially available. See, for example, the lists of enzymes available from [Thermo-Fisher](#) or [New England Biolabs](#). [RestrictionMapper](#) is extremely helpful for determining which restriction enzymes will cut a given DNA sequence.

The names of restriction enzymes are based on the species they were originally isolated from. EcoRI and EcoRV are the first and fifth restriction enzymes isolated from *E. coli* strain R, for example, and HindIII was isolated from *Haemophilus influenza* strain Rd.

Since we are able to read DNA sequences, we can reliably predict where a restriction enzyme will cut any given DNA molecule. Some restriction enzymes break both DNA strands at the same base pair, generating a blunt ended cut. Others, which are typically more useful for molecular biology, break the two strands in a staggered way, generating "sticky ends" with short single-stranded overhangs at the end of the cleaved DNA molecule, as illustrated below:

DNA fragment 1:

**CATATGTTAAAAAATCTGTTTATTGCAACACTATTATCTGGCGTTATGGCATTTCCACCAATGCAGATGATAAAATAATTCTGATAAGGATCC
GTATACAATTTTAGACAAAATAACGTTGTGATAAGACCGCAATACCGTAAAGGTGGTTACGTACTATTTTATTAAGACTATTCTTAGG**

DNA fragment 2 (part of plasmid pET-11a):

...GAAGGAGATATA**CATATGGCTAGCATGACTGGTGGACAGCAAATGGGTGCGGGATCCGGCTGTAACAAA**...
...CTTCCTCATAT**GTATACCGATCGTACTGACCACCTGTCGTTACCCAGGCCTAGGCCAGCATTGTTT**...

DNA fragment 1, digested with *Nde*I and *Bam*HI:

TATGTTAAAAATCTGTTATTGCAACACTATTATCTGGCGTATGGCATTTCCACCAATGCGATGATAAAATATTCTGATAAG
ACAAATTAGACAAAATAACGTTGTATAATAGACCGCAATACCGTAAAGGTGGTACGTACTATTATTAAGACTATTCCCTAG

DNA fragment 2 (part of plasmid pET-11a), digested with *Nde*I and *Bam*HI:

...GAAGGAGATATA
...CTTCTCTATATGTAT TATGGCTAGCATGACTGGTGGACAGCAAATGGGTCGCG GATCCGGCTGCTAACAAA...
ACCGATCGTACTGACCACCTGTGCGTTACCCAGCGCCTAG GCGCACGATTGTTT...

I will explain the value of DNA sticky ends in the section on Paste steps below.

The most recent major addition to the molecular biology tool kit is a technology for cutting DNA *in vivo*. CRISPR (which stands for **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats, referring to the genomic context in which the relevant genes were discovered) takes advantage of a nuclease called Cas9 that can be targeted to a specific DNA sequence *in vivo* by a short guide RNA. This confers great specificity to Cas9 and allows it to introduce double-stranded DNA breaks at very precise locations in the chromosomes of living cells. Applications of CRISPR are in very active development and are allowing previously impossible genetic engineering procedures in a wide range of species. The biotech company Genscript has a very nice summary of the history and uses of CRISPR available for download [here](#), and we will discuss applications of CRISPR in bacterial genetics in **Lecture 8**.

Both restriction enzymes and CRISPR are derived from naturally occurring systems bacteria use to defend themselves against infection by viruses. Since restriction enzyme recognition sites are short and occur commonly in the genomes of the bacteria encoding those enzymes, each restriction enzyme is paired *in vivo* with a DNA methylase that methylates the recognition site and protects the host cell's DNA against restriction. Unmethylated DNA, such as the genome of an invading virus, is therefore cut by the restriction enzyme, preventing infection. (Infection by phage is what is restricted by restriction enzymes.) Practically speaking, this means that we can protect a DNA molecule from digestion by a particular restriction enzyme by treating it with the corresponding methylase *in vitro* or by copying it *in vivo* in a strain expressing that methylase. PCR products are always unmethylated, which is important to remember and is useful in some cloning and mutagenesis procedures (as we will discuss later in this chapter).

CRISPR targets longer (approximately 20 bp), less common DNA sequences, and bacteria defend themselves against their own CRISPR systems by simply not encoding the target sequences anywhere in their genomes, or occasionally by encoding CRISPR-inhibiting proteins, which can also be found in some phage genomes.

Paste

Recombinant DNA technology depends on being able to paste two or more DNA molecules together into a single molecule. This reaction is catalyzed by an enzyme called *DNA ligase*.

Ligase forms a phosphodiester bond between the 5' phosphate of one linear single DNA strand and the 3' hydroxyl of another. *In vivo*, this is part of a cell's DNA repair mechanism, and repairs nicks or breaks in a single strand of double-stranded DNA molecules. If a molecular biology protocol results in a DNA molecule with a single nick or a few widely-spaced nicks, this will be repaired when that molecule replicates *in vivo*. Most bacteria do not typically ligate double strand breaks efficiently *in vivo*, although some have pathways for doing so. (Double strand break repair in eukaryotic cells is called non-homologous end-joining.)

Generating recombinant DNA *in vitro* with purified ligase is more versatile. The most common enzyme used for this in labs is the DNA ligase from the bacteriophage T4. At high enzyme concentrations, T4 DNA ligase can join blunt-ended linear double stranded DNA fragments into linear or circular products, but does so at fairly low efficiency. Sticky-ended DNA fragments allow more precision, since fragments with complementary sticky ends will anneal to each other, in essence creating loosely fused DNA molecules with two nearby nicks, one on each strand. T4 ligase efficiently forms phosphodiester bonds to repair these nicks, allowing construction of composite DNA molecules with their components joined in a particular orientation and order.

Insert, digested with *NdeI* and *BamHI*:

**TATGTTAAAAAAATCTGTTTATTGCAACACTATTATCTGGCGTTATGGCATTTCACCAATGCAGATGATAAAAATTCTGATAAG
ACAAATTAGACAAAATAACGTTGATAATAGACCGCAATACCGTAAAGGTGGTTACGTCTACTATTTATTAAGACTATTCTAG**

Vector, digested with *NdeI* and *BamHI*:

**...GAAGGAGATATACA
...CTTCCTCTATATGTAT** **GATCCGGCTGCTAACAAA...
GCCGACGATTGTT...**

Ligated product:

**...GAAGGAGATATACA
...CTTCCTCTATATGTAT** **TATGTTAAAAAAATCTGTTTATTGCAA...CACCAATGCAGATGATAAAAATTCTGATAAG
ACAAATTAGACAAAATAACGTT...GTGGTTACGTCTACTATTTATTAAGACTATTCTAGGCCGACGATTGTT...**

Essentially all genetic engineering protocols involve a Paste step, although for some procedures that step is relatively invisible since it happens *in vivo* at the same time as the final Copy step before sequencing.

Swap

Finally, in some protocols you will take advantage of the ability of cells to Swap sequences from one DNA molecule to another. This is dependent on another DNA repair mechanism called *homologous recombination*, and only occurs *in vivo*. We will discuss the mechanism and use of recombination in [Lecture 8](#).

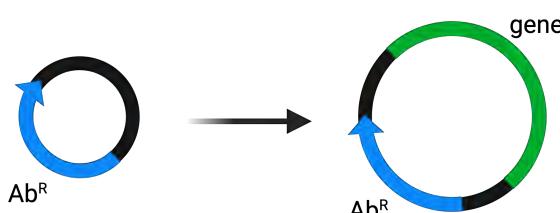
! Beware of confusing terminology here: **recombinant DNA** and **DNA recombination** are not the same thing!

For an alternative discussion of the principles of molecular biology, from a different perspective and with some more technical details, see [this very nice article](#) from Addgene.

SCREENING FOR PRODUCTS

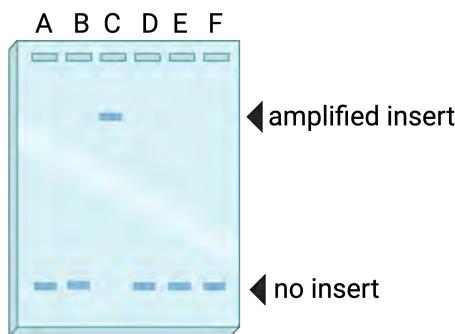
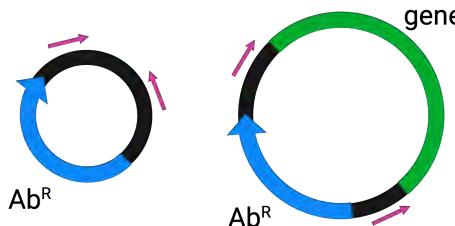
One important note to consider when looking at the protocols below is that none of the steps occurs with 100% efficiency, and it is essentially always necessary to screen multiple candidates to find the final product you are looking for. I have left out the screening steps from the detailed protocols below for simplicity's sake, but in the lab you will definitely need to include them. They typically occur just before the final Read step.

To illustrate, say you are trying to construct a plasmid containing a gene for complementation analysis. You start with an expression vector and you are hoping to end up with a final product like this:

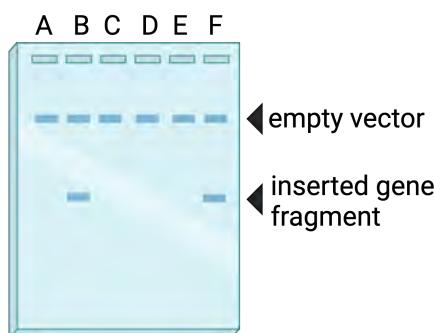
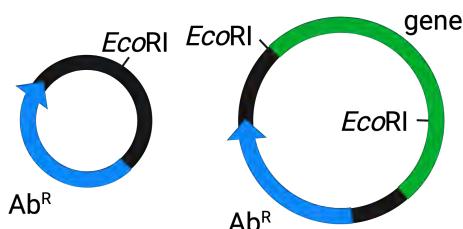


You will inevitably end up with a bunch of different plasmids, some of which will be the empty vector you started with and some of which will contain the gene of interest. How can you tell them apart? You have a few different choices:

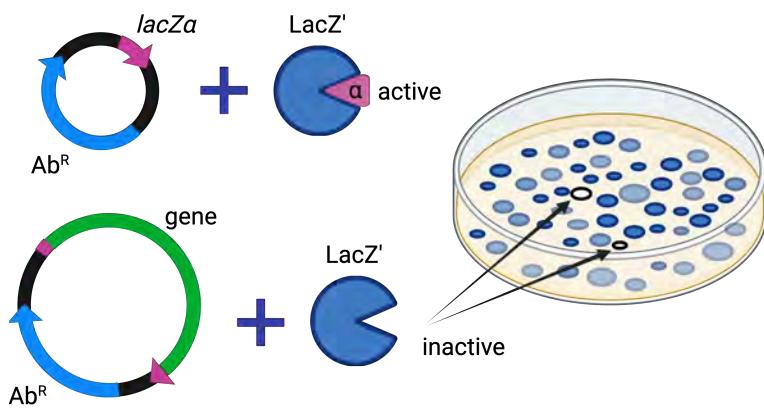
1. **Sequence them all.** Just jump right to the Read step. This is the simplest, perhaps, conceptually, but is also the most labor-intensive and expensive way to do it. Sometimes it's what you've got to do, though, especially if you're making more subtle changes to a plasmid, like single base pair mutations.
2. **Screen by PCR.** To save on sequencing costs, you can use PCR primers (purple in the image below) flanking either side of the expected gene insert to amplify the part of the plasmid that contains your gene of interest, then separate the PCR products on an agarose gel to screen by size. In this example, candidate C is the one to send for sequencing.



3. **Screen by restriction digestion.** If there are conveniently-located restriction sites in your plasmid (such as the EcoRI sites in the example below), it's even easier and cheaper to digest the plasmids and screen for plasmids that give the expected size fragments (candidates B and F in this example):



4. **⚠ Blue-white screening.** This is an old, but occasionally still useful method that uses specialized *E. coli* cloning strains (such as [DH5α](#)) encoding an inactive form of the LacZ enzyme paired with specific plasmids (such as [pUC18](#)) encoding a peptide which can restore LacZ activity in those strains (the LacZα fragment). When used together on media containing a dye that turns blue in the presence of active LacZ (called X-gal), this allows you to pick out white colonies that contain plasmids with inserts that disrupt the LacZα fragment:



For more details, see [this Wikipedia page](#).

EXAMPLES OF COMMON MOLECULAR BIOLOGY PROTOCOLS

In this section, I will break down a series of protocols into their component steps, both in outline and graphical form. I will proceed from fairly simple procedures to more complex ones. Notice, however, that many of the steps are the same for all or almost all of the protocols. For example, essentially every protocol ends with an *in vivo* copy step and an *in vitro* read step to confirm the sequence of your engineered DNA product.

The key principle I want you to take away from this chapter is this: **complicated protocols are just combinations of simple procedures.**

Molecular biology is essentially a creative endeavor. Like any artist, you are using the tools at your disposal to solve problems in a creative way. This is your toolbox.

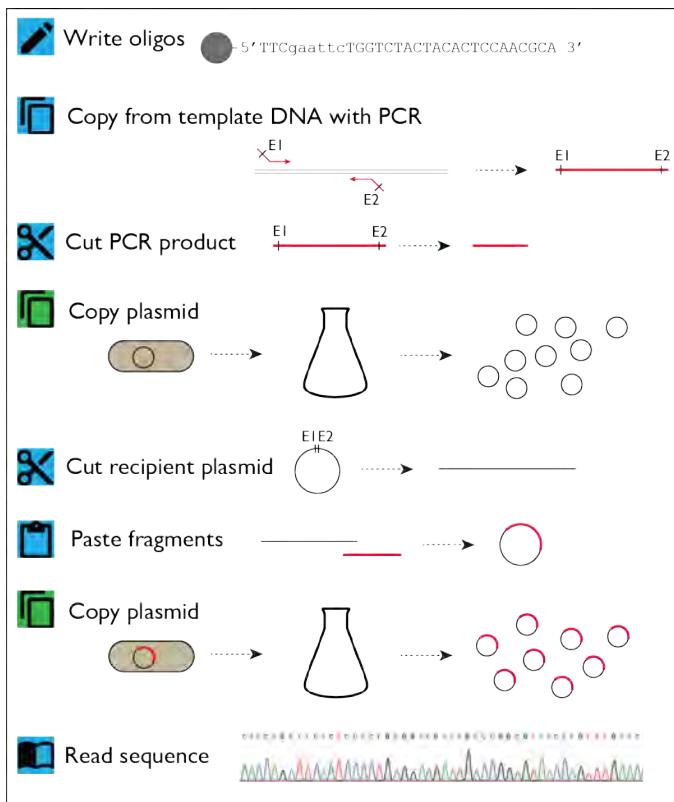
Icons on a **green** background indicate *in vivo* steps, while those on a **blue** background indicate *in vitro* steps.

CLONING

The protocol most frequently referred to as cloning in a modern molecular biology lab involves generating a gene sequence by PCR and then inserting it into a plasmid's multiple cloning site *in vitro* with compatible restriction enzymes and DNA ligase.

⚠ I am going to raise some hackles here, but while this classic method of cloning with restriction enzymes is still used by many labs, I think it is almost always a poor choice and is basically obsolete, given advances in restriction-free cloning and gene synthesis technologies (see below). Nevertheless, you should certainly know how it works.

Because PCR primers can be synthesized directly, this allows you to place any restriction site you like at the ends of the DNA to be inserted and means you do not have to depend on whatever restriction sites are naturally present in the original source of that DNA. (If you are just cutting out a piece of one plasmid and cloning it into another, without a PCR step, this is generally referred to as *subcloning*.)



Protocol:

- 1. Write – *in vitro***
 - design PCR primers that amplify your DNA of interest
 - add desired restriction site sequences to the 5' end of the primers (with a few extra nucleotides, since many restriction enzymes don't cut well at the very end of a DNA fragment)
- 2. Copy – *in vitro***
 - PCR amplify the insert DNA from a template (for example, genomic DNA) using the primers designed in step 1
- 3. Cut – *in vitro***
 - digest the PCR-amplified insert DNA with the restriction enzymes whose sites you added to the primers
- 4. Copy – *in vivo***
 - grow cells containing recipient plasmid and make a large amount of it
- 5. Cut – *in vitro***
 - digest the recipient plasmid with the same restriction enzyme(s) used in step 3
 - optionally, treat with a phosphatase to remove the 5' phosphate from the DNA
- 6. Paste – *in vitro***
 - mix the digested recipient plasmid and donor gene and treat with ligase to covalently join the sticky ends
- 7. Copy – *in vivo***
 - transform the ligation mixture into a fresh bacterial strain, select for strains containing complete plasmids, and grow to make a large amount of the recombinant product
- 8. Read – *in vitro***
 - sequence the recombinant plasmid to confirm that it has the desired sequence

Since PCR is generally limited to amplifying no more than 10 kb of DNA, it is difficult to obtain enough PCR product for cloning fragments larger than this, but most plasmids are not very good at maintaining inserts that large anyway. For large inserts, specialized vectors (e.g. BACs) are a better choice.

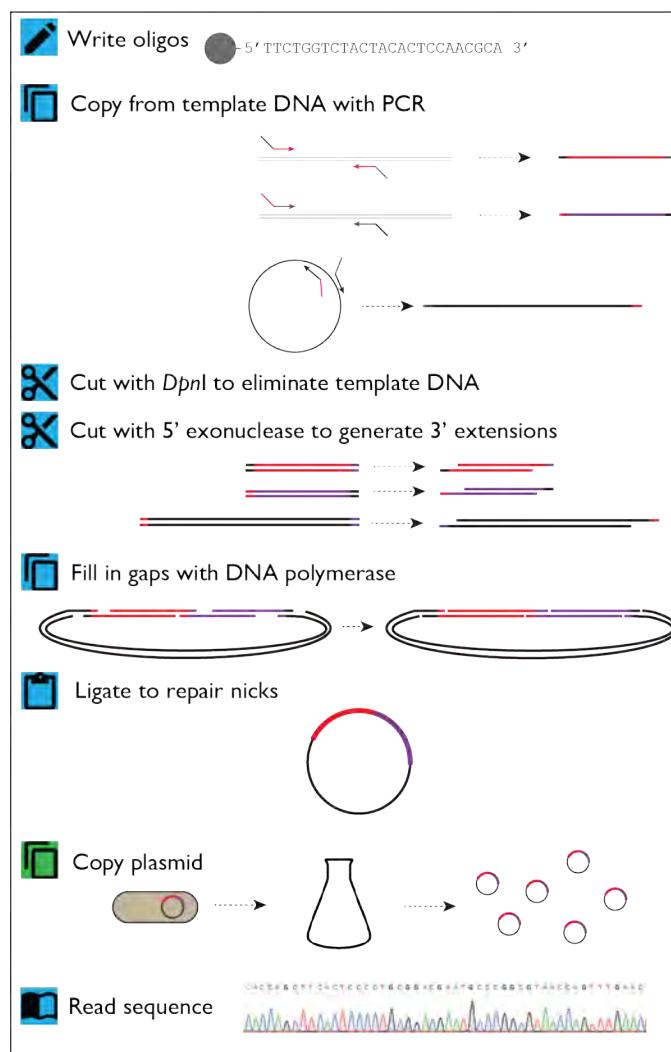
Exactly the same procedure can be done with completely or partially restriction-digested genomic DNA instead of a PCR product, which results in a pool of plasmids containing a variety of different inserts. This is a *genomic library* and is useful for many kinds of mutant hunts, as we discussed in **Lecture 5**. (Partial digestion just means a restriction digest that is stopped before allowing the enzyme to cut at every possible site in the DNA.) If you modify the PCR conditions so that the DNA

polymerase is not very accurate (changing the nucleotide ratio or Mg concentration are common ways to do this), you can simultaneously clone and randomly mutagenize a gene, which is a useful kind of localized mutagenesis.

GIBSON ASSEMBLY

Restriction-free cloning is a general term for cloning procedures that do not depend on restriction enzymes (or, in some cases, ligase). There are many of these types of methods available, but perhaps the most widely used is called Gibson Assembly, after its inventor, Daniel Gibson. [Wikipedia](#) has a good description of the procedure, and Dr. Gibson has written a detailed description of the history and different variations on the protocol which is available, at time of writing at [this link](#).

In a Gibson Assembly procedure, you generate PCR products designed to have 20 – 40 bp of sequence overlap at the ends. Then you add a 5' exonuclease, which essentially generates 3' sticky ends on all of those linear DNA molecules. They then can anneal, the gaps can be filled with DNA polymerase, and ligase then repairs the remaining nicks to generate a circular plasmid DNA product.



Protocol:

- 1. Write – *in vitro***
 - design PCR primers that amplify your DNA of interest; this will include both your vector backbone and the insert(s) to be assembled
 - make sure that your primers include sequences at each end so that the resulting linear DNA fragments have overlapping ends (20 – 40 bp)
- 2. Copy – *in vitro***
 - PCR amplify the insert and vector DNA from templates (plasmid or genomic DNA) using the primers designed in step 1
- 3. Cut – *in vitro***

- treat with the restriction enzyme *Dpn*I, which cuts **methylated** DNA at GATC; this eliminates the original vector, methylated by the natural Dam methylase of *E. coli*, while leaving the unmethylated PCR-synthesized DNA intact

4. Cut – *in vitro*

- mix the DNA fragments to be assembled together
- add 5' exonuclease, which degrades **only** one strand of each DNA molecule, leaving all of the fragments with single-stranded 3' overhangs on both ends
- the overlapping ends will anneal to one another, forming a loosely-linked version of the final product

5. Copy – *in vitro*

- add DNA polymerase to fill in the gaps in the annealed DNA molecules

6. Paste – *in vitro*

- treat with ligase to repair the nicks that remain in the resulting product, resulting in a covalently-joined double stranded circular DNA product

7. Copy – *in vivo*

- transform the ligation mixture into competent bacteria, select for strains containing complete plasmids, and grow to make a large amount of the recombinant product

8. Read – *in vitro*

- sequence the recombinant plasmid to confirm that it has the desired sequence

Note that while steps 4, 5, and 6 are written here in sequential order, they are actually performed simultaneously in a tube containing all three enzymes: 5' exonuclease, DNA polymerase, and DNA ligase. This simplifies the benchwork considerably.

A key reagent in this procedure is the restriction enzyme *Dpn*I. As noted above, *Dpn*I cleaves at a very commonly occurring tetranucleotide sequence, but **only** when that DNA has been methylated by the Dam methylase. From a practical standpoint, this means that it will digest any genomic or plasmid DNA purified from *E. coli* into many small fragments but will **not** digest PCR products. This is an extremely useful thing to be able to do and comes up in a lot of protocols.

One of the key advantages of Gibson Assembly and related methods, besides not being limited by the presence or absence of restriction sites, is that they allow you to assemble vectors with multiple inserts or complex genetic structures in a single cloning reaction. This means you can pretty rapidly construct a recombinant DNA molecule with essentially any sequence you want. Gibson Assembly can reportedly be used, with some modification of the protocol, to assemble plasmids from as many as 15 individual DNA fragments in a single reaction.

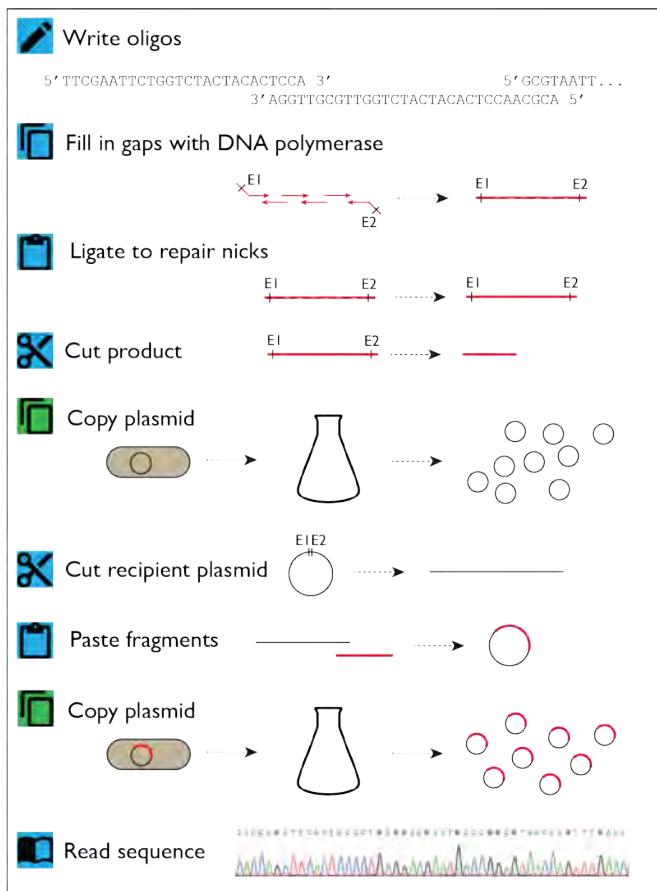
Most of the steps of Gibson Assembly take place *in vitro*, but essentially exactly the same procedure can be carried out almost entirely *in vivo*, in a protocol called *In Vivo* Assembly Cloning or *In Vivo* *Escherichia coli* Cloning (iVEC). We will read much more about this technique, which has almost completely replaced other cloning technologies in my lab, in this week's Critical Reading ([Lecture 9](#)). iVEC is, however, only useful for plasmids that can replicate in *E. coli*, so Gibson Assembly is still necessary when constructing plasmids for use in other species.

GENE SYNTHESIS

PCR isn't the only way to generate a large DNA fragment to be cloned into a plasmid. It is possible to build up DNA molecules of any sequence and, in theory, any length by synthesizing a series of overlapping oligonucleotides and stitching them together in a process called overlap extension.

Many companies will synthesize DNA for you in this way fairly inexpensively. ([Twist Bioscience](#), for example, will synthesize DNA fragments and clone them into any one of a set of useful plasmids for 9¢ a base pair, or only 7¢ a base pair if you just want the synthesized linear DNA pieces without having them cloned.) In many cases, this is now the most cost-effective way to obtain a plasmid or other DNA molecule, since you are outsourcing the work to a specialist company and can spend your time on more interesting problems.

The practical length for a single DNA synthesis reaction is about 2 kb, and you always obtain a linear DNA molecule from this procedure. The resulting DNA fragment can then be cloned into a plasmid as usual. The diagram below shows the product being cloned with restriction enzymes, but obviously Gibson Assembly or iVEC works too (and is probably a better choice if you are doing the assembly yourself).



Protocol:

1. Write – *in vitro*

- design 40-50 nucleotide oligos that overlap at their ends and together encode your desired sequence, with restriction sites as desired at the ends of the final product
- anneal the oligos to each other

2. Copy – *in vitro*

- add DNA polymerase to fill in the gaps in the annealed oligo chain

3. Paste – *in vitro*

- treat with ligase to repair nicks and form a single double stranded linear DNA product

4. Cut – *in vitro*

- digest the synthesized DNA with the restriction enzymes whose sites you added to the primers

5. Copy – *in vivo*

- grow cells containing recipient plasmid and make a large amount of it

6. Cut – *in vitro*

- digest the recipient plasmid with the same restriction enzyme(s) used in step 4
- optionally, treat with a phosphatase to remove the 5' phosphate from the DNA

7. Paste – *in vitro*

- mix the digested recipient plasmid and donor gene and treat with ligase to covalently join the sticky ends

8. Copy – *in vivo*

- transform the ligation mixture into a fresh bacterial strain, select for ligated plasmids, and grow to make a large amount of the recombinant product

9. Read – *in vitro*

- sequence the recombinant plasmid to confirm that it has the desired sequence

One advantage to having a gene synthesized rather than cloning it directly from genomic DNA is that it allows you to optimize the gene's codon usage for expression in your target organism. Different species translate the various codons for

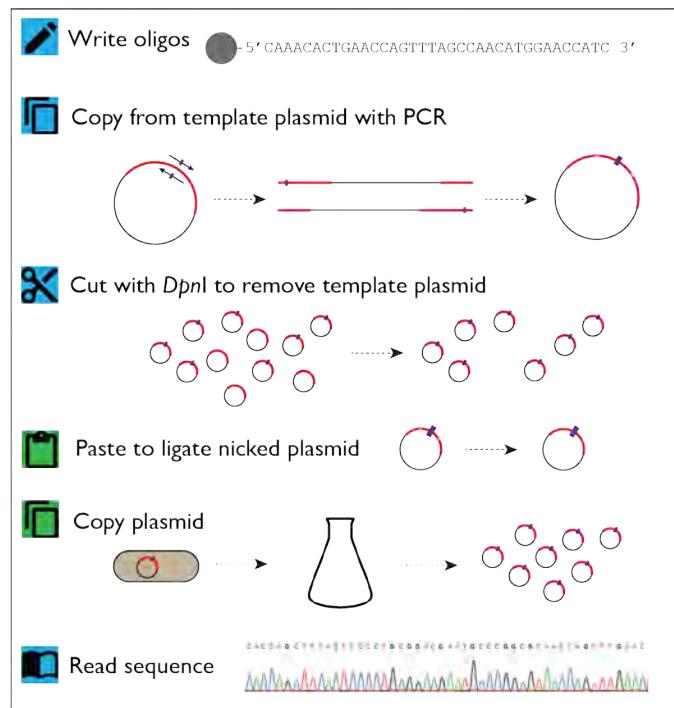
specific amino acids at different efficiencies, and this can strongly affect how much protein is produced. For example, *E. coli* very rarely uses the AGG codon for arginine and has low levels of the tRNA for that codon.

Another useful application is to obtain genes from organisms which are difficult to grow or for which genomic DNA is not readily available. It is also a simple way to obtain complex protein fusions that would be labor-intensive to construct by cloning the parts individually.

It is worth noting that the technology described here is not the be-all and end-all of gene synthesis, and new technologies that allow ever larger DNA molecules to be synthesized *in vitro* are in development. Here's [a 2023 paper](#), for example, that describes a method for synthesizing megabase-scale products.

SITE-DIRECTED MUTAGENESIS OF PLASMIDS

The methods I've described so far focus on constructing plasmids from large component parts, which is a very common molecular biology procedure. However, you will often want to make more subtle changes to a DNA molecule, including changing single base pairs or codons. There are a variety of ways to accomplish this. Here is one of the most common. It works well for small mutations of all kinds.



Protocol:

1. Write – *in vitro*

- design a pair of oligos that are complementary to your plasmid, with the desired mutation centered in the oligo sequences (the [PrimerX](#) tool is very useful for this)

2. Copy – *in vitro*

- using the oligos designed in step 1 as primers, use PCR to amplify the entire plasmid; this will require using a high-fidelity DNA polymerase with high enough processivity to generate a full-sized plasmid product
- the resulting single stranded products will anneal into a nicked, double stranded circular DNA molecule

3. Cut – *in vitro*

- treat with the restriction enzyme *DpnI*, which cuts methylated DNA at GATC; this eliminates the original vector, methylated by the natural Dam methylase of *E. coli*, while leaving the unmethylated PCR-synthesized DNA intact

4. Paste – *in vivo*

- transform the resulting nicked circular DNA product into a fresh bacterial strain; the DNA repair system of the recipient strain will repair the nicks in the plasmid

5. Copy – *in vivo*

- grow up the transformed strain to make a large amount of the recombinant product

6. Read – *in vitro*

- sequence the recombinant plasmid to confirm that it has the desired sequence

This is the basis of the QuikChange mutagenesis kit, available from Agilent. You don't need to buy their reagents to do this kind of mutagenesis, but they do have a nice description of how it works in the manual available [here](#).

Surprisingly, it is not actually necessary to generate a double-stranded DNA product for this type of site-directed mutagenesis to work. The procedure above works very well with only a single primer. This generates a single-stranded, linear mutated DNA product, which *E. coli* is able to repair into a circular double-stranded DNA, probably by first synthesizing the second strand and then circularizing the resulting DNA by recombination (*in vivo* copy and swap steps).

Note that, because the site-directed mutagenesis procedure happens almost entirely *in vitro*, there is no opportunity for the cell's efficient natural DNA repair mechanisms to reverse the mutation you are incorporating into the plasmid. This is why it's so much easier to make precise mutations of plasmids than of the chromosome.

DISCUSSION PROBLEM SET #14: DESIGNING CONSTRUCTS FOR GENETIC EXPERIMENTS

Design a detailed protocol for generating a plasmid which will allow **inducible** expression of a protein fusion between the YcaL periplasmic protein of *E. coli* and the red fluorescent protein mCherry, one of the most commonly used of the [mFruit](#) family of fluorescent proteins.

As raw materials, you have wild-type [E. coli MG1655](#) genomic DNA, any plasmid available from [Addgene](#), a reasonable gene synthesis budget, and whatever standard genetic tools you care to use.

Be sure to include **all** of the necessary steps, stating whether they are Read, Write, Copy, Cut, or Paste steps (you need not worry about exact primer sequences and such, but if you are going to use restriction enzymes in any part of the procedure, explain which ones and why).

Draw a map of the plasmid product you intend to construct and explain how it will be used.

What phenotype might you expect to observe when you express the YcaL-mCherry fusion protein in *E. coli*?

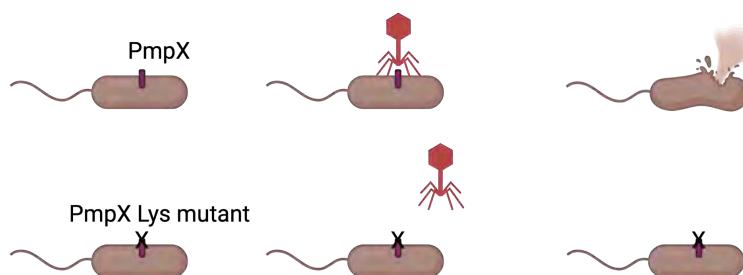
EXAMPLE OF AN EXPERIMENTAL DESIGN: CLONING AND MUTAGENESIS

From now on, when you are designing experiments for this class, I want you to provide a plan for engineering any plasmids that you might need to carry out those experiments, although I don't need you to break those procedures down in quite as step-by-step a way as in Discussion Problem Set #14 above. Here's an example of how that might look, using some of the methods described above.

Many labs are now working to develop bacteriophage-based treatments to kill bacterial pathogens. This is an old idea which has gained more traction as antibiotic resistance becomes more common. However, bacteria can also develop phage resistance by a variety of mechanisms, not all of which are well understood.

Pectobacterium carotovorum is a Gram-negative plant pathogen that causes soft rot disease in a variety of food crops. Application of antibiotics to field crops is problematic for a lot of reasons, so you isolate a phage that can lyse and kill your lab strain of *P. carotovorum*. In infection experiments, the phage work all right for a while, but you consistently isolate spontaneous phage-resistant mutant bacteria which are able to survive the infections.

Sequencing identifies a common feature present in many of the phage-resistant strains: missense mutations changing a lysine to either uncharged or negatively charged amino acids in a predicted membrane protein (which you therefore decide to call PmpX).



Made with Biorender.

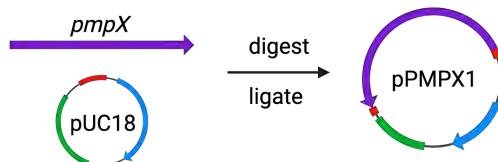
Based on this, you propose a model in which PmpX is the receptor bound by the phage, and that the positive charge of the conserved lysine is important for phage binding. You would like to use the genetic techniques you have learned to test this model.

Hypothesis: Phage infection of *P. carotovorum* requires a positively-charged amino acid at the site of the conserved lysine in membrane protein PmpX.

Experimental Design:

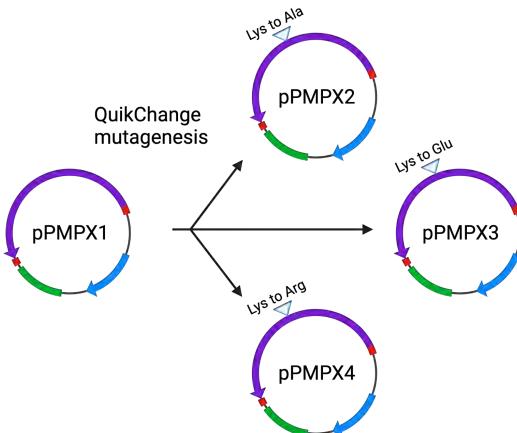
- 1) Construct a $\Delta pmpX$ mutant of *P. carotovorum*. This will eliminate the chromosomally-encoded PmpX protein, and simplify the interpretation of experiments in which PmpX is expressed from a plasmid. (We will discuss methods for generating chromosomal mutations in **Lecture 8**.)
- 2) Amplify *pmpX* with PCR and clone it into a plasmid that will replicate and allow expression of the gene in *P. carotovorum*, calling the resulting plasmid pPMPX1.

A Google Scholar search of the literature suggests that the *E. coli* cloning vector pUC18 is likely to be suitable for use in *P. carotovorum*. In this case, we will add EcoRI and HindIII restriction sites to the *pmpX* PCR product to allow us to do a standard, old-fashioned restriction enzyme / ligase-based cloning procedure, and we will also include the native *pmpX* promoter in our PCR product to ensure that the gene will be expressed in *P. carotovorum*.



Made with Biorender.

- 3) Use the QuikChange site-directed mutagenesis protocol to mutate the conserved lysine of PmpX in pPMPX1 to alanine, glutamate, or arginine, yielding plasmids pPMPX2 through pPMPX4.



Made with Biorender.

- 4) Transform *P. carotovorum* $\Delta pmpX$ with pUC18, pPMPX1, pPMPX2, pPMPX3, or pPMPX4.

- 5) Infect the resulting strains with phage, testing for susceptibility.

Independent Variable: (what will you change?)

- The allele of *pmpX* expressed in each strain.

Dependent Variable(s): (what will you measure?)

- Susceptibility to phage infection (probably a quantitative measurement, since you're likely to count the number of bacteria that survive exposure to phage)

Negative Control(s): (eliminate false positive results)

- 1) The empty-vector (pUC18) control is key here to confirm that phage cannot infect the $\Delta pmpX$ mutant.
- 2) It will also be important to test whether overexpression of any of the *pmpX* alleles has any toxic effect on *P. carotovorum* in the absence of phage infection.

Positive Control(s): (eliminate false negative results)

- Wild-type PmpX (pPMPX1) should restore the ability of phage to infect *P. carotovorum* to the same level seen in wild-type bacteria.

Potential Outcomes:

- 1) The following results would support the original hypothesis:
 - PmpX wild-type: infection
 - PmpX Lys to Ala: no infection
 - PmpX Lys to Glu: no infection
 - PmpX Lys to Arg: infection
- 2) If none of the mutants are susceptible to infection, then that will suggest that it is something specific to Lys that is required for phage binding, not just a positive charge (since Arg is positively charged).
- 3) If all of the mutants are susceptible to infection, then that will suggest that the Lys residue is not important for phage binding, at least in the absence of other mutations.
- 4) If the Glu and Arg mutants are susceptible to infection, but the Ala mutant is not, then that will suggest that the size of the amino acid is more important than its charge (Lys, Glu, and Arg are all large, while Ala is small).

Does this experiment establish correlation or causation?:

- Because we are fulfilling Falkow's postulates, this design tests **causality**.

MODULAR DE NOVO PLASMID DESIGN

In the previous chapter ([Lecture 5](#)), I gave a list of the kinds of sequence features that are found in plasmids. There are many examples of each of these that are more or less well-characterized, but whose sequences are known and can be either amplified or synthesized. In combination with the Gibson Assembly and iVEC cloning protocols, this means that it is possible, when necessary, to construct plasmids entirely from scratch, without relying on someone else having already constructed a vector that is useful for the experiment we are designing.

The idea that biological systems are basically modular and that we can rationally design and engineer them to accomplish particular goals is the basis of the field of *synthetic biology*. The goals of synthetic biologists are much more ambitious than just plasmid engineering (see [these recent papers](#), for example), but we can use the tools and resources developed by them to improve our experimental designs and build vectors useful for achieving our goals.

The most common bacterial platforms for synthetic biology are *E. coli* and *Bacillus subtilis*, and therefore it is easiest to find well-characterized biological parts for those organisms. There are a number of online repositories of such parts. For example, the Joint BioEnergy Institute has one [here](#), but I have found the international Genetically Engineered Machine (iGEM) competition database ([parts.igem.org](#)) to be a particularly helpful resource, since it contains over 20,000 DNA sequences with documented biological functions (mostly in *E. coli*, but also in some other species).

For example, here is a screenshot (taken Oct. 28, 2022) of a table of plasmid origin of replication sequences available in the iGEM Registry of Standard Biological Parts:

Registry of Standard Biological Parts					
Replication origins					
The genetic element responsible for the replication of plasmids during cell growth and division is called a replication origin (also "origin of replication" or simply "origin"). There are several different replication origins and they differ in their plasmid copy number per cell (how many molecules of the plasmid are maintained in the cell), mechanism of copy number control, cell-to-cell copy number variation, and even the degree of coiling of the physical DNA. Thus, BioBrick parts, devices and systems can operate very differently from one plasmid backbone to another.					
Several replication origins have been designed in the Registry. Although most of them are not available as individual BioBrick parts, some can be obtained by PCR of existing plasmid backbones in the Registry using appropriate primers. In fact, when making new BioBrick vectors using the BioBrick base vector, Reshma simply assembled a PCR-amplified linear DNA encoding the replication origins with the antibiotic resistance markers. (The PCR primers included the BioBrick prefix and suffix so that the linear DNA encoding the origin was ready for digestion with EcoRI and SpeI.)					
NOTE:					
Name	Description	Replicon	Copy number	Chassis	Length
BBa_I50000	F plasmid backbone with BioBrick sites removed				4640
BBa_I50001	F plasmid backbone with BioBrick sites removed, reverse				4640
BBa_I50002	Minimal pUC19-derived high copy replication origin				674
BBa_I50003	pBR322 replication origin				2018
BBa_I500032	p15A replication origin				1594
BBa_I50040	nonfunctional pSC101 origin of replication				2226
BBa_I50041	pSC101 origin of replication, reverse				2037
BBa_I50042	pSC101 origin of replication				2037
BBa_I50050	RSF replication origin				312
BBa_I50052	pSC101 temperature sensitive origin of replication				2037
BBa_JR1001	[RSF] Origin of replication				406
BBa_K12125	OrV origin of replication				846
BBa_K125020	lac promoter/insPRT/lac promoter/insPRT/lac				4551
BBa_K125340	Origin of Vegetative Replication				424
BBa_K200000	Asota replication origin				1391
BBa_K220004	Origin for Asota and resistance to Kanamycin				2374
BBa_K524203	pSC101 heat sensitive origin of replication with deletion in par locus				1941
BBa_K533000	p15A replication origin		relaxed, high copy number		914
BBa_K664000	pRSF101 low copy origin of replication				2216
BBa_K664001	pSC101ts temperature sensitive low copy origin of replication				2216

http://parts.igem.org/Plasmid_backbones/Construction/Parts#Replication_origins

Each part is given a systematic name (e.g. BBa_I50000), and can be inspected to get more information (although in some cases that's kind of minimal). For an example, let's take a closer look at BBa_K125340, described in the table as Origin of Vegetative Replication (not a very useful description):

Registry of Standard Biological Parts					
Part:BBa_K125340					
Designed by: Margaret Ruzicka	Organ: iGEM08_Hawaii	(2008-08-29)	Rating: Not Released	DNA: Barcode Not in Stock	
			Experience: None	1 Uses	-Get This Part
Origin of Vegetative Replication					
Plasmids containing oriV and the corresponding Rep proteins (from RSF1010) are autonomously replicated in a broad range of hosts. It has been reported that RSF1010 derived plasmids containing oriV and the Rep proteins are stably maintained in Pseudomonas (Begdassarian 1981; Caulobacter (Umbao-Njaka et al. 2001); Erwinia, Serratia (Lemire 1987), and several cyanobacteria strains (Mermel-Bouvier 1993) including Synechocystis PCC6803 and PCC6714 and Synechococcus PCC7942 and PCC6301..					
Sequence and Features					

http://parts.igem.org/Part:BBa_K125340

Now this is actually fairly informative, and lets us know that this origin of replication would be useful for plasmids in several different Gram-negative genera of bacteria. We can then click through to obtain the actual DNA sequence:

```
>BBa_K125340
AACCCTGCAATAACTGTACGCCCTGCATAACTGTACGAACCCCTGCAATAACTGTACGCCAAACCTGCAAACCGCAGGGGGGGCT
GGCGGGTGTGGAAAAATCATCCATGATTATCTAAAGATAATCCACTAGGCCTGTTACAGGCCCTTGCGCGCTGCTGCCCTGCCAATATGC
CCGGCCAGAGGCCGATAGCTGGCTATTGCGCTAGGCTACACACGCCAACCGCTGCGCCGAGGGGAAAGGCCGAAAGGCCGCTAAACCCC
ACACCAAACCCCGAGAAATACGCTGGAGCGTTTAGCCGCTTACCCGAAGGGTGGGGCGCGTGTGCAGCCCCGAGGGCCT
GTCGGTCGATCATTCAGC
```

I will note, though, that if you read this page closely, this sequence on its own is **not** sufficient for plasmid replication, and you would need to include the corresponding RSF1010 repABC genes as well. You can get the sequences of those from [Genbank](#) easily enough, but this does illustrate the need to be careful about what you are and are not including when engineering a plasmid from scratch and the limitations of this kind of database.

The iGEM Registry is a useful resource for obtaining the sequences of [antibiotic resistance genes](#), [promoters](#), and [terminators](#) for use in *E. coli*, but is a lot less helpful for other species.

IGEM wiki tools search PRODUCTION 2017 SERVER

Registry of Standard Biological Parts

tools catalog repository assembly protocols help search

Promoters/Catalog

Browse by function

- Constitutive promoters:** These promoters are active independent of transcription factors, and are "on" by default.
- Cell signalling:** The registry has a set of promoters related to sending and receiving signals between different cells.
- Metal sensitive:** This set includes promoters that are sensitive to various metals. The promoters are typically regulated by a receptor protein that binds to the metal ion or complex.
- Phage promoters:** A collection of all phage promoters available from the registry. The promoters are often used for very high expression of a protein. These promoters work in *E. coli* and other chassis but typically require a particular RNA polymerase to be present.
- IIT Madras Stresskit promoters:** a well-characterized collection of negatively regulated *E. coli* promoters that have been engineered to be recognized by alternative σ factors. This collection was developed by the 2008 IIT Madras iGEM team.
- USTC logic promoters:** a collection of multiple input promoters all based on a similar template. These promoters were developed by the 2007 USTC iGEM team.

Browse by regulation and RNA polymerase

Most promoters are designed for a particular RNA polymerase (RNAP) holoenzyme, e.g. *E. coli* RNAP bound to a particular σ factor (σ factor is a protein that confers specificity on E. coli RNAP for particular promoter sequences), or the RNAP from the T7 bacteriophage. Promoters are also categorized by the manner in which they are regulated. For example, a promoter can be unregulated (constitutive), positively regulated (activatable) or negatively regulated (repressible).

Regulation	RNA Polymerase						
	<i>E. coli</i>	<i>E. subtilis</i>	Prokaryotic	Bacteriophage	Eukaryotic	Miscellaneous eukaryotic	
Positive (+)	53	1	6	1	14	5	
Constitutive(?) 0	63	9	2	12	1	10	11
Negative(?) -	89	1	-	6	-	3	7
Multiple (+/-)	111	-	1	+	-	5	4

We also have a collection of miscellaneous promoters available that are provided "as is". If you can contribute more information about these promoters, for example, by providing the information that will allow us to add them into one of the categories in the table above, please do so.

<http://parts.igem.org/Promoters/Catalog>

You can, of course, scavenge for useful sequence parts in Addgene or anywhere else plasmid sequences are stored. Software exists to streamline this process (e.g. as described [here](#) and [here](#)). New sets of useful biological parts for synthetic biology are constantly being developed and published for different bacterial species, so it is well worth performing Google Scholar searches for papers describing biological parts that may be useful for your own experiments.

For example, a search for tightly-controlled promoters in *E. coli* might yield this paper describing the optimization of 12 different inducible promoters which can be used simultaneously with minimal interference: [Meyer et al. \(2019\)](#). You might also find the crystal violet-inducible Jungle Express system of [Ruegg et al. \(2018\)](#) useful for certain applications.

If you were building a plasmid for use in *Staphylococcus aureus*, you might find this paper characterizing promoter activity in that species helpful: [Liu et al. \(2022\)](#).

A search for origins of replication that function in a very broad range of bacteria might lead you to this paper describing the construction of an entire toolkit for the honey bee microbiome based on the RSF1010 replicon: [Leonard et al. \(2018\)](#).

If you were looking for easily-visualized reporters, this set of brightly colored proteins might be just the thing: [Lijeruhm et al. \(2018\)](#).

For examples of general-purpose genetic toolkits built this way, take a look at the BglBrick family of plasmids described in [Lee et al. \(2011\)](#), which consists of all possible combinations of 4 origins of replication, 8 promoters, and 3 antibiotic resistance cassettes, designed for use in *E. coli*, but compatible with many Gram-negative species, the similar, but more modern pGinger vector set ([Pearson et al. 2023](#)), or the RhoClo system described in [Jansen et al. \(2024\)](#) for use in a variety of soil-dwelling Gram-positive environmental bacteria.

The possibilities are basically endless, and there is the opportunity here for a lot of creativity. It's not always worth your time to build the Perfect Plasmid for an experiment, but sometimes it has to potential to make your life much easier.

EXAMPLE OF A SYNTHETIC BIOLOGY DESIGN: BUILDING A BETTER VECTOR

To take an example that's come up for my lab recently, let's say you're doing a genetic analysis in the Gram-positive commensal *Streptococcus sanguinis*, and you discover a knockout mutation that has an interesting phenotype. Being a good bacterial geneticist, you want to complement that mutation to establish causality. A literature search identifies a few shuttle vector candidates, and you are able to get in contact with a lab which agrees to send you one, but when you finally receive and sequence it you discover that they have sent you something different from what you expected.

The plasmids available for complementation and gene expression in oral streptococci, like *S. sanguinis* or the cavity-causing pathogen *S. mutans*, are very limited and, in most cases, very old. This means that they are not available from common plasmid repositories and do not have readily available sequences, making it challenging to design experiments using those plasmids, if you can even get them. They also commonly encode only erythromycin resistance, which is a very effective antibiotic for selection of Gram-positive bacteria, but works quite poorly as a selectable marker in *E. coli* and many other Gram-negative species. They're just not very good.

This kind of problem is common when working in less commonly studied organisms, where the genetic tools are often not well optimized. In cases like this, it may be simpler to build an entirely new vector from scratch than to try to get the old plasmid and force it into doing what you want.

Goal: Construct an *E. coli* – streptococcus shuttle vector for complementation of mutants in *S. sanguinis*.

Design Plan:

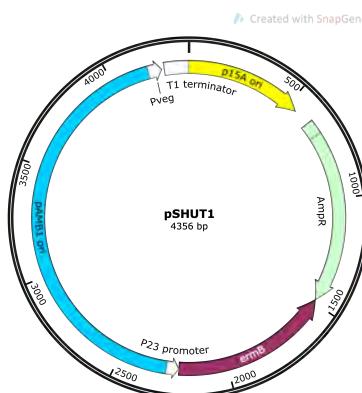
1) First, make a list of the features you would like this plasmid to have:

- a fairly high-copy number origin of replication in *E. coli*
- a reliable *E. coli* antibiotic resistance marker
- a Gram-positive origin of replication known to function in *Streptococcus* species
- a reliable *Streptococcus* antibiotic resistance marker
- a constitutive promoter to drive inserted gene expression in *Streptococcus*

2) Turn to the literature and databases to find sequences for all of those components:

- the *E. coli* p15a origin of replication (1554 bp): iGEM part BBa_J5003
- *E. coli* ampicillin resistance cassette (including a promoter) (943 bp): iGEM part BBa_P1002
- the pAMβ1 Gram-positive origin of replication (1888 bp): from [this paper](#) and [this sequence file](#)
- an erythromycin resistance gene (*ermB*) that should work in *Streptococcus* (738 bp): [this protein](#)
- constitutive P₂₃ promoter (61 bp) to drive *ermB* expression and P_{veg} promoter (62 bp) for inserted gene expression: from [this paper](#)
- a transcriptional terminator to prevent readthrough transcription from P_{veg} (113 bp): iGEM part Bba_J61048

3) Assemble a sequence file with the final desired vector sequence (I used SnapGene to do this, but any equivalent software would work):



I called this construct pSHUT1, since it's a **shuttle** vector.

4) Order synthetic DNA fragments to construct pSHUT1 by Gibson Assembly cloning.

[Twist Biosciences](#) will synthesize linear DNA fragments up to 1800 bp long for 8¢ a base pair, so three of those should be enough. Dividing the 4.4 kb pSHUT1 sequence into three pieces, with 40 bp overlaps at the ends of each fragment to allow assembly, means we'll be ordering three 1.5 kb fragments and paying about \$360. Perfectly reasonable. (A 10-reaction Gibson Assembly kit from New England Biolabs costs \$200, to put this in some kind of context.)

5) Use Gibson Assembly to construct pSHUT1 in *E. coli*, selecting for ampicillin resistance and checking to make sure the resulting plasmid has the sequence you expect (by [Plasmidsaurus](#) full-length sequencing).

Once pSHUT1 is constructed, we should be able to use Gibson Assembly (or iVEC) to insert any gene we want under the control of the P_{veg} promoter and transform the resulting plasmid into *S. sanguinis* for complementation experiments. Be-

cause of the components we chose, it will probably work as a shuttle vector in almost any Gram-positive bacterium and we should submit it to [Addgene](#) so that others can use it too.

Throughout the rest of this course, if you really want to build a new plasmid to accomplish some experimental goal and you can identify likely components to construct it out of, feel free to propose that. I **will**, however, expect you to explain your protocol for constructing the plasmid in question.

DISCUSSION PROBLEM SET #15: DESIGNING CONSTRUCTS FOR GENETIC EXPERIMENTS 2

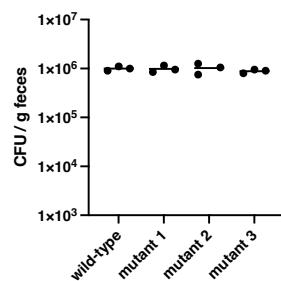
Problem #1

Citrobacter rodentium is a pathogenic Gram-negative bacterium that infects mice and is used as a model for enterohemorrhagic *E. coli* infections, since the diseases caused by the two are very similar, but *C. rodentium* is much less able to cause disease in humans.



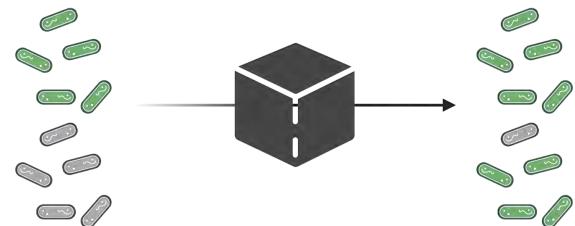
A healthy mouse on the left, and a mouse with a serious stomachache on the right, caused by *C. rodentium* infection. Made with Biorender.

You have identified a series of *C. rodentium* genes that, based on their regulation, you suspect might be involved in colonization of the mouse intestine, but infecting mice with each individual mutant results in about the same CFU / g of bacteria in the feces as with wild-type (determined by plate count).

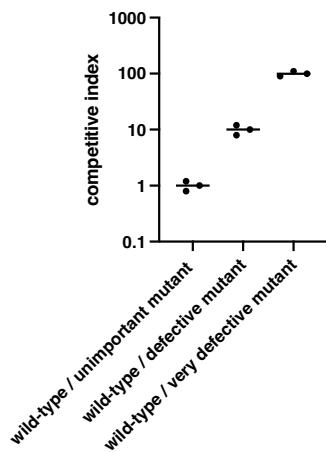


This is actually fairly common in cases where a phenotype is impacted by multiple factors, since other genes may be able to compensate for the loss of one gene. We need a more sensitive method to assess whether individual genes play a role in complex phenotypes.

This is where **competitive index** (CI) experiments come in handy. In a CI experiment, we take an equal mixture of mutant and wild-type cells, compete them against each other under the conditions of interest, then measure the **ratio** of mutant to wild-type cells afterwards. If the mutant is less able to compete, then the wild-type will have a growth advantage. This can amplify subtle phenotypic differences:



An equal mixture of green and grey bacteria pass through a black box, after which the ratio of green to grey bacteria increases. This is good evidence that the grey bacteria have a competitive disadvantage under whatever condition is represented by the box. Made with Biorender.



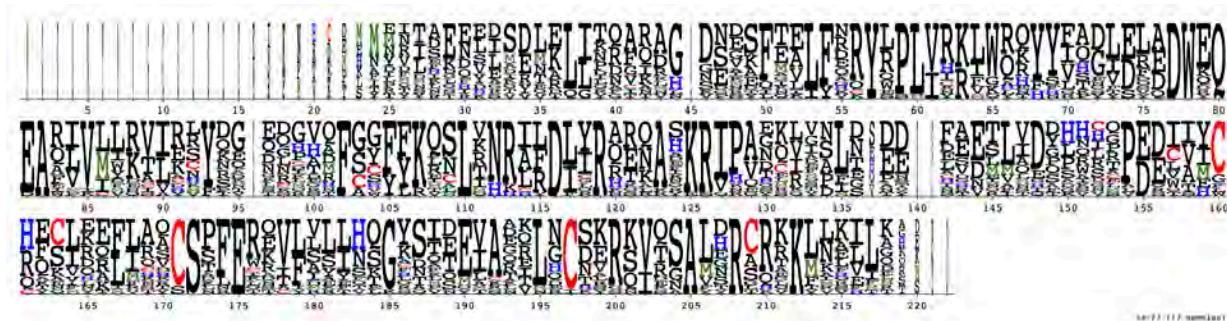
In this example figure, you can see that the wild-type and the first mutant are equally able to compete in this environment, while the other mutants have progressively more serious defects in competition with the wild-type strain.

In order to use a CI approach to testing your *C. rodentium* mutants, you need a way to easily distinguish between wild-type and mutant *C. rodentium* bacteria in mouse fecal samples and count them. Propose an approach to solving this problem **using plasmids**, and state a detailed description of how you will construct those plasmids.

Most plasmids, reporters, and promoters that work in *E. coli* will also work in *C. rodentium*.

Problem #2

The gut-inhabiting lactic acid bacterium *Lactobacillus reuteri* has only one alternative sigma factor: SigH, which you suspect controls gene expression in response to changes in oxygen levels. SigH homologs are found in many lactobacilli, so you generate an alignment of SigH sequences from 17 different species:



Based on this alignment and knowing that the redox state of cysteine residues often regulates protein activity (see [Lecture 4](#)) you hypothesize that Cys171 and Cys197 are required for oxygen sensing by SigH.

Propose an experiment **using plasmids** to test this hypothesis. State:

- a detailed description of how you will construct the necessary plasmids

(Note that there are a handful of useful *Lactobacillus - E. coli* shuttle vectors available. For the purposes of this experiment, you can use [pTRKH2](#) or build your own.)

- the independent and dependent variables
- both positive and negative controls
- whether your proposed experiment tests correlation or causation
- potential outcomes of your experiment, and how you will interpret them

LECTURE 8: GENE TRANSFER AND RECOMBINATION

INTRODUCTION

In this lecture, we will discuss how genetic material can be transferred between bacterial cells, both naturally and in the lab. This will lead to a discussion of homologous recombination and techniques for genetic engineering that depend on gene transfer and recombination. We will design experiments using these techniques and discuss the benefits and disadvantages of such approaches.

GENE TRANSFER IN BACTERIA

There are three main ways by which a bacterial cell can take up new DNA: *transformation*, *conjugation*, and *transduction*. All of these occur in nature, and are mechanisms by which bacteria can acquire new genetic material from other more or less distantly related organisms (*horizontal gene transfer*).

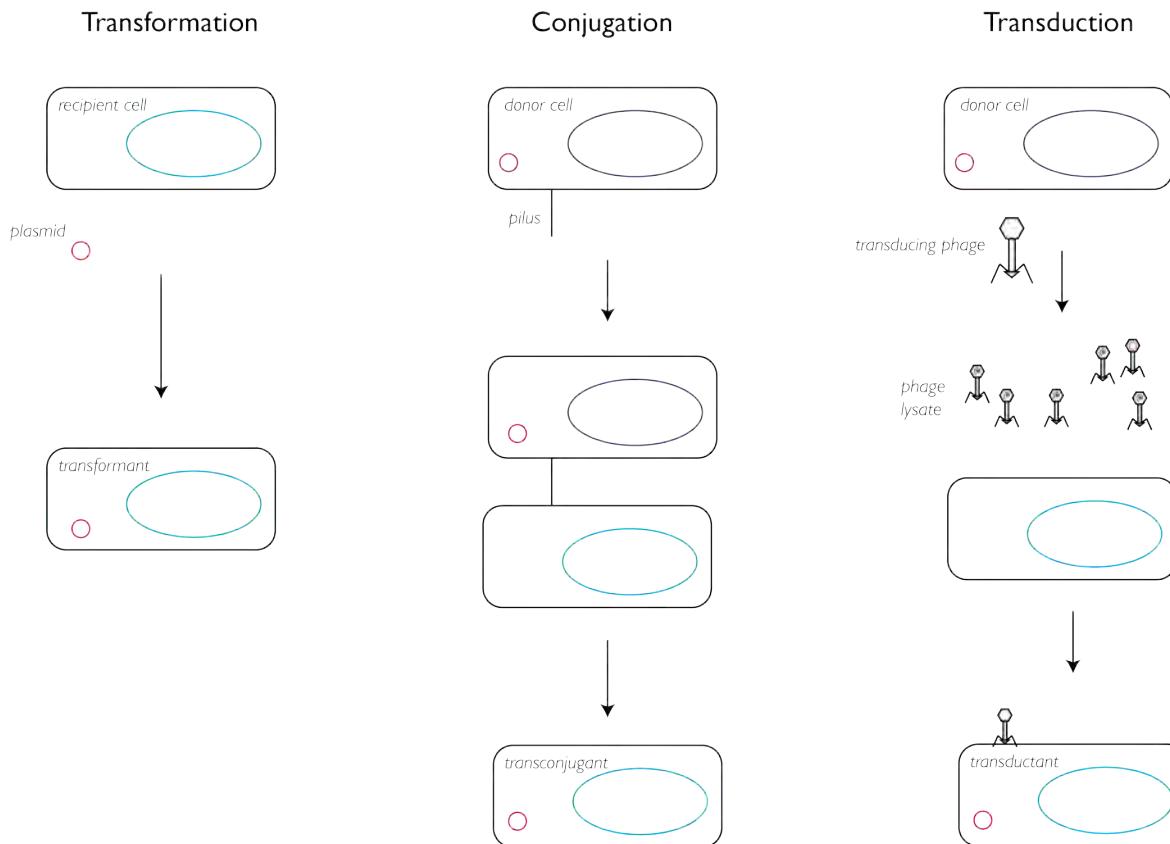


Figure 8.1. Moving a plasmid (red circle) into a recipient cell (blue chromosome) by three different methods. Note that transduction also results in phage particles containing fragments of chromosomal DNA from the donor cell (purple chromosome), which may then be transferred into the recipient cell. See the Common Protocols section of this chapter for more details.

Transformation is a process in which bacterial cells directly take up DNA from their environment and incorporate it into their genetic material. Cells that can do this are called *competent cells*. It's called transformation because the uptake of new genes transforms the phenotype of a strain. In fact, Oswald Avery's 1944 experiments showing that adding very pure DNA could change the colony morphology phenotype of the pneumonia-causing pathogen *Streptococcus pneumoniae* were among the first pieces of evidence that DNA is the genetic material of cells. (Note that for eukaryotic cells, direct uptake of DNA is called "transfection" instead of "transformation".)

Some species are *naturally competent* (e.g. *S. pneumoniae*, *Bacillus subtilis*, *Neisseria gonorrhoeae*) and will take up DNA from their environment on their own, but most species require special treatment to allow transformation. *E. coli* and some other Gram-negative bacteria can be made *chemically competent* by resuspending them in very cold CaCl_2 solutions and then briefly heat shocking them at 42°C (a method based on a 1970 discovery by Morton Mandel and Akiko Higa). Many types of cells can be transformed by *electroporation*, in which cells are mixed with DNA in a cold, low ionic-strength solution then subjected to an electric shock. These methods are thought to work by disrupting the cell membrane enough to allow DNA through.

Conjugation is a process in which bacterial cells form hair-like tubular structures (called *pili*, which is the plural of *pilus*) on their surfaces and transport single-stranded DNA (copied from a double-stranded DNA molecule, usually a plasmid, in the

donor cell) through those pili into other cells. The genetic elements that allow specific DNA molecules to be conjugated are called *tra* factors (short for transfer). Only DNA molecules containing an origin of transfer (*oriT*) for the particular *tra* system in a donor bacterium can be conjugated.

Note that conjugation is not species-specific and is in fact a common method used in the lab for transferring DNA from easy-to-work-with species (like *E. coli*) into species that are more challenging to transform, which do not even necessarily have to be bacteria. In nature, the plant pathogen *Agrobacterium tumefaciens* (which causes crown gall disease) conjugates a genetic element called T-DNA into plant cells, causing formation of tumors in the host plant, and T-DNA-based vectors are historically common tools for genetic engineering of plants.

Conjugation was discovered in *E. coli* in 1947 by Joshua Lederberg and Edward Tatum in the first demonstration that bacteria can mate and exchange genes, a discovery which really made bacterial genetics (and, eventually, molecular biology) possible. Tatum, Lederberg, and George Beadle later won the 1958 Nobel Prize for this and other contributions to molecular genetics.

Transduction is the use of bacteriophage (recall that these are viruses that infect bacteria, often just called phage) to transfer DNA from one bacterial strain to another. Generalized transducing phage are phage which are able to package plasmids or random fragments of DNA from the chromosome of their host cell into virus particles. These particles can then attach to and inject that DNA into another bacterial cell, where it can potentially be incorporated into the host chromosome by homologous recombination (see below). This is contrasted with specialized transducing phage, which are less useful and are only able to package host genes directly adjacent to the single site where the phage integrates into the host chromosome.

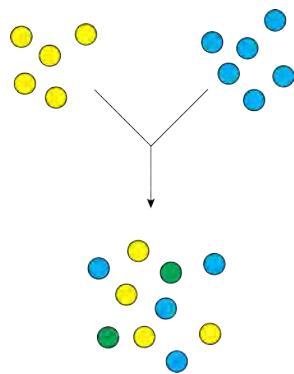
Most wild-type phage normally only package viral DNA, of course, but many lab strains of generalized transducing phage have been selected to package host DNA at higher frequency. (About 1 in 30 phage particles produced during a P1 vir infection of *E. coli* contains a random fragment of host DNA instead of phage DNA, for example.)

Like other viruses, phage are typically extremely species- or even strain-specific. The P1 phage will only work for transductions in certain strains of *E. coli* and *Klebsiella pneumoniae*, for example, while the P22 phage is specific for *Salmonella enterica*. Generalized transduction (by phage P22) was discovered by Joshua Lederberg and Norton Zinder in 1951, and specialized transduction (by phage λ) was discovered by Esther Lederberg in 1956.

These three mechanisms may not be the **only** ways bacteria can exchange genetic information. A recent report from Eleftherios Papoutsakis' lab (linked [here](#)) describes a startling phenomenon that occurs in co-cultures of *Clostridium acetylbutylicum* and *Clostridium ljungdahlii* in which cells of the two different species are able to fully fuse, mixing their cytoplasms, proteins, and DNA. How this happens and how the hybrid cells are resolved back into their constituent species is not known.

DISCUSSION PROBLEM SET #16: GENE TRANSFER

While studying antibiotic resistance in *S. aureus*, you discover that incubating an erythromycin-resistant strain together with a chloramphenicol-resistant strain in liquid media **without** antibiotics results in the appearance of some cells that are resistant to **both** antibiotics.



Design an experiment or experiments to determine whether this genetic exchange occurs via transformation, transduction, or conjugation. State:

- the hypotheses you are testing
- the independent and dependent variables
- both positive and negative controls
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiment, and how you will interpret them

PRACTICAL CONSIDERATIONS

When working in the lab, there are some practical considerations you should take into account when attempting to move a particular piece of DNA into a bacterial strain:

1. Is the gene you want to move on the chromosome or on a smaller DNA element like a plasmid? Highly competent cells with efficient recombination systems (see below) may be able to take up and incorporate genomic DNA, but this is likely to result in incorporation of a lot of genetic material from the donor. Transduction can move fragments of chromosomal DNA, but generally only between fairly closely related strains. It is important to remember that, while plasmids **can** be moved from one cell to another by transduction, it is usually less convenient than the other two methods.
2. Are generalized transducing phage or conjugative plasmid systems available for your model organism? While these tools exist for many species, they have not been developed for all bacteria. It has become particularly unfashionable to identify generalized transducing phage for new model organisms. (It's a lot of work to do, and not that many labs are equipped to do it anymore.)
3. Can you easily make the bacteria you are working with competent? If so, transformation is likely to be the most convenient method to move a plasmid into those cells.

Moving DNA between species can present a particular challenge. Most bacteria possess defense mechanisms that will attempt to break down any foreign DNA molecules that enter their cells. These include restriction enzymes, which we have discussed as molecular tools, and which recognize and cut specific DNA sequences. As mentioned in [Lecture 7](#), in nature these function to protect bacteria against attack by phage, and what they restrict is the ability of particular phage to infect that strain. The bacterium protects its own DNA from restriction digestion with a sequence-specific *restriction methylase* that adds a methyl group to the DNA sequence recognized by its cognate restriction enzyme, preventing them from being cut. If you are trying to move DNA into a cell with a restriction enzyme system from a cell without the appropriate methylase, the transformation efficiency will be very low. Daisy Dussoix, a graduate student in Werner Arber's lab, was the first to recognize the existence of restriction-modification enzyme systems and their effects on DNA transfer around 1960.

HOMOLOGOUS RECOMBINATION

When a bacterial cell takes up a plasmid with an appropriate origin of replication, the plasmid is able to replicate and be maintained in that cell and its descendants. DNA molecules **without** their own origin of replication have to be incorporated somehow into the host chromosome in order to be passed down to the next generation. One very common mechanism by which this can occur is known as *homologous recombination*. (For many more details and links, [this Wikipedia article](#) is a great place to start.)

The main physiological function of homologous recombination in cells is in DNA damage repair, and the complex details of its mechanism are well beyond the scope of this course. However, it is important to have a general sense of what it does, since many genetic engineering procedures depend on it.

As shown in Figure 8.2, when there are two pieces of DNA in a cell with similar sequences, the RecA single-stranded DNA binding protein and RecBCD recombinase proteins can recognize single- or double-strand breaks in those DNA molecules, bind to them, and create stretches of hybrid base-paired DNA with crossover points called *Holliday junctions* (after Robin Holliday, who first proposed their existence in 1964). This requires stretches of DNA with very similar sequences at the crossover points, which is why this is called **homologous** recombination. The higher the homology between the two DNA molecules, the more likely RecABCD is to be able to generate Holliday junctions.

Homologous Recombination To Incorporate A Linear DNA Fragment

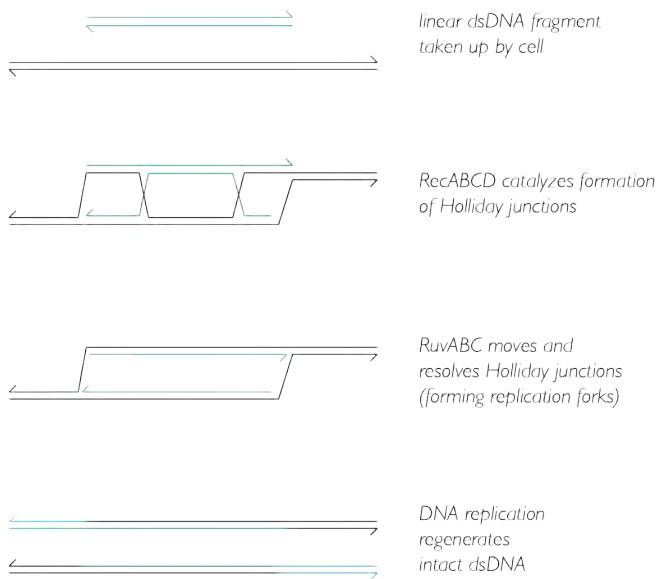


Figure 8.2. A very simplified diagram illustrating incorporation of a linear DNA fragment into a bacterial chromosome by homologous recombination.

The Holliday junctions are resolved into replication forks by the RuvABC complex, and subsequent DNA replication results in two intact chromosomes, each of which incorporates the new DNA on one strand, where it may be inherited by some of the cell's progeny or repaired by other DNA repair mechanisms (e.g. the mismatch repair system). As you might expect, mutants lacking any one of the *rec* or *rvu* genes are unable to carry out homologous recombination and are extremely sensitive to DNA-damaging chemicals and radiation.

Many cloning strains of *E. coli*, which contain mutations to make them more competent and easier to work with in the lab (common examples are the strains DH5 α and JM109), are *recA* mutants. This is to prevent the possibility of recombination between genes on plasmids and genes on the chromosome and also improves plasmid stability. It does, however, mean that no genetic engineering protocols that rely on RecA-dependent homologous recombination will work in these strains.

Lysogenic bacteriophage encode their own recombinases, which they use to integrate their genomes into the host chromosome at a specific site (the *att* site, which is different for each phage). The phage recombinases that are most useful for molecular biology are not sequence-specific and require much shorter regions of homology than RecABCD to stimulate recombination, which has made them powerful tools for molecular genetics, as we will discuss below.

USING HOMOLOGOUS RECOMBINATION FOR GENETIC ENGINEERING

Swap

As mentioned in the last chapter, one of the six fundamental procedures in molecular biology is recombination, which I abbreviated as Swap, since it results in swapping or exchanging sequences from one DNA molecule to another. In the next section of this chapter, I will describe a variety of genetic engineering protocols that depend on recombination, to illustrate what is possible. Swap steps always happen *in vivo*. The DNA recombination machinery is very complex, and reconstituting it *in vitro* is impractical for general purposes.

The protocols described in the previous chapter were useful for engineering plasmids, which are relatively easy to manipulate. Recombination allows us to expand our toolkit and generate site-directed mutations in the bacterial **chromosome**. This means that these protocols allow us to make site-directed mutations in the chromosome, and are especially useful for reverse genetics, in which we observe the phenotypes caused by specific genetic changes.

It is important to know the difference between *single-crossover* and *double-crossover* recombination events, which are distinguished by requiring either one or two independent homologous recombination events. Single-crossover recombination occurs much more frequently, and combines two DNA molecules into a single product. This is another reason that cloning strains are usually *recA* mutants, since recombination between two circular plasmids can result in concatemer formation (difficult to detect most of the time, but surprisingly common).

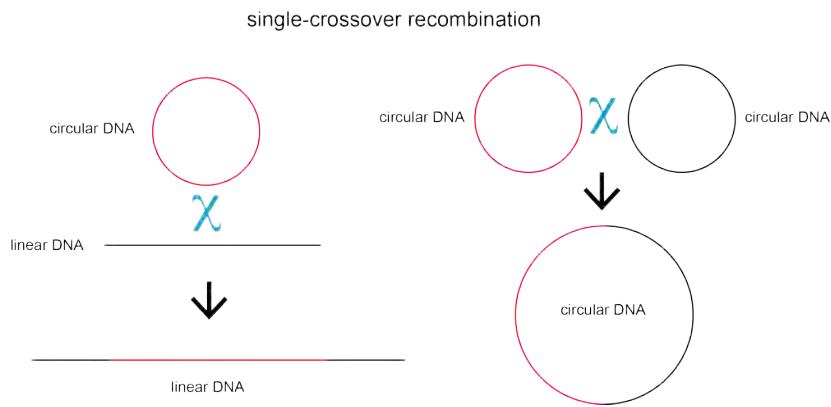


Figure 8.3. The results of single-crossover homologous recombination events between a circular and a linear DNA molecule (on the left) or between two circular DNA molecules (on the right). Recombination between two identical plasmids can result in concatemer formation.

More relevant to genetic engineering approaches that use recombination, single-crossover recombination between one linear and one circular DNA molecule results in a linear product, which, if the circular DNA was the chromosome, constitutes a lethal double strand break. The protocols described in this chapter that rely on recombination between linear and circular DNA almost exclusively rely on double-crossover recombination, which generates two products, but no DNA strand breaks in the circular molecule.

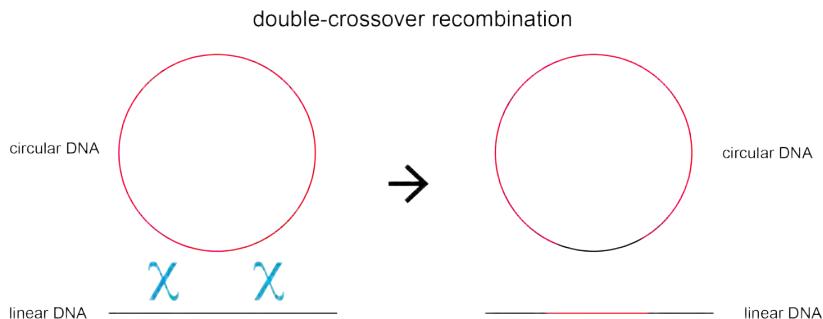


Figure 8.4. The results of a double-crossover homologous recombination event between a circular and a linear DNA molecule. Double crossover recombination between two circular DNA molecules results in two circular DNA molecules.

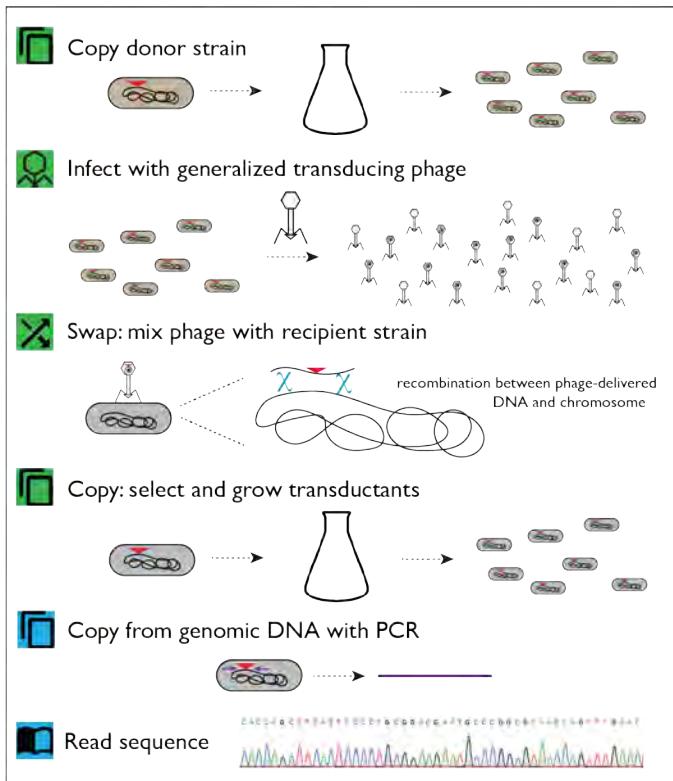
In either case, but especially when demanding a double-crossover product, recombination is a rare event, so having a strong selection for strains that contain the desired final product is essential. This is usually accomplished by including an antibiotic resistance gene (commonly encoding chloramphenicol, kanamycin, or tetracycline resistance, but any antibiotic resistance gene should work in principle) in the DNA that you want to incorporate into the chromosome.

EXAMPLES OF COMMON MOLECULAR BIOLOGY PROTOCOLS

In the next section, I will break down a series of protocols into their component steps, both in outline and graphical form. The protocols in this section all depend on recombination, although most of them **also** require some amount of plasmid engineering, which can be done using the methods described in the previous chapter.

TRANSDUCTION

As described above, generalized transduction uses transducing phage to transfer selectable markers between bacterial strains, which are then incorporated into the chromosome of the recipient cell by homologous recombination.



Protocol:

1. Copy – *in vivo*

- grow donor cells containing the selectable marker you plan to transduce

2. Infect – *in vivo* (not really one of the six steps...)

- infect donor cells with a generalized transducing phage for the species you are working with and harvest phage particles, some of which will contain DNA from the donor cell chromosome

3. Swap – *in vivo*

- add phage containing donor cell DNA to recipient cells, and allow time for DNA injection and recombination to occur
- you will need to include a step to **stop** the phage infection, since most of the phage particles you added will be virulent; for P1 and P22, this is done by chelating away calcium, which many (but not all) transducing phage require for attachment, but different methods may be appropriate depending on the transducing phage in question

4. Copy – *in vivo*

- select for recombinants and grow them

5. Copy – *in vitro*

- use PCR to amplify the region of the chromosome containing the desired mutation

6. Read – *in vitro*

- sequence the PCR product to confirm that the selected transductant has the desired sequence derived from the donor strain

Instead of steps 5 and 6, you could (and arguably should) just purify genomic DNA from the cells grown in step 4 and have the entire genome sequenced, which would still be a Read step, but is less focused on the particular locus you're modifying.

Since successful transfer and incorporation is a relatively low-frequency event, a selection is required to identify successful *transductants*. When the mutation you want to move is itself selectable (e.g. an insertion containing an antibiotic resistance gene), this is straightforward. However, since transducing phage package large fragments of host chromosomal DNA (100 kb in the case of the *E. coli* transducing phage P1, for example, or 40 kb for the *Salmonella* phage P22), just having a selectable marker **near** your mutation of interest (a *linked marker*) is sufficient. This could be, for example, a transposon insertion in a nearby gene or intergenic region. The closer two mutations are on the chromosome, the more frequently they will be *cotransduced*, although you will certainly need to screen to make sure that your mutation of interest was actually incorporated into the recipient strain along with the linked marker.

Note that step 5, using PCR to amplify the genomic region containing the putative mutation for sequencing, can be done with purified genomic DNA or, for many species, simply by suspending some cells in the PCR reaction mix. This is called colony PCR, and works because the 95–98°C melting step of the PCR cycle lyse some of the bacteria, releasing their DNA into solution.

ALLELIC EXCHANGE

Allelic exchange procedures involve the construction of plasmids containing the desired mutant allele, which are then recombined into the chromosome of the recipient strains using the native RecA-dependent recombinase activity of that strain (Figure 8.5). RecA usually requires long regions of homology for recombination to occur (500 to 1000 bp).

Normally, allelic exchange templates will consist of a suicide vector containing an *antibiotic resistance cassette* (a gene encoding a product that confers antibiotic resistance, along with all of the additional sequences needed to ensure its expression) flanked by sequences homologous to the target region in the host chromosome. This makes it straightforward to select for recombinants on plates containing the relevant antibiotic. Any of the plasmid construction methods described in the previous chapter can be used to construct this vector.

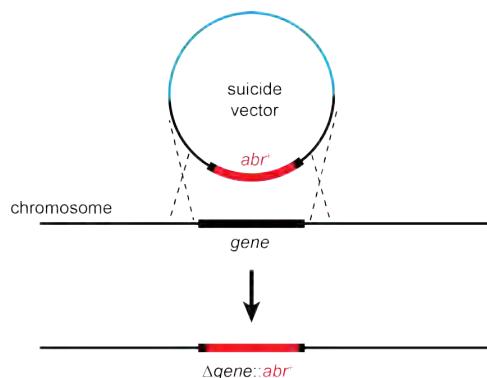
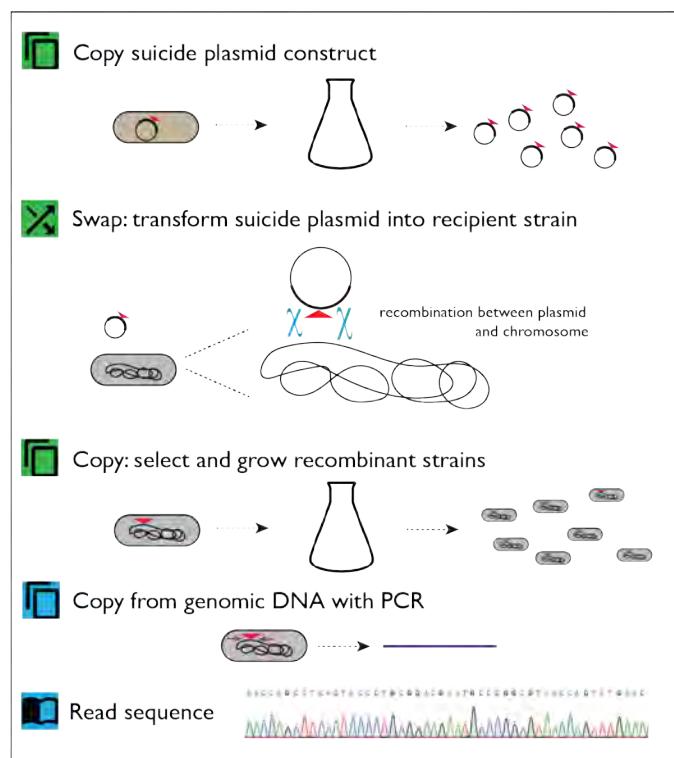


Figure 8.5. Double-crossover allelic exchange mutagenesis, replacing gene with antibiotic resistance marker *abr*⁺. The desired gene:*abr*⁺ mutant is selected for with the antibiotic for which *abr* encodes resistance. In practice, this is usually tetracycline, kanamycin, or chloramphenicol.

Use of a suicide vector that **cannot** replicate in the recipient strain is essential. Otherwise selecting for antibiotic resistance will only yield transformants (cells containing replicating plasmids), not recombinants. Transformation is orders of magnitude more efficient than recombination. See **Lecture 5** for some more details on suicide vectors.



Protocol:

1. Copy – *in vivo*

- grow cells containing suicide plasmid with desired allele and make a large amount of it

2. Swap – *in vivo*

- transform the suicide plasmid into recipient cells and allow time for recombination to occur

3. Copy – *in vivo*

- select for recombinants and grow them

4. Copy – *in vitro*

- use PCR to amplify the region of the chromosome containing the desired mutation

5. Read – *in vitro*

- sequence the PCR product to confirm that the selected strain has the desired mutation

While the protocol above illustrates allelic exchange with double-crossover recombination, do note that this is the only protocol in this chapter which **can** (and honestly, in practice, almost always **does**) use single-crossover recombination, since the suicide vector is a circular DNA molecule and no double-stranded DNA breaks are made in the chromosome when single-crossover recombination occurs.

In this case, the Swap step involves one recombination step to integrate the plasmid into the chromosome, followed by a second single-crossover recombination step to remove the integrated plasmid from the chromosome, which will (about 50% of the time) result in the chromosome containing the allele that was originally in the vector (Figure 8.6). This has the very significant advantage that the antibiotic resistance gene need not end up in the final product, and you can therefore make non-polar in-frame deletions this way (or other more subtle genetic changes, like single amino acid mutations).

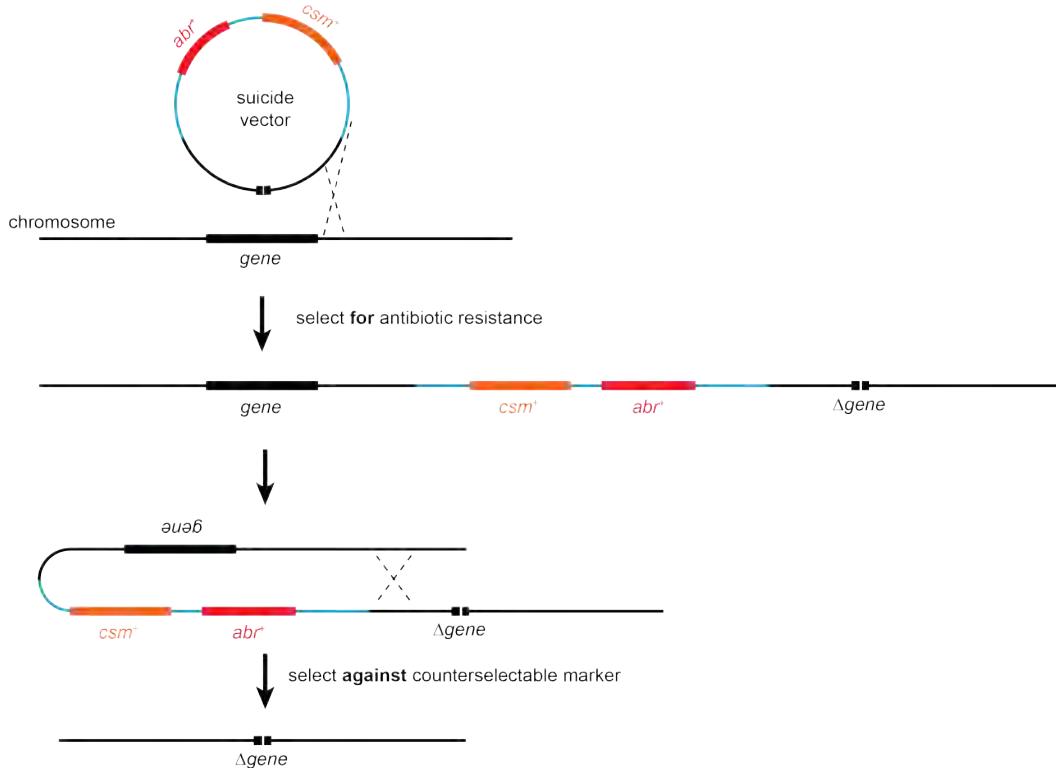


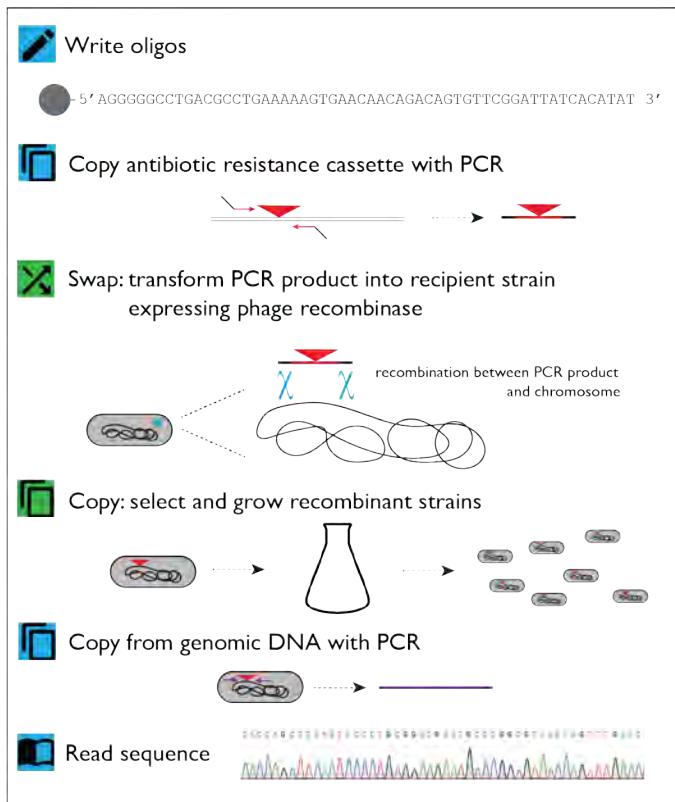
Figure 8.6. Single-crossover allelic exchange mutagenesis. to construct a deletion of gene. In the first step, antibiotic resistance is used to select for recombination of the suicide vector into the chromosome. In the second step, strains in which recombination between homologous sequences at the site where the vector integrated has occurred are selected for by selecting against the counterselectable marker *csm⁺*. In practice, this is often the sucrose-sensitivity gene *sacB*, but many other counterselectable markers are available. Only cells which retain *abr'* and then subsequently lose *csm⁺* will survive this selection gauntlet. Roughly 50% of the time, this will result in the deletion remaining in the chromosome instead of the wild-type gene.

This selection / counterselection allelic exchange method is a very common procedure for site-directed mutagenesis of chromosomal genes in many bacteria (perhaps the **most** common, outside of those few model species with well-developed recombineering tools – see the next few sections).

RECOMBINEERING

Recombineering uses double-stranded linear DNA fragments (typically PCR products) as templates for recombination in cells expressing highly active phage recombinases that can integrate DNA fragments with as little as 40 to 50 bp of se-

quence homology to the host chromosome. The PCR products used for recombineering almost always contain an antibiotic resistance gene to allow selection of recombinants.



Protocol:

1. Write – *in vitro*

- design PCR primers that amplify an antibiotic resistance cassette
- add sequences homologous to the desired insertion site in the chromosome to the 5' end of the primers

2. Copy – *in vitro*

- PCR amplify an antibiotic resistance cassette using the primers designed in step 1

3. Swap – *in vivo*

- transform the PCR product into recipient cells expressing a phage recombinase and allow time for recombination to occur

4. Copy – *in vivo*

- select for recombinants

5. Copy – *in vitro*

- use PCR to amplify the region of the chromosome containing the desired mutation

6. Read – *in vitro*

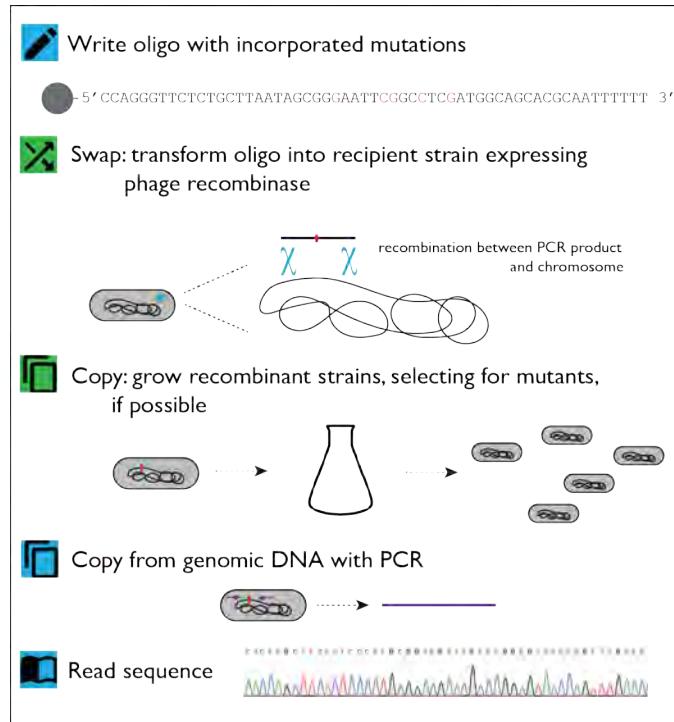
- sequence the PCR product to confirm that the selected strain has the desired mutation

The phage recombinase (usually the Red recombinase from phage λ , especially in Gram-negative bacteria) must be expressed from an inducible promoter on a plasmid, which can be constructed using any of the plasmid construction methods described in **Lecture 7**. Recombinase expression plasmids for recombineering also often have temperature-sensitive origins of replication or counterselectable markers to make curing that plasmid easy after the desired chromosomal mutation(s) have been made. It is not generally healthy for bacteria to constitutively express recombinases, which can lead to unwanted chromosome rearrangements.

OLIGO-DIRECTED RECOMBINEERING

The PCR products used as recombination templates in the previous protocol are double stranded DNA molecules. The phage recombinases used for recombineering also work with single-stranded DNA templates, like oligos. It is only practical to order oligos up to about 100 bp long, so oligo-directed recombineering cannot be used to insert large sequences (like

antibiotic resistance genes), but if a point mutation you're interested in has a selectable or easily screenable phenotype, this approach can work very well.



Protocol:

1. Write – *in vitro*

- design an oligo homologous to the bacterial chromosome with the desired mutation near its center
- adjust the sequence of the oligo to avoid mismatch repair and increase recombination efficiency (see below)

2. Swap – *in vivo*

- transform the mutagenic oligo into recipient cells expressing a phage recombinase and allow time for recombination to occur

3. Copy – *in vivo*

- select for recombinants, if possible, or dilute and plate for individual colonies to screen

4. Copy – *in vitro*

- use PCR to amplify the region of the chromosome containing the desired mutation

5. Read – *in vitro*

- sequence the PCR product to confirm that the selected strain has the desired mutation

When recombineering primers are carefully designed to avoid triggering the host cell's DNA repair mechanisms it is sometimes possible to generate non-selectable alleles, including point mutations, using this method. Of course, in this case, you need to screen the resulting colonies to determine which ones contain your desired mutation, usually by sequencing the affected gene. Efficiency may be quite low (< 1%), however, making this a labor-intensive approach. See below for how we can use CRISPR to greatly increase the efficiency of oligo-directed recombineering.

The bacterial mismatch repair system does not work well on mutations that change 5 or more sequential nucleotides or on several closely spaced point mutations (as shown in the figure above), although the details of **why** this is true are beyond the scope of this course. From a practical genetic engineering standpoint, though, this means that making 3 or 4 silent mutations directly adjacent to the mutation of interest can improve mutation efficiency greatly, as can synthesizing the oligo with more-stable phosphorothioate linkages between the nucleotides at the 5' or 3' ends (as opposed to the naturally-occurring phosphodiester bonds). Mutagenic oligos are also more efficiently incorporated into the chromosome when they are complementary to the lagging strand during DNA replication, possibly because the cell mistakes them for Okazaki fragments.

CRISPR

The most recent major addition to the molecular genetics toolkit is CRISPR, which stands for **c**lustered **r**egularly **i**nter-spaced **p**alindromic **r**epeats. The name is derived from the fact that *CRISPR arrays* of short, repetitive DNA sequences were observed in many bacteria and archaea long before their function was known. In the early 2000's, Philippe Horvath and

Rodolphe Barrangou, working for the food company Danisco, realized that CRISPR was involved in protecting the yogurt-fermenting bacterium *Streptococcus thermophilus* from bacteriophage, and in fact, functioned as a kind of adaptive immune system for those bacteria. The CRISPR array contains short pieces of DNA derived from parasitic phage or plasmids, and the CRISPR-associated (Cas) proteins are then able to recognize and bind to the matching sequences in those parasites and cause double strand breaks in their DNA, protecting the bacterium from infection.

Extraordinary (and [recently Nobel-winning](#)) work from many labs (including those of Jennifer Doudna, Emmanuelle Charpentier, and Feng Zhang) has turned this bacterial defense system into a bioengineering tool that can efficiently introduce double- or single-strand breaks in nearly **any** targeted DNA sequence. In the framework I have laid out for genetic engineering processes, this is an *in vivo* Cut step. The nuclease most often used in genetic engineering protocols is called Cas9, which is directed to a roughly 20 nucleotide target sequence by a guide RNA (gRNA) that base pairs to the target.

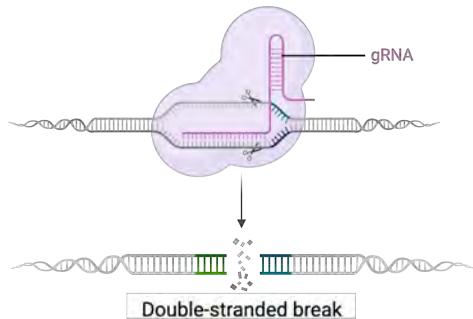


Figure 8.7. The basic function of CRISPR. Directed by the guide RNA (gRNA), Cas9 makes a double strand break at a specific DNA sequence. Made with Biorender.

There are many other variations on CRISPR, which we do not have space to discuss in detail here. For example, using CRISPR with inactive Cas proteins that bind DNA but do not cut it (dCas9) allows precise targeting of proteins fused to the inactive Cas protein to specific DNA sequences. This has been used with fluorescent fusions to dCas9 to visualize where particular DNA elements are found in cells, and, since dCas9 bound to a gene or promoter blocks the progress of RNA polymerase, has also been used as a tool to inducibly repress gene expression (resulting in a gene knock-down instead of a gene knockout). More and more sophisticated CRISPR-based tools are being developed all the time. Here's [a 2023 paper](#), for example, that describes a very flexible CRISPR-driven system for rapid engineering of the *Bacillus subtilis* genome.

CRISPR is a tremendously versatile and powerful tool. It works in both bacteria and in eukaryotic organisms, and is far simpler and faster than other techniques for genetic manipulation of eukaryotes (although this may not be true for bacterial systems, where many different genetic engineering technologies exist). The development of CRISPR-based technologies has stimulated an active debate in the scientific community about the ethics of genetic engineering in higher organisms.

CRISPR-ASSISTED RECOMBINEERING

The most common use of CRISPR in bacterial genetics is in combination with oligo-directed recombineering, where it is used as a selection after the recombination step. As discussed above, many recombineering procedures have low efficiency, especially when they are used to generate point mutations. Combining recombineering with a CRISPR system that targets the wild-type sequence for double strand breaks efficiently kills any cells that are not mutated.

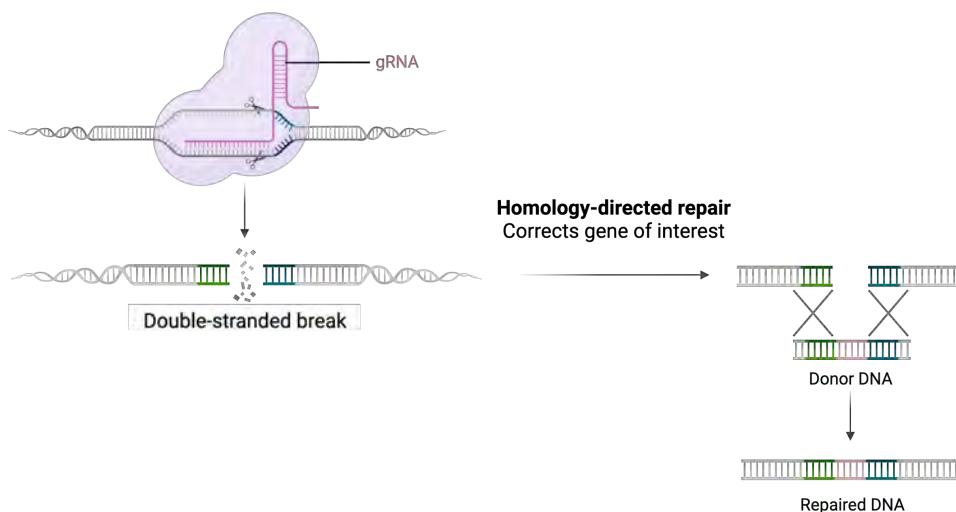
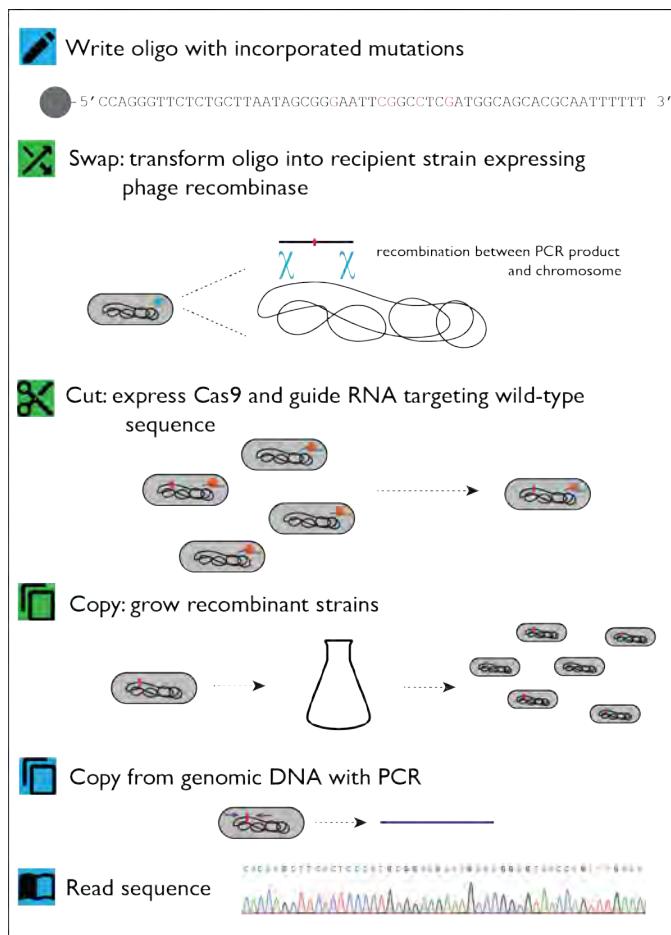


Figure 8.8. Using CRISPR in combination with recombineering. Directed by a guide RNA (gRNA) specific to the wild-type sequence, Cas9 makes a double strand break that kills any cells in which homology-directed repair (i.e. recombination of the mutated template with the chromosome) has **not** occurred, selecting for successful mutagenesis. Note that the donor DNA template is illustrated here as dsDNA, but ssDNA oligos also work, when an appropriate phage recombinase is present. Made with Biorender.



Protocol:

1. Write – *in vitro*

- design an oligo homologous to the bacterial chromosome with the desired mutation near its center
- adjust the sequence of the oligo to avoid mismatch repair and increase recombination efficiency

2. Swap – *in vivo*

- transform the mutagenic oligo into recipient cells expressing a phage recombinase and allow time for recombination to occur

3. Cut – *in vivo*

- simultaneously, express Cas9 and a guide RNA targeting the wild-type sequence

4. Copy – *in vivo*

- plate recombinant colonies, most of which will have the desired mutation

5. Copy – *in vitro*

- use PCR to amplify the region of the chromosome containing the desired mutation

6. Read – *in vitro*

- sequence the PCR product to confirm that the selected strain has the desired mutation

In most of the bacterial systems I have seen, the recombinase and Cas9 are expressed from one plasmid, while the guide RNA is expressed from another. One additional cloning step is therefore required compared to the previous protocol: constructing a plasmid that will express the guide RNA to target Cas9. As with recombinase expression plasmids, it is important that the plasmid(s) expressing Cas9 and the guide RNA be easily curable so that the final strain construct does not contain unnecessary plasmids. This is often accomplished with temperature-sensitive origins of replication or by adding guide RNAs targeting the Cas9-expressing plasmids themselves.

It is well worth searching the literature (and the plasmid repository [Addgene](#)) to see if anyone has developed a CRISPR-based mutagenesis system for your organism of interest, although since it is relatively new, many species do not yet have such a system. Depending on what you need to do, in that case it might be a good idea to make one yourself.

SINGLE-COPY INSERTION ELEMENTS

 (mostly) The last type of site-directed mutagenesis I want to mention briefly is somewhat old-fashioned and not much used any more, but you may encounter examples of it in older papers or strains or have specific applications where it might be useful.

Some transposons and most lysogenic bacteriophage do not insert into the bacterial genome randomly, but always insert at the same *attachment site*, which is usually between genes or within a conserved non-essential gene. Cloning genes into such insertion elements can be useful when you want to integrate a single copy of a gene or operon into a strain in a very stable way. Plasmids have higher and sometimes variable copy numbers and are less stable than a chromosomal insertion. Single-copy insertions can be a very clean method to use for complementation experiments.

For phage-based single-copy insertion systems, it is important that the inserted sequence not include the genes required for production of live phage particles, since cultures with active viral infections often behave very differently from uninfected cells. The resulting irreversible insertion is called a *defective prophage* or *stable lysogen*. The most common stable lysogen encountered in molecular biology is the defective λ phage DE3, which carries a *lac* promoter driving expression of the powerful RNA polymerase from phage T7. This is found in the protein overexpression *E. coli* strain BL21 DE3, for example, which is used in many protein purification procedures.

The transposon Tn7, which unlike other transposons we've discussed, does **not** insert randomly throughout a genome, is a widely-used system for making single-copy insertions, and is much more relevant in the modern literature than defective prophage. It can be relatively easily engineered to carry a sequence of interest and integrates into a wide variety of bacteria at the end of the highly conserved *glmS* gene, so is less species-specific than phage integrants. Variants of Tn7 containing tightly regulated promoters and lacking transposase ("mini-Tn7") are excellent platforms for complementation experiments.

For most purposes, at least in more tractable model organisms, it is now usually simpler to use recombineering to construct single-copy chromosomal gene insertions. Recombineering also has the advantage that genes can be inserted into the chromosome anywhere you desire, instead of only at the specific attachment site for a given transposon or prophage.

EXAMPLE OF AN EXPERIMENTAL DESIGN: SITE-DIRECTED MUTAGENESIS

[Seef et al. \(2021\)](#) have recently described a mechanism by which the predatory bacterium *Myxococcus xanthus* (which we will discuss in more detail in [Lectures 11](#) and [15](#)) is able to kill its prey (other bacteria) upon contact. They named the genes involved in this contact-dependent killing process *kilA* through *kilM*, and noticed that homologs of a subset of these genes are also present in other predatory *delta*-proteobacteria. Of particular interest is the *kil* operon in *Bdellovibrio bacteriovorus*, a well-known predatory bacterium that can attack and replicate in the periplasm of other Gram-negative bacteria and which has been explored as a treatment for infections with antibiotic-resistant Gram-negative pathogens.



A scanning helium-ion microscope image of small, curved *B. bacteriovorus* cells attacking a larger rod-shaped *E. coli* cell, from [Schmidt et al. \(2021\)](#).

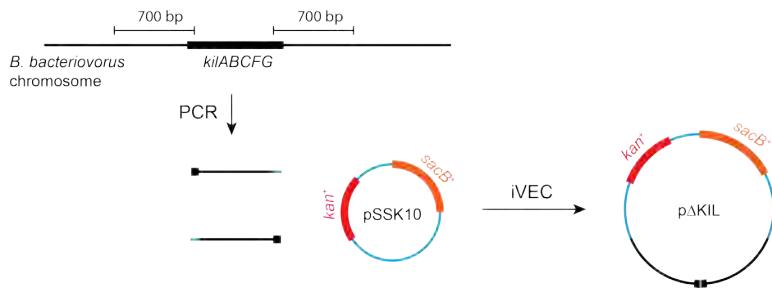
Seef et al. propose, based on this observation, that the *kil* genes may encode important prey killing machinery in predatory bacteria more generally. How can we test if that's true?

Hypothesis: The *kil* genes of *B. bacteriovorus* are important for its ability to kill prey cells.

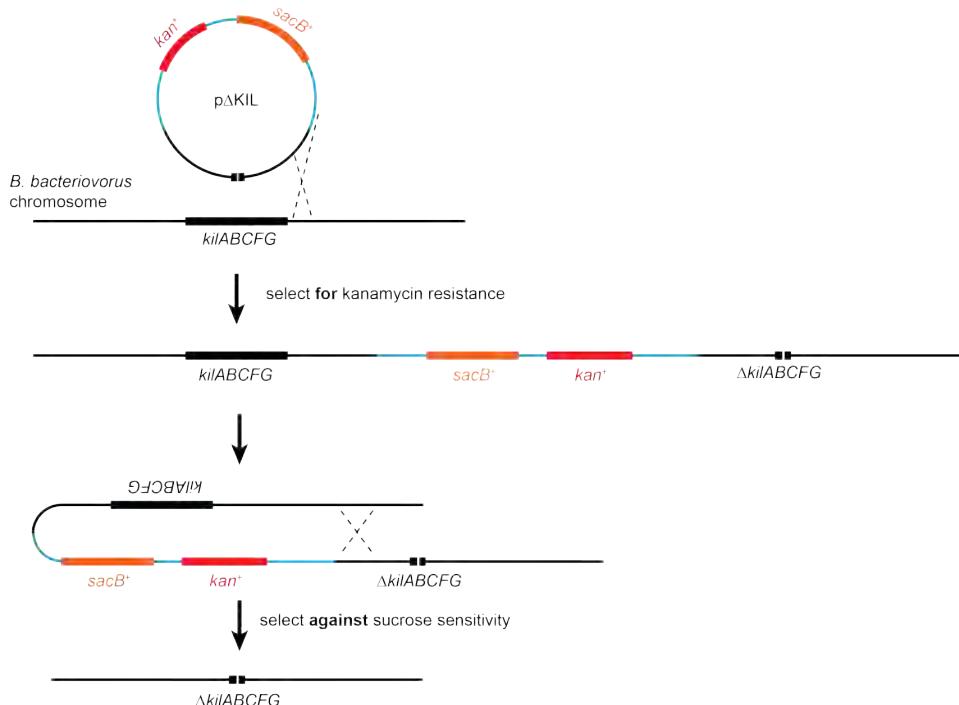
Experimental Design:

- 1) Construct a $\Delta kilABC$ mutant of *B. bacteriovorus*.

Based on a literature search of available methods for this species, we will use allelic exchange to do this, first using Gibson Assembly to clone 700 bp upstream of *kilA* and 700 bp downstream of *kilG* into the suicide vector [pSSK10](#), generating the plasmid $\Delta kilABC$, containing a $\Delta kilABC$ allele.

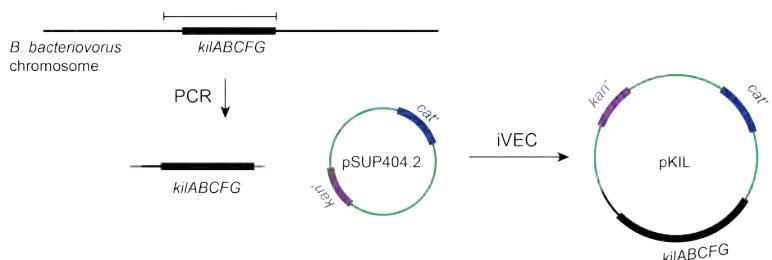


We will then conjugate pΔKIL from *E. coli* into *B. bacteriovorus*, selecting for kanamycin resistance transconjugants. Once those are obtained, we will passage the resulting strain in media containing sucrose to select against the integrated plasmid and screen for the loss of the chromosomal *kilABC* genes by PCR, followed by genome sequencing to confirm that no off-target mutations were made.



2) Construct a *kilABC* expression vector (pKIL).

Best-practice experimental design requires complementation, so we will also use Gibson Assembly to construct a shuttle vector that will allow us to express the *kilABC* genes in *B. bacteriovorus*. There aren't a ton of options in the literature (this might be a candidate situation for building a new shuttle vector), but it looks like the pSUP404.2 plasmid described [here](#) should work, although it is pretty big (11 kb), which isn't ideal. PCR amplification of the vector backbone might require some optimization.



Note that we amplify some sequence upstream of the *kiA* gene, so as to include the native promoter sequence for the *kilABC* operon in pKIL.

3) Test the ability of *B. bacteriovorus* *ΔkilABC* / pSUP404.2 and *B. bacteriovorus* *ΔkilABC* / pKIL to kill *E. coli*.

Independent Variable: (what will you change?)

- The presence or absence of the *kilABC_G* genes in *B. bacteriovorus*.

Dependent Variable(s): (what will you measure?)

- CFU of *E. coli* before and after exposure to *B. bacteriovorus* (a quantitative measurement).

Negative Control(s): (eliminate false positive results)

- 1) The vector-only control will be important to confirm that pSUP404.2 has no effect on killing on its own.
- 2) *E. coli* with no *B. bacteriovorus* added should not die.

Positive Control(s): (eliminate false negative results)

- 1) It will be important to include the wild-type *B. bacteriovorus* strain to tell how much *E. coli* killing to expect in our experiment.
- 2) qRT-PCR to confirm expression of the *kil* genes in pKIL will be important to interpret the results of the complementation experiment.

Potential Outcomes:

- 1) The following result would support the hypothesis and show a causal relationship between the presence of the *kilABC_G* genes and prey killing:
 - wild-type *B. bacteriovorus* / pSUP404.2 kills *E. coli*
 - *B. bacteriovorus* Δ*kilABC_G* / pSUP404.2 does not kill *E. coli*
 - *B. bacteriovorus* Δ*kilABC_G* / pKIL kills *E. coli*
- 2) If *B. bacteriovorus* Δ*kilABC_G* still kills *E. coli*, this will mean at least that the *kil* system is not **essential** for prey killing, although there could be several possibilities within this conclusion. Comparing the rates of killing might be more informative here. Does Δ*kilABC_G* kill fewer *E. coli* or does it kill them more slowly? Is it important for killing other, non-*E. coli* prey species? It is also possible, of course, that these genes have an entirely different function in *B. bacteriovorus* than they do in *M. xanthus*.
- 3) If the mutant has a killing defect, but complementation with pKIL does not restore that defect, it will be important to examine the expression of the genes flanking the *kilABC_G* locus. It's possible that expression of one or more of those genes has been disrupted and **that** is responsible for the observed phenotype.

Does this experiment establish correlation or causation?:

- Because we are fulfilling Falkow's postulates, this design tests **causality**.

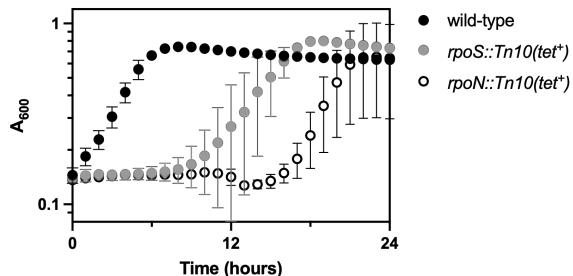
DISCUSSION PROBLEM SET #17: STRAIN CONSTRUCTION WITH RECOMBINATION

Problem #1

For many model organisms, *knockout collections* have been generated that consist of thousands of individual mutants, generally one in each non-essential gene for that organism. These may be generated by isolating individual transposon insertion mutants (like the [Nebraska Transposon Mutant Library](#) for *Staphylococcus aureus* strain USA300_FPR3757) or by recombining (like the [Keio collection](#) for *E. coli* strain BW25113), but regardless of how they are constructed, are a tremendously useful resource.

(See [this paper](#) for a clever method called Knockout Sudoku that can be used to relatively rapidly construct a knockout library with transposons.)

While studying responses to starvation stress in *E. coli*, you find that Tn10(tet^r) transposon insertions in either *rpoS* or *rpoN* reduce the ability of *E. coli* to grow in minimal media with a poorly utilizable nitrogen source.



The *rpoS* and *rpoN* genes encode alternative sigma factors (Lecture 4), and you hypothesize that they each drive the expression of different sets of genes needed under these growth conditions. As part of a series of experiments to test this hypothesis, you decide to construct a double mutant lacking both *rpoS* and *rpoN*.

Design a detailed protocol to build an *E. coli* *rpoS rpoN* mutant. You have access to the Keio collection (in which each mutation is a replacement of the gene in question with a kanamycin resistance gene – i.e. Δ gene::kan⁺) and all standard genetic tools.

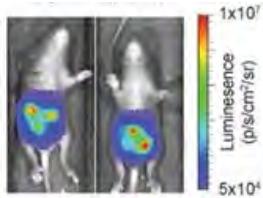
Problem #2

Streptococcus agalactiae, also known as group B *Streptococcus*, is a Gram-positive bacterium commonly found in the gastrointestinal and vaginal microbiota, but capable of causing a variety of dangerous infections (most notoriously, sepsis in newborns, but also meningitis, endocarditis, and osteomyelitis in elderly or immunocompromised adults).



It can be challenging to study the course of an invasive bacterial infection in an animal model in a dynamic way, since counting the number of bacterial cells in an animal or organ usually requires killing the animal. One way around this is to engineer the bacterial pathogen to express luciferase, a generic term for an enzyme that produces light. With sensitive enough light detectors, you can then visualize the spread of the bacteria in a living animal.

Here's an example of what that kind of experiment looks like, visualizing live neonatal mice infected with luciferase-expressing *Klebsiella pneumoniae* (from [this paper](#)):



There are several types of luciferase. Firefly luciferase is one of the most commonly used, but it is rather unstable and requires addition of a small molecule cosubstrate for activity (luciferin), as does Renilla luciferase (RLuc), from the coral-like sea pansy *Renilla reniformis* (which requires coelenterazine), and the Nanoluc luciferase from the deep-sea shrimp *Ophophorus gracilirostris* (which requires furimazine). This is inconvenient for use in whole animals, since you have to somehow deliver those cosubstrates to the location of the bacteria expressing the luciferase.

One common way around this (and the way that *K. pneumoniae* strain works) is to engineer bacteria to constitutively express the *Photobacterium luminescens luxCDABE* operon, which encodes the LuxAB luciferase along with the LuxCDE enzymes needed to synthesize its cosubstrate (a long chain fatty aldehyde derived from a lipid molecule).

Design a detailed protocol for constructing a strain of *S. agalactiae* that constitutively expresses *luxCDABE* from the chromosome for use in mouse infection experiments.

(Expression from the chromosome rather than a plasmid allows you to carry out the infection experiments without having to give the animals antibiotics, which would have an obvious impact on the natural microbiota of the mice. In the absence of selection, as we have discussed, plasmids tend to be lost by bacteria, which would seriously impact our ability to track an infection *in vivo* this way.)

SCIENTIFIC PROCESS 6: SCIENTIFIC REASONING

Logical reasoning is key to the scientific process, but scientists rarely have training in formal logic. What I will do here is to try to give a concise overview of some of the most important principles. This is especially important when you are interpreting the results of experiments, and trying to determine what you can **actually** conclude from a given set of results. Beware of *confirmation bias*, the tendency all humans have to interpret new information as supporting their own prior ideas about how things work.

There are three types of reasoning, all of which have a place in science: *deduction*, *induction*, and *abduction*.

Deduction starts from known principles (true statements) and identifies the logical conclusions of those statements.

general rule → specific conclusion

Particularly popular with philosophers and mathematicians, deduction has the advantage that, if your premises are true, your conclusions must also be true. The obvious disadvantage is that, at least in biology, it is extremely difficult to be certain that your premises are **always** true, and if that general rule doesn't hold, then your deduced conclusion will be wrong every time.

Induction starts with observations and generalizes them to make predictions about broader events and general rules. The likelihood that the conclusions are true depends entirely on the quality and comprehensiveness of the observations.

specific observation → general rule

This is a much more practical form of reasoning for most scientists, and is the basis of the cycle of hypothesis testing and model building that was introduced in previous Scientific Process sections. What **is** worth noting is that making deductions based on induced general rules is, in fact, how new hypotheses are generated. If a particular model was true, what would be the consequences of that truth? Those deductions can then be tested empirically.

Less commonly considered, I think because philosophers and logicians find it unsatisfying, is **abduction**, in which you draw the best possible conclusion from incomplete observations.

incomplete observations → best prediction

This is closely related to induction, but is, in my opinion, a more realistic description of the scientific process as I define it. Observations of real systems are **always** incomplete, and models always have some element of uncertainty. However, it is extremely useful when designing experiments to think about what new data you could collect to fill in the logical holes in your model.

For an alternate description of these different types of logical reasoning, [this link](#) from Mirriam-Webster may be useful.

The final logical principle I want to introduce is that of **parsimony**.

Parsimony, also called "Ockham's Razor", after 14th century English theologian William of Ockham, is a principle that emphasizes preferring the simplest **possible** explanation for a phenomenon. Note that this does not mean that everything **is** simple. It is a way of choosing between models that explain the available data equally well. If a simple model cannot explain all of the data, then it is not valid, and more complexity is necessary to make that model useful and more reflective of the real situation.

From a practical standpoint, the idea of parsimony is to prevent you from introducing unnecessary complexity to your thought process. Start with the most straightforward model with the fewest possible steps, and only add additional elements as the data demand it.

"As simple as possible, but no simpler." – Albert Einstein (c. 1950)

LECTURE 9: CRITICAL READING (GENETIC ENGINEERING)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
 - What is the central **question** of this paper?
 - Is the experimental design clear and appropriate to address that question?
 - Do you understand the methods used?
 - Are the data clearly presented, with appropriate statistics?
 - Do you agree with the conclusions the authors came to based on their data?
 - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Nozaki & Niki (2019) "Exonuclease III (XthA) Enforces *In Vivo* DNA Cloning of *Escherichia coli* to Create Cohesive Ends." J Bacteriol 201:e00660-18.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

You may also find the following minireview / methods paper interesting or relevant, although we will not be discussing it in detail in class:

Watson & García-Nafria (2019) "*In vivo* DNA assembly using common laboratory bacteria: A re-emerging tool to simplify molecular cloning." J Biol Chem 294(42):15271-15281.

LECTURE 10: BACTERIAL CELL ENVELOPES

INTRODUCTION

The second half of this course will focus on bacterial *physiology*: the structure, metabolism, energetics, and development of bacteria as living organisms. In each of the following chapters, we will review the fundamentals of one aspect of bacterial physiology, and design molecular genetics experiments to answer questions about those topics. You will have plenty of opportunities to practice using the experimental design principles we have discussed (summarized on page 241), the methods for measuring gene and protein expression and activity we explored in **Lecture 4**, and the genetic engineering techniques described in **Lectures 7** and **8**, as well as learning more about some more specialized techniques and applications.

My goals for the rest of this class are really two-fold:

1. I want you to be familiar with the basic principles of bacterial physiology, so that you will be prepared to take on more advanced treatment of those topics in future classes (for example, the upcoming Bacterial Pathogenesis module directed by Drs. Scoffield and Swords) and in your own research career.
2. I want you to get lots of practice thinking about and developing ideas for experiments on a variety of different aspects of microbial biology. The problem sets are going to become more complicated and (I hope) realistic from here on forward.

I will only really be able to include the basics for each physiological topic in each upcoming chapter, any of which could easily be expanded into an entire course on their own, and there are certainly whole important topics I will miss entirely due to time constraints. If you're interested in a comprehensive and detailed reference on the fundamentals of bacterial physiology, I've found the following book very helpful, although it is now about 10 years old, so there are probably good newer options out there:

[The Physiology and Biochemistry of Prokaryotes](#) by David White et al. (4th edition)

I will also note once again that the examples and topics I have chosen are inevitably biased towards those that I personally think are the most broadly important or interesting (or happen to have run across recently!). This is an idiosyncratic selection process, and it's important to remember that other scientists would certainly make different choices. My approach is not the only right way, and my real goal is to give you the tools to build your own scientific literacy and proficiency. In other classes, you will encounter teachers with different approaches and emphases, and that's great! Learn as much as you can from as many people as you can.

In this chapter in particular, we will discuss the major components of bacterial *cell envelopes*, including *lipid bilayer membranes* and *cell walls*, that separate the interior of bacterial cells from their external environments. Different groups of bacteria organize their cell envelopes differently, and we will discuss the structures of four major groups of bacteria in some detail.

For a **much** more in-depth treatment of microbial cell structure, the [Atlas of Bacterial and Archaeal Cell Structure](#) by Catherine Oikonomou and Grant Jensen is a phenomenal resource.

BACTERIAL MEMBRANES

All cellular organisms have a lipid membrane that encloses their *cytoplasm* and defines the inside and outside of the cell. The lipids that make up these membranes vary considerably in structure, but have key features that allow them to assemble into bilayer membranes. First, they are *amphipathic*, meaning they have a polar or charged hydrophilic group as well as hydrophobic fatty acid chains. This means that, in aqueous media, the polar headgroups form outer layers around an inner hydrophobic layer (Figure 10.1).

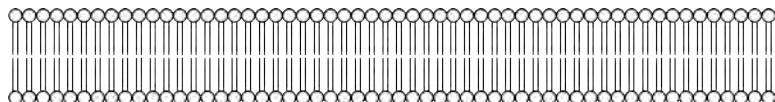


Figure 10.1. A cartoon of a lipid bilayer membrane, made up of amphipathic lipids with polar headgroups (circles) and hydrophobic tails (lines).

Some examples of *E. coli* membrane lipids are shown in Figure 10.2. The *phospholipids* phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) are the most abundant lipids in the *E. coli* inner membrane (roughly 70% and 20%, respectively), with cardiolipin (CL) making up an additional 5-10% of the total. Note that since the headgroups of these lipids contain phosphate, the aqueous surfaces of typical bacterial bilayer membranes are negatively charged at roughly neutral pH.

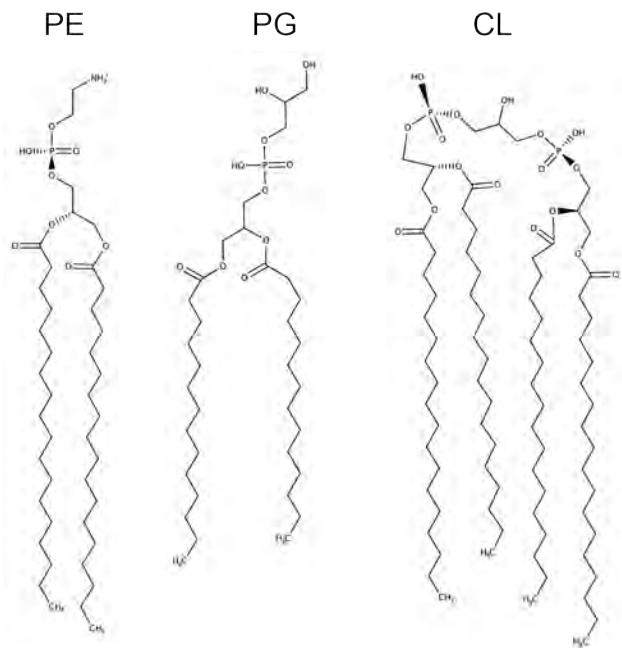


Figure 10.2. The chemical structures of phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and cardiolipin (CL), common lipids in the membrane of *E. coli*. The fatty acids shown are all unsaturated, but bacterial membrane lipids often contain a single double bond or, occasionally, a cyclopropane group. Bacterial lipids do not typically contain more than one double bond.

As in eukaryotes, most bacterial membrane lipids consist of fatty acids linked by ester bonds to glycerol. The length of the fatty acid chains can vary, as can their *saturation* (the number and position of double bonds in the fatty acid chain). Different species have different repertoires of membrane lipids they can synthesize, and the membrane composition of individual cells is regulated in response to environmental conditions.

A key property for membrane function is *fluidity*. The cell membrane must be fluid enough to allow free movement of lipids and proteins within the membrane in order for the cell to function. Think of how cooking fats can be solid (e.g. butter, shortening) or liquid (e.g. vegetable oils), and how easily changes in temperature can soften or harden those fats. The same holds true for membrane lipids. The fluidity or melting temperature of a particular fat depends on the mixture of fatty acids in it: double bonds tend to decrease the melting temperature (butter has more saturated fatty acids, and therefore fewer double bonds, and vegetable oil has more unsaturated ones). Bacteria adjust the mixture of fatty acids they produce to keep their membranes at the correct level of fluidity. For much more detail on this topic, see [this 2023 review article](#).

Bacterial membranes have lots of proteins associated with them, including proteins involved in transporting hydrophilic compounds and proteins from one side of the membrane to the other (see **Lectures 13, 15, and 16**), adhesion to surfaces and to other cells (**Lecture 14**), generating energy (**Lecture 16**), and a wide variety of other functions.

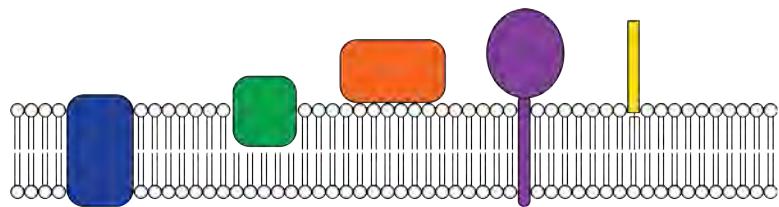


Figure 10.3. Cartoon showing the different ways in which proteins can be associated with membranes. In order, from left to right: two kinds of integral membrane proteins, a peripheral membrane protein, a protein anchored by a hydrophobic α -helix, and a lipoprotein.

As illustrated in Figure 10.3, proteins can associate with membranes in a variety of ways. *Integral membrane proteins* have hydrophobic domains that insert into or span the membranes completely, and these may be as little as a single α -helix or as much as a complete hydrophobic β -barrel. *Peripheral membrane proteins* are more loosely attached to membranes, and may interact largely via charge-charge interactions with the polar headgroups of membrane lipids. *Lipoproteins* have covalently attached lipid groups that insert into the membrane and anchor the protein there. Membrane-associated proteins may also, of course, associate with each other to form complexes or arrays in or on the membrane.

Figure 10.3, like essentially every cartoon depiction of a biological membrane, **drastically** under-represents the amount of protein embedded in biological membranes. Atomic force microscopy images of the surface of *E. coli* (an example drawn from [this review article](#) is shown below) reveal a much different actual situation, at least for the outer membrane of Gram-negative species:

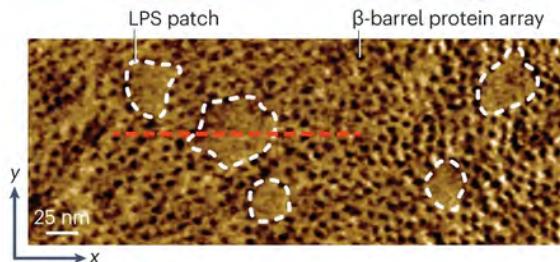


Image from Lithgow et al. 2023 Nat Rev Microbiol 21(8):502-518.

Most of the surface area of a Gram-negative bacterium is covered with dense arrays of integral membrane β -barrel proteins, with only small patches of lipids (i.e. LPS). Keep this in mind when we discuss the components of Gram-negative bacterial membranes below, and be aware that all biological "lipid bilayer" membranes are probably actually mostly protein.

CELL WALLS

Most bacteria have rigid cell walls that define their cell shape. These are composed of crosslinked polymers of sugars and amino acids called *peptidoglycan* or *murein*. The sugar polymers are long chains of alternating *N*-acetylmuramic acid (MurNAc) and *N*-acetylglucosamine (GlcNAc) amino sugars. In *E. coli*, as shown in Figure 10.4, each MurNAc sugar has a pentapeptide of *L*-alanine, *D*-glutamate, meso-diaminopimelate (*mDAP*, a lysine derivative), and two *D*-alanines attached. These are cross-linked to other pentapeptides by the formation of peptide bonds between *mDAP* and *D*-alanine to form the mesh-like cell wall. The nature and abundance of the cross-linked peptide component of the cell wall, as well as the extent of cross-linking, varies considerably among different bacteria. In *S. aureus*, for example, MurNAc has an *L*-alanine, *D*-glutamine, *L*-lysine, *D*-alanine tetrapeptide, and the peptidoglycan includes a 5-glycine peptide bridge between linked lysine and *D*-alanine residues. Note the presence of abundant *D*-amino acids in peptidoglycan, in contrast to the *L*-amino acids found in proteins.

Different bacteria also contain different amounts of peptidoglycan. This can range from a 7-nm thick single layer making up about 10% of the dry weight of Gram-negative bacteria to a 30- to 100-nm thick shell making up 20-25% of the dry weight of many Gram-positive bacteria. In either case, the peptidoglycan mesh is generally permeable to molecules less than 2 nm in diameter, or proteins of less than 25-50 kDa.

The entire crosslinked shell of peptidoglycan is called a *sacculus*, and can be purified intact by careful biochemical procedures. Inversely, bacteria from which the cell wall has been removed by enzymatic treatment are called *sphaeroplasts*. As you might expect, such cells lose their shape and become extremely sensitive to lysis by osmotic stress (changes in the concentration of solutes in the environment).

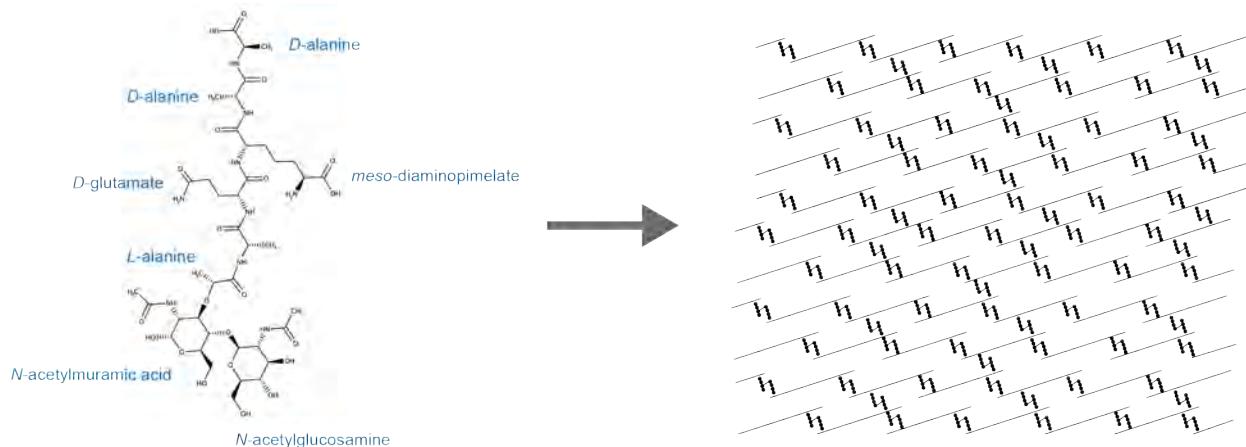


Figure 10.4. The chemical structure of an *E. coli* peptidoglycan monomer, which polymerizes and crosslinks to form the peptidoglycan sacculus of the bacterial cell wall. The amino acids and type of crosslinking varies greatly among different bacterial species.

The monomers of peptidoglycan are synthesized in the cytoplasm. They are exported across the membrane by a process we'll explore in more detail in [Lecture 14](#) and added to growing glycan chains by *glycosyltransferases*. These are then cross-linked by *transpeptidases* (of both the *DD*- and *LD*- variety), which belong to a family of enzymes also called *penicillin-binding proteins* (PBPs). *Endopeptidases*, *amidases*, and *lytic transglycosylases* are all enzymes that break different bonds in peptidoglycan, which is necessary to allow new material to be incorporated into the sacculus as the cell grows and divides.

Many bacteria (including *Clostridium difficile*, *Bacillus anthracis*, *Deinococcus radiodurans*, and *Caulobacter crescentus*) also have an outermost *S*-layer composed of a regular lattice-like shell of identical proteins or glycoproteins that cover the entire sur-

face of the cell (hence the name). The S-layers of the extremely stress-resistant bacterium *D. radiodurans* and the marine archaeon *Nitrosopumilus maritimus*, for example, are both composed of extremely stable hexameric glycosylated immunoglobulin-domain proteins ([links here](#)). S-layers can be important for protection against environmental stresses, attachment, structural stabilization, antigenic variation, or a variety of other functions, and can vary dramatically among even closely related species, with different structures, permeability, charges, and symmetries.

GRAM-NEGATIVE BACTERIA

In 1884, Danish scientist Hans Christian Gram published a method for differentially staining bacteria with the dyes crystal violet and safranin, which results in some kinds of bacteria taking on a dark purple color (*Gram-positive*) and others turning pink (*Gram-negative*) when viewed through a microscope. The *Gram stain* remains a useful tool in diagnostic microbiology. The differences in cell structure that distinguish Gram-negative and Gram-positive bacteria were not discovered until 80 years later, in 1964, when Howard Bladen and Stephen Mergenhagen used thin-section electron microscopy to show that Gram-negative bacteria have two membranes, separated by a narrow *periplasm* (also called the *periplasmic space*). Because of this two-skinned structure, Gram-negative bacteria are also called *diderms* or *diderm bacteria* ("derm" being the Greek root word for "skin").

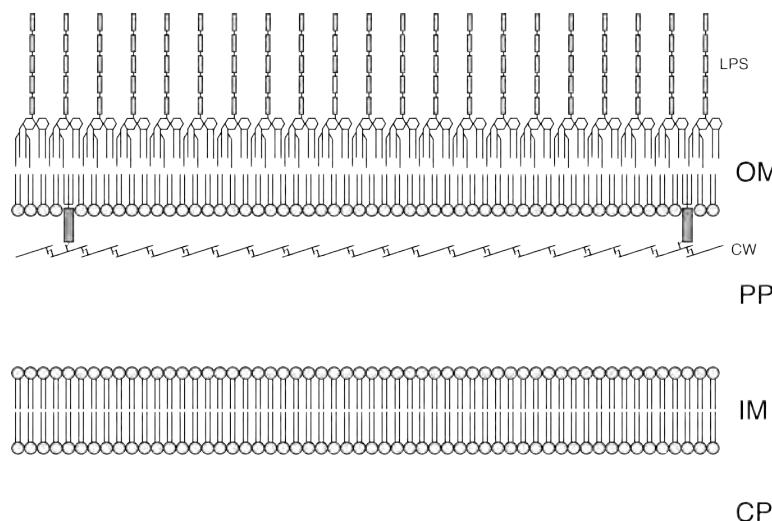


Figure 10.5. A schematic view of the cell envelope of a Gram-negative bacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; PP, periplasm; CW, peptidoglycan cell wall; OM, outer membrane; LPS, lipopolysaccharide. Grey rectangles represent Braun's lipoprotein (Lpp).

The periplasm is about 20 nm across, and is estimated to constitute between 5 and 20% of the total volume of a Gram-negative cell. It contains a distinct set of proteins, including many involved in nutrient acquisition and stress resistance. In *E. coli*, roughly 15% of the proteome is targeted to the periplasm (see [Lecture 13](#)). However, it does **not** contain ATP or other small, physiologically important cytoplasmic molecules like NADH, FADH₂, or glutathione, so periplasmic proteins do not have access to these as cofactors or substrates.

The inner and outer membranes are quite different in both lipid and protein content. Major **outer membrane proteins** (OMPs) include the *porins* OmpC, OmpF, OmpD, etc., which are β -barrel proteins that form channels through the outer membrane and allow diffusion of solutes and small molecules into the periplasm. The outer membrane is therefore much more permeable than the inner membrane. This means that, for example, while the pH of the cytoplasm is maintained at near neutrality, the pH of the periplasm is the same as that of the outside environment. The exact permeability of the outer membrane is regulated by the relative expression of different porins under different environmental conditions.

The most abundant protein in *E. coli* by number (about 500,000 copies per cell) is the **lipoprotein Lpp** (also called Braun's lipoprotein, after its discoverer; Volkmar Braun), which is attached to the inner surface of the outer membrane by a covalently-attached lipid group. (Lpp was the first lipoprotein ever discovered.) About 1 in 3 Lpp proteins are also covalently bound to peptidoglycan (by an LD-transpeptidase), securely attaching the outer membrane to the cell wall. In many Gram-negative bacteria, the cell wall is covalently bound to outer membrane β -barrel proteins instead, which has the same general functional role.

Notably, the outer membrane is asymmetric with respect to its lipid content. The *inner leaflet* (facing the periplasm) is similar in composition to the inner membrane, but the *outer leaflet* is composed almost entirely of **lipopolysaccharide** (LPS), a complex molecule with both lipid and carbohydrate components. LPS is also sometimes called *endotoxin*, because the human immune system can mount an extremely aggressive immune response to LPS, leading to septic shock. (This term was coined by Richard Pfeiffer while working in Robert Koch's lab in the late 1880's, to distinguish it from *exotoxins*, which are secreted away from the bacterial cell; see [Lecture 13](#) for more on secretion.)

The innermost component of LPS, called *Lipid A*, consists of two phosphorylated glucosamine sugars with four to six ester-linked fatty acids as the lipid component that forms the hydrophobic interior of the membrane. Attached to one of the glucosamines is the core region, a branched and phosphorylated polysaccharide that is divided into inner and outer cores. The core region of *E. coli* contains two keto-deoxyoctulosonate, four *L*-glycero-*D*-manno-heptose, three glucose, and one galactose sugar in a characteristic arrangement, but the composition of the LPS core varies widely from one species to another. The rest of LPS is known as the *O-antigen*, and is extremely variable, even within a single species. It is typically a repetitive polysaccharide composed of up to 40 three- to five-sugar repeat units (see **Lecture 14** for more on carbohydrate biochemistry). At least 20 sugars are known to occur in O-antigens in different bacteria, including some that are very rare in other contexts. More than 160 antigenically-distinct O-antigens are known for *E. coli* alone, and are a key tool for distinguishing pathogenic strains in a clinical context (recall enterohemorrhagic *E. coli* O157:H7, where O157 is the serotype of that strain's O-antigen).

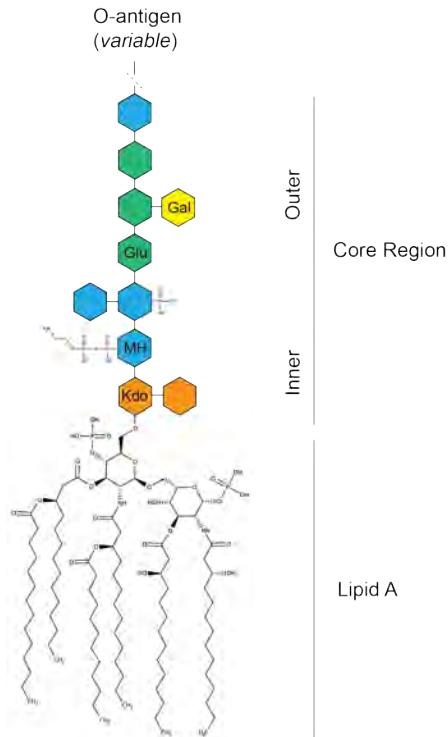


Figure 10.6. The structure of lipopolysaccharide from *E. coli*. Kdo = keto-deoxyoctulosonate, MH = *L*-glycero-*D*-manno-heptose, Glu = glucose, Gal = galactose. The O-antigen is a polymer of hexose and hexosamine repeats and is highly variable from strain to strain.

As part of the outermost surface of the Gram-negative cell, LPS is important for interactions between bacteria and host microbes (both pathogenic and symbiotic), ranging from inhibition of phagocytosis in *Salmonella*, to mimicry of human Lewis blood group antigens by *Helicobacter pylori*, to a role in establishment of symbiotic interactions with leguminous plants by nitrogen-fixing *Rhizobium* bacteria.

Many bacteria important in human, animal, and plant health and the environment have what we now know to be a typical Gram-negative bacterial cell envelope. These include human commensals like *Escherichia coli* and *Bacteroides thetaiotaomicron*, important pathogens like *Salmonella enterica*, *Pseudomonas aeruginosa*, *Vibrio cholerae*, and *Yersinia pestis*, plant pathogens like *Pectobacterium* (formerly *Erwinia*) *carotovora*, and vast numbers of others, including the green photosynthetic cyanobacteria and the predatory bacterium *Bdellovibrio bacteriovorus*. Basically, in any environment where bacteria are found, you will find Gram-negative species.

GRAM-POSITIVE BACTERIA

The Gram-positive bacteria, members of the phyla *Firmicutes* and *Actinobacteria*, have cell envelopes quite different from those of Gram-negative bacteria. They are monoderms and have only a single cell membrane with a (typically) very thick peptidoglycan cell wall outside of it.

In addition to peptidoglycan, Gram-positive cell walls contain large amounts of *teichoic acids*, which are polymers of repeating sugars and phosphate groups that are either anchored in the cell membrane by lipid groups (*lipoteichoic acids* or LTAs) or covalently attached to peptidoglycan (*wall teichoic acids* or WTAs). Teichoic acids can be highly charged, with both positively charged amino groups and negatively charged phosphate groups. There are also uncharged teichoic acids in some species. Teichoic acids in general are important for cell wall rigidity, maintenance of cell shape, attachment of S-layer proteins,

chelating cations like sodium and magnesium, and also play roles analogous to those of LPS in interactions with the host immune system and other kinds of host-microbe interactions.

Lacking an outer membrane to keep extracellular proteins from diffusing away, many secreted proteins in Gram-positive bacteria are covalently bound to either peptidoglycan (by enzymes called *sortases*) or teichoic acids. There is some data that argues that the innermost volume of the cell wall space in Gram-positives, between the outer surface of the cytoplasmic membrane and the bottom of the peptidoglycan matrix, is different enough in composition from the environment to constitute a kind of Gram-positive “periplasm”, but this is somewhat controversial and probably mostly semantic.

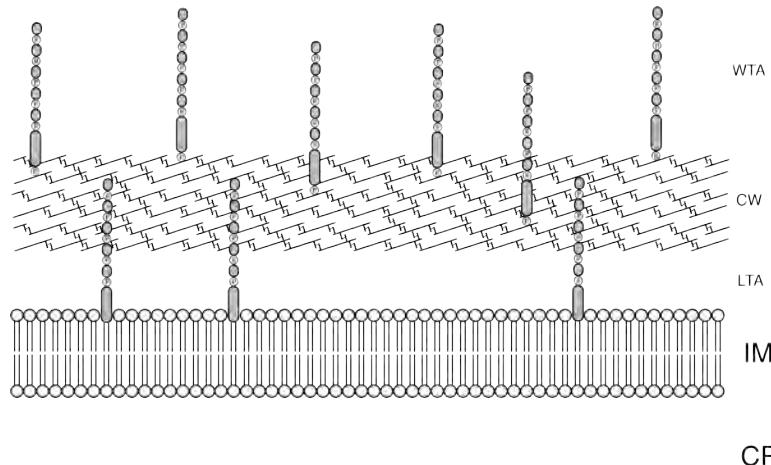


Figure 10.7. A schematic view of the cell envelope of a Gram-positive bacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; LTA, lipoteichoic acids; CW, peptidoglycan cell wall; WTA, wall teichoic acids.

Gram-positive bacteria are also ubiquitous in humans, animals, and the environment. They include human pathogens like *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Clostridium difficile*, and *Bacillus anthracis*, as well as commensals like *Bifidobacterium bifidum* and *Faecalibacterium prausnitzii*. The Gram-positive lactic acid bacteria (commonly found on plants, especially fruit) are important in food fermentation and preservation (e.g. *Lactococcus lactis*, *Streptococcus thermophilus*, *Staphylococcus carnosus*) as well as as health-promoting probiotics (e.g. *Lactobacillus acidophilus* or *Lactobacillus rhamnosus*). In the environment, the filamentous Gram-positive *Streptomyces* species and their relatives (the actinobacteria) are very abundant and are notable for producing many complex chemical compounds, including most of the known kinds of natural antibiotics (see [Lecture 18](#)) and petrichor, the chemical responsible for the characteristic aroma of damp soil. Many Gram-positive bacteria are able to differentiate into extremely stress-tolerant spores (see [Lecture 11](#)). Before the advent of more sophisticated taxonomic methods, aerobic spore-forming rods were given the genus name *Bacillus* while anaerobic spore-forming rods were called *Clostridium*. Neither of those phenotypically-defined genera turned out to be *monophyletic*, so each has now been divided into many more phylogenetically meaningful genera.

I will note here that not every microbe that stains Gram-positive with crystal violet is a monoderm bacterium (the diderm *Deinococcus radiodurans* stains Gram-positive, as do eukaryotic yeast cells), and some monoderm bacteria do not stain strongly Gram-positive (late stationary phase cultures of *Clostridium* species are notorious for appearing Gram-negative). Modern phylogenetic methods do not rely on dye staining, and the current *phylogenetic tree* of the bacteria, a rough version of which is illustrated below, is based on the sequences of highly conserved genes (most commonly 16S ribosomal RNA, although more comprehensive methods incorporate other genes as well). See [this brief article](#) for a summary of the state of the field as of 2024.

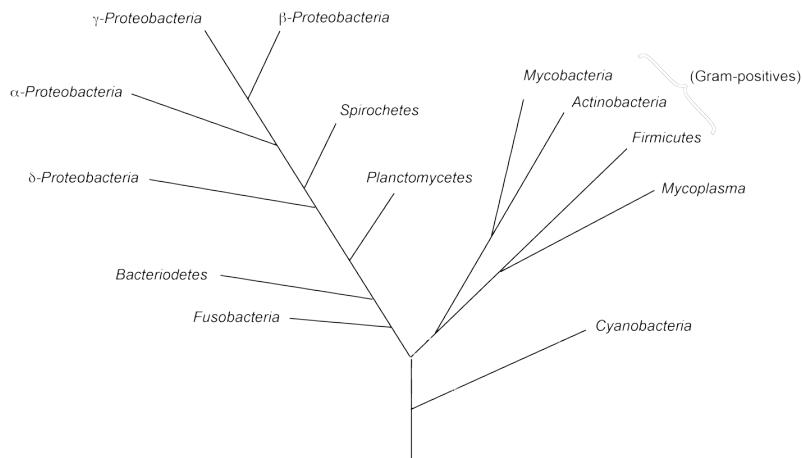
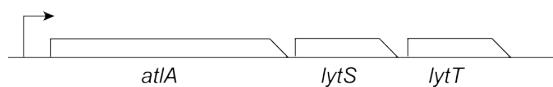


Figure 10.8. The approximate phylogenetic relationships between major groups of bacteria, as determined by the sequence of their 16S rRNA and other highly conserved genes. See this [2024 special issue](#) of the Journal of Molecular Evolution for recent developments in ancient phylogeny.

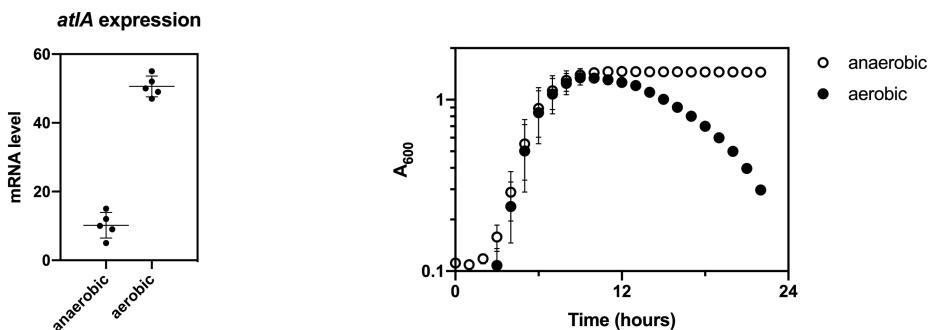
DISCUSSION PROBLEM SET #18: REGULATION OF AUTOLYSIS IN *STREPTOCOCCUS GORDONII*

Streptococcus gordonii is a Gram-positive lactic acid bacterium that is an early colonizer of the oral cavity and member of the healthy oral microbiome. It is able to antagonize the growth of the oral pathogen *S. mutans*.

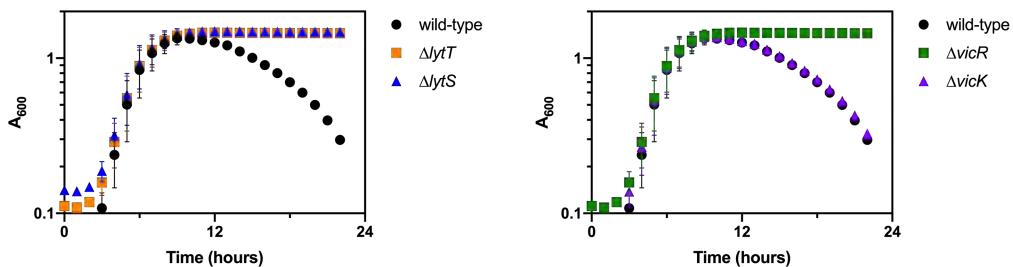
The genome of *S. gordonii*, like that of many streptococci, encodes an *autolysin*, an enzyme with peptidoglycan-degrading N-acetyl-muramidase activity that can degrade the cell wall and cause cell lysis. The *S. gordonii* autolysin is encoded by the *atlA* gene, which is in an operon with genes encoding a two-component system histidine kinase (*lytS*) and response regulator (*lytT*) (recall TCS regulators from [Lecture 4](#)). The signal that *LytS* responds to is unknown.



Expression of *atlA* increases in the presence of oxygen, and aerobic cultures lyse more rapidly than anaerobic ones:



There is a second known TCS in *S. gordonii* that responds to oxygen, composed of the *VicK* histidine kinase and *VicR* response regulator. Mutants lacking each of the four TCS regulators have the following **aerobic** growth phenotypes:



Propose a model to explain the regulation of *AtlA* transcription by the *LytST* and *VicRK* two-component systems in response to oxygen. Design an experiment to test this model. State:

- your hypothesis, and how your experiment will test that hypothesis
- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiments, and how you will interpret them

For discussion in class: What advantage(s) might bacteria gain from expressing a protein whose sole purpose seems to be to break down the cell's own cell wall? (Or from bacterial programmed cell death in general.)

MYCOBACTERIA

Mycobacteria are members of the genus *Mycobacterium*, and include both free-living environmental species and obligate pathogens. The most important of these in human medicine are *Mycobacterium tuberculosis*, the cause of tuberculosis, and *Mycobacterium leprae*, the causative agent of leprosy. Phylogenetically, mycobacteria are members of the Gram-positive actinobacteria (Figure 10.8), but they have a unique cell wall structure that sets them apart from other members of that clade, and in fact, they cannot be stained by the Gram stain procedure at all. They do not take up crystal violet very well, and have long been known as acid-fast bacilli, due to their resistance to destaining by acids, both of which properties are due to their unusual cell envelope structure.

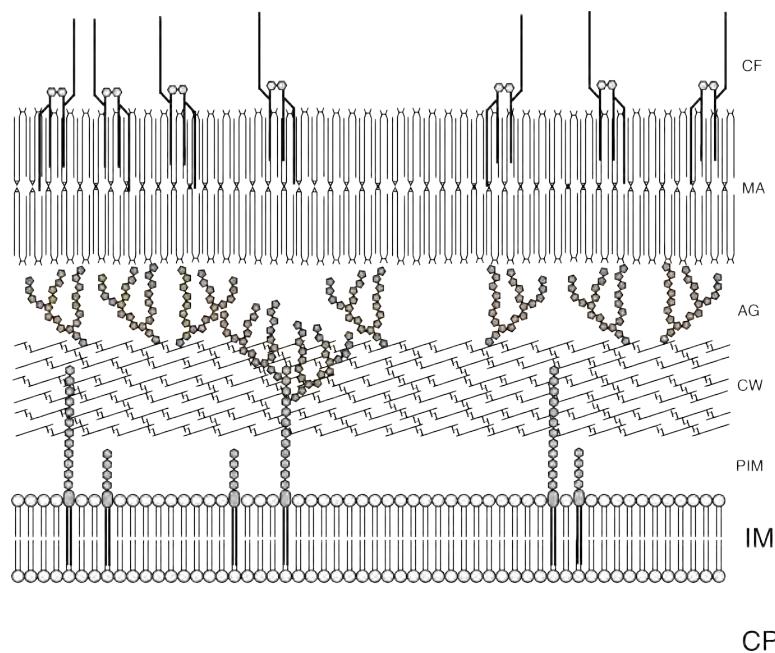


Figure 10.9. A simplified schematic view of the cell envelope of a mycobacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; PIM, phosphatidylinositol mannosides; CW, peptidoglycan cell wall; AG, arabinogalactan; MA, mycolic acids; CF, cord factor (trehalose dimycolate).

As in other bacteria, the inner membrane of mycobacteria is composed mostly of phospholipids, including phosphatidylethanolamine, phosphatidylinositol, and cardiolipin, but with a very high abundance of phosphatidylinositol mannosides (PIMS) that are not found in other groups of bacteria. These glycolipids span the space between the inner membrane and the peptidoglycan cell wall, and both the PIMs and the peptidoglycan are decorated with highly branched polymers of arabinogalactan that form an additional layer outside of the cell wall. Attached to the arabinogalactan layer is another lipid bilayer that is chemically very different from that of the inner membrane or of the outer membrane of Gram-negative bacteria. It is composed of *mycolic acids*, which are long-chain (70 – 90 carbon atoms), extremely hydrophobic lipids unique to the mycobacteria.

Mycolic Acids

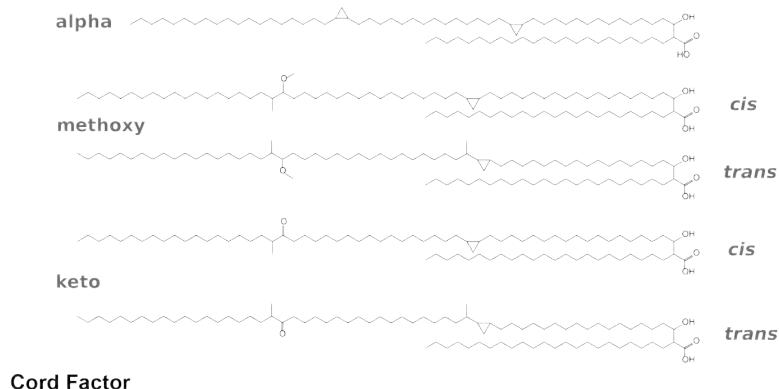


Figure 10.10. The chemical structures of mycolic acid-containing lipids from *Mycobacterium tuberculosis*. (Images from Wikimedia Commons.)

Mycolic acids make up at least 50% of the dry weight of a mycobacterial cell, and their waxy, hydrophobic properties are largely responsible for the unique nature of the mycobacterial cell envelope. Cord factor (trehalose-6,6'-dimycolate) is an abundant mycolic acid-containing glycolipid associated with virulence and resistance to antimicrobial compounds in *M. tuberculosis*. There are a complex variety of other lipids in the outer membrane of mycobacteria, including other glycolipids, phthiocerol dimycocerosate, and sulfolipids, all of which contribute to the stability and overall impermeability of mycobacterial cells.

Compared to other groups of bacteria, mycobacteria are generally very resistant to toxins, stressful environmental conditions, and antibiotics. This is largely due to their cell envelopes, and the fact that very few compounds can penetrate the many layers to reach the interior of the cell. This is particularly important when considering how pathogenic mycobacteria resist attack by the host immune system.

Mycobacteria face an interesting physiological challenge in transporting compounds across their impermeable cell envelopes. In order to survive, they must import nutrients and export waste products, surface-associated proteins, and (as has recently been discovered by the Niederweis lab here at UAB) toxins. The mechanisms by which they accomplish these transport functions are not completely understood and are an area of active research. See, however, the discussion of type VII secretion in **Lecture 13** for one important mechanism.

MYCOPLASMAS

Mycoplasmas are another group of bacteria that are phylogenetically related to the Gram-positive bacteria (Figure 10.8), but which have distinct cell envelope structures. In the case of the mycoplasmas, they lack cell walls entirely. They are also extremely small, both physically (less than half a micron in diameter in many cases) and genetically. The genome of *Mycoplasma genitalium*, a pathogen that causes urethritis, contains fewer than 600 kilobases of DNA, and that of *Mycoplasma pneumoniae*, the causative agent of walking pneumonia, is just over 800. This makes the mycoplasmas the smallest and simplest free-living cellular organisms currently known.

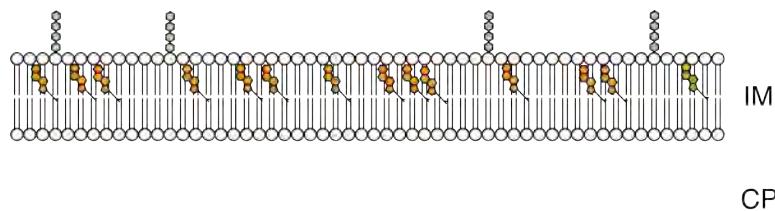


Figure 10.11. A schematic view of the cell envelope of a mycoplasma. Abbreviations: CP, cytoplasm; IM, inner membrane. Cholesterol is indicated in orange, and cell surface glycolipids are indicated in grey. Note that there is **much** more glycosylation of the lipids on the outer leaflet of the membrane than is implied in this image.

Unlike other bacterial membranes, most mycoplasma membranes contain large amounts (25-30% by weight) of the sterol lipid cholesterol, which many species of mycoplasma cannot synthesize, but must obtain from eukaryotic cells (using a [specific extraction protein](#)). The function of cholesterol is to help maintain membrane stability and fluidity. The phospholipids in the outer leaflet of the mycobacterial membrane are heavily glycosylated, which is also important for maintaining membrane integrity. Since this is the outer surface of the cell, these glycolipids are also highly immunogenic and are a major point of interaction between pathogenic mycoplasmas and the host immune system.

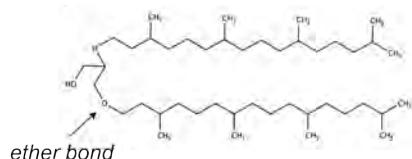
Because they are small and flexible, mycoplasma cells pass easily through filters designed to sterilize growth media. The pores of these filters are usually 0.2 or 0.45 μm in diameter, large enough to block passage of most bacteria, but not mycoplasmas. Mycoplasmas are also immune to antibiotics that target cell wall synthesis, like ampicillin. This is a serious problem in laboratories that use animal or human cell culture systems, which can easily become contaminated by mycoplasma species, and since mycoplasma cells are so small, this can be difficult to detect by normal microscopic examination.

OTHER PROKARYOTIC CELL ENVELOPES

The four examples described above, which include most of the important groups of human bacterial pathogens, do not encompass the entire diversity of bacterial cell envelopes, the evolution of which appears to have been quite complicated. If you're interested in more details, I'd recommend [these review articles](#), which give many examples of bacteria that do not fit neatly into any of the categories described above and argue convincingly that the common ancestor of all bacteria was a diderm, and that modern monoderm bacteria are examples of diderms that have lost their outer membranes. The existence of bacteria like the oral commensal *Veillonella parvula*, a deep-branching member of the *Firmicutes* clade with an outer membrane and periplasmic space, is one example that supports this hypothesis.

The cell envelopes of archaea are dramatically different from those of bacteria or eukaryotes, and are well worth reviewing briefly here, although most of you are unlikely to study archaea intensively during your microbiology careers. Unlike the ester-linked lipids found in bacteria and eukaryotes, archaeal membranes contain **ether**-linked lipids (Figure 10.12). This, along with the occurrence in some species of tetraether lipids that form a lipid **monolayer** (rather than the more typical lipid bilayer), means that archaeal membranes have very distinctive chemical properties, including the potential for much higher heat tolerance than bacterial membranes. This is probably why the most heat-resistant living organisms are all archaea. (The current record-holder, to the best of my knowledge, is the deep-sea volcanic vent archaeon *Methanopyrus kandleri*, which can grow at 122°C.) It is worth remembering that most archaea are not thermophilic, and archaea can be found in almost any environment where bacteria are present (although the reverse is not always true).

archaeol



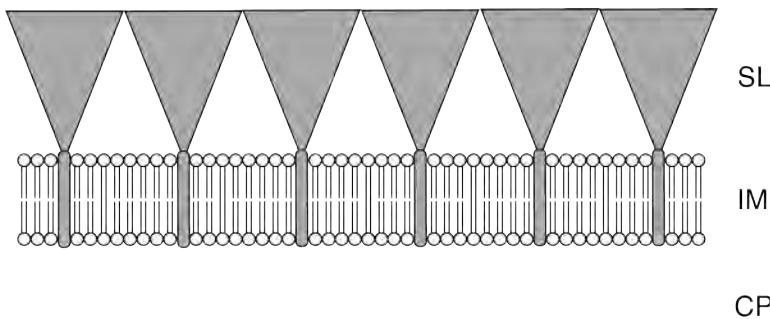


Figure 10.13. A cartoon of a representative archaeal cell envelope, showing the inner membrane composed of both bi- and tetraether lipids and an S-layer (SL), a crystal-like array of glycosylated surface proteins. This arrangement is found in *Sulfolobus* spp., for example, but many archaea have more complicated cell envelopes, which may contain additional protein, carbohydrate, or lipid layers.

Archaea are weird and diverse, and their biology is fascinating, but because there are no known archaeal pathogens, studies of archaea are not funded as intensively as those of bacteria. Some recent hints that methanogenic archaea may play a role in the health and function of the gut and oral microbiomes may help to change this in the future, though.

DISCUSSION PROBLEM SET #19: NOT ALL GENETIC SYSTEMS ARE CREATED EQUAL

Up to this point in the class, we've been assuming that all of the tools of molecular biology are available for any organism you might be interested in working on. This is 100% **not** the case, and you will often be limited in what techniques and tools you have available to you in any particular species.

For future problems, I will list what genetic tools are available for the species in question (to the best of my knowledge), and notable limitations for working with that species in the lab, which may limit the kinds of experimental approaches you can take. This is, however, a great opportunity for you to be creative in thinking about ways around the limitations of a particular study organism. If you want to propose **developing** a tool to help you solve a particular problem, that's great, but I will want to hear exactly how you plan to do that.

Here is an incomplete list of techniques and tools which may or may not be available for a particular species:

Essentially available for every species:

- can extract total DNA, RNA, and proteins
- has a complete genome sequence
- is susceptible to chemical or radiation mutagenesis

Available in many or most commonly-studied bacteria:

- can grow in pure culture and/or outside of host cells
- is or can be made competent
- can take up DNA by conjugation
- efficiently carries out homologous recombination with foreign DNA
- has suitable cloning, shuttle, or suicide vectors available
 - has usable selectable and/or counter-selectable markers
 - has known inducible promoters
- has a compatible transposon mutagenesis system

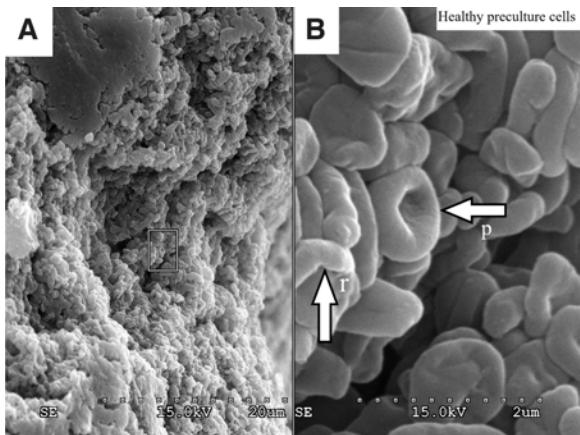
Not available for very many bacteria:

- can carry out oligo-directed recombination
- has a CRISPR-based mutagenesis system (or variations thereof)
- has a genome-scale knockout collection
- has a generalized transducing phage

For discussion in class: which of these do you think is **most important** for being able to do genetic experiments in a bacterial species? In what order would you prioritize developing these tools for a new study organism?

EXAMPLE OF AN EXPERIMENTAL DESIGN: S-LAYER GLYCOSYLATION IN HALOFERAX VOLCANII

Haloferax volcanii is a halophilic archaeon originally isolated from the Dead Sea that has emerged as a tractable model organism. It grows at 40–45°C in media containing between 1 and 4 M salt, and has no cell wall. Instead, it has a proteinaceous S-layer made up mostly of the Csg protein, which is N-glycosylated at several asparagine (Asn) residues (as described [here](#)).



Scanning electron microscope image of *H. volcanii* cells grown in laboratory media. From [this paper](#).

When grown in 3.4 M NaCl (high salt), Csg Asn-13 and Asn-83 are heavily glycosylated. This glycosylation is much less abundant in 1.75 M NaCl (low salt). In contrast, Asn-498 is glycosylated at low salt, but not at high salt. This leads you to wonder whether glycosylation of Csg is important for the ability of *H. volcanii* to adapt to different salt concentrations.

The following methods are available for *H. volcanii* (summarized [here](#)):

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (4 Mbp)	✓
susceptible to mutagens	✓
plasmids can be introduced by transformation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
inducible and constitutive promoters available	✓
compatible transposon	✓
gene knockout library (here)	✓
native CRISPR system for gene knockdown (here)	✓

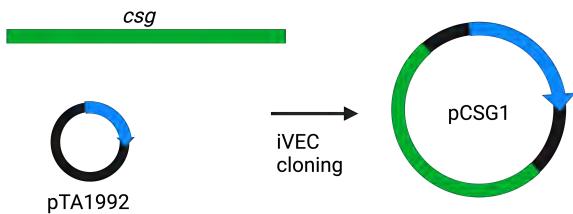
Design an experiment to determine whether glycosylation of Asn residues in Csg affects the ability of *H. volcanii* to grow at different salt concentrations.

Hypothesis: Glycosylation of Asn-13 and Asn-83 is important for growth in high salt, and glycosylation of Asn-498 is important for growth in low salt.

Experimental Design:

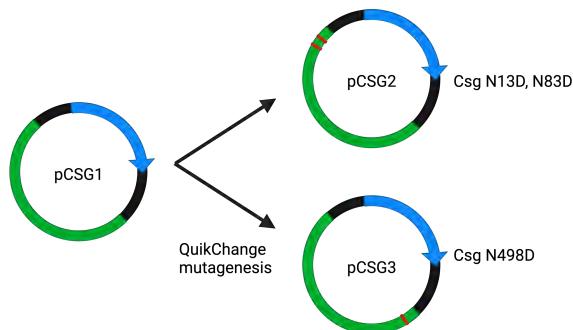
This is a little tricky, since the S-layer protein is essential in *H. volcanii*. We therefore need to jump through a few extra hoops in our experimental design. The basic plan is to first express either wild-type or mutant csg from a plasmid, and **then** knock out the chromosomal copy. This is a common strategy to approach genetic analysis of essential genes.

- I) Use iVEC to clone csg into the shuttle vector [pTA1992](#), under control of a strong constitutive synthetic promoter, generating plasmid pCSG1.



Made with Biorender.

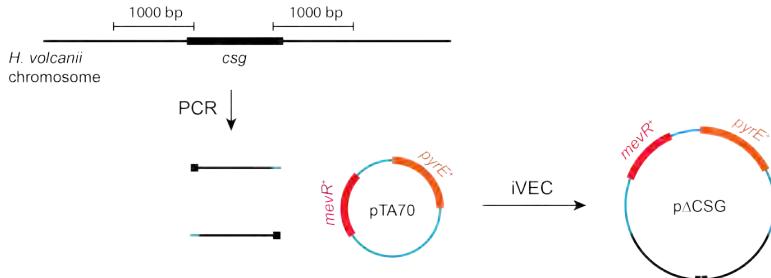
- 2) Use QuikChange mutagenesis to generate variants of pCSG1 encoding Csg^{N13D, N83D} or Csg^{N498D} variants (pCSG2 and pCSG3, respectively). (Aspartate is chemically similar to asparagine, but cannot be N-glycosylated.)



Made with Biorender.

- 3) Transform *H. volcanii* with pCSG1, pCSG2, or pCSG3.

- 4) Use iVEC to construct an allelic exchange vector to knock out *csg* from *H. volcanii* (using the [pTA70 suicide vector](#), which encodes mevinolin resistance and a counterselectable marker based on the *pyrE* gene).



- 5) Delete *csg* by allelic exchange in *H. volcanii* strains containing pCSG1, pCSG2, or pCSG3.

Note that this should be probably done at a moderate NaCl concentration (perhaps 2.5 M), to try to minimize any negative impacts the mutations may have during strain construction. If one or the other Csg mutant-expressing strains does not allow you to knock out the chromosomal copy of *csg*, that will suggest that those Asn residues themselves are essential.

- 6) Assess growth of *H. volcanii* Δ*csg* / pCSG1, pCSG2, and pCSG3 strains at 1.75 M and 3.4 M NaCl.

Independent Variable(s): (what will you change?)

- 1) The allele of Csg expressed (wild-type, Csg^{N13D, N83D}, or Csg^{N498D})
- 2) NaCl concentration (1.75, 2.5, or 3.4 M)

Dependent Variable(s): (what will you measure?)

- Growth rate (a quantitative measurement)

Negative Control(s): (eliminate false positive results)

- A wild-type *H. volcanii* / pTA1992 vector-only control will be important to determine what the normal growth rate at different salt concentrations is.

Positive Control(s): (eliminate false negative results)

- The Δcsg / pCSG1 strain is the key positive control here. It is expected to grow at the same rate as the wild-type strain, and if it doesn't, then that will need to be taken into account in interpreting the Asn mutant results.

Potential Outcomes:

- The following results would be the most intuitive, and would fully support the stated hypothesis:

	1.75 M NaCl	2.5 M NaCl	3.4 M NaCl
wild-type / pTA1992	+	+	+
Δcsg / pCSG1 (wild-type Csg)	+	+	+
Δcsg / pCSG2 (Csg ^{N13D, N83D})	+	+	-
Δcsg / pCSG3 (Csg ^{N498D})	-	+	+

- Since we're changing two variables at once, there are a lot of alternative outcomes. (This is why only changing one variable at a time is better practice, generally.) Assuming our controls look as expected, here are a few possibilities, with their interpretations:

	1.75 M NaCl	2.5 M NaCl	3.4 M NaCl
wild-type / pTA1992	+	+	+
Δcsg / pCSG1 (wild-type Csg)	+	+	+
Δcsg / pCSG2 (Csg ^{N13D, N83D})	+	+	+
Δcsg / pCSG3 (Csg ^{N498D})	+	+	+

- No impact of Asn mutations on growth rate. Would lead us to reject the hypothesis.

	1.75 M NaCl	2.5 M NaCl	3.4 M NaCl
wild-type / pTA1992	+	+	+
Δcsg / pCSG1 (wild-type Csg)	+	+	+
Δcsg / pCSG2 (Csg ^{N13D, N83D})	+	+	-
Δcsg / pCSG3 (Csg ^{N498D})	+	+	+

- N13 and N83 glycosylation is important at high salt, but N498 glycosylation has no impact on growth

	1.75 M NaCl	2.5 M NaCl	3.4 M NaCl
wild-type / pTA1992	+	+	+
Δcsg / pCSG1 (wild-type Csg)	+	+	+
Δcsg / pCSG2 (Csg ^{N13D, N83D})	+	+	+
Δcsg / pCSG3 (Csg ^{N498D})	-	+	+

- N13 / N83 glycosylation has no impact, but N498 glycosylation is important at low salt

Of course, there's always the possibility that we'd see impacts of one or the other mutation at 2.5 M NaCl, or that there would be a gradient of growth rates more complex than is represented by the qualitative + and - used here. It's also possible that phenotypes other than growth rate might be more relevant (cell shape or size, maybe?).

This experiment illustrates how quickly interpretation becomes more complicated as you add additional independent variables, and why it's almost always better to limit yourself to just one independent variable per experiment.

DISCUSSION PROBLEM SET #20:ANTIBIOTIC RESISTANCE IN FAECALIBACTERIUM PRAUSNITZII

Faecalibacterium prausnitzii is a strictly anaerobic, non-sporulating, Gram-positive commensal bacterium that is very abundant in the human gut microbiome (making up as much as 5% of the bacteria in the large intestine). It digests dietary fiber and produces butyrate and other short-chain fatty acids which have significant effects on the metabolism of intestinal epithelial cells. Low levels of *F. prausnitzii* are correlated with some diseases, including Crohn's disease, asthma, and major depressive disorder.

The β -lactam antibiotics kill bacteria by inhibiting the DD-transpeptidases (PBPs, or penicillin-binding proteins) necessary for peptidoglycan synthesis (more on those in [Lecture 14](#)). Surprisingly, some isolates of *F. prausnitzii* are sensitive to the first-generation β -lactam ampicillin but resistant to the more modern β -lactams of the cephalosporin and cefoxitin classes. This suggests that *F. prausnitzii* may be a source of clinically important antibiotic resistance genes that could be transferred to pathogens by horizontal gene transfer. The genome of *F. prausnitzii* encodes only 3 PBPs. (Compare this to the model Gram-positive organism *B. subtilis*, which encodes 11, and is typically sensitive to all β -lactam antibiotics.)

The following limited set of methods are available for *F. prausnitzii*, which is extremely resistant to transformation (for reasons that are not entirely clear, but may have to do with highly active restriction enzymes and/or a native CRISPR system):

growth in pure culture (strictly anaerobic)	✓
can extract DNA/RNA/protein	✓
complete genome sequence (2.9 Mbp)	✓
susceptible to mutagens	✓

Design an experiment to identify gene(s) or gene features required for either ampicillin sensitivity or cephalosporin resistance in *F. prausnitzii*. State:

- your hypothesis, and how your experiment will test that hypothesis
 - the independent and dependent variables
 - both positive and negative controls
 - a description of how you will construct any necessary strains
 - whether your proposed experiment(s) test correlation or causation
 - potential outcomes of your experiments, and how you will interpret them
-

LECTURE 11: BACTERIAL CYTOSKELETON AND DEVELOPMENT

INTRODUCTION

This chapter is concerned with the broad topic of bacterial *cell biology*, the study of the molecular mechanisms that determine the shape, size, division, differentiation, and development of bacterial cells. This includes the cytoskeleton, composed of different protein filaments with a variety of functions, as well as the complex machinery of cell division. We will also discuss development and differentiation in different bacterial species and conclude with a discussion of the use of fluorescent protein fusions, which have been invaluable tools in the study of these systems.

THE BACTERIAL CYTOSKELETON

Different species of bacteria contain homologs of all of the main cytoskeletal proteins found in eukaryotes, with a few extras for good measure. These typically have very little primary sequence similarity to their eukaryotic equivalents, but the homology becomes clear when their three-dimensional structures are compared, and they carry out many of the same kinds of functions in maintaining cell shape, organizing organelles within the cell, segregating DNA molecules, and driving cell division. All of the different types of cytoskeletal proteins are characterized by an ability to form filaments (at least *in vitro*), although the physical properties and dynamics of the resulting filaments differ.

Most rod-shaped bacteria contain one or more [actin](#) homologs, usually called MreB after the *E. coli* protein (encoded by the second gene in the *mre* murein formation gene cluster **E** operon). Gram-negative rods typically encode one MreB homolog, while many Gram-positive rods encode multiple copies. Mutations in *mreB* affect the shape of the bacterial cell, with single amino-acid changes in MreB able to change the width, evenness, or curvature of the cells, and even lead to branching morphologies. MreB is essential in most bacteria that encode it, but inhibition or depletion of MreB leads to rounding of cells, loss of rod shape, and eventually (after several rounds of cell division) lysis.

In vitro, MreB assembles into filaments structurally very similar to those of eukaryotic actin. This is affected by the presence of magnesium, other salts, and in some species, ATP or GTP concentrations. *In vivo*, MreB filaments are of varying length, associated with the inside of the cell membrane, and **move** around the circumference of the cell. In *B. subtilis*, the three MreB homologs are colocalized and distributed along the length of the cell throughout growth (Figure 11.1). In the α -proteobacterium *Rhodobacter sphaeroides*, MreB is found primarily as a ring in the middle of the cell. In *E. coli* and *Caulobacter crescentus*, these two patterns interconvert, depending on the stage of cell division.

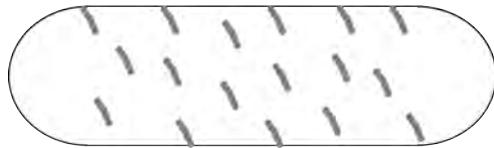


Figure 11.1. MreB filaments distributed around the length of an elongating rod-shaped bacterial cell. The MreB filaments move around the circumference of the cell, and are colocalized with the peptidoglycan synthesis machinery. Early microscopy data suggested that MreB formed a continuous helix, but higher-resolution images clarified that MreB filaments are relatively short, but highly mobile.

MreB is the central organizer of peptidoglycan synthesis enzyme complexes. MreB colocalizes with the sites of new peptidoglycan synthesis, and peptidoglycan synthesis is necessary to drive the motion of MreB filaments. The protein complex involved is called the **elongasome**, and links the cytoplasmic and periplasmic components of the cell wall synthesis machinery ([Lecture 10](#)). The diameter of each species' cells correlate with the degree of curvature of that species' MreB filaments, and MreB localization depends on the local geometry of the cell surface. By directing peptidoglycan synthesis to regions of appropriate curvature, MreB creates a feedback loop that maintains the desired cell shape.

MreB and MreB homologs can also have other functions, which vary from species to species, but include roles in chromosome segregation, motility, development, and the positioning of different cellular components.

Many bacteria are not simple spheres or rods, and one of the simpler variations in the rod shape is the presence of varying degrees of curvature or helicity (which are actually the same thing, depending on the degree of curvature relative to the length of the cells)(Figure 11.2).

In *C. crescentus*, cell curvature depends on a protein called *crescentin* (or CreS) that is homologous to eukaryotic [intermediate filaments](#). Crescentin forms non-mobile, stable filaments anchored to the inside of the inner membrane along the short axis of the gently-curved cells and maintains curvature by a combination of mechanical force and local effects on peptidoglycan synthesis. The localization of crescentin depends on MreB activity, but the interactions between these proteins must be in well-conserved regions of MreB, since expressing CreS in *E. coli* is sufficient to curve *E. coli* cells.

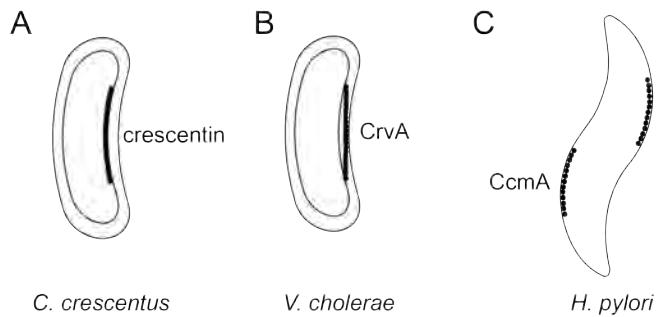


Figure 11.2. The localization of cytoskeletal proteins involved in cell curvature in *Caulobacter crescentus*, *Vibrio cholerae*, and *Helicobacter pylori*.

In *Vibrio cholerae* (and presumably other *Vibrio* spp., although to my knowledge this has not been studied extensively), cell curvature is maintained by an unrelated filament-forming protein called CrvA. CrvA is not related to any known eukaryotic cytoskeletal proteins, and, unlike crescentin, assembles into a stable filament in the **periplasm**. CrvA localizes to the inner curve of the cell, where it biases the relative rates of peptidoglycan synthesis and turnover; so that the cell wall grows faster on the other side of the cell.

Another cytoskeletal protein family common in bacteria but absent from eukaryotes are the *bactofilins*. These have a variety of functions, but as one example, in the tightly coiled helical bacterium *Helicobacter pylori*, the bactofilin CcmA localizes to the major axis of the helix and **enhances** peptidoglycan synthesis (by a mechanism that is still unclear). This is essentially the opposite of the negative effects that CreS and CrvA have on cell wall synthesis, but has the same effect of adjusting cell shape by biasing peptidoglycan synthesis to one side of the cell.

Finally, the most widely conserved cytoskeletal element in bacteria is the *tubulin* homolog FtsZ (filamentous temperature-sensitive mutant Z), which plays a central role in cell division, a topic that certainly deserves its own section heading:

BACTERIAL CELL DIVISION

Cell division is an exceptionally complicated process, in which DNA replication, chromosome segregation, growth, cytokinesis (the separation of the cytoplasm of daughter cells from one another), cell wall synthesis, and daughter cell separation all must be carefully coordinated. I'm not going to go into great detail here, since that would take a textbook all on its own, but the role and function of FtsZ is critical and is an important topic to cover in any discussion of the bacterial cytoskeleton.

With only a few exceptions, all bacteria contain FtsZ, and the *ftsZ* gene is essential. Like eukaryotic tubulin, FtsZ is a GTPase whose assembly into filaments is dynamic and regulated by GTP binding, which stimulates assembly, and GTP hydrolysis, which stimulates disassembly. FtsZ is the key component of the *Z-ring* that forms at the septum between dividing bacteria cells, recruiting a cascade of proteins to direct cell division (forming a complex called the *divisome*) and providing at least part of the contractile force that shrinks the *Z-ring* and separates the two daughter cells (a process known as cytokinesis; Figure 11.3).

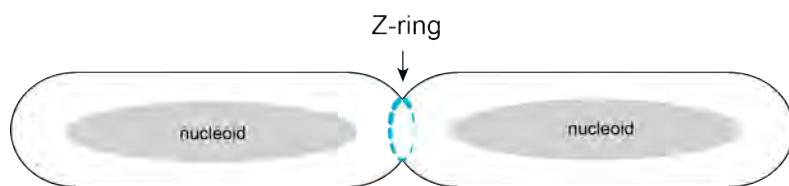


Figure 11.3. FtsZ is the central organizer of the Z-ring and divisome, the protein complex that carries out the process of cell division.

FtsZ in the Z-ring is not contiguous, but forms highly dynamic short filaments that move around the circumference of the cell by *treadmilling*, rapidly polymerizing at one end while simultaneously depolymerizing at the other end. FtsZ filaments are anchored to the cell membrane by a protein called FtsA (an actin homolog, and therefore related to MreB) and, in some species (including *E. coli*), an additional linker protein called ZipA. FtsA, in turn, recruits a complex of peptidoglycan synthesis enzymes that synthesize the septum dividing the two daughter cells. The divisome contains many proteins that are recruited in a very specific order, the details of which are beyond the scope of what we want to discuss here. [This review](#) is a good place to start for more information, if you're interested.

Z-rings only form at the site of cell division (mid-cell in bacteria that divide by binary fission), and there are several mechanisms by which this localization is enforced. Here, I will discuss the three partially redundant systems that operate simultaneously in *E. coli* as well as a separate system found in *C. crescentus*.

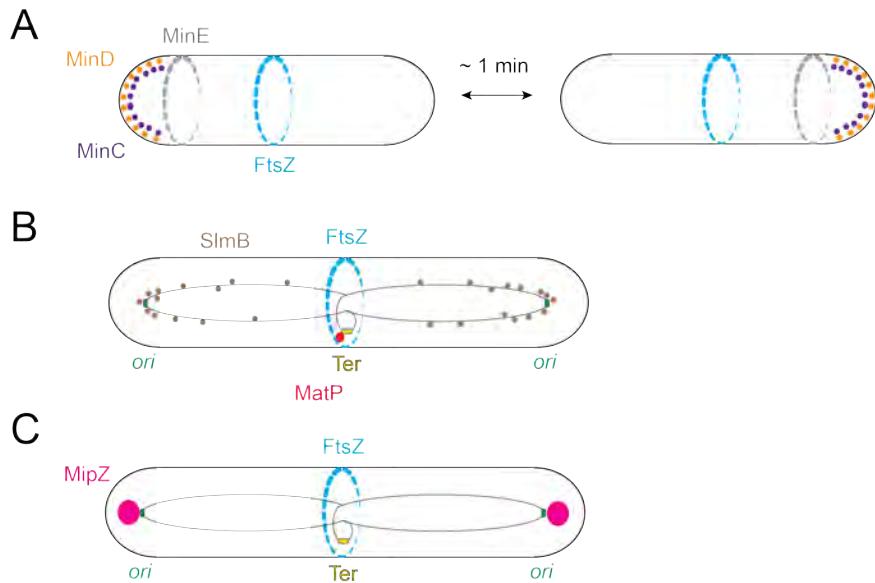


Figure 11.4. Mechanisms of Z-ring positioning. (A) In *E. coli*, MinC (purple) is an inhibitor of FtsZ (blue) polymerization. MinD (orange) localizes MinC to the membrane near the poles of the cell and, due to the activity of MinE (grey) oscillates from one pole of the cell to the other about once every 60 seconds, keeping the concentration of MinC at the center of the cell low. (B) Also in *E. coli*, the FtsZ inhibitor SlmB (brown) binds throughout the genome, but is enriched near *ori* and not present near *Ter*, while the FtsZ binding protein MatP (red) tethers FtsZ directly to *Ter*. (C) In *C. crescentus*, the FtsZ inhibitor MipZ (pink) binds to *ori*.

The *E. coli* *min* (mini-cell) mutants were discovered in 1967, and are characterized by the abundant production of tiny, DNA-free cells that cannot replicate, but do remain metabolically active for many hours. Ultimately this phenotype was linked to the *minCDE* operon, disruption of which causes the Z-ring to form off-center in many dividing cells.

MinC is an inhibitor of FtsZ polymerization. MinD binds to MinC and tethers it to the cell membrane (Figure 11.4 A). MinE is a small (88 amino acid) protein that has multiple functions: it forms filaments and rings, it binds to both the cell membrane and MinD, and it stimulates the ATPase activity of MinD, resulting in MinD monomerization, detachment from the membrane, and the release of MinC. MinD immediately rebinds to ATP, allowing it to reform the complex that tethers MinC to the membrane, but it tends to reassemble as far as possible from the MinE ring. The ultimate result of these dynamic interactions is that the MinCD proteins **oscillate** back and forth from one pole of *E. coli* to the other with a regular period of about 1 minute. This means that, on average, the concentration of MinC is highest at the poles of the cell, and Z-rings tend to form at mid-cell, where MinC is, on average, at its lowest concentration.

B. subtilis lacks MinE, and the MinCD proteins are tethered to the cell poles by proteins called MinJ and DivIVA. There is therefore no oscillation in the Min system of *B. subtilis*, but the principle of minimizing MinC concentration at mid-cell to allow Z-ring formation equidistant from the poles of the dividing cell is the same.

E. coli has two additional known systems for Z-ring positioning, both of which link Z-ring formation with chromosome replication. Most bacteria have circular chromosomes with a single origin of replication (*ori*) and, on the opposite side of the circle, a somewhat larger DNA region called the *Ter macrodomain*, which is where the two replication forks meet and replication terminates. The DNA replisome, containing DNA polymerase and both replication forks, tends to be found near the center of the cell, and as the two daughter chromosomes are produced, they are pushed outward, towards the poles. This means that, as the DNA is being replicated, *ori* (which was replicated first) will tend to be found near the poles and *Ter* (replicated last) will be in mid-cell.

MatP (**macrodomain Ter protein**) binds to the chromosome at *matS* sites in the *Ter* macrodomain and also forms a protein-protein complex with ZapAB (**Z** ring-associated **p**roteins), which are FtsZ-binding proteins (Figure 11.4 B). This directly links FtsZ and the Z-ring to *Ter*, providing a positive localization signal. In *C. crescentus*, ZapA forms a complex with proteins called ZauP and ZapT to anchor the Z-ring to the *Ter* macrodomain, achieving the same goal.

SlmA (**synthetically lethal with a defective min system**) is a negative regulator of FtsZ polymerization that binds to specific DNA sequences that are found throughout the chromosome, but are enriched close to *ori* and not present in *Ter* (Figure 11.4 B). The activity of SlmA is the basis of a phenomenon called *nucleoid occlusion*, the observation that Z-rings in *E. coli* cannot form around regions of the cytoplasm containing large amounts of DNA. Nucleoid occlusion in *B. subtilis* is mediated by the Noc protein, which is unrelated to SlmA and works by sterically inhibiting the diffusion of FtsZ away from the Z-ring, resulting in the same ultimate effect: formation of a tightly localized Z-ring structure.

C. crescentus lacks homologs of both the *E. coli* Min and nucleoid occlusion systems, but does have a protein called MipZ (**mid-cell positioning of FtsZ**) which *E. coli* and *Bacillus* lack. MipZ binds to *ori* and interacts with proteins involved in chro-

mosome segregation, and is therefore localized to the poles (Figure 11.4 C). It is also an FtsZ inhibitor. MipZ is therefore more or less the functional opposite of MatP, and provides an *ori*-specific Z-ring inhibition signal.

Other Z-ring localizing systems have been discovered in other bacteria, including filamentous and spherical species, but the mechanisms by which these function are much less well understood.

PLASMID PARTITIONING

As a final example of bacterial cytoskeletal elements, I want to discuss two distinct systems that are involved in the segregation of low copy number plasmids into daughter cells during cell division. High copy number plasmids do not typically have dedicated partitioning systems, since the odds of a daughter cell not containing any plasmids is relatively low.

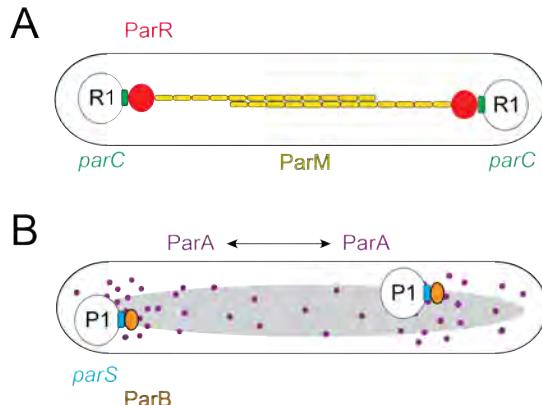


Figure 11.5. Examples of plasmid partitioning mechanisms. (A) For plasmid R1, ParR (red) binds the *parC* site on the plasmid (green). ParM (yellow) polymerization is stabilized by binding to ParR and by interactions with other ParM filaments, pushing the plasmids to the poles of the cell. (B) For plasmid / phage P1, ParB (orange) binds the *parS* site on the plasmid (blue). The MinD homolog ParA (purple) oscillates along the nucleoid (grey), creating concentration gradients along the length of the cell. ParB binds transiently to ParA, and moves P1 to regions of higher ParA concentration.

The R1 plasmid, which confers multi-drug antibiotic resistance to *Salmonella* and other *Enterobacteriaceae*, depends on the actin-like protein ParM to partition replicated plasmids between daughter cells (Figure 11.5 A). ParM very dynamically polymerizes and depolymerizes at about equal rates. ParM also interacts with ParR, a protein that specifically binds the *parC* site, which is a DNA sequence near the origin of replication of R1 and is functionally similar to the centromere of eukaryotic chromosomes (which is what the C in *parC* stands for). When ParM binds to ParR, the ParM filament is partially stabilized. If two ParM filaments encounter each other, they can interact to form an antiparallel complex that is much more prone to assemble into longer filaments than to disassemble. The end result is that the ParM filaments extend, pushing the attached ParR protein and attached plasmids as far apart as possible within the cell (i.e. to the poles).

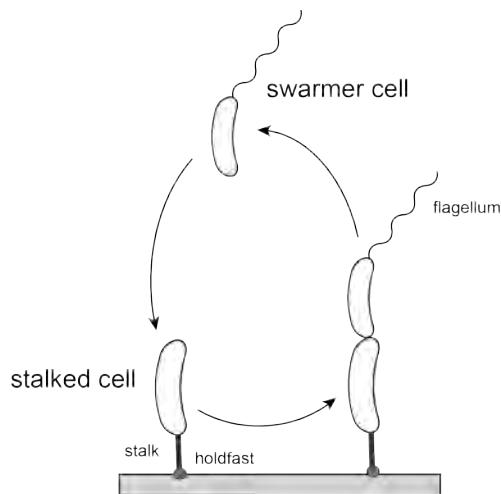
Bacteriophage P1 exists as a single-copy plasmid in P1 lysogenic *E. coli*, and has the best studied example of a very common plasmid segregation system consisting of the DNA-binding protein ParB and the MinD homolog ParA (Figure 11.5 B). Like ParR, ParB binds to a specific DNA sequence on the plasmid, in this case called *parS*. ParA interacts both with ParB and, in a non-specific, ATP-dependent manner, with the DNA of the nucleoid. ParA oscillates along the nucleoid, creating concentration gradients along the length of the cell, and the ParB / plasmid complexes are dragged along these gradients, ultimately settling in positions as far apart from one another as possible. The exact mechanism by which this occurs is an area of [active research](#).

These are by no means the only mechanisms for DNA segregation in bacteria, which are extremely diverse. Some species bacteria use Par systems similar to the plasmid systems described above to segregate their chromosomes. The *C. crescentus* chromosome, for example, has a ParAB system, with *parS* near *ori*, that is required for proper chromosome partitioning between daughter cells.

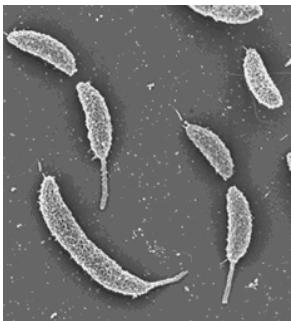
DISCUSSION PROBLEM SET #21: BACTOFILINS FROM CAULOBACTER CRESCENTUS

As you can perhaps tell from the mention of crescentin, MipZ, and ParAB above, *Caulobacter crescentus* is one of the flagship model organisms of bacterial cell biology, largely due to the pioneering work of Lucy Shapiro and many of her trainees. See [this review](#) for a great summary of this fascinating and important organism, whose cell biology differs substantially from that of the more conventional model organisms *E. coli* and *B. subtilis*.

C. crescentus is an α -proteobacterium found in freshwater streams and other dilute environments under flow. To thrive in these environments, it has a relatively complex cell cycle, alternating between motile swarmer cells and surface-attached stalked cells, as shown in the figure below.



You can see both types of cells in this electron micrograph (from [MicrobeWiki](#)):



Only the stalked cells divide, budding motile swarmer cells off from the non-stalked pole. This asymmetric cell division from one pole is typical of α -proteobacteria, and distinct from the binary fission found in *E. coli* and *B. subtilis*. Swarmer cells swim until they encounter a nutrient-rich environment, then differentiate into stalked cells, growing a stalk (or *prostheca*, if you're fancy), which is a membrane-enclosed, cytoplasm-containing appendage tipped with an adhesive *holdfast*. The holdfast allows them to stick firmly to surfaces, anchoring them in place in the new location.

The genome of *C. crescentus* encodes two bactofilins: BacA, encoded by the *bacA* gene (locus tag CC1873) and BacB, encoded by *bacB* (locus tag CC3022). Purified BacA and BacB can form filaments *in vitro*. In *C. crescentus*, these proteins form flat sheet-like structures and localize to the stalked pole by an unknown mechanism. Mutants lacking either bactofilin have defects in stalk formation, and there seems to be some interaction between BacAB and MreB.

In order to study the *in vivo* function and assembly of BacA and BacB in isolation, you decide to express these proteins in *E. coli*, which has no native bactofilin homologs. While codon usage is similar in these two organisms, promoters are quite different, and you cannot count on expression signals from one species functioning properly in the other.

All molecular methods are available for *E. coli*, and the following methods are available for *C. crescentus*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (4 Mbp)	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
inducible promoters available	✓
compatible transposon	✓

Design a genetic engineering strategy to express BacA and BacB in *E. coli*. Explain the necessary features of any plasmids you intend to use (**Lecture 5**), the molecular methods you will employ (**Lectures 7 & 8**), and the rationale behind your choices.

Design an experiment using your BacAB expression construct(s) to determine the effect(s) of BacAB expression in *E. coli*. State:

- what observations you plan to make
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- whether your proposed experiment tests correlation or causation
- potential outcomes of your experiments, and how you will interpret them

DEVELOPMENT

Development in bacteria can be defined in a number of different ways, including *asymmetric cell division*, like that of *Caulobacter*, where each daughter cell is genetically identical to but has a different phenotype than the mother cell, or *differentiation*, where a proportion of cells in a population change their phenotypes (and sometimes genotypes, in cases of irreversible *terminal differentiation*) according to a defined program.

By its nature, development of either variety involves sophisticated regulation and coordination of gene expression within and between cells, and we will examine just a few fairly well-characterized examples here. Note that the regulatory pathways involved are quite complex, and that I do not expect you to memorize the details. The general patterns of how cell differentiation is regulated in different bacteria is the important take-home message here.

SPORULATION

We will begin by discussing *sporulation* in Gram-positive rod-shaped bacteria, a form of asymmetric cell division where one daughter cell is a metabolically-inactive, highly stress-tolerant spore. Spores from these species (technically, *endospores*, since they are formed inside of the *mother cell*) are exceptionally resistant to killing. To kill *Clostridium botulinum* spores and prevent the possibility of food-borne botulism, for example, food canning protocols must ensure that every part of the product spends at least 10 minutes at 121°C. This is why we sterilize bacteriological growth media in an autoclave. Boiling is not sufficient to sterilize any solution that contains spores, which germinate to produce growing vegetative cells when conditions are permissive for growth. Bacterial spores can remain viable for centuries.

The process of sporulation has been studied in the model organism *Bacillus subtilis* for many decades, and the terminology used here is from that organism. Sporulation in other species differs slightly, but the basic process and regulatory mechanism is conserved. The species that can form endospores have classically been divided into the genera *Bacillus*, for aerobes, and *Clostridium*, for anaerobes, but both of those genera have been subdivided extensively based on modern phylogenetic methods, as mentioned in **Lecture 10**.

Sporulation has classically been divided into a 7-step process (Figure 11.6):

1. Stage 0 is regular cell division by binary fission, generating two identical vegetative daughter cells.
2. Stage I is characterized by the formation of a Z-ring off-center in the cell, and the beginning of constriction of the peptidoglycan septum in between the *mother cell* and the forespore, the cell which will eventually become the spore.
3. In stage II, the cytoplasm of the mother cell and forespore are completely separated and DNA replication is completed, with a complete chromosome in each cell.
4. In stage III, the mother cell engulfs the forespore, leading to a double membrane around the forespore.
5. In stage IV, a thick peptidoglycan cortex is synthesized in between the two membranes of the forespore.
6. In stage V, the outer layers of the spore (the *spore coat*) are added, which consist of a thick modified peptidoglycan layer and several different protein layers.
7. In stage VI, the spore is dehydrated and matures into its stress-tolerant final state.
8. In stage VII the mother cell lyses to release the spore.

The nomenclature of genes involved in sporulation is based on the 7-stage process described above. Genes which, when mutated, cause the cell to be stuck in stage 0 are called *spo0*, so, for example *spo0A*, *spo0B*, and *spo0F* are all genes that are

necessary for the initiation of the sporulation process. Similarly, *spolI*D and *spolI*M mutants have engulfment defects and never reach stage III. There are over 100 genes absolutely required for sporulation in *B. subtilis* and many more that influence the process, which illustrates the complexity of even this relatively simple developmental program.

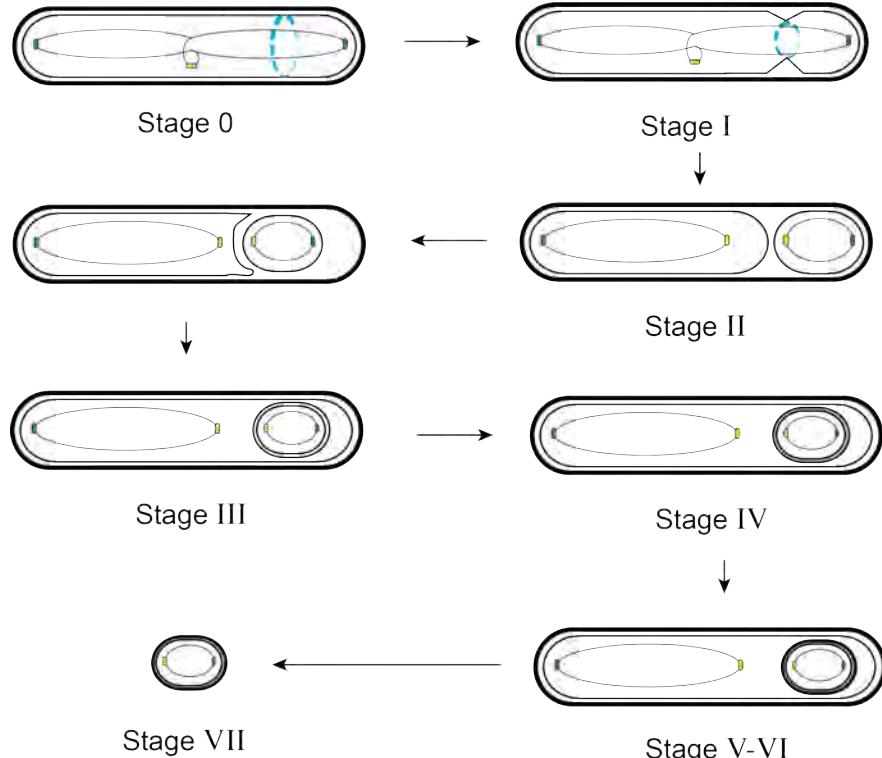


Figure 11.6. The stages of sporulation in *B. subtilis*. Stage 0 is vegetative cell growth, and sporulation begins with the formation of an off-center Z-ring (blue). Stage I begins the separation of the mother cell and forespore, with about 1/3 of the replicated chromosome trapped in the forespore. At stage II, septation is complete and the spore chromosome is completely translocated into the forespore. The mother cell engulfs the forespore at stage III, and a peptidoglycan cortex is assembled between the double membranes of the forespore in stage IV. Stages V and VI involve the assembly of the outer coat layers and final maturation of the spore, and stage VII is the death and lysis of the mother cell, releasing a mature endospore.

The signal that triggers sporulation is nutrient limitation, and in *B. subtilis*, starvation in stationary phase causes about half of the bacterial population to sporulate. Many of the remaining cells lyse, about 10-20% become competent, and some increase their mutation rate, all of which is presumably a bet-hedging strategy to ensure that the population as a whole has the best chance of surviving starvation. We will only consider the regulation of the sporulation cascade here, but many of the same regulatory pathways do feed into the regulation of this population-wide phenotypic heterogeneity.

The master regulator of sporulation is the response regulator Spo0A, which is the endpoint of a phosphorelay consisting of 5 histidine kinases (KinA-E), two phosphotransferases (Spo0B and Spo0F) and several phosphatases and kinase inhibitors (Figure 11.7). Note that phosphorelays in general are composed of more than two of the same histidine kinase and response regulator domains that make up two-component regulatory systems ([Lecture 4](#)). Phosphorylated Spo0A (Spo0A-P) regulates the expression of about 120 genes and is the trigger that begins the sporulation process.

Among the five histidine kinases, KinA and KinE are cytoplasmic and KinB, KinC, and KinD are transmembrane proteins. They coordinate to sense environmental conditions, but the mechanism by which they do so and exactly what signal(s) they sense is not fully known. KinA and KinB are the most important activators of sporulation, and some evidence suggests that they may be responding directly somehow to slow growth.

KinA and KinB phosphorylate Spo0F, which, in turn, transfers that phosphate to Spo0B, from where it can be transferred to Spo0A. KinC and KinE can directly phosphorylate Spo0A, and KinD acts to reduce Spo0A phosphorylation, although it is not known whether this is direct or indirect.

Acting against the kinases that phosphorylate Spo0A are phosphatases, including Spo0E, which dephosphorylates Spo0A-P, and RapA, RapB, RapE, and RapH, which dephosphorylate Spo0F-P. There are also kinase inhibitors like KipI and Sda that inhibit KinA and/or KinB, and, in turn, negative regulators of those inhibitors (KipA and ClpXP, respectively).

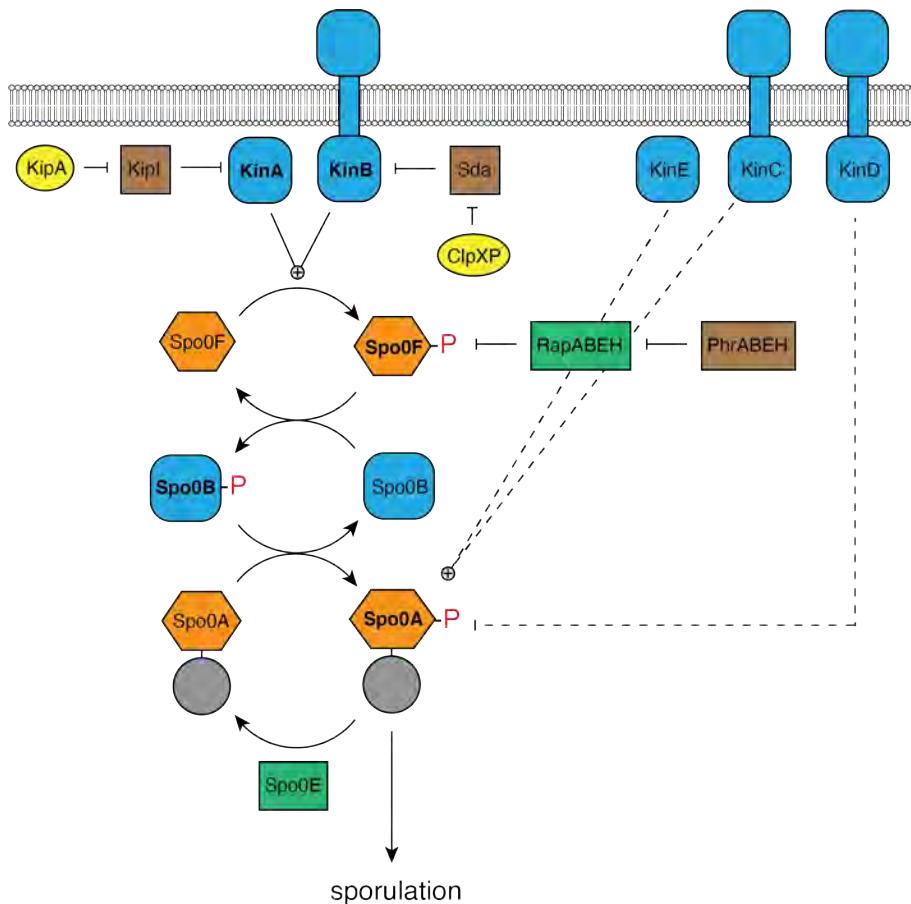


Figure 11.7. The phosphorelay that regulates initiation of sporulation in *B. subtilis*, as described in the text. This is, by necessity, an oversimplification of the actual phosphorelay, which incorporates many additional signals and proteins.

What is the advantage of this intricate phosphorelay? The general idea is that a more complex input system allows many different signals and regulators to be integrated into a single output: in this case, the phosphorylation state of Spo0A. Sda, for example, inhibits sporulation during periods of active chromosome replication, while the Rap phosphatases are inhibited by secreted quorum sensing peptides derived from the PhrA, B, E, and H proteins, therefore ensuring that sporulation only occurs at high cell density.

(Like other quorum sensing regulatory signals or *autoinducers*, the Phr peptides are secreted by bacterial cells and only exert their regulatory effects when they pass a threshold concentration in the extracellular milieu, therefore providing a chemical signal of culture density. This is a common mechanism for regulating population-level behaviors in bacteria, although the chemical signals used vary widely. Gram-positive quorum sensing systems often use peptide autoinducers. See **Lecture 14** for more on quorum sensing.)

Once Spo0A reaches a threshold level of phosphorylation and sporulation begins, the developmental progression is driven by a series of alternative sigma factors (**Lecture 4**) that progressively direct different transcriptional programs in the mother cell and the spore (Figure 11.8). The first two of these are σ^E and σ^F , expression of both of which is activated by Spo0A-P. Both of these sigma factors are normally kept in an inactive state by their respective anti-sigma factors.

At stage I of sporulation, the septum between the mother cell and forespore is almost complete, separating the cytoplasm of the two cells, but only about a third of the replicated chromosome is present in the forespore. The gene for σ^F itself is located near *ori*, but the gene for SpollAB (the σ^F anti-sigma factor; which is a proteolytically unstable protein) is located near Ter. This means that in the forespore SpollAB levels drop while σ^F levels rise, leading to expression of the σ^F regulon (about 50 genes) **only** in the forespore. Among the genes activated by σ^F is *spollR*, also close to *ori* and encoding the SpollR protein, which is secreted from the forespore into the mother cell. SpollR activates a protease (SpollGA) that degrades the σ^E anti-sigma factor, leading to expression of the σ^E regulon (about 300 genes) **only** in the mother cell. This is the beginning of the gene expression asymmetry that underlies differentiation.

As development continues, expression of σ^E -dependent genes in the mother cell generates a signal that leads to activation of the sigma factor σ^G in the forespore at stage III, which in turn, causes expression and secretion of a signal that activates

the σ^K regulon in the mother cell at stage V. This criss-cross mechanism of compartment-specific sigma factor activation is well-conserved among endospore-forming bacteria, even when many of the details of structural sporulation genes differ.

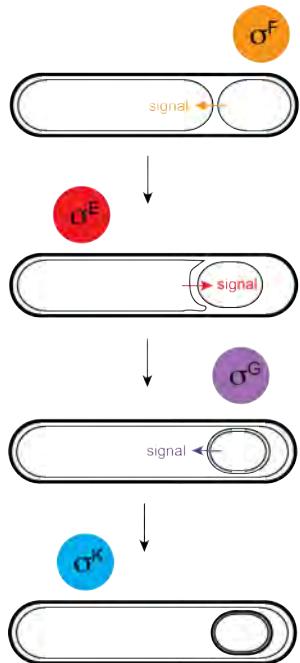


Figure 11.8. Alternating sigma factor activity in the forespore and mother cell that controls the progress of sporulation in *B. subtilis*. Each sigma factor activates expression of a signal that is transmitted to the other cell compartment, leading to activation of the next sigma factor in the cascade.

DEVELOPMENT IN MYXOBACTERIA

The myxobacteria are Gram-negative δ -proteobacteria that exhibit a variety of group or multicellular behaviors, including a multicellular developmental process that lead to the production of stress-tolerant *myxospores*. Myxospores are resistant to UV light, desiccation, and freezing, but are not especially heat-resistant, unlike Gram-positive endospores. (Myxospores die slowly at 60°C.) As we will discuss below, the pathway regulating their development is also quite different.

Myxobacteria are *predatory bacteria* that get their nutrients by killing other microbes. They do not do this as individual cells, but rather as a *swarm* of millions of cells that move collectively across surfaces and within the soil environment where they are found. We will discuss the mechanisms of myxobacterial motility in **Lecture 15**. The swarm secretes antibiotics and hydrolytic enzymes to kill, lyse, and digest prey bacteria. Before their bacterial nature was discovered, the myxobacteria were thought to be a type of slime mold, multicellular eukaryotic soil organisms with a similar appearance and behavior.

Myxospore formation, like that of endospores, is triggered by starvation, which, in the case of the myxobacteria, means that the density of potential prey bacteria has dropped below the threshold necessary to maintain growth of the myxobacterial swarm. Under those conditions, rather than spreading, the individual members of the swarm will move together, clumping together to form *fruiting body* structures. Within the fruiting body, 10% of the cells (usually about 100,000 per fruiting body) differentiate into myxospores, which are stuck together by a dense extracellular polysaccharide (EPS) matrix (see **Lecture 14** for more on EPS). This ensures that when environmental conditions improve and the myxospores germinate, there will be a large enough population of bacteria for efficient predation and growth. Note that, instead of forming within a mother cell during cell division, vegetative cells of myxobacteria differentiate directly into myxospores. Myxospores are metabolically inactive, dehydrated, spherical, and have thicker cell envelope structures than vegetative cells, all of which protect them from adverse environmental conditions.

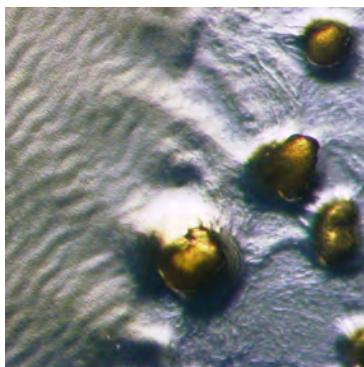


Figure 11.9. Low-magnification image of *Myxococcus xanthus*. When prey becomes limiting, fruiting bodies (yellow) form. Zalman Vaksman and Heidi Kaplan, University of Texas Medical School. CC BY 4.0. Obtained from Wikipedia.

Myxococcus xanthus is the best-studied model species of myxobacterium (Figure 11.9), and the regulation and development of myxospores is best understood in that species, although there are still many unknowns. It involves several interacting regulatory pathways, including two second messengers (Lecture 4), the alternative sigma factor σ^{54} , and multiple intercellular communication signals.

There are at least 5 signaling systems that are used by *M. xanthus* for intracellular communication during development, called the A-, B-, C-, D-, and E-signals. While genes required for the production of each signal have been identified (the *asg*, *bsg*, *csg*, *dsg*, and *esg* genes, respectively), only the chemical identity of the A-, C-, and E-signals have been identified, and the functions of only the A- and C-signals are reasonably well-understood (Figure 11.10). The E-signal is the branched lipid iso-15:0 O-alkylglycerol, but its role in development and regulation is currently unclear.

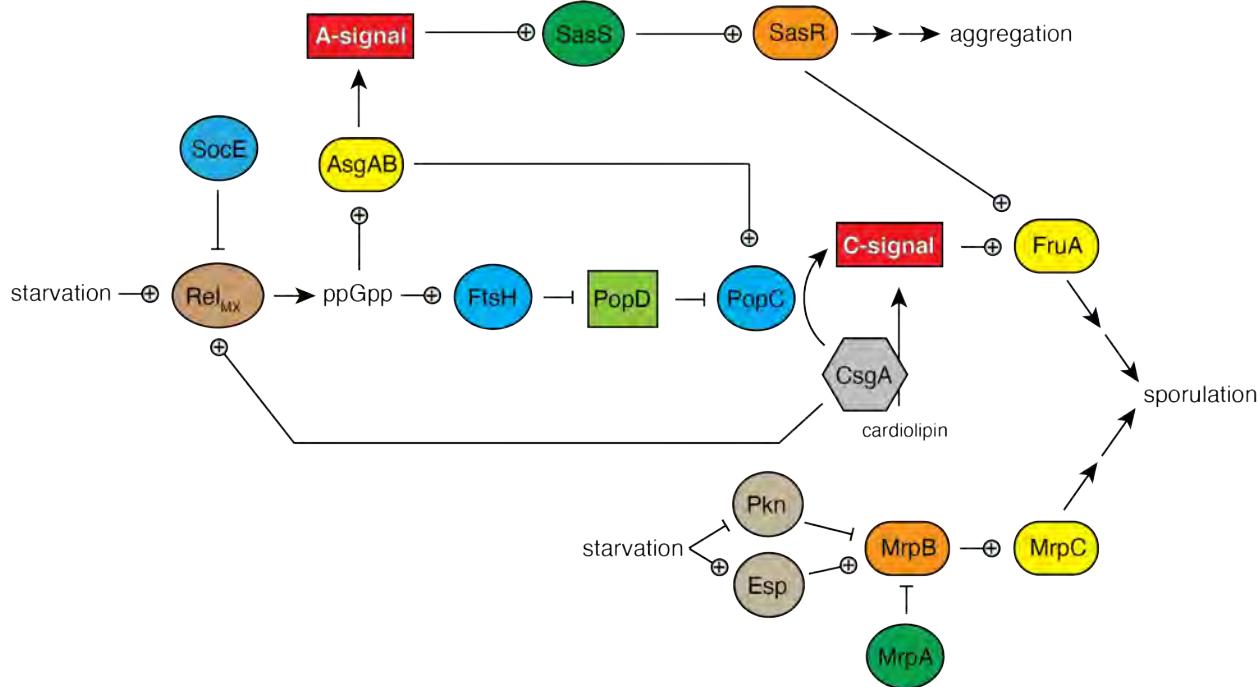


Figure 11.10. Stress-sensing pathways involved in the production of A- and C-signal in response to starvation stresses, leading to sequential expression of genes required for aggregation of fruiting bodies and those necessary for myxospore formation.

Starvation signals are sensed by multiple regulators in *M. xanthus*, but one of the key factors required for development is the second messenger guanosine tetraphosphate (ppGpp), which is a stress-responsive signaling molecule conserved in species across the bacterial domain. This signaling pathway, commonly called the *stringent response*, relies on ppGpp synthases and hydrolases (called RSH enzymes) that control growth rate and metabolism via the interaction of ppGpp with a wide range of proteins, including nucleotide synthases, translation factors, and, notably, RNA polymerase. The transcriptional signature of the stringent response typically includes upregulation of, for example, amino acid biosynthesis genes and downregulation of ribosomal RNAs, and has been extremely well-studied in *E. coli*, *B. subtilis*, and other model organisms.

In *M. xanthus*, ppGpp is synthesized by the RSH protein RelMX, which is associated with the ribosome and senses amino acid starvation by detecting uncharged tRNAs in the A-site of the ribosome. When ppGpp accumulates, the AsgAB transcript-

tional activator responds and upregulates genes involved in the production of the A-signal, which accumulates outside the cells early in development. The A-signal is a mixture of a specific set of amino acids (primarily Y, P, F, W, L, and I) produced by proteolytic cleavage of cell surface proteins. The presence of these amino acids at low concentration (about 10 μ M) acts as a quorum sensing signal, detected by SasS, which, in turn, activates SasR, an activator of genes necessary for the aggregation of swarming cells into a fruiting body. The A-signal is thought to ensure that development can only begin when there is high cell density, although, since the A-signal is composed of amino acids, additional regulators (including SocE and CsgA) are needed to maintain ppGpp production by RelMx during development.

The C-signal is related to a proteolytically cleaved form of the CsgA protein, an enzyme which localizes to the poles of the *M. xanthus* cell and itself has cardiolipin cleavage activity. It is unclear whether cleaved CsgA itself or the products of cardiolipin cleavage are the C-signal, but these possibilities are not mutually exclusive. Physical interaction between the C-signal-expressing poles of adjacent cells in a fruiting body are necessary for successful development. The production of the C-signal, which is important for later stages of development and differentiation into spores, is also dependent on ppGpp. CsgA is cleaved by the protease PopC, whose expression is activated by the ppGpp-activated regulators AsgAB, and whose activity is inhibited by the PopD protein. PopD is degraded by the protease FtsH in the presence of ppGpp, so ultimately ppGpp affects C-signal production via two different pathways.

The C-signal is detected by the transcription factor FruA, which activates expression of the genes for sporulation and development. Many of those genes **also** require activation by the MrpC transcription factor, which is activated by ppGpp-independent starvation signals from the Pkn and Esp sensing systems, by way of the MrpB activator of *mrpC* expression.

A more recently-discovered pathway, required for synthesis of the EPS matrix, depends on production of a different second messenger, cyclic di-GMP (often abbreviated c-di-GMP)(Figure 11.11):



Figure 11.11. The c-di-GMP-dependent pathway leading to expression of genes needed for extracellular polysaccharide synthesis in the fruiting body.

In this pathway, starvation signals are detected by the DifACE sensor system, which leads to activation of the DmxB enzyme. DmxB is a diguanylate cyclase that converts GTP into c-di-GMP. Accumulation of c-di-GMP leads to activation of the transcriptional activator Nla24, which is required for production of EPS. The genes regulated by Nla24 have not been exhaustively identified.

Like ppGpp, c-di-GMP is nearly universally conserved among bacteria, and plays diverse roles in controlling bacterial development, cell cycle progression, virulence, and biofilm formation. We will return to c-di-GMP in **Lecture 14**, when we will be discussing biofilm formation in depth.

Finally, layered on top of the regulatory pathways described above is a cascade of alternative sigma factor-dependent regulons (Figure 11.12). However, unlike *B. subtilis*, which employs alternative sigma factors related to the housekeeping sigma factor σ^{70} , *M. xanthus* depends on its homolog of the alternative sigma factor σ^{54} , which is **not** homologous to σ^{70} . The key feature of σ^{54} -dependent transcription is that σ^{54} on its own cannot activate transcription by RNA polymerase. That means that σ^{54} -dependent promoters require additional activators, which are transcription factors called **bacterial enhancer binding proteins** or bEBPs. *M. xanthus* contains a very high number of bEBPs (53, compared to 11 in *E. coli* K-12), meaning that it is able to very precisely control the activity of multiple σ^{54} -dependent regulons. bEBPs are often activated by phosphorylation (**Lecture 4**), so both their expression and activity can be regulated. *M. xanthus* is somewhat unusual among bacteria in that the gene encoding σ^{54} is essential in this species.

Three of the regulators we've already discussed are bEBPs: SasR, MrpB, and Nla24, all of which are required for successful sporulation. However, there are at least 7 more that regulate and are required for different aspects of the developmental process. These form what is called the bEBP cascade, since they are activated sequentially to drive a developmental program. This is analogous to the sigma factor cascade involved in *B. subtilis* sporulation, where each regulon includes activators of the next regulon in the sequence. In the case of the bEBP cascade, each bEBP's regulon includes the gene encoding the bEBP for the next step in the developmental sequence. However, the exact role(s) of the other genes controlled by each bEBP are much less well understood.

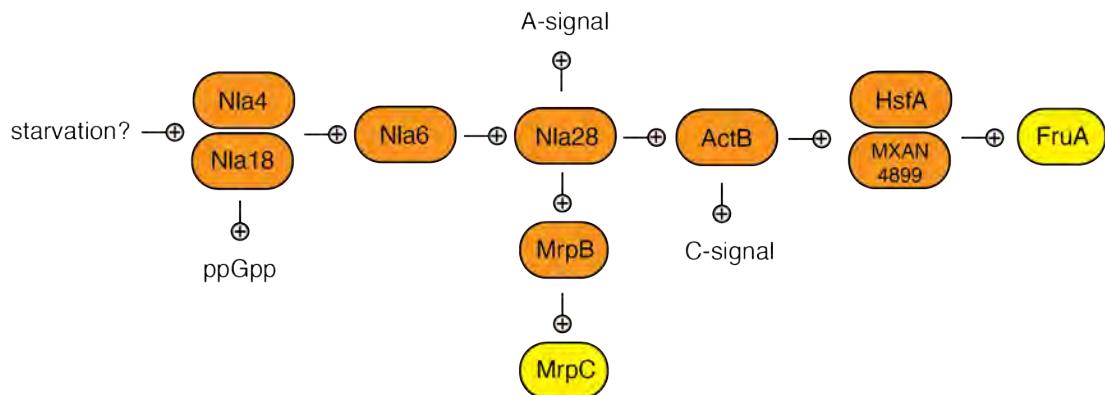


Figure 11.12. The bacterial enhancer binding protein cascade, consisting of sequentially activated σ^{54} -dependent regulons, all of which are required for successful development.

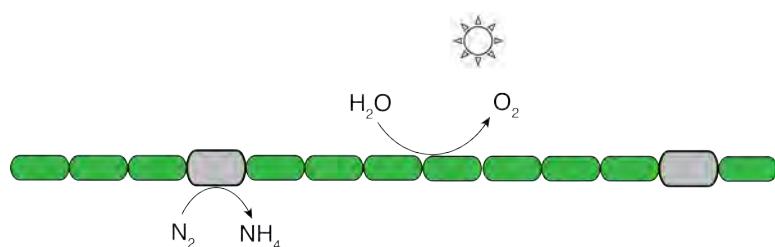
By an unknown mechanism, starvation leads to expression and activation of the bEBPs Nla4 and Nla18. Mutants lacking Nla4 and Nla18 do not accumulate ppGpp, although how this regulation works is unknown. Nla4 and Nla18 drive expression of Nla6, which is necessary for expression and activation of Nla28. Nla28 is required for production of both the A-signal and the bEBP MrpB (and therefore of MrpC). Nla28 is also required for expression of the bEBP ActB. ActB, in turn, is necessary for production of the C-signal and the final bEBPs in the cascade: HsfA and MXAN4899 (which really ought to be given a proper name, and honestly, the nomenclature in this field is a real mess). HsfA and MXAN4899 are required for FruA expression. As you might expect, each bEBP drives expression of quite a few genes, some of which have known functions and many of which do not. However, nearly a quarter of the genes in the *M. xanthus* genome are differentially regulated over the course of development.

There remain many unknowns in the field of myxobacterial development, and the regulators of differentiation overlap considerably with the regulators of multicellular predatory behavior and motility. This is an area of active research, although, since myxobacteria are not human or animal pathogens, it is also an area in which progress is relatively slow and limited to a small number of labs worldwide.

DISCUSSION PROBLEM SET #22: CYANOBACTERIAL HETEROCYSTS

Not all bacterial developmental processes lead to the formation of metabolically inactive resting cells.

Cyanobacteria are the Gram-negative oxygen-generating photosynthetic bacteria formerly known as the blue-green algae, and share a common ancestor with the chloroplasts of plants. In filamentous cyanobacteria belonging to the order Nostocales, some cells in a filament are able to differentiate into heterocysts, specialized non-photosynthetic cells that fix atmospheric nitrogen (N_2) into bioavailable ammonia (NH_4).

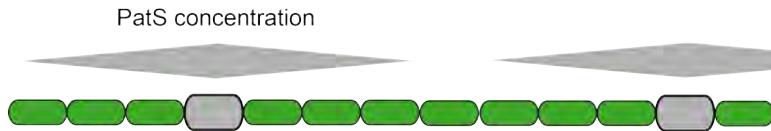


The problem that nitrogen-fixing cyanobacteria face is that nitrogenase, the enzyme that fixes nitrogen, is **very** sensitive to inactivation by oxygen, and oxygenic photosynthesis produces a **lot** of O_2 . Under nitrogen starvation conditions, therefore, some individual cells in each filament differentiate, losing the ability to photosynthesize, upregulating O_2 -degrading enzymes, and developing a thick, gas-impermeable cell envelope. This is an irreversible process, and heterocysts cannot revert back into photosynthetic cells. The photosynthetic cells provide carbon, in the form of sucrose, to the heterocysts, and the heterocysts, in turn, export fixed nitrogen in the form of glutamine.

The master transcription factor controlling heterocyst formation in the model cyanobacterium *Anabaena* sp. strain PCC 7120 is called HetR, and activation of HetR is required to initiate differentiation. HetR regulates a very large number of genes involved in the formation of heterocysts.

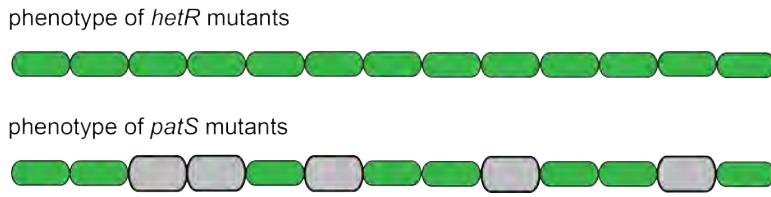
In wild-type cells, the spacing of heterocysts in a filament is extremely regular. This is regulated by the expression of a diffusible peptide (PatS) from mature heterocysts. PatS expression is activated by HetR and PatS itself inhibits heterocyst for-

mation. In a filament of *Anabaena*, the concentration of PatS is highest closest to a mature heterocyst and lower the further away from a heterocyst you get.

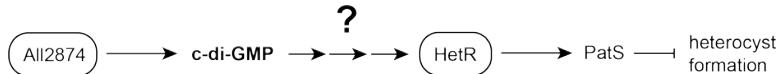


There is some debate about whether PatS diffuses through channels connecting the cytoplasm of adjacent cells or in the contiguous periplasm shared by all of the cells in a filament. Regardless, PatS inhibits formation of new heterocysts close to ones that already exist.

Null mutants lacking *patS* make too many heterocysts without even spacing between them, and mutants lacking *hetR* make no heterocysts at all.



There are 14 diguanylate cyclases (c-di-GMP synthases) in *Anabaena* sp. strain PCC 7120, but, of these 14, **only All2874** is required for heterocyst formation. An *all2874* null mutant does not form heterocysts or express *patS*. However, over-expressing HetR in the *all2874* mutant restores normal heterocyst formation. This indicates that c-di-GMP is in some way acting **upstream** of HetR in the differentiation signaling pathway, but does not clarify whether or not regulation of PatS is involved in the mechanism by which c-di-GMP influences heterocyst formation.



The following methods are available for *Anabaena* sp. strain PCC 7120:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (7 Mbp)	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposon	✓

Design an experiment to identify the mechanism by which c-di-GMP controls heterocyst formation in *Anabaena*. State:

- your hypothesis
- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment tests correlation or causation
- potential outcomes of your experiment, and how you will interpret them

FUSION PROTEINS AND BACTERIAL CELL BIOLOGY

In **Lecture 5**, I mentioned the construction of fusion proteins that combine domains from different proteins into a single polypeptide chain. This technique has been very important in the study of the internal structures of bacteria, especially using fusions with fluorescent proteins (most often derivatives of GFP, the green fluorescent protein from the jellyfish *Aequorea victoria* or of DsRed, a red fluorescent protein from the cnidarian coral *Discosoma*). Note that the term **fluorophore** is used

to refer to fluorescent molecules, including both proteins and small molecule dyes, which are commonly used together in microscopy experiments.

In this section, I will discuss how fluorescent protein fusions are constructed in more detail, some considerations about their use, and caveats that must be taken into account when designing experiments with protein fusions.

There are a wide variety of fluorescent proteins available with different properties that make them useful for different kinds of applications. The [FPbase](#) database is an excellent resource that describes the sequences and characteristics of fluorescent proteins. Key properties include color (both of the light needed to excite the proteins and of the fluorescent emission), brightness, and maturation time (fluorescent proteins take time to fold into fully active fluorescent forms, and this can vary from a few minutes up to several hours). Most, but not all, fluorescent proteins also require oxygen to mature into their active form, which limits the usefulness of fluorescent protein fusions in anaerobic bacteria (see [this review](#) for a discussion of progress in this area).

Fusions between bacterial proteins and fluorescent reporters (Figure 11.13) can be constructed using the techniques described in **Lectures 7** and **8** for expression in bacterial cells, either from a plasmid or from the chromosome. Fluorescent proteins can be fused to either the N- or C-terminus of a protein or into a surface-exposed loop within the protein.

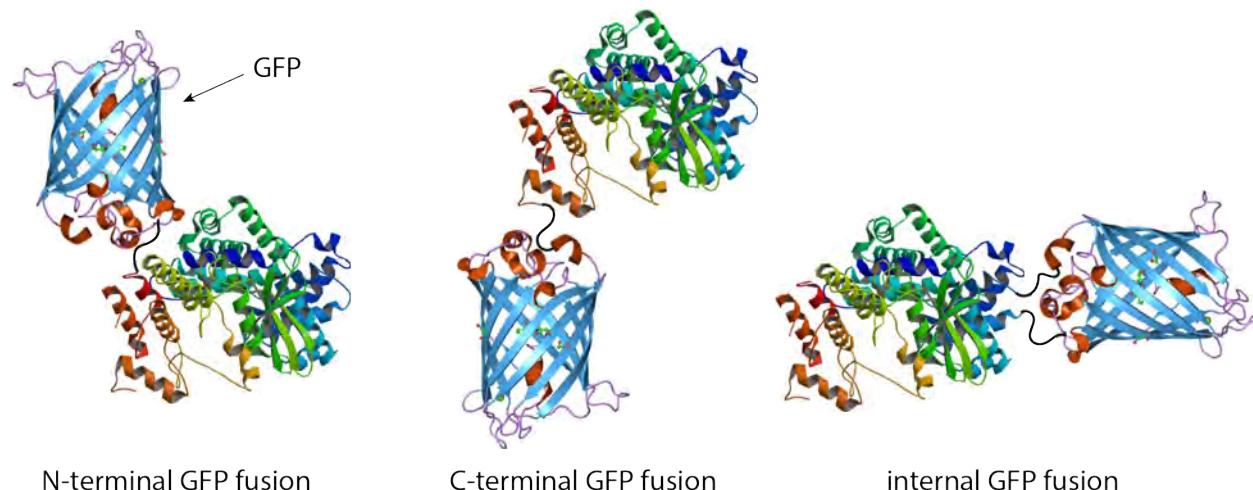
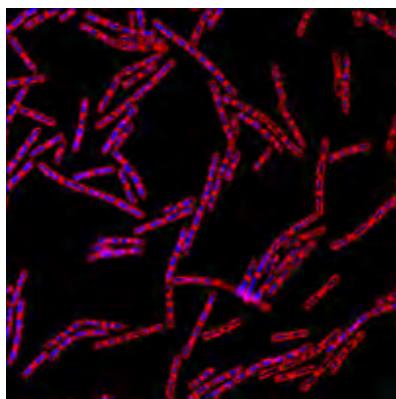


Figure 11.13. Fluorescent protein fusions. Fluorescent proteins, which are usually β -barrel proteins, can be fused to either terminus of the protein of interest or within a surface-exposed loop, as long as the presence of the fluorescent domain does not influence the folding or activity of the protein of interest. Linkers are represented by black lines. Ribbon structures from Wikimedia Commons.

Normally, flexible linkers made up of relatively inert amino acids (serine, alanine, etc.) are included between the proteins to minimize possible steric effects of the fluorescent protein (typically large, stable β -barrel domains) on the structure and function of the protein of interest. However, it is absolutely essential to confirm experimentally that the fusion protein retains the key function(s) of the wild-type protein.

The expression level of fusion proteins is also an important variable in this kind of experiment. Especially when using plasmids, it is important to not express fusion proteins at much higher levels than the wild-type protein. Not only could this lead to artifacts in visualizing the presence and localization of the protein of interest, high level expression of fluorescent proteins is often toxic to bacterial cells (due to increased production of reactive oxygen species). Careful selection of promoters is important, and Western blotting (see **Lecture 4**) to confirm that protein levels are in the expected range is a common control. (Antibodies to GFP and other fluorescent protein markers are commercially available.) It is also a good idea to confirm that the protein of interest remains stably fused with the fluorescent protein *in vivo*, since proteases may degrade either the poorly-structured linkers or the protein of interest itself. Fluorescent protein domains themselves are typically extremely stable and protease-resistant.

If the correct conditions are met, fluorescence microscopy can be used to visualize where protein fusions and, by extension, presumably also where the wild-type proteins localize within the bacterial cell. Since bacteria are so small (*E. coli* is approximately 1 μm long and about 0.6 μm in diameter), this is far less precise than equivalent experiments in eukaryotic cells, although larger bacterial species (e.g. *B. subtilis*, which are about twice as long and slightly wider than *E. coli*) can be used to make this somewhat easier.



UC Berkeley

Figure 11.14. A fluorescent micrograph of *Bacillus subtilis* cells, illustrating how cellular structures in bacteria can be visualized with fluorophores of different colors.

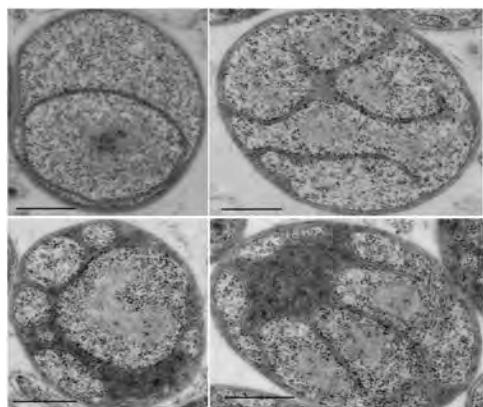
Confocal microscopy takes successive images of the same bacteria at different planes, allowing the reconstruction of 3D images of fluorescence for more precise localization of fluorescent proteins within a cell. At the absolute cutting edge of fluorescence microscopy, various methods for *super-resolution microscopy* have been developed that allow the detection of single molecules of fluorescent proteins within cells (see [this Wikipedia article](#) for a nice summary, and [these papers](#) for some applications in practice). These require extremely specialized equipment and sophisticated data analysis but are allowing genuinely astonishing advances in our understanding of how living cells are structured.

For a review on the use of fluorescence microscopy in bacterial cell biology, see [this paper](#). Considerations for the design of dual fluorescence reporter systems are discussed [here](#).

EXAMPLE OF AN EXPERIMENTAL DESIGN: PLANCTOMYCETES

There is one major group of bacteria that **lack** FtsZ: the members of the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) superphylum, which includes the pathogenic *Chlamydia* species and the ubiquitous environmental *Planctomycetes*. In addition to an unconventional (for bacteria) budding replication mechanism, *Planctomycetes* also have complex cell structures with internal membranes derived from invaginations of the cytoplasmic membrane, organelle-like structures, and at least some of them appear to be capable of both endocytosis and phagocytosis. (See [this review](#) for a summary.)

The best-studied model organism among the *Planctomycetes* is *Gemmata obscuriglobus*, found in freshwater environments around the world. It has, for example, been studied for its ability to synthesize sterols, which are not made by most bacteria.



(Wikipedia)

Electron micrograph of the planctomycete *G. obscuriglobus*, illustrating representative morphologies. Scale bar = 500 nm. Cropped top 4 panels from Figure 3 of: Santarella-Mellwig R, Franke J, Jaedicke A, Gorjanacz M, Bauer U, Budd A, et al. (2010) PLoS Biol 8(1): e1000281.

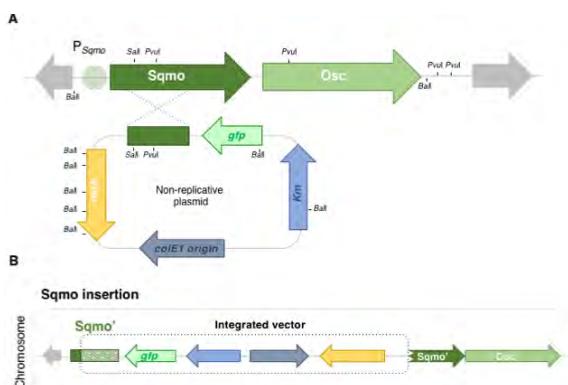
What is really not clear is how cell division in *G. obscuriglobus* and other PVC bacteria occurs in the absence of FtsZ. There are some recent hints in the literature for *Chlamydia* that might suggest a mechanism, though. The cell walls of PVC bacteria do not contain extensive peptidoglycan networks, but PVC bacteria are sensitive to penicillin (which specifically inhibits peptidoglycan synthesis) and **do** appear to synthesize at least some peptidoglycan during cell division. This, in combination with the observation that MreB is conserved in PVC bacteria, has led to the hypothesis that MreB and the associated peptidoglycan synthesis machinery somehow substitutes for FtsZ in PVC bacteria.

Can we develop an experiment to test this hypothesis in *G. obscuriglobus*? It's tricky, since genes required for cell division are certainly going to be essential, but perhaps we can employ some eukaryotic-style cell biology strategies in this rather large and complex bacterial cell.

The following methods are available for *G. obscuriglobus*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (9.2 Mbp)	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
suicide vector available	✓
selectable markers available	✓

Note that as far as I can tell, no one has developed shuttle vectors, expression vectors, or transposons for Planctomycetes, but it is possible to disrupt individual genes in a targeted way by introducing a suicide vector (by conjugation) that recombines into the *G. obscuriglobus* chromosome by single-crossover homologous recombination.



Supplemental Figure from [Rivas-Marin et al. 2019](#), showing the use of a suicide vector to disrupt the *G. obscuriglobus* Sqmo gene (involved in sterol synthesis). Note the presence of a kanamycin-resistance gene and a GFP marker in the integrative plasmid.

Hypothesis: MreB localizes to the site of cell division in *G. obscuriglobus*.

Experimental Design:

- 1) Construct a mutant of *G. obscuriglobus* that expresses a GFP-tagged MreB fusion protein.

This is complicated slightly by the fact that both N- and C-terminal fusions tend to inactivate MreB homologs or disrupt their functions, but there are multiple examples of functional "sandwich fusions" with fluorescent proteins inserted into loops within MreB. We will use iVEC to construct a suicide vector that will replace the chromosomal copy of *mreB* with an *mreB'-gfp-mreB* allele based on those fusions, conjugate it into *G. obscuriglobus*, and select for kanamycin resistance.

- 2) Use fluorescence microscopy to identify where in *G. obscuriglobus* MreB is localized in dividing cells.

Independent Variable: (what will you change?)

- The point in the cell cycle at which images of individual *G. obscuriglobus* cells are taken.

Dependent Variable(s): (what will you measure?)

- The location of MreB in individual *G. obscuriglobus* cells.

Negative Control(s): (eliminate false positive results)

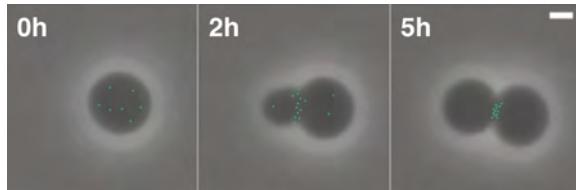
- Making a separate mutant or mutants expressing GFP-tagged protein(s) with alternative predicted localizations would be helpful in this case. For example, simply expressing cytoplasmic GFP or tagging an unrelated membrane protein with GFP would help eliminate the possibility that **any** fluorescent protein might tend to accumulate in the division site.

Positive Control(s): (eliminate false negative results)

- It will be important to confirm that your *mreB'-gfp-mreB''* fusion is being expressed and has retained a functional GFP domain attached to MreB. You could do this by qRT-PCR of the *mreB* and *gfp* RNA sequences and by Western blot with anti-GFP antibodies (to make sure the right-sized protein is being produced).

Potential Outcomes:

- 1) MreB localizes to the site of cell division, supporting the hypothesis that MreB is a functional substitute for FtsZ in *G. obscuriglobus*.



Adapted (by adding little green circles) from a phase-contrast image of dividing *G. obscuriglobus* in [Rivas-Marin et al. 2019](#).

- 2) MreB localizes elsewhere, arguing against the hypothesis.

Does this experiment establish correlation or causation?:

- This design tests the **correlation** between MreB localization and the site of cell division.

This is obviously only the first step in a more complicated series of experiments exploring cell division in *G. obscuriglobus*. It might, for instance, be interesting to label peptidoglycan in these cells during the cell cycle, to fluorescently tag other predicted cytoskeletal proteins, or to see how addition of penicillin affects any of those measurements over the course of the cell cycle.

LECTURE 12: CRITICAL READING (BACTERIAL CELL STRUCTURE)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
 - What is the central **question** of this paper?
 - Is the experimental design clear and appropriate to address that question?
 - Do you understand the methods used?
 - Are the data clearly presented, with appropriate statistics?
 - Do you agree with the conclusions the authors came to based on their data?
 - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Wan et al. (2022) "McaA and McaB control the dynamic positioning of a bacterial magnetic organelle." Nature Communications 13:5652.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

LECTURE 13: PROTEIN SECRETION

INTRODUCTION

Proteins are synthesized in the cytoplasm, but many of them need to be targeted to other places in the cell in order to function, including the cytoplasmic membrane, the periplasm, the outer membrane, or exported out of the cell entirely, including, for many pathogens, into host cells. This is called *protein secretion*, and bacteria have a wide variety of mechanisms to accomplish this goal. In this lecture, I will describe the basic operation of most of the currently known systems for protein secretion in bacteria (although new ones do keep getting discovered; since I first wrote this chapter there are now two additional named secretion systems in bacteria: types [X](#) and [XI](#)). Some of these secretion systems can also secrete other substrates, including DNA or small molecules.

Since most protein secretion systems are large multi-protein complexes, I will also discuss some genetic methods useful for detecting and studying protein-protein interactions *in vivo*.

THE SEC PATHWAY

The general secretion (Sec) pathway is conserved among bacteria, archaea, and eukaryotes, and is a very common mechanism by which proteins are transported across the cytoplasmic membrane. In Gram-negative bacteria, this results in substrate proteins being translocated into the periplasm or into the inner membrane itself, while in Gram-positive bacteria substrate proteins are excreted into the extracellular environment, modified to attach them to the cell surface, or inserted into the inner membrane. Sec substrates include many key metabolic proteins, as well as a variety of virulence factors. More than a third of the bacterial proteome is found in the cell envelope (membranes or periplasm), and a large proportion of those proteins utilize the Sec pathway to reach their final destinations.

The Sec system (Figure 13.1) translocates unfolded proteins, which are recognized by the presence of an N-terminal signal sequence, about 20 amino acids long with a positively-charged N-terminal sequence, a hydrophobic central sequence, and a polar C-terminal region. For proteins destined to be secreted into the periplasm or further exterior cellular compartments, the newly synthesized proteins are recognized and maintained in an unfolded state by the SecB chaperone protein. SecB targets the substrate to the ATPase SecA, which binds to the SecYEG protein channel in the inner membrane. ATP hydrolysis by SecA provides the energy for secretion of the substrate through SecYEG. A signal peptidase (also called LepB) proteolytically cleaves off the signal sequence, and the substrate protein folds into its final tertiary structure in the periplasm.

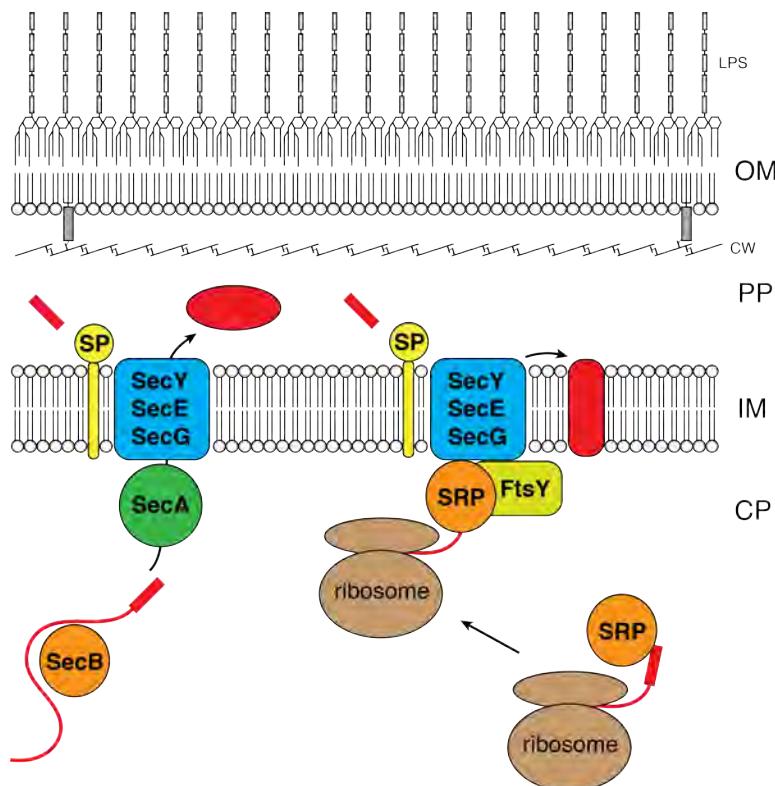


Figure 13.1. The Sec pathway. Proteins targeted to the periplasm (PP) or for secretion out of the cell are expressed with an N-terminal signal sequence (red rectangle) and are maintained in an unfolded state by the chaperone SecB. SecB delivers substrates to SecA, an ATPase which provides the energy to secrete the unfolded substrate through the SecYEG channel in the inner membrane (IM). The signal sequence is removed by signal peptidase (SP; also called LepB), and the secreted protein folds in the periplasm. Proteins targeted to the inner membrane have an N-terminal signal sequence that is recog-

nized as soon as it is translated by the signal recognition particle (SRP; 4.5S RNA and the Ffh protein), which recruits the docking protein FtsY. This complex interacts with SecYEG and the protein is translocated into the IM as it is being translated.

Integral inner membrane proteins exported by the Sec system are recognized by a different mechanism. The signal sequence for these proteins is recognized as soon as it is translated by the **signal recognition particle** (SRP), a riboprotein complex between 4.5S RNA (encoded by the *fif* gene) and the Ffh protein. The SRP, along with the docking protein FtsY, assembles a complex between the ribosome and SecYEG, so that these proteins are simultaneously translated and transported, with protein synthesis providing the energy needed to drive secretion. There appears to be a mechanism by which integral inner membrane proteins are released somehow from the side of the SecYEG channel, so that they end up inserted fully into the membrane.

Although Figure 13.1 illustrates the Sec system in a Gram-negative bacterial envelope, as mentioned above, Gram-positive bacteria also contain this secretion system. In fact, they often have two distinct copies of SecY or sometimes of the entire Sec system, each of which secretes a different set of proteins. Since there is no outer membrane in Gram-positive bacteria, in order to prevent proteins from diffusing away from the cell, Gram-positives use membrane-anchored enzymes called sortases that recognize a conserved LPXTG motif and covalently attach secreted proteins containing that motif to the peptidoglycan cell wall. Some species have multiple sortases with different specificities.

Since many inner membrane proteins are required for bacterial growth, the Sec system is essential, although some individual components (e.g. SecB, SecG, signal peptidase) can be knocked out without immediately lethal effects.

THE TAT PATHWAY

Unlike the Sec pathway, which secretes **unfolded** proteins, the **twin arginine translocation** (Tat) pathway secretes **folded** proteins (Figure 13.2). This is necessary when the proteins to be secreted contain cofactors, metal ions, or post-translational modifications that can only be synthesized in the cytoplasm or for substrates that are secreted as multi-protein complexes. Phospholipase C, which is a virulence factor secreted by *Pseudomonas aeruginosa*, *Mycobacterium tuberculosis*, and some other pathogens, is one example of an enzyme secreted via the Tat pathway.

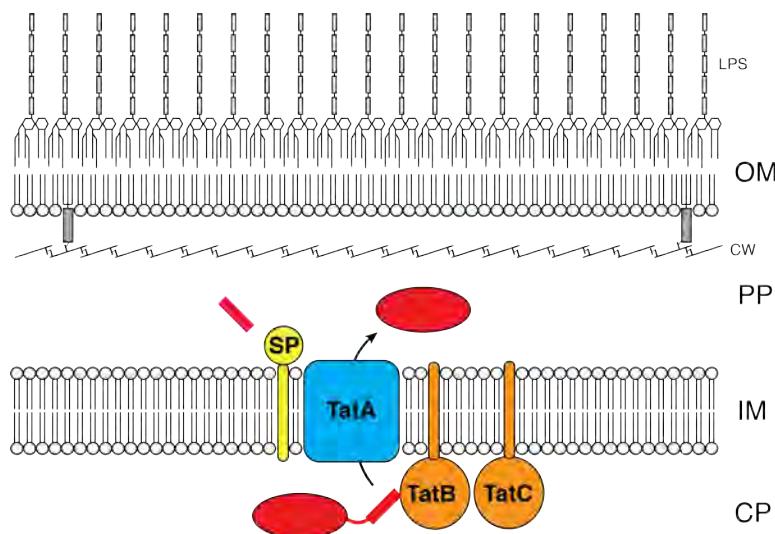


Figure 13.2. The Tat pathway. Folded Tat substrate proteins have an N-terminal signal sequence (containing a twin-arginine motif) that is recognized and bound by the TatB and TatC proteins, then secreted through the TatA channel. The signal sequence is removed by signal peptidase (SP; LepB).

Tat pathway substrates have an N-terminal signal sequence similar to the Sec signal peptide, but which contains the Ser-Arg-Arg motif that gives the system its name. This signal sequence is recognized by TatB and TatC, which deliver the substrate to the TatA pore, through which it is secreted into the periplasm or (in some cases) the inner membrane. The signal sequence is cleaved by the same LepB signal peptidase used in the Sec system. No ATP is needed for Tat-dependent secretion, and the energy necessary appears to be derived directly from the *proton motive force* (we will discuss bacterial energetics in [Lecture 16](#)).

As with the Sec system, both Gram-negative and Gram-positive bacteria contain the Tat secretion system, and homologous systems are found in archaea and some eukaryotes (although not animals). Perhaps surprisingly, the *tatABC* genes are not essential (at least in *E. coli*), although many Tat substrates **are** redox cofactor-containing proteins required for respiration (also see [Lecture 16](#)).

EXAMPLE OF AN EXPERIMENTAL DESIGN: TAT-DEPENDENT VIRULENCE FACTORS IN PSEUDOMONAS AERUGINOSA

Pseudomonas aeruginosa is an opportunistic pathogen that causes a wide variety of both chronic and acute infections, and is notable for acquiring extensive antibiotic resistance. Mutants of *P. aeruginosa* lacking the Tat secretion system ($\Delta tatC$ mutants are commonly used) are not as virulent as the wild-type. (While there are many *in vivo* animal models of *P. aeruginosa* infection, this particular phenotype has been reported in a rat pneumonia model.)

[Gimenez et al. \(2018\)](#) used a combination of *in silico* analysis of the *P. aeruginosa* genome to identify proteins with likely Tat signal sequences and an *E. coli*-based Tat-dependent screening system to validate which of those signal sequences could actually promote Tat secretion. They identified 34 putative Tat-dependent proteins in the *P. aeruginosa* genome (Table 2 of that paper), and proposed that one or more of these proteins was responsible for the impact of Tat on *P. aeruginosa* virulence. Some of them are known virulence factors (e.g. the phospholipases PlcN and PlcH or the iron-enterobactin transport protein FecB), but most of them have no obvious link to pathogenesis, and some have no predicted function at all.

The following methods are available for *Pseudomonas aeruginosa*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (6.3 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
generalized transducing phage ($\varphi PA3$)	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

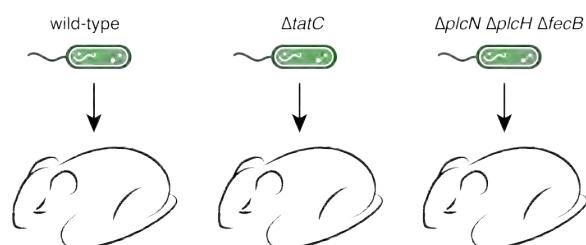
Design an experiment to determine the relative contribution of Tat-secreted proteins to *P. aeruginosa* virulence.

Hypothesis: Tat substrates other than PlcN, PlcH, and FecB contribute to *P. aeruginosa* virulence.

Experimental Design:

It would certainly be possible to generate all 34 individual mutants and assay them for virulence in a rat model, but this is tedious, inelegant, and would be both expensive and unethical. Let's get at this question a different way.

- 1) Using 3 cycles of allelic exchange mutagenesis (as described [in this paper](#)), construct a $\Delta plcN \Delta plcH \Delta fecB$ triple mutant of *P. aeruginosa*. Construct the necessary suicide vector plasmids using iVEC in *E. coli*.
- 2) Infect rats with wild-type, $\Delta tatC$, and $\Delta plcN \Delta plcH \Delta fecB$ strains of *P. aeruginosa*.



Made with Biorender. Unhappy rodent picture by Patrick Molina (it's a mouse, but you get the idea).

Independent Variable: (what will you change?)

- The strain of *P. aeruginosa* used for infection.

Dependent Variable(s): (what will you measure?)

- "Virulence", in this case as many parameters as possible, including animal weight, infectious burden (CFU of bacteria), cytokine panels to assess immune response, etc. These are all quantitative measurements.

Negative Control(s): (eliminate false positive results)

- The $\Delta tatC$ mutant is the negative control in this experiment.

Positive Control(s): (eliminate false negative results)

- The wild-type *P. aeruginosa* is the positive control.

Potential Outcomes:

- 1) The $\Delta tatC$ and $\Delta plcN \Delta plcH \Delta fecB$ strains have identical virulence defects. This will indicate that secretion of these three known virulence proteins fully explains the contribution of Tat secretion to virulence.
- 2) The $\Delta tatC$ mutant has a more severe virulence defect than the $\Delta plcN \Delta plcH \Delta fecB$ strain. This will indicate that there are additional virulence factors secreted via the Tat system (presumably one or more of the 31 other candidates with predicted Tat signal sequences).

Does this experiment establish correlation or causation?:

- This design tests the **correlation** between the presence of *plcN*, *plcH*, and *fecB* and virulence.

TYPE I SECRETION

Type I secretion systems (TISS) are found in a wide variety of Gram-negative bacteria and secrete proteins across both membranes in a single step, without involving either the Sec or Tat pathways (Figure 13.3). The substrates secreted by TISS include many toxins (e.g. HlyA from uropathogenic *E. coli* and MARTX from *V. cholerae*) nutrient acquisition proteins (e.g. the iron scavenger HasA from *Serratia marcescens*), and antimicrobial effectors (e.g. colicin V and other microcin antimicrobial peptides from *E. coli*). TISS are therefore important for virulence in many pathogens.

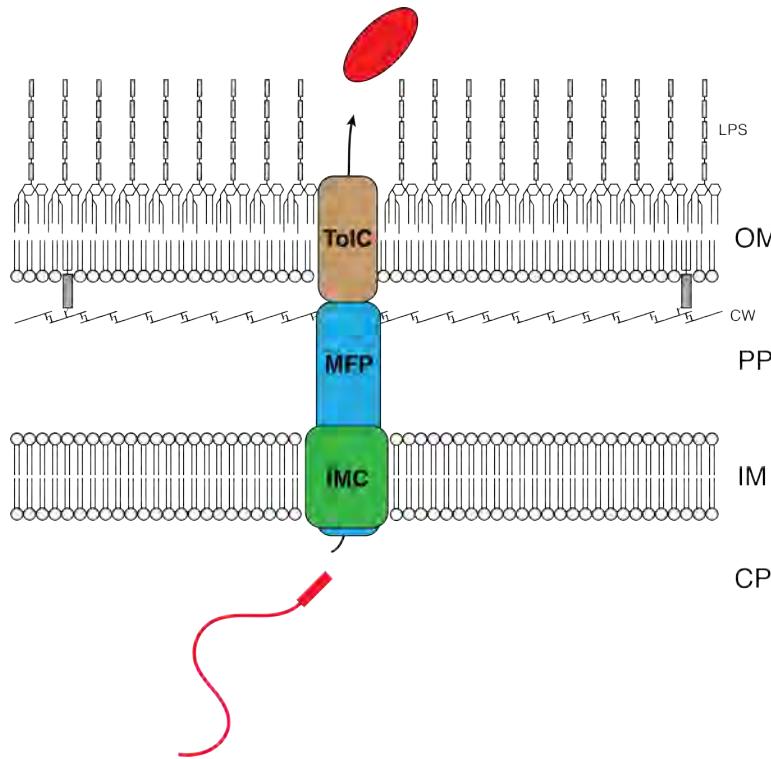


Figure 13.3. Type I secretion. TISS consist of three components: an inner membrane component (IMC) which is a member of the ABC transporter ATPase family, a membrane fusion protein (MFP) which creates a channel spanning the periplasm, and an outer membrane factor (usually TolC). TISS substrates are secreted in an unfolded state, and are typically recognized by a C-terminal GGXGXD repeat motif, which is not cleaved off of the secreted protein.

TISS consist of three subunits: an **inner membrane component** (IMC), a **periplasmic membrane fusion protein** (MFP), and an **outer membrane factor** (OMF). The IMC is an ATPase of the ABC transporter protein family (more about these in Lec-

ture 16), the MFP forms a channel that spans the width of the periplasm, and the OMF (usually a protein called TolC) forms a pore in the outer membrane. TISS substrates are recognized by the IMC and MFP by the presence of a repeated GGXGXDX motif in their C-terminus and are transported through the TISS complex in an unfolded state. Energy for this process is provided by the ATPase activity of the IMC.

TISS are homologous to RND efflux pump proteins, which use a related mechanism and the same TolC OMF to secrete small molecules (such as antibiotics) out of the cell.

TYPE II SECRETION

Most Gram-negative bacteria encode homologs of the type II secretion system (T2SS), which exports folded proteins from the periplasm into the extracellular environment (Figure 13.4). T2SS substrates must be secreted through the inner membrane by either the Sec or Tat pathways, and it is not currently known how the T2SS recognizes its substrates within the periplasm, although presumably there are structural features that act as targeting signals. T2SS substrates include a wide variety of toxins (e.g. cholera toxin, *Pseudomonas aeruginosa* exotoxin A) and enzymes (e.g. proteases, lipases, cellulases, nucleases, etc.), so T2SS are important for both virulence and adaptation to diverse environments.

The T2SS is a large protein complex, consisting of up to 15 different proteins. These are called Gsp (general secretion pathway) proteins in *E. coli*, which is the nomenclature I will use here, but in other species they are given other names, including Xcp in *Pseudomonas* and Eps in *Vibrio*. The outer membrane pore, or secretin, is composed of a dodecamer (12-mer) of the GspD protein. The lipoprotein pilin GspS recruits GspD to the outer membrane and assists in its proper assembly there. GspF, GspL, and GspM form a complex in the inner membrane which is tethered to the secretin by GspC. A cytoplasmic ATPase, GspE, associates with this complex and provides the energy necessary to drive secretion.

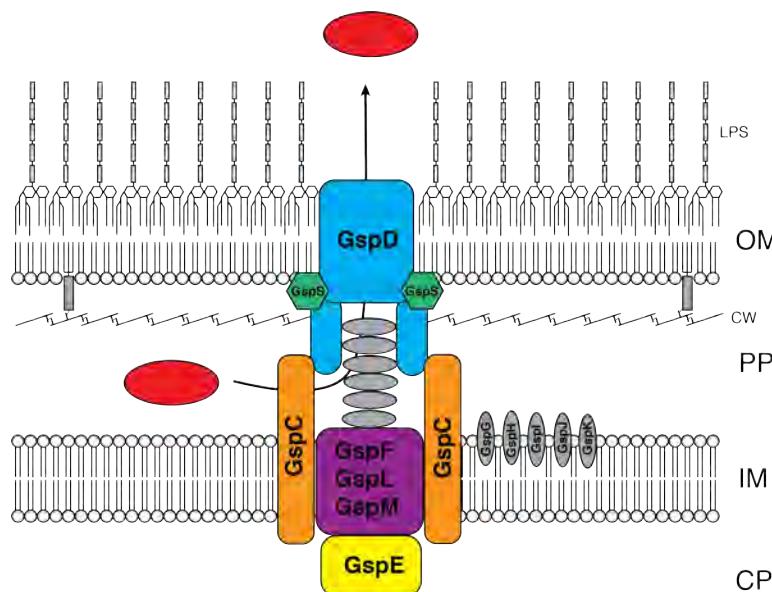


Figure 13.4. Type II secretion. Folded substrates in the periplasm are pushed out of the cell through the GspD secretin channel by polymerization of GspGHJK monomers into a pseudopilus. The energy for pseudopilus assembly and disassembly is provided by ATP hydrolysis by GspE.

Secretion by the T2SS is driven by the assembly and disassembly of a short polymeric *pseudopilus* within the channel formed by the GspCD proteins. The pseudopilus is composed of the major *pseudopilin* GspG and the minor pseudopilins GspHJK, and is thought to act as a piston that physically pushes substrate proteins out of the cell. T2SS are very closely related to type 4 pili (involved in adherence and twitching motility; [Lectures 14 and 15](#), respectively; see [this review](#) article for more details), but unlike a true *pilus*, the pseudopilus of T2SS never extends past the cell surface. In a handful of species (including some cyanobacteria) the T2SS and type 4 pili use the **same** secretion machinery to both secrete proteins and construct pili.

TYPE III SECRETION

Type III secretion systems (T3SS) have been extensively studied for more than 20 years, and are instrumental in the pathogenesis of many disease-causing Gram-negative bacteria. T3SS can deliver effector proteins directly into eukaryotic cells, and these effectors can have dramatic effects on the physiology of host cells. For example, in *Salmonella enterica*, two virulence-associated T3SS deliver at least 6 proteins into epithelial cells that remodel the host cell cytoskeleton and membrane, resulting in the uptake of *Salmonella* into non-phagocytic cells. Additional T3SS effector proteins orchestrate the formation of a *Salmonella*-containing vesicle (SCV), in which *Salmonella* replicates. Other pathogens, including *Shigella*, *Yersinia*, pathogenic *E. coli*, *Vibrio*, and *Pseudomonas* species also deliver multiple effectors and toxins into host cells with T3SS, and there is extensive evidence for the horizontal transfer of T3SS between species during the evolution of different pathogens.

The T3SS is a very large protein complex (often called a needle complex, because of its syringe-like structure and function), comprised of as many as 30 proteins, that spans both membranes of the Gram-negative cell envelope and extends well beyond the outer membrane (Figure 13.5). The proteins indicated in that figure use the *Salmonella* nomenclature, but homologous proteins in other species have often been given other names. Effector proteins are exported in a single step as unfolded proteins that pass through the channel in the center of the PrgI needle, driven by the ATPase activity of InvC.

It is not clear how the T3SS recognizes substrates for export, but most of them are associated with specific chaperone proteins that keep them unfolded in the cytoplasm. These chaperones associate with the SpaA, OrgA, and OrgB proteins, which are thought to be involved in substrate recognition. It is also not clear how the T3SS is able to detect that it has come in contact with a suitable host cell, although this presumably has something to do with conformational changes in the needle upon interaction with host cell surface receptors.

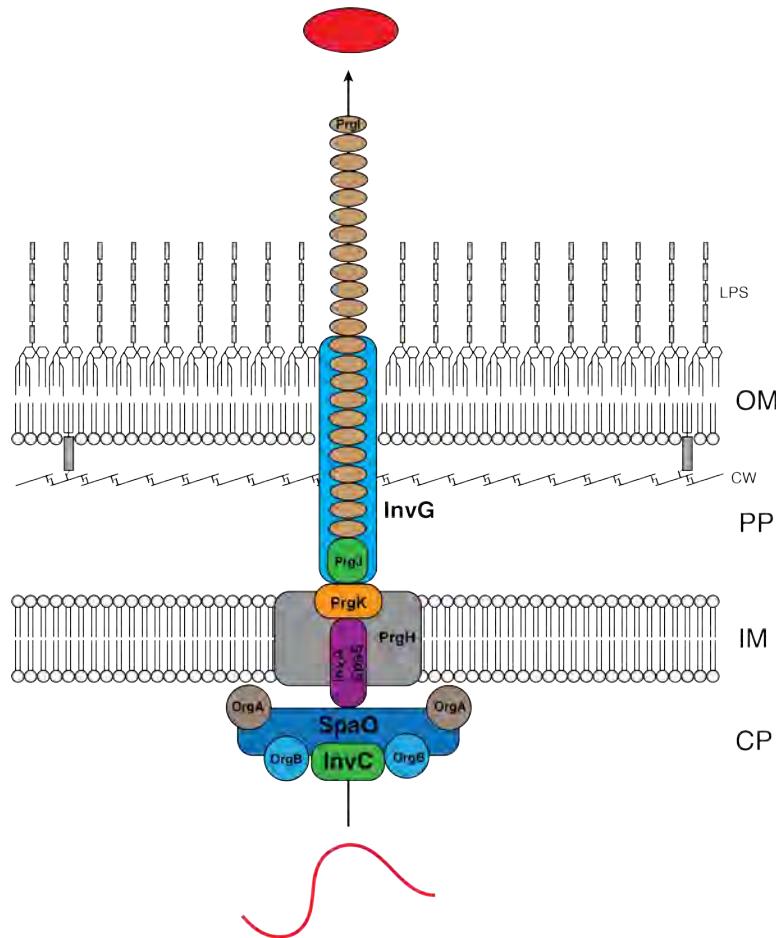


Figure 13.5. Type III secretion. Unfolded substrates are exported in a single step through the needle complex and directly into the cell membrane or cytoplasm of a eukaryotic cell. The energy for export comes from ATP hydrolysis by InvC.

Notably, the T3SS is very closely related to bacterial flagella (see **Lecture 15** for more on flagella and motility). The core proteins and structure of the T3SS share extensive homology with the flagellar basal body and PrgI is closely related to the flagellar hook protein. The T3SS lacks the motor proteins that rotate flagella, but the mechanism of export for PrgI and effector proteins is very similar to that for flagellin subunits. While Gram-positive bacteria can certainly have flagella, I am not aware of any examples of T3SS in a Gram-positive species.

TYPE IV SECRETION

Type IV secretion systems (T4SS) are widely distributed among Gram-negative and Gram-positive bacteria as well as among the archaea (Figure 13.6). They are involved in toxin and effector protein secretion in a few pathogenic bacteria (e.g. *Bordetella pertussis* and *Legionella pneumophila*). They are unique among the secretion systems discussed in this chapter in that they can mediate the translocation of DNA into adjacent cells (pro- or eukaryotic) as well as proteins. They also can carry out both secretion and import of substrates, unlike most of the systems we're discussing in this chapter. In fact, many naturally-occurring plasmids encode T4SS, and T4SS are the most common mechanism by which DNA export and import for bacterial conjugation occurs (**Lecture 8**), so they are also referred to as *competence pili*. Because of this, T4SS are important for understanding how antibiotic resistance genes spread, for example.

Remarkable work from Ankur Dalia's lab (linked [here](#)) has shown how T4SS pili in *Vibrio cholerae* are directly involved in DNA uptake from the media. They used fluorescent microscopy to observe the binding of the tips of pili to extracellular double-stranded DNA, then observed the pili being retracted, pulling the DNA into the cell, where it was expressed. David Bikard and Matxalen Llosa published a [paper in PNAS](#) this past year reporting a method for using T4SS to deliver large DNA-protein CRISPR complexes for gene editing ([Lecture 8](#)) into recipient bacteria.

The nomenclature for T4SS is mostly derived from that of the T4SS encoded by the Ti plasmid of the plant pathogen *Agrobacterium tumefaciens*, which is conjugated into host cells to deliver genes that lead to the formation of crown gall tumors. A large complex of proteins (called VirB1-11 and VirD4) spans both membranes (of Gram-negative cells), with VirB2 and VirB5 forming the long, filamentous pilus and its tip, respectively. VirB4, VirB11, and VirD4 are ATPases, and are presumably involved in the assembly and retraction of the pilus and the secretion of effector proteins and DNA.

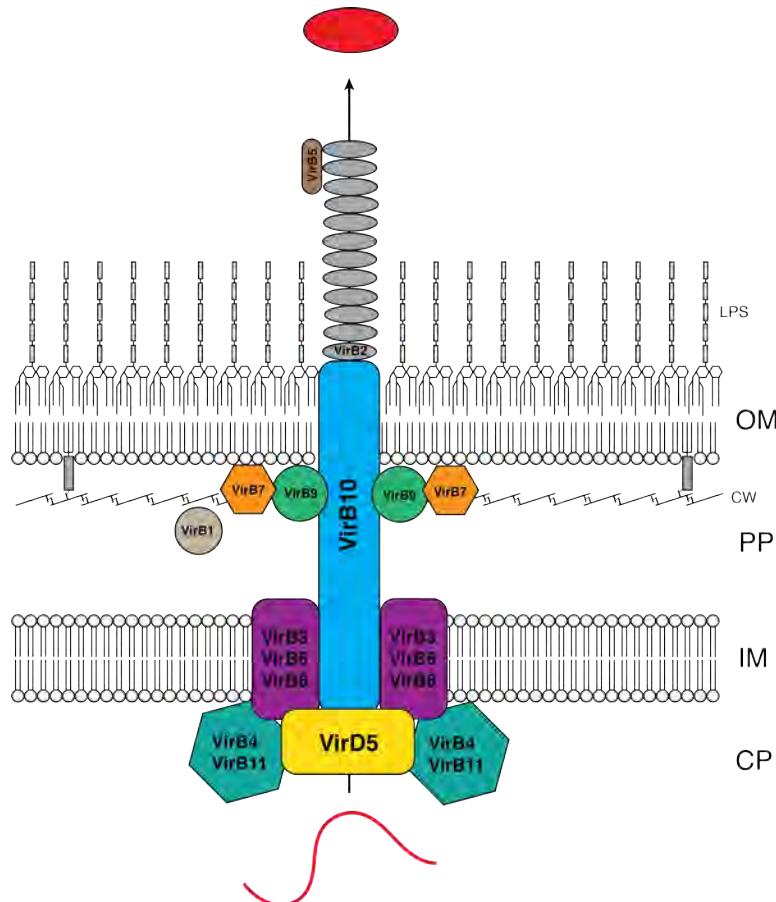


Figure 13.6. Type IV secretion. Protein or DNA substrates are secreted through the pilus, which is a dynamic structure that extends and retracts, driven by the ATPase activities of VirB4, VirB11, and VirD5.

Be very careful not to confuse the T4SS with type 4 pili, which are involved in adherence and twitching motility ([Lectures 14 and 15](#), respectively). As mentioned above, type 4 pili are related to the **T2SS**, not the T4SS. This is particularly tricky since both type 4 systems involve filamentous structures called pili, but they are otherwise quite different and are not evolutionarily related. There's really no excuse for this to be so confusing, but the two nomenclatures were established independently of one another and it's too late to disentangle them in the literature at this point.

For a recent review of T4SS functions, see [Costa et al. 2024](#).

TYPE V SECRETION

Compared to the complex machinery involved in T2, T3, and T4SS, type V secretion systems (T5SS) are refreshingly simple (Figure 13.7). Also called *autotransporters*, T5SS proteins secrete themselves across the outer membrane with no other specialized components necessary other than those required for assembly of all outer membrane proteins. Autotransporters have a β -barrel domain which inserts into the outer membrane, forming a channel through which the remainder of the protein (the *passenger domain*) is extruded into the extracellular space. Many T5SS proteins contain a protease domain which cleaves and releases the passenger domain, but some remain displayed on the cell surface. In some cases, the β -barrel domain left behind in the outer membrane after export of the passenger domain is known to have its own distinct function, for example as a porin.

Most of the well-studied T5SS substrates are virulence factors, and include a variety of toxins and other proteins. These include the VacA vacuolating cytotoxin from *Helicobacter pylori*, the IcsA actin-assembling protein from *Shigella flexneri*, the *Neisseria gonorrhoeae* immunoglobulin A protease, and the large autoaggregation protein from *E. coli* called antigen 43 (encoded by the *flu* gene, see **Lecture 14**). Since T5SS are dedicated to moving proteins across the outer membrane, they are found only in Gram-negative bacteria and mycobacteria.

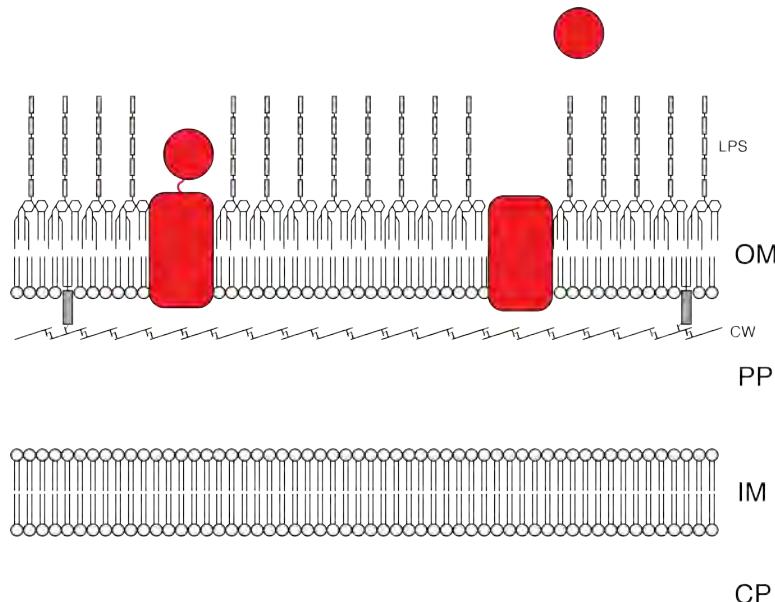


Figure 13.7. Type V secretion. The β -barrel domain of T5SS substrates assembles in the outer membrane (OM), and extrudes the passenger domain out of the cell. In some cases, the passenger domain is released by intrinsic protease activity of the T5SS protein.

T5SS proteins are always secreted across the inner membrane by the Sec pathway, and rely on the periplasmic chaperones and Bam complex which assemble integral outer membrane proteins. There are variants of T5SS which are slightly more complicated and involve two or sometimes three proteins (the two-partner and chaperone-usher secretion systems), but the core mechanism remains extremely similar.

TYPE VI SECRETION

Type VI secretion systems (T6SS) are another very large multiprotein complex, but they make up for it by being, for my money, the coolest of the protein secretion mechanisms (Figure 13.8). T6SS are evolutionarily related to the tails of *contractile phage*, and work by launching a protein-tipped harpoon that physically penetrates adjacent bacterial or eukaryotic cells.

Discovered in *Pseudomonas aeruginosa* in 2006 by Joseph Mougous, working in the lab of John Mekalanos, T6SS are widely conserved among Proteobacteria and deliver effector proteins that are typically toxins of one kind or another. Some bacteria use T6SS to deliver proteins into eukaryotic cells, including immune-modulating proteins from *Salmonella* and *Legionella*, the VgrG2b microtubule disruptor from *P. aeruginosa*, phagosome escape factors by *Francisella tularensis*, or anti-amoebe effectors in *Xanthomonas citri*, but most T6SS seem to be involved in competition between bacteria and therefore carry anti-bacterial toxins. The genetic loci encoding these kinds of T6SS typically also encode immunity proteins that prevent the T6SS-encoding cells from being killed by their own toxins, and T6SS often mediate a kind of self-recognition among very otherwise closely related bacterial strains.

Like the contractile phage tails they are related to, the T6SS sheath (composed mostly of the TssBC proteins) does not require ATP hydrolysis in order to contract, and this contraction appears to be triggered by the physical impact of the surface-exposed portion of the T6SS with another cell. Indeed, another common term for T6SS is **contact-dependent inhibition** (CDI). The TssBC tail sheath can be very long, extending across the entire width of a bacterial cell (a micron or more). In most species, after the projectile is launched, the baseplate and tail sheath are disassembled by the ClpV ATPase, so that the Tss protein monomers can be reassembled into new T6SS complexes, making T6SS very dynamic structures within the bacterial cell.

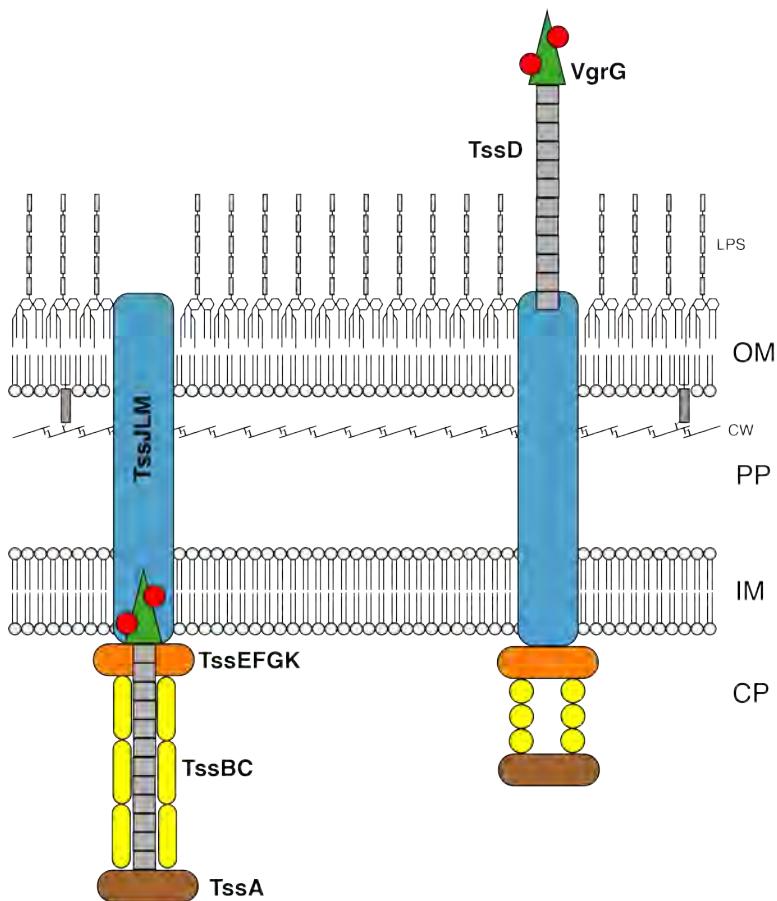


Figure 13.8. Type VI secretion. Effector proteins are bound to VgrG and are physically driven into target cells, along with the TssD shaft, by contraction of the TssBC sheath. Note that the VgrG / TssD protein harpoon does not remain attached to the original cell and is expelled into the medium (or target).

TYPE VII SECRETION (ESX)

Unlike the rest of the secretion systems in this chapter, type VII secretion systems (T7SS, also called ESX systems, which purportedly stands for “6-kDa early secretory antigenic target protein family secretion”, although don’t ask me how **that** acronym works) are only found in monoderm Gram-positive and mycobacteria (Figure 13.9). First identified in 2003 in *Mycobacterium bovis*, T7SS are best characterized in pathogenic mycobacteria, and are required for virulence of *M. tuberculosis*, for example, where the EsxG and EsxH secreted proteins are involved in damaging host phagosomes. However, it seems clear that T7SS are more versatile and can be involved in nutrient acquisition, conjugation, development (in *Streptomyces* spp.), and other general transport processes. The roles of the T7SS homologs found in more conventional Gram-positive bacteria (e.g. *B. subtilis*, *S. aureus*, or *Listeria monocytogenes*) have not been as well-studied to date, although [progress](#) is being made.

The core channel of T7SS through the inner membrane is formed from the EccBCDE proteins. EccA is a cytoplasmic ATPase which associates with the pore, and is probably responsible for providing the energy necessary for secretion. There is a cytoplasmic chaperone (EspG) which delivers substrates to the complex, and MycP is a membrane-associated protease involved in substrate processing. T7SS substrates are recognized by a relatively large N-terminal signal sequence consisting of a helix-turn-helix structure followed by a conserved YxxxD/E motif.

There are outer membrane proteins belonging to the PE-PPE protein family that may form conduits for T7SS substrates in the mycobacteria. Alternatively, or additionally, there is some evidence that the EspC or EsxEF proteins may form a filamentous pore that penetrates the thick outer membrane layers of the mycobacterium and provide a channel for secreted substrates to escape the cell (although this model remains unproven). An additional surface-associated protease called PecA is also thought to be involved in removing remaining signal sequences from T7SS-secreted proteins.

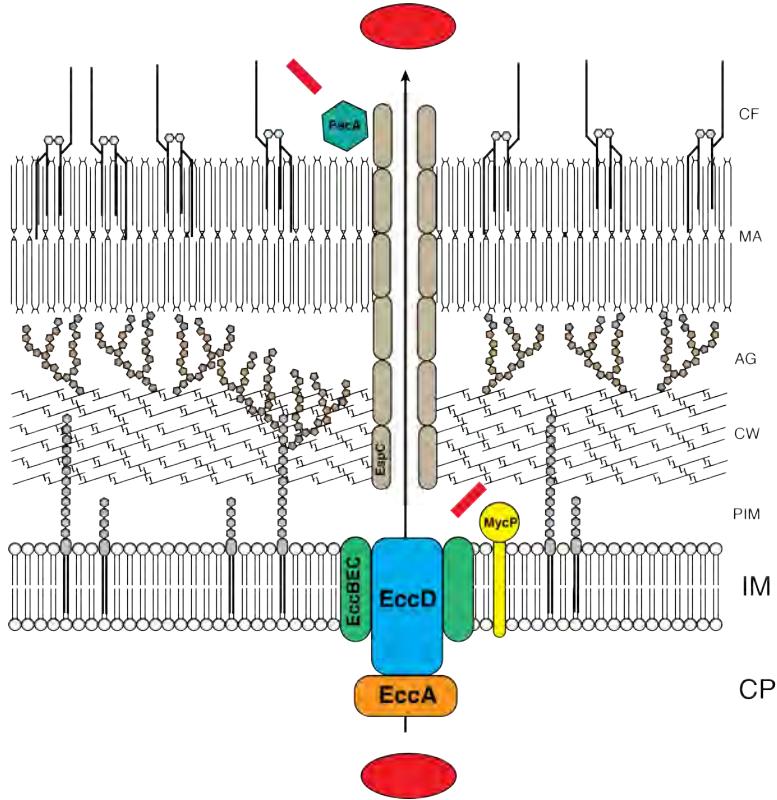


Figure 13.9. Type VII secretion. Substrate proteins are exported through the EccBECD channel in the inner membrane, powered by ATP hydrolysis by EccA. They may pass through the outer layers of the cell through a filamentous EspC / EsxEf pore, and signal sequences are removed from secreted proteins by the proteases MycP and PecA.

T7SS remain relatively poorly understood, and their mechanism, substrates, and function in bacterial physiology and virulence are very active areas of research. See [this review](#) for reasonably up-to-date information.

TYPE VIII SECRETION (CURLI)

Type VIII secretion systems (T8SS), best characterized in *Enterobacteriaceae* (e.g. *E. coli* and *Salmonella*), are specialized for the secretion and assembly of curli, which are proteinaceous fibers composed of the protein CsgA (Figure 13.10). Curli are attached to the outer surface of the cell and are unusual in that they are *functional amyloid proteins* that form a characteristic extremely stable cross- β -sheet secondary structure. Like other amyloids (including the amyloid- β and tau proteins involved in formation of neuronal plaques in human Alzheimer's disease), curli are extremely sticky and prone to aggregation, and are important for biofilm formation and adherence. *Salmonella* use curli to bind to the surfaces of eukaryotic cells, which contributes to the pathogenesis and spread of disease-causing strains. They can also use curli to bind to abiotic surfaces, including Teflon and stainless steel, which is problematic in food processing factories, for example. We will discuss biofilms in detail in [Lecture 14](#).

The T8SS consists of three proteins (called CsgEFG in *E. coli* and AgfEFG in *Salmonella*) that form a complex in the outer membrane. The curli monomer itself (CsgA or AgfA) is secreted into the periplasm by the Sec pathway, then exported across the outer membrane by the T8SS. Once outside the cell, the curli monomers rapidly aggregate and form curli fibers, which are normally anchored to the outer membrane by the minor curli monomer CsgB/AgfB, which also nucleates the aggregation of CsgA into fibers. A review of curli secretion and assembly can be found [here](#).

Curli-producing strains express a dedicated periplasmic chaperone called CsgC that prevents CsgA aggregation in the periplasm. Interestingly, work by Margery Evans (in Matthew Chapman's lab at the University of Michigan) has shown that *E. coli* CsgC can also prevent the amyloid aggregation of variants of human α -synuclein associated with neurodegeneration in Parkinson's disease, at least *in vitro*.

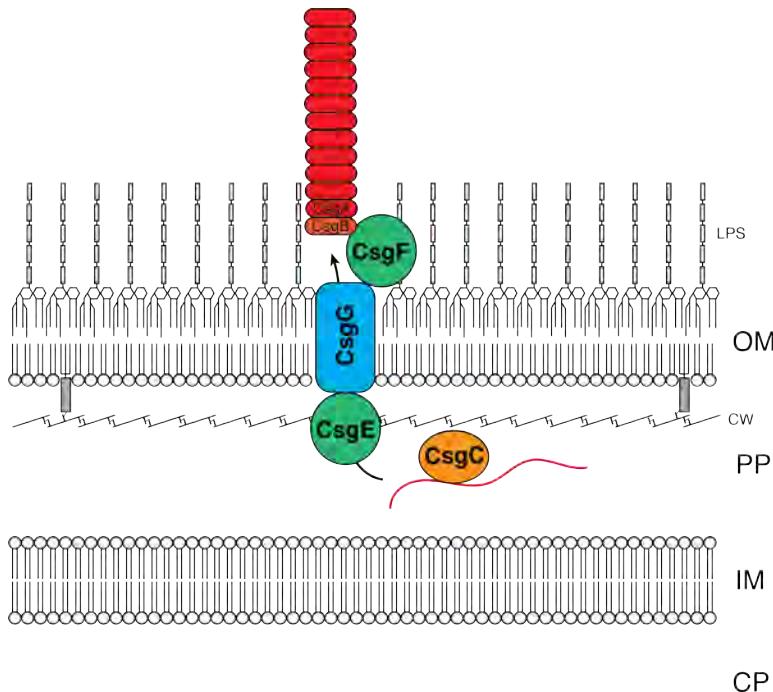


Figure 13.10. Type VIII secretion. Curli monomers (CsgA and CsgB) are secreted into the periplasm by the Sec pathway. While in the periplasm, they are maintained in an unfolded state by the chaperone CsgC. The CsgEG complex is required for secretion of curli, while CsgF is required for attachment of CsgB and CsgA to the outer membrane.

TYPE IX SECRETION

Type IX secretion systems (T9SS) are found only in certain species of the *Bacteroidetes* phylum, and are best characterized in the periodontal pathogen *Porphyromonas gingivalis* and the soil bacterium *Flavobacterium johnsoniae*. The terminology I will use here comes mostly from the *F. johnsoniae* model.

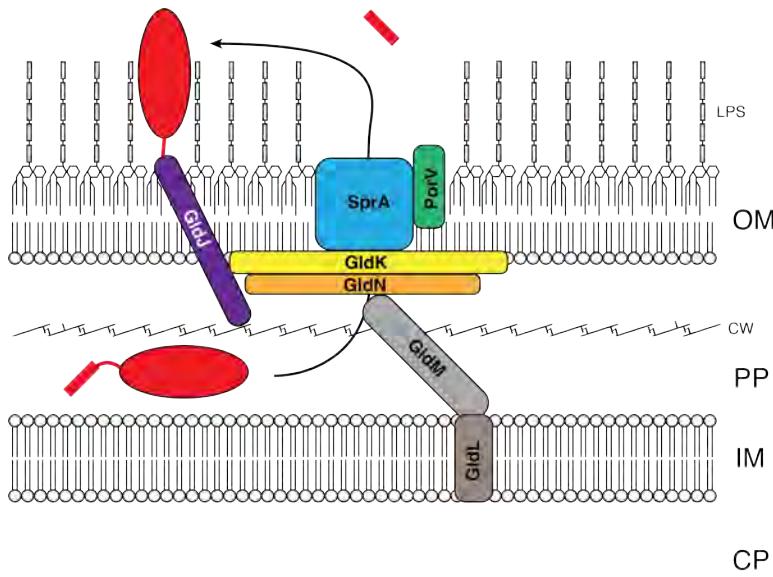


Figure 13.11. Type IX secretion. Substrates are translocated into the periplasm by the Sec system. After removal of the N-terminal Sec signal peptide, a C-terminal T9SS signal peptide directs the folded protein to the T9SS apparatus. After translocation across the outer membrane, the C-terminal signal sequence is cleaved off by the sortase PorU (not shown) and, for many substrates, the protein is covalently attached to LPS for anchoring to the cell surface. In this figure, the secreted protein is shown covalently bound to GldJ, which is the case for the SprB adhesin of *F. johnsoniae*. Other substrates may be secreted without being covalently attached to the cell surface.

The T9SS is a complex structure, consisting of at least 18 proteins, not all of which have known functions (and not all of which are shown in Figure 13.11). Key among these are the inner membrane ion channel protein GldL, the periplasmic proteins GldM, GldN, and GldK, and the outer membrane or surface-associated proteins SprA and PorV. Extensive recent struc-

tural characterization has made it clear that the T9SS is a large protein complex that spans the inner and outer membranes, with the outer membrane complex including at least 7 integral membrane β -barrel proteins.

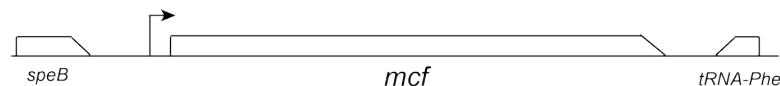
T9SS substrates are secreted to the periplasm by the Sec system, at which point they fold into their final structure and are recognized by the T9SS based on a C-terminal signal sequence and translocated across the outer membrane, at which point the C-terminal signal sequence is removed by the sortase-like protease PorU. Many T9SS substrates are then covalently modified by the attachment of an anionic LPS lipid group, which causes the resulting proteins to be anchored into the outer leaflet of the outer membrane. T9SS are also required for a remarkable mechanism of *gliding motility* in *F. johnsoniae*, which we will discuss in more detail in **Lecture 15**, and which requires the covalent binding of the T9SS-secreted adhesin SprB to the GldJ protein (as shown in Figure 13.11).

The energy necessary for type IX secretion comes from the proton motive force across the inner membrane, and is transduced across the periplasm to the outer membrane complex by GldM. The T9SS rotates as it works, and is one of only three known ion-driven rotary motors in biology, the others being the bacterial flagellum (**Lecture 15**), and the F₁F₀ ATP synthase (**Lecture 16**).

Type IX secretion is a very active subject of study right at the moment, and [this review](#) gives a reasonable summary of the state of the field.

DISCUSSION PROBLEM SET #23: TOXIN SECRETION BY PHOTORHABDUS LUMINESCENS

Photobacterium luminescens is a Gram-negative insect pathogen which secretes a large (324 kDa) soluble protein toxin into its growth medium that kills insect larvae and induces apoptosis in cultured insect cells. The *mcf* (makes caterpillars floppy) gene encoding this protein is monocistronic.



(*speB* encodes the enzyme agmatinase, which is not involved in protein secretion.)

The following methods are available for *P. luminescens*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (5.7 Mbp)	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposon	✓

Describe a series of observations and experiments that will allow you to identify the mechanism by which the *P. luminescens* toxin is secreted. State:

- observations you can make to narrow down the possible mechanisms
- experiments you will use to further narrow the possibilities
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiments, and how you will interpret them

Note that for the well characterized secretion systems described above, it is typically possible to identify homologs of the genes required bioinformatically from whole-genome sequences, although this cannot generally identify the secreted substrates of those systems, except in a few cases (Sec and Tat signal sequences are pretty obvious, for example).

GENETIC METHODS TO DETECT PROTEIN-PROTEIN INTERACTIONS IN VIVO

As we have seen, protein secretion often involves large protein complexes. It is certainly possible to directly measure the interactions of two or more purified proteins *in vitro*, but this is primarily a bacterial **genetics** class, and in this section I will introduce you to genetic methods that can be used to study the interactions of proteins *in vivo*. These commonly involve the construction of fusion proteins with different kinds of reporters, and depend on a general principle called *protein fragment complementation*.

Protein fragment complementation depends on the observation that some proteins can be divided into two or more amino acid chains that individually have no activity, but when brought into close physical proximity to one another are able to interact and form an active protein. The individual reporter fragments are expressed as fusions with the proteins of interest (typically called the bait and prey proteins), and any resulting reporter activity is interpreted as evidence that the bait and prey proteins physically interact (or are at least found close to one another at some point) in the cell.

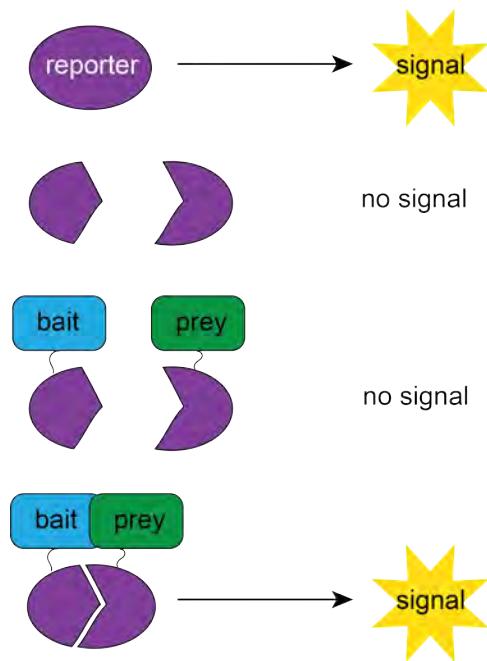


Figure 13.12. The basic principle of protein fragment complementation. The inactive reporter fragments form an active complex when brought into physical proximity, for example by being fused to two proteins that bind to one another. If the bait and prey proteins do not bind to one another, no signal will be produced.

This kind of assay was first developed in yeast by Stanley Fields and Ok-kyu Song (as reported in [this 1989 paper](#)). They referred to their system as **two-hybrid screening**, a term which is still commonly used for protein fragment complementation procedures. In fact, in many cases, the interactions between bacterial proteins have been screened by expressing those proteins in the yeast two-hybrid system. However, here I will discuss three protein fragment complementation systems used in bacteria. These methods include the bacterial two-hybrid screen (described in [this paper](#)), which uses adenylate cyclase as a reporter gene, the split murine dihydrofolate reductase (mDHFR) system (described [here](#)), and the use of tripartite split GFP (described [here](#)). I will also briefly describe a recently-developed system with fluorescent reporters that can be used to study the interaction of proteins in the cell envelope, which is difficult with the other systems.

The original and most commonly-used **bacterial two-hybrid screen** (BACTH; Figure 13.13) depends on the division of adenylate cyclase into two inactive fragments, called T18 and T25. Adenylate cyclase, encoded by the *cya* gene, is an enzyme which generates the second messenger cyclic AMP (cAMP). *E. coli* *cya* null mutants are unable to ferment lactose or maltose, because expression of the promoters for the *lac* and *mal* operons requires the transcriptional activator CAP (**catabolite activator protein**, also called CRP, which stands for **cyclic AMP receptor protein**, since it acts as a repressor at some promoters), and CAP is only active when bound to cAMP.

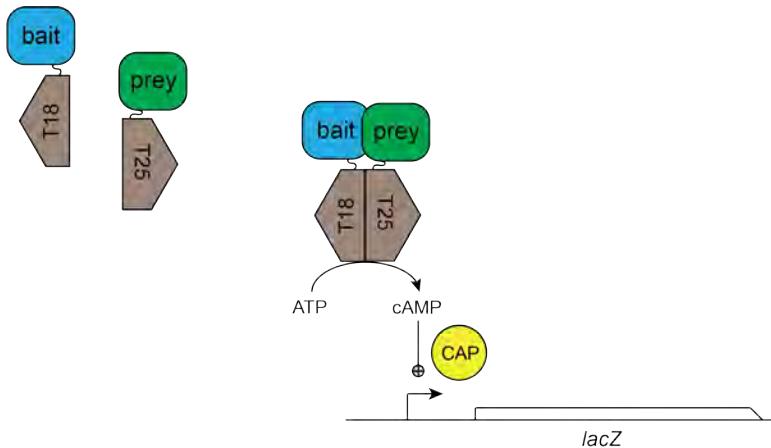


Figure 13.13. Bacterial two-hybrid screen, using split adenylate cyclase as a reporter, with (usually) a colorimetric *lacZ* expression readout.

The formation of active adenylate cyclase when the T18 and T25 domains are brought close to one another can be screened for on indicator plates (e.g. LB containing the colorimetric lactose analog X-gal, which turns blue when cleaved by LacZ) or selected for on media with maltose as a sole carbon source. BACTH is usually used as a colorimetric screen, since weak interactions may not result in production of enough cAMP to allow robust growth under selective conditions, but do result in the formation of visibly colored colonies on indicator plates.

The most common application of BACTH is to look at a single pairwise protein-protein interaction, for which the bait and prey proteins are cloned individually as T18 and T25 fusions (using any of the methods from **Lecture 7**). Some very large screens of *interactomes* (the set of protein-protein interactions in a cell) have been performed this way, but this is very labor-intensive, requiring individually cloning thousands of genes. If you are interested in identifying new interaction partners of a particular protein in a more elegant way, you can clone the bait protein as one fusion fragment and clone a library of random fragments of genomic DNA into the plasmid for the other fragment. Most of the resulting clones will, of course, encode protein fragments that are out of frame with the fusion, but if you generate a large enough library you can screen for proteins that have physiologically-interesting interactions with the bait protein.

Split mDHFR (Figure 13.14) is a system that allows more sensitive protein fragment complementation **selections** than BACTH. DHFR is an enzyme in the synthesis pathway for the essential cofactor folate, without which *E. coli* cannot grow. Bacterial DHFR is inhibited by the antibiotic trimethoprim, but mammalian DHFR (including mDHFR from mice) is not. Therefore, when the inactive fragments of mDHFR (called mDHFR[1] and mDHFR[2]) are brought together in *E. coli* to form active mDHFR, the resulting protein confers trimethoprim resistance, which is easily selected for. Stronger interactions result in resistance to higher concentrations of trimethoprim.

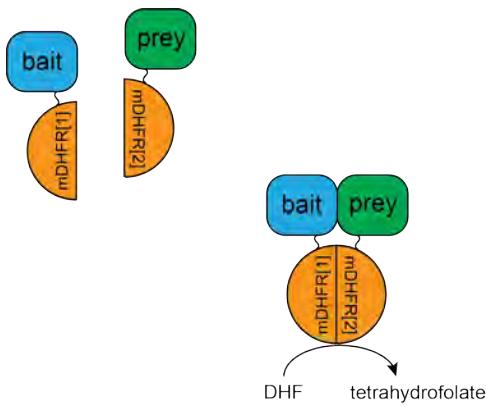


Figure 13.14. Split murine dihydrofolate reductase reporter, with trimethoprim-resistant tetrahydrofolate production as an output.

For many screening applications, especially those that use microscopy or flow cytometry, fluorescence is a sensitive and easily detected phenotype. Tripartite split GFP (Figure 13.15) can be used in these cases. In this assay, individual strands of the GFP β-barrel (called GFP10 and GFP11) are fused to the bait and prey proteins with short, flexible linkers and expressed in the same cell as the remainder of GFP (GFP1-9). None of these individual fragments of GFP are fluorescent, but when GFP10 and GFP11 are brought close together, they associate into a structure that can assemble with GFP1-9 into complete, active GFP. This method also has the advantage that the GFP10 and GFP11 fragments are quite small (20 and 21 amino acids, respectively), and therefore are less likely to interfere with the structure or interactions of the bait and prey proteins than larger fusion proteins.

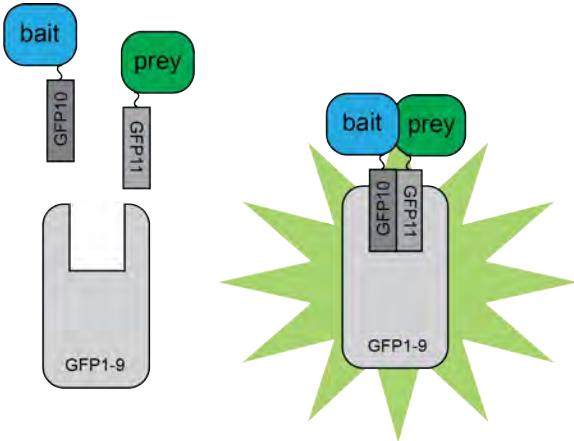


Figure 13.15. Split GFP reporter, with fluorescence signal as an output.

Note that split GFP has the distinct advantage over BACTH and split mDHFR in that the screening will work in any cell, not just in specifically engineered strains of *E. coli*. It does require expression of at least 3 different protein fragments, though, and it is common to also engineer the strains used for screening to also constitutively express a red fluorescent protein as a normalization control, so constructing appropriate reporter strains is more complicated than for BACTH or mDHFR.

One limitation of the above methods, especially using adenylate cyclase or mDHFR, is that the reporter proteins must be in the cytoplasm. Many protein complexes, as we have seen in this chapter, assemble in the membranes and periplasm. A fairly recent publication from Thomas Bernhardt's lab ([linked here](#)) describes POLAR (PopZ-linked apical recruitment; Figure 13.16) a clever approach for studying protein interactions in the cell envelope which I suspect will be extremely useful, but has not yet been widely used.

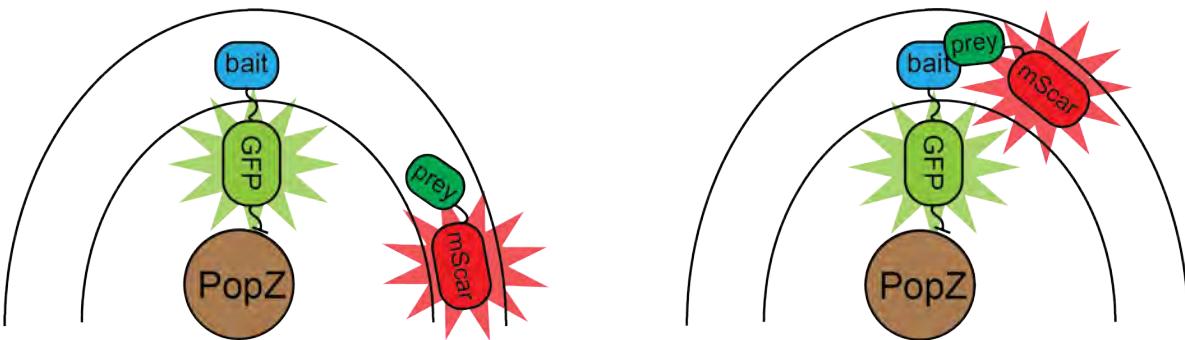


Figure 13.16. PopZ-linked apical recruitment assay, with colocalization of fluorescent GFP and mScarlet at the cell pole as an output signal. In this example, the bait protein is an inner membrane protein and the prey is periplasmic, with the interaction occurring in the periplasm.

POLAR takes advantage of the *Caulobacter crescentus* PopZ protein, which, when expressed in *E. coli*, localizes to the poles of the bacterial cell (by a mechanism that is not fully understood). Bait proteins are expressed in PopZ-expressing *E. coli* as fusions with a PopZ-binding tag and GFP. This results in the recruitment of the bait protein, which can be either a cytosolic or an inner membrane protein, to the poles of the cell, where it can be visualized by fluorescence microscopy. The prey protein is then expressed in the same cells fused to the red fluorescent protein mScarlet. The prey can be present in the cytoplasm, the inner membrane, the periplasm (as shown in Figure 13.16), or even the outer membrane. If the bait and prey proteins interact, the GFP and mScarlet fluorescence will colocalize at the cell poles. Otherwise, the mScarlet signal will be distributed throughout the cell, and this can be observed microscopically.

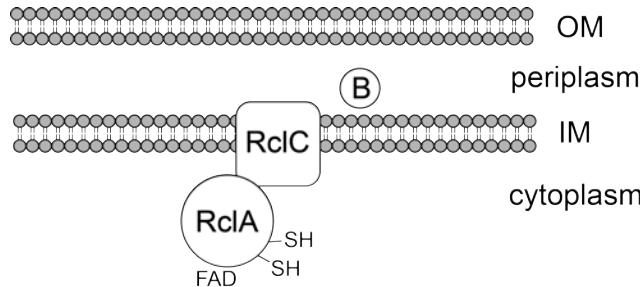
Note that most of the above methods test for protein-protein interactions *in E. coli*, even when examining proteins from other organisms. This is not typically a problem, and is sometimes a benefit, both because *E. coli* plasmids are straightforward to engineer and because there will be less interference from interactions with other proteins in the host. However, if a particular protein of interest is poorly expressed or toxic in *E. coli*, this may present a challenge.

It is also important to remember that bacteria are **much** smaller than eukaryotic cells (on the order of 1 μm in width). This makes fluorescence colocalization experiments significantly trickier than you may be used to thinking about if you have previously worked in eukaryotic systems. POLAR works because it is typically obvious whether a fluorescent protein is at the poles of a bacterial cell or not, but finer-scale localization is substantially more difficult. As noted in **Lecture 11**, this can be addressed to some extent by using larger species (*B. subtilis* is about twice the size of *E. coli*) or, more effectively but more expensively, with super resolution microscopy.

DISCUSSION PROBLEM SET #24: PROTEIN COMPLEX FORMATION IN *ESCHERICHIA COLI*

The *rclABC* genes of *E. coli* are conserved among enterobacteria (like *Salmonella* and *Shigella*) that thrive in the gut during inflammatory diseases. This operon is strongly upregulated when *E. coli* is exposed to antimicrobial oxidants produced by neutrophils, and mutants lacking these genes are sensitive to killing by those oxidants.

RclA is a cytoplasmic enzyme with conserved cysteine residues and an FAD cofactor-binding site, RclB is a small periplasmic protein with a typical Sec signal sequence, and RclC is predicted to be an integral inner membrane protein. You hypothesize, based on this, that they might form a complex at the inner membrane.



All genetic and molecular biology methods are available for *E. coli*.

Design an experiment or series of experiments to determine whether RclA, RclB, and RclC interact with each other *in vivo*. State:

- the independent and dependent variables of each experiment
 - both positive and negative controls for each experiment
 - a description of how you will construct any necessary strains and plasmids
 - whether your proposed experiment tests correlation or causation
 - potential outcomes of your experiment(s), and how you will interpret them
-

Finally, in some cases, protein interactions can also be studied more directly using genetics.

Protein-protein interactions are determined by the surface properties of proteins, with interacting surfaces having complementary shapes and chemical properties. For example, there may be a positively charged amino acid in one protein that interacts with a negatively charged amino acid on the surface of its interaction partner, or two proteins may have complementary patches of surface-exposed hydrophobic amino acids. Missense mutations of key amino acids forming the interaction surfaces can disrupt these interactions.

If missense mutations in one protein are found that prevent complex formation, it can often be interesting to look for suppressors of those mutations in the **other** protein (recall the discussion of intergenic suppressors in **Lecture 3**). For example, if mutating a positively-charged Arg residue to Glu in one protein prevents interaction, perhaps there is a negatively-charged amino acid in the other protein that can be mutated into a positively-charged residue to restore the interaction. This would be strong evidence that the charge-charge interaction at that site is important for the interaction of the two proteins involved. These kinds of mutants could be made by site-directed mutagenesis, if there was structural or other evidence to implicate specific amino acids in an interaction, or by random mutagenesis, if there is a screenable or selectable phenotype for the interaction between the proteins in question.

LECTURE 14: CAPSULE AND BIOFILMS

INTRODUCTION

Most, if not all, bacteria are able to synthesize and secrete complex polysaccharide carbohydrate polymers. This includes the GlcNAc-MurNAc polymer that forms the backbone of peptidoglycan ([Lecture 10](#)), but also a very wide range of other extracellular polysaccharides (EPS) with diverse functional roles. Capsules are EPS that (usually) remain attached to the surface of the bacterial cell, while slimes and gums are EPS that are released into the extracellular environment. EPS are also a key component of bacterial biofilms, which are multicellular bacterial communities attached to surfaces.

In this chapter, we will discuss the synthesis of EPS, the importance of capsule in pathogenesis, how bacteria adhere to surfaces, and the structure, formation, and regulation of biofilms. I will note that the topics of this chapter are the subjects of research by many labs here at UAB, including those of Drs. Moon Nahm, Janet Yother, Carlos Orihuela, Jessica Scofield, Ed Swords, and Megan Kiedrowski, all of whom are, of course, far more expert than I am in this area.

Thanks in particular to Drs. Janet Yother and Megan Kiedrowski for their suggestions and comments on this chapter, which helped improve it immensely and saved me from some embarrassing errors! Any remaining mistakes are my fault.

POLYSACCHARIDE BASICS

Strictly speaking, carbohydrates are any organic compounds composed only of carbon, hydrogen, and oxygen, but for biological purposes, the term is mostly synonymous with sugars, which are 3- to 7-carbon carbohydrates distinguished by the arrangement and chirality of their various hydroxyl groups, and polymers of those sugars. Simple sugars, or monosaccharides, are the monomers that make up polysaccharides, including EPS.

The most common sugars in biological polysaccharides are hexoses (6 carbons) and pentoses (5 carbons), which are typically found as rings in biological systems. Some examples are shown in Figure 14.1. The *anomeric carbon* of a monosaccharide, always adjacent to the oxygen atom, is numbered 1. The other carbons in the ring are numbered consecutively from the anomeric carbon. In an α sugar, the substituent groups on the anomeric carbon and the carbon on the other side of the oxygen atom in the sugar ring are facing **opposite** directions, while in a β sugar, they are facing the **same** direction. The substituent group of the anomeric carbon is the one whose orientation can change without changing the name of the sugar (i.e. both α - and β -glucose are still glucose).

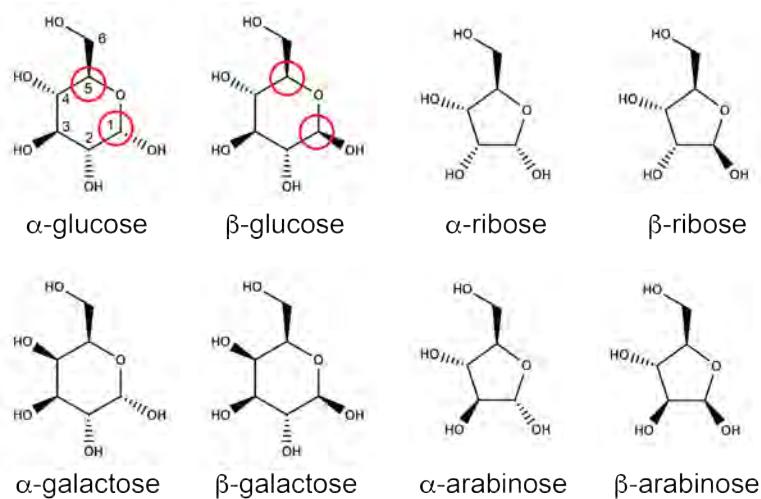


Figure 14.1. A few common hexoses and pentoses found in bacterial polysaccharides, with each shown in its α and β form. Carbon numbering starting at the anomeric carbon is illustrated for α -glucose, and the carbons whose substituents determine the difference between α - and β -glucose are circled in red.

In polymers, monosaccharides are linked by *glycosidic bonds*, which are carbon-oxygen-carbon bonds connecting two sugars. When the two linked carbons have the same relative stereochemistry, that is called an α -glycosidic bond. When they have opposite stereochemistries, it is a β -glycosidic bond. The nomenclature for describing specific glycosidic bonds indicates the stereochemistry of the bond as well as the carbons in the two sugars being linked. The disaccharide lactose, for example, contains a β -1,4 linkage between the hexoses galactose and glucose (Figure 14.2).

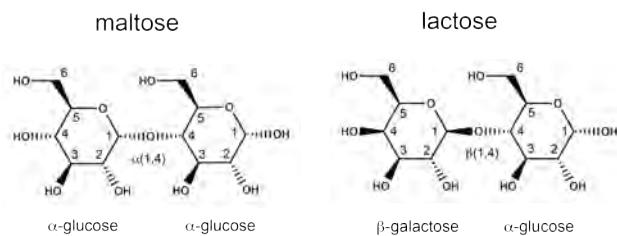


Figure 14.2. Examples of disaccharides containing different kinds of glycosidic bonds. Maltose has an $\alpha(1,4)$ bond, lactose has a $\beta(1,4)$ bond.

In many biologically important polysaccharides, the monosaccharide subunits are chemically modified. Common modifications are the replacement of a hydroxyl group with an amino group to form an *amino sugar* and *acetylation*, which is the addition of a $\text{CH}_3\text{CO}-$ acetyl group to either a hydroxyl group (*O-acetylation*) or an amino group (*N-acetylation*) (Figure 14.3). These and other modifications can have dramatic effects on the chemical properties of a sugar or polysaccharide.

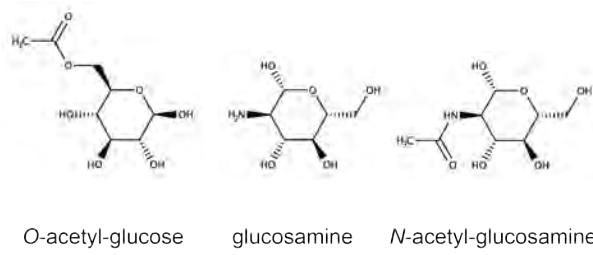


Figure 14.3. Common sugar modifications found in bacterial polysaccharides, illustrated on a glucose monosaccharide.

There is a nearly unlimited amount of diversity possible in polysaccharide biochemistry. Any number of different monosaccharides, with different modifications, can be linked by different kinds of glycosidic bonds, in linear or branched forms, and with widely varying lengths. This imparts tremendous functional and chemical diversity, but also makes the study and categorization of polysaccharides a complicated topic.

Luckily, most polysaccharides are constructed from repeating subunits, so we can categorize them at least partially by reducing them to their minimal repeat unit (Figure 14.4). The simplest polysaccharides have repeat units of a single monosaccharide: cellulose is linear polymer of β -1,4-linked glucose, and amylose is composed of α -1,4-linked glucose. This seemingly minor difference results in polymers with dramatically different chemical properties. The repeat unit of the polysaccharide component of peptidoglycan is a disaccharide: β -1,4-linked N-acetyl-glucosamine and N-acetyl-muramic acid (**Lecture 10**). However, many EPS have more complicated, often branching repeat units.

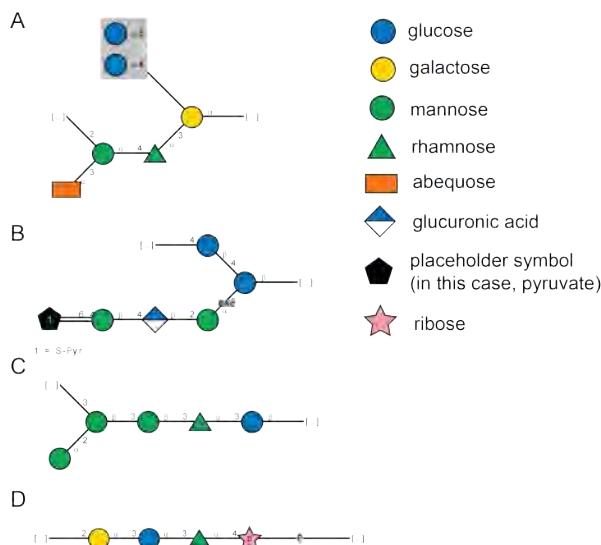


Figure 14.4. Examples of the polymeric repeat units of some bacterial polysaccharides. (A) *Salmonella enterica* O-antigen, with an example of a sugar that can have either of two possible glycosidic linkages. (B) Xanthan gum from *Xanthomonas campestris*, with an example of an acetylated sugar. (C) The *Pseudomonas aeruginosa* Psl secreted polysaccharide. (D) A *Streptococcus pneumoniae* type 6B capsular antigen. Schematics obtained from the Carbohydrate Structure Database.

The [Carbohydrate Structure Database](#) is a resource hosted by the Zelinsky Institute of Organic Chemistry in Moscow that stores all of the known biological carbohydrate structures, although it suffers from a somewhat outdated design and user interface.

interface and (in my experience) frequent downtime. Nevertheless, it does provide a comprehensive system for visualizing and reporting polysaccharide structures. Some examples drawn from that database are shown in Figure 14.4.

The function of biological polysaccharides is determined not only by their structure, of course, but also by their localization. Obviously, peptidoglycan is an integral component of the cell wall, and LPS and O-antigen are major constituents of the outer leaflet of the Gram-negative outer membrane ([Lecture 10](#)). EPS are more peripherally associated with the cell. EPS that are physically attached to the surface of the bacteria cell are generally referred to as capsule. In Gram-positive bacteria, capsules are often covalently bound to peptidoglycan. Surprisingly, in Gram-negative bacteria, the mechanism(s) by which capsules are attached to the cell surface are not currently well understood. EPS that are released from the cell into the supernatant are sometimes called either *slime* or *gum*, depending on whether they're generally slippery or sticky. (The food-safe thickening agent xanthan gum is a secreted EPS from *Xanthomonas campestris*, for example.)

Note that, because EPS form the outermost layers of bacterial cells, the mammalian immune system is very good at recognizing and differentiating between subtly different polysaccharides. Serotyping is a method of distinguishing bacteria by what antibodies recognize them, and the antigens involved are very often polysaccharides. Many bacterial EPS have names that reflect this (e.g. the O- and K-antigens of *E. coli*, Vi-antigen of *Salmonella*, etc.).

There are four main pathways by which polysaccharides are synthesized in bacteria. We will discuss two of them in some detail, and mention the others in passing before moving on to exploring the functions of EPS in host-microbe interactions and the formation of surface-attached bacterial communities.

WZY-DEPENDENT EPS SYNTHESIS PATHWAY

The most prevalent type of EPS synthesis pathway depends on a class of enzymes called Wzy polymerases. This includes the pathways for the synthesis of most enterobacterial O-antigens, the Psl polysaccharide from *Pseudomonas aeruginosa*, and most *Streptococcus pneumoniae* capsule types, among many, many others. It's also very closely related to the pathway by which peptidoglycan is synthesized, as we will see shortly.

In this pathway, polysaccharide repeat units are assembled in the cytoplasm on a phospholipid carrier molecule called undecaprenyl phosphate (Und-P) (Figure 14.5), which is inserted into the cytoplasmic face of the inner membrane.

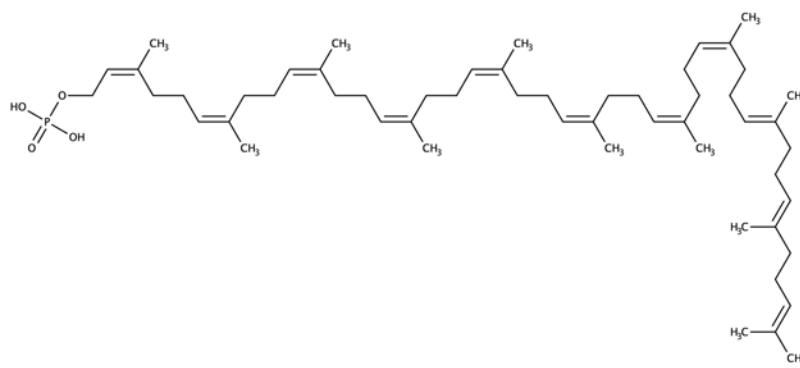


Figure 14.5. The structure of undecaprenyl phosphate (Und-P), the lipid carrier on which Wzy-dependent polysaccharides are assembled.

The Wzy polymerase dependent EPS synthesis pathway is illustrated schematically in Figure 14.6. The Wz* nomenclature used here indicates protein homolog families. Each separate polysaccharide synthesis pathway has its own dedicated equivalent of each enzyme.

The monosaccharide subunits for polysaccharide assembly are in the form of nucleotide diphosphate (NDP) sugars, and assembly of the repeat unit begins with a membrane-bound polyprenol phosphate phosphoglycosyl transferase that takes Und-P and a specific NDP-sugar to generate an Und-PP-monosaccharide molecule, releasing nucleotide monophosphate (NMP). Subsequent sugars are added to the repeat unit by sequentially-acting cytoplasmic glycosyltransferase enzymes, until the complete repeat unit is assembled.

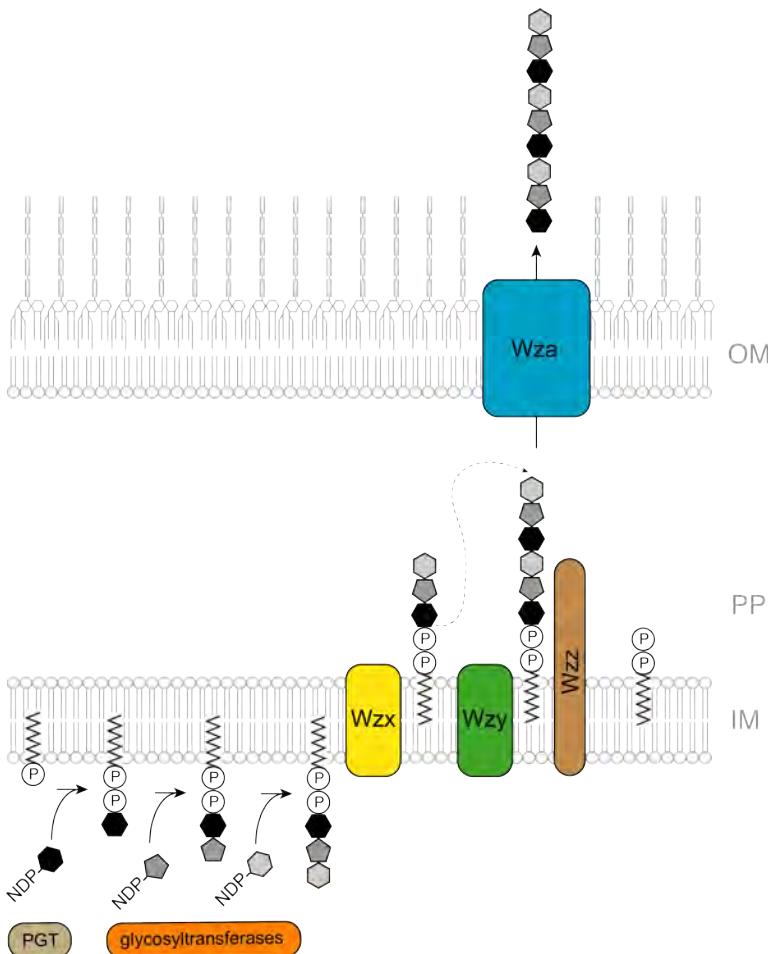


Figure 14.6. Schematic of Wzy-dependent EPS biosynthesis in a Gram-negative bacterium. EPS biosynthesis is initiated on an undecaprenyl phosphate (Und-P) carrier by a PGT (polyprenyl phosphate phosphoglycosyltransferase) and repeat units are synthesized by sequential addition of monosaccharides by glycosyltransferases. Completed Und-PP-linked repeat units are flipped across the membrane by a Wzx flippease and added to growing polymer chains by a Wzy polymerase. Polymer length is often regulated by Wzz. In Gram-negative bacteria, released EPS polymers are exported through the outer membrane via the transporter Wza.

Next, the completed repeat units, still anchored into the membrane by the lipid carrier, are moved through the membrane to the **outer** leaflet by a *Wzx flippease*. At that point, the Und-PP-repeat unit is the substrate for a *Wzy polymerase*, which assembles chains of repeat units. The polysaccharide chain is constructed while still anchored to an Und-PP lipid carrier, and as each new repeat unit is added, the attached Und-PP is released. In many pathways, the length of the polysaccharide product, by which I mean the number of repeat units, is controlled by a *Wzz* protein.

In Gram-negative bacteria, completed EPS must normally be exported through the outer membrane, which is accomplished by a protein complex whose central member is called *Wza*. There are at least some Wzy-dependent EPS that are retained in the periplasm (e.g. the cyclic form of [enterobacterial common antigen](#)). In Gram-positive bacteria, assembled EPS may be released from the cell completely or may be covalently attached to the peptidoglycan cell wall by *Wzd*/*Wze* proteins.

There are additional proteins that regulate EPS synthesis and export, as well as dedicated enzymes that catalyze chemical modifications of NDP-sugar precursors, but these differ from pathway to pathway.

The O-antigen of LPS ([Lecture 10](#)) is synthesized by a Wzy-dependent mechanism in most Gram-negative bacteria, and is covalently attached to the LPS core in the periplasm by an enzyme called *WaaL* before the complete LPS is flipped to the outer leaflet of the outer membrane. It is worth noting here that K-12 strains of *E. coli*, which includes most laboratory strains, do not synthesize O-antigen, due to a mutation (*rfb-50*) in the O-antigen synthetic pathway. Among other things, this makes K-12 considerably more sensitive to hydrophobic antibiotics than other strains that **do** synthesize O-antigen.

POLYSACCHARIDE GENE NOMENCLATURE

I want to take a moment here to try to explain the very confusing state of gene nomenclature in the bacterial polysaccharide world. The fundamental problem is that, for each different polysaccharide, there is usually a different, specific set of *Wzx*, *Wzy*, and *Wzz* proteins, each individual bond between two monosaccharides has to be formed by its own specific glycosyl-

transferase, and individual monosaccharides may be modified by sugar-specific acetyltransferases, aminotransferases, or other enzymes. This results in a vast number of genes and proteins, all of which need names.

As an example, assembly of the *E. coli* enterobacterial common antigen (ECA) polysaccharide, which has a relatively simple linear three-sugar repeat unit, requires three glycosyltransferases (WecA, WecG, and WecF), six sugar-modifying enzymes (WecB, WecC, WecD, WecE, RffH, and RffG), and dedicated WzxE flippase, WzyE polymerase, and WzzE length determination proteins. For ECA, at least the names of the flippase, polymerase, and length determination protein contain “Wzx”, “Wzy”, and “Wzz”, but this is by no means universal. Many polysaccharide synthesis gene names do follow the *w**** format (proposed in [this paper](#) in 1996), but others may use *cps*, *cap*, *eps*, *rbf*, or a variety of other gene symbols, and there's very little in the way of standardization from one species to another.

On the other hand, for the most part, all of the genes necessary for synthesis of a particular polysaccharide are **usually** encoded in the same operon or genetic locus in bacterial genomes, and it is possible to examine the sequence of those genes and get a pretty good idea of how many glycosyltransferases, etc. are involved in any particular pathway, which is helpful.

PEPTIDOGLYCAN SYNTHESIS PATHWAY

The main difference between the peptidoglycan synthesis pathway and other Wzy-dependent polysaccharide synthesis pathways is that the pentapeptide cross-linking stem (**Lecture 10**) is assembled on UDP-MurNAc by the MurA-F enzymes before the phosphoglycosyltransferase MraY attaches that modified sugar to Und-P, forming an intermediate called *Lipid I*.

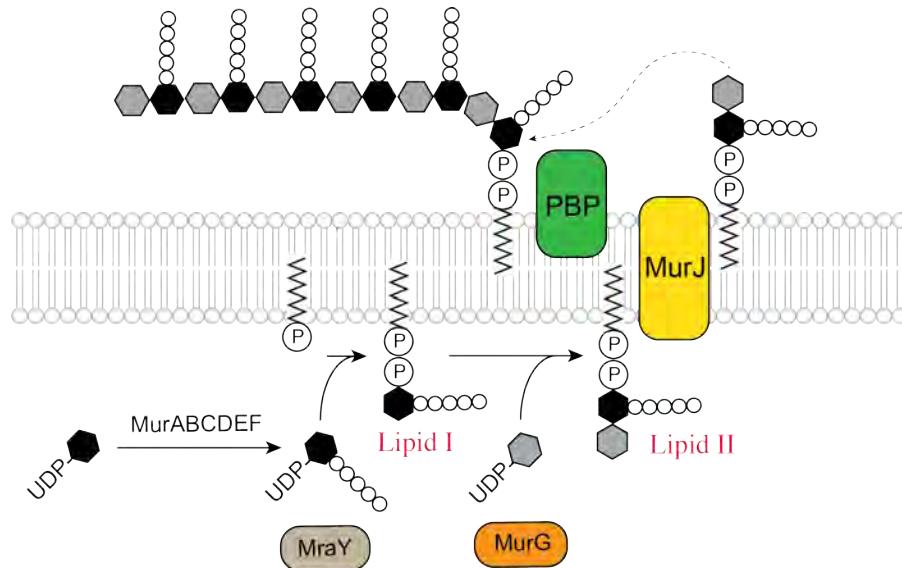


Figure 14.7. Schematic of peptidoglycan polysaccharide backbone synthesis. UDP-MurNAc (black hexagons) is synthesized and the pentapeptide stem (white circles) added by the MurA-F enzymes. The phosphoglycosyltransferase MraY combines this with Und-P to generate Lipid I. The glycosyltransferase MurG adds GlcNAc (grey hexagons), generating Lipid II, which is flipped across the membrane by the flippase MurJ. Lipid II is added to growing peptidoglycan chains by the glycosyltransferase activity of PBPs (penicillin-binding proteins).

To form the repeat unit of peptidoglycan, the glycosyltransferase MurG adds UDP-GlcNAc to Lipid I, forming the intermediate *Lipid II*, which is flipped across the membrane by MurJ. The assembly of peptidoglycan chains is then catalyzed by the glycosyltransferase activity of penicillin-binding proteins (PBPs), releasing Und-PP.

Peptidoglycan synthesis is the major drain on the cell's fairly limited pool of Und-P, and the recycling of Und-PP after the repeat unit has been transferred to the growing polysaccharide chain is essential for continued bacterial growth and survival, as well as to supply Und-P for other polysaccharide synthesis pathways. *E. coli* has three partially redundant Und-PP phosphatases, BacA, YbjG and PgpB. It is currently unknown how the Und-P generated by these enzymes returns to the inner leaflet of the inner membrane, but [this 2022 Nature paper](#) identified a DedA domain protein in *V. cholerae* and *S. aureus* that may be involved, and [subsequent work](#) has shown that these putative DedA flippases are conserved in all clades of bacteria that produce peptidoglycan.

SYNTHASE-DEPENDENT EPS SYNTHESIS PATHWAY

In contrast to the Wzy polymerase pathways that assemble complex polysaccharides attached to a lipid carrier molecule, some polysaccharides are directly assembled as linear glycans by synthase-translocase catalytic enzyme complexes. These are called synthase-dependent systems, and the basic components of such a system are illustrated in Figure 14.8.

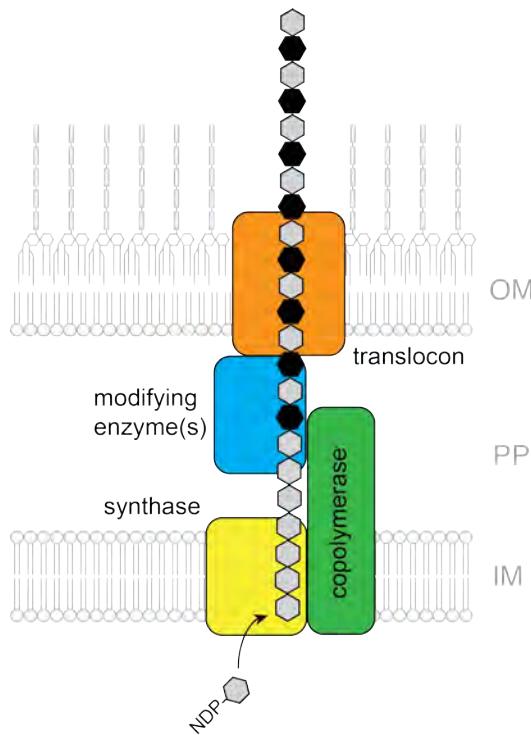


Figure 14.8. Schematic of a generic synthase-dependent EPS biosynthesis complex in a Gram-negative bacterium. The synthase is the enzymatically-active member of the polymerization complex, and the copolymerase is required for activity and regulation of the synthase.

In these EPS synthesis pathways, a synthase-copolymerase complex in the cell membrane constructs linear polysaccharides from NDP-monosaccharide precursors at the same time that the polysaccharide chain is translocated through the membrane. Once outside of the cytoplasm, there may be a variety of accessory proteins that modify some or all of the monomers in the polysaccharide chain, including acetylases, epimerases, and methylases. Finally, in Gram-negative bacteria, there are dedicated translocon proteins that allow the completed polysaccharide to cross the outer membrane. There do not seem to be any consistent naming conventions for the component proteins of synthase-dependent polysaccharide pathways.

The synthase is the enzymatically-active component of the inner membrane complex, while the copolymerases seem to play important roles in activation and regulation of synthase activity.

As you might imagine, the polysaccharides assembled by synthase-dependent pathways are, on average, not as structurally complex as those assembled by Wzy-dependent polymerases. In fact, probably the most abundant EPS assembled by a synthase-dependent pathway is cellulose, which is simply a linear chain of β -1,4-linked glucose. This and some other examples of EPS synthesized by synthase-dependent pathways are illustrated in Figure 14.9.

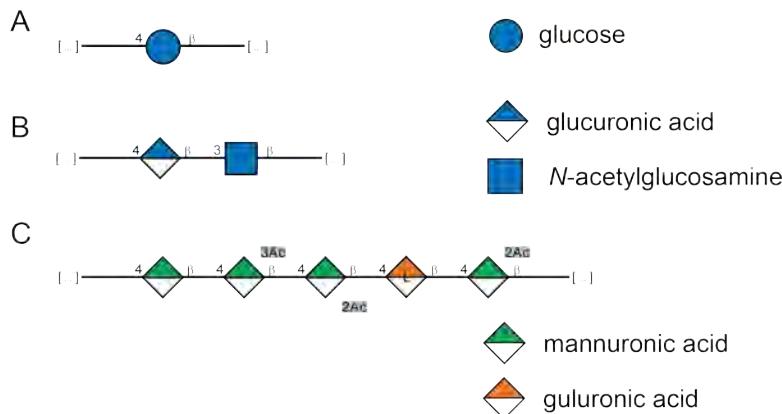


Figure 14.9. Examples of polysaccharides synthesized by synthase-dependent EPS biosynthesis. (A) Cellulose, synthesized by a very wide range of bacteria. (B) Hyaluronan, synthesized by *Streptococcus dysgalactiae* subsp. *equisimilis*. (C) Alginate, synthesized by *P. aeruginosa*. Schematics obtained from the Carbohydrate Structure Database.

Many EPS assembled by synthases are important for biofilm formation (as we will see below), while others play key roles in bacterial pathogenesis. Hyaluronan is structurally identical to a major mammalian polysaccharide and allows group A streptococci to evade recognition by the host immune system. Alginate, a thick, sticky polysaccharide produced by *P. aeruginosa*, is a major virulence determinant in the lungs of patients with cystic fibrosis.

ALTERNATIVE EPS SYNTHESIS PATHWAYS

There are two other major pathways by which EPS are assembled, which we will not discuss in detail here due to space limitations. This should not be taken to mean that the EPS made by these pathways are not important!

Some Gram-negative bacteria synthesize capsular glycolipids by a pathway that is characterized by the cytoplasmic synthesis and polymerization of polysaccharides on a phosphatidylglycerol lipid carrier and export via a dedicated ATP-dependent ABC transporter (**Lecture 16**). Unlike the previous pathways, the polysaccharides are completely assembled in the cytoplasm before export, and tend to be composed of relatively simple, unbranched repeat units. These include some capsular antigens from *E. coli*, *Neisseria meningitis*, *Haemophilus influenzae*, and *Salmonella typhi*.

Finally, some bacteria simply secrete extracellular transglycosylases (by the Sec pathway, in the examples I could find) so that EPS biosynthesis happens entirely outside of the cell. Glucansucrases polymerize starch or sucrose into a variety of α -linked glucans, and fructansucrases assemble β -linked fructans. The products are more heterogeneous than those synthesized by other pathways, but are particularly prominent products made by cultures of lactic acid bacteria. In fermented foods, EPS synthesized this way can have significant effects on food texture, while in biofilms (see below), they can play important roles in adhesion and colonization, for example by the dental pathogen *Streptococcus mutans*.

FUNCTIONS OF CAPSULES

Capsules are EPS attached to the surfaces of bacteria, and are common among many different species. In different species and strains they have different compositions, different thicknesses, and different mechanisms of synthesis, but their physiological roles fall generally into the categories of protection and/or adherence. We will discuss adherence more generally below and will focus here on their role in protecting the cell against environmental stress.

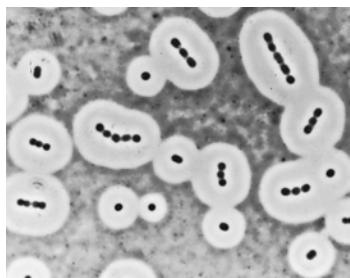


Figure 14.10. An India ink stain of *Acinetobacter calcoaceticus*, allowing visualization of the capsule as areas lacking stain under a light microscope. From Taylor & Juni (1961) J Bacteriol 81(5):688-693.

Capsules protect bacterial cells from abiotic stresses, like desiccation, UV light, and oxidative stress, but they have been more extensively studied for their roles in resistance to biotic stresses. This includes antibiotics and antimicrobial peptides, which may not be able to penetrate through the capsule, bacteriophage, whose receptors may be concealed under the capsule, as well as resistance to consumption by protists or phagocytic cells of the animal immune system (presumably by making them too big to be effectively phagocytosed).

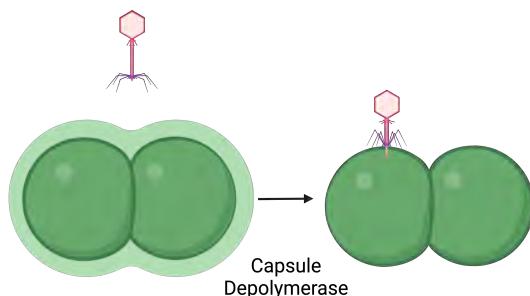
Immune recognition of capsular polysaccharides, as I mentioned above, is the basis of serotyping, and capsule is one of the main targets of the antibodies the adaptive immune system generates against many bacterial pathogens. Different strains of the same species often produce very different capsular polysaccharides, and capsule biosynthesis operons are often horizontally transferred from one strain to another. Selective pressure by the immune system drives the evolution of new, serologically distinct capsules. There are also bacteria (e.g. *S. pneumoniae*) where antibody binding to capsule can **enhance** colonization under certain circumstances, analogous to viral antibody enhancement.

Some bacteria are capable of *antigenic variation*, in which the same strain encodes two or more capsular polysaccharide synthesis pathways and is able to switch among them. This is a mechanism for evading the immune system, and allows infections to persist much longer, since antibodies need to be generated against all of the possible capsule types encoded by the infecting strain.

It should be noted, however, that while capsule's role in pathogenesis has received a lot of attention, many commensal bacteria **also** produce capsule, and those capsules seem to play important roles in establishing productive host-microbe interactions. The mechanism(s) by which they might do so, and how and whether the immune system can distinguish between capsular EPS produced by pathogens and those produced by commensals remains poorly understood.

EXAMPLE OF AN EXPERIMENTAL DESIGN: PHAGE-ENCODED CAPSULE DEPOLYMERASES

The evolutionary arms race between bacteria and phages has resulted in many proteins and systems that are potentially useful to humans. One that is fairly widespread among tailed phages is a class of enzymes called capsule depolymerases. As the name implies, these are enzymes that break down the polysaccharides that make up bacterial capsules, and the phages that encode them depend on them to penetrate the capsule and gain access to the bacterial outer membrane.



Encapsulated bacteria are protected from phage, but if a phage particle includes a depolymerase capable of degrading that capsule and exposing the cell surface, the bacteria become vulnerable to infection. Made with Biorender.

There have been a few studies showing that purified capsule depolymerases can be used in animal infection models to sensitize pathogens to attack by the immune system, so this is a potentially useful therapeutic tool.

Like other phage-derived antimicrobials, though, capsule depolymerases tend to have very high specificity. The phage KP36, for example, infects and degrades the capsule of serotype K63 *Klebsiella pneumoniae*, but not of other serotypes found on infectious strains (e.g. K1, K54, or K57). KP36 encodes a single depolymerase called Gp50, and purified Gp50 only degrades serotype K63 polysaccharides. This very high specificity limits the usefulness of depolymerases in treating infections.

The following genetic tools are available for *K. pneumoniae* (surprisingly, the same P1 transducing phage used for *E. coli* also works in many strains of *K. pneumoniae*):

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (5.5 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
generalized transducing phage (P1)	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

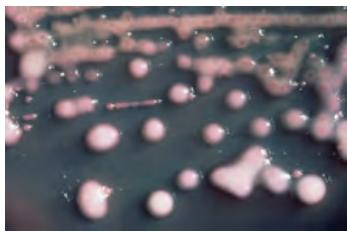
Design an experiment to isolate mutants of Gp50 able to degrade capsule polysaccharides other than K63.

Hypothesis: Gp50 can be mutated to change its substrate specificity.

Experimental Design:

This is somewhat complicated by the fact that phage KP36 is a lytic phage, not a lysogenic one, and is therefore not able to be genetically manipulated using the regular toolkit for *K. pneumoniae*.

The other problem is that a mutant hunt for gain-of-function Gp50 variants will require a lot of point mutations, and a low-throughput screen is unlikely to be very productive. It is certainly possible to tell if *K. pneumoniae* colonies have capsule on an agar plate, but this will only allow us to screen a couple of hundred colonies at a time at most.

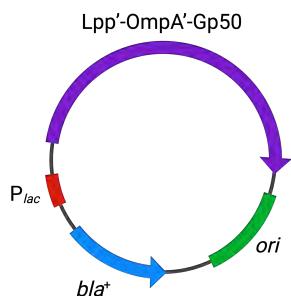


K. pneumoniae colonies on MacConkey lactose agar. Image from Wikipedia.

So, what can we do here? Since *K. pneumoniae* is closely related to *E. coli*, it's relatively straightforward to generate an expression plasmid that will allow us to express Gp50 on the *K. pneumoniae* cell surface ([this paper](#) describes a good tool for doing this, by expressing proteins as fusions to parts of the outer membrane proteins Lpp and OmpA), and we can use any number of methods to randomly mutate that plasmid, but how do we enrich for the mutants we want?

My idea is to use a phenotype described in *E. coli* in which bacterial strains that produce more capsule become more resistant to being spun down at low speeds in a centrifuge (paper [linked here](#)). Essentially, the capsule acts as a kind of parachute, and the more capsule a cell has, the more difficult it is to pellet.

- 1) Have an expression plasmid synthesized encoding an Lpp'-OmpA'-Gp50 fusion protein under control of a *lac* promoter. GenScript, for example, has pBluescript as one of the vectors they will synthesize genes in, so this should be straightforward.

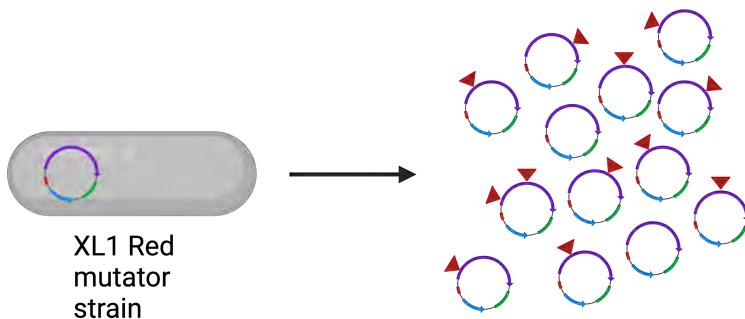


Made with Biorender.

I'm going for gene synthesis here, rather than cloning it myself, because it A) requires construction of a reasonably complex fusion protein, B) I don't need to purify KP36 phage DNA to PCR amplify from, and C) the standard *E. coli* vectors provided by DNA synthesis companies should work just fine in *K. pneumoniae*.

It would be easier to make the Gp50 secreted (see [Lecture 13](#)), but that would allow active Gp50 variants to cleave the capsule polysaccharides of other cells *in trans*, and we don't want that. By attaching the depolymerase to the cell surface, I hypothesize that I will limit its activity mostly to the cell that is expressing that particular variant.

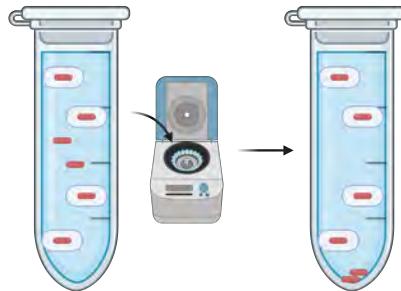
- 2) Randomly mutate the Lpp'-OmpA-Gp50 expression plasmid by passage through a mutator strain of *E. coli* (e.g. XL1-Red).



After an *in vivo* Copy step in XL1 Red, which lacks several important DNA repair enzymes, a plasmid prep will result in a pool of randomly mutated plasmids (as indicated by red triangles). These will mostly be point mutations. Some plasmids will not be mutated, and some may contain multiple mutations. Made with Biorender.

3. Transform *K. pneumoniae* serotype K1 with the mutant plasmid library and grow in media containing lactose (to induce expression of the Lpp'-OmpA-Gp50 fusion).

4. Centrifuge culture and separate pelleted cells from those that remain in suspension.



Centrifugation of a mixture of encapsulated and unencapsulated cells will result in the cells without capsule pelleting at the bottom of the tube more easily, allowing enrichment for cells expressing a capsule-degrading variant of Gp50. Made with Biorender.

It will certainly be necessary to do a pilot study to figure out the speed to spin the cells at here to actually separate bacteria with different amounts of capsule.

5. Repeat enrichment by growth and centrifugation several times.

6. Plate pelleted cells on agar containing lactose and screen for colonies that do not have as much capsule as the wild-type, which should look much less slimy.

7. Isolate plasmids from those colonies and sequence.

Independent Variable: (what will you change?)

- The sequence of the Gp50 variant in each plasmid.

Dependent Variable(s): (what will you measure?)

- The amount of capsule accumulated by each strain (indirectly, so this is a qualitative measurement).

Negative Control(s): (eliminate false positive results)

1) *K. pneumoniae* serotype K1 expressing wild-type Lpp'-OmpA'-Gp50. This should not degrade the K1 capsule and should not pellet.

2) Wild-type *K. pneumoniae* serotype K1 and K63. These should both have capsule and should not pellet.

Positive Control(s): (eliminate false negative results)

1) *K. pneumoniae* serotype K63 expressing wild-type Lpp'-OmpA'-Gp50. This should have its capsule degraded and should pellet well.

2) A *K. pneumoniae* mutant lacking capsule entirely. This should pellet.

Potential Outcomes:

1) Mutants of Gp50 are isolated that degrade the K1 capsule. This is what we want, and supports the hypothesis. We can then test it on other capsules and see if we have changed the specificity or (with luck) broadened it.

2) No mutants of Gp50 are isolated that are capable of degrading the K1 capsule. This could mean that our mutagenesis was inadequate, that K1 is too different from K63 for Gp50 homologs to ever act upon it, or that we simply didn't screen enough mutants.

Does this experiment establish correlation or causation?

- This experiment tests the **correlation** between specific Gp50 mutations and capsule accumulation in the test construct. Additional experiments will be needed to establish which changes in those mutant proteins might be causing that phenotype.

BIOFILMS

In most environments, the vast majority of microbial cells are not free-swimming *planktonic cells*, but are found attached to surfaces, often in multi-cellular aggregates called *biofilms*. The main exception to this rule is the open ocean, although some authors argue that the relatively high concentration ($\sim 10^5$ to 10^6 per milliliter) of bacteria in the top millimeter of the water column (the *bacterioneuston* or *surface micro-layer*), many of which are found in small aggregates, constitutes a kind of loose

biofilm. The same kind of debate about whether any aggregate of bacteria constitutes a biofilm is actively discussed in medical fields, as well. Defining what is and isn't a biofilm is, in many ways, a matter of semantics.

Of course, nearly all of our knowledge about bacterial physiology comes from studies of planktonic cells, which are more homogenous in their growth rates and far easier to measure and manipulate in the lab. This is a problem for those of us trying to figure out what bacteria are doing in the real world. It is, for example, very clear that biofilm formation is important in the virulence of many pathogens.

In this second half of the chapter, we will discuss the properties and development of biofilms, how bacteria sense and adhere to surfaces, and some of what is known about the regulation of biofilm formation. As you will see, the extracellular polymers produced by bacteria (including EPS) play an important role in biofilm biology.



Figure 14.11. A microbial mat, a kind of macroscopic biofilm, on the surface of a hot spring in Yellowstone National Park, Wyoming (Image from Wikimedia Commons, taken by Penny Higgins.)

Before I begin, though, I want to make a couple of important points. One is that most studies of the molecular biology of biofilm formation have been done using pure cultures and mostly with a fairly limited set of model organisms (notably *Pseudomonas aeruginosa*, for which you can largely thank George O'Toole, a pioneer in the field). Nearly all bacteria can form biofilms of some kind, and actual biofilms usually contain multiple species of bacteria and often other microbes. The second point is that, while biofilms have general properties in common, as we will see, the **details** of how bacteria attach to surfaces, what polymers they secrete to bind the biofilm together, and the effects of biofilm growth on their metabolism can vary widely, even between different strains of the same species. There is nothing simple about biofilms, and our understanding of their biology is actually fairly limited.

Biofilms form at interfaces, including those between liquids and solids, liquids and gases, and solids and gases. They can be macroscopically visible, and the names we give to different kinds of biofilms is often due to their visible appearance. Probably the best-studied kind of biofilm is the slimy kind that forms along solid-liquid surfaces, and most of the molecular biology we will discuss below was done on this kind of biofilm. Biofilms at liquid-gas interfaces are called *pellicles*. You could argue that bacterial colonies on an agar plate are a kind of biofilm forming at a solid-gas interface, and certainly many kinds of bacteria form elaborate multicellular structures under these conditions. When bacteria clump together in liquid cultures so that the aggregate of bacteria itself is the surface to which additional bacteria adhere, it is called *autoaggregation* (especially in pure cultures) or *flocculation*, and individual aggregate particles are sometimes called *flocs*. In a biofilm, bacterial cells are encased in a matrix composed of biopolymers, a species- and strain-specific combination of polysaccharides, proteins, and eDNA.

There is no hard and fast rule about how many adherent bacteria you need to constitute a biofilm. In reality, there is a continuum between planktonic cells, single adherent bacteria, microcolonies, flocs, etc. all the way up to very large and complex microbial mats many centimeters thick.

MEASURING BIOFILM FORMATION

There are a variety of ways of visualizing and quantifying biofilm formation in the lab. The challenge is making measurements of biofilm formation quantitative instead of qualitative, but there are a couple of approaches that are commonly used to do that. Some images of biofilms visualized with different techniques are shown in Figure 14.12.

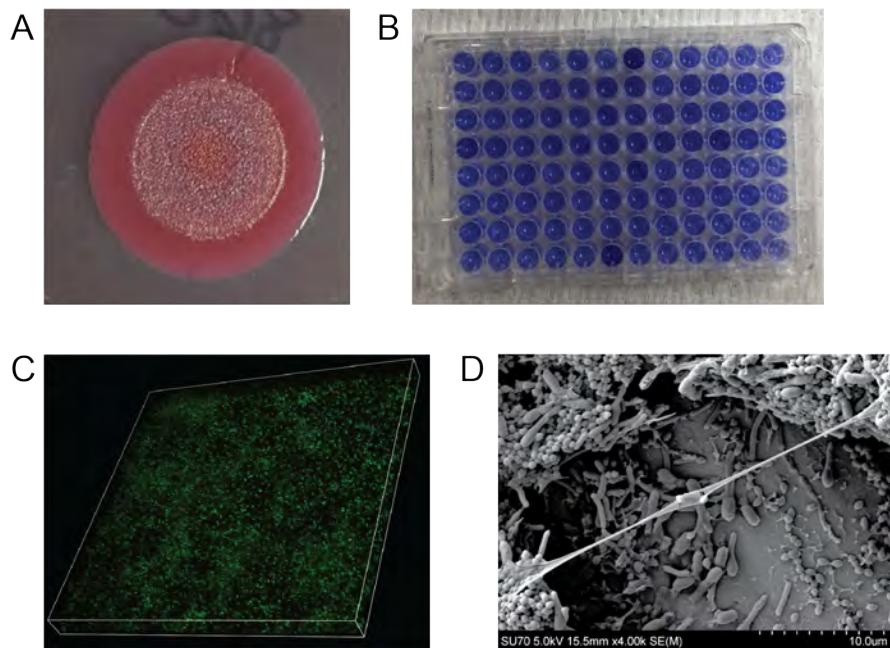


Figure 14.12. (A) Wrinkled colony *E. coli* biofilm on an agar plate containing the dye Congo red. (B) 96-well plate with crystal violet dye used to quantify biofilm formation. (C) 3-D confocal micrograph of an *E. coli* biofilm on a glass coverslip, stained with a green fluorescent dye. Images in A-C by Rhea Derke and Leanna Crafford (Gray lab). (D) Scanning electron micrograph of a polymicrobial biofilm on a stainless steel surface. Image from Wikimedia Commons, taken by Krzysztof A. Zacharski.

The simplest way to measure a biofilm is simply to take a picture of it. If one strain makes a lot of biofilm (at the gas-liquid interface in a test-tube, for example) and another does not, then it may be very easy to just show what it looks like. This is qualitative, but if there's a big enough difference, that might be fine.

A common assay to quantify differences in biofilm formation between strains uses the purple dye crystal violet. Cultures are grown, often in 96-well plates, under conditions where biofilms are expected to appear. The liquid media are then removed, and the plates are rinsed thoroughly to remove any cells that are not firmly adhered to the solid substrate. Next, crystal violet is added, which stains the adherent biofilm. Excess dye is rinsed away (this assay is very messy!), then resolubilized in ethanol. The dye in the resulting extract can be quantified in a spectrophotometer. [Here](#) is a link to a video protocol showing how this assay is done. The crystal violet assay has pretty high variability, but does a reasonable job of quantifying total biofilm mass. It does not give any information about more subtle structural differences between biofilms or distinguish between living cells and matrix materials.

Fluorescent microscopy, and especially *confocal microscopy*, which generates a 3D image by taking images at several different focal planes, is very useful for obtaining more subtle assessments of biofilm structure. Cells in a biofilm can be stained with various fluorescent dyes or engineered to express fluorescent proteins. There is a variety of software available designed to analyze and extract quantitative data from these kinds of images. Some students in my lab have gotten reasonable results using [BiofilmQ](#), but there are lots of other options, including [Comstat](#), [Imaris](#), and [Volocity](#) (those last two cost money, while BiofilmQ and Comstat are both free).

The highest resolution images of biofilms are obtained with electron microscopy, but it is difficult to extract quantitative information from electron micrographs, and the samples need to be fixed in such a way that the cells will certainly be dead when the images are taken.

Some labs that study biofilms intensively use sophisticated *microfluidics* technology which can maintain carefully-controlled flow rates, media composition, and growth conditions under constant microscopic observation. These kinds of experimental setups allow the best observations of biofilm formation and development over time, but [require specialized equipment](#).

BIOFILM DEVELOPMENT

The growth of a biofilm goes through a number of developmental stages, which are illustrated in Figure 14.13:

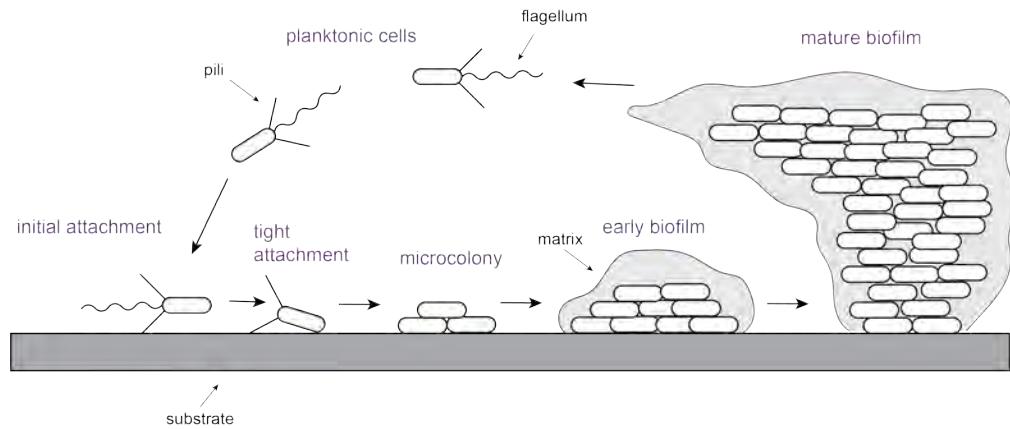


Figure 14.13. The developmental stages of a single-species biofilm. Planktonic cells swim freely until they encounter a surface, which may be sensed by a variety of mechanisms, commonly involving flagella and / or pili. Initial attachment is reversible, but is followed by irreversible tight attachment. Attached cells divide to form microcolonies, and produce extra-cellular matrix polymers to become an early, or immature biofilm. Mature biofilms contain many more cells, often have complex 3-dimensional structures, and are able to shed planktonic cells to begin the cycle again.

Planktonic cells, if they are motile, swim through the liquid phase of their environment until they encounter a surface. They may recognize that this has happened by a variety of mechanisms, including changes in the properties of liquid flow, distortions in their cell shape, increased resistance to flagellar rotation (because the surface gets in the way), or by binding of adhesive pili (usually, but not always, type 4 pili) to the surface. This initial encounter is reversible, and cells may spend some time close to the surface before swimming away. Non-motile bacteria (e.g. *Staphylococcus* and *Streptococcus* spp.) have a simpler attachment process which mostly involves settling by gravity or fluid flow.

However, on the path to a biofilm, the cell eventually becomes more permanently attached to the surface. It generally loses any motility appendages it might have had, and begins to replicate, forming a *microcolony*. As this continues to grow, matrix compounds are produced and the cells begin to enter into a state that is more biofilm-like. As it matures, the biofilm grows in height and width, supported by the matrix and forming more complex structures.

Finally, *biofilm dispersal* is the release of planktonic cells from the biofilm, or even the complete reversal of biofilm formation, and involves the active modification and breakdown of matrix components. See [this review](#) for more information. Dispersal is a regulated process, and since bacteria in biofilms are more resistant to antibiotics and disinfectants than planktonic cells, there has been quite a bit of interest in developing drugs that stimulate dispersal for use in combination with chemicals that kill bacterial cells.

ADHERENCE

The first step in biofilm formation is adherence. Bacteria can adhere to abiotic surfaces, like rocks, metal, and plastic, as well as biotic ones, like plant leaves, chitin, and animal epithelia (both internal and external).

Bacteria can adhere to surfaces by simple physical interactions, like charge-charge interactions (the outer surfaces of bacteria are usually negatively charged) or hydrophobic interactions, but also produce a variety of *adhesins*, which are molecules produced by bacteria that mediate the attachment of those bacteria to surfaces. They include proteins and carbohydrates, and can either bind to specific receptors or be just generally sticky (Figure 14.14).

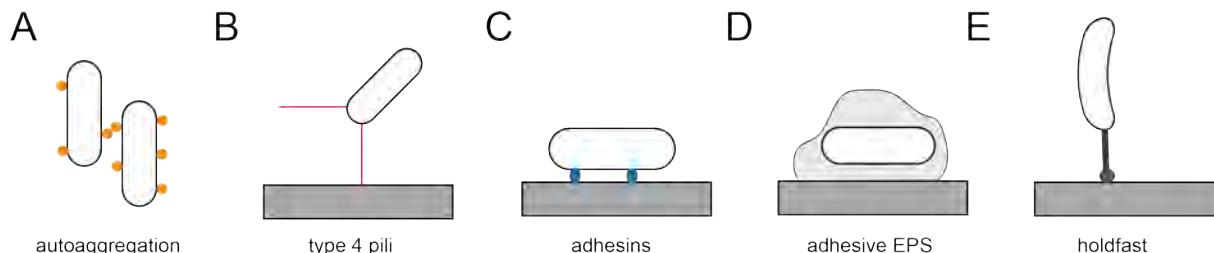


Figure 14.14. Different bacterial structures involved in adhesion. (A) Autoaggregation proteins bind to each other, causing bacteria to clump together. (B) Type 4 pili bind to surfaces and retract, pulling cells closer to the surface. (C) Adhesins are bacterial surface proteins that bind to surfaces directly, often to a specific receptor. (D) Some EPS are sticky, forming a glue that attaches cells to surfaces. (E) The holdfast of *Caulobacter* and similar polar adhesion foci in other α -proteobacteria combine adhesive EPS and proteins into a very strong attachment point.

When the surface that bacteria are adhering to is made up of other bacteria, and especially when it is other bacteria of the same species, this may depend on *autoaggregation* adhesins. One example is *E. coli* antigen 43, a 1099-amino acid autotransported (T5SS) protein encoded by the *flu* gene (which stands for **f**luffing, a description of the clumpy, flocculent phenotype

of antigen 43-expressing cells). The interaction between antigen 43 molecules on adjacent *E. coli* cells is tight and specific. Many bacteria autoaggregate, and this can be an important step in the early development of biofilms.

We discussed curli in **Lecture 13**, and these secreted amyloid proteins play an important adhesive role in mature biofilms produced by enterobacteria. Curli are very sticky and extremely stable, the exact properties that make amyloid aggregates in the brain so toxic in human neurodegenerative diseases. For a biofilm, however, this reinforces the matrix and makes it extremely resistant to physical stress.

Other adhesin proteins are divided generally into *fimbrial* and *non-fimbrial adhesins*. Fimbrial adhesins are fibrous and extend well away from the cell, and are also often called pili. There are several types, but many bacteria employ retractile type 4 pili (recall that these are homologous to type 2 secretion systems, **Lecture 13**). The initial attachment of *C. crescentus* swimmer cells, for example, depends on the type 4 Tad pili, which bind non-specifically to surfaces. Other kinds of pili can also be involved in attachment. The type 1 pili of uropathogenic strains of *E. coli* (UPEC), which are not retractile, but are exported by a type 5 chaperone-usher secretion mechanism (**Lecture 13**) are tipped with an adhesin called FimH. FimH binds tightly and specifically to mannose sugars on glycoproteins found on mammalian epithelial cell surfaces.

Non-fimbrial adhesins are a very diverse group of proteins, and again, range from non-specific sticky proteins to proteins with extremely specific binding partners.

Some Gram-positive and mycobacteria seem to export the metabolic enzyme glyceraldehyde-3-phosphate dehydrogenase (GapDH, see **Lecture 17**) to their surfaces, where it is involved in adherence to epithelial and endothelial cells and host proteins like plasminogen and fibrinogen. This kind of dual protein function is called *moonlighting*, and this particular case has required particularly extensive evidence to convince the scientific community that it's real, since GapDH's role as a central metabolic enzyme is so important.

Another, much more specialized example of a non-fimbrial adhesin is intimin, a surface protein produced by *enteropathogenic* and *enterohemorrhagic* *E. coli* strains (EPEC and EHEC, respectively). Intimin forms a very specific and tightly-bound complex with a protein called Tir (translocated intimin receptor). Tir is a substrate of the type 3 secretion system (T3SS, **Lecture 13**) of EPEC and EHEC, and is injected from the bacteria into host cells, where it is displayed on the host cell surface. These strains therefore deliver their own adhesin receptor into the cells to which they want to bind, allowing them to form extremely tight attachments to those cells.

Of course, polysaccharide EPS also play roles in attachment, since many EPS can act like glue to stick cells to surfaces and to each other. We've already discussed the *Caulobacter* holdfast, which contains a polysaccharide that is one of the strongest biological adhesives ever described. Surprisingly, the repeat unit structure of the holdfast EPS is not currently known (polysaccharide biochemistry is difficult!), but recent work from Sean Crosson's lab at the University of Chicago has shown that it contains a 1,4-linked backbone of glucose, mannose, N-acetylglucosamine, and xylose monosaccharides with branches at the C-6 positions of glucose and mannose.

As another example, *Pseudomonas aeruginosa* produces several EPS, including Psl, Pel, and alginate, all of which play different roles in adhesion and biofilm structure. Different strains produce different amounts of these polysaccharides, with most laboratory strains and wound isolates producing Psl and Pel, but very little alginate, while isolates from the lungs of cystic fibrosis patients produce very large amounts of alginate, but very little Pel. Psl is anchored to the bacterial cell surface, is found throughout *P. aeruginosa* biofilms, and is involved in both autoaggregation and attachment to surfaces. Pel, on the other hand, is required for formation of pellicle biofilms, is localized to the periphery and vertical mushroom stalk-like structures of surface-attached *Pseudomonas* biofilms, and is positively charged, which allows it to interact with extracellular DNA in the biofilm matrix.

If you want to explore this topic in a little more depth and see another example of the variety of polysaccharides bacteria use for adhesion and biofilm structure, [here](#) is a good review of EPS and biofilms in staphylococci.

STRUCTURAL AND METABOLIC PROPERTIES OF BIOFILMS

Bacteria in biofilms are much more resistant to physical stresses, disinfectants, antibiotics, predation by protists, and attack by the immune system. Some of this has to do with the physical properties of the biofilm and its matrix, but there are also important metabolic and regulatory differences between planktonic and biofilm-grown bacteria that contribute to stress tolerance.

The matrix components we discussed above as adhesins are often also important structural components that give biofilms their physical properties and contribute to the ability of bacteria in biofilms to resist stress. EPS and other adhesins interact with cells, surfaces, and each other, as well as with other compounds. An important component of the matrix of many biofilms is **extracellular DNA** (eDNA), first described by Cynthia Whitechurch in 2002 (in [this paper](#)). As mentioned above, positively-charged EPS can cross-link with eDNA to form a tight, gel-like matrix, and non-specific DNA-binding proteins can play a similar role. In some cases, eDNA seems to be released from dead cells in the biofilm, but in other cases chromosomal DNA is actively secreted by living bacteria. See [Sharma et al. 2024](#) for a summary of known eDNA functions.

Biofilms can also become mineralized, with the deposition of carbonate minerals solidifying the matrix into a rock-like state. This is clinically relevant in calcification of dental plaque and in the formation of kidney stones, which can be the product of mineralized biofilms produced by uropathogenic bacteria. The more cross-linked and impermeable the matrix, the more difficult it is for predators, immune cells, and disinfectants to penetrate or remove the biofilm from a surface. Biofilms of *Streptococcus mutans* are notorious for demineralizing tooth enamel, mostly by producing acid that dissolves the hydroxyapatite that makes up the hard outer layer of teeth.

The exact composition of a particular biofilm matrix depends on the strain(s) and species of the bacteria that live in it, as does the arrangement and interactions between the organism(s) involved. Since most biofilms are *polymicrobial*, there is a lot of complexity to deal with in real biofilms. See [these papers](#) for very cool visualizations of polymicrobial biofilms, to get an idea of what we're dealing with here. Some labs (including ones here at UAB) are beginning to address this complexity with laboratory models in which biofilms composed of two or more genetically-manipulatable species are studied, and this is an exciting area of contemporary research.

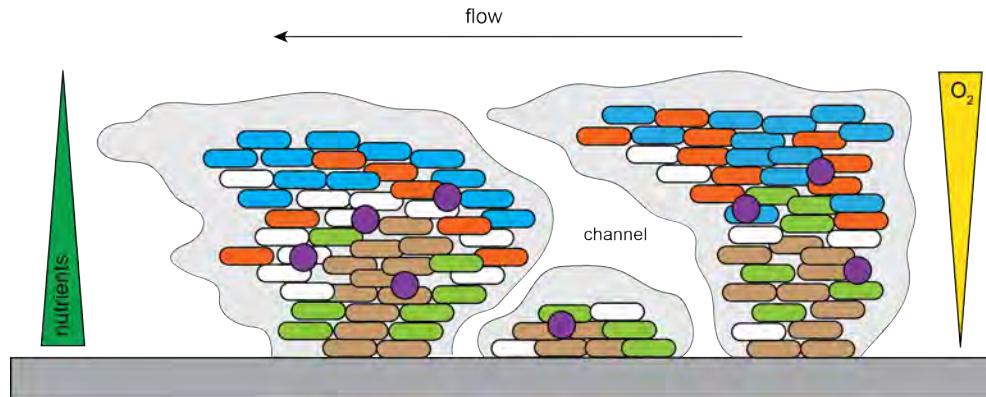


Figure 14.15. The structure of a polymicrobial biofilm, shaped by gradients of nutrients, oxygen, and flow. Different bacteria are found in different parts of the biofilm, with (for example) aerobes (blue and orange) more abundant near the surface and anaerobes (brown and green) near the substrate. Channels through the biofilm allow the exchange of molecules among different parts of the biofilm or the removal of waste products. Biofilms often contain dead or metabolically-inactive cells (white).

Bacteria growing in biofilms experience a number of gradients different from those generally experienced by planktonic cells, and unlike planktonic cells, bacteria in a biofilm are not able to move freely to respond to chemical gradients ([Lecture 15](#)). These gradients may be driven by both physical and biological factors. The cells near the top of a biofilm may experience higher oxygen levels, higher flow rates, greater light intensity, or greater exposure to environmental toxins, while cells near the bottom may experience lower levels of these stimuli, but higher levels of nutrients or available electron acceptors, depending on the species and the nature of the substrate to which they are attached. Biofilms in liquid environments often contain channels through which liquids can circulate, allowing nutrients and waste products to diffuse through the biofilm, but the permeability and density of the matrix can vary dramatically.

The physical properties of the biofilm matrix are protective, but that is not the only reason that bacteria in biofilms are resistant to stress. Unlike planktonic cells, which can be in a rapidly-replicating exponential growth phase, bacteria in biofilms grow and replicate much more slowly, if at all. This is analogous to the stationary phase of planktonic cultures, and slow-growing cells are usually **much** more resistant to killing by antibiotics and other stress conditions.

This is a complex topic, but the general concept is that antibiotics typically kill bacteria by inhibiting cellular processes essential for **growth**, like RNA polymerase, ribosomes, or cell wall synthesis. If the cell isn't growing and dividing, then inhibition of these processes isn't nearly as big a deal, and the cell can essentially wait out the antibiotic exposure. This is called **tolerance**.

Persister cells are an extreme example of tolerance which can be observed in both biofilm and planktonic cultures. A small proportion of the cells in a culture contain very low levels of ATP ([Lecture 16](#)), are not metabolizing or growing, and are therefore not susceptible to killing by antibiotics. Those cells can then recover and regrow after the antibiotics are removed, but will become susceptible to antibiotics once they have started growing.

Tolerance and persistence are distinct from antibiotic resistance, which results from heritable genetic mutations and depends on specific degradation or export proteins. The cultures regrown from a sample of persister cells will be just as sensitive to antibiotics as the original strain. As many as 1% of the cells in a biofilm or stationary phase culture may be persisters.

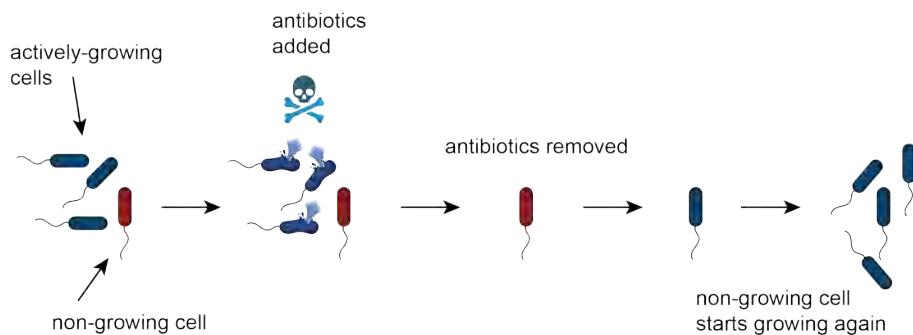


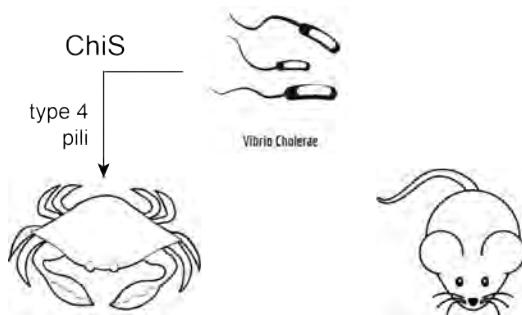
Figure 14.16. Antibiotic persistence / tolerance. In any population of bacteria, some proportion of the cells will be metabolically inactive and not growing (red). Upon addition of antibiotics, only actively-growing cells (blue) are killed. Once the antibiotic is removed, eventually some of those inactive cells will recover and regrow. Made with Biorender.

DISCUSSION PROBLEM SET #25: BIOFILMS IN THE VIBRIO CHOLERAE LIFE CYCLE

The causative agent of cholera, *Vibrio cholerae*, has two distinct phases to its life cycle. In brackish water, *V. cholerae* forms biofilms on the chitinous exoskeletons of copepods, a kind of microscopic marine crustacean, and doesn't cause any particular harm to those organisms.

However, when *V. cholerae* is ingested by a human, it grows to extremely high density in the small intestine, produces a potent toxin, and causes devastating rice water stool diarrhea, in which a patient loses up to 20 liters of water per day. There is good evidence that *V. cholerae* also forms biofilm-like aggregates in the human gut.

V. cholerae senses chitin (a polysaccharide not found in mammals, but common in the cell walls of fungi and the exoskeletons of crustaceans and insects) with a two-component regulatory system driven by the histidine kinase ChiS. ChiS activation leads to upregulation of a variety of genes for chitin utilization, including a type 4 pilus that mediates attachment to chitin. The chitin-regulated pilus is required for biofilm formation on chitin, but not for virulence in mouse models of cholera.



You wonder if there is any overlap between the pathways involved in biofilm formation in the two very different environments inhabited by *V. cholerae*.

The following methods are available for *V. cholerae*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (4 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
generalized transducing phage (CP-TTs; not commonly-used, though, as far as I can tell)	✓
compatible transposons	✓

oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

Design an experiment or series of experiments to address this question and identify genes or proteins involved in biofilm formation by *V. cholerae* A) on crustacean shells, B) in the mammalian gut, and C) in both environments. State:

- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiment(s), and how you will interpret them

REGULATION OF BIOFILM FORMATION

There are complex dedicated regulatory pathways that control the formation and dispersal of biofilms, and we will only be able to touch on a couple of these briefly. The first is quorum sensing, which we mentioned in [Lecture 11](#), and which plays an important role in biofilm regulation in a wide range of Gram-negative and Gram-positive bacteria. To review, bacteria synthesize and export quorum sensing signal molecules or *autoinducers*, and when those signals reach a threshold concentration in the medium they activate expression of specific genes.

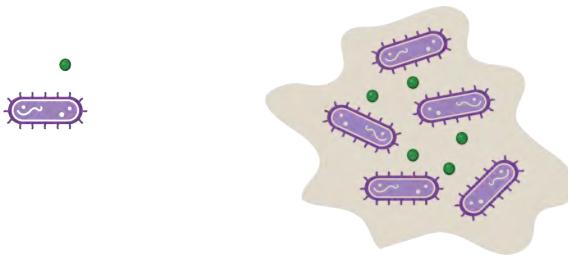


Figure 14.17. The basic principle of quorum sensing. Bacteria secrete diffusible signal molecules (green spheres). The concentration of these signal molecules correlates with the density of the bacterial population. Once the concentration of signal crosses a threshold, the bacteria respond to that signal by regulating population-level phenotypes, in this example, expressing an EPS matrix. Made with Biorender.

Quorum sensing systems are a mechanism by which bacteria sense cell **density**, and bacteria in a biofilm are in physical contact with each other, which is essentially the maximum possible density. The diffusion of quorum sensing signaling molecules may also be limited by the permeability of the matrix, increasing the local concentration. Quorum sensing is involved in biofilm formation in *Pseudomonas*, *Vibrio*, *Staphylococcus*, and a variety of other bacteria. For the sake of space, I'm going to focus on quorum sensing in Gram-negative biofilms, but you should be aware that the role(s) of quorum sensing in Gram-positive biofilms is complicated and quite different.

The initial work describing quorum sensing was done in *Vibrio* (now *Aliivibrio*) *fischeri* by Bonnie Bassler's lab, and in that Gram-negative organism, one of the systems under quorum sensing control is the expression of *luciferase*, which produces light. In *A. fischeri*, the signaling molecule N-3-oxohexanoyl homoserine lactone is synthesized by the enzyme LuxI and detected by the transcription factor LuxR. Homologous systems using different acyl homoserine lactone (AHL) signaling molecules are common in Gram-negative bacteria (we will discuss their synthesis pathway in [Lecture 18](#)). *P. aeruginosa*, for example, produces four distinct AHLs (one of which, butyryl-homoserine lactone, is illustrated in Figure 14.18), each of which regulates a different set of genes involved in biofilms, virulence, and other population-level behaviors.

Due to the generally hydrophobic nature of AHLs, they typically pass freely through cell membranes, so the concentration of AHL inside the cell is the same as that outside of the cell. This is not universally true for all combinations of AHLs and cell membranes, though. One of the **other** *P. aeruginosa* AHLs (N-3-oxohexanoyl homoserine lactone, generated and sensed by LasI and LasR, respectively) requires a dedicated efflux pump (MexAB-OprM) to be excreted from the cell. Note that this is the same molecule produced and sensed by LuxI/LuxR of *A. fischeri*, in which organism it diffuses freely through the membrane.

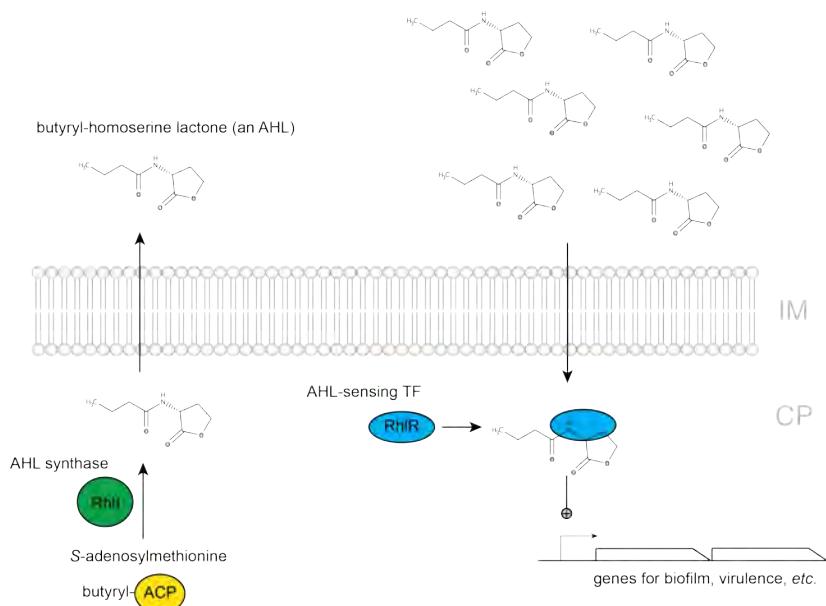


Figure 14.18. An example of an acyl homoserine lactone (AHL)-dependent quorum sensing signaling system, in this case the RhII/RhIR system of *Pseudomonas aeruginosa*. AHL synthases synthesize their cognate AHL from S-adenosylmethionine and an acylated acyl carrier protein (ACP), in this case with a 3-carbon butyryl group. AHLs diffuse freely through cell membranes, and once they reach a threshold concentration they bind to AHL-sensing transcription factors (in this case, RhIR) that regulate expression of genes involved in population behaviors, including biofilm development.

Many Gram-negative and some Gram-positive bacteria produce a conserved quorum signaling molecule called **autoinducer 2** (AI-2; a furanosyl borate diester) which appears to function as a signal of overall species-nonspecific bacterial population density, and is also involved in regulating genes required for biofilm production. As mentioned in **Lecture 11**, Gram-positive bacteria usually use peptide signals for quorum sensing.

For an in-depth review of quorum sensing in biofilms, I recommend [this paper](#).

Another important regulator of biofilm production is the second messenger (**Lecture 4**) cyclic-di-GMP (c-di-GMP). This molecule is widely conserved among bacteria, and is involved in regulating developmental processes in diverse species. It was first discovered in 1987 in *Komagataeibacter xylinus*, where it is required for cellulose production and formation of a pellicle biofilm (see Discussion Problem Set 7 in **Lecture 3**).

C-di-GMP is synthesized from GTP by guanylate cyclases, which contain a characteristic GGDEF motif, and is degraded by specific phosphodiesterases containing either HD-GYP or EAL domains (Figure 14.19). It is fairly common for GGDEF and EAL domains to be present in the same protein, indicating that these proteins can both synthesize and degrade c-di-GMP.

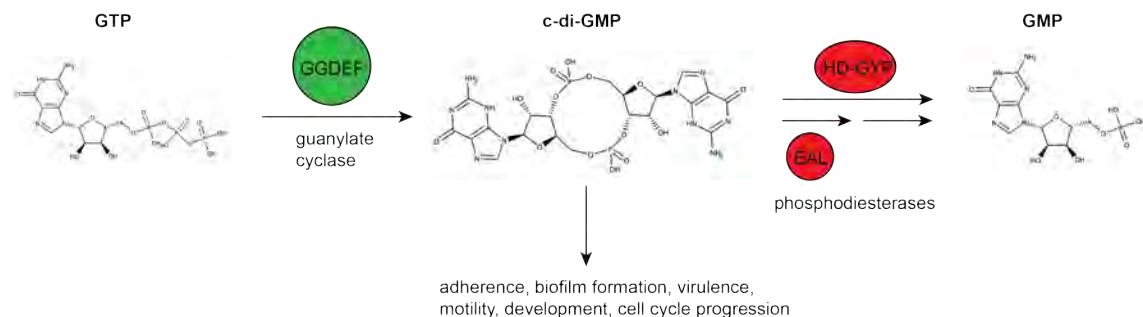


Figure 14.19. Cyclic di-GMP is synthesized from GTP by guanylate cyclases, which contain a characteristic GGDEF motif. C-di-GMP interacts with effector proteins that regulate a variety of cellular functions, as discussed in the text. C-di-GMP is degraded to GMP by specific phosphodiesterases, of which there are two types: those containing an HD-GYP domain, and those with an EAL motif.

Guanylate cyclases have a variety of sensor domains which respond to environmental stimuli to activate c-di-GMP production. C-di-GMP is then bound by different effector molecules, whose activity is modulated by that binding. This can regulate bacterial processes at transcriptional, post-transcriptional, and post-translational levels (**Lecture 4**).

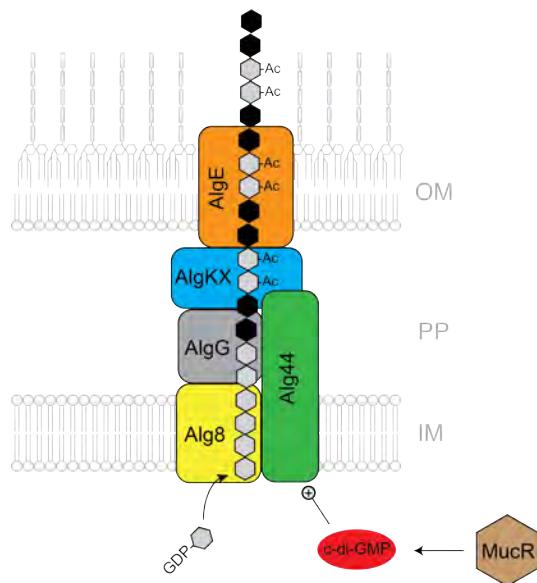
Many of the effectors regulated by c-di-GMP are relevant to biofilm formation. In *E. coli*, for example, c-di-GMP-bound YcgR inhibits the flagellar motor, inhibiting motility at the same time that c-di-GMP binds to BcsA, a protein involved in activating cellulose synthesis and to PdeR and DgcM, which ultimately lead to expression of CsgD, the transcription factor that activates the csg operon encoding curli. This theme of c-di-GMP repressing motility while inducing attachment / biofilm is con-

served in many bacterial systems. C-di-GMP is involved in regulation of type 4 pili in many bacteria, including *Myxococcus*, *Vibrio*, *Pseudomonas*, *Clostridium*, and *Caulobacter*, and is required for production of specific EPS in many biofilm-forming bacteria.

For a review of the diverse roles of c-di-GMP in bacterial regulatory networks, [this paper](#) is a great place to start, with an older, but more exhaustive review [here](#).

DISCUSSION PROBLEM SET #26: CYCLIC-DI-GMP SIGNALING SPECIFICITY

Cystic fibrosis-associated strains of *Pseudomonas aeruginosa* produce large amounts of alginate using a synthase-dependent EPS synthesis pathway encoded by the *alg* locus.



Alg8 and Alg44 polymerize GDP-mannuronate (grey hexagons), some of which are epimerized into glucuronate (black hexagons) by AlgG or acetylated by AlgK or AlgX as it crosses the periplasm to the translocase AlgE. C-di-GMP produced by the diguanylate cyclase MucR is sensed by Alg44 and activates polymerization. Note that alginate synthesis is regulated at multiple additional stages, and this figure is a dramatic simplification of the process.

C-di-GMP production **by MucR** is required for alginate production (i.e. a *mucR* null or enzymatically-inactive mutant does not produce alginate), but the genomes of *P. aeruginosa* strains also encode 15 or 16 **other** GGDEF domain-containing guanylate cyclases that produce chemically identical c-di-GMP.

The following methods are available for *P. aeruginosa*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (6.3 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
generalized transducing phage (ϕ PA3)	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓

a genome-wide knockout collection ([link](#))

✓

Propose a model to explain why c-di-GMP synthesis by MucR **specifically** is required for alginate production in *P. aeruginosa*. Design an experiment or series of experiments to test your model. State:

- your model
- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiment(s), and how you will interpret them

LECTURE 15: MOTILITY

INTRODUCTION

Not all bacteria are able to move under their own power, but in many environments the ability to do so (*motility*) provides a significant advantage. In this chapter, I will discuss several different mechanisms by which different bacterial species move. I will also cover the basics of how bacteria direct their movement towards or away from particular stimuli, a phenomenon generally called *chemotaxis*.

OVERVIEW OF BACTERIAL MOTILITY

Motility is essential for the survival of many bacteria both in the environment and for the ability of some pathogens to cause disease. It allows cells to spread into new environments, either as individuals or as populations. The mechanisms of motility are varied, and different mechanisms are necessary in different environments, depending on the viscosity of the medium, whether the cells are moving as individuals or as a group, whether they are moving through a liquid or along a surface, and the nature of that surface.

As we will see, motility depends on complex multi-protein machines, and these are among the largest and most complicated structures found in bacterial cells. (An individual flagellum consists of 30 or so different proteins, with copy numbers ranging from 1 to 30,000, a total length of as much as 15 μm , with a total molecular mass of more than 1000 MDa, and some bacterial species can have dozens of flagella per cell.) In the interests of space, we will not go into great detail about the genes and proteins that make up any individual motility complexes here, although we **will** briefly discuss the molecular machinery underlying chemotaxis in *E. coli*.

Howard Berg, a giant in the field of bacterial motility who sadly passed away in 2021, posthumously published a [review](#) that you may find of interest. You can watch a two-part lecture by Dr. Berg on YouTube [here](#) and [here](#) (from 10 years ago, but still worthwhile, especially on the basic principles of chemotaxis, which I will describe below).

SWIMMING

In relatively low-viscosity liquid environments, many bacteria are able to actively swim using *flagella*. Flagella are fairly rigid helical filaments (either right- or left-handed, depending on the species) that rotate to act as propellers to push or pull bacterial cells through the medium. This is unlike eukaryotic flagella, which are completely unrelated evolutionarily and move in a whip-like manner, not by rotation.

Different bacterial species have different numbers and arrangements of flagella (Figure 15.1). Some species are *monotrichous*, and have a single *polar flagellum* at one end. Others are *amphitrichous* and have one flagellum at each pole. Species with *lophotrichous* and *amphilophotrichous* flagella have multiple polar flagella at one or both poles, respectively, and those with *peritrichous* flagella have multiple flagella distributed around the cell surface. *E. coli*, the organism in which flagellar motility is best studied, has an average of 4 to 6 peritrichous flagella per cell. In some Gram-negative bacteria, the flagella are *sheathed*, which means they are covered by a lipid bilayer derived from the outer membrane. *Aliivibrio fischeri*, the bioluminescent symbiont of Hawaiian bobtail squid, has recently been shown to coil its single sheathed polar flagellum tightly around itself to burrow through narrow channels filled with mucus to enter the light organ of its host.

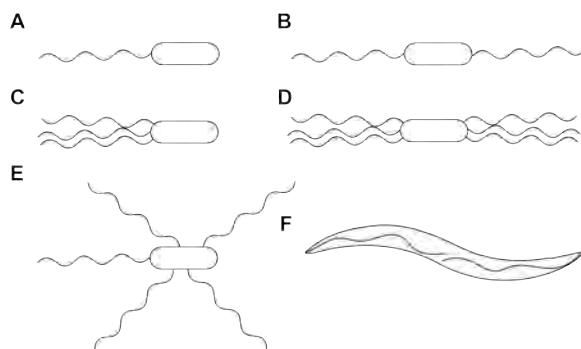


Figure 15.1. Representative arrangements of bacterial flagella, including (A) polar or monotrichous, (B), amphitrichous, (C) lophotrichous, (D) amphilophotrichous, and (E) peritrichous flagella, as well as (F) a very poorly-drawn rendition of the internal periplasmic flagella of spirochetes.

The *flagellar basal body* is a specialized type III secretion system (see [Lecture 13](#)) (Figure 15.2) that is dedicated to the export of flagellar proteins. The basal body and attached flagellum is spun by *motor proteins* (called MotA and MotB) that are held in place by interactions with the cell wall and use either a proton or sodium gradient to power the rotation (see [Lecture 16](#)). The flagellum is essentially a very small rotary electric motor. MotAB makes up the *stator*, or stationary part of the motor, and turns the *rotor* made up of the rest of the basal body. The speed of rotation of bacterial flagella is variable, but generally appears to be tuned to maintain constant torque under physiological conditions.

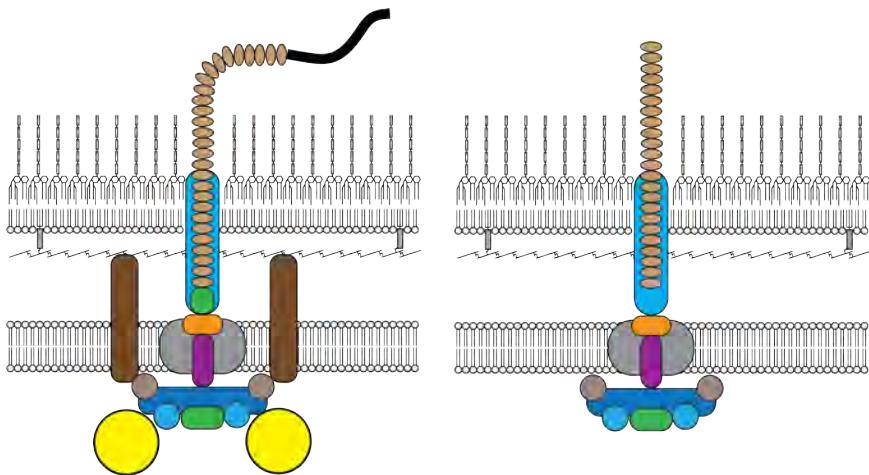


Figure 15.2. Diagram illustrating the structural similarities between the bacterial flagellum (on the left) and the type 3 secretion system (on the right).

The flagella of archaea (also called *archaella*) are not homologous to bacterial flagella, and are evolutionarily related to type 4 pili, but are also rotating helical filaments, so function similarly in cellular propulsion. (They are also involved in attachment, at least in some species, which is an interesting example of opposite functions for the same organelle under different conditions.) Unlike bacterial flagella, which are powered by ion gradients directly, archaella are powered by ATP hydrolysis ([Lecture 16](#)).

The maximum swimming speed of different bacteria varies a great deal, and is affected by the viscosity of the medium they are moving in (unsurprisingly, bacteria move slower in more viscous solutions). *E. coli*, whose cells are about 1 μm long, has a maximum speed of about 30 μm per second, while the freshwater predatory bacterium *Bdellovibrio bacteriovorus* moves at about 160 μm per second. Some marine bacteria have been reported to swim at up to 400 μm per second, which may be an adaptation to life in the very dilute, sparsely-populated ocean environment.

Flagella are too thin to see with normal light microscopy (about 20 nm in diameter), but can be stained or visualized in a variety of ways. Most of these, unfortunately, involve killing the cell. However, many recent studies have taken advantage of a technique in which the *flagellin* protein (which makes up the body of the flagellar filament) is mutated to contain an extra cysteine residue. A fluorescent dye attached to a cysteine-reactive group (usually a [maleimide](#)) can then be added to fluorescently label the flagella of living cells. Analogous methods have been applied to visualize a variety of different proteinaceous cell surface structures.

Swimming motility in the spirochetes is unlike that of any of the bacteria mentioned above, although it does still depend on flagella. Spirochete cells are long, corkscrew-like spirals that can burrow through extremely viscous environments, notably the connective tissue of animal hosts (Figure 15.3).

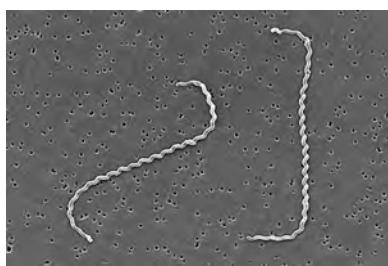


Figure 15.3. Individual cells of the spirochete *Leptospira interrogans* (Wikipedia).

Spirochetes cause a number of important diseases (e.g. leptospirosis, Lyme disease, and syphilis), and their ability to penetrate host tissues is key to their pathogenesis. Spirochetes have amphitrichous flagella anchored near the cell poles (as few as one per pole up to hundreds, depending on the species), but the flagella are wrapped close to the cell **inside** the periplasm (sort of shown in Figure 15.1 F, although that is an appallingly bad drawing). As they rotate, they turn the whole cell body, drilling it through whatever medium they are embedded in. The flagella of spirochetes are also sheathed in a complex, asymmetric protein layer not found in other bacteria, and which appears to be involved in maintaining a supercoiled structure important for determining overall cell shape and driving motility.

SWARMING

The other form of bacterial motility that depends on flagella is *swarming*, which is one way in which bacteria move along solid surfaces (Figure 15.4). Swarming is a group activity that requires many cells working together, and only occurs on soft, moist surfaces (in the lab, freshly-made rich media plates containing 0.4 – 0.6 % agar; although this is obviously highly artifi-

cial). Swarming is often, but not always, associated with the secretion of surfactants, which are amphipathic lubricants that reduce the surface tension, and therefore the drag, between bacteria and a solid surface. Many surfactants (like surfactin from *Bacillus subtilis* and serratine from *Serratia liquefaciens*) are small, cyclic, non-ribosomally assembled peptides (see [Lecture 18](#)).

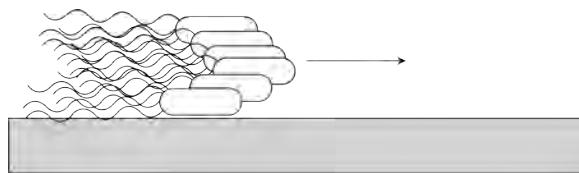


Figure 15.4. Swarming motility is a group behavior in which populations of bacteria move along moist surfaces. It is dependent on the concerted action of many flagella per cell.

One notable feature of swarming is that it requires (for reasons that are not completely understood) the interaction of many flagella, so that bacteria often increase the expression and number of flagella they produce under swarming conditions. In some species that normally produce only a single polar flagellum, an entire second flagellar operon system is expressed for swarming, producing large numbers of peritrichous flagella in addition to the normal polar one. Many *Vibrio* species do this, for example.

Swarming bacteria can move very rapidly, at close to the same speed as swimming (i.e. about 30 μm per sec for *E. coli*), and swarming cells can therefore spread across the surface of a soft agar plate very quickly. For this reason, many domesticated lab strains of bacteria have been selected for the loss of swarming, since fast-spreading colonies are inconvenient for microbiologists. This is especially noticeable in *Bacillus subtilis* and some *Pseudomonas* strains, where “undomesticated” wild isolates often form elaborate and [beautiful swarms](#) on the surface of growth media.

TWITCHING

Twitching is another type of motility along moist surfaces, but in this case it is not dependent on the presence of flagella. Instead, it depends on the presence of type 4 pili, which, if you recall from [Lecture 13](#), are homologous to the **type II** protein secretion system and are important adhesins cells use to attach to surfaces. Cells extend pili that adhere to surfaces via proteins at their tips, then retract those pili, pulling the cell along like a grappling hook (Figure 15.5). The jerky movement of cells using this system led to the name.

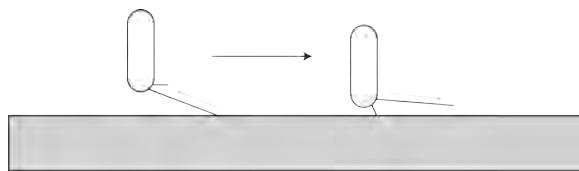


Figure 15.5. Twitching motility in bacteria by extension and retraction of surface-binding type 4 pili.

Twitching motility can be performed by individual cells (for example, those of *Pseudomonas aeruginosa*) or by populations (e.g. *Myxococcus xanthus* S- [social] motility, [Lecture 11](#)). *Pseudomonas aeruginosa* type 4 pili are localized to the poles of the cells, so they tend to stand up on end when moving via twitching, and move relatively slowly (~0.2 μm per sec).

GLIDING

Gliding motility is a catch-all term for bacteria moving smoothly along a surface with no obvious external appendages. Swarming and twitching both require fairly soft, moist surfaces, while gliding bacteria can often move across firmer, drier substrates. Many different kinds of bacteria can do this, and there are at least three radically different mechanisms by which they do so. These are exemplified by gliding motility in the mycoplasmas, in the myxobacteria, and in the *Bacteroidetes*.

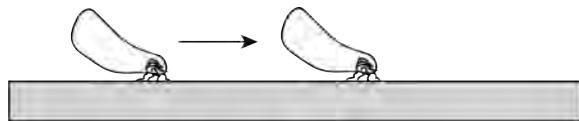


Figure 15.6. Gliding motility in *Mycoplasma mobile* depends on surface binding by many large proteins that change conformation to pull the cells along a sialic acid-coated surface.

The gliding motility of mycoplasmas has been best studied in the fish pathogen *Mycoplasma mobile*. *M. mobile* glides at 2 to 4 μm per sec along the surfaces of host cells by the repeated binding to, pulling, and release of host sialic acid residues of up to four hundred and fifty 50-nm long leg-like structures composed of the very large proteins Gli123, Gli349, and Gli521 (the numbers in their names are each protein's mass in kDa) (Figure 15.6). These legs are attached to a cytoskeletal structure called the “jellyfish” (composed of many copies of each of 10 JSP jellyfish-structure proteins) which defines the front of the mycoplasma cell. As you may remember from [Lecture 10](#), mycoplasmas have no cell wall and are small, flexible, and

blob-like. Gliding proceeds in the direction of the jellyfish structure, as the gliding machinery pulls the cells along the surface. Mycoplasma gliding is powered by ATP hydrolysis catalyzed by the JSP MMOB1670. See [this paper](#) for more details and electron microscopy of the jellyfish structure.

Myxobacteria, as we discussed in **Lecture 11**, are predatory bacteria with a complex social lifestyle exemplified by the model organism *Myxococcus xanthus*. When moving as a swarm, their S- or social-motility is (as mentioned above) type 4 pili-dependent twitching. However, myxobacteria are also capable of exploring surfaces by gliding as single cells, and this is called A-motility (for adventurous). Both S- and A-motility are required for both efficient predation and myxospore development.

The mechanism of A-motility has been mysterious for some time, but a recent model (described in [this paper](#)) has emerged to explain the phenomenon (Figure 15.7). Genes required for gliding motility were first identified in the late 1970's, but little real progress was made in understanding how A-motility is powered until 2011, when the AglRQS protein complex required for gliding was identified as a proton channel composed of a MotA homolog (AglR) and two MotB homologs (AglQS). This suggested that the energy driving gliding was the proton motive force, and that it was, in some way, descended from or related to the flagellar motor system. Unlike MotA and MotB, however, AglRQS are **not** attached to the cell wall, and in fact, they are able to move freely within the inner membrane (*M. xanthus* is a Gram-negative species).



Figure 15.7. Gliding (A- or adventurous-) motility in *Myxococcus xanthus* depends on the production of extracellular polysaccharide slime and a standing wave distortion of the cell shape, the physical interactions between which pull the cells forward along the surface.

High-resolution microscopy eventually revealed that the hundreds of individual AglRQS motor complexes in an *M. xanthus* cell are in continuous motion in rotating helices around the length of the cell, with “traffic jam”-like accumulations along the cell’s ventral surface. Exactly how this led to gliding motility was obscure, however; until the recent work cited above showed that, in combination with the production of a thin slime layer underneath the cells, the deformations of cell shape driven by AglRQS create a net forward capillary force or pressure gradient. The exact properties of this force depend on the nature of the surface the cells are on, and it will be fascinating to see if other groups of bacteria take advantage of the same kinds of physical forces for movement along surfaces.

As mentioned briefly in **Lecture 13**, some members of the Bacteroidetes phylum are capable of gliding motility via a type IX secretion system-dependent mechanism (Figure 15.8).

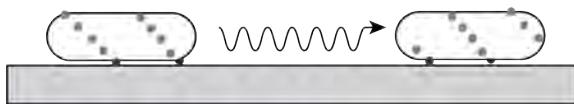


Figure 15.8. Gliding motility in *Flavobacterium johnsoniae* is driven by T9SS-dependent movement of the adhesin SprB (grey circles) along a helical track on the exterior of the bacterial cell.

This is best studied in *Flavobacterium johnsoniae*, where the T9SS secretes many proteins, including the adhesin SprB. SprB is distributed along a fixed helical track around the outside surface of the cell, and movement of the cell is driven by the movement of SprB along this track. The track is made up of the cell-wall anchored GldJ protein (Figure 13.11). The body of the cell rotates along its long axis as it moves. *F. johnsoniae* glides relatively quickly, at up to 4 μm per sec. The T9SS is also responsible for powering the movement of SprB along the cell surface, and does so by rotating at a constant speed of about 1 Hz, apparently acting analogously to the pinion of a [rack and pinion](#) motor (see [this paper](#) for more details). The rotation of the T9SS is driven by the proton motive force. The oral bacterium *Capnocytophaga gingivalis* uses T9SS-dependent gliding motility to penetrate biofilms of other bacteria, and can carry bacteriophage along with it as it does so (as described [here](#)).

Many filamentous cyanobacteria are able to move along surfaces, and this was referred to as gliding motility in the older literature, but more recent results show that cyanobacterial gliding is powered by type 4 pili and is actually a form of twitching motility, although they do also have to produce a polysaccharide slime to lubricate the surface they move along. This slime-dependence caused some early models to suggest that the force of slime extrusion was pushing the cells along a surface. This hypothesis has been discarded now that new evidence shows that slime extrusion is not sufficient for motility.

KINKING

Spiroplasma species are cell wall-less bacteria closely related to the mycoplasmas (**Lecture 10**). Many *Spiroplasma* species are symbionts of insects, some are pathogens of crustaceans, and others are plant pathogens, causing diseases like corn stunt disease and citrus stubborn disease. Unlike the blob-like *Mycoplasma* spp., *Spiroplasma* cells are helical, and they are able to swim in liquid media without flagella or other appendages.

The helical shape of *Spiroplasma* is maintained by a cytoskeletal protein called Fib that polymerizes into a helical bundle called the fibril. It associates with the cell membrane in an interaction that requires an MreB homolog ([Lecture 11](#)) called MreB5. Dynamic changes in the helicity of the fibril create kinks or bends in the helical cells which propagate processively along the length of the cell, creating wave-like motions that propel the cells forward. The exact mechanism by which this is accomplished is unknown, although *Spiroplasma* motility is inhibited by chemicals that disrupt the proton motive force ([Lecture 16](#)), so this is presumably the energy source that drives this form of motility.

CHEMOTAXIS: INTRODUCTION

Chemotaxis, strictly speaking, is the ability of an organism to move along a chemical gradient, either towards chemical *attractants* or away from chemical *repellants*. It is often used as a catch-all term for any movement towards or away from stimuli, though, although this isn't quite technically correct. Many bacteria, especially photosynthetic ones, exhibit *phototaxis*, or movement towards light, motile aerobic bacteria are almost universally capable of *aerotaxis*, movement towards oxygen, and a subset of bacteria exhibit *magnetotaxis*, or movement along magnetic field lines. All of these kinds of directional motility have some features in common, especially with regards to how cells regulate their movements, sense changes in their environments, and adapt to new conditions.

Chemotaxis is important for many motile bacteria to spread efficiently, acquire nutrients, and avoid predation, both in environmental bacteria and in bacteria that live in animal hosts. For example, in the human gut, enterohemorrhagic *E. coli* are chemotactically attracted to the human hormone norepinephrine, which is thought to allow the bacteria to move rapidly towards host epithelial cells. Many plant pathogens and symbionts are attracted by chemical compounds secreted by the roots of their plant hosts. There are a vast array of chemicals that either attract or repel different bacteria, and each bacterial species has its own repertoire of chemical responses. Not all motile bacteria have chemotaxis systems, and some (like the actin-polymerizing intracellular bacteria described in one section below) simply move randomly, presumably just so that they will spread more rapidly through their environments.

In this section, we will focus mostly on the molecular mechanism of chemotaxis in *E. coli*, since that is a well-understood and relatively simple model system. Be aware, as always, that other species have somewhat different, and in many cases more complex, pathways to accomplish chemotaxis, but the basic principles laid out here apply broadly.

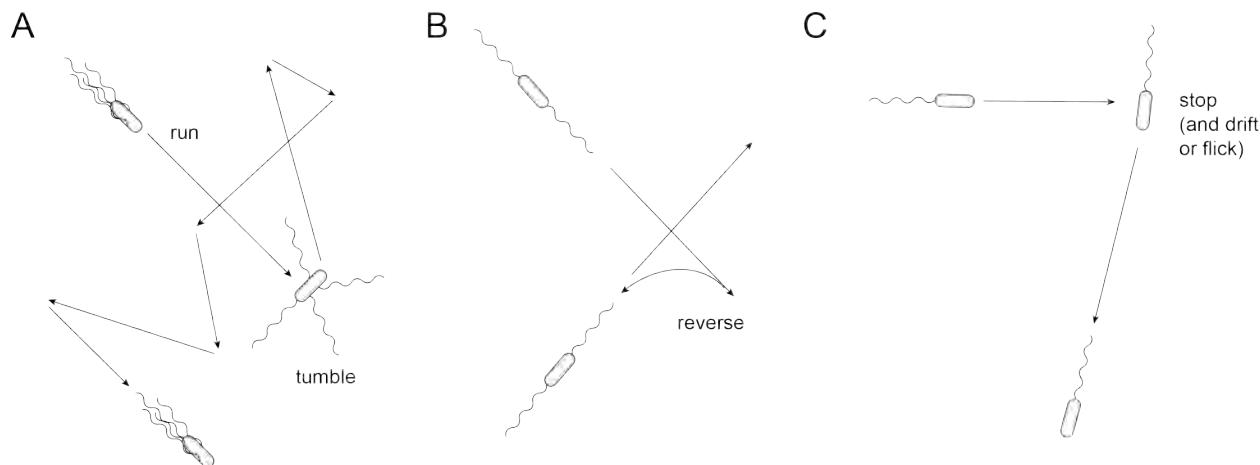


Figure 15.9. Patterns of movement by swimming bacteria. (A) Many bacteria with peritrichous flagella (including *E. coli*) move in more or less linear runs, interspersed with random tumbles, distinguished by the direction of rotation of the flagellar motors. (B) Amphitrichous and some monotrichous bacteria move in runs, interspersed with reverses in direction which reorient the bacteria in a random plane. (C) Many monotrichous bacteria intersperse runs with stops, during which they reorient randomly due to Brownian motion. Some also are able to perform a flagellar flick to actively switch their direction of travel.

Bacterial movement is characterized by *random walk* patterns, in which the cells alternate periods of linear movement with random reorientations and changes in direction. As shown in Figure 15.9 A, for bacteria like *E. coli* with peritrichous flagella, the linear movements are called "runs" and the reorientations are called "tumbles". In the absence of chemotaxis, this results in entirely random movement through three-dimensional space. However, when attractants or repellants are present, bacteria change the proportion of runs and tumbles such that their **average** movement is in a favorable direction.

Some bacteria with different arrangements of flagella are not able to actively tumble and reorient themselves by reversing direction (Figure 15.9 B) or by simply stopping and allowing Brownian motion to jostle them around (Figure 15.9 C). Some monotrichous bacteria also perform a flick with their flagellum to reorient themselves. On two-dimensional surfaces, the same general idea of interspersing periods of linear movement with random reorientations applies, although constrained, of course, by the surface itself.

HOW CHEMOTACTIC SIGNALS ARE SENSED

But how are these runs and tumbles regulated in response to changes in the environment? Bacteria are, in general, too small to sense chemical gradients over **space** (the length of their bodies), so they typically sense changes in their environment over **time**. (See [this recent paper](#) for evidence to the contrary, at least in *P. aeruginosa* moving on surfaces, though.) For *E. coli*, a general rule of thumb is that environmental conditions are constantly being compared to conditions from about 2 seconds ago. If conditions are **more** favorable than they were 2 seconds ago, the length of runs between tumbles **increases**, biasing the random walk in a favorable direction. If conditions are **less** favorable, the frequency of tumbling increases, as the cell attempts to reorient itself into moving in a more favorable direction.

In *E. coli*, the default direction of rotation for the flagella is clockwise (CW), and when the flagella are rotating CW, the cells make linear runs. However, the flagellar basal body contains a *motor switch complex*, which includes the FliM protein, that can reverse the direction of rotation to counterclockwise (CCW), resulting in tumbling. The frequency of CCW rotation of the flagella is directly controlled by the phosphorylation state of a response regulator called CheY (Figure 14.10 A)(recall the discussion of two-component regulatory systems in [Lecture 4](#)). Phosphorylated CheY (CheY-P) interacts directly with FliM, and **the more CheY-P is present in the cell, the more frequently it will tumble**. To counteract CheY-P accumulation, the phosphatase CheZ dephosphorylates CheY-P to inactive CheY more or less constitutively, although some data suggests that CheZ activity may also be regulated under some conditions.

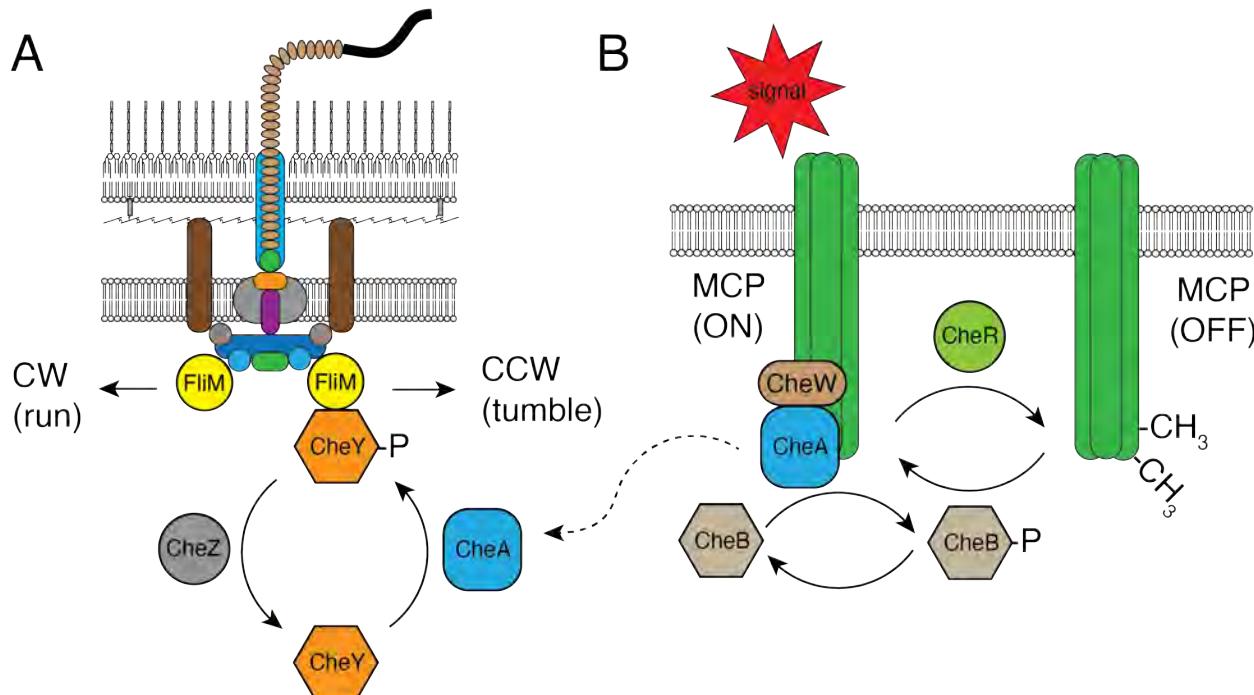


Figure 15.10. Chemotaxis in *E. coli*. (A) Flagella can rotate either clockwise (CW) or counterclockwise (CCW), which bias the cell towards either runs or tumbles, respectively. The switch in direction is regulated by an interaction between the motor switch protein FliM and the response regulator CheY. The histidine kinase CheA phosphorylates CheY and the phosphatase CheZ dephosphorylates CheY-P. Phosphorylated CheY increases the rate of switching to CCW rotation. (B) Methyl-accepting chemotaxis proteins (MCPs) interact with CheA (via the adaptor protein CheW) to drive the phosphorylation of both CheY and the response regulator CheB. Phosphorylated CheB is a demethylase that removes inhibitory methyl groups from MCPs. CheR is a constitutive methyltransferase that continuously methylates MCPs, inactivating them.

The first key question, then, is this: how do environmental conditions affect the phosphorylation state of CheY?

The genome of *E. coli* K-12 encodes five sensor proteins that indirectly control CheY phosphorylation, and therefore chemotaxis: the **methyl-accepting chemotaxis** (MCP) proteins Tsr, Tar, Trg, and Tap, and the non-methylated MCP homolog Aer. (See the section below for a discussion of the role of MCP methylation in chemotaxis.) Tsr is responsible for sensing the attractants serine, cysteine, alanine, glycine, and warm temperatures (37°C) and the repellants acetate, benzoate, indole, and leucine. Tar is responsible for sensing the attractants aspartate, asparagine, glucose, maltose, and phenol and the repellents nickel and cobalt. Trg senses the attractants ribose and galactose, and Tap is responsible for attraction towards dipeptides. Aer senses the energy state of the cell, and is responsible for attraction towards terminal electron acceptors ([Lecture 16](#)), especially oxygen.

The MCPs are integral membrane proteins, and form large clusters or arrays at the poles of the bacterial cell. Their interactions with each other are thought to allow amplification of weak signals. Note that different species of bacteria (and, indeed, other strains of *E. coli*) have different repertoires of MCP proteins and therefore are able to sense and respond to different sets of environmental signals. *Azospirillum* sp. B510, a nitrogen-fixing bacterium that is associated with the roots of rice, has

89 different MCPs, for example, and we have very little idea what each of them responds to or why it's important for *Azospirillum* sp. B510 to know so much about its chemical environment.

The histidine kinase CheA is responsible for phosphorylating CheY. The activity of CheA is controlled by the MCPs, through an adaptor protein called CheW (Figure 15.10 B). Some signals increase CheA activity, and others decrease it, modulating the total amount of CheY-P present in the cell, and therefore integrating all of the individual chemotactic signals into a single output that controls how frequently the flagella switch to rotating CCW, and therefore the length of runs between tumbles.

ADAPTATION AND MEMORY IN CHEMOTAXIS

The second key question of chemotaxis is this: how do the cells remember what their previous environment was like?

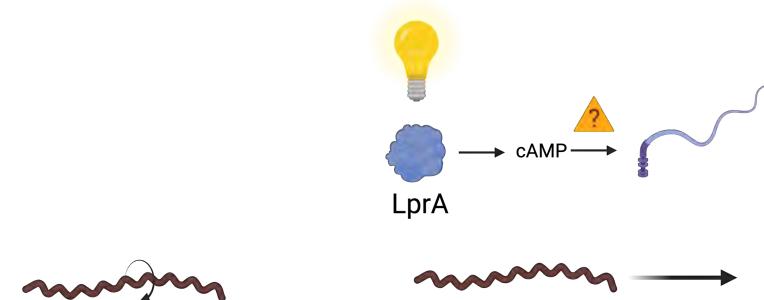
This memory or adaptation feature of chemotaxis is important to allow cells to continue moving along gradients across a very wide range of concentrations. The chemotaxis system continuously adjusts back to a baseline state, so that it is sensing **changes** in attractants or repellants, and not just their presence, absence, or a threshold concentration. This is where the methylation of MCPs comes into play.

The constitutively-active methyltransferase CheR adds methyl groups to MCPs, progressively reducing their ability to affect CheA activity (the OFF state) (Figure 15.10 B). This is counteracted by the activity of CheB-P, the phosphorylated form of a **second** response regulator that is phosphorylated by CheA. This means that activation of CheA in response to signals has **two** simultaneous effects: increasing tumbling (via production of CheY-P) **and** increasing the sensitivity of the MCP sensors (the ON state). CheB-P is rapidly autodephosphorylated, which means that, due to the constitutive activity of CheR, in the absence of activated CheA, the MCPs are constantly adjusted back towards the less-sensitive baseline OFF condition. The rate at which this occurs is responsible for the length of *E. coli*'s memory, which, as I mentioned above, is about 2 seconds. The net result is that *E. coli* cells can adapt to their current conditions and sense when those conditions change, no matter what the actual concentrations of attractants or repellants might be.

Question for in-class discussion: The O₂ sensor Aer is not methylated. What effect do you expect this to have on aerotaxis (movement towards O₂), and why might this have been selected for?

EXAMPLE OF AN EXPERIMENTAL DESIGN: REGULATION OF SPIROCHETE MOTILITY

Leptospira kobayashii is a non-pathogenic spirochete isolated from soil in Japan with an [interesting motility phenotype](#). In the dark, *L. kobayashii* rotates in place without moving but when exposed to light, it swims smoothly and rapidly, going faster at higher light intensities. This phenotype depends on a protein called LprA, which is a light-responsive cyclic AMP synthase that localizes near the poles of the *L. kobayashii* cell (i.e. near the flagellar basal bodies).



Made with Biorender.

According to this model, LprA-catalyzed cyclic AMP (cAMP; **Lecture 4**) production activates swimming motility, but the molecular mechanism by which cAMP affects flagellar rotation is not known. Does cAMP directly modulate the activity of the motor proteins? Or are there other regulatory proteins (perhaps homologs of the chemotaxis machinery) that conduct that signal to the flagellar motor?

One useful observation from the original paper describing this phenomenon is that exogenous cAMP (cAMP added to the growth media) can activate swimming even in the dark and even in an *lprA* mutant. This means we can separate the cAMP-dependent phenotype from the light activation of LprA.

The following methods are available for *L. kobayashii*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓

complete genome sequence (4.2 Mbp)	✓
susceptible to mutagens	✓
shuttle vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓

Design an experiment to identify proteins involved in cAMP-dependent regulation of flagellar activity in *L. kobayashii*.

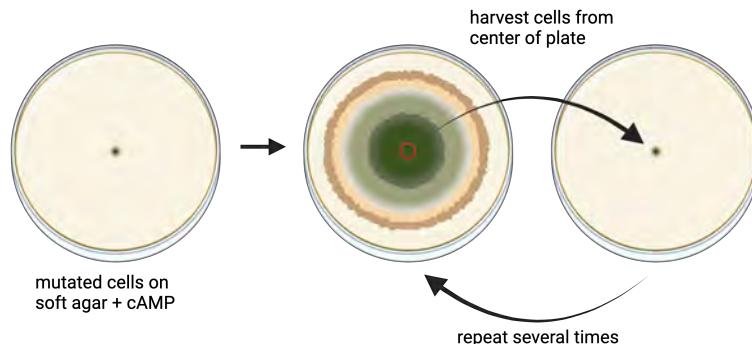
Hypothesis: There is a cAMP receptor protein that modulates swimming motility in *L. kobayashii*.

Experimental Design:

The mutants we're looking for will be non-motile in the presence of exogenous cAMP. However, we're not really interested in mutants that have simply lost the ability to make flagella at all, so we need to have a way to eliminate those. The challenge here is that the only way to really look for that is through the microscope, where strains without flagella will not rotate in place (as we expect interesting cAMP non-responders to do). Therefore, we will implement an enrichment to reduce the amount of screening we need to do.

It's reasonably likely that cAMP works by allosterically regulating a central component of the flagellar motor or basal body, so transposon mutagenesis (which results in the generation of gene knockouts) is not likely to be an effective strategy here. This means we need to use a random mutagen to make point mutations. We're hoping to disrupt cAMP signaling specifically, without destroying the entire motility machinery.

- 1) Treat *L. kobayashii* wild-type cells with a chemical mutagen (say, nitrosoguanidine).
- 2) Spot 10^8 mutated *L. kobayashii* at the center of a plate containing low-percentage agar (to allow motility) and incubate in media containing cAMP, allowing the cells to grow and spread.



Made with Biorender.

- 3) Harvest cells that remain at the center of the plate, which will be enriched for non-motile mutants, regrow, and repeat step 2 several times. This will progressively increase the proportion of mutants in the pool which are unable to swim.
- 4) Isolate individual mutants from the enrichment and screen under the microscope in media with cAMP added. Eliminate those which do not rotate (and any remaining strains that are still able to swim).
- 5) Sequence the genomes of multiple cAMP non-responder mutants, determining the point mutations present in those strains and looking for specific mutations or mutated genes that occur frequently in those strains. These are candidates for the cAMP receptor protein(s) which will then warrant further study and confirmation.

Independent Variable: (what will you change?)

- The point mutations present in the genome of each strain.

Dependent Variable(s): (what will you measure?)

- 1) Swimming motility in the presence of cAMP (a qualitative measurement)
- 2) Ability to rotate without moving (a qualitative measurement)

Negative Control(s): (eliminate false positive results)

- A mutant lacking flagellin will not be able to swim or rotate under any conditions.

Positive Control(s): (eliminate false negative results)

- 1) The wild-type will rotate in place in the dark and swim in the light or when given cAMP.
- 2) An *lprA* mutant will rotate in place in the light, but swim when given cAMP.

Potential Outcomes:

- 1) You identify point mutations that confer a cAMP non-responder phenotype. This supports the hypothesis and will give you a list of candidate genes and amino acid residues that may be involved in this process for further study.
- 2) You do not identify point mutations that confer a cAMP non-responder phenotype. If such mutants are very rare or selected against in some way, identifying them by microscopic screening of non-motile strains may be very difficult, even after enrichment. This is a significant limitation of this approach, and not being able to find mutants does not necessarily invalidate the hypothesis.

Does this experiment establish correlation or causation?:

- Since we have not fulfilled Falkow's postulates for any of these mutations, this design tests **correlation**..

If this mutant hunt is unsuccessful, there are other ways to approach this kind of problem, but they are more model-driven and have their own failure modes. For example, since we know that LprA is localized to the cell poles, where the flagellar basal bodies are, we might consider attaching a epitope tag to LprA and isolating proteins that co-purify from *L. kobayashii* with LprA, on the theory that LprA might form a complex with the relevant cAMP receptor protein. We could also try localized mutagenesis of the known components of the flagellar basal body or motor to see if that allows us to identify cAMP non-responders. Either of those approaches might work, if those models happen to be correct, but I like the idea of a non-biased, genome-wide search.

It's also probably possible to biochemically isolate cAMP-binding proteins from a cell lysate of *L. kobayashii* (maybe using a column with cAMP attached to the resin), which would presumably include (among other things) the receptor relevant to this phenotype. That could end up being a fairly large list of proteins, though, and sorting out which one(s) are involved in motility might not be trivial.

ACTIN POLYMERIZATION

None of the mechanisms of motility described above are able to function inside of host cells. Several different intracellular pathogens have independently evolved the ability to hijack the eukaryotic cytoskeleton to drive motility within the host cytoplasm (Figure 15.11). They do this by expressing surface proteins that mimic host actin nucleation factors, causing actin to polymerize around the bacterial cell and push it rapidly through the host cytoplasm. This is thought to facilitate the spread of pathogens from one host cell to another without exposing themselves to the host immune system. Bacteria that move by actin polymerization move quickly within host cells, but apparently at random, with no known mechanisms of steering or chemotaxis.

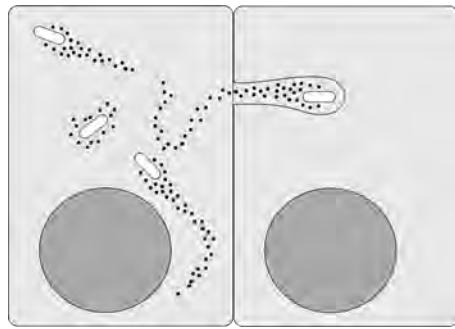


Figure 15.11 Intracellular bacteria moving within and between eukaryotic host cells by actin polymerization. Actin, indicated by black dots, forms comet-like clouds and tails as it polymerizes around and behind bacteria. Bacterial cells that push into neighboring host cells are generally able to lyse the layers of host cell membrane with lipases to release themselves into the cytoplasm of the second cell.

The classic examples of pathogens that move by actin polymerization are *Listeria monocytogenes* and *L. ivanovii*, which cause listeriosis in humans and ruminants, respectively, and *Shigella flexneri*, one of the causative agents of dysentery. However, actin-based motility is also found in *Burkholderia pseudomallei* and *B. mallei*, which cause melioidosis in humans and glanders in horses, respectively, as well as in *B. thailandensis*, a pathogen of fruit flies, in the fish pathogen *Mycobacterium marinum*, and in many species of *Rickettsia*, including those that cause spotted fevers and typhus. Each of these groups expresses different,

unrelated proteins that lead to actin polymerization (e.g. ActA in *Listeria*, BimA in *Burkholderia*, and IcsA in *Shigella*), and indeed, the mechanism by which *M. marinum* stimulates actin polymerization remains unknown.

DISCUSSION PROBLEM SET #27: INTRACELLULAR MOTILITY IN *MYCOBACTERIUM MARINUM*

M. marinum causes skin ulcers and is an economically important pathogen of fish, but unlike other pathogenic mycobacteria, is highly motile inside the host cell cytoplasm. This motility is actin-dependent, and actin tails can be seen forming around motile *M. marinum* cells in appropriately stained host cells, but the genome of *M. marinum* does not encode homologs of any known bacterial or mammalian proteins that stimulate actin polymerization.

The following methods are available for *M. marinum*, which has gained some popularity as a model mycobacterium that grows considerably faster than *M. tuberculosis* and is able to infect zebrafish, amoebas, and cultured human macrophages (as well as, occasionally, whole humans):

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (6.5 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
inducible promoter known	✓
compatible transposon	✓
CRISPR-assisted recombineering	✓

Design an experiment or series of experiments to identify factors specifically required for driving or regulating actin-dependent motility in *M. marinum*. State:

- a model to explain this phenomenon
- a hypothesis to test that model
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiments, and how you will interpret them

FLOATING

Several different types of aquatic bacteria are able to adjust their buoyancy by producing gas vesicles, allowing them to float upward towards oxygen (for aerobic heterotrophs) or to a particular depth where the light intensity and wavelength is optimal (for phototrophs) (Figure 15.12). This preferred depth varies among different organisms, and therefore floating allows photosynthetic bacteria stratify into layers in the water column.

Gas vesicles are hollow gas-filled shells (or organelles) made up of a few protein components; largely the major protein GvpA (for **g**as **v**esicle **p**rotein) and the structural strengthening protein GvpC. They are typically small, biconical structures that assemble into large, hexagonally-packed gas vacuoles, which can take up very large proportions of the cytoplasmic volume. Gasses diffuse into the vesicles from the cytoplasm, so the gas present in a vesicle will depend on the gasses present in a cell's environment or produced by its metabolism.

The genes for gas vesicle production in cyanobacteria, halophilic archaea, and other types of bacteria are homologous, indicating that they evolved once and have been spread by extensive horizontal gene transfer during the course of evolution.

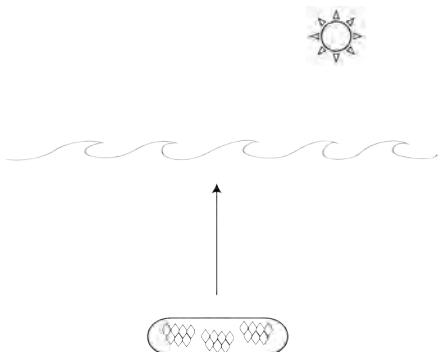


Figure 15.12. Gas vesicles providing buoyancy to an aquatic bacterium, allowing it to move upwards towards oxygen and/or light.

Mikhail Shapiro's lab at Cal Tech has developed a system to express gas vesicles in *E. coli* and other species where they are not normally present as an acoustic reporter gene to track bacteria inside living human or animal hosts. If enough bacteria are present, gas vesicles can be detected by non-invasive ultrasound methods, and they've shown that visualizing bacteria this way can give 3-dimensional images with resolution down to 100 µm in mice. Read more [here](#).

SYMBIOTIC MOTILITY INTERACTIONS

Even bacteria that do not produce their own motility machinery can sometimes take advantage of other motile bacteria to move in their environments. There are recent data from a couple of different research groups showing that non-motile *S. aureus* can adhere to and hitchhike on motile *E. coli* or *P. aeruginosa* cells (Figure 15.13 A), and that this can contribute to the spread of *S. aureus* during polymicrobial infections. [This review](#) discusses this phenomenon in more detail.

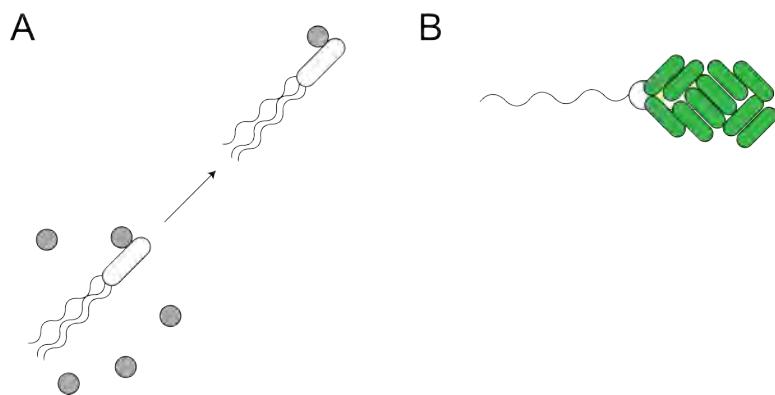


Figure 15.13. (A) Non-motile *S. aureus* (grey circles) can hitchhike on motile *E. coli* or *P. aeruginosa* cells. (B) Photosynthetic *C. chlorochromatii* (green) in close symbiotic association with a motile *Candidatus S. mobilis* cell.

In a more complex example, the photosynthetic green sulfur bacterium *Chlorobium chlorochromatii* forms a close symbiotic relationship with the motile heterotroph *Candidatus Symbiobacter mobilis* (a consortium originally called *Chlorochromatium aggregatum*) (Figure 15.13 B). Between 20 and 70 cells of *C. chlorochromatii* attach to a central *Candidatus S. mobilis* cell, which uses its flagellum to propel the consortium.

DISCUSSION PROBLEM SET #28: PHOTOTAXIS BY A BACTERIAL CONSORTIUM

The *Chlorobium aggregatum* consortium shown in Figure 15.13 B is, as a whole, able to swim towards light, and if it enters a region of darkness, is able to reverse direction back into the light. As described above, however, only *Candidatus Symbiobacter mobilis* has flagella.

As you can tell from the prefix *Candidatus*, the motile partner cannot be grown in pure culture, and requires its photosynthetic partner. The reverse is not true, and *C. chlorochromatii* can be grown on its own.

The following methods are available for each species in this consortium (or at least for closely related species):

	<i>C. chlorochromatii</i>	<i>Candidatus S. mobilis</i>
growth in pure culture	yes	no
can extract DNA/RNA/protein	yes	yes

complete genome sequence	yes (2.6 Mbp)	yes (3 Mbp)
susceptible to mutagens	yes	yes
can be made competent	yes	no
shuttle vectors available	yes	no
inducible promoter known	yes (light-activated)	no
compatible transposon	yes	no

Design an experiment or series of experiments to determine how the two partners in this consortium coordinate phototaxis towards light conditions optimal for *C. chlorochromatii* photosynthesis. State:

- a model to explain this phenomenon
 - a hypothesis to test that model
 - the independent and dependent variables of each experiment
 - both positive and negative controls for each experiment
 - a description of how you will construct any necessary strains and plasmids
 - whether your proposed experiment(s) test correlation or causation
 - potential outcomes of your experiments, and how you will interpret them
-

LECTURE 16: BACTERIAL ENERGETICS

INTRODUCTION

Living cells must do a substantial amount of work, both physical (e.g. flagellar rotation or pilus retraction) and chemical (e.g. catalyzing endergonic reactions), and both of these require the input of energy. This chapter is about the forms that energy takes and, to some extent, where that energy comes from. We will discuss the *chemiosmotic theory* and how the *proton motive force* is generated and provides energy for many cellular processes. We will also discuss *high-energy phosphate bonds* and how nucleotide phosphates provide the other main source of energy to drive chemical reactions in the cell. Finally, we will cover respiration, fermentation, and photosynthesis, three mechanisms by which bacteria generate and conserve energy.

I suspect much of the core content of this chapter and the next chapter on central metabolism will be review of material you've covered in your undergraduate biochemistry classes. I'll cover the basics, but will try to focus on illustrating aspects of these topics that are specific to prokaryotic systems, which have considerable diversity.

PROTON MOTIVE FORCE

The *chemiosmotic theory* describes how potential energy is stored in biological systems as an electrochemical gradient of ions across a lipid bilayer membrane. Since the ions involved are usually protons (H^+), this energy is also called the *proton motive force* (PMF), and is the sum of the force derived from the difference in **concentration** in ions on either side of the membrane and the force derived from the difference in **charge**. Any molecule will tend to diffuse towards a region of lower concentration, and charged molecules will tend to move towards an area of opposite charge.

Lipid bilayer membranes are **not** permeable to charged ions, and the PMF in bacteria is due to cells actively maintaining a higher proton concentration and / or positive charge **outside** of the cell than **inside**:

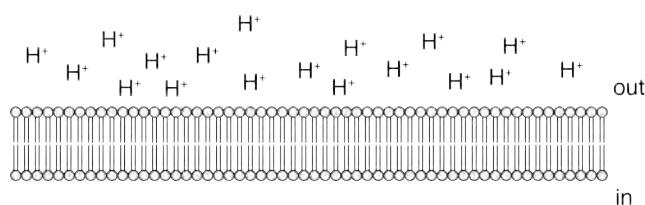


Figure 16.1. The proton motive force is a combination of the force derived from the difference in proton concentration on either side of a membrane and the force from the difference in charge (positive outside, negative inside).

The forces involved can be expressed mathematically, although we will not be worrying about calculating the exact amounts of energy involved in any given system in this course:

$$PMF = \Delta p = \Delta\Psi + \Delta pH$$

The PMF (or Δp) is equal to the difference in charge ($\Delta\Psi$) plus the difference in proton concentration (ΔpH). In bacteria living at neutral pH, the ΔpH across the membrane contributes about 70-80% of the PMF, but in bacteria that live in very high or very low pH environments the relative contributions change. In *alkaliphilic bacteria* living at pH 10 or 11, for example, membranes must maintain a very high $\Delta\Psi$ to overcome the negative contribution of ΔpH to the PMF.

Another thing that is good to know is that many bacteria also maintain a **sodium** ion gradient in addition to the PMF. Those Na^+ ions can do all of the same kinds of work that protons can, but require specialized protein components to do so (i.e. a transporter that is driven by the PMF cannot derive energy from the Na^+ gradient and vice versa). [This review](#) goes into more detail about the sodium motive force (SMF), including a discussion of data that suggests that the SMF may predate the PMF evolutionarily.

The potential energy of the PMF is used by bacteria to drive a variety of biological processes, as we will see below, but the first issue to address is how the PMF is generated.

HOW THE PMF IS GENERATED

Bacteria use a variety of mechanisms to generate a PMF, with the relative importance of different mechanisms varying among species. We will discuss some examples here (as illustrated in Figure 16.2). This is by no means an exhaustive list, and we will, in fact, cover some additional pathways that generate PMF later in this chapter. What they all have in common is that protons are removed from the cytoplasm and transported, by one means or another, across the cell membrane. They use different sources of energy to accomplish this goal. Remember that the PMF is a way of **storing** potential energy, and therefore creating it requires an energy input.

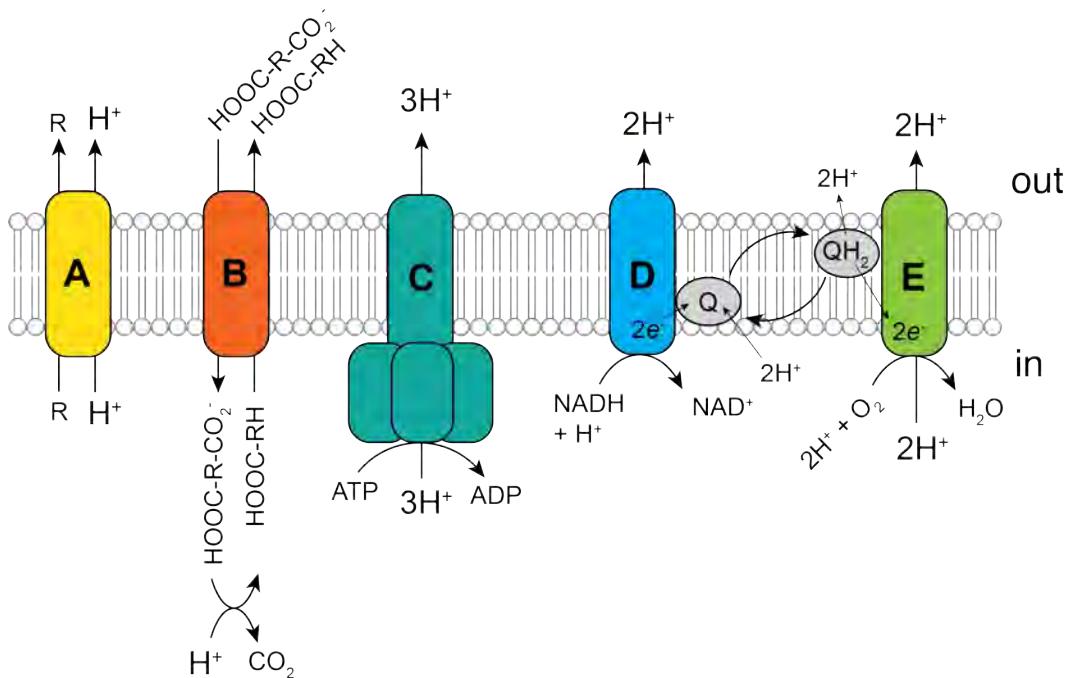
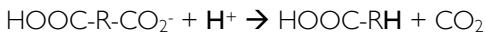


Figure 16.2. Representative mechanisms by which bacteria generate a PMF. (A) Symport of protons with fermentation end products. (B) Coupling decarboxylation of a dicarboxylic acid to antiport. (C) Proton pumping by the F_1F_0 ATPase. (D) Proton pumping and quinone reduction by NADH:quinone oxidoreductase. (E) Proton pumping and reduction of oxygen by cytochrome quinol oxidase.

Figure 16.2A illustrates a very simple mechanism by which some bacteria generate a PMF. If there is a concentration gradient of some other compound (typically a fermentation end product like lactate; see below) with that compound much more concentrated inside the cell than outside, the potential energy of **that** gradient can drive the export of protons through a **symporter** (a transporter that transports two molecules across a membrane in the same direction). This depends, of course, on maintaining the gradient of the other compound. Organisms that use this mechanism to generate a PMF usually rely on other species in their environment to rapidly degrade that compound. This is a kind of *syntrophy* or metabolic symbiosis.

Figure 16.2B illustrates a less common mechanism for generating a PMF, but it's one that demonstrates an important point. The decarboxylation of a dicarboxylic acid consumes a proton and releases CO_2 :



In combination with an **antiporter** (a transporter that transports two molecules across a membrane in opposite directions) that links import of the dicarboxylic acid to export of the decarboxylated product, the net result is the loss of a proton from the cytoplasmic side of the membrane. Removing protons from the cytoplasm generates PMF even without any proton transport *per se*, since the PMF results from the **difference** in proton concentration on either side of the membrane and, as mentioned above, the difference in charge. The result of this mechanism is to remove a negative charge from outside the cell and remove a positive charge from inside the cell, therefore impacting both $\Delta\Psi$ and ΔpH .

Figure 16.2C shows a very common mechanism by which bacteria generate a PMF using the reversible proton-pumping F_1F_0 ATPase (a large protein complex encoded by the *atpCDEFG* operon). This complex can use binding and hydrolysis of a molecule of ATP (see below) to power the export of 3 protons across the membrane. As we will see shortly, this is a reversible reaction and the same protein complex can use the import of 3 protons to power the synthesis of a molecule of ATP. Which direction the F_1F_0 ATPase runs in a particular organism depends on its metabolic needs at any given time. If it has sufficient ATP but insufficient PMF, then it will run in the proton-pumping direction.

Finally, Figures 16.2D and E illustrate **respiration**, a mechanism by which bacteria link the generation of a PMF to oxidation-reduction reactions by way of an **electron transport chain**. The particular form of respiration illustrated in Figure 16.2 is aerobic respiration, powered by the reduction of oxygen to water, which yields a very large amount of energy. The electron transport chain depends partly on membrane-soluble electron shuttles called **quinones** to move electrons and protons between respiratory enzymes. We will discuss respiration in more detail below, but note that oxidation of one NADH and reduction of one O_2 during aerobic respiration results in direct transport of 6 protons across the membrane and consumption of 2 more in the cytoplasm. Respiration is an extremely potent mechanism for generating PMF and is used by most fast-growing organisms.

HOW THE PMF IS USED

The PMF is used to power a variety of important processes in bacteria. Some examples are shown in Figure 16.3:

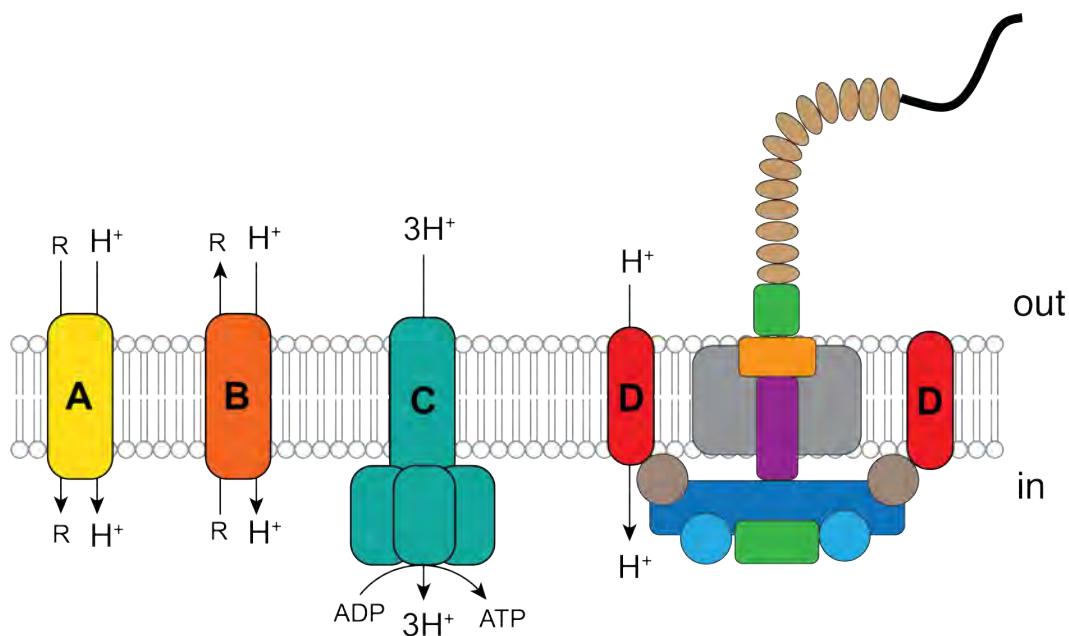


Figure 16.3. Examples of how the PMF is used to power bacterial metabolism. (A) Symporters couple the import of a proton to the simultaneous import of another molecule. (B) Antiporters couple the import of a proton to the simultaneous export of another molecule. (C) F₁F₀ ATP synthase couples the import of 3 protons to the synthesis of ATP. (D) The MotAB motor proteins couple the import of protons to the rotation of flagella.

The transport of small molecules across the cell membrane is often powered by the PMF. As shown in Figure 16.3A and B, H⁺ symporters couple the import of a proton to the **import** of another molecule, while H⁺ antiporters couple the import of a proton to the **export** of another molecule. In *E. coli*, for example, H⁺ symporters are responsible for the import of many amino acids, sugars, nucleotides, metals, and a variety of carbon and nitrogen sources ([link](#)), while PMF-driven antiporters catalyze the export of various ions, toxins, antibiotics, metabolic byproducts, and metals ([link](#)).

As I mentioned above, the F₁F₀ ATPase is reversible, and in fact, its primary function in respiration bacteria is to use the PMF to generate ATP (Figure 16.3C). When operating in this direction, it is often called ATP synthase. We will discuss how ATP and other nucleotide phosphates are used to power bacterial processes in the next section.

As a final example, the MotAB motor proteins use the PMF to power rotation of flagella ([Lecture 15](#)). This is a major drain on the PMF in motile bacteria, with more than 500 protons needed to power one rotation of a single flagellum in *E. coli*.

See [Lecture 13](#) for some additional examples of PMF-powered systems in bacteria, and remember that any of these kinds of PMF-powered protein complexes can, in some species, be driven by a Na⁺ gradient. Na⁺ gradients are generated by different mechanisms in different species, but Na⁺ / H⁺ antiporters are common, making the generation of the Na⁺ gradient dependent on the PMF. Chemiosmotic theory suggests that any ion gradient could, in principle, be used to store energy. There are known examples of flagellar motors powered by K⁺, Ca²⁺, and Mg²⁺ gradients (in some *Bacillus* and *Paenibacillus* spp.).

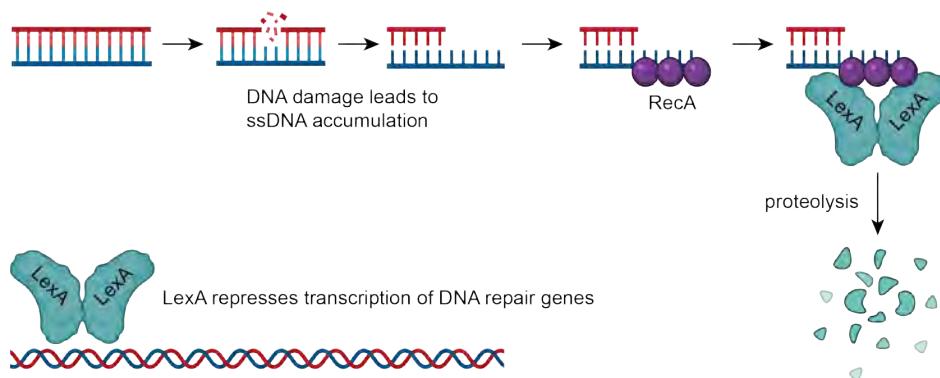
Experimentally, the addition of **ionophores** or **uncouplers**, chemicals that bind to cations and allow them to diffuse through membranes, can be used to collapse ion gradients. This can be useful to determine whether the PMF is the driving energy source for a particular process, and since there are ionophores specific for particular cations, they can also be used to determine if H⁺, Na⁺, or some other cation gradient is involved. Carbonyl cyanide m-chlorophenyl hydrazone (CCCP) is a commonly used proton-specific ionophore, and the antibiotic monensin (produced by *Streptomyces cinnamomeus*) is a sodium ionophore. Gramicidin, the first commercially manufactured antibiotic (circa 1939), is an antimicrobial peptide produced by *Brevibacillus brevis* that forms channels in bacterial cell membranes permeable to both H⁺ and Na⁺, thereby acting as a general-purpose uncoupler. Oligomycin A (produced by *Streptomyces diastatochromogenes*) is an inhibitor of ATP synthase, which can also be useful in these kinds of experiments. All of these compounds can be extremely toxic to human cells, so care needs to be taken when using them in the lab.

DISCUSSION PROBLEM SET #29: RESPONDING TO STRESS BY COLLAPSING THE PMF

Stress response is the general term for pathways that organisms use to respond to harmful changes in their environments, from increased temperature or dehydration to accumulation of reactive or otherwise toxic chemicals. There are *general stress response* pathways, that are activated under multiple stress conditions, as well as more specific responses to particular stresses.

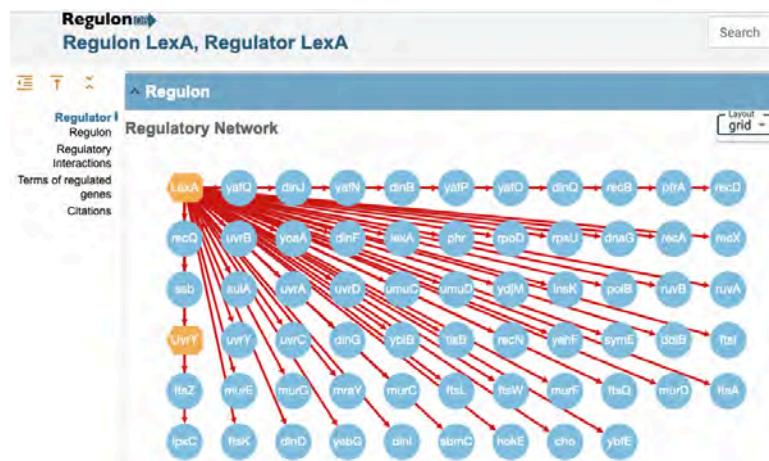
The bacterial *SOS* response, for example, is a well-characterized stress response pathway conserved among many bacteria that responds to DNA damage or inhibition of DNA replication. It was the first stress response whose molecular mechanism was deciphered in work pioneered by Evelyn Witkin and Miroslav Radman starting around 1971.

The SOS response depends on the fact that many different kinds of DNA damage cause single-stranded DNA (ssDNA) to accumulate in the cell. Single-stranded DNA is bound by RecA ([Lecture 8](#)), and the RecA-ssDNA complex is bound by the LexA transcriptional repressor, leading (through an allosteric mechanism; [Lecture 4](#)) to proteolytic self-cleavage of LexA and derepression of the LexA regulon.

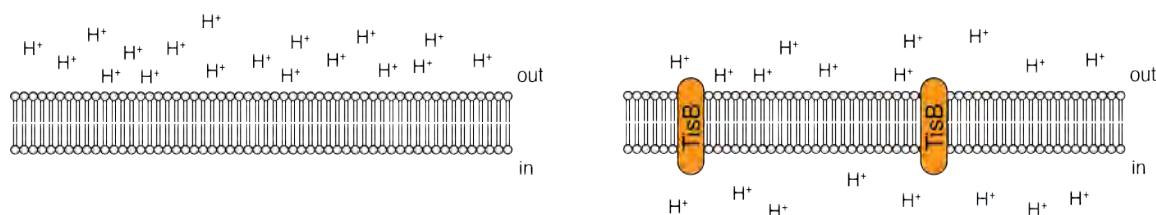


The LexA-dependent bacterial SOS response to DNA damage. Made with Biorender.

LexA controls the transcription of a large number of genes in *E. coli* (as illustrated in this figure from [RegulonDB](#)):



Most of these gene products are directly involved in recombination or DNA damage repair. However, one of the most strongly-induced genes in the LexA regulon encodes a small protein called TisB, which has a very different function: TisB inserts into the cytoplasmic membrane, depolarizing it and eliminating the proton gradient that underlies the PMF.



What isn't so obvious is **why** TisB is induced as part of the SOS response. What advantage could collapsing the PMF have in the presence of DNA damage?

Propose a model for the physiological role of TisB in DNA damage response. Design an experiment to test your model. State:

- a model to explain this phenomenon
- a hypothesis to test that model
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiments, and how you will interpret them

All genetic and molecular biology methods are available for *E. coli*. The details of how exactly to measure PMF and different kinds of DNA damage are beyond the scope of this course, but for the sake of this problem, assume that you can quantify those things however you'd like.

HIGH-ENERGY PHOSPHATE BONDS

The PMF is a useful source of energy for processes that take place at the cell membrane. The main source of energy in the **cytoplasm** of cells, however, is the hydrolysis of nucleotide triphosphates (NTPs), usually adenosine triphosphate (ATP). The same kinds of chemical energy can be derived from other NTPs, and it is fairly common for certain specific enzymes to use GTP instead, but ATP is the most commonly used, and the one we will focus on here.

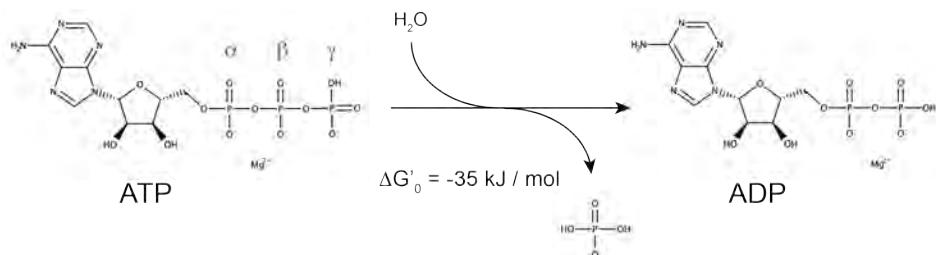


Figure 16.4. The structure of ATP, and its hydrolysis to ADP and P_i. This reaction is exergonic, and release of the γ phosphate results in the release of 35 kJ / mol of free energy.

As shown in Figure 16.4, ATP has three phosphate groups (the α, β, and γ phosphates, in order outward from the ribose sugar of the adenine nucleoside). The bonds between these phosphate groups are often referred to as "high-energy phosphate bonds", which is a misnomer. The actual chemical energy of these bonds is the same as for any other phosphate bond. What is notable about the phosphate bonds in ATP and other NTPs is their high phosphoryl **group transfer potential**. We will not get into the detailed chemistry here, but essentially this means that hydrolysis of these bonds releases an unusually large amount of free energy (35 kJ / mol). This is largely due to the fact that the negative charges of the sequential phosphate groups repel each other, so that hydrolysis reduces the electrostatic strain within the molecule.

Note also in Figure 16.4 the presence of a Mg²⁺ ion chelated by the phosphates of ATP and ADP. *In vivo*, essentially all NTPs are chelated to magnesium, and *in vitro* all ATP-dependent enzymes require both ATP and Mg²⁺.

How do enzymes access the free energy of ATP hydrolysis to drive chemical reactions? Some examples of ATP-driven group transfer reactions are shown in Figure 16.5:

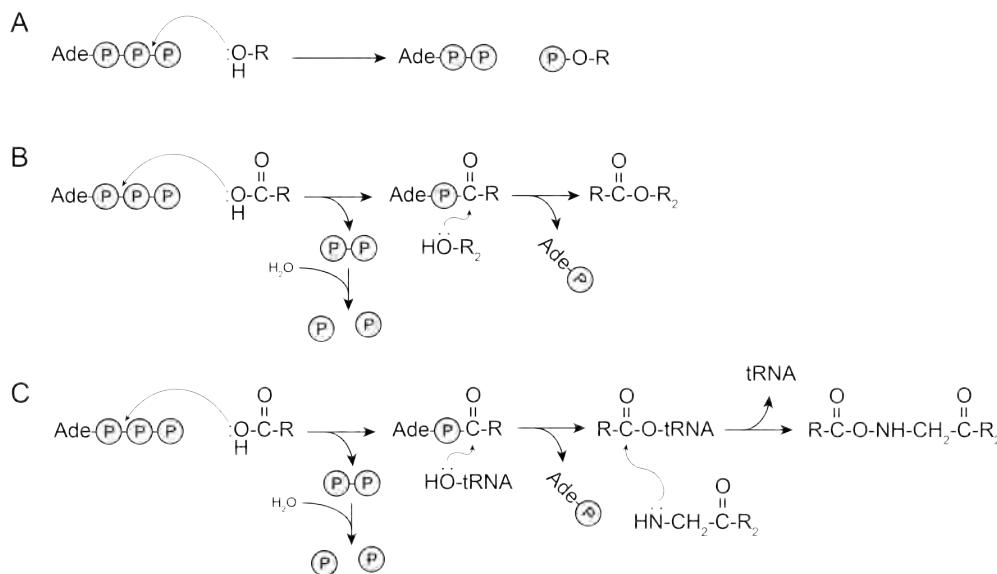


Figure 16.5. Chemical reactions using ATP (Ade-P-P-P). (A) Kinases catalyze the group transfer of the γ phosphate of ATP, resulting in ADP and a phosphorylated product. (B) Ester bonds are formed by enzymes that catalyze the attack of a carboxyl group on the α phosphate of ATP, releasing pyrophosphate (PP_i) and forming an AMP derivative. Hydrolysis of PP_i by a pyrophosphatase provides additional free energy to drive the reaction to completion. The AMP is displaced by a hydroxyl group, releasing AMP and the ester product. (C) Formation of a peptide bond, a specific and physiologically important example of the kind of group transfer reaction shown in (B).

Conceptually, the simplest group transfer from ATP is phosphorylation (Figure 16.5A). Enzymes that catalyze phosphorylation reactions are called *kinases*, and simply direct the attack of a hydroxyl group in their substrate to the γ phosphate of ATP, generating a phosphorylated product and releasing ADP. This is exactly what ATP hydrolysis is, when the R-OH substrate molecule is H_2O (Figure 16.4).

Group transfer reactions can also use ATP to generate other kinds of chemical bonds (Figure 16.5B). In these cases, enzymes catalyze the attack of a carboxyl group on the α phosphate of ATP, releasing pyrophosphate (PP_i) and generating an AMP derivative of the substrate. Hydrolysis of PP_i by a *pyrophosphatase* contributes additional free energy to drive the reaction to completion and make it irreversible. Attack of a hydroxyl, amino, or sulfhydryl group on the AMP-derived product releases AMP and results in ester, amide, or thioester bonds, respectively.

Figure 16.5C shows a very important example of chemical bond formation energized by ATP. The formation of a *peptide bond* in the active site of the ribosome depends on a series of group transfer reactions, first between ATP and an amino acid, which is transferred to a tRNA, then between the tRNA-derived amino acid and the growing protein chain. Protein synthesis is one of the main consumers of ATP in a growing bacterial cell. (Ribosomes are not the only enzymes that can catalyze the formation of peptide bonds. We will discuss others in **Lecture 18**.)

ATP is also important for controlling the activity of many proteins in the cell, even ones that are not enzymes catalyzing biosynthetic reactions. To illustrate, one important example is the ABC (**ATP-binding cassette**) transporters, a very widely conserved family of proteins that are responsible for import and export of diverse substrates across membranes, including both small molecules and proteins. (ABC transporters are related to type I protein secretion systems; **Lecture 13**.)

Figure 16.6 illustrates a generic ABC importer:

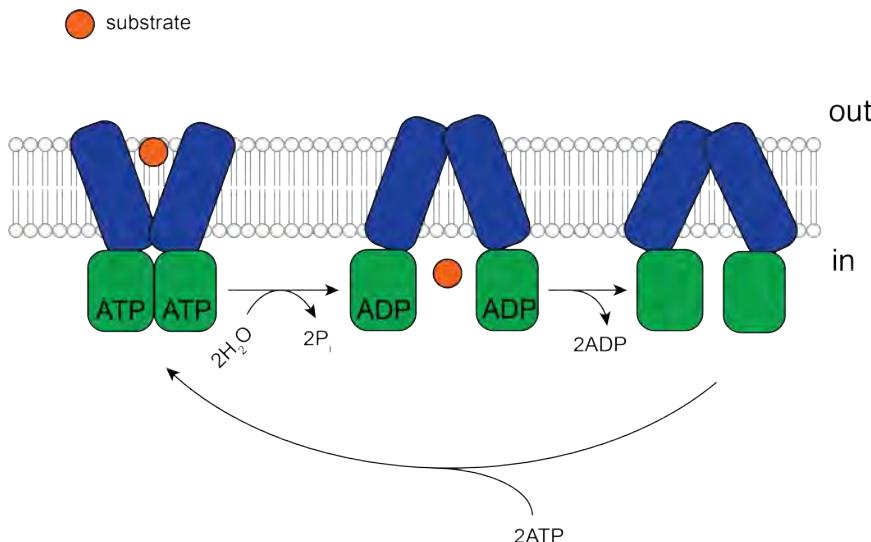


Figure 16.6. The mechanism of an ABC transporter; illustrating how ATP hydrolysis can be linked to protein conformational change.

ABC transport complexes consist minimally of dimers of two proteins: a transmembrane protein and a cytoplasmic nucleotide-binding protein. These form a complex in the membrane which can take two conformations. In one, a cavity in the complex faces **outward** and is open to the periplasm, while in the other, the cavity faces **inward** and is open to the cytoplasm. ATP binding to the nucleotide-binding protein changes the conformation of the transporter complex to face outward, allowing the substrate to enter the cavity. The nucleotide-binding protein then catalyzes hydrolysis of ATP to ADP, and the ADP-bound form of the complex changes conformation to face inward, releasing the substrate into the cytoplasm. The ADP is released and a new ATP is bound, allowing the cycle to continue.

E. coli has about 70 different ABC transport complexes, responsible for the import of various amino acids, monosaccharides, iron, autoinducer-2, vitamin B₁₂, and other nutrients and for the export of antibiotics and toxic metabolites like putrescine in addition to TISS-dependent protein secretion. See [this review](#) for more information. The repertoire of ABC transporters is species-specific and varies widely between strains, but *E. coli*'s number is not unusual among bacteria. ABC transporters are found in all three domains of life. It is not possible to predict the exact substrate(s) of an ABC transporter from sequence alone, although you may be able to determine the general class of molecule it is likely to transport (sugars, amino acids, etc.).

Somewhat surprisingly, it's **not** the free energy of ATP hydrolysis that provides the energy for the conformational change in ABC transporters, but rather the energy of ATP **binding**. Hydrolysis simply converts ATP, which binds tightly to the nucleotide-binding protein, to ADP, which binds very weakly. This mechanism is thought to be widely conserved among proteins that use ATP hydrolysis to control conformational changes, at least among the many that contain homologs of the Walker A and Walker B ATP-binding motifs found in ABC transporters.

AEROBIC RESPIRATION

As we have seen, the PMF is a very important energy source for bacteria, and, when working in concert with the F₁F₀ ATP synthase, is a major driver of ATP synthesis. The most efficient mechanism for generating PMF is to link proton pumping to oxidation-reduction reactions and electron transport chains, by the process of respiration. We introduced this process in Figure 16.2 above, and in this and the following section we will explore bacterial respiration in somewhat more detail.

In electron transport, electrons flow through a series of electron carriers down an energy gradient, so that they move from molecules with low electrode potentials to those with relatively higher electrode potentials. This means that electrons move from **electron donors** (which become oxidized) to **electron acceptors** (which become reduced). The electrical current generated can then be linked to generation of a PMF at specific points in the electron transport chain called **coupling sites**.

A very common electron donor for respiration is the reduced nicotinamide cofactor NADH. We will discuss how cells synthesize NADH in [Lecture 17](#). The NAD⁺ / NADH redox pair has a standard electrode potential of -320 mV. During aerobic respiration, the **terminal electron acceptor** of the electron transport chain is O₂, which is reduced to H₂O, a redox pair with a standard electrode potential of +815 mV. This electron transport chain therefore has a redox potential of (ΔE_h) of 1.135 V, which is energy that can be used to extrude protons from the cytoplasm at coupling sites. The O₂ / H₂O redox pair has a very high electrode potential, and aerobic respiration is therefore an extremely efficient source of energy. We will discuss other terminal electron acceptors in the next section.

Eukaryotic mitochondria (which are descended from bacteria, of course) have a single pathway for aerobic respiration which generates a constitutively very high PMF for ATP synthesis. You are probably familiar with this pathway from your pre-

vious classes. Bacteria have additional pathways, and in fact, many of them have **branched** electron transport chains that allow them to adjust the flow of electrons and pump different numbers of protons depending on which coupling sites they express under a given set of conditions. Figure 16.7 shows the options possessed by *E. coli*:

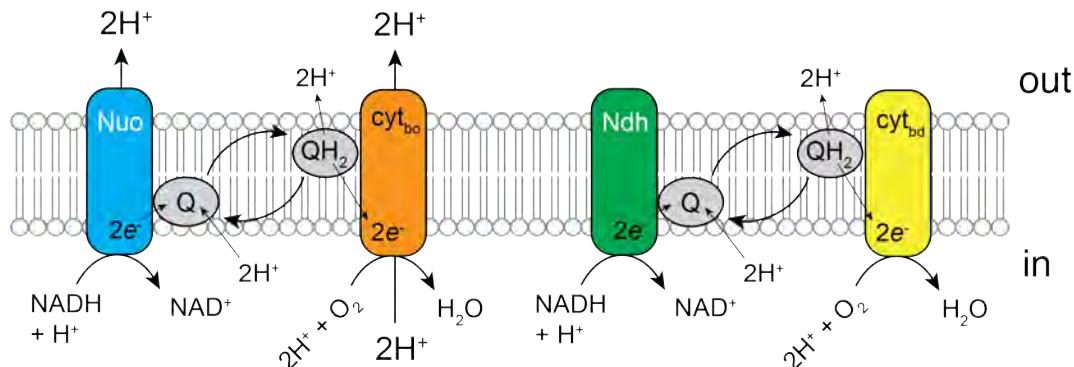


Figure 16.7. Complexes involved in aerobic respiration in *E. coli*, including the NADH:quinone oxidoreductases Nuo (NDH-1) and Ndh (NDH-2) and the cytochrome bo and bd quinol oxidase complexes (encoded by the *cyo* and *cyd* operons, respectively). Ndh and cytochrome bd do not pump protons. Cytochrome bo has a low affinity for O_2 , while cytochrome bd has a very high affinity for O_2 . Quinones reduced by Nuo or Ndh can be oxidized by either cytochrome bo or bd.

E. coli has NADH:quinone oxidoreductase and cytochrome bo quinol oxidase complexes, encoded by the products of the *nuo* and *cyo* operons, respectively, that are coupling sites and pump protons across the membrane, as discussed above. The Q loop of reduced and oxidized quinones is also a coupling site. However, *E. coli* has an additional NADH:quinone oxidoreductase complex, encoded by the *ndh* operon, that is **not** a coupling site and an alternative cytochrome bd quinol oxidase, the product of the *cyd* operon, that is **also** not a coupling site. *E. coli* can mix and match among these complexes depending on its growth conditions, thereby tuning the amount of PMF generated by oxidizing an NADH and reducing an O_2 . Cytochrome bo (*Cyo*) has a low affinity for O_2 , and is therefore active at very high O_2 concentrations, while the higher-affinity cytochrome bd (*Cyd*) is active under microaerobic conditions. In general, *E. coli* uses Ndh during aerobic and nitrate respiration, and Nuo during fumarate respiration (see below).

Question for discussion in class: Why would a bacterium ever want to generate less than the maximum possible amount of PMF?

As usual, the pathways present in *E. coli* are not the only pathways found among bacteria more broadly, but the general pattern of oxidation of a reduced cofactor (NADH or $FADH_2$), the passage of electrons through a series of electron carriers (both membrane-soluble quinones and protein-bound flavin, iron-sulfur cluster, and cytochrome cofactors), some of which act as coupling sites to extrude protons to generate PMF, and the final reduction of O_2 to H_2O are conserved among all organisms capable of aerobic respiration.

O_2 and the various **reactive oxygen species** (ROS) derived from it are potent oxidizing agents that can be quite toxic, but the very high energy yield of aerobic respiration means that it is the energy-generation pathway of choice for many fast-growing bacteria, some of which are **obligate aerobes** that cannot grow **without** O_2 . Aerobic organisms have a variety of mechanisms for resisting and detoxifying ROS, regulated by a broad range of O_2^- - and ROS-sensing stress response proteins. In at least some bacteria, the elimination of toxic O_2 by aerobic respiration seems to be as important as its energy-generating function.

For a more detailed review of aerobic respiration in bacteria, see [this paper](#).

EXAMPLE OF AN EXPERIMENTAL DESIGN:AEROBIC RESPIRATION IN KLEBSIELLA PNEUMONIAE

Klebsiella pneumoniae is a Gram-negative, facultatively aerobic enterobacterium that is found as a commensal in the human intestine, but which can cause serious disease when it spreads to other parts of the body, which typically occurs in immunocompromised patients. This includes pneumonia (in adults with diabetes or chronic obstructive pulmonary disease, for example) and intestinal dysbiosis and sepsis (in premature infants). Mouse models of both these diseases exist.

Like *E. coli*, *K. pneumoniae* has both a low-affinity cytochrome bo (*Cyo*) and a higher-affinity cytochrome bd (*Cyd*) quinol oxidase dedicated to aerobic respiration, and, like *E. coli*, *K. pneumoniae* does not **require** oxygen to survive, at least under laboratory conditions. (It does grow much faster when oxygen is present, though.)

The following genetic tools are available for *K. pneumoniae*:

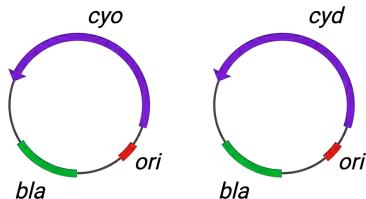
growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (5.5 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
generalized transducing phage (P1)	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

Design an experiment to determine the role of aerobic respiration in lung infections by *K. pneumoniae*.

Hypothesis: Aerobic respiration is required for lung colonization by *K. pneumoniae*.

Experimental Design:

- 1) Obtain *cyo*::T30(*cat*⁺) and *cyd*::T30(*cat*⁺) transposon mutants from the genome-wide knockout collection linked above.
- 2) Use P1 transduction to move the *cyo*::T30(*cat*⁺) and *cyd*::T30(*cat*⁺) transposon mutants from the knockout collection strain background into your virulent test strain of *K. pneumoniae*.
- 3) Using *in vivo* assembly cloning, construct plasmids for expression of *cyo* and *cyd* in *K. pneumoniae*.



Made with Biorender.

Note the use of an ampicillin-resistant plasmid (*bla*⁺), so as to be compatible with the chloramphenicol-resistant chromosomal markers. It will also be important to include a promoter to drive *cyd* / *cyo* expression. In this case, I would probably choose to clone the native promoter for each gene, to keep expression as close to wild-type levels as possible.

- 4) Construct complemented and vector-only control strains for each mutant:

1. wild-type + vector
2. *cyo*::T30(*cat*⁺) + vector
3. *cyo*::T30(*cat*⁺) + p*cyo*⁺
4. *cyd*::T30(*cat*⁺) + vector
5. *cyd*::T30(*cat*⁺) + p*cyd*⁺

- 5) Infect the lungs of several mice with each of the 5 strains. Use the same number of bacteria for each mouse.

Independent Variable: (what will you change?)

- The genotype of *K. pneumoniae* used for each infection.

Dependent Variable(s): (what will you measure?)

- Bacterial load in the mouse lung after 48 hours (a quantitative measurement).

Negative Control(s): (eliminate false positive results)

- Uninfected control mice that receive no bacteria.

Positive Control(s): (eliminate false negative results)

- 1) Infect mice with wild-type bacteria with no vector, to confirm that the vector itself does not impact bacterial load in the lung.
- 2) Grow each strain in laboratory media with and without oxygen, to confirm that the mutations have the expected effect on aerobic growth and that the plasmids are able to complement those effects.

Potential Outcomes:

- 1) Neither of the mutants has any defect in lung colonization. This might indicate that either aerobic respiration is not required for lung colonization or that the *cyo* and *cyd* systems are redundant in that environment. The chloramphenicol resistance cassette in the T30(*cat*⁺) transposon can be removed by [Cre recombinase](#), which would allow you to construct a *cyo* *cyd* double mutant to test this possibility. You could also use allelic exchange or CRISPR-assisted recombineering to make the double mutant.
- 2) Loss of *cyo* reduces colonization and complementation with *cyo* restores it. This supports a model in which respiration of high O₂ concentrations is important for lung colonization.
- 3) Loss of *cyd* reduces colonization and complementation with *cyd* restores it. This supports a model in which respiration of low O₂ concentrations is important for lung colonization.
- 4) Both outcomes 2 & 3, which would support a role for respiration across a range of O₂ concentrations.
- 5) If any of the plasmids fail to complement defects in the respective mutants, this would suggest that those phenotypes might be due to off-target mutations or polar effects. The strains should be genome sequenced to eliminate the first possibility. Complementation with potential downstream genes could be used to address the second possibility.

Does this experiment establish correlation or causation?:

- Since we have fulfilled Falkow's postulates for *cyd* and *cyo*, this design tests **causation**.

This kind of experiment is another good candidate for a competitive index approach (recall this from [Lecture 7](#)), which increases the sensitivity and includes the wild-type as an internal control for each mouse.

ANAEROBIC RESPIRATION

Anaerobic respiration refers to any electron transport pathway in which the terminal electron acceptor is **not** O₂. There are a vast number of these used by different microbes in different environments, and we will not be able to do more than scratch the surface of the topic here. One key fact, though, is that the standard electrode potentials of all of the *alternative electron acceptors* are lower than that of O₂ / H₂O, so less energy is derived from reduction of those compounds. A very small list of examples of terminal electron acceptors used by some prokaryotes are listed below, with their respective standard electrode potentials:

O ₂ / H ₂ O	+815 mV
Fe ³⁺ / Fe ²⁺	+771 mV
NO ₃ ⁻ (nitrate) / NO ₂ ⁻ (nitrite)	+421 mV
fumarate / succinate	+33 mV
CO ₂ / CH ₄	-250 mV
S ⁰ / H ₂ S	-270 mV

(methanogenesis, specific to certain archaea)

Remember that the energy drawn from respiration is the ΔE_h between the terminal electron acceptor and the NAD⁺ / NADH redox pair (standard electrode potential of -320 mV), so you can see from this list that the possible energy yield of anaerobic respiration can vary widely.

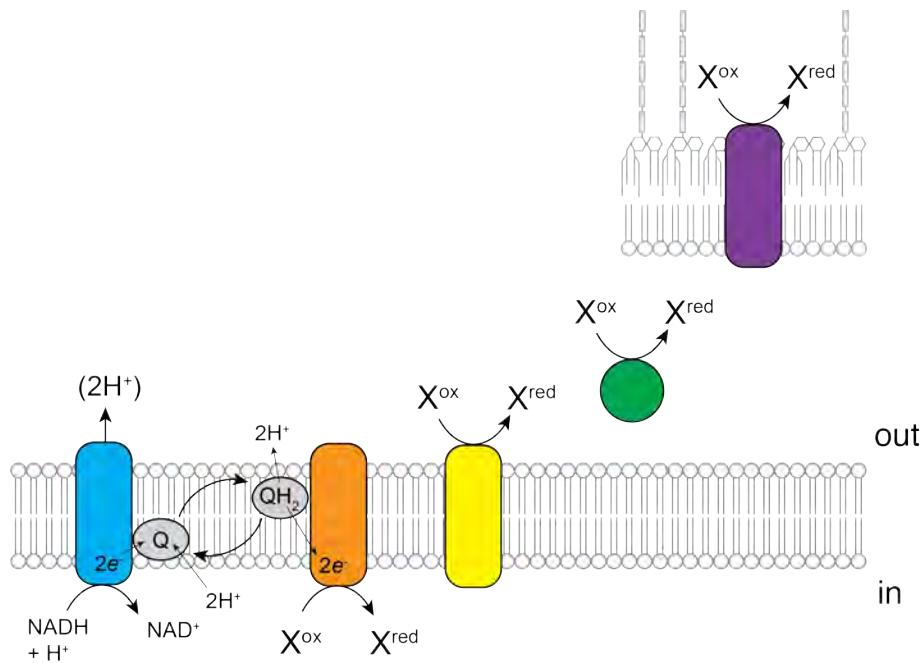


Figure 16.8. Anaerobic respiration, in which oxidation of NADH by a NADH:quinone oxidoreductase (which may or may not be a coupling site) is linked to reduction of any molecule that is not O₂. The respiratory reductases that reduce various terminal electron acceptors used by different bacteria may be in the inner membrane, the periplasm, the outer membrane, or on the surface of the cell.

As shown in Figure 16.8, anaerobic respiration begins with oxidation of NADH by an NADH:quinone oxidoreductase, which may or may not be a proton-pumping coupling site, passes through a quinone intermediate, and ultimately ends up at a reductase that uses the electrons from NADH to reduce the terminal electron acceptor. Depending on the specific terminal electron acceptor, the reductase may be in the inner membrane, the periplasm, the outer membrane, or on the surface of the cell, with additional electron shuttling proteins existing to carry the electrons to their final destination.

From an evolutionary standpoint, anaerobic respiration is much older than and ancestral to aerobic respiration. [This 2023 minireview](#) does a good job of summarizing what we know about the relationships between these pathways.

Returning once again to *E. coli*, an organism with a conveniently large respiratory repertoire for illustrative purposes, the respiratory nitrate and fumarate reductases are in the inner membrane and carry out their reactions in the cytoplasm. The trimethylamine oxide (TMAO) and dimethylsulfoxide (DMSO) reductases are also in the inner membrane, but carry out their reduction reactions in the periplasmic space. Nitrite reductase is a soluble periplasmic protein. None of these proteins is a coupling site, and they do not contribute directly to generating PMF except via the Q loop. Expression of the respiratory reductases for anaerobic respiration in *E. coli* is typically repressed by the presence of oxygen and stimulated by the presence of their cognate electron acceptors. Nitrate, nitrite, TMAO, and DMSO are all compounds abundant in the mammalian gut, especially during inflammation. Respiration of these compounds gives *E. coli* and related enterobacteria (e.g. *Salmonella*) a significant advantage over other bacteria in the intestine during inflammation.

DISCUSSION PROBLEM SET #30: METAL REDUCTION BY SHEWANELLA ONEIDENSIS

Shewanella oneidensis is an environmental γ -proteobacterium that is notable for being able to respire using metal oxides as terminal electron acceptors (although they can also respire more conventional compounds like oxygen and sulfate). They are essentially capable of breathing rocks, deriving energy from the reduction of ions like Fe³⁺, Mn⁴⁺, U⁶⁺, and Tc⁷⁺. The metal oxides containing these ions and their reduced products are extremely insoluble solids, which means that *Shewanella* must carry out their terminal electron transfer reactions outside of the cell, using reductases on their cell surfaces, with additional protein complexes needed to move electrons from cytoplasmic NADH across the cell envelope.

When grown anaerobically in the presence of soluble HAuCl₄ and AgNO₃ salts, *S. oneidensis* uses the gold and silver ions as terminal electron acceptors, producing insoluble particles a couple of hundred nanometers in diameter composed of pure metallic Au⁰ and Ag⁰.

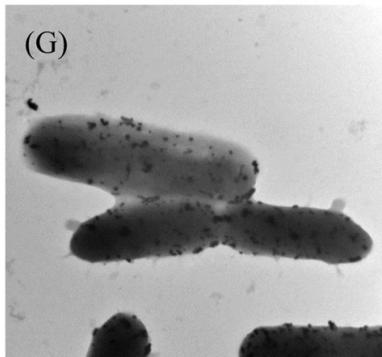


Image from Huang et al. (2019) Sci Rep 9:7589.

These nanoparticles, when purified, turn out to have potent antimicrobial and anti-biofilm activity which can be used to kill drug-resistant pathogens, including *Pseudomonas aeruginosa* and *Enterococcus faecalis*.

The following genetic tools are available for *S. oneidensis*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (4.8 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (Link)	✓

Design an experiment to identify proteins specifically required for respiration of Au and Ag in *S. oneidensis*. State:

- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiments, and how you will interpret them

Question for discussion in class: Metal oxide nanoparticles are thought to kill bacteria by a variety of mechanisms, but a primary one is disrupting cell membranes by electrostatic interactions. This is a much more difficult thing for bacteria to evolve resistance to than antibiotics, which is good for us, but less good for *Shewanella*. What strategies might you hypothesize that *Shewanella* could use to cope with the toxic byproducts of their respiration?

FERMENTATION

Not all bacteria are able to respire, and even many of those that are can often grow in the absence of available terminal electron acceptors. They do so by the process of *fermentation*.

The fundamental problem faced by fermenting organisms is recycling the cell's limited pool of nicotinamide, since, as we will see in **Lecture 17**, central metabolism generates a very large amount of reduced NADH. Respiring organisms put that NADH to work generating PMF, as described above, but without respiration, NADH and other *reducing equivalents* (like phosphorylated NADPH) must be oxidized by alternative pathways. It is a little tricky to separate the discussion of fermentation from central metabolism, but I will do my best, and some aspects of this will hopefully become more clear after to-

morrow's chapter. There are a very wide range of fermentation strategies used by different bacteria, and I will only describe a few of the most common here, but they all are ways to achieve the same goal of having some place to put the electrons that would be fed into an electron transport chain in a respiring organism.

The lactic acid bacteria (*Lactobacillus*, *Streptococcus*, *Lactococcus*, *Leuconostoc*, *Enterococcus*, etc.) are very abundant and common Gram-positive bacteria that survive almost entirely by fermentation. As their name suggests, they produce abundant lactic acid (lactate) as a fermentation byproduct, by one of two pathways: *homolactic fermentation*, which produces only lactate, and *heterolactic fermentation*, which produces lactate, ethanol, and CO₂.

Homolactic fermentation is very simple, and serves to illustrate the principal function of fermentation pretty clearly:

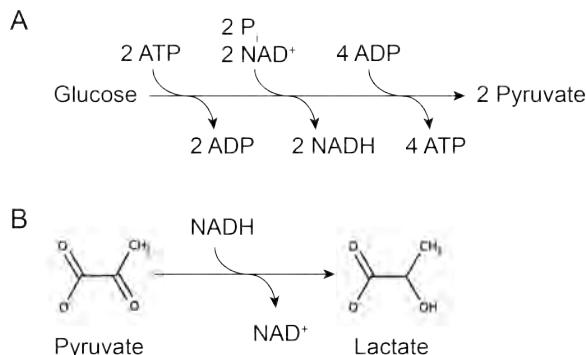


Figure 16.9. Homolactic fermentation. (A) Glycolysis consumes a glucose molecule, generating net products of 2 ATP, 2 NADH, and 2 pyruvate. (B) Lactate dehydrogenase catalyzes the reduction of pyruvate to lactate, with the concomitant oxidation of NADH.

As shown in Figure 16.9A, glycolysis converts glucose to pyruvate, generating 2 ATP and 2 NADH in the process. In order to regenerate the NAD⁺ needed for more cycles of glycolysis, NADH is used by the enzyme lactate dehydrogenase to reduce pyruvate to lactate (Figure 16.9B). The lactate is secreted from the cell, rapidly acidifying the growth media of bacteria using this pathway. The substrate-level phosphorylation steps in glycolysis are the primary source of ATP in bacteria growing by homolactic fermentation, and we will return to them in the next chapter. ATP production by glycolysis is actually up to 10 times **faster** than ATP production by aerobic respiration (as reported [here](#)), even though the final yield is much lower, so under glucose-rich conditions many organisms will prioritize glycolysis and some kind of fermentation over respiration, even if they are capable of both.

Some bacteria that are not able to tolerate the high acidity generated by homolactic fermentation use a different, more flexible fermentative pathway called *mixed acid fermentation* that reduces pyruvate to a mixture of lactate, formate, acetate, ethanol, CO₂, and H₂ (Figure 16.10). This pathway is found in enterobacteria, including *E. coli*, *Salmonella*, and *Shigella*.

Mixed acid fermenters have lactate dehydrogenase, like homolactic fermenters, but can also use an enzyme complex called pyruvate-formate lyase and a sulfur-containing cofactor called *cofactor A* (usually just called CoA) to divide pyruvate into acetyl-CoA and formate. Some mixed acid fermenters use formate-hydrogen lyase to convert formate to CO₂ and H₂ gases, but others just excrete formate. Acetyl-CoA can then follow one of two pathways. It can be reduced in two steps to acetaldehyde and ethanol, consuming an NADH at each stage. It can also be converted into acetyl-phosphate (acetyl-P) by phosphotransacetylase. Acetate kinase can use acetyl-P to generate an ATP and acetate, but does not consume any NADH in the process. Acetyl-CoA is a key intermediate in metabolism, and we will be seeing it again in **Lecture 17**.

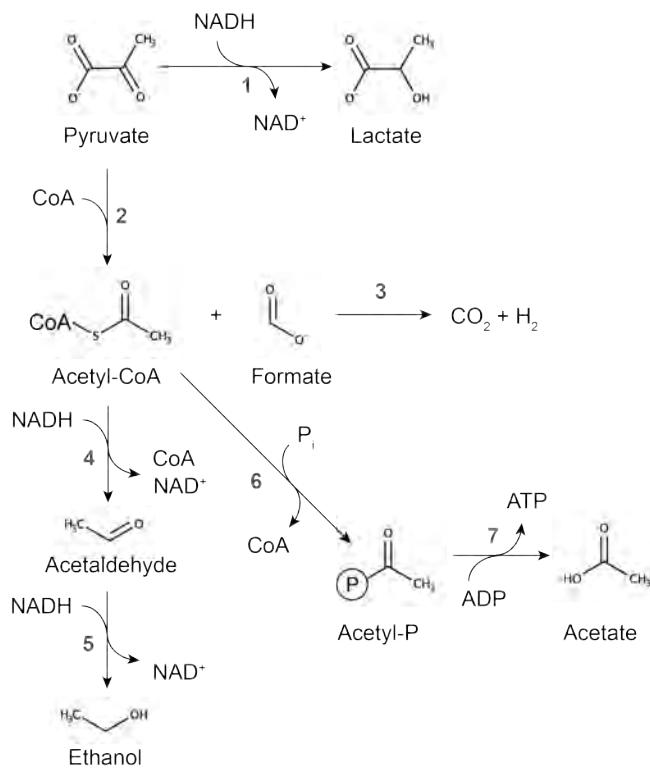


Figure 16.10. Mixed acid fermentation. Pyruvate is reduced to lactate by lactate dehydrogenase (1) or converted to acetyl-CoA and formate by pyruvate-formate lyase (2). Formate is degraded to CO_2 and H_2 gases by formate-hydrogen lyase (3). Acetyl-CoA can either be reduced to acetaldehyde by acetaldehyde dehydrogenase (4) and then to ethanol by alcohol dehydrogenase (5) or converted to acetyl-P by phosphotransacetylase (6). Acetyl-P is used as a phosphate donor for ATP synthesis by acetate kinase (7).

Mixed acid fermenters can balance the different branches of this pathway to produce different ratios of products, including the possibility of some additional ATP that homolactic fermenters cannot produce and a useful acetyl-CoA intermediate. Exactly what products are produced depends on the growth conditions of the organism. For example, secreted acetate and lactate acidify the media, while ethanol does not. However, all of the products of mixed acid fermentation except the gases are toxic to cells at high concentrations.

Butyrate and other **short chain fatty acids** (SCFA) produced by bacterial fermentation in the large intestine are a major carbon source for intestinal epithelial cells and are an important factor in the health of the gut. SCFA have systemic health effects on the host, influencing the gut-brain axis, sleep cycles, epithelial barrier functions, and lowering inflammation. Generally speaking, if your microbiome is fermenting a lot of complex carbohydrates to butyrate in your large intestine, you will be healthier. The most abundant butyrate producing bacteria in the human gut are extremely oxygen-sensitive Gram-positive obligate anaerobes related to the *Clostridia*, including members of the genera *Faecalibacterium*, *Eubacterium*, and *Roseburia*.

Butyrate fermentation is illustrated in Figure 16.11. Two pyruvates, 2 ATP, and 2 NADH are produced from a glucose by glycolysis in these organisms, exactly as in the fermentations we discussed above.

The two pyruvates are decarboxylated to acetyl-CoA, then condensed to form a 4-carbon acetoacetyl-CoA product. Acetoacetyl-CoA is reduced twice, consuming an NADH each time, to yield butyryl-CoA, which, in a direct parallel to the acetyl-CoA to acetyl-phosphate to ATP pathway we saw in mixed acid fermentation, is used to generate an ATP. The resulting butyrate is excreted from the cell, along with 2 CO_2 and 2 H_2 produced in the initial decarboxylation step, which uses a protein called ferredoxin as an electron and hydrogen carrier molecule. Note how the fermentation is balanced, with no NADH left over, and yielding a total of 3 ATP from each molecule of glucose.

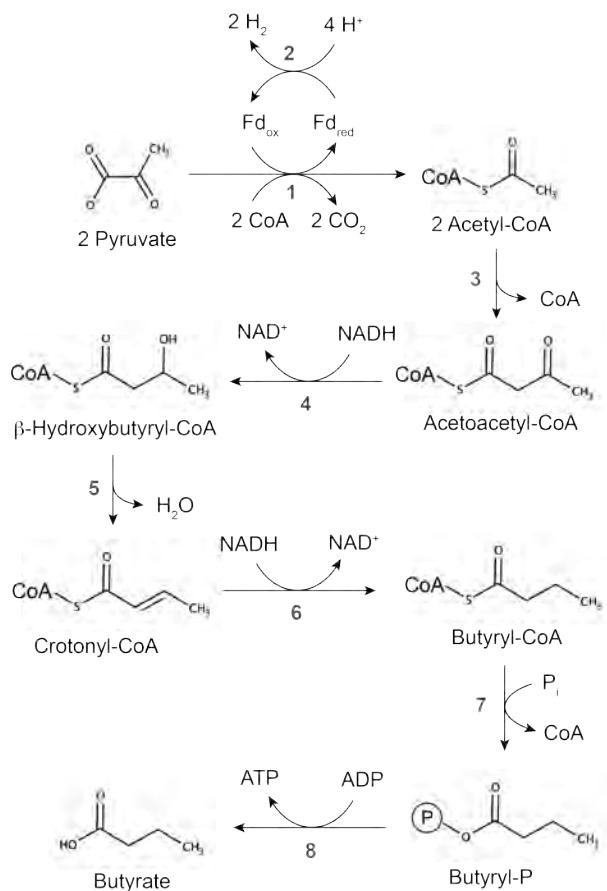


Figure 16.11. Butyrate fermentation. Pyruvate is oxidatively decarboxylated by pyruvate-ferredoxin oxidoreductase (1) to acetyl-CoA and CO₂. Oxidized ferredoxin is reduced by a hydrogenase (2), which produces H₂ gas. Two molecules of acetyl-CoA are condensed into acetoacetyl-CoA by acetyl-CoA acetyltransferase (3), which is then reduced by β-hydroxybutyryl-CoA dehydrogenase (4) to crotonyl-CoA, consuming an NADH. Crotonyl-CoA dehydrogenase (6) consumes another NADH and produces butyryl-CoA. Phosphotransbutyrylase (7) displaces CoA with phosphate, forming butyryl-phosphate, which butyrate kinase (8) uses as a phosphate donor to produce ATP and butyrate.

I do need to point out that not all fermentation pathways start with glycolysis and its end product pyruvate. For example, heterolactic fermentation begins with the conversion of glucose to xylulose-5-phosphate via the pentose phosphate pathway ([Lecture 17](#)), and many *Clostridium* species ferment lactate to acetate or propionate.

Since fermentation does not result in proton pumping, fermenting bacteria need to use alternative mechanisms to generate a PMF. The most common is to simply run the F₁F₀ ATP synthase in reverse (Figure 16.2C), although if the organism is living in an environment with another species that very rapidly consumes a secreted fermentation end product, a proton symporter can be used to generate PMF (Figure 16.2A). Obviously, organisms growing by fermentation are usually more energy-limited than respiring species, both in terms of PMF and ATP, and therefore often grow more slowly and to lower final cell density on a given nutrient source.

PHOTOSYNTHESIS

Most of you are probably most interested in disease-causing bacteria (given the expertise of the UAB Microbiology Department), but I do want to devote some space here to *photosynthesis*, the process by which **light** is used as an energy source for generating a PMF or reducing equivalents (in this case, usually NADPH rather than NADH). No pathogenic bacteria I'm aware of are photosynthetic, but many environmental bacteria are, and sunlight is, of course, the major energy input into the biosphere of Earth. Roughly 50% of the oxygen we breathe is produced by marine photosynthetic cyanobacteria, the same general kind of bacteria which are the ancestors of plant chloroplasts (source of the other 50%). The biochemistry of photosynthesis is very complicated, and I will only touch on the essentials from a few major groups of prokaryotes here.

The simplest photosynthetic pathway depends on a single transmembrane protein called *bacteriorhodopsin* that absorbs light and uses the energy of that light to pump protons across the membrane (Figure 16.12).

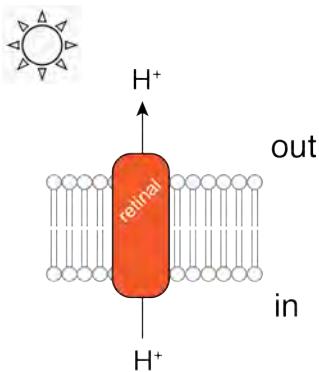


Figure 16.12. Light-driven proton pumping by bacteriorhodopsin, powered by light absorption by the covalently-bound retinal cofactor.

Bacteriorhodopsin was actually first identified in extremely *halophilic* (salt-loving) archaea belonging to the confusingly-named genus *Halobacterium*, but we now know that homologous proteins are found in some bacteria, including the halophilic Bacteroidetes clade member *Salinibacter ruber* and the cyanobacterium *Gloeobacter violaceus*. Each bacteriorhodopsin protein contains a covalently bound retinal (vitamin A) cofactor. When retinal absorbs a photon, it undergoes a conformational change which drives a change in the structure of the bacteriorhodopsin protein, resulting in transport of a proton across the membrane. This is a very direct method of linking light energy to PMF generation, but most photosynthetic organisms use considerably more complex pathways that draw more energy out of each captured photon.

Plants and cyanobacteria use H_2O as an electron donor during photosynthesis, generating abundant O_2 . This is called *oxygenic photosynthesis* (Figure 16.13). This produces lots of energy, but requires a way of dealing with oxygen toxicity, which they usually do at least partly by being capable of aerobic respiration.

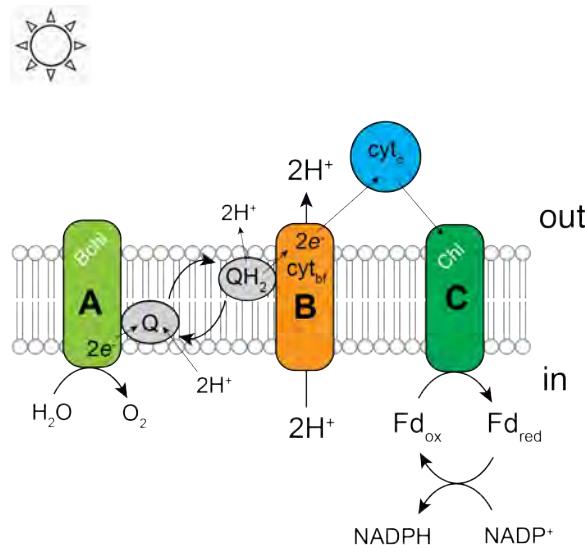


Figure 16.13. Oxygenic photosynthesis. Electrons are donated from water, producing O_2 , powered by light absorption by the bacteriochlorophyll (Bchl) cofactor of photosystem II (A). This is linked via the quinone pool to a membrane-bound proton-pumping cytochrome bf protein (B) and a soluble cytochrome c protein, with the electrons ultimately being used to reduce the protein ferredoxin in a reaction powered by light absorption by the chlorophyll (Chl) cofactor of photosystem I (C). Reduced ferredoxin is oxidized to produce NADPH.

The first step in oxygenic photosynthesis is absorption of light by a bacteriochlorophyll-containing photosystem II (PSII) complex (Figure 16.13A), which draws electrons from water and passes them, via a quinone intermediate, to a cytochrome bf proton pump (Figure 16.13B), generating PMF. From there, the electrons are transferred via a cytochrome c protein to a chlorophyll-containing photosystem I (PSI) complex which, with the absorption of another photon, uses them to reduce ferredoxin. The resulting reduced ferredoxin is used to reduce NADP^+ to NADPH (Figure 16.13C). Oxygenic photosynthesis thereby directly generates both PMF and additional reducing equivalents for respiration, biosynthesis, or other pathways (Lecture 17).

Like chloroplasts and, indeed, most phototrophs, cyanobacteria contain complex internal membrane structures to increase the surface area available for photosynthetic reaction centers. These are called *thylakoid membranes* in cyanobacteria and chloroplasts.

Diverse types of bacteria are able to carry out various forms of *anoxygenic photosynthesis*, which is a general term for photosynthetic pathways that use molecules other than H_2O as an electron donor and therefore do not produce O_2 . These

include the well-studied purple non-sulfur bacteria, which are α - and β -proteobacteria belonging to genera like *Rhodospirillum*, *Rhodobacter*, *Rhodopseudomonas*, or *Rhodococcus* ("rhodon" is Greek for "rose", and these bacteria are generally red to purple in color), that use H₂ or organic molecules as electron donors. The purple sulfur bacteria are γ -proteobacteria of the order Chromatiales that use sulfur-containing molecules (largely H₂S and S⁰, but also thiosulfate [S₂O₃²⁻]) as electron donors. The green sulfur bacteria are distantly related to the Bacteroidetes and mainly use H₂S as an electron donor. The heliobacteria, including the genera *Heliobacterium*, *Heliophilum*, and *Heliobacillus*, are spore-forming anaerobic Firmicutes with a photosynthetic pathway similar to but somewhat simpler than that found in green sulfur bacteria.

To illustrate some of the properties of anoxygenic photosynthesis, we will first look at a simplified version of how it works in purple non-sulfur bacteria (Figure 16.14A, B). In these bacteria, the photosynthetic reaction center (Figure 16.14A) is a membrane protein complex that contains a bacteriochlorophyll cofactor, which is able to absorb light. When this happens, it powers electron transfer from bacteriochlorophyll to a quinone. Very similarly to what we saw earlier when discussing respiration (Figure 16.7), the resulting reduced quinone is oxidized by a cytochrome-containing, proton-pumping membrane protein (Figure 16.14B), resulting in the generation of a PMF. Unlike during respiration, there is cyclic electron flow in this form of photosynthesis, with a soluble cytochrome c protein serving as an electron carrier to return the electrons to the reaction center. This means that no NAD(P)H is produced during this reaction. The photosynthetic reaction center of purple non-sulfur bacteria is homologous to photosystem II (PSII) of plants and cyanobacteria.

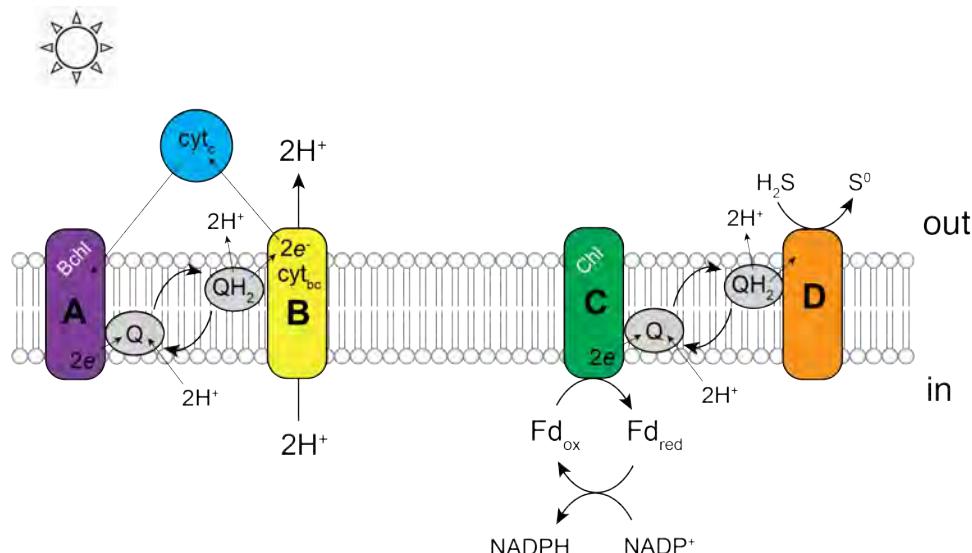


Figure 16.14. Anoxygenic photosynthesis. In purple non-sulfur photosynthetic bacteria (A, B), light absorption by the bacteriochlorophyll (Bchl) cofactor of photosystem II is used to set up a cyclic electron flow (A). This is linked via the quinone pool to a proton-pumping cytochrome bc protein (B). Electrons are returned to (A) by a soluble cytochrome c protein. In green sulfur bacteria (C, D), light absorption by the bacteriochlorophyll (Bchl) cofactor of photosystem I (C) is used to reduce ferredoxin and is linked via the quinone pool to a membrane-bound sulfide-quinone reductase (D), which reduces hydrogen sulfide to elemental sulfur in the periplasm.

Anoxygenic photosynthesis as carried out by the green sulfur bacteria, on the other hand, is non-cyclic, and uses a variety of possible electron donors, including H₂S, H₂, or some organic molecules. Green sulfur bacteria basically have only photosystem I. Figure 16.14 C and D illustrates a simplified example of this kind of pathway.

As I said earlier, this is a **very** brief introduction to the world of photosynthesis. There are many more cofactors, light-harvesting molecules, and proteins involved, and many variations on the pathways. Hopefully what I've presented here will at least give you a place to start if you ever find yourself studying a photosynthetic organism.

Note that, while in many photosynthetic organisms, photosynthesis is linked to *autotrophy* (the ability to incorporate or fix CO₂ into organic molecules), this is by no means universally true. Most plants and many cyanobacteria are *obligate photoautotrophs*, but a lot of photosynthetic bacteria are *photoheterotrophs* that obtain carbon from the breakdown of organic molecules, and there are lots of autotrophs that use non-photosynthetic pathways and energy sources to fix CO₂. As fascinating as these pathways are, due to space and time limitations, we will **not** be discussing any of them during this class.

LECTURE 17: CENTRAL METABOLISM

INTRODUCTION

Central metabolism refers to both the *catabolic* biochemical pathways by which organisms break down molecules into their component parts and the *anabolic* pathways by which the components of new and growing cells are synthesized. We will explore some of the pathways by which organic molecules are degraded and link these to the energy-generating mechanisms we discussed in the previous chapter. We will also discuss representative pathways by which biochemical building blocks like amino acids and nucleotides are synthesized from their constituent parts, with an attempt to illustrate some of the general principles common to such pathways. The goal is **not** to memorize pathways, but rather to give you a sense of the **patterns** that unify metabolic processes.

Many definitions of central metabolism focus entirely on carbon (C) metabolism, but for the sake of branching out a bit and going beyond material you've probably covered several times in prior classes, I'm going to include discussions of some of the pathways of nitrogen (N) and sulfur (S) metabolism as well.

CATABOLISM

When discussing metabolism, there are a few key pathways that we really can't avoid, including glycolysis, the *pentose phosphate pathway*, and the *tricarboxylic acid* (TCA) cycle. You have almost certainly encountered these before, since they are conserved in both eukaryotes and *E. coli*, but let's take a closer look, starting with glycolysis (Figure 17.1).

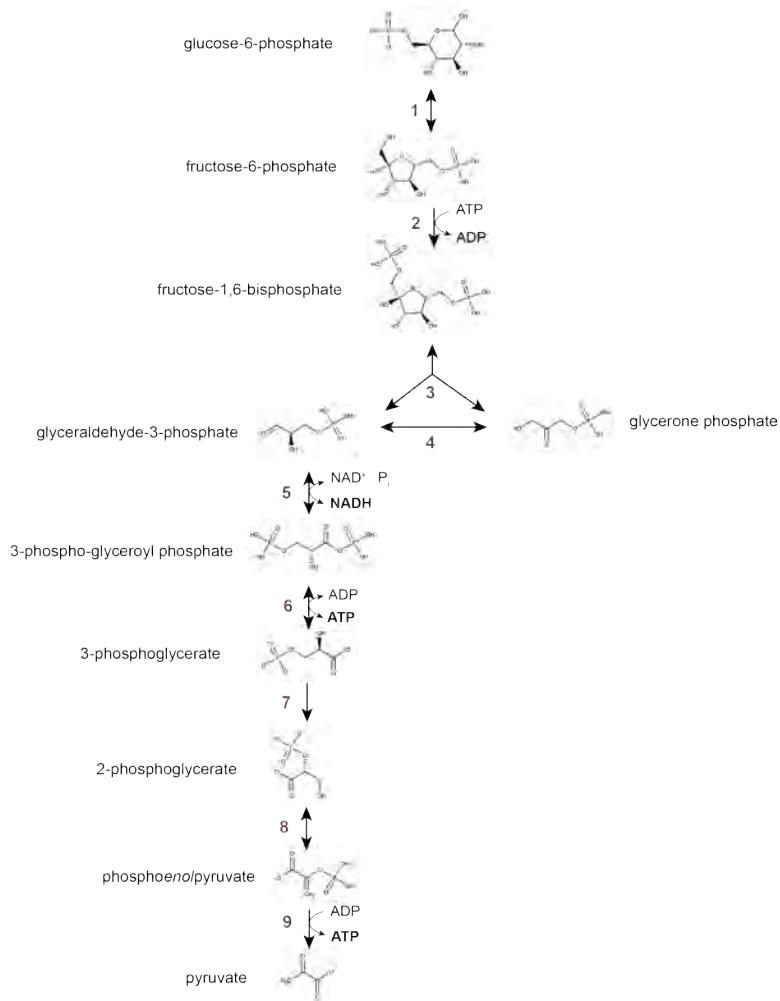


Figure 17.1. Glycolysis. Enzymes: 1, glucose-6-phosphate isomerase (Pgi); 2, 6-phosphofructokinase (Pfk); 3, fructose-bisphosphate aldolase (Fba); 4, triose-phosphate isomerase (Tpi; 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pgk); 7, phosphoglycerate mutase (Gpm); 8, enolase (Eno); 9, pyruvate kinase (Pyk). Note that glyceraldehyde-3-phosphate is also sometimes known as dihydroxyacetone phosphate (DHAP).

Starting with **glucose-6-phosphate** (G6P), glycolysis is catalyzed by a series of enzymes that result in the production of 2 molecules of pyruvate, along with 4 ATP and 2 NADH. One of those ATP equivalents is needed for the phosphorylation of fructose-6-phosphate to fructose-1,6-bisphosphate, so the actual energy yield per glucose is 3 ATP. We have already discussed how fermenting organisms recycle the NADH and dispose of the pyruvate generated by glycolysis (**Lecture 16**).

The net reaction balance of glycolysis is:



A key enzyme in glycolysis that deserves specific attention is **glyceraldehyde-3-phosphate dehydrogenase (GapDH)**, step 5 in Figure 17.1. Note that this enzyme catalyzes the phosphorylation of **glyceraldehyde-3-phosphate (G3P)** **without** the involvement of an ATP, which is what allows glycolysis to be a net energy-generating process. It is also important to notice that most of the enzymes catalyzing steps in glycolysis are reversible and operate close to equilibrium (**Lecture 4**), so that the direction of flux through the pathway is controlled by the activity of just a few enzymes (2, 7, and 9 in Figure 17.1).

So far, nothing terribly new and exotic, but bacteria have much more metabolic diversity than eukaryotes and glycolysis is not the **only** way to break down glucose-6-phosphate to pyruvate. *P. aeruginosa* and many other bacteria use a somewhat different pathway, called the **Entner-Douderoff pathway** (Figure 17.2.).

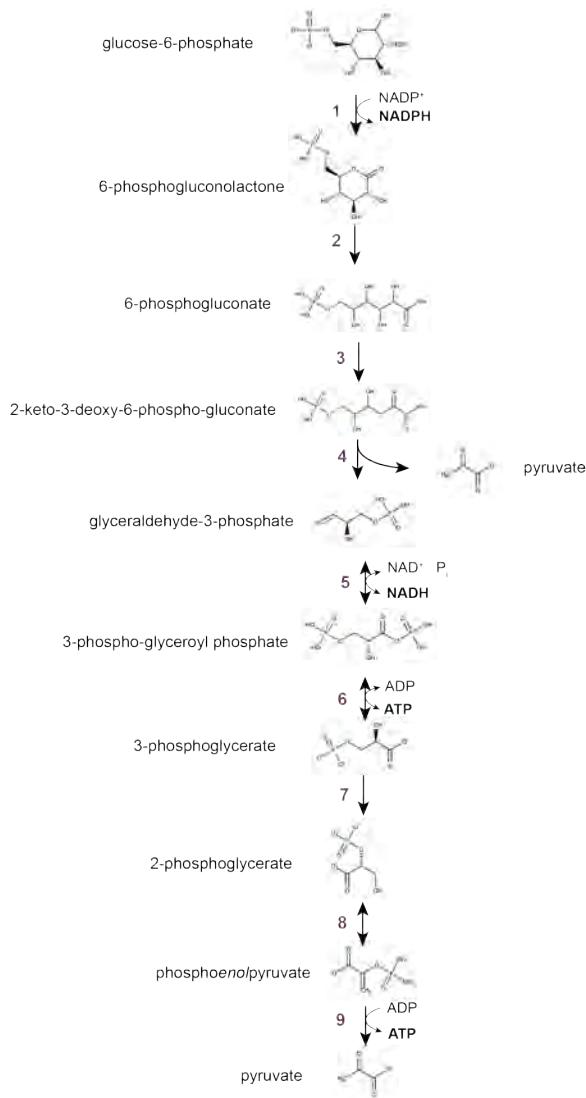


Figure 17.2. Entner-Douderoff pathway. Enzymes: 1, glucose-6-phosphate dehydrogenase (Zwf); 2, 6-phosphogluconolactonase (Pgl); 3, phosphogluconate dehydratase (Edd); 4, 2-keto-3-deoxygluconate-6-phosphate aldolase (Eda); 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pfk); 7, phosphoglycerate mutase (Gpm); 8, enolase (Eno); 9, pyruvate kinase (Pyk).

The net reaction balance of the Entner-Douderoff pathway is:



The Entner-Douderoff pathway converts G6P via four irreversible enzyme steps to one molecule of G3P and one of pyruvate, generating a reduced NADPH in the process. The G3P is oxidized to pyruvate by the same enzymes as in glycolysis, generating 2 ATP and an NADH in the process.

The Entner-Douderoff pathway combines enzymes from glycolysis and another extremely well-conserved pathway you have almost certainly encountered before: the **pentose phosphate pathway (PPP)** (Figure 17.3). The PPP is yet another path-

way by which G6P is oxidized to G3P, but in a more complicated way that generates a large amount of NADPH and passes through several intermediates which are precursors of essential metabolites (see below).

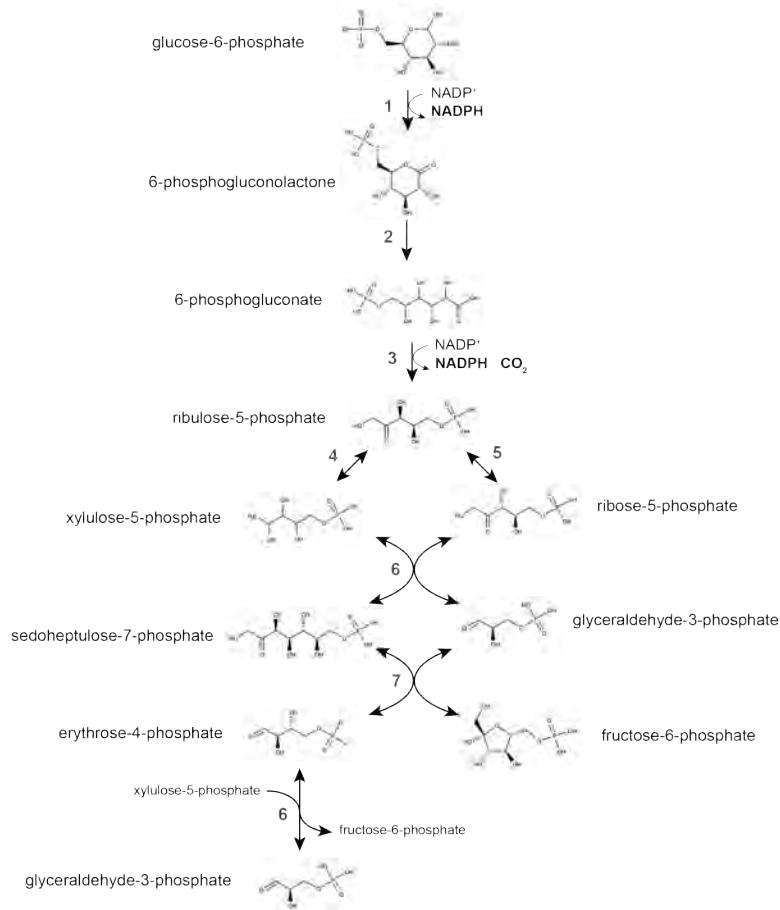


Figure 17.3. The pentose phosphate pathway. Enzymes: 1, glucose-6-phosphate dehydrogenase (Zwf); 2, 6-phosphogluconolactonase (Pg); 3, phosphogluconate dehydratase (Edd); 4, ribulose-phosphate 3-epimerase (Rpe); 5, ribose-5-phosphate isomerase (Rpi); 6, transketolase (Tkt); 7, transaldolase (Tal).

Enzymes 1 and 2 in Figure 17.3 are the same as the first two enzymes of the Entner-Douderoff pathway (Figure 17.2), with the fate of 6-phosphogluconate as the branch point that distinguishes one pathway from the other.

The net reaction balance of the PPP is:



The **fructose-6-phosphate** (F6P) and G3P are utilized by glycolysis, at least in *E. coli*, so that ultimately this will yield 5 NADH, 8 ATP, and 5 pyruvate from 3 G6P, which is slightly less of each than the yield of glycolysis alone, with one C lost as CO₂. Note once again how most of the enzymes of this pathway are reversible.

This is all very well for bacteria growing on glucose, of course, but bacteria can use lots of different sugars as carbon sources. Indeed, some bacteria do not have transporters capable of taking up glucose at all. How are these other sugars catabolized? The simple answer is that they are broken down into intermediates in one of the above pathways. We will look at a couple of examples to illustrate, remembering of course that we can't even begin to scratch the surface of the number of pathways that exist (Figure 17.4).

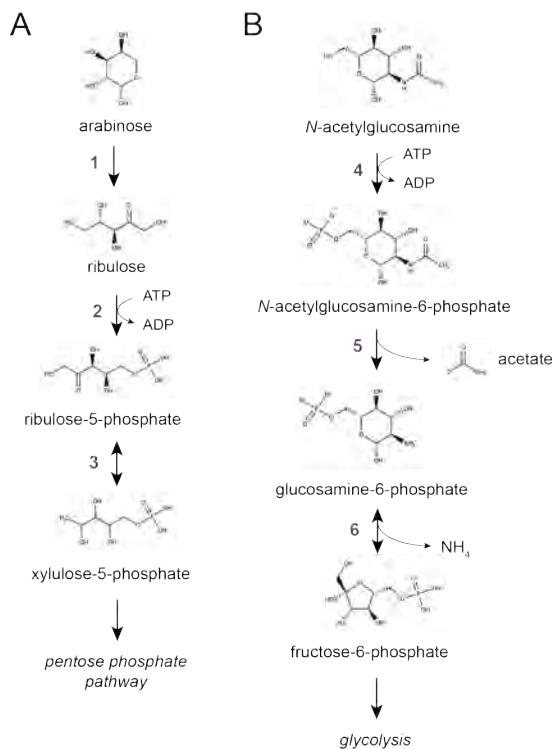


Figure 17.4. Sample monosaccharide utilization pathways. (A) Arabinose utilization. Enzymes: 1, arabinose isomerase; 2, ribulokinase; 3, ribulose-5-phosphate epimerase. (B) N-acetylglucosamine (GlcNAc, **Lectures 10 and 14**) utilization. Enzymes: 4, N-acetylglucosamine kinase; 5, N-acetylglucosamine-6-phosphate deacetylase; 6, glucosamine-6-phosphate deaminase.

The 5C sugar arabinose is isomerized into ribulose, phosphorylated, and enters the PPP as xylulose-5-phosphate (Figure 17.4A). The peptidoglycan component N-acetylglucosamine (GlcNAc, **Lectures 10 and 14**) is phosphorylated, then deacetylated and deaminated, yielding acetate, ammonia, and F6P, which can be degraded by glycolysis (Figure 17.4B).

Polysaccharides are degraded into mono- or disaccharides (by secreted enzymes, as a rule, **Lecture 13**) before uptake into bacteria, and disaccharides are cleaved into their constituent monosaccharides as the first step in their catabolism. LacZ (β -galactosidase), for example, cleaves the disaccharide lactose into glucose and galactose monosaccharides, which are then degraded individually.

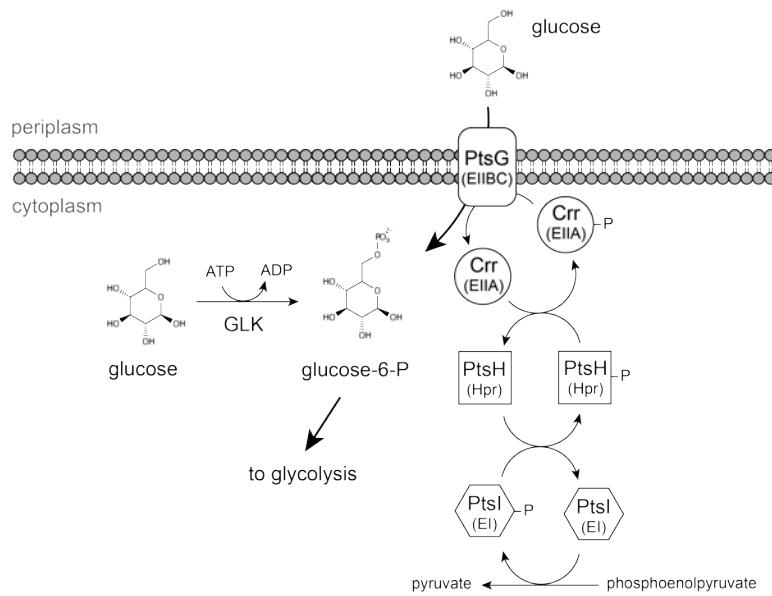
DISCUSSION PROBLEM SET #31: PTS AND NON-PTS SUGAR PHOSPHORYLATION

As we've seen in Figure 17.1 – 17.4, the intermediates of the central metabolic pathways are nearly all phosphorylated. There are a number of reasons for this, including not only the need to derive ATP from substrate-level phosphorylation, but also to make sure that the intermediates stay **inside** the cell. Charged molecules cannot easily cross lipid bilayers, so phosphorylated intermediates are retained in the cytoplasm.

Phosphorylation of sugars after uptake into the cytoplasm serves a second function in maintaining a concentration gradient of the un-phosphorylated sugar with higher concentrations outside the cell than inside, which provides an energetic boost for the transport of those sugars into the cell (**Lecture 16**).

Let's take, as an example, the incorporation and phosphorylation of glucose. Many bacteria contain an enzyme called *glucokinase* (GLK), which uses an ATP to phosphorylate glucose to glucose-6-phosphate, the first intermediate in glycolysis, Entner-Douderoff, and the PPP. This is a very straightforward way to solve this problem, and is typical of the catabolism of many sugars in many species (see Figure 17.4 for a couple of other sugar kinases with analogous roles). However, it is not the **only** way that sugars can be phosphorylated, and indeed, in many bacteria some sugars are phosphorylated **during** transport, by pathways called **phosphotransferase systems** (PTS).

The glucose PTS of *E. coli* is illustrated below:



In PTS transport, the source of the phosphate group added to the sugar is PEP (phosphoenolpyruvate) rather than ATP. This phosphate is transferred to the substrate by a series of carrier proteins. The EI (PtsI) and Hpr (PtsH) proteins are shared among all PTS systems, while different EI_A, EI_B, and EI_C homologs are specific to particular sugars (mono- or disaccharides). The EI_C component is an integral membrane protein, and the substrate is phosphorylated as it passes through into the cytoplasm. The EI_A, EI_B, and EI_C components of a PTS transporter can be separate or can be fused into 1 or 2 combined polypeptides. The glucose-specific PTS transporter proteins in *E. coli* are, as indicated, called Crr (EI_A) and PtsG (a fusion of EI_B and EI_C), respectively.

PTS transporters are common in bacteria, but different species have different repertoires of PTS and non-PTS sugars. *E. coli* has 21 EI_C homologs and *B. subtilis* has 16, although some of these are cryptic, meaning we don't know what their substrates are.

Xylose is a common 5-carbon sugar, ubiquitous in plants, that serves as a good carbon source for many bacteria. While studying the fermentation of traditional Greek olives (as you do), you find that the potential probiotic *Lactobacillus pentosus* uses xylose as its main carbon source for growth in that environment. In the lab, it can also grow on arabinose, which is likely to be fermented by a similar pathway ([Lecture 16](#)), although you expect there to be separate transporters for xylose and arabinose.

The following methods are available for *L. pentosus* (or at least for closely related species):

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (3.9 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓

Design an experiment to either determine whether xylose is phosphorylated by a PTS or a non-PTS system in *L. pentosus*, or to identify the gene(s) encoding the xylose transporter in this species. State:

- the hypothesis tested by each experiment
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation

- potential outcomes of your experiments, and how you will interpret them

Non-sugar carbon sources with more than 2 carbon atoms per molecule are generally catabolized using a similar strategy, by converting them into intermediates of the glycolytic pathways. Two pathways for glycerol (3C) assimilation, using different reductants and phosphodonorers but resulting in the same product (glycerone phosphate), are shown in Figure 17.5 as an example.

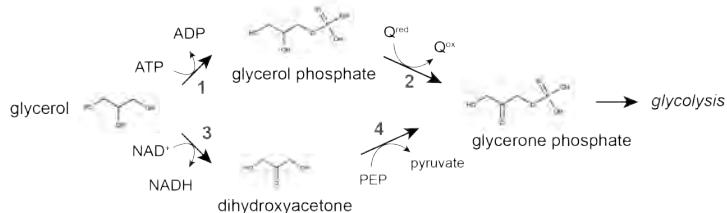


Figure 17.5. Glycerol utilization pathways found in bacteria. Enzymes: 1, glycerol kinase; 2, glycerol-3-phosphate dehydrogenase; 3, glycerol dehydrogenase; 4, dihydroxyacetone kinase.

As we will see shortly, assimilating molecules with fewer than 3 carbon atoms as carbon sources requires somewhat different strategies.

The last highly-conserved central pathway that we need to address is another one found in eukaryotes and *E. coli*, but which is by no means universal among bacteria: the tricarboxylic acid (TCA) or Krebs cycle (Figure 17.6):

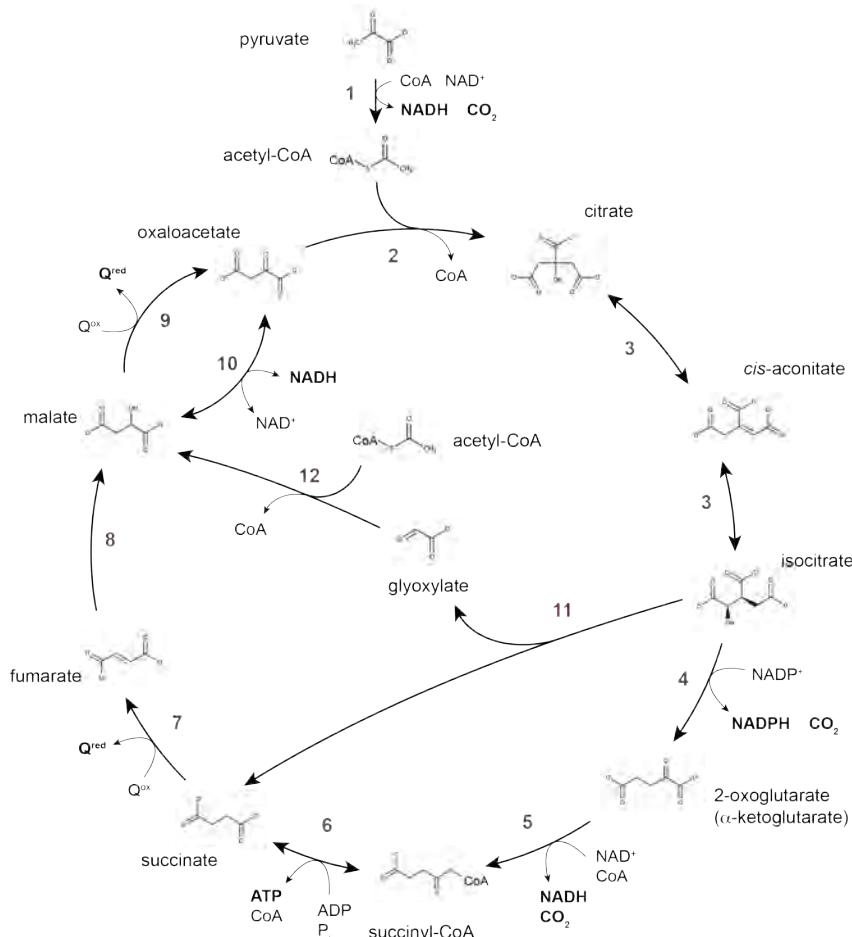


Figure 17.6. The TCA cycle and glyoxylate bypass. Enzymes: 1, pyruvate dehydrogenase (Pdh); 2, citrate synthase (GltA); 3, aconitase (Acn); 4, isocitrate dehydrogenase (Icd); 5, 2-oxoglutarate dehydrogenase; 6, succinyl-CoA synthetase (SucCD); 7, succinate:quinone oxidoreductase (Sdh); 8, fumarase (Fum); 9, malate:quinone oxidoreductase (Mqo); 10, malate dehydrogenase (Mdh); 11, isocitrate lyase (AceA); 12, malate synthase (AceB).

The TCA cycle oxidizes pyruvate to CO_2 , generating a little bit of ATP and a large number of reducing equivalents (reduced NADH, NADPH, and quinones). In respiring bacteria, those reducing equivalents are converted into PMF using the electron transport chain ([Lecture 16](#)). Bacteria growing by fermentation may carry out some TCA cycle reactions to generate

biosynthetic intermediates (see below), but typically have little flux through the cycle as a whole, and many obligate fermenters lack the enzymes of the TCA cycle entirely. Generating large amounts of NADH is not desirable in organisms that must dispose of reducing equivalents by fermentation.

The first reaction in Figure 16.6, which deserves some additional attention, is decarboxylation of pyruvate, the end product of glycolysis and the Entner-Douderoff pathways, to form acetyl-CoA. In aerobically-growing *E. coli*, this is done by the pyruvate dehydrogenase protein complex, which is essential under aerobic conditions. Anaerobically and in many fermentative organisms, pyruvate-formate lyase (Lecture 16, Figure 16.10) converts pyruvate to acetyl-CoA, releasing formate instead of CO₂.

A close look at the TCA cycle reveals that there is no net C assimilation from this pathway. A 2C compound (acetyl-CoA) enters the cycle by being condensed with a 4C compound (oxaloacetate) to form a 6C compound (citrate). In the course of the cycle, 2C are lost as CO₂. This is problematic for organisms trying to use the TCA cycle to grow on carbon sources containing only 2C, such as acetate, ethanol, or ethanolamine. Where do they obtain the carbon necessary for biosynthesis?

We will look at growth on acetate (the shortest fatty acid) to illustrate how bacteria can resolve this problem.

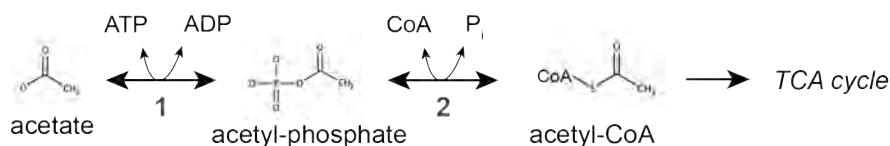


Figure 17.7. Acetate assimilation. Enzymes: 1, acetate kinase; 2, phosphotransacetylase.

The first steps of acetate assimilation, in which acetate is converted first to acetyl-phosphate and then to acetyl-CoA, are shown in Figure 17.7. Longer-chain fatty acids are also broken down into acetyl-CoA for catabolism (by β -oxidation, a pathway we do not have time to discuss), so the pathways necessary for growth on acetate are also necessary for growth on lipid carbon sources. (Note that the reactions of this pathway, like so many we've examined in this chapter, are reversible, meaning that acetyl-CoA synthesized by other pathways can, in fact, be used to generate ATP. This happens during mixed acid fermentation, for example; Figure 16.10.)

When growing on carbon sources that yield acetyl-CoA as a breakdown product, *E. coli* and many other bacteria bypass the decarboxylation steps of the TCA cycle (reactions 5 and 6 in Figure 17.6), instead using isocitrate lyase (reaction 11 in Figure 17.6) to split isocitrate into succinate (4C) and glyoxylate (2C). Glyoxylate is condensed with another acetyl-CoA to produce a 4C TCA cycle intermediate (malate). This pathway is called the glyoxylate bypass or glyoxylate shunt, and, as you can see from Figure 17.6, reduces the energy yield of the cycle considerably (by 2 reducing equivalents and an ATP), but prevents the loss of 2C as CO₂. Problem solved.

However, not every bacterium that can grow on acetate has the enzymes of the glyoxylate bypass. So, how do they solve the problem of carbon loss?

One mechanism is the ethylmalonyl-CoA mutase pathway, found in the purple nonsulfur photosynthetic α -proteobacterium *Rhodobacter sphaeroides*, a species which lacks isocitrate lyase but grows well with acetate as a sole carbon source (Figure 17.8). The ethylmalonyl-CoA mutase pathway consumes 3 acetyl-CoA and 2 CO₂ to generate the TCA cycle intermediates malate and succinate, using 11 enzymes to achieve roughly the same outcome as the two enzymes of the glyoxylate bypass, but with the added advantages of generating an ATP and actually fixing CO₂ into bioavailable carbon. The ethylmalonyl-CoA pathway consumes an NADPH to do so, and requires the energy-intensive synthesis of the complex cobalt-containing tetrapyrrole cofactor B₁₂ (required by enzymes 4 and 9) but as photosynthetic, respiring bacteria, *R. sphaeroides* are not generally limited for either energy or reducing power (Lecture 16).

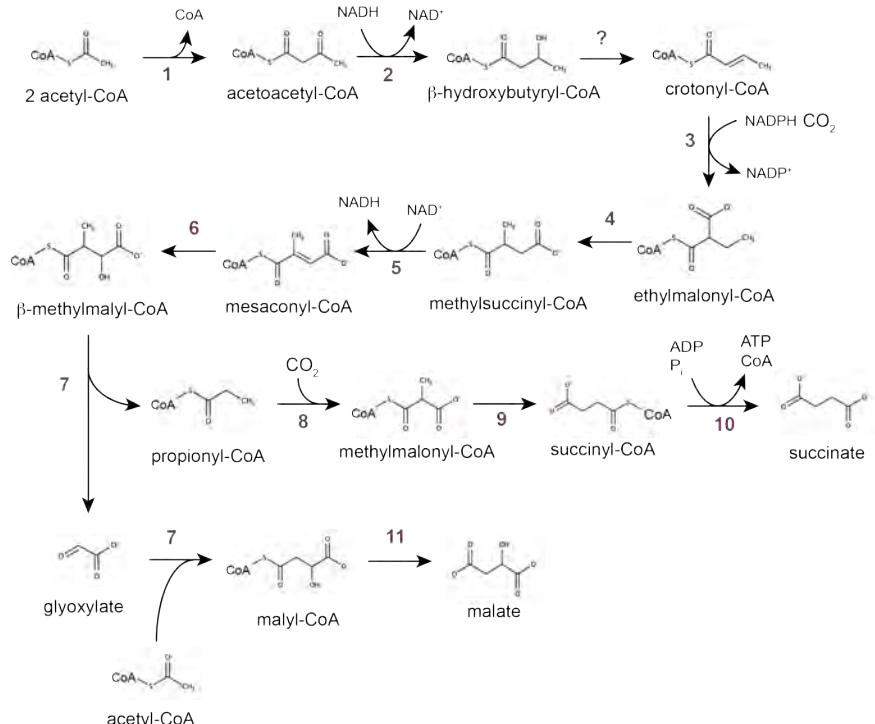


Figure 17.8. Acetate assimilation in *R. sphaeroides* via the ethylmalonyl-CoA mutase pathway. Enzymes: 1, β -ketothiolase; 2, acetoacetyl-CoA reductase; 3, crotonyl-CoA carboxylase/reductase; 4, ethylmalonyl-CoA mutase; 5, methylsuccinyl-CoA dehydrogenase; 6, mesaconyl-CoA hydratase; 7, β -methylmalyl-CoA / maryl-CoA lyase; 8, propionyl-CoA carboxylase; 9, methylmalonyl-CoA mutase; 10, succinyl-CoA synthetase; 11, maryl-CoA thioesterase. The enzyme that catalyzes conversion of β -hydroxybutyryl-CoA into crotonyl-CoA in *R. sphaeroides* has not been identified.

The contrast between acetate assimilation in enterobacteria and *R. sphaeroides* is a good example of the diversity of bacterial metabolism. There is rarely only one way to accomplish any particular metabolic goal, and a pathway being conserved in both eukaryotes and *E. coli* is no indication that that pathway is universally conserved among bacteria as a whole.

Another slightly problematic category of carbon sources are 4-carbon TCA cycle intermediates (succinate, malate, and fumarate). These are preferred carbon sources for many organisms, but since they enter directly into the TCA cycle, bacteria growing on 4CTCA cycle intermediates (or compounds that are broken down into 4CTCA cycle intermediates) require a mechanism to prevent the buildup of oxaloacetate (Figure 17.6). Most commonly, this is accomplished by the activity of phosphoenolpyruvate carboxykinase, which catalyzes the following reaction:

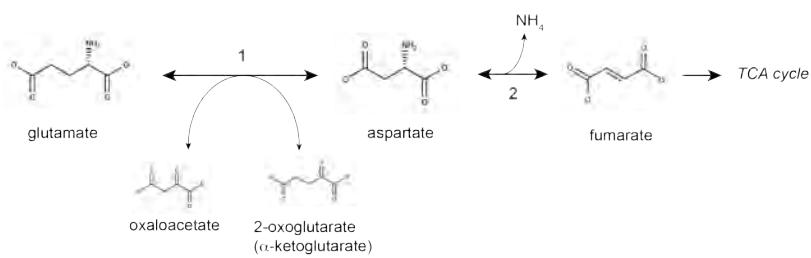


PEP can be converted into acetyl-CoA as in glycolysis (Figures 17.1, 17.6), with the concomitant loss of another CO_2 , allowing the TCA cycle to continue, or used to generate biosynthetic intermediates, which we will discuss in the second part of this chapter on anabolism. It is also, as noted above, an important phosphodonor in its own right for certain kinase reactions.

Not all carbon sources are sugars or intermediates in central metabolic pathways. Many bacteria can use amino acids as carbon sources, and in fact, in most rich laboratory media, these are the primary carbon sources available. Lysogeny Broth (LB, sometimes incorrectly called Luria-Bertani medium), a common rich medium for *E. coli* and other fast-growing bacteria, contains essentially no carbohydrates, for example. There's little point in plowing through all of the possible amino acid catabolic pathways, but I'll present a couple of representatives to illustrate the basic principles, which should not, at this point, be too surprising.

The pattern of carbon sources being broken down into intermediates of one of the central metabolic pathways continues: glutamate, for example (Figure 17.9A), is first converted to aspartate, which is then deaminated to release an NH_4 (ammonia) and fumarate, which is catabolized via the TCA cycle. Cysteine (Figure 17.9B) is converted into an unstable 2-aminoacrylate intermediate, releasing H_2S (hydrogen sulfide gas). 2-aminoacrylate spontaneously deaminates, releasing NH_4 and pyruvate, which can then be catabolized by any of the pyruvate-consuming pathways we've discussed, including the TCA cycle.

A



B

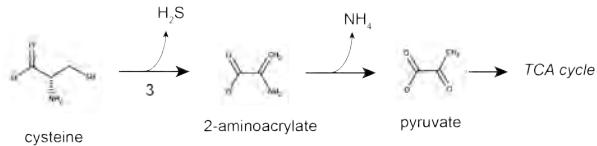


Figure 17.9. Representative amino acid degradation pathways. (A) Glutamate and aspartate degradation. Enzymes: 1, aspartate transaminase; 2, aspartate-ammonia lyase. (B) Cysteine degradation. Enzyme: 3, cysteine desulfhydrase. Deamination of 2-aminoacrylate to pyruvate is spontaneous.

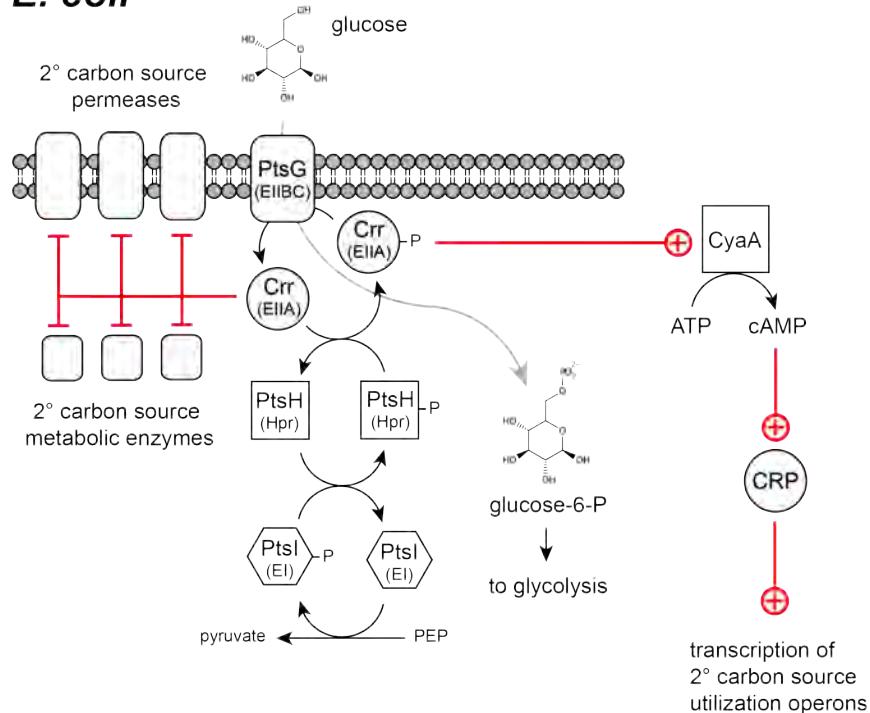
As we will see shortly, NH₄ and H₂S can be used as nitrogen or sulfur sources for biosynthesis or, if they are not needed, secreted out of the cell, where they will make the medium more basic or smellier, respectively.

DISCUSSION PROBLEM SET #32: CARBON CATABOLITE REPRESSION

Bacteria typically prefer to catabolize particular carbon sources (often, but not always, glucose), and repress the expression of proteins needed for the catabolism of other, secondary carbon sources when their preferred carbon source is present. This is called **carbon catabolite repression (CCR)** or just catabolite repression, and is a major global regulatory system in most bacteria.

The mechanisms of CCR have been well-studied in *E. coli* and *B. subtilis*, and to some extent in other bacteria. In both *E. coli* and *B. subtilis*, CCR depends on the glucose-specific PTS (recall Discussion Problem Set #31 above), but in distinctly different ways. The descriptions here are necessarily simplified somewhat, since CCR is a complex process and individual secondary carbon sources have their own different variations on the details, but I will try to explain the general principles involved.

E. coli

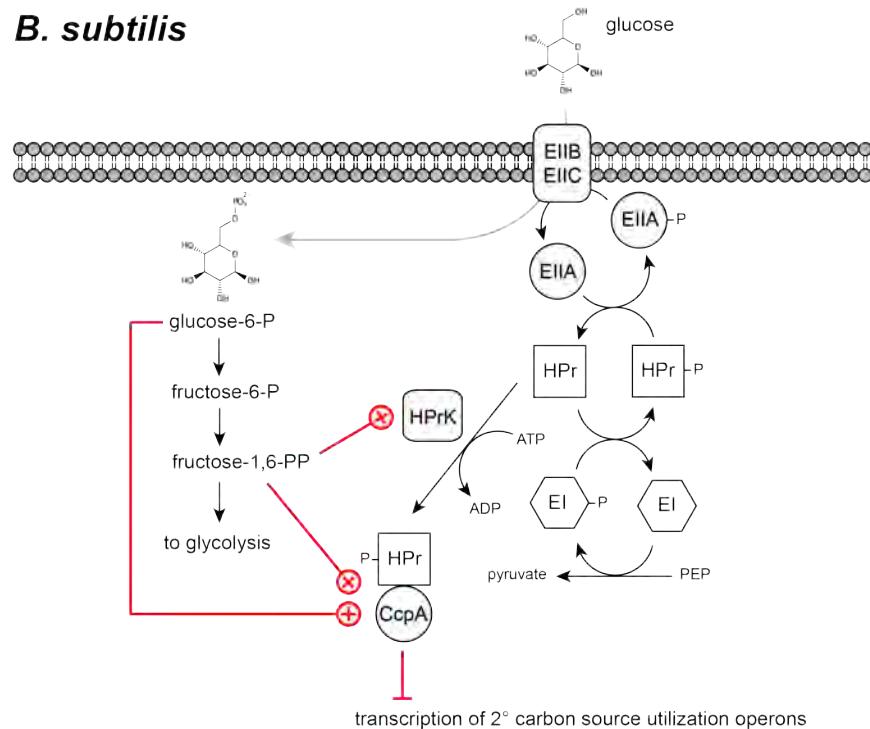


In *E. coli*, the presence of glucose results in increased flux through the glucose PTS, with the phosphate group on EIAGlc (Crr) being transferred to a glucose molecule as soon as it is available. This means that, during growth on glucose, most of the Crr is **dephosphorylated**. Dephosphorylated Crr interacts with and inhibits the activity of a wide range of metabolic enzymes and permeases required for catabolism of secondary carbon sources. This effect, and in particular the prevention of secondary carbon source import during growth on glucose, is called *inducer exclusion*.

In the **absence** of glucose, on the other hand, phosphorylated Crr (Crr-P) accumulates. Crr-P interacts with and **activates** adenylate cyclase, producing the second messenger cyclic AMP (cAMP). The global transcriptional activator CRP, when bound to cAMP, is necessary for transcription of the genes for catabolism of secondary carbon source. Finally, each operon for utilization of a secondary carbon source typically has its own local transcriptional regulation, so that the genes for catabolism of a particular carbon source (e.g. lactose or arabinose) are only actually expressed when that particular carbon source is present.

CCR in *B. subtilis* (and most other Firmicutes) also depends on the glucose PTS, but in this case it is the phosphorylation state of the **HPr** protein that exerts a regulatory effect. When levels of fructose-1,6-bisphosphate are high (indicating high glycolytic activity; see Figure 16.1), the regulatory kinase HPrK phosphorylates HPr on a site **different** from the one used for phosphotransfer to EI. This phosphorylated HPr forms a complex with the transcription factor CcpA, activating it for DNA binding. The CcpA-HPr-P complex is also stabilized directly by the glycolytic intermediates fructose-1,6-bisphosphate and glucose-6-phosphate, meaning that there are at least 3 inputs from glycolysis into CcpA activity.

B. subtilis



CcpA, unlike CRP is a **repressor** of gene expression, and, in complex with HPr-P inhibits the transcription of genes for utilization of secondary carbon sources. In the absence of glucose, the levels of glycolytic intermediates drop, HPrK becomes a phosphatase that dephosphorylates the regulatory site of HPr; the HPr-CcpA complex dissociates, and free CcpA is no longer able to bind to DNA. This frees up the genes for catabolism of secondary carbon sources to be expressed, dependent again on the activity of local regulators specific to each individual carbon source.

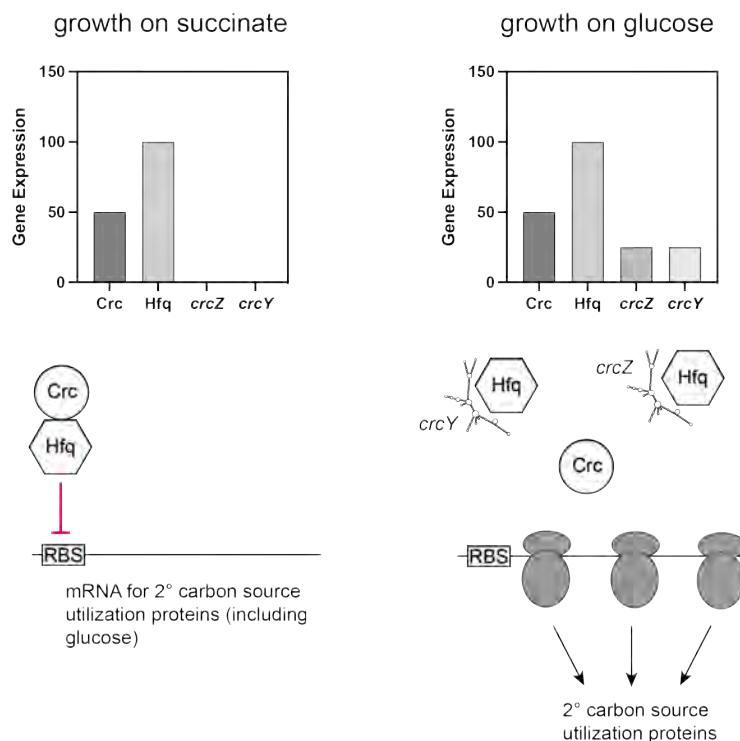
CCR does not depend on the PTS in all bacteria. For example, glucose is the preferred carbon source for *Streptomyces*, but *Streptomyces* do not use a PTS to transport glucose, instead depending on a glucose permease and glucokinase. In these bacteria, CCR depends (in a not very well-understood way) on glucokinase instead of on the PTS.

Pseudomonas species, unlike *E. coli*, *B. subtilis*, or *Streptomyces*, prefer to grow on short-chain fatty acids like succinate rather than glucose, and have a CCR system that represses utilization of other carbon sources (including glucose) when succinate is present (this is sometimes called *reverse catabolite repression*). The *Pseudomonas* CCR system is completely unlike that of *E. coli* and *B. subtilis*.

In *Pseudomonas*, in the presence of succinate the RNA-binding protein Crc, in complex with the RNA chaperone Hfq (which we mentioned way back in **Lecture 4**), binds to the ribosome binding sites of mRNAs encoding proteins necessary for utilization of secondary carbon sources, blocking their **translation**. Expression of both Crc and Hfq is constitutive, but in

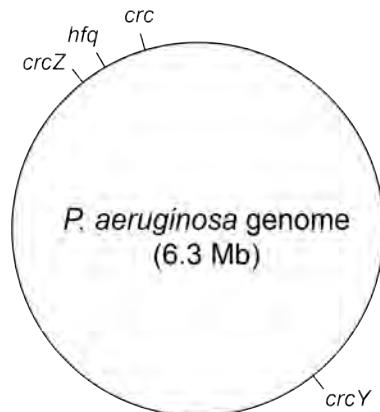
the absence of succinate *Pseudomonas* expresses two sRNAs (*crcY* and *crcZ*) that disrupt the Crc-Hfq-mRNA complex, allowing translation of those proteins, including for example, the enzymes necessary for growth on glucose.

P. aeruginosa



What **isn't** clear in *Pseudomonas* is how the transcription of the small RNAs *crcY* and *crcZ* is controlled. How does growth on succinate prevent their expression? Or does the **lack** of succinate somehow activate their expression?

Note that the *crcY* and *crcZ* genes are not located near the *hfq* or *crc* genes in the *Pseudomonas* genome.



The following methods are available for *Pseudomonas aeruginosa*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (6.3 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓

selectable & counter-selectable markers available	✓
generalized transducing phage (ϕ PA3)	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

Propose a model to explain how growth on succinate regulates *crcY* and *crcZ* expression in *P. aeruginosa*. Design an experiment to test your model. State:

- your model
- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment(s) test correlation or causation
- potential outcomes of your experiment(s), and how you will interpret them

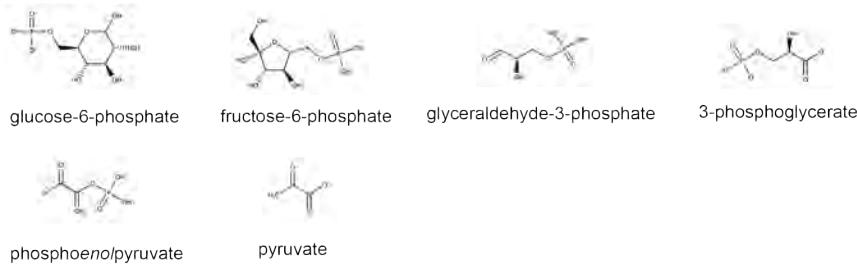
Question for discussion in class: *Mycoplasma* species do not seem to have catabolite repression, and just catabolize everything they're capable of catabolizing simultaneously. Why do you hypothesize these particular bacteria might have lost the capacity for catabolite repression?

ANABOLISM

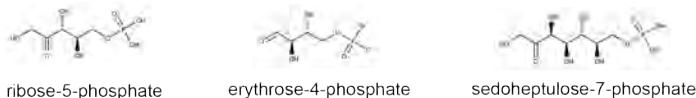
Where catabolism is the process of breaking molecules down into their component parts, anabolism is the process of synthesizing those molecules. *Prototrophic* bacteria are able to synthesize all of the amino acids, nucleotides, lipids, carbohydrates, and other cellular components they need from simple precursors (sugars, phosphate salts, and inorganic S and N compounds). *Auxotrophic* bacteria must rely on other organisms in their environment to synthesize one or more of the basic building blocks of their cells. However, even prototrophic bacteria will, given the opportunity, often salvage components from their environment rather than expend the energy to synthesize them from scratch.

Fortunately for human comprehension, the carbon backbones of **all** biomolecules are derived from just 13 precursor intermediates, which are shown in Figure 17.10. Six of these are intermediates in glycolysis, three from the pentose phosphate pathway, and four from the TCA cycle.

Glycolysis / Entner-Douderoff



Pentose Phosphate Pathway



TCA Cycle

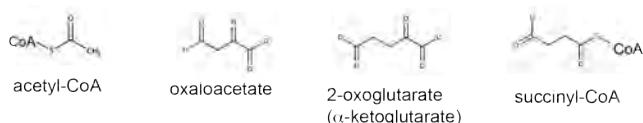


Figure 17.10. The 13 precursor intermediates needed to synthesize all biomolecules, and the central pathways from which they are derived.

Any prototrophic organism must, therefore, be able to use a simple carbon source to synthesize all 13 precursor intermediates, as well as having the additional anabolic pathways to construct other biomolecules from those precursors (see below). This is why glycolysis, the pentose phosphate pathway, and the TCA cycle are so central to metabolism, and why so many organisms have all three pathways. Fermentative bacteria ([Lecture 16](#)), which often lack the TCA cycle, are also notoriously multiply auxotrophic and typically must scavenge many amino acids, nucleotides, and cofactors from their environments.

While Figure 17.10 indicates that the precursor intermediates derived from glycolysis can also come from the Entner-Douderoff pathway, this is not **quite** true. As you can see in Figure 16.2, fructose-6-phosphate (F6P) is **not** an intermediate in that pathway, so bacteria that employ Entner-Douderoff instead of glycolysis either cannot synthesize the biomolecules derived from F6P (including many polysaccharides; [Lecture 14](#)) or must employ a separate pathway to generate F6P for biosynthetic purposes (the pentose phosphate pathway will do, since it also produces F6P as a product; Figure 17.3).

I made a point of mentioning above that the catabolic pathways of central metabolism are mostly reversible or cyclic, which means that they mostly allow cells to interconvert the precursor intermediates. That “mostly” requires a little bit of unpacking, though, and reversing central metabolism does sometimes require additional enzymes to convert catabolic pathways into anabolic ones.

In Figure 17.11, for example, you can see that anabolic gluconeogenesis, the pathway that synthesizes G6P from pyruvate, uses almost entirely the same set of enzymes as catabolic glycolysis, with two critical exceptions. In glycolysis, F6P is phosphorylated by phosphofructokinase to form fructose-1,6-bisphosphate, while in gluconeogenesis the opposite reaction is catalyzed by fructose-1,6-bisphosphatase (Figure 17.11 reaction 2). Similarly, in glycolysis, PEP is converted to pyruvate by pyruvate kinase, while in gluconeogenesis, pyruvate is converted to PEP by PEP synthetase (Figure 17.11 reaction 9). When growing on any carbon source that feeds into central metabolism below G6P (e.g. acetate, glycerol, succinate, glutamate, arabinose, etc., etc.), gluconeogenesis is required for synthesis of the upstream precursor intermediates.

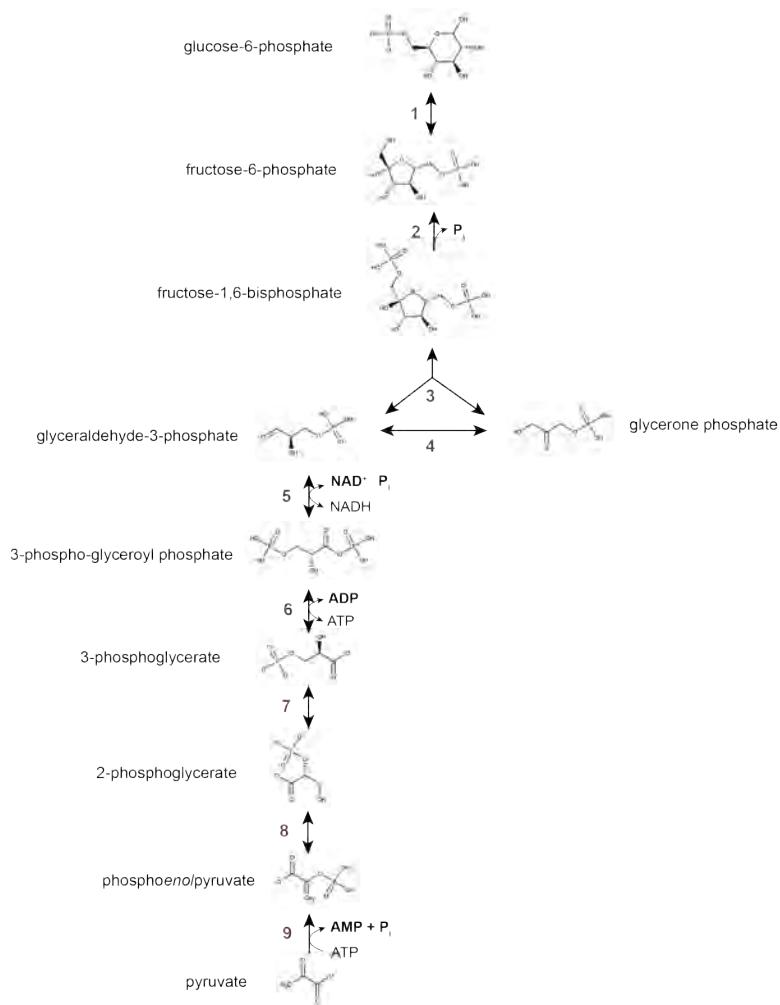


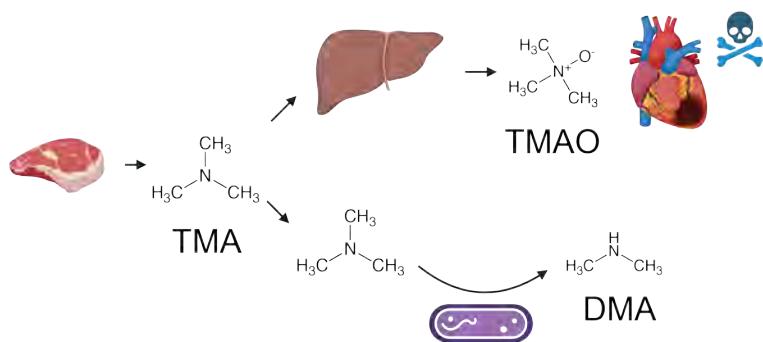
Figure 17.11. Gluconeogenesis. Enzymes: 1, glucose-6-phosphate isomerase (Pgi); 2, fructose-1,6-bisphosphatase (Fbp); 3, fructose-bisphosphate aldolase (Fba); 4, triose-phosphate isomerase (Tpi); 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pgk); 7, phosphoglycerate mutase (Pgm); 8, enolase (Eno); 9, phosphoenolpyruvate synthetase (PpsA).

All of the precursor intermediates derived from the pentose phosphate pathway are made by reversible enzymes using G3P and F6P as inputs (Figure 17.3), so no additional enzymes are required to run that pathway in reverse. The TCA cycle (Figure 17.6) is cyclic, so also does not need to run in reverse to generate precursor intermediates, although many of the enzymes involved are reversible, so in some cases, parts of the cycle can flow in the opposite of the usual direction. (For example, in organisms like *Listeria* spp. that do not have the complete set of TCA cycle enzymes.)

EXAMPLE OF AN EXPERIMENTAL DESIGN: TMA CATABOLISM IN BILOPHILA WADSWORTHIA

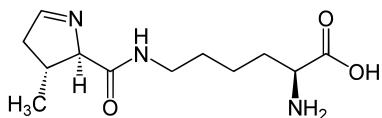
Trimethylamine (TMA) and various TMA precursors are abundant in red meat. Dietary TMA is converted in the liver into trimethylamine oxide (TMAO), a strong promoter of cardiovascular disease, which is, of course, a major cause of death worldwide.

Most bacteria cannot use TMA as a carbon or nitrogen source, but some methanogenic archaea contain a pathway for catabolism of TMA which has been pretty well characterized. The first step of that pathway involves an enzyme called TMA methyltransferase, which converts TMA into dimethylamine (DMA).



TMA is derived from a diet high in red meat, and is converted into TMAO by the liver. TMAO accumulation is a significant risk factor for cardiovascular disease. Some archaea contain a pathway for catabolism of TMA that begins with conversion of TMA into DMA. Made with Biorender.

The really unusual thing about TMA methyltransferase is that its catalytic activity depends on the incorporation of the unusual amino acid pyrrolysine in the active site.



Structure of pyrrolysine. Image from Wikipedia.

Pyrrolysine is synthesized from lysine by the PylBCD enzymes and then activated by a pyrrolysine-specific tRNA synthetase, which generates a Pyr-tRNA that recognizes the codon TAG, which is normally a stop codon. In some pyrrolysine-containing archaea, the TAG stop codon appears to have completely disappeared, while in others there appear to be some other, possibly structural, signals that determine which TAG codons will be read as stops and which will be translated by the ribosome as pyrrolysine.

The Gram-negative anaerobic bacterium *Bilophila wadsworthia* is a common inhabitant of the human intestine, and is best characterized for its ability to tolerate high bile concentrations and produce hydrogen sulfide gas (H_2S). It also increases in abundance in the microbiota of people who eat a diet high in red meat.

Genome sequencing has revealed that *B. wadsworthia* encodes homologs of TMA methyltransferase and all of the genes necessary for pyrrolysine synthesis. This has led to the hypothesis that perhaps *B. wadsworthia* is capable of using the TMA methyltransferase pathway to grow on TMA, potentially reducing the amount of toxic TMAO that accumulates in the gut.

The following methods are available for *Bilophila wadsworthia*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (4.8 Mbp)	✓
susceptible to mutagens	✓

As far as I can tell, there are no examples of gene knockouts or plasmid transformation in *B. wadsworthia* in the literature. Whether this is because no one has tried to develop such tools or because *B. wadsworthia* is particularly recalcitrant I don't know.

Design a genetic experiment to test the model that the pyrrolysine-dependent TMA methyltransferase pathway is involved in TMA catabolism in *B. wadsworthia*.

Hypothesis: The pyrrolysine-dependent TMA methyltransferase pathway is involved in TMA catabolism in *B. wadsworthia*.

Experimental Design:

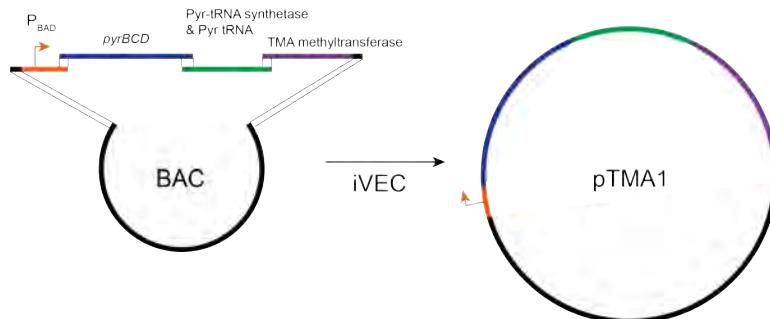
This is tricky, since we really don't have any useful genetic tools in our study organism. However, we **do** have a full set of candidate *B. wadsworthia* genes that are homologous to the known archaeal pathway for TMA catabolism.

So, rather than trying to make mutants in *B. wadsworthia* and screen for loss of TMA catabolism, we can try to reconstitute the *B. wadsworthia* TMA pathway in a more tractable organism. *E. coli* is probably the simplest, since it's easy to

manipulate and has no known pathways for TMA breakdown. In fact, under some circumstances *E. coli* produces TMA, since it can use trimethylamine oxide (TMAO) as a terminal electron acceptor in anaerobic respiration ([Lecture 16](#)).

We do have to bring together quite a few different genes here, so in order to make the cloning process somewhat simpler, I'm choosing to use a BAC vector ([Lecture 6](#)), which will allow me to stably maintain a large number of genes in *E. coli* in one single-copy plasmid.

- 1) Have DNA fragments synthesized with each of the predicted *B. wadsworthia* genes necessary for TMA catabolism (TMA methyltransferase, PyrBCD, pyrrolysine tRNA synthetase, and Pyr-tRNA), with *E. coli* RBSs and a P_{BAD} arabinose-inducible promoter, and with ~30 bp overlaps to allow iVEC or Gibson Assembly cloning into a BAC vector. Call the resulting plasmid pTMA1.



Gene synthesis is a practical approach here, if only because it means we don't have to purify a bunch of *B. wadsworthia* DNA to amplify genes from, and it lets us generate overlaps very precisely.

- 2) Transform *E. coli* with pTMA1 and incubate in media containing arabinose and TMA.
- 3) Measure DMA accumulation over time (this turns out to be easiest to do by gas chromatography).

Independent Variable: (what will you change?)

- Expression of the putative *B. wadsworthia* TMA pathway.

Dependent Variable(s): (what will you measure?)

- Production of DMA (a quantitative measurement)

Negative Control(s): (eliminate false positive results)

- 1) An *E. coli* vector-only control should not produce any DMA.
- 2) *E. coli* / pTMA1 without arabinose should not produce any DMA either (or at least much less than a culture with arabinose added).

Positive Control(s): (eliminate false negative results)

- 1) *B. wadsworthia* should produce DMA from TMA when grown under these conditions.
- 2) Confirm that *E. coli* is not killed by the concentration of TMA in the test conditions (TMA is potentially toxic).
- 3) Use qRT-PCR to confirm transcription of the genes in pTMA1 under these growth conditions.

Potential Outcomes:

- 1) *E. coli* transformed with pTMA1 converts TMA into DMA. This supports the hypothesis.
- 2) *E. coli* transformed with pTMA1 does not convert TMA into DMA. The interpretation of this result is harder, since it could mean that the hypothesis is wrong **or** that the plasmid construct is not working somehow (perhaps translation is poor, the pyrrolysine tRNA is not recognized by *E. coli* ribosomes, or there are additional genes needed that we could not predict from sequence homology). Confirming protein expression and correct Pyr-tRNA assembly might end up being useful in troubleshooting this, but it's also possible that the approach of reconstituting the pathway in *E. coli* will simply not work.

Does this experiment establish correlation or causation?:

- Because we are adding specific genes to *E. coli* and seeing if they confer a very specific phenotype, this design does test **causation**, although it is not necessarily testing whether the genes in question cause that phenotype in *B. wadsworthia*. Kind of an edge case, honestly.

This is a pretty common synthetic biology approach to this kind of problem, but it is one with significant limitations. A positive result is informative, but a negative result is much less so. If you were to propose an experiment like this in a grant (or a qualifying exam), expect the reviewers to give you a hard time about what you will do if this strategy fails.

In the last portion of this chapter, I will explore a few anabolic pathways in detail, to show how the precursor intermediates are used as the basis of essential cellular components like sugars, lipids, amino acids, and nucleotides. As with catabolism, I will only be able to touch on a few examples, and the underlying principles are more important than the details of each pathway.

G6P and F6P are the precursors of many other sugars. For biosynthetic purposes, G6P is usually first activated by conversion into a nucleotide-bound form (usually uridine diphosphate, forming UDP-glucose), then modified by other enzymes. As an example, the synthesis of UDP-galactose is illustrated in Figure 17.12. Notably, UDP-glucose is also a precursor of the LPS core ([Lecture 10](#)), and various nucleotide-activated sugars are the substrates for many pathways of polysaccharide biosynthesis ([Lecture 14](#)).

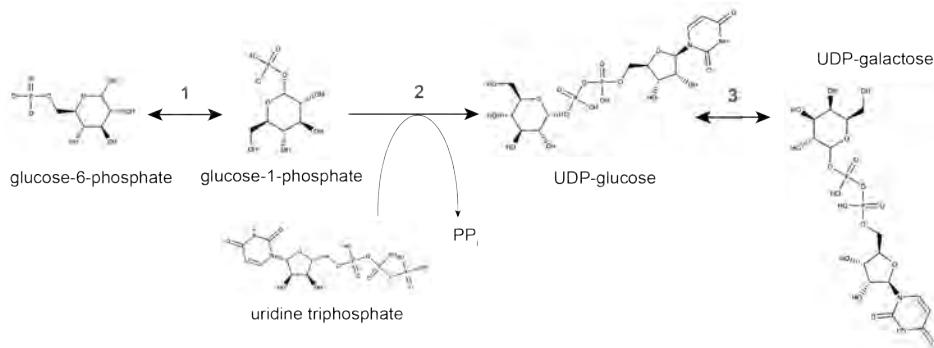


Figure 17.12. Galactose biosynthesis from G6P. Enzymes: 1, phosphoglucomutase; 2, UTP:glucose-1-phosphate uridylyltransferase; 3, UDP-glucose 4-epimerase.

Fatty acids and other lipids are synthesized from acetyl-CoA, as shown in Figure 17.13. There is an initiation pathway (Figure 17.13A), in which two acetyl-CoA are combined to generate an acetoacetyl chain, and then an elongation cycle (Figure 17.13B) in which the products of the initiation pathway are concatenated into longer fatty acid chains.

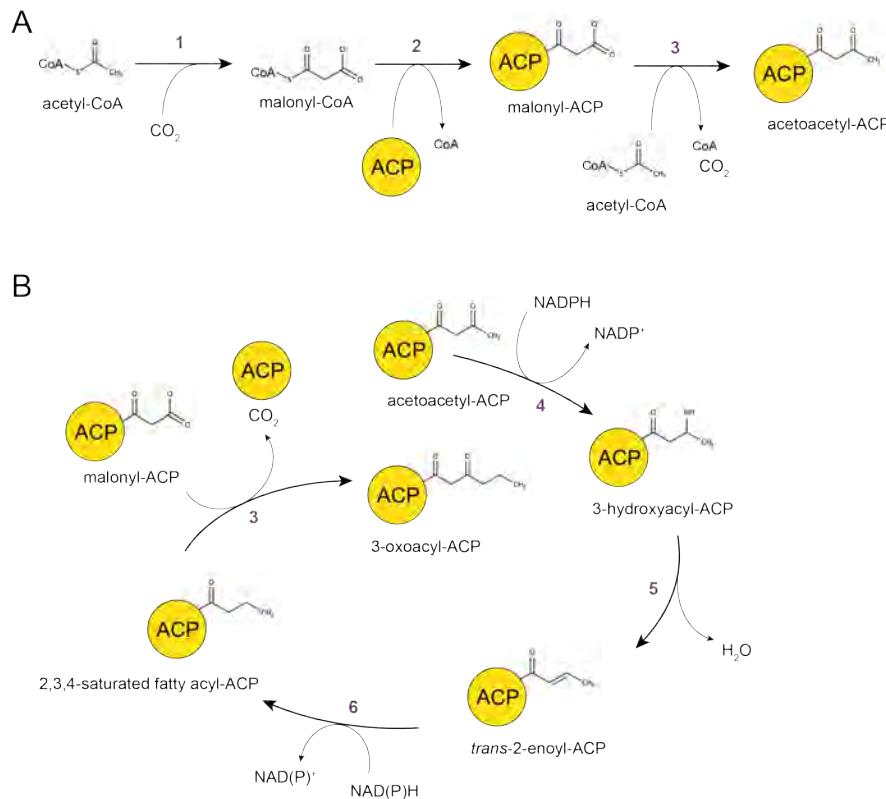
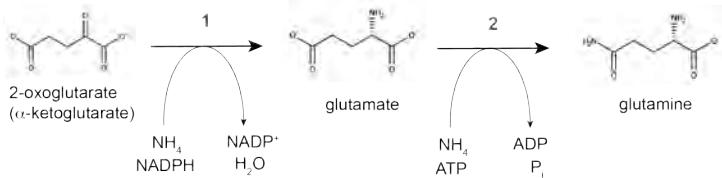


Figure 17.13. Fatty acid biosynthesis from acetyl-CoA. (A) Initiation and (B) elongation cycle. ACP = acyl-carrier protein. Enzymes: 1, acetyl-CoA carboxyl-transferase; 2, ACP S-malonyltransferase; 3, β -ketoacyl-ACP synthetase; 4, 3-oxoacyl-ACP reductase; 5, 3-hydroxyl-ACP dehydratase; 6, enoyl-ACP reductase.

The growing fatty acid chain is bound to a small **acyl-carrier protein** (ACP) while it is being synthesized, and each turn of the elongation cycle adds 2 carbons to the ACP-bound fatty acid. ACP is needed to keep the hydrophobic lipid soluble in the cytoplasm and accessible to the synthesis enzymes. Once the fatty acid is complete, it is cleaved off of the ACP at the inner membrane, where those lipids are then incorporated. Fatty acids of different lengths are generated with different numbers of elongation cycles, and additional enzymes are involved in inserting double bonds and other modifications to generate different fatty acid types (**Lecture 10**). Fatty acid biosynthesis consumes a large number of reducing equivalents (mostly NADPH).

Synthesizing amino acids requires incorporation of N and S atoms in addition to the C backbones derived from the precursor intermediates. This requires reduced N and S donors, typically NH_4^+ and H_2S , which are initially used to synthesize glutamate and cysteine, respectively (Figure 17.14). Glutamate is derived from the precursor intermediate 2-oxoglutarate, and cysteine from 3-phosphoglycerate and acetyl-CoA, by way of serine. Many organisms can reduce other inorganic S sources in their environments (sulfate, sulfite, thiosulfate, etc.) into H_2S for incorporation into cysteine.

A



B

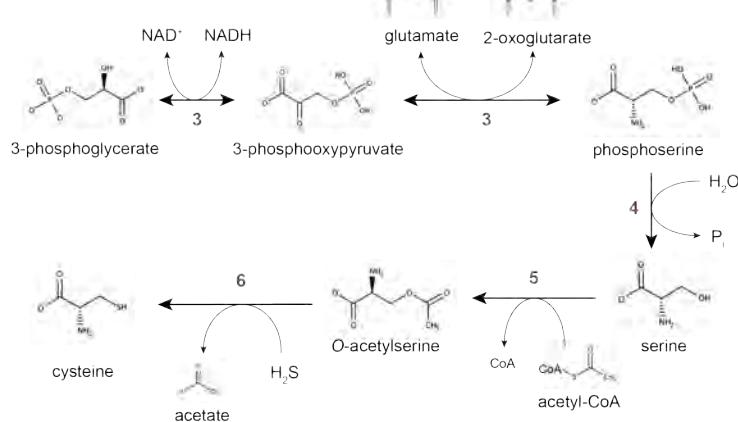


Figure 17.14. Examples of important amino acid biosynthetic pathways. (A) Glutamate and glutamine biosynthesis. Enzymes: 1, glutamate synthase; 2, glutamine synthetase. (B) Serine and cysteine biosynthesis. Enzymes: 3, 3-phosphoglycerate dehydrogenase (yes, it really does both of these reactions); 4, 3-phosphoserine aminotransferase; 5, phosphoserine phosphatase; 6, serine acetyltransferase; 7, O-acetylserine sulfhydrylase.

Glutamate is by far the most abundant metabolite in bacterial cells (present at about 100 mM in the cytoplasm of *E. coli*), and either glutamate or glutamine are the N donors for most biosynthetic pathways (note glutamate donating an amino group in reaction 3 in Figure 17.14 B). Cysteine is the S donor for most biosynthetic pathways that require S, and while cysteine itself is not especially abundant in most cells, cysteine-derived low molecular weight thiols, such as the tripeptide glutathione (glutamine-cysteine-glycine), are present at millimolar concentrations as reducing agents that help maintain the redox potential of the cytoplasm.

As a last example of a biosynthetic pathway, we'll look at the somewhat more complicated mechanism by which nucleotides are assembled, and specifically at the pathway that leads to the pyrimidine nucleotide cytidine-5'-triphosphate (CTP) (Figure 17.15). This is mostly to illustrate that anabolic pathways can become quite involved, with large numbers of enzymes needed to convert precursor intermediates into their final products.

The pyrimidine base is assembled from bicarbonate, glutamine, and aspartate (itself derived from an oxaloacetate and a glutamate; not shown), and then combined with the ribose-5-phosphate derivative 5-phospho-ribose-1-diphosphate (PRPP) to form the nucleotide precursor orotidine-5'-phosphate. Orotidine-5'-phosphate, in turn, is decarboxylated to form uridine-5'-phosphate (UMP), the precursor of all the pyrimidines in DNA and RNA (CTP, dCTP, TTP, dTTP) and an important player in carbohydrate biosynthesis in its own right (see Figure 17.12 and **Lecture 14**).

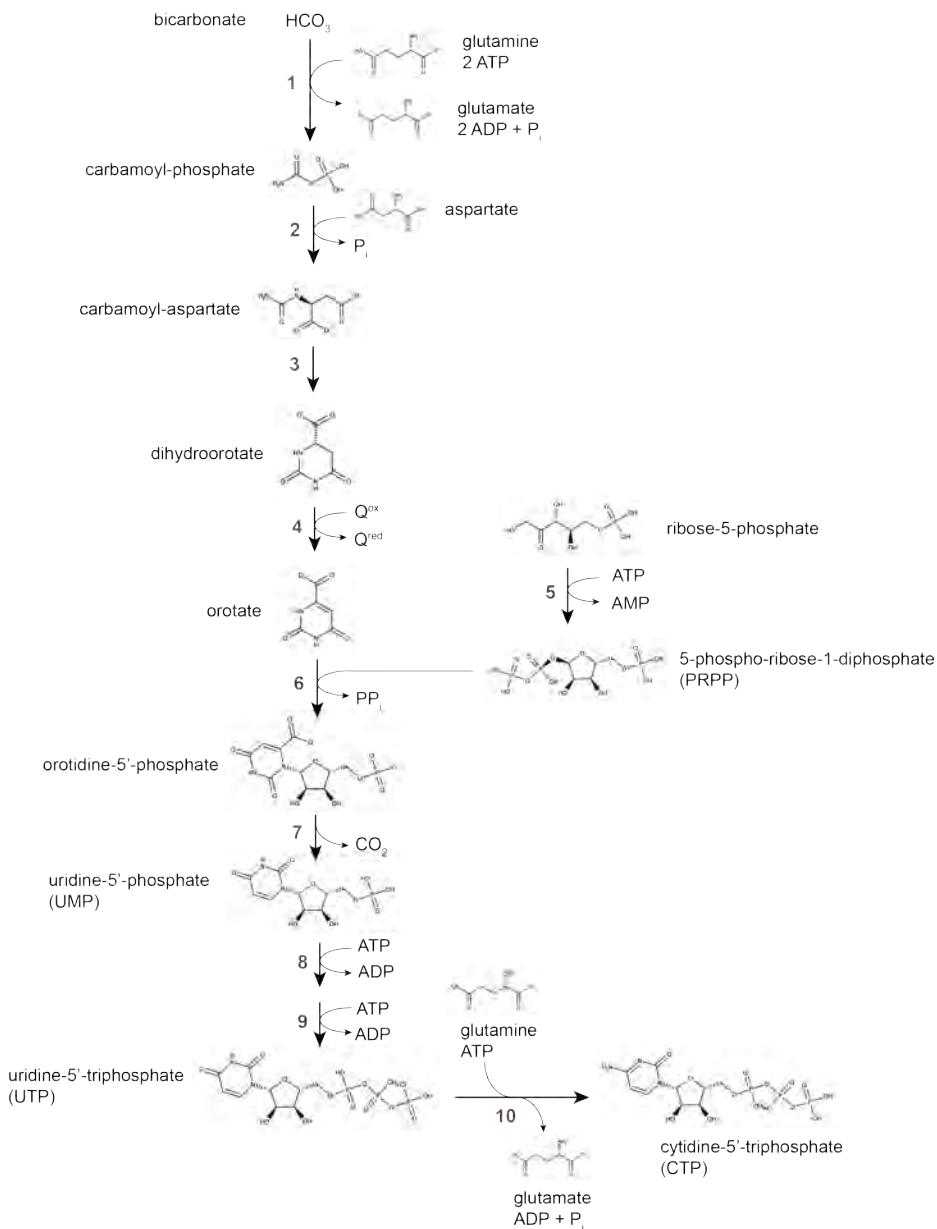


Figure 17.15. Pyrimidine nucleotide biosynthesis. Enzymes: 1, carbamoyl-phosphate synthetase; 2, aspartate carbamoyl transferase; 3, dihydroorotase; 4, dihydroorotate dehydrogenase; 5, ribose-phosphate diphosphokinase; 6, orotate phosphoribosyltransferase; 7, orotidine-5'-phosphate decarboxylase; 8, UMP kinase; 9, UDP kinase; 10, CTP synthetase.

PRPP is an important biosynthetic intermediate, and is a precursor of not only pyrimidines, but purines, the amino acids histidine and tryptophan, and the nicotinamide cofactors (NAD and NADP), among other compounds.

LECTURE 18: SECONDARY METABOLISM

INTRODUCTION

Secondary metabolism (a subset of anabolism) encompasses the synthesis of an enormous range of biological molecules that do **not** make up the bulk constituents of cells (i.e. are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids). These include a wide range of biologically interesting but non-essential compounds like antibiotics, quorum sensing autoinducers, electron shuttles, metal-binding siderophores, and many more.

In this chapter, we will examine the pathways by which a representative set of these secondary metabolites are synthesized, although the nature of the topic means that this will necessarily be a **very** incomplete survey.

WHAT IS SECONDARY METABOLISM?

The original terminology for secondary metabolism was drawn from the fact that these molecules are species- or strain-specific, as opposed to the primary metabolites that are found in all cells ([Lecture 17](#)). Some people have argued that secondary metabolites are important to the organisms that produce them, and they should therefore be called “specialized metabolites” instead, but I will use the more commonly-accepted term here. Another term you will encounter in the field is *natural products*, an extremely vague category which includes essentially any molecule produced by a living organism. It is useful in comparison to the (usually) simpler molecules generated by organic chemists or to human-modified variants of naturally occurring compounds.

Essentially all bacteria produce at least a few secondary metabolites, and there is great diversity in the metabolite repertoire of strains within a single species. Different strains of *B. subtilis* produce widely differing sets of surfactants and anti-fungal compounds, for example, and different strains of *E. coli* produce different siderophores that bind to and allow the incorporation of various metal ions (especially iron), which can contribute to their ability to cause infections.

Secondary metabolites have a vast range of biological functions, and we actually have very little idea of what the physiological function of many bacterial metabolites might be, even those which have turned out to be extremely useful for human purposes, for example as anti-cancer or anti-inflammatory drugs. (More than 30% of all FDA-approved drugs are natural product derivatives.) *Cryptic metabolites* are secondary metabolites with no known function. You will often see secondary metabolites with known effects on cells referred to as *bioactive compounds*, but of course all metabolites are presumably bioactive in some context for the organisms that produce them.

Many secondary metabolites are synthesized by *biosynthetic gene clusters* (BGCs), which are genetic loci dedicated to the production of a specific secondary metabolite. These can be very large (100 kb or more), and contain multiple operons with complex regulation. *Silent BGCs* are BGCs that are not expressed under laboratory conditions, a problem we will return to later in the chapter. The grouping of secondary metabolite synthesis genes into BGCs means that production of secondary metabolites can fairly easily be horizontally transferred between strains and species, although this can sometimes require the transfer of very large DNA fragments.

Of course, as microbiologists, the secondary metabolites that probably have the most direct relevance to our work are antibiotics. Most of the families of antibiotics in current use are derivatives of natural products, almost all from different species of actinomycetes, especially members of the genus *Streptomyces*. The evolution of pathogenic bacteria resistant to clinical antibiotics is a problem which has led to a great deal of discussion and concern, with wide interest in discovering novel secondary metabolites that may be useful as new antimicrobial treatments. Kim Lewis, director of the Antimicrobial Discovery Center at Northeastern University, has written a thorough, thoughtful [perspective article in Cell](#) on the history and current state of antibiotic discovery science (as of 2020), and explains several reasons to be optimistic about our ability to overcome the looming antibiotic resistance crisis. This is, of course, a very active area of research.

There are many resources available for the study of secondary metabolism. The website secondarymetabolites.org is a convenient clearinghouse for many useful bioinformatic tools and databases.

PHENAZINE SYNTHESIS

Some secondary metabolites are synthesized by pathways very much like the anabolic central metabolic pathways we discussed in [Lecture 17](#). As our representative example, we will look at the synthesis of pyocyanin, the blue phenazine that lends a characteristic greenish tinge to stationary phase cultures of *Pseudomonas aeruginosa*.

The initial steps of pyocyanin synthesis are the same as those for the aromatic amino acids phenylalanine, tryptophan, and tyrosine, and begin with the precursor intermediates phosphoenolpyruvate and erythrose-4-phosphate (Figure 17.10). Phenazine synthesis branches off from those amino acid synthesis pathways in the fate of chorismate (Figure 18.1, which is somewhat simplified to make everything fit on the page). The basic conjugated three-ring phenazine scaffold (in the form of phenazine-1-carboxylic acid) is formed by combination of two chorismate molecules, and subsequent enzymes modify this scaffold to form specific phenazines, in this case, of course, pyocyanin.

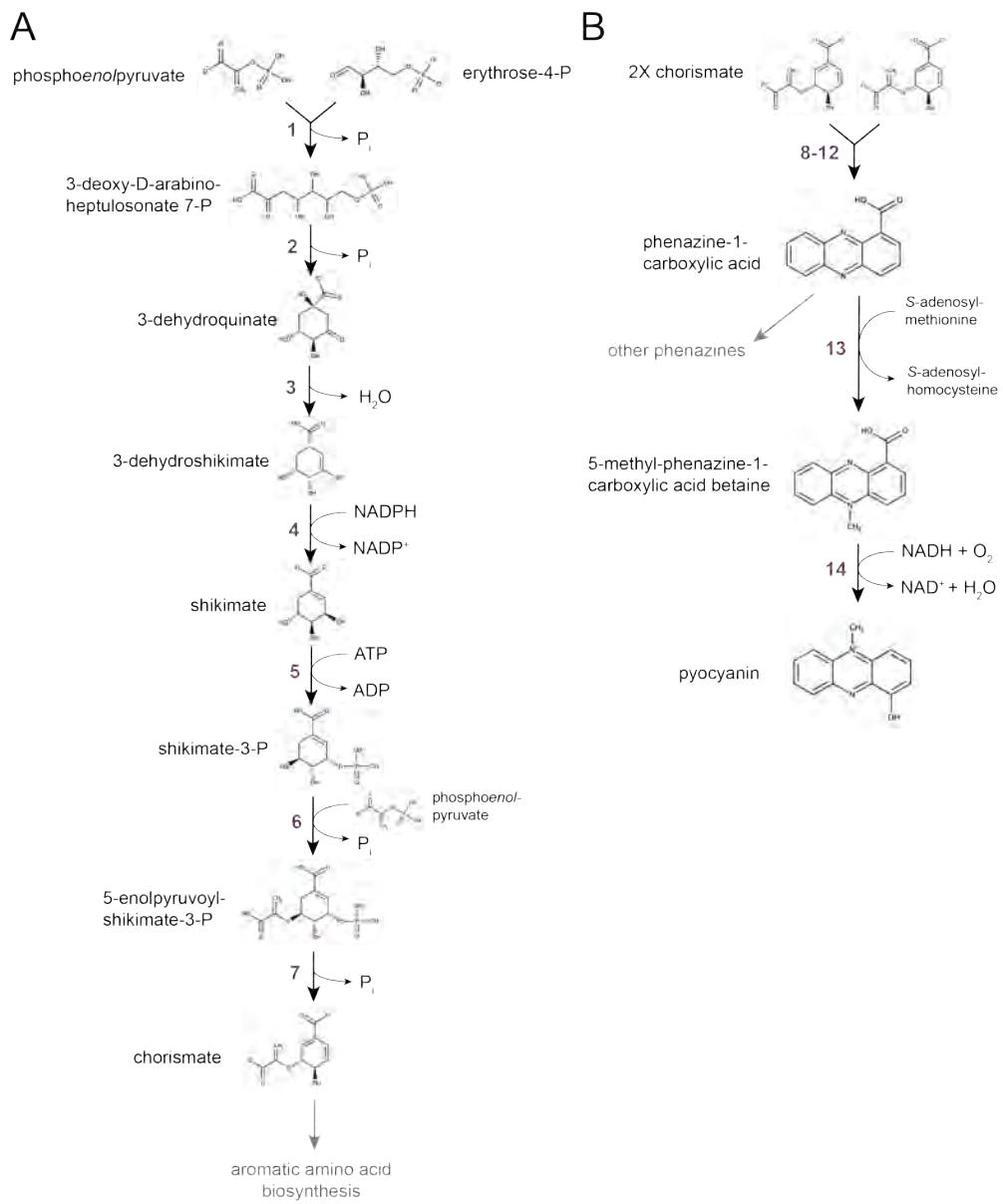


Figure 18.1. Synthesis of the phenazine pyocyanin in *Pseudomonas aeruginosa*. A = shared steps with aromatic amino acid biosynthesis; B = phenazine synthesis pathway. Enzymes: 1, 3-deoxy-7-phosphoheptulonate synthase (AroGHH); 2, 3-dehydroquinate synthase (AroB); 3, 3-dehydroquinate dehydratase (AroD); 4, shikimate dehydrogenase (AroE); 5, shikimate kinase (AroKL); 6, 3-phosphoshikimate-1-carboxyvinyltransferase (AroA); 7, chorismate synthase (AroC); 8 – 12, five enzymatic steps (PhzEDFBG); 13, phenazine-1-carboxylate N-methyltransferase (PhzM); 14, 5-methylphenazine-1-carboxylate 1-monooxygenase (PhzS).

Pyocyanin is a particularly interesting metabolite because it illustrates the difficulty of assigning a single physiological function to secondary metabolites. Pyocyanin is often described as a virulence factor, and indeed it **is** highly toxic to human cells and makes a major contribution to tissue damage during *P. aeruginosa* infections, but this is only one of its functions, and (from the point of view of *P. aeruginosa*), probably not the most important.

In work pioneered by the lab of Dianne Newman at CalTech, the roles of pyocyanin and other phenazines in *P. aeruginosa* biofilms have recently become more clear. *P. aeruginosa* is an obligate respiring organism (**Lecture 16**), but characteristically forms thick biofilms during infections (**Lecture 14**). Oxygen is not able to penetrate to the inner layers of the biofilms, but phenazines can diffuse freely through the matrix and cell membranes. Phenazines are redox-active electron shuttles, and *P. aeruginosa* can use oxidized pyocyanin as a terminal electron acceptor for respiration (**Lecture 16**). The resulting reduced pyocyanin diffuses towards the outer surface of the biofilm, where it is oxidized by O₂. Phenazines therefore allow deeply-buried cells in biofilms to respire O₂ at a distance. That's not all that pyocyanin does, though.

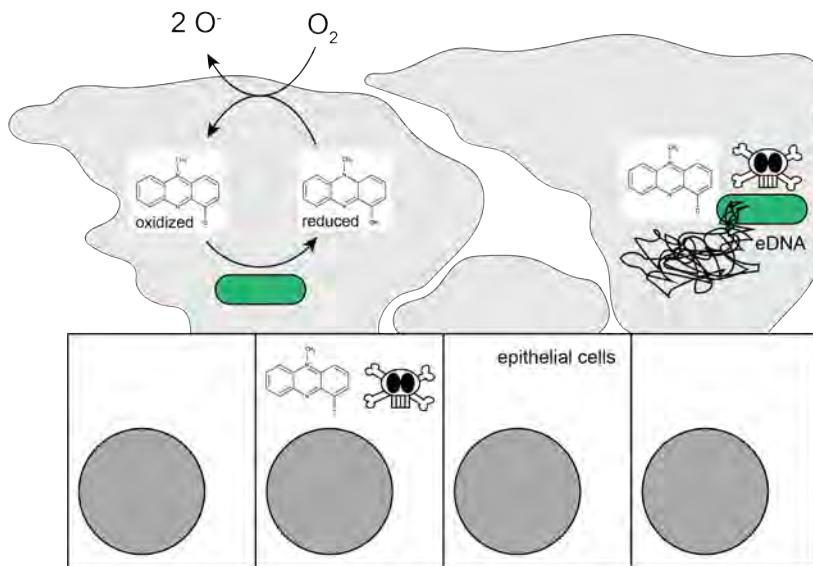


Figure 18.2. Distinct functions of pyocyanin in *P. aeruginosa* biofilms and infections. Redox cycling of pyocyanin allows deeply-buried *P. aeruginosa* to respire, but starved bacteria may be killed by pyocyanin toxicity, releasing eDNA to reinforce the biofilm matrix. Pyocyanin is also a potent toxin that kills host cells, contributing directly to pathogenesis.

The most abundant polymer in the matrix of a *P. aeruginosa* biofilm is eDNA, which is released from bacteria by lysis of part of the population. Pyocyanin also plays a key role in this process. When *P. aeruginosa* is starved for carbon (and therefore for ATP; [Lecture 16](#)), *P. aeruginosa* can no longer cope with the toxic effects of pyocyanin, which mostly result from production of reactive oxygen species, leading to the death of about 90% of the starved cells. This lysis releases large amounts of eDNA, which stabilizes the biofilm as a whole, making it more resistant to physical disruption.

The Newman lab published [a paper](#) in Science in 2021 that shows that phenazines **also** function as phosphorus-solubilizing agents that are important for growth in phosphate-limiting conditions, showing that neither respiration nor toxicity fully encompass the physiological roles of these metabolites. To make things more complicated yet, phenazines can play a regulatory role, controlling bioenergetics and antibiotic tolerance (reported [here](#)), and Lars Dietrich published [a paper](#) in 2023 showing that *P. aeruginosa* regulates the activity of phenazine by methylating it, in a pathway under carbon catabolite repression control (recall Discussion Problem # 32 in [Lecture 17](#)). Methylated phenazine appears to be both more effective as a respiratory electron shuttle, but also more toxic, so the bacteria must balance these activities depending on their metabolic state and position in a biofilm.

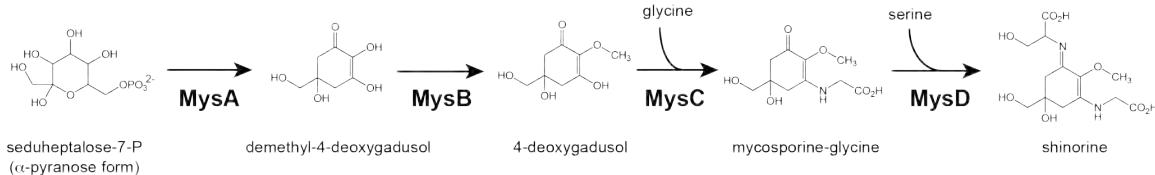
Pyocyanin clearly plays a dynamic physiological role in *P. aeruginosa* biofilms and infections, with different functions under different growth conditions and stages. There is no particular reason to think that a given metabolite, secondary or otherwise, ever has just one physiological function.

EXAMPLE OF AN EXPERIMENTAL DESIGN: IDENTIFYING MECHANISMS OF STRESS RESISTANCE

Many cyanobacteria which live in the open water are exposed to high doses of UV radiation from the sun and produce secondary metabolites which act as protective UV-absorbing sunscreens. One such molecule is shinorine, which is a member of the mycosporine-like amino acid (MAA) chemical family. Shinorine has an absorbance maximum of 333 nm and a molar extinction coefficient of $44,700 \text{ M}^{-1} \text{ cm}^{-1}$, which makes it one of the strongest UV-absorbing molecules found in nature.



Shinorine is synthesized from sedoheptulose-7-phosphate (an intermediate in the pentose phosphate pathway; [Lecture 17](#)), glycine, and serine. The *mysABCD* genes encoding the enzymes needed for this pathway have been identified in the model cyanobacterium *Nostoc punctiforme*.



Mutants that lack shinorine (e.g. *mysA* knockouts) are, unsurprisingly, very sensitive to UV radiation, but they are also sensitive to other stresses, including osmotic stress (high salt) and heavy metals (e.g. chromium). The connection between shinorine production and survival of these stresses is not obvious.

The following methods are available for *N. punctiforme*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (9.1 Mbp)	✓
susceptible to mutagens	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
CRISPR and related technologies (e.g. CRISPRi)	✓

Design an experiment to identify potential mechanisms by which shinorine might protect *N. punctiforme* against non-UV stresses.

This is tricky problem, and not one that is easily addressed in a systematic way. Different proposals have been made, for example suggesting that shinorine might quench free radicals or chelate toxic metals, but these seem to mostly be guesses based on the chemical properties of purified shinorine. The difficult question is this: can we get unbiased information about what role(s) shinorine might be playing *in vivo*?

Here's one possible way to address that question:

Hypothesis: Shinorine-defective *N. punctiforme* will upregulate different stress response pathways in response to non-UV stress than wild-type *N. punctiforme*, and what these pathways are will indicate what kind of protective effect(s) shinorine has under those conditions.

This probably requires a little unpacking. When bacteria are exposed to stress, they upregulate different responses depending on what that stress is. We've already mentioned the SOS response to DNA damage (in **Lecture 16**) and (very briefly) the heat shock response to protein folding damage (in **Lecture 4**), but there are also reasonably well-characterized stress response pathways to membrane damage, metal accumulation, oxidative stresses, and different kinds of starvation, among others. Each of these pathways includes well-conserved characteristic genes specific to each stress condition, and many of them are controlled at the transcriptional level.

The idea here is to remove shinorine from the equation, then expose the cells to a stress and examine genome-wide gene transcription in those cells. Whatever category of stress shinorine prevents will now be more severe, and therefore that specific response will be upregulated more strongly. So, for example, if shinorine quenches reactive oxygen species, a ΔmysA mutant would be expected to upregulate its oxidative stress defense pathways more strongly than the wild-type under ROS-producing stress conditions.

Experimental Design:

- 1) Harvest mRNA from *N. punctiforme* wild-type and ΔmysA cells exposed to no stress, UV light, high salt, or chromium.
- 2) Use RNA sequencing to characterize the transcriptome of those cells.

Independent Variable: (what will you change?)

- 1) The presence or absence of stress
- 2) The presence or absence of shinorine synthesis

Note that this is a case where we have more than one independent variable in our experiment. This means we will have to be particularly careful about controls and interpretations.

Dependent Variable(s): (what will you measure?)

- Transcription of each gene in the *N. punctiforme* genome (a quantitative measurement)

Negative Control(s): (eliminate false positive results)

- 1) The no-stress condition for the wild-type cells is a key control here, to establish a baseline of expression for each gene to compare everything to.
- 2) Comparing the no-stress transcriptome of the wild-type to the Δ mysA mutant is also important. Ideally, with no stress applied, there would be no effect of the loss of shinorine on the transcriptome, but this is honestly unlikely. What is critical is to make sure that we know what that effect is, so that we can account for it when examining different stress conditions. It might also tell us something about what role shinorine plays in supposedly non-stressful conditions.

Positive Control(s): (eliminate false negative results)

- 1) The comparison between the wild-type and mutant cells under UV light stress is a key positive control in this experiment. UV light is primarily a DNA-damaging stress (**Lecture 3**), so we can make a strong prediction that the Δ mysA mutant should upregulate the SOS response more than the wild-type will under these conditions. If that's not true, we will need to reevaluate our experimental design.
- 2) It will be important to titrate the dose of each stress. We want to make sure to have a condition where the Δ mysA mutant is inhibited more than the wild-type is, but is still able to mount some kind of stress response and recover. If they just die, we will not get any useful information out of them.

Potential Outcomes:

- 1) We identify stress response pathways more strongly upregulated in salt- or chromium-treated Δ mysA mutant *N. punctiforme* than in the wild-type. This information will allow us to develop models for what kind of cellular damage is prevented by shinorine under those stress conditions, which we can then test in future experiments.
- 2) We do not identify any stress response pathways more strongly upregulated in salt- or chromium-treated Δ mysA mutant *N. punctiforme* than in the wild-type. This might indicate that shinorine protects against a kind of damage that *N. punctiforme* has no other defense against, or that whatever alternative stress response pathways are compensating for the lack of shinorine are not transcriptionally regulated.

Does this experiment establish correlation or causation?:

- This experiment tests the **correlation** between gene expression and salt and chromium stress in *N. punctiforme*.

This is primarily a hypothesis-generating experiment, and as noted, even positive results will need to be validated with other methods. However, this kind of approach can be extremely helpful for coming up with models for how complex responses might function.

ACYL-HOMOSERINE LACTONE SYNTHESIS

The AHL quorum sensing signal molecules (previously mentioned in **Lecture 14**) are another large class of secondary metabolites whose structures vary widely among bacteria. However, their synthesis proceeds by a fairly simple pathway.

The hydrophobic acyl portion of the AHL is synthesized on ACP by the same kind of pathway used to synthesize other lipids, which we have already discussed (Figure 17.13). An AHL synthase enzyme then catalyzes the reaction of that acyl-ACP with S-adenosylmethionine, yielding ACP, 5'-methyl-thioadenosine, and the desired AHL (Figure 18.3).

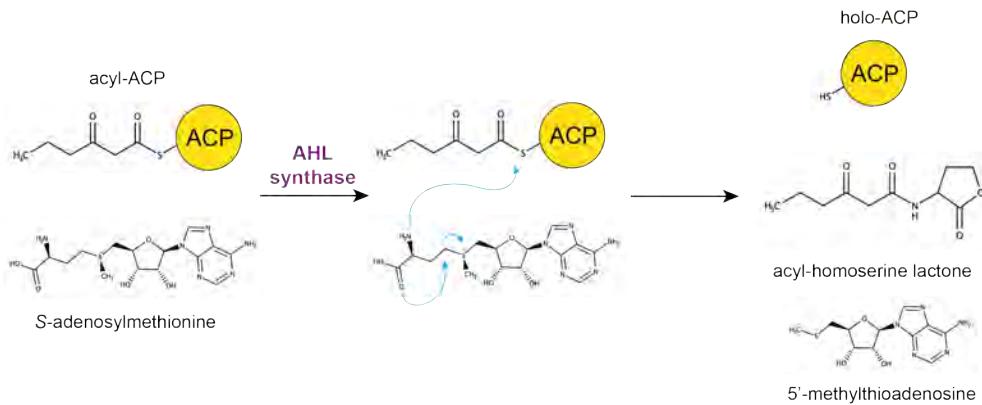


Figure 18.3. Acyl-homoserine lactone synthesis by AHL synthase, using an acyl-ACP and S-adenosylmethionine as substrates.

S-adenosylmethionine (often abbreviated SAM or AdoMet), which is synthesized from a methionine and an ATP, is a common cosubstrate involved many kinds of anabolic reaction. EcoCyc lists [87 known reactions](#) that consume SAM in *E. coli* K-12, a strain that lacks AHL-based quorum sensing pathways. The reaction catalyzed by AHL synthases is an unconventional SAM-dependent process, in that it destroys the SAM, and is not at all typical of the catalytic methyl transfer reactions SAM is normally involved in.

POLYKETIDE SYNTHASES

The really complicated secondary metabolites, though, are synthesized by two types of modular biosynthetic machines: the **polyketide synthases** (PKS) and the **non-ribosomal peptide synthetases** (NRPS), which construct large, complex molecules out of simple carboxylic acid and amino acid components, respectively. The **megasynthase** enzymes involved in some PKS and NRPS pathways are among the largest proteins synthesized by bacteria, with the largest known being proteins over 2 MDa in molecular weight, more than 17,000 amino acids long, encoded by genes spanning more than 50 kb (> 1% of an entire bacterial genome). Both PKS and NRPS pathways are broadly conserved across the tree of life, and are capable of synthesizing a vast array of compounds. Here, we will touch briefly on how each of these pathways typically works in bacteria, with the understanding that, once again, I am **vastly** oversimplifying the true complexity and diversity of mechanisms that exist.

There is, thankfully, a more or less standardized nomenclature for PKS and NRPS components, which I will use throughout this section.

Like fatty acid synthases ([Lecture 17](#)), PKS construct their products from acyl-carrier protein (ACP)- or CoA-bound carboxylic acid subunits. These can include malonyl-ACP (the precursor used for fatty acid synthesis), but also a variety of other acyl-ACPs, like acetyl-ACP, benzoyl-ACP, and ethylmalonyl-ACP.

PKS are large, highly modular enzymes (Figure 18.4). They contain several ACP-like domains on which the products are assembled, as well as additional enzymatic domains that carry out key steps in synthesis. These include:

- acyltransferases (abbreviated AT)
- β -ketoacylsynthetases (KS)
- ketoreductases (KR)
- dehydratases (DH)
- enoylreductases (ER)
- methyltransferases (MT)
- cyclases (CYC)
- aromatases (ARO)
- thioesterases (TE)

In any PKS, the first step is transfer of an initiating acyl group from acyl-ACP to the first ACP domain of the PKS by an AT domain. Each subsequent acyl-CoA building block is added by the next module's KS and AT domains in an elongation stage, transferring the growing polyketide chain to the next ACP domain. Optionally, at each step there may be any combination of DH, ER, MT, CYC, or ARO domains which chemically modify the resulting product. Once the polyketide reaches the final ACP domain, the TE domain cleaves it off of the PKS, often including a cyclization step.

The smallest possible PKS, which would combine two acyl group-containing compounds with no additional modifying domains, therefore consists of the following domain structure, with a minimal module enclosed in brackets:

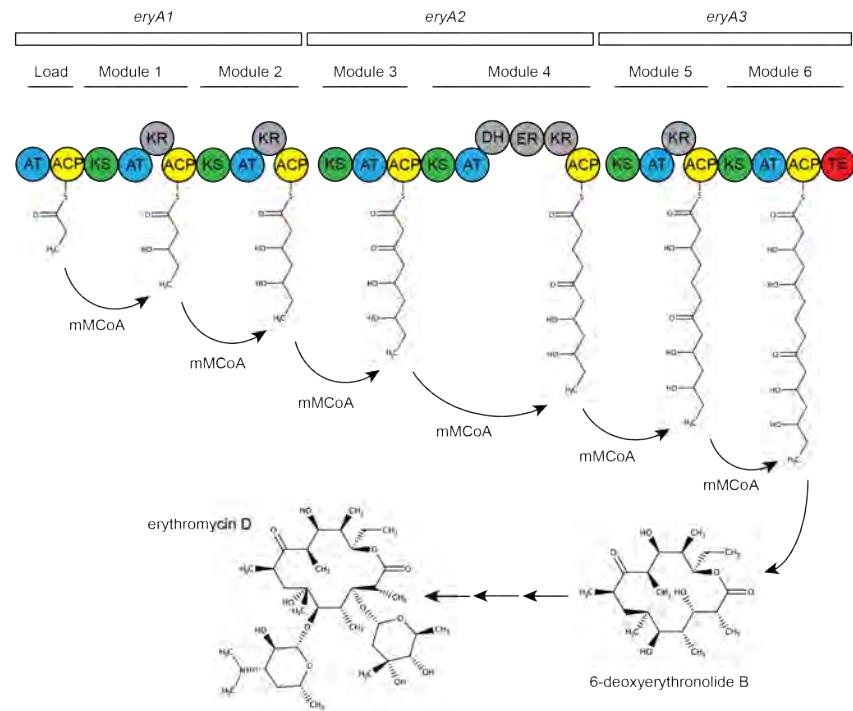
AT-ACP-[KS-AT-ACP]-TE

The initiating AT domain loads the first acyl group onto the first ACP domain, each subsequent module adds another acyl group, and the final TE domain releases the product. The resulting product will then almost always be modified further to generate the final metabolite, which can involve many additional enzymatic steps.

There are many types and variations of PKS, of which we only have space to cover a couple. In bacterial type I PKS, like the very well-studied one responsible for synthesis of the antibiotic erythromycin by *Saccharopolyspora erythraea* (Figure 17.4A), the PKS modules are arranged in one or a few enormous megasynthases and synthesis proceeds linearly from N-terminal to C-terminal ACPs. While erythromycin happens to be constructed by addition of 6 sequential methylmalonyl-CoA groups to a malonyl-ACP starter, the acyl groups added to growing type I PKS substrates can vary widely.

There are examples of PKS enzymes in which modules are skipped or act out of order, but these are relatively rare and the mechanisms by which this happens are not extremely well understood.

A. Type I PKS: erythromycin



B. Type II PKS: oxytetracycline

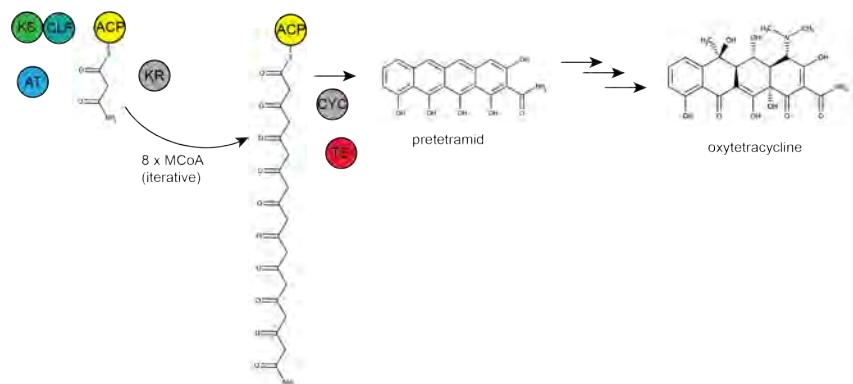


Figure 18.4. Polyketide synthase-dependent secondary metabolism. (A) Erythromycin biosynthesis by linear type I PKS, encoded by the eryA1, eryA2, and eryA3 genes. (B) Tetracycline biosynthesis by iterative type II PKS. Abbreviations: AT, acyltransferase; ACP, acyl carrier protein; KS, β -ketoacylsynthetase; KR, ketoreductase; DH, dehydratase; ER, enoylreductase; TE, thioesterase; CLF, chain length factor; CYC, cyclase; mMCoA, methylmalonyl-CoA; MCoA, malonyl-CoA.

Bacterial type II PKS, on the other hand, like fatty acid synthesis (Figure 17.13), are **iterative**, in that they re-use the same enzyme modules repeatedly to progressively add the **same** acyl-CoA unit to a growing polyketide chain. In bacterial type II PKS, the individual domains are **not** typically part of the same polypeptide, although they do form multi-enzyme complexes,

and there is an additional protein, related to KS domains, called *chain length factor* (CLF) that determines how many cycles will proceed before the TE protein releases the product.

Iterative type II PKS is illustrated in Figure 18.4B with the pathway for synthesis of the oxytetracycline precursor pretetramid by *Streptomyces rimosus*, which begins with a malonyl-mate group and proceeds with successive addition of 8 malonyl-CoA groups.

NON-RIBOSOMAL PEPTIDE SYNTHETASES

PKS use acyl-CoA subunits as the building blocks of secondary metabolites. NRPS use **amino acids** instead, generating peptides between 3 and 15 amino acids long. However, the modular logic of the two synthetic pathways is very similar. There are both linear (type A) NRPS megasynthases, analogous to type I PKS, and iterative (type B) NRPS pathways, analogous to type II PKS, and NRPS pathways are also constructed from a modular set of enzymatic domains, in this case:

- peptidyl carrier protein domains (abbreviated PCP)
- adenylation domains (A); specific for particular amino acids
- condensation domains (C)
- thioesterases (TE)

There are also, as for PKS, optional modification domains, like cyclases (CY) or epimerases (E), that occur in some pathways.

In a minimal NRPS, an A domain transfers an activated amino acid to the first PCP domain, in much the same way that tRNA synthetases attach amino acids to tRNA. Note that NRPS are not limited to the 22 amino acids found in ribosomally-synthesized proteins, and there are A domains specific for a very wide range of both D- and L-amino acids. The C domain catalyzes the formation of the peptide bond between each consecutive amino acid. There is no RNA template equivalent to an mRNA which encodes the order of amino acids in a NRPS-synthesized peptide.

A minimal NRPS with no additional modification domains has the following domain architecture:

A-PCP-[C-A-PCP]-TE

The synthesis of the *B. subtilis* surfactant surfactin is catalyzed by a linear type A NRPS (Figure 18.5), which also incorporates a fatty acid chain in the first module.

Type A NRPS: surfactin

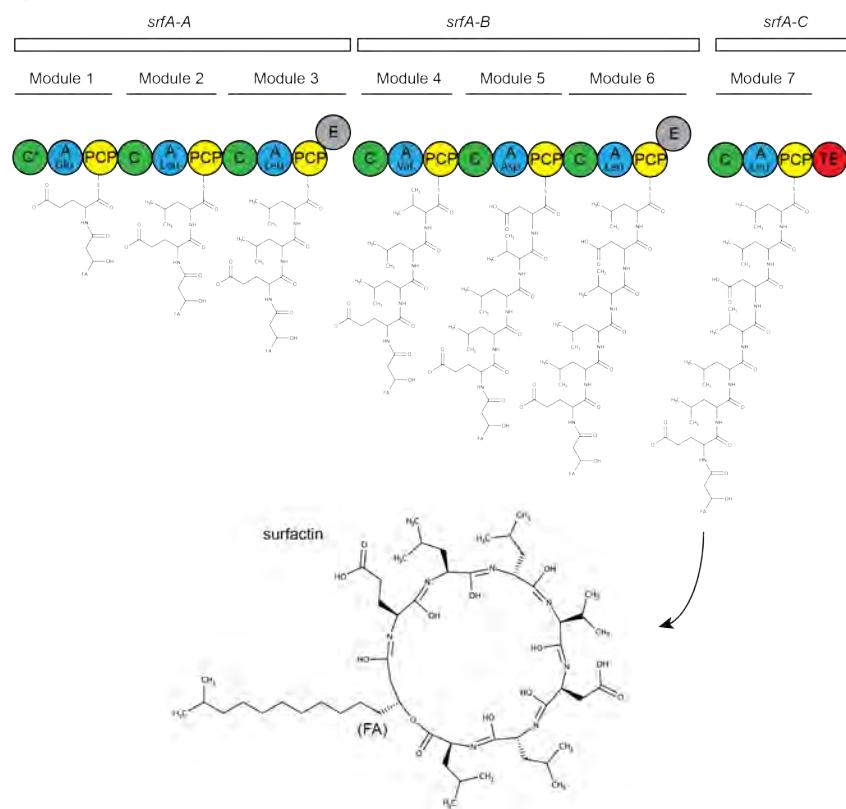


Figure 18.5. Surfactin biosynthesis by linear type A non-ribosomal peptide synthesis, encoded by the srfA-A, srfA-B, and srfA-C genes. The first C domain, indicated with an asterisk (C*) transfers a fatty acid to the glutamate added by the first adenylation domain.

The synthesis of the *E. coli* siderophore enterobactin is catalyzed by an iterative type B NRPS (Figure 18.6), combining three dihydroxylbenzoyl-seryl amino acid dimers into a cyclic hexameric peptide.

Type B NRPS: enterobactin

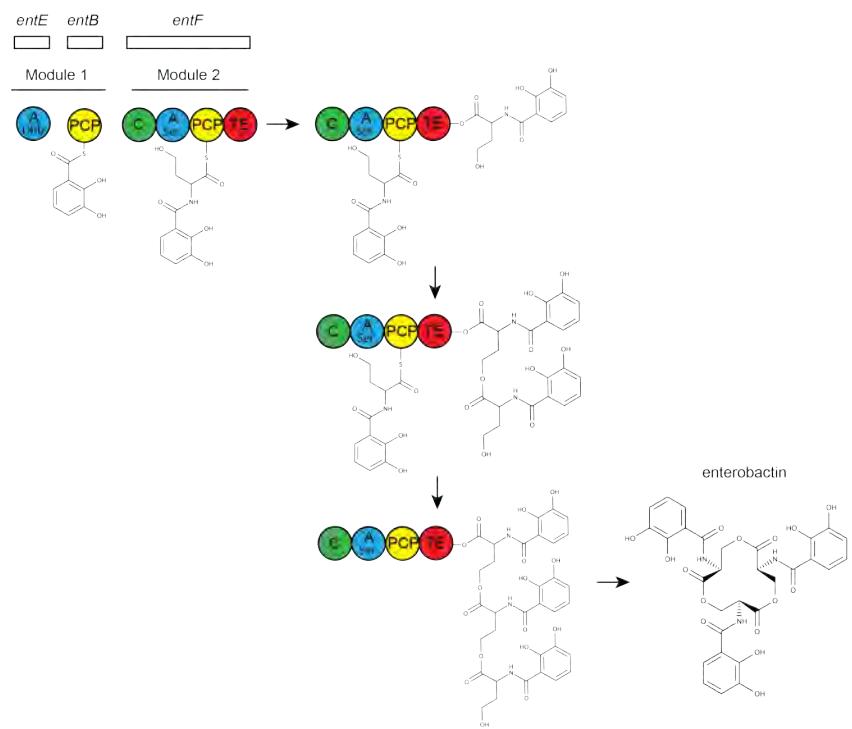


Figure 18.6. Enterobactin biosynthesis by iterative type B non-ribosomal peptide synthesis, encoded by the *entE*, *entB*, and *entF* genes. Abbreviations: C, condensation domain; PCP, peptidyl carrier protein; A, adenylation domain (amino acid-specific); TE, thioesterase; Dhb, 2,3-dihydrobenzoic acid; Ser, serine.

I will reiterate that not all PKS and NRPS systems are as "simple" as the ones illustrated above. There are type III PKS, which are iterative megasynthase-type enzymes and non-linear type C NRPS which skip or repeat individual modules, as well as variations and combinations of all of the different possibilities.

Indeed, because of the similarity between the PKS and NRPS biosynthetic strategies, it turns out to also be possible to **combine** PKS and NRPS modules in a single hybrid PKS/NRPS biosynthetic pathway. This is fairly common, and is how compounds like bleomycin, bacillaene, and myxothiazol are assembled from a combination of acyl and amino acid precursors. The PKS and NRPS modules may be on separate polypeptides or combined in hybrid megasynthases that incorporate both PKS and NRPS modules.

The highly modular nature of PKS and NRPS synthesis pathways is central to their ability to generate diverse natural products, and presumably encourages the rapid evolution of new secondary metabolites. Certainly, human bioengineers have put a lot of effort into rearranging and swapping out modules to engineer pathways that synthesize novel compounds, some of which have been extremely valuable.

It is important to note that not **all** peptide-derived secondary metabolites are synthesized by NRPS. Some are made by processing fragments of proteins transcribed as mRNA and translated by ribosomes. These secondary metabolites (including nisin, which we looked at in Discussion Problem Set 8, a wide range of small anti-bacterial peptides, and many quorum sensing signal peptides in Gram positive bacteria) are called *RiPPs* (ribosomally-synthesized and post-translationally modified peptides).

DISCUSSION PROBLEM SET #33: ACTIVATING SILENT BIOSYNTHETIC GENE CLUSTERS

One of the most widely-discussed problems in natural product discovery is that of cryptic and silent BGCs. The genomes of most actinomycetes and of many other bacteria encode BGCs whose products are unknown and which are not expressed under laboratory growth conditions. Many researchers consider these cryptic BGCs an untapped source of natural products that may have useful therapeutic properties.

Photobacterium luminescens is a Gram-negative bacterium that, in symbiosis with a nematode host, is a pathogen of insects. Its genome encodes at least 23 large BGCs, mostly of the NRPS type, but the products of only 5 of these have been identified (two siderophores, two antibacterial compounds, and one compound which repels ants and birds that might eat an infected insect). However, none of the BGCs are transcribed during growth of *P. luminescens* as a pure culture in laboratory media.

The following genetic tools are available for *P. luminescens*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (5.7 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
inducible promoters known (arabinose-inducible)	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
oligo-directed recombineering	✓

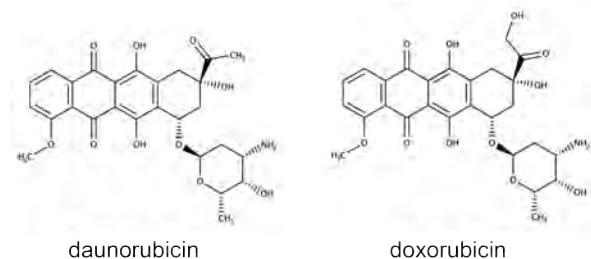
Design an experiment to determine whether the secondary metabolites produced by any of the silent BGCs of *P. luminescens* might be useful antibiotics. State:

- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment tests correlation or causation
- potential outcomes of your experiments, and how you will interpret them

P. luminescens is the source of the most commonly used bacterial luciferase reporter (the *luxCDABEG* operon). It is not especially obvious why this insect pathogen is so strongly bioluminescent, but occasionally human wounds can become infected with *Photorhabdus*, especially when the patient is also experiencing hypothermia (*Photorhabdus* does not survive well at normal human body temperature). After the Battle of Shiloh during the American Civil War, some soldiers' wounds were reported to light up at night with a faint blue "angel's glow", that scientists now hypothesize might have been due to infestation with *Photorhabdus*-harboring nematodes. The glowing wounds were reported to heal faster, which may have been due to the release of anti-microbial compounds from *Photorhabdus*. It's a nice story, but I don't know that I'd count on it as a treatment plan.

DISCUSSION PROBLEM SET #34: PHAGE-INHIBITING SECONDARY METABOLITES

Karen Maxwell's lab at the University of Toronto has discovered that some secondary metabolites produced by *Streptomyces* species function to protect the bacteria against infection by bacteriophage. Most of the metabolites they identified in their initial screens for this activity were members of a chemical family called the anthracyclines, including daunorubicin and doxorubicin, which are currently in clinical use as anti-cancer drugs.



Anthracyclines are DNA intercalating agents whose anti-cancer effects are thought to be due to inhibition of topoisomerase in rapidly-replicating cancer cells. The doses needed to prevent phage infection of bacteria are much lower than the doses of these drugs that are toxic to the bacteria themselves. Other *Streptomyces* secondary metabolites that affect DNA (like the DNA cleavage stimulator bleomycin) have no effect on phage replication.

Strikingly, daunorubicin and doxorubicin not only prevent *Streptomyces*-specific phages from killing *Streptomyces* strains, but also prevent the unrelated λ , T5, T6, and T7 phages from propagating in *E. coli*, suggesting that they function by interfering with a mechanism common to all phage.

The following genetic tools are available for *S. peucetius*, the species from which both daunorubicin and doxorubicin were originally isolated:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (8 Mbp)	✓
susceptible to mutagens	✓

Many more tools are available for the model organism *S. coelicolor*, which does not produce either daunorubicin or doxorubicin (it produces a variety of other secondary metabolites, including the antibiotics actinorhodin and methylenomycin, the anti-malarial undecylprodigiosin, and the anti-fungal perimycin):

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence (8.7 Mbp)	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

And, of course, you can do anything you want, genetically speaking, in *E. coli* and its phages. Both daunorubicin and doxorubicin are commercially available in pure form.

Design an experiment or experiments to determine the mechanism by which anthracyclines protect bacteria against phage infections. State:

- a model to explain how daunorubicin and doxorubicin might inhibit phage propagation
- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- whether your proposed experiment tests correlation or causation
- potential outcomes of your experiments, and how you will interpret them

LECTURE 19: CRITICAL READING (BACTERIAL METABOLISM)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
 - What is the central **question** of this paper?
 - Is the experimental design clear and appropriate to address that question?
 - Do you understand the methods used?
 - Are the data clearly presented, with appropriate statistics?
 - Do you agree with the conclusions the authors came to based on their data?
 - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Shukla et al. (2023) "An antibiotic from an uncultured bacterium binds to an immutable target". Cell 186:4059-4073.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

Note that the supplemental information of this particular paper contains a lot of NMR data that was used to determine the structures of the compound they discovered. This well beyond the scope of this class, so don't worry about being asked to present those figures.

SUMMARY OF EXPERIMENTAL DESIGN PRINCIPLES

This page is simply a compilation of the rules for experimental design discussed in the Scientific Process sections of the previous chapters. **This is the single most important thing I want you to take away from this class.** Proposing good experiments to test valid hypotheses is the key to good science, the central element of grant writing, and is something you will have to do for your qualifying exam (and the rest of your career), so it's an important skill to cultivate.

OBSERVATIONS

When describing a set of observations that you plan to make, you should explain:

- What will you be measuring, and how will you measure it?
- Is it a qualitative or a quantitative measurement?
- When and how often will you measure it?

MODELS

When proposing a model, it should:

- incorporate all of the available data
- propose a mechanism that explains the behavior of the system
- make testable predictions about the system being studied

HYPOTHESES

When proposing a hypothesis:

- it should test a specific aspect of a model
- it should be falsifiable
- you should be able to propose a set of observations that can be used to test that hypothesis

EXPERIMENTS

When designing an experiment, you should:

- define the dependent and independent variables
- explain what you will measure and how (i.e., what observations will you make?)
- describe both positive and negative controls
- describe the possible outcomes of the experiment and what they would mean for your hypothesis
- state whether the experiment will determine correlation or causation

ALTERNATIVE APPROACHES

There is more than one way to answer any scientific question. You should be able to:

- design more than one distinct experiment to test a given hypothesis
- be able to explain the strengths and weaknesses of each approach

GRADING RUBRIC FOR WEEKLY QUIZ PROBLEMS

Each week you will be responsible for solving an experimental design problem on your own, and this page describes how those quizzes will be graded.

Each week's problem is worth **20 points** in total, and will involve examining a data set, proposing a model to explain those data, and then designing an experiment to test a hypothesis based on that model. Each week you also have the possibility of earning **1 bonus point** for a particularly clever, creative, or elegant solution to the experimental problem.

	4 points	3 points	2 points	1 point	0 points
Model	Incorporates all of the data available and proposes a mechanism.	Lose 1 point each for: <ul style="list-style-type: none">• does not incorporate all available data• does not propose a mechanistic explanation of the data• is biologically implausible		No model proposed.	
Hypothesis	Falsifiable hypothesis that tests a specific and important aspect of the model.	Lose 1 point each for: <ul style="list-style-type: none">• hypothesis is not falsifiable• hypothesis does not test a specific aspect of the model• hypothesis does not test an important part of the model		No hypothesis proposed.	
Experimental Design	Will effectively test the hypothesis, is practical, and has clearly-described observations.	Lose 1 point each for: <ul style="list-style-type: none">• will not effectively test the stated hypothesis• unclear description of the observations or measurements necessary• is not practical with standard laboratory techniques		No experiment described.	
Controls	All necessary positive and negative controls clearly described.	Lose 1 point each for: <ul style="list-style-type: none">• no positive controls• no negative controls• missing controls essential to the interpretation of the proposed experiment		No controls described.	
Interpretation of Results	All possible results described, with explanations of the meaning of those results for the model. State clearly whether expected results establish correlation or causation.	Lose 1 point each for: <ul style="list-style-type: none">• only some possible results described• unclear description of interpretation• not stating clearly or correctly whether results establish correlation or causation		No possible results described.	

Note: Week 2's quiz will be graded slightly differently, since it will be a genetic engineering problem rather than an experimental design problem. The rubric for that quiz will be included in the quiz itself and will be somewhat simpler.

GLOSSARY

2-dimensional gel electrophoresis	a largely obsolete method for direct quantification of proteins that works by separating proteins by both size and isoelectric point
abduction	logical reasoning in which you draw the best possible conclusion from incomplete observations
abstract	a short summary of a paper
acetylation	modification of a molecule by addition of an acetyl group
acidophilic bacteria	bacteria that live at very low pH (< 3)
acyl carrier protein	a protein with a CoA cofactor which is used as the platform for synthesis of fatty acids or polyketides
adherence	the process by which bacteria stick to surfaces
adhesin	a protein or polysaccharide produced by a bacteria specifically for the purpose of attaching to a surface
aerobic respiration	a respiratory pathway in which the terminal electron acceptor is oxygen
erotaxis	movement towards air; or oxygen
alignment	a visual representation of homology between DNA, RNA, or protein sequences
alkaliphilic bacteria	bacteria that live at very high pH (> 10)
allele	a version of a gene, typically differing from other alleles by only a small number of mutations
allele number	a notation used to distinguish between different mutations of the same gene
allosteric effector	a molecule that controls the activity of a protein by interacting with it at a site distant from its active site
allostery	a regulatory mechanism by which a molecule controls protein activity by non-covalently binding to a site that is not the active site of that protein
alternative electron acceptors	catch-all term for terminal electron acceptors in anaerobic respiratory pathways
alternative sigma factors	sigma factors responsible for recognizing promoters other than those recognized by the housekeeping sigma factor; often involved in stress response or development
amidase	an enzyme that breaks an amide bond
amino sugar	a monosaccharide modified by attachment of an amino group
amphilophotrichous	multiple flagella at both poles
amphipathic	a molecule that has both hydrophobic and hydrophilic functional groups
amphitrichous	a single flagellum at each pole
anabolism	the biosynthesis of molecules from precursor intermediates
anaerobic respiration	any respiratory pathway in which the terminal electron acceptor is not oxygen
anomeric carbon	the carbon of a monosaccharide adjacent to the oxygen atom and whose orientation varies in α and β versions of a sugar
anoxygenic photosynthesis	photosynthetic pathways which use a molecule other than H_2O as an electron donor

anti-Shine-Dalgarno sequence	the sequence of the 3' end of the 16S ribosomal RNA, which binds to the ribosome binding site in mRNA
anti-terminator	a regulator that prevents transcription termination
antibiotic resistance	acquisition of mutations that allow bacteria to detoxify or export antibiotics
antibiotic resistance cassette	a gene encoding a product that confers antibiotic resistance, along with all of the additional sequences needed to ensure its expression
antigenic variation	the ability of some bacteria to alternate between expression of different surface components, as a method of avoiding the immune system
antiporter	a transporter that transports two molecules across a membrane in opposite directions
archaella	the structure in archaea analogous to bacterial flagella, homologous to type 4 pili
asymmetric cell division	a developmental process in which the two daughter cells are genetically identical but have different phenotypes
attachment site	(also "att site") the specific DNA sequence at which lysogenic bacteriophage (and some transposons) insert themselves into their host chromosomes
attractant	a chemical or other stimulus that causes bacteria to move towards higher concentrations of that stimulus
autoaggregation	(also "flocculation") the ability of bacteria to stick to other cells of the same species and strain
autoinducer	the secreted signal molecule secreted by cells in quorum sensing regulatory systems
autoinducer 2	(also "AI-2") a specific boron-containing quorum sensing molecule produced by many Gram-negative and some Gram-positive bacteria
autolysin	an enzyme that disrupts the cell envelope of the organism whose genome encodes it
autotransporters	another name for type V secretion systems; proteins that catalyze their own transport across the outer membrane of Gram-negative bacteria
autotroph	an organism that can fix CO ₂ into organic carbon
auxotroph	a mutant that requires a particular nutrient
bacterial artificial chromosome	(also "BAC") a plasmid based on the F factor that can be used to clone very large inserts
bacterioneuston	(also "surface micro-layer") the top millimeter of the ocean, inhabited by a relatively dense population of bacteria
bacteriophage	a virus that infects bacteria
bacteriorhodopsin	a transmembrane protein that absorbs light and uses the energy of that light to pump protons across the membrane
bactofilin	a family of filamentous cytoskeletal proteins found only in bacteria
binning (of data)	grouping quantitative data into categories for ease of analysis
bioactive compounds	secondary metabolites that have a known effect of some kind on cells
biochemistry	the study of the physical properties of biological molecules
biofilm	a multi-cellular bacterial community attached to a surface

biofilm dispersal	the regulated process by which planktonic cells are released from biofilms
biosynthetic gene cluster	(also "BGC") genetic loci dedicated to the production of a specific secondary metabolite
blunt end	a double strand break with no sticky ends, produced by some restriction enzymes
bradytroph	a mutant that grows slowly without a particular nutrient
brute-force approach	an inelegant, labor-intensive experimental design
capsule	the polysaccharide coating of many bacteria, generally attached to the cell surface and often important for recognition by the adaptive immune system
carbohydrate	organic compound composed only of carbon, hydrogen, and oxygen
carbon catabolite repression	(also "CCR" or "catabolite repression") a regulatory system that represses expression of genes for catabolism of non-preferred carbon sources in the presence of preferred carbon sources
catabolism	the breakdown of molecules, yielding their component parts and energy
causation	proof that one phenomenon directly leads to another
cDNA library	a pool of plasmids containing many different cloned DNA inserts derived by reverse transcription from an organism's mRNA
cell biology	the study of the molecular mechanisms that determine the shape, size, division, differentiation, and development of (bacterial) cells
cell envelope	the boundary separating the inside of a cell from the outside of a cell, consisting of membrane(s), cell wall structures, S-layers, etc.
cell wall	a rigid polymeric shell outside of a cell's cytoplasmic membrane that maintains the cell's shape and integrity
chain length factor	a component of iterative PKS pathways that determines how many long the final product will be
chaperone	a protein that binds to other proteins and modulates their folding or solubility
chaperone-usher secretion system	a variation on type V protein secretion systems that involves three proteins
chemical mutagen	a chemical that damages DNA, resulting in mutations
chemically competent cells	bacteria treated (often by rinsing in cold CaCl ₂ followed by brief heat shock) to make them capable of taking up DNA directly from their environment (transformation)
chemiosmotic theory	a biophysical theory that describes how potential energy is stored in biological systems as an electrochemical gradient of ions across a lipid bilayer membrane
chemoautotroph	a non-photosynthetic organism that can fix CO ₂ into organic carbon
chemotaxis	the ability of an organism to move along a chemical gradient
chimeric protein	see "protein fusion"
ChIP	(also "chromatin immunoprecipitation") a technique that uses antibodies to a protein to purify both that protein and any protein or DNA molecules that are bound to it
ChIP-seq	(also "chromatin immunoprecipitation sequencing") an <i>in vivo</i> technique to identify all of the genomic binding sites of a DNA binding protein using next generation sequencing

chromosome	a large DNA molecule containing essential gene(s) and usually present in single copy
cistron	an obsolete synonym for gene
cloning	incorporating a gene into a plasmid for expression
cloning strain	a strain of <i>E. coli</i> (usually) that contains mutations to improve competence and plasmid stability, making molecular cloning easier
cloning vectors	plasmids used to express genes in bacteria
codon optimization	changing the sequence of a gene so that it uses only the most abundant codon for each amino acid; species specific
codon usage	a measure of how well particular codons are translated in a given organism or how frequently they occur in a given genome
cofactor A	(also "CoA") a thiol-containing cofactor involved in many metabolic reactions
compatibility group	see "origin of replication"
competence pili	a subset of type IV secretion system pili that are involved in DNA uptake
competent cells	bacteria capable of taking up DNA directly from their environment (transformation)
competitive index	a kind of experiment that assesses the role of genes in complex phenotypes by competing mutants against wild-type under the conditions of interest
complementation analysis	an experimental design that establishes genetic causation by removing and replacing individual genes
conditionally essential	a gene that is required for growth under some conditions, but which is dispensable under others
conditional phenotype	a phenotype that is only observed under specific growth conditions
confocal microscopy	microscopy technique that constructs a 3D image from sequential images taken in different focal planes
confirmation bias	the problematic tendency to interpret results as supporting your preconceived notions, whether or not they actually do
conjugation	DNA transfer between cells via pili; requires <i>tra</i> factors, an origin of transfer, and physical contact between cells
consensus sequence	the most common or average sequence for a particular gene or locus
conserved residues or nucleotides	(also "conservation") protein, RNA, or DNA sequence features that do not change (or change slowly) over evolutionary time
constitutive promoter	a promoter that is always active and expresses genes under its control at a constant level
constitutively active	always expressed or functioning at a constant level
constitutively inactive	never expressed or never functional
contact-dependent inhibition	another term for type VI protein secretion
contractile phage	bacterial viruses whose tails contract upon contact with a host cell, driving DNA injection
control	a treatment included in an experiment to make sure that the experiment is working as intended

copy number	how many of a DNA molecule (typically a plasmid) are present per cell
correlation	the observation that two or more phenomena appear or change together
corresponding author	the person who gets contacted about a paper if there are any questions, typically the head of the lab where the work was done
cortex	the thick peptidoglycan cell wall layer of an endospore
cos site	a site that allows a plasmid to be packaged in λ phage particles
cosmid	a plasmid with a cos site
cotranscribed	genes adjacent to each other on the chromosome, and transcribed in the same direction
cotransduction frequency	how often two genes or mutations will be transferred simultaneously by transduction, a function of transducing phage packaging size and the distance between the genes or mutations on the chromosome
counter-selectable marker	a gene encoding a product which allows you to select for cells that don't contain that gene; a conditionally lethal gene
coupling sites	points in a respiratory electron transport chain at which a PMF is generated
crescentin	(or "CreS") a cytoskeletal protein responsible for the curved shape of <i>Caulobacter crescentus</i> cells, homologous to eukaryotic intermediate filaments
CRISPR	(also "clustered regularly interspaced short palindromic repeats") a system that uses short guide RNAs to direct the activity of a nuclease (usually Cas9) to specific sites in a DNA (or RNA) molecule
CRISPR array	the series of repetitive DNA sequences that incorporate guide RNAs in natural CRISPR systems
CRISPR-associated proteins	the various proteins that are part of natural CRISPR systems; Cas9 nuclease is the most important for biotechnological purposes
cryptic metabolites	secondary metabolites with no known function
curls	proteinaceous fibers attached to the outer surface of Gram-negative bacteria, a kind of functional amyloid
cytokinesis	the separation of the cytoplasm of two daughter cells during cell division
cytoplasm	the aqueous interior of a living cell
cytoskeleton	filamentous proteins involved in cell shape, chromosome segregation, etc.
data	high-quality, carefully recorded observations
daughter cell	the cells derived from a bacterial cell division event
deduction	logical reasoning which starts from known principles (true statements) and identifies the logical conclusions of those statements
defective prophage	see "stable lysogen"
degeneracy	the fact that multiple codons can encode the same amino acid
degron	protein sequences recognized by proteases as signals for protein degradation
deletion	the removal of DNA sequence from a gene
dependent variable	the variable(s) measured by the experimenter during an experiment
derepression	the effect of inactivating a negative regulator

development	a process by which genetically-identical cells express different phenotypes
diderm	a bacterial cell with two lipid membranes (an outer and an inner), typically Gram-negative
differentiation	a developmental process in which a proportion of cells in a population change their phenotypes
divergently transcribed	genes adjacent to each other on the chromosome, but transcribed in opposite directions
divisome	the protein complex responsible for the process of cell division
DNA ligase	an enzyme that joins two DNA molecules together
DNA methylase	(see "restriction methylase") an enzyme that methylates specific sequences in DNA
DNA microarray	a direct method to detect RNA by hybridizing it with an array of oligo probes of known sequence; largely obsolete
DNA recombination	see "homologous recombination"
domain	a structurally-conserved element of a protein, usually with a specific function
double-crossover recombination	a recombination event that requires two independent homologous recombinations, such as integrating a linear DNA fragment into a circular chromosome
downstream gene	a gene encoded 3' of the gene being discussed on an mRNA
duplication	a mutation that results in multiple copies of a DNA sequence
effector proteins	secreted proteins that have effects on other cells, often host cells or competing bacteria
electron acceptors	molecules that become reduced during respiration
electron donors	molecules that become oxidized during respiration
electron shuttles	(also "electron carriers") small molecules that can be oxidized and reduced to carry electrons between proteins
electron transport chain	a pathway that links oxidation reduction reactions in a bilayer membrane to generation of a PMF
electronic table of contents	a service that emails you the list of papers published in a journal when each issue becomes available
electrophoretic mobility shift assay	(also "EMSA" or "gel shift assay") a direct measurement of the binding affinity of a protein for a nucleic acid molecule, using gel electrophoresis to separate bound and unbound nucleic acids by size
electroporation	a method for transformation in which cells are mixed with DNA and subjected to an electric shock
ELISA	(also "enzyme-linked immunosorbent assay") an assay that uses immobilized antibodies to detect and quantify antigenic substrates
elongasome	the protein complex responsible for organizing synthesis of new cell wall material as the cell grows, specifically in rod-shaped bacteria
endonuclease	a nuclease that cleaves within a DNA or RNA molecule
endonuclease cleavage site	a DNA or RNA sequence that is recognized by an endonuclease
endopeptidases	proteases that break peptide bonds within proteins

endospore	a spore formed inside of a mother cell, typical of Gram positive <i>Bacillus</i> and <i>Clostridium</i> species
endotoxin	see "lipopolysaccharide"
enrichment	a procedure that increases the proportion of mutants of interest in a population
Entner-Douderoff pathway	catabolic pathway that breaks down glucose into pyruvate
enzyme activity assay	a direct biochemical measurement of protein activity, specifically for proteins that catalyze chemical reactions
epigenetic	modifications of DNA or other cellular components that result in a (usually heritable) change in phenotype without a change in the DNA sequence
episome	see "plasmid"; obsolete
epitope tag	a short peptide sequence that can be fused with proteins of interest to allow their detection or purification with commercially available antibodies
essential gene	a gene that cannot be knocked out; encodes a function the cell depends on
e-value	a measure of the statistical significance of a BLAST search; a smaller e-value indicates a more significant degree of similarity between two sequences
exonuclease	a nuclease that degrades a DNA or RNA molecule from one end
exotoxin	a toxic protein or compound secreted by a bacterial cell
experiment	a test of the effects of a specific manipulation on a system
extracellular DNA	(also "eDNA") DNA released from bacteria for functional purposes, for example as a component of a biofilm matrix
extracellular polysaccharides	(also "EPS") polymers of sugars produced by bacteria and secreted out of the cell
fI origin	a site that allows a plasmid to be packaged as concatenated single-stranded DNA when the host bacterium is infected with bacteriophage fI
false negative result	an erroneous result that looks like nothing happened when something did
false positive result	an erroneous result that looks like something happened when it did not
falsifiable	a property of a useful hypothesis – can it be proved wrong?
fatty acid	a carboxylic acid with a long, hydrophobic hydrocarbon chain; a linear lipid
fermentation	a metabolic pathway in which the byproducts of initial catabolic pathways are reduced using the reducing equivalents generated during those pathways
fimbrial adhesin	fibrous adhesins that extend well away from the cell, usually called pili
first author	typically the person who did most of the experiments on a paper; may have multiple "first authors" who contributed equally to the work
flagella	helical rotary filaments used by many bacteria for motility
flagellar basal body	a specialized type III secretion system dedicated to the export and assembly of flagella
flagellin	the protein monomer that makes up the bulk of the flagellar filament
flippase	an enzyme that moves a lipid from one side of a bilayer membrane to the other side
floc	a clump of bacteria stuck to each other in suspension

flow cytometry	see "fluorescence-activated cell sorting"
fluidity	a property of lipids that depends on their melting temperature
fluorescence-activated cell sorting	(also "FACS") a technology that is able to rapidly separate individual cells on the basis of their fluorescence
fluorophore	a fluorescent molecule (either a small molecule or a protein)
forespore	the specialized cell that will become an endospore during sporulation
forward genetics	identifying mutations that lead to a specific phenotype of interest
frameshift mutation	insertion or deletion of 1 or 2 nucleotides (or any number not divisible by 3)
fruiting body	a multicellular structure formed during development (e.g. in myxobacteria) which contains spores or other resting cells
FtsZ	bacterial tubulin homolog essential for cell division in most bacteria
functional amyloid proteins	proteins that aggregate into characteristically extremely stable, sticky β -sheet rich structures, often involved in adherence and biofilm formation
functional redundancy	two genes products that carry out the same or overlapping functions
functional RNA	RNA that is not mRNA; includes ribosomal RNA, transfer RNA, small regulatory RNAs, and ribozymes
fusion protein / tag	see "protein fusion"
gain-of-function mutation	a mutation that gives a gene product new or enhanced abilities
gas vesicles	bacterial organelles that contain gas, involved in floating
gene	a DNA sequence encoding a functional product
gene knockdown	artificially reducing the expression of a gene without constructing a null mutation; useful for studying essential genes, for example
gene knockout	see "null mutation"
gene product	an RNA or protein encoded by a gene
generalized transducing phage	phage which are able to package random fragments of DNA from the chromosome of their host cell into virus particles
general stress response	a regulatory program induced by many different kinds of harmful changes in an organism's environment
genes of unknown function	genes with no currently known role in the cell
genetic drift	accumulation of non-selected mutations in the genomes of bacterial strains over time
genetic toolkit	ways to put new DNA into an organism or to change the DNA that it already has
genetics	the science of how heritable characteristics are passed from one organism to another
genome	the complete DNA sequence of a cell
genomic library	a pool of plasmids containing many different cloned inserts derived from an organism's genomic DNA
genotype	the sequence of the genome of an organism
germinate	the process by which a spore turns into a growing vegetative cell

gliding motility	any of several mechanisms by which bacteria move along surfaces without obvious external appendages
global regulator	a regulator that controls many genes or gene products from around the genome
glucokinase	an enzyme that phosphorylates glucose to glucose-6-phosphate
gluconeogenesis	anabolic pathway for the synthesis of glucose; mostly the reverse of glycolysis
glycolysis	catabolic pathway that breaks down glucose into pyruvate
glycosidic bonds	covalent bonds between monosaccharides in a polysaccharide
glycosyltransferase	an enzyme that adds a monosaccharide to one end of a polysaccharide chain
glyoxylate bypass	(also "glyoxylate shunt") catabolic variation of the TCA cycle that bypasses several steps to allow catabolism of 2-carbon compounds
Gram stain	a method for differential staining of bacteria
Gram-negative	a cell that stains pink in the Gram stain, often a diderm bacterium
Gram-positive	a cell that stains purple in the Gram stain, often a monoderm bacterium
guide RNA	a short sequence that serves to direct Cas9 nuclease to a specific target site
gum	a secreted polysaccharide that is sticky
hairpin	a DNA or RNA structure that is folded into a small, stable loop
heterocyst	a terminally-differentiated cell specialized in nitrogen fixation formed by filamentous cyanobacteria
heterolactic fermentation	a fermentative pathway in which pyruvate is reduced to lactate, ethanol, and CO ₂
hexose	a monosaccharide containing 6 carbons
high-energy phosphate bonds	the phosphate bonds in ATP and other NTPs, notable for their high phosphoryl group transfer potential
histidine kinase	an enzyme that phosphorylates a histidine, but more specifically usually refers to the sensor component of a two-component regulatory system
holdfast	the sticky anchor at the end of the stalk in <i>Caulobacter crescentus</i>
Holliday junction	the crossover point between two homologous DNA sequences that is the essential intermediate in homologous recombination
homolactic fermentation	a fermentative pathway in which pyruvate is reduced to lactate
homologous recombination	a DNA repair mechanism that allows the exchange of sequences from one DNA molecule to another; requires sequence homology
homologs	(also "homologous genes" or "homologous proteins") genes with a common evolutionary ancestor, inferred from sequence homology
homology	a measure of how similar two DNA, RNA, or protein sequences are
horizontal gene transfer	the acquisition of genetic material from a phylogenetically distant organism
host range	the list of different species a particular plasmid can replicate in
housekeeping sigma factor	the most abundant sigma factor in the cell, and the one responsible for recognizing most promoters
hypothesis	a prediction made by a scientific model, a possible answer to a scientific question

immunity proteins	proteins encoded by bacteria which prevent those strains from killing themselves, for example with type VI secretion systems or bacteriocins
immunoblot	see “western blot”
impact factor	the number of citations of papers in a journal over the previous 2 years, divided by the number of papers published in that journal in that time
in-frame	denotes DNA sequences whose codons are lined up with each other so that a continuous protein is produced from them during translation
incompatible plasmids	plasmids with the same origin of replication and / or the same selectable marker
independent variable	the variable(s) changed by the experimenter during an experiment
inducer	a compound that can be added to cells to control the activity of an inducible promoter
inducer exclusion	a mechanism of catabolite repression in which the import of alternative carbon sources is inhibited in the presence of preferred carbon sources
inducible promoter	a promoter that can be turned on or off by the addition of inducers; this term is usually used in reference to promoters in plasmids
induction	logical reasoning process which starts with observations and generalizes them to make predictions about broader events
initiating nucleotide	the first nucleotide of a transcribed RNA
inner leaflet	of a lipid bilayer; the phospholipids in the half of the membrane facing the cytoplasm (or periplasm of an outer membrane)
insertion	the addition of extra DNA sequence into the chromosome
integral membrane protein	a protein that is embedded in a lipid bilayer and crosses it at least once
interactome	the complete set of protein-protein interactions in an organism
intragenic suppressor	a second mutation in a mutated gene that reverses the phenotype of the mutant
intrinsic terminator	a stable, GC-rich stem-loop RNA structure, followed by several uracil residues, that leads to transcription termination
ionophore	chemical that bind to cations and allows them to diffuse through membranes, can be used to collapse ion gradients
isogenic strains	strains that are identical except for the specified mutations
isozymes	non-homologous enzymes in the same organism that catalyze the same reaction
journal club	a group meeting in which papers from the (usually) recent scientific literature are discussed in detail
kilobase pair	1,000 base pairs
kinase	an enzyme that adds phosphate groups to a substrate
knockout collection	a complete set of null mutants in a particular strain, each lacking one non-essential gene
leader peptide	a short protein encoded at the beginning of an operon, often as part of a transcriptional attenuation regulatory mechanism
lethal mutation	a mutation that kills the cell
linkage mapping	an obsolete method of determining the location of mutations by how often different genes are cotransduced by generalized transducing phage

linked marker	a selectable marker located in the genome close to a mutation of interest
Lipid A	the innermost lipid-disaccharide component of lipopolysaccharide
lipid bilayer membrane	a biological membrane made up of amphipathic lipids that assemble into sheets with their hydrophobic regions inside
lipopolysaccharide	(or "LPS") the complex sugar-lipid molecule that makes up most of the outer leaflet of the outer membrane in diderm bacteria
lipoprotein	a protein with a covalently-attached lipid group (a post-translational modification)
lipoteichoic acid	(or "LTA") polymer of repeating sugars and phosphate groups anchored in the cell membrane of monoderm bacteria by lipid groups
local regulator	a regulator that controls only a small number of genes or loci, often including the regulator itself
localized mutagenesis	random mutagenesis of a single gene or locus, as opposed to the entire genome
locus	a location on a chromosome; could be a gene, an operon, a regulatory site, etc.
locus tag	a unique identifier for a gene, used in genome sequencing projects
lophotrichous	multiple flagella at one pole
lysogeny	a bacterial cell containing a prophage
lysogenic phage	a bacteriophage able to integrate itself into the chromosome of a host cell
lytic transglycosylase	an enzyme that breaks the bonds between MurNAc and GlcNAc in peptidoglycan
magnetotactic bacteria	bacteria that are able to orient themselves along magnetic field lines
magnetotaxis	movement along magnetic field lines
mass spectrometry	a powerful technique for determining the molecular weight of molecules
material transfer agreement	(also "MTA") paperwork necessary to transfer research materials from one university to another
matrix	the extracellular components (polysaccharides, eDNA, and proteins) of a biofilm
megabase pair	1,000,000 base pairs
megasynthase	a very large modular enzyme for synthesis of secondary metabolites by PKS, NRPS, or hybrid PKS/NRPS pathways
merodiploid	a strain that contains two copies of a gene (often one on the chromosome and one on a plasmid, but potentially both in the chromosome), usually two different alleles
metabolic flux	a measurement of how active a particular enzyme or pathway is within a cell
metabolite	a small molecule produced by a cell or used as an intermediate in a cellular pathway
metabolome	the set of all small molecules (metabolites) in a cell
metabolomics	methods for measuring large numbers of metabolites in a cell simultaneously
metagenome	the DNA sequences of a community of organisms
metagenomic library	a pool of plasmids containing many different cloned inserts derived from DNA purified from a mixture of different organisms
metatranscriptome	mRNA sequences derived from a community of organisms

methylation	covalent addition of a methyl group to a protein or DNA molecule
Michaelis constant	(also "K _m ") the concentration of substrate at which an enzyme's reaction rate V is half of V _{max}
microcolony	a small group of cells clumped together on a surface, can be a precursor of biofilm formation
microfluidics	a technology that examines bacteria in very small volumes and with precisely-controlled flow conditions
mineralization	the deposition of carbonate minerals in a biofilm, solidifying the matrix into a rock-like state
minimal media	growth media that contains only the compounds a particular species needs to grow
minireview	a short review, either giving a brief introduction or reporting recent progress in a field
missense mutation	a mutation of an amino acid encoding codon to a different amino acid encoding codon
mixed acid fermentation	a fermentative pathway in which pyruvate is reduced to lactate, formate, acetate, ethanol, CO ₂ , and H ₂
model	a mechanistic explanation of a system, based on data from observations and experiments
model organism	an easily-studied species, the properties of which are used to infer the properties of less easily-studied (or just less studied) organisms
molecular biology	(see "molecular genetics")
molecular genetics	genetics with an understanding of the biochemical nature of genes
monocistronic	an mRNA encoding one gene
monoderm	a bacterial cell with one lipid membrane, typically Gram-positive
monophyletic	a group of organisms descended from a single common ancestor
monosaccharide	(also "sugar") 3- to 7-carbon carbohydrates distinguished by the arrangement and chirality of their various hydroxyl groups
monotrichous	a single flagellum at one pole
mother cell	the cell that nurtures the formation of an endospore during sporulation
motor proteins	proteins that convert proton motive force into rotary motion of the flagellar basal body
motor switch complex	the components of the flagellar basal body which are involved in changing the direction of rotation of the motor
MreB	bacterial actin homolog
mRNA stability	how long a particular mRNA remains in the cell before being degraded
multicopy suppressor	a gene that reverses the phenotype of a mutation in a different gene when over-expressed
multiple alignment	an alignment of more than two sequences
multiple cloning site	(also "MCS") a small region of a plasmid with several closely spaced restriction sites

murein	see "peptidoglycan", somewhat outdated term
mutagen	a treatment that damages DNA, resulting in mutations
mutagenesis	the act of making mutations in an organism
mutant	an organism containing a mutation
mutant hunt	an experiment intended to identify mutations that affect a particular phenotype
mutation	a change in the DNA sequence of an organism
mutation rate	how quickly mutations accumulate in a population
mutator strain	a bacterial strain defective in DNA repair; useful for random mutagenesis of plasmids
mycolic acid	long-chain, extremely hydrophobic lipid unique to the mycobacteria
myxospores	stress-tolerant resting cells produced during development of myxobacteria
N-acetylation	modification of a molecule by addition of an acetyl group to a nitrogen atom
N-acylation	covalent addition of acyl groups to lysine residues in proteins
natural products	any molecule produced by a living organism, usually means secondary metabolites
naturally competent cells	bacteria capable of taking up DNA directly from their environment (transformation) without special treatment
negative control	a control that tests for the possibility of false positive results in an experiment
negative regulator	a regulator that represses the system being studied
N-end rule	the impact of the N-terminal amino acid(s) of a protein on that protein's proteolytic stability
next generation sequencing	(also "NGS") any of a variety of methods of DNA sequencing that read the sequence very large numbers of (typically) very short DNA fragments
non-fimbrial adhesin	adhesins that are not pili
non-ribosomal peptide synthetase	(also "NRPS") a biosynthetic pathway that construct complex molecules out of simple amino acid components
nonsense mutation	a mutation of an amino acid encoding codon to a stop codon
northern blot	a direct method to detect RNA by probing with radioactively labeled oligos; obsolete
nuclease	an enzyme that degrades DNA or RNA by breaking the bonds between nucleotides
nucleoid occlusion	the prevention of Z-ring formation in some bacteria in regions of the cell containing large amounts of DNA
null mutation	a mutation that inactivates a gene product
O-acetylation	modification of a molecule by addition of an acetyl group to an oxygen atom
O-antigen	the outermost polysaccharide chain component of lipopolysaccharide
observation	a measurement of some feature of the objective universe
oligonucleotide	(also "oligo") a short, artificially synthesized DNA molecule
open reading frame	the protein-coding sequence of a gene

operator sequence	the DNA sequence to which a regulator binds
operon	several genes encoded on the same mRNA
origin of replication	(also “ <i>ori</i> ” or “ <i>oriC</i> ”) the site which determines the ability of a plasmid to replicate within a cell, its copy number, and host range
origin of transfer	(also “ <i>oriT</i> ”) a DNA sequence allowing a plasmid to be mobilized by conjugation
orthologs	(also “orthologous genes” or “orthologous proteins”) homologs in different genomes
outer leaflet	of a lipid bilayer; the phospholipids in the half of the membrane facing the outside of the cell
overexpression strain	a strain for use with overexpression vectors, optimized for very high level production of cloned gene products
overexpression vector	a plasmid specifically designed to allow very high level production of a cloned gene product
oxygenic photosynthesis	photosynthetic pathways which use a H ₂ O as an electron donor
pairwise alignment	an alignment between two sequences
paralogs	(also “paralogous genes” or “paralogous proteins”) homologs in the same genome
parent strain	see “wild-type”; could also denote a strain from which a particular mutant strain was constructed
parsimony	the principle that the simplest explanation for a phenomenon that incorporates all of the available data is likely to be the best
passenger domain	the domain of a T5SS autotransporter which is passed through the outer membrane
pellicle	a biofilm that forms at a liquid-gas interface
penicillin-binding protein	(also “PBP”) see “transpeptidase”
pentose	a monosaccharide containing 5 carbons
pentose phosphate pathway	(also “PPP”) catabolic pathway that breaks down glucose into fructose-6-phosphate and glyceraldehyde-3-phosphate
peptide	a short protein
peptide bond	the amide bond (C-N) linking two amino acids in a protein
peptidoglycan	polymer of MurNAc and GlcNAc sugars, cross-linked by peptides; the main component of the bacteria cell wall
percent identity	what percentage of positions in an alignment of two homologous protein or nucleic acid sequences contain the same amino acid or nucleotide in both sequences
percent similarity	what percentage of positions in an alignment of two homologous proteins contain amino acids with similar chemical properties in both sequences
peripheral membrane protein	a protein that is associated with a lipid membrane, but not embedded in it
periplasm	(also “periplasmic space”) the space between the inner and outer membranes of a diplococcus bacterium
peritrichous	flagella distributed across the cell surface
permissive temperature	for temperature sensitive mutants, the temperature at which the gene functions

persister cells	a proportion of a bacterial population, that due to their low levels of ATP, are able to survive and eventually recover from antibiotic treatment
phage recombinase	a highly efficient recombinase derived from a lysogenic bacteriophage
phagemid	a plasmid with an fI origin
phenomenon	a measurable event in objective reality
phenotype	the measurable physical properties of an organism
phenotypic heterogeneity	a property of many clonal bacterial populations, where proportions of a genetically identical community will express different phenotypes
phosphatase	an enzyme that removes phosphate groups from a substrate
phospholipid	a hydrophobic lipid that contains one or more hydrophilic phosphate groups at one end
phosphorelay	a signaling network related to two-component regulatory systems, but containing more than two components
phosphorylation	covalent addition of a phosphate group to a molecule
phosphotransferase	an enzyme that transfers a phosphate group, some are involved in phosphorelay-type signaling pathways
phosphotransferase system	(also "PTS") sugar transport system that phosphorylates the substrate as it is imported
photoautotroph	a photosynthetic organism that can fix CO ₂ into organic carbon
photoheterotroph	a photosynthetic organism that does not fix CO ₂ into organic carbon (they use organic carbon sources instead)
photosynthesis	the process by which light is used as an energy source for generating a PMF or reducing equivalents
phototaxis	movement towards light
phylogenetic tree	a visualization of the evolution of organisms from common ancestors, typically based on the sequence homology of highly conserved genes
phylogeny	the evolutionary relationship between organisms or genes, inferred from homology
physiology	for the purposes of this class, the structure, metabolism, energetics, and development of bacteria as living organisms
pilot experiment	a quick experiment, meant to test the practicality of a more complex experiment
pilus	(plural "pili") a fiber or tube-like structure in which DNA is transferred from one bacterial cell to another (conjugation), or is involved in attachment to surfaces
planktonic cells	bacteria not currently attached to a surface
plasmid	a small DNA molecule capable of replicating in a bacterial cell
plasmid library	a pool of plasmids containing many different cloned inserts
plasmid map	a visual representation of a plasmid, with indications of important features and sites
pleiotropic phenotype	multiple, apparently unrelated phenotypes resulting from a single mutation
point mutation	a change in a single nucleotide in a genome

polar flagellum	a flagellum located at the pole of a rod-shaped bacterial cell
polarity	the fact that mutations of one gene in an operon can have effects on the expression of downstream genes in that operon
polycistronic	an mRNA encoding several genes
Polyketide synthase	(also "PKS") a biosynthetic pathway that constructs complex molecules out of simple carboxylic acid components
Polymerase chain reaction	(also "PCR") a very common method that uses DNA polymerase to amplify large amounts of a specific DNA molecule <i>in vitro</i>
polymicrobial	a biofilm or infection containing bacteria of more than one species
polyphyletic	a group of organisms descended from different ancestors
polysaccharide	a polymer of monosaccharides
porin	an integral membrane protein, usually in the outer membrane of diderms, that forms a channel through the membrane
positive control	a control that tests for the possibility of false negative results in an experiment
positive regulator	a regulator that activates the system being studied in response to a signal
post-translational modification	(also "PTM") a covalent modification of a protein that affects its activity
predatory bacteria	bacteria that obtain nutrients by attacking and killing other bacteria
predatory journal	a journal with no scientific standards that exists solely to make money
predictive power	the ability of a model to predict the behavior of reality
preprint	a scientific manuscript made publically available before being peer reviewed or accepted for publication in a journal
prestige journal	a journal which only publishes "high-impact" science; <i>Nature</i> , <i>Science</i> , <i>Cell</i> , etc.
primary literature	published papers directly reporting the results of scientific research
primer	see "oligonucleotide"
product inhibition	a property of some enzymes, whose reactions are slowed by high concentrations of product
programmed cell death	a developmental pathway which ends in the death of a cell
promoter	a DNA sequence that binds RNA polymerase and, potentially, regulators to control transcription of a gene
prophage	a bacteriophage that is integrated into a bacterial chromosome
prostheca	a membrane-enclosed, cytoplasm-containing bacterial appendage
protease	an enzyme that breaks peptide bonds in proteins
protein	a linear chain of amino acids, encoded by an mRNA and produced by a ribosome
protein fragment complementation	a technique for studying protein-protein interactions <i>in vivo</i> by dividing a reporter into two or more segments that only regain activity when brought into physical proximity
protein fusion	a single polypeptide encoded by sequence derived from more than one gene, or a protein artificially modified to add a small peptide sequence to its C- or N-terminal end

protein secretion	transport of proteins to cellular compartments other than the cytoplasm, including outside of the cell entirely
protein stability	how long a particular protein remains in the cell before being degraded
proteome	the complete set of proteins in a cell
proteomics	methods to quantify the entire set of proteins in a cell
proton motive force	(also "PMF") the energy source derived from the difference in H ⁺ concentration on either side of the cell membrane
prototroph	a strain that does not require a particular nutrient (compare to auxotroph and bradytroph)
pseudomurein	a peptidoglycan-like cell wall polymer found in some archaea
pseudopilus	a short pilus homolog involved in type II protein secretion systems
pulse-chase experiment	an experiment that briefly labels proteins and then follows their stability over time
pupylation	a posttranslational modification added to proteins in actinobacteria to direct their degradation by the bacterial proteasome
pyrophosphatase	an enzyme that hydrolyzes pyrophosphate to two orthophosphates
qRT-PCR	(also "quantitative reverse transcriptase PCR") a direct method to detect RNA by reverse transcribing it to DNA and amplifying it by PCR
qualitative measurement	a measurement that results in a categorical (non-numerical) value
quantitative measurement	a measurement that results in a numerical value
quinones	membrane-soluble electron shuttles essential for respiration (among other reactions)
quorum sensing	a mechanism depending on secretion of soluble signal molecules (autoinducers) that bacteria use to control gene expression in response to culture density
radiation	electromagnetic energy or energetic particles that damage DNA, resulting in mutations
random mutagenesis	any of a variety of methods of making mutations throughout a DNA molecule with no (or little) predetermined targeting
random walk	a pattern of movement that depends on random changes in direction
rare codons	codons that are not translated efficiently in an organism due to low numbers of tRNAs for that codon
reaction center	a chlorophyll-containing light-absorbing protein complex central to photosynthesis
reactive oxygen species	(also "ROS") H ₂ O ₂ , superoxide, singlet oxygen, and oxygen radicals; potent and toxic oxidants derived from oxygen
read-through transcription	transcription from the promoter of one gene that drives (often unwanted) expression of a downstream gene
recombinant DNA	a DNA molecule constructed with sequences from two or more different organisms
recombinase	an enzyme or enzyme system that catalyzes homologous recombination
recombination	see "homologous recombination"
recombineering	a method of constructing chromosomal mutations using phage recombinases and PCR products or oligos as templates

reducing equivalents	NADH, NADPH, reduced quinones, and other sources of electrons in a cell
regulon	the set of genes or proteins controlled by a regulator
repellant	a chemical or other stimulus that causes bacteria to move towards lower concentrations of that stimulus
replica printing	using a sterile piece of velvet as a printing block to transfer colonies to several different plates; useful for screens
replisome	the protein complex that catalyzes replication of the chromosome
reproducibility	a desirable property of experiments: they give the same result each time
respiration	a mechanism by which bacteria link the generation of a PMF to oxidation-reduction reactions by way of an electron transport chain
response regulator	the effector protein of a two-component regulatory system or phosphorelay
restriction enzyme	a nuclease that makes double strand breaks in or near a specific sequence in a DNA molecule
restriction methylase	a DNA methylase that blocks the activity of a particular restriction enzyme
restriction site	the DNA sequence recognized by a restriction enzyme
restrictive temperature	for temperature sensitive mutants, the temperature at which the gene does not function
reverse catabolite repression	catabolite repression in which the preferred carbon source is not glucose
reverse genetics	identifying phenotypes that arise from a mutation of interest
reverse transcription	the production of DNA from an RNA template by reverse transcriptase
revertant	a mutation that reverses the phenotype of a different mutation
review	a paper summarizing previous research on a particular topic
Rho-dependent transcription termination	transcription termination driven by the Rho protein, which recognizes single-stranded RNA with no ribosomes attached
Rho-independent transcription termination	transcription termination at intrinsic terminators
ribonuclease	a nuclease that specifically degrades RNA
ribosomally synthesized and post-translationally modified peptides	(also "RiPPs") peptide-based secondary metabolites synthesized by processing of a ribosomally-translated protein
ribosome binding site	(also "RBS") a short AG-rich sequence required for ribosomes to interact with mRNA and start translation
ribosome profiling	an indirect method to measure protein abundance using next-generation sequencing to quantify the proportion of each mRNA in a cell which is bound by ribosomes
riboswitch	a regulator formed entirely from RNA structures in an mRNA
RNA sequencing	(also "RNA-seq") a direct method to detect RNA by next-generation sequencing
rotor	what is turned by a stator; in the case of flagella, the flagellar basal body
S-layer	a protein or glycoprotein layer making up the outer surface of a cell envelope
sacculus	the purified cell wall of a bacterium, with no cell inside it

Sanger sequencing	a common and inexpensive way of sequencing several hundred to 1000 bp of DNA
saturation	for lipids, the number of double bonds present; a fatty acid with no double bonds is "saturated", while a fatty acid with several is "polyunsaturated"
scientific literature	the whole body of published scientific work
scientific method	a systematic approach to uncover truths about objective reality
screen	a mutant hunt in which each cell or colony must be individually analyzed to determine whether it contains a mutation of interest
second messenger	a small molecule that allosterically regulates multiple proteins, often produced in response to stressful changes in the cell's environment
secondary metabolism	anabolic pathways for the biosynthesis of molecules that are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids
secondary metabolites	molecules that are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids
secondary mutation	(see "revertant" and "multicopy", "intra-", and "intergenic suppressor") a mutation that is selected for by the presence of a primary mutation
selectable marker	a gene encoding a product which allows you to select for cells containing that gene; most often a gene for antibiotic resistance
selection	a mutant hunt in which the wild-type dies and only mutants of interest survive
selective pressure	conditions which favor the growth or survival of certain mutants over their parent strain
septum	the barrier that forms between two dividing cells
sequence logo	a visual representation of an alignment in which the relative frequency of particular nucleotides or amino acids is represented by letter size
serotype	classification system for bacteria based on reactivity to specific antibodies
serotyping	a method of distinguishing bacterial strains by the antibodies they react with
Shine-Dalgarno sequence	see "ribosome binding site"
short chain fatty acids	(also "SCFA") acetate, propionate, butyrate, and other small carboxylic acid lipids
shuttle vector	a plasmid used to move genes from one species to another; may have separate origins of replication for each species
sigma factor	(also "sigma subunit") a small protein component of RNA polymerase that determines the promoter sequence that will be bound
signal peptidase	an enzyme that cleaves signal peptides off proteins
signal recognition particle	(also "SRP") riboprotein complex that, with the Sec secretion system, is required for assembly of integral membrane proteins in the inner membrane
signal sequence	an N-terminal protein sequence that is recognized by cellular export machinery and directs the cell to secrete the protein
silent biosynthetic gene cluster	a BGC that isn't expressed under laboratory growth conditions
silent mutation	a mutation of an amino acid encoding codon to a different codon encoding the same amino acid
single nucleotide polymorphism	see "point mutation"

single-crossover recombination	a recombination event that requires only one homologous recombination event, such as integrating a circular plasmid into a circular chromosome
site-directed mutagenesis	(also “targeted mutagenesis”) constructing a specific mutation at a specific site in a DNA molecule
slime	a secreted polysaccharide that is slippery
society journal	a journal published by a scientific professional society
sortase	an enzyme that covalently attaches secreted proteins to peptidoglycan in the cell wall
SOS response	the LexA-dependent response to DNA damage stress
specialized transducing phage	lysogenic phage which are able to package some DNA from near their site of insertion into the chromosome of their host cell into virus particles
sphaeroplast	a cell from which the cell wall (and outer membrane, if it has one) has been removed
spontaneous mutagenesis	random mutations resulting from natural mistakes made by DNA polymerase during replication
spore	a metabolically-inactive, highly stress-tolerant resting state cell formed by some bacteria
spore coat	the outermost layers of an endospore, consisting of modified peptidoglycan and protein layers
sporulation	a process by which bacteria form metabolically inactive, stress-tolerant spores
sRNA	(also “small non-coding RNA”) a regulatory RNA that interacts with mRNA to change its expression, often by targeting it for degradation
stable lysogen	a DNA element incorporated into the bacterial chromosome that is derived from a lysogenic bacteriophage, but lacks the ability to re-enter the lytic lifecycle
stator	the MotAB motor proteins that turn the flagellar basal body
sticky end	a staggered double strand break produced by some restriction enzymes
stress response	a regulatory program induced by harmful changes in an organism’s environment
stringent response	a widely conserved starvation stress response pathway depending on the second messenger (p)ppGpp
subcloning	a protocol in which a DNA fragment from one plasmid is moved into another plasmid by restriction digestion and ligation
substrate analog	a non-natural molecule that can be acted on by an enzyme, often resulting in products that are easier to measure than the natural products
substrate inhibition	a property of some enzymes, whose reactions are slowed by high concentrations of substrate
substrate-level phosphorylation	reactions that produce ATP without being linked to electron transport (for example, during glycolysis)
suicide vector	a plasmid which can be introduced into a species, but does not replicate there, or one whose replication can be blocked under certain conditions (see temperature-sensitive origin of replication)
super-resolution microscopy	any of a variety of microscope technologies that allow visualization of individual fluorescent molecules in bacterial cells

swarm	a motile group of bacteria, moving across a surface
swarming	flagellum-dependent group motility along a surface
symporter	a transporter that transports two molecules across a membrane in the same direction
synthetic biology	the field of engineering that works to create artificial biological systems
synthetic lethality	two genes which can be knocked out individually, but not simultaneously
syntrophy	metabolic symbiosis in which one organism consumes the metabolic endproduct of another organism
teichoic acid	polymer of repeating sugars and phosphate groups that makes up a substantial portion of the cell envelope of monoderm bacteria
temperature-sensitive mutant	a mutant that grows at low temperature, but not at high temperature; typically due to mutations that destabilize essential proteins
temperature-sensitive origin of replication	an origin of replication that only functions at low temperature, typical of some suicide vectors
Ter macrodomain	the region of the bacterial chromosome where DNA replication terminates
terminal differentiation	a developmental process in which a proportion of cells in a population change their phenotypes irreversibly
terminal electron acceptor	the last molecule to be reduced in a respiratory electron transport chain
terminator	a sequence which stops transcription
testable	see "falsifiable"
thylakoid membrane	an internal membrane structure, rich in photosynthetic reaction centers, in cyanobacteria
tolerance	the ability of non- or slowly-growing bacteria to survive treatment with antibiotics
<i>tra</i> functions	genes encoding the machinery that allows transfer of plasmids with an appropriate <i>oriT</i> by conjugation
transconjugant	a cell that has incorporated DNA delivered by conjugation
transcription	the production of mRNA from a DNA template by RNA polymerase
transcription elongation	the activity of RNA polymerase actively producing mRNA
transcription factor	a protein that binds to the promoter of a gene to control its transcription
transcription initiation	the process by which RNA polymerase begins transcribing a gene into mRNA
transcription termination	the process by which RNA polymerase releases DNA and stops transcribing
transcriptional activator	a transcription factor that increases transcription of a gene
transcriptional attenuation	a regulatory mechanism in which an mRNA can take on more than one structural conformation, one of which is an intrinsic terminator
transcriptional pause site	a DNA or RNA sequence where RNA polymerase briefly stops producing mRNA
transcriptional reporter fusion	an indirect method to measure transcription by placing an easily-measured gene product under control of a promoter of interest
transcriptional repressor	a transcription factor that reduces transcription of a gene

transcriptional start site	the point in a promoter sequence where RNA polymerase begins producing mRNA
transcriptome	the entire set of mRNAs in a cell
transcriptomics	methods to quantify the entire set of mRNAs in a cell
transductant	a cell that has incorporated DNA derived from a transducing phage
transduction	DNA transfer between cells mediated by bacteriophage
transformant	a cell that has incorporated DNA delivered by transformation
transformation	bacterial cells taking up DNA directly from their environment
transition	a mutation of a purine (A or G) to a purine or of a pyrimidine (T or C) to a pyrimidine
translatability	a measure of how easily an mRNA is translated into protein in a particular organism
translation	the production of protein from an mRNA template by ribosomes
translation elongation	the activity of ribosomes actively producing protein
translation initiation	the process by which ribosomes bind to mRNA and begin producing protein
translational reporter fusion	an indirect method to measure translation by placing an easily-measured gene product under control of the promoter and translation initiation signals of a gene of interest
transpeptidase	enzyme that catalyzes the formation of peptide crosslinks in peptidoglycan
transposon	(also “insertion element”) a DNA sequence capable of inserting itself into another DNA sequence, often at random
transposon library	a pool of transposon mutants, each cell containing only one transposon, but with a total of tens or hundreds of thousands of different insertion sites
transposon sequencing	(also “Tn-seq”, “INSeq”, “TraDIS”, or “HITS”) a technique that uses next-generation sequencing technology to identify all of the insertion sites in a transposon library
transversion	a mutation of a purine (A or G) to a pyrimidine (T or C) or vice versa
treadmilling	by cytoskeletal proteins; rapidly polymerizing at one end while simultaneously depolymerizing at the other end
treatment	see “independent variable”
tri-carboxylic acid cycle	(also “TCA cycle” or “Krebs cycle”) catabolic cycle that breaks down pyruvate to CO ₂
twitching motility	a form of motility along surfaces dependent on the sequential attachment and retraction of type 4 pili
two-component system	a family of bacterial regulators consisting of histidine kinases and response regulators whose activity is controlled by phosphorylation
two-hybrid screening	see “protein fragment complementation”
two-partner secretion system	a variation on type V protein secretion systems that involves two proteins
uncoupler	chemical that bind to cations and allows them to diffuse through membranes, can be used to collapse ion gradients
untranslated region	the parts of an mRNA which do not encode protein; often include regulatory elements

UP element	AT-rich sequence upstream of the -35 site of a promoter that increases transcription 30 to 70-fold
upstream gene	a gene encoded 5' of the gene being discussed on a polycistronic mRNA
vector	see "plasmid"
vector-only control	a type of negative control in which a strains containing an empty plasmid is compared to the same plasmid containing a gene of interest
vegetative cell	a growing bacterial cell, as opposed to a spore
wall teichoic acid	(or "WTA") polymer of repeating sugars and phosphate groups covalently attached to the cell wall of monoderm bacteria
western blot	a direct method of detecting proteins using antibodies specific to those proteins
wild-type	a strain that does not contain a particular mutation of interest
Z-ring	the ring of proteins, organized by FtsZ, that is the necessary precursor to formation of the divisome and initiation of cell division