BACTERIAL GENETICS (GBSC 716)

University of Alabama at Birmingham, Microbiology Theme
January 2020
Course Packet for Lectures 1 – 9
Dr. Michael J. Gray

TABLE OF CONTENTS

LECTURE 1: 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 lecture will be the primary determinant of your grade**, and most of the "lecture" time (at least for the first half of the 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 go through more advanced material and discussion problems during the "lecture" itself. I do not plan to spend very much time actually lecturing at you.

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, our 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 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. We want to give you the tools to effectively access and add to that knowledge.

By the end of the first nine lectures 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
- 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 83. See page 82 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 for lectures 1 through 9 using the following scale:

3 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.

2 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.

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

0 points
Has not prepared for class or does not contribute to discussion; displays disrespect towards students and/or faculty, or is absent without explanation.

If you are concerned about your grade or class status at any point during the class, please contact me immediately. I am happy to talk to you outside of class to try to clear up any confusing points. Dr. Yother also has tutors available, if you feel that you need extra help.

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 hypothesis to the test, **discarding ideas 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 fit 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. By reiterating this self-correcting process, scientific knowledge converges on truth.

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 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 particular 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.

"Molecular Genetics of Bacteria", by Larry Snyder and Wendy Champness and "Fundamental Bacterial Genetics" by Nancy Trun and Janine Trempy are excellent textbooks 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 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:

en.wikipedia.org/wiki/Transcription_(genetics)

en.wikipedia.org/wiki/Translation_(biology)

(Wikipedia is a surprisingly reliable source for information on biochemistry. There's nothing particularly controversial about, say, the molecular weight of salicylic acid.)

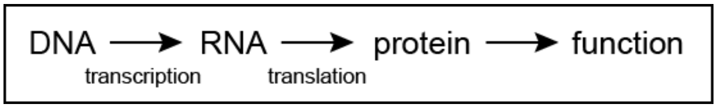
Different organisms are more or less easy to manipulate, and it's important to understand what is technically possible in the species you're studying. An experiment that takes a week in a well-established *model organism* like *Escherichia coli* may take months, years, or be impossible in a slow-growing, 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 modern methods for manipulating DNA molecules in **Lectures 6 - 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.

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, 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. 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. 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.

GENETIC NOMENCLATURE (IN BACTERIA)

For bacteria and archaea, there is a straightforward and 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* Δ*ppk 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 investigator followed by a number. For example, strain MG1655 was isolated by Mark Guyer in 1981, and was probably 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 "OR: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.)

A typical *E. coli* genome contains about 4000 genes in total, which is near the middle of the range for most types of bacteria. *Streptomyces* species can have more than 8000 genes, 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. In *E. coli*, *genes of unknown function* (and there are still many hundreds of these) have names starting with the letter "y" (*e.g. ydjA* or *yeaG,* or the *yci* genes in Figure 1.1), which indicates the location of that gene on the chromosome, and 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.)

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 separate 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.

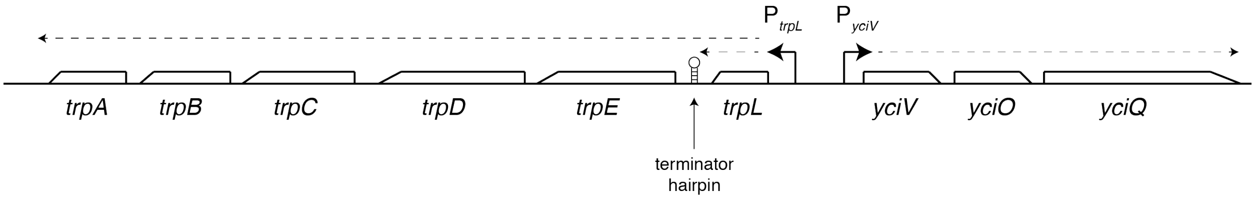


Figure 1.1. The *trp* locus of *E. coli*, to illustrate conventions of drawing genes and operons. The *trpA-E* genes are required for tryptophan synthesis. They and the tryptophan-rich leader peptide encoded by *trpL* are expressed as an operon from the P_{trpL} promoter. Two mRNA transcripts are possible 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 later in the course). The *yci* genes are *divergently transcribed* (that is, transcribed in the opposite direction) from the P_{yciV} 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.

When looking at the genotype of a bacterium, the general rule is that any gene not mentioned is assumed to have 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 *Coli* Genetic Stock Center (cgsc.biology.yale.edu/) 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 (scholar.google.com) 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 Δ*ppk* 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. 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 ("Fertility factor") is a large circular DNA element (~ 100 kb) found in natural *E. coli* isolates which is capable of transferring itself (by conjugation, which we will talk about in **Lecture 7**) to other *E. coli* strains. F$^+$ strains are sometimes called "male" strains, while F$^-$ strains like MG1655 lack the F plasmid and are sometimes called "female". (It's not actually a very good analogy, since only the F plasmid is transferred and the "female" recipient then becomes "male".) "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. 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. 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. 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.

For more details on the rules for writing bacterial genotypes, see the instructions on genetic nomenclature from the Journal of Bacteriology (http://jb.asm.org/site/misc/journal-ita_nom.xhtml - 03). You can find the genotypes of many lab strains of *E. coli* at openwetware.org/wiki/E._coli_genotypes. 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 (http://www.atcc.org) and the German DSMZ collection (www.dsmz.de/home.html). These collections maintain stocks of thousands of strains of bacteria and other organisms that have been deposited by researchers around the world. For *E. coli* strains, the *Coli* Genetic Stock Center mentioned above is also a great resource. Many of the most 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 those labs, and most researchers are happy to share published strains with their fellow scientists, and in fact, 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.

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. 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. *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 (www.ncbi.nlm.nih.gov/pubmed/). PubMed allows you to automate searches of the literature for particular keywords, which is a great way to make sure you don't miss any publications directly relevant to your interests. **Google Scholar** (scholar.google.com) 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: www.benchfly.com/video/121/performing-a-pubmed-lit-search/

Video supplement: www.benchfly.com/video/114/gene-searching-on-pubmed/

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 *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 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 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 http://www.eigenfactor.org for an alternative, more robust measure of journal quality. While the quality of a paper is not necessarily linked to the quality of the journal in which it is published, higher quality journals will usually have more rigorous peer review and standards for publication, and will 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, the Journal of Biological Chemistry, 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.

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. The largest of these is **GenBank**, from the NCBI (www.ncbi.nlm.nih.gov/genbank/), which contains all publicly available DNA sequences. A favorite of mine is the **Integrated Microbial Genomes** system (img.jgi.doe.gov/cgi-bin/w/main.cgi), which contains all of the information obtained from the full genome sequences of over 70,000 organisms and close to 20,000 *metagenomes* from different environments or bacterial communities (as of November 2019, but that number will rapidly become outdated as more become available).

Additional databases and resources that you may find useful include:

EcoCyc (ecocyc.org): a very well curated repository of information on the model organism *E. coli*, combining large amounts of manually compiled 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. (*Subti*Wiki for *Bacillus subtilis*, for example.) **MetaCyc** (metacyc.org) 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*. The BioCyc app (available for iOS) accesses these databases and is a convenient tool for quickly looking up genes of interest.

PATRIC (www.patricbrc.org): a very comprehensive database of bacterial 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 PATRIC much myself, but it has a lot of very powerful tools for genome comparison and analysis.

RegulonDB (http://regulondb.ccg.unam.mx): 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. **PRODORIC** (http://prodoric.tu-bs.de) is similar, but contains regulatory network information from a much broader assortment of bacterial species.

BioNumbers (bionumbers.hms.harvard.edu/default.aspx): a remarkably useful database that collects biological "trivia" 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 references you need.

KEGG (http://www.genome.jp/kegg/), the Kyoto Encyclopedia of Genes and Genomes: *the* database for 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 (http://www.rcsb.org/pdb/home/home.do): 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 PyMOL (http://www.pymol.org) or CCP4 (http://www.ccp4.ac.uk).

BRENDA (http://www.brenda-enzymes.org/index.php): 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.

EcoSal Plus (http://www.asmscience.org/content/journal/ecosalplus): is the online descendent of the book "*Escherichia coli* and *Salmonella*: Cellular and Molecular Biology", and is an exceptional collection of reviews summarizing all aspects of these important model organisms. It is regularly updated and definitive, and if your institution has a subscription (and UAB does) it is 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 "lecture" time will be devoted to talking about and working through these problem sets as a class. 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 SMc00166?

- what species / strain is this gene found in?
- what is its common name / gene symbol?
- draw the SMc00166 locus, indicating genes and operons near SMc00166 in the chromosome and their functions (if known)
- what is known about the function of this gene? (give 2 citations)

BLAST SEARCHES

Databases 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*" (Basic Local Alignment Search Tool)). 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 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 matches to each other, and calculates the statistical significance of those matches (as an e-value; like p-values, a smaller number indicates higher statistical significance). BLAST is commonly used to identify members of gene families or to infer evolutionary or functional relationships between sequences.

The most common place to do BLAST searches is via the BLAST page at the National Center for Biotechnology Information (blast.ncbi.nlm.nih.gov/Blast.cgi). This will allow you to search nucleotide or protein sequences against GenBank, NCBI's database of sequence information. 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 (img.jgi.doe.gov/cgi-bin/m/main.cgi), which contains only sequences from complete genomes and can filter searches in a variety of ways. This 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 (http://www.biopython.org) as a very accessible and flexible system for writing bioinformatics programs. Many professional bioinformaticians seem to prefer R (https://www.r-project.org), a programming language that provides very powerful tools for statistical analysis.

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 RclR):

```
>646312216 NP_414839 transcriptional regulator, AraC family [Escherichia coli str. K-12 substr.
MG1655 chromosome: NC_000913]
MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHALTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDNESTCIVCGTLRLQHS
ARYFLTSLPETLFLAPVNHSVEYNWLREAIPFLQQESRSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGYASESSFHKAFVREFGCTPGEYRERVRQLAP
```

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 itself. A FASTA-formatted sequence file can contain any number of sequences in this format. Here, for example, are three *E. coli* genes involved in hydrogenase activity (note that FASTA format can contain either protein or DNA sequences, but not both):

```
>hyfA
ATGAACCGCTTTGTGGTGGCCGAACCACTGTGGTGTACAGGATGTAATACCTGTCTCGCTGCCTGTTCGGACGTGCATAAAACGCAAGGTTTACAGC
AACACCCGCGCCTGGCCCTGGCGAAGACGTCAACAATCACTGCCCCTGTCGTGTGTCATCACTGTGAGGAAGCCCCTTGCCTGCAGGTCTGCCCGGT
CAATGCCATCTCTCAGAGGGATGATGCGATCCAACTCAACGAAAGCCTCTGTATTGGCTGCAAGCTTTGCGCCGTGGTCTGCCCATTTGGCGCAATC
AGCGCTTCAGGAAGCCGTCCGGTGAATGCCCATGCGCAATATGTTTTTCAGGCTGAAGGCTCACTCAAAGACGGCGAAGAAAACGCGCCAACACAAC
ATGCTTTGCTGCGCTGGGAACCTGGTGTCCAGACCGTCGCGGTGAAATGCGACCTGTGTGATTTCTTGCCAGAAGGTCCGGCCTGCGTTCGCGCTTG
CCCGAATCAGGCGTTACGGCTGATCACCGGTGATAGCCTGCAACGTCAGATGAAAGAAAAACAGCGCCTTGCCGCAAGCTGGTTTGCCAATGGCGGG
GAGGATCCCCTTTCCCTCACTCAGGAGCAACGCTAA
```

```
>hyfC
ATGAGACAAACTCTTTGCGACGGATATCTGGTCATTTTTGCGTTAGCACAGGCCGTGATTCTGCTGATGCTAACCCCACTTTTTACGGGTATTTCCC
GGCAGATACGCGCGCGTATGCACTCCCGCCGCGGGCCGGGGATCTGGCAGGATTATCGCGATATCCACAAACTGTTTAAACGCCAGGAAGTTGCGCC
GACATCTTCAGGTCTGATGTTCCGCCTGATGCCGTGGGTATTAATCAGCAGCATGCTGGTGCTGGCGATGGCCTTACCACTGTTTATTACCGTTTCC
CCTTTTGCGGGCGGCGGCGATCTGATCACCCTTATCTATCTTCTTGCCCTGTTTCGTTTTTTCTTTGCTCTTTCCGGGCTGGATACCGGAAGTCCGT
TTGCGGGAGTCGGTGCCAGTCGCGAGTTGACGCTCGGCATTCTGGTCGAACCAATGCTTATTCTCTCACTGCTGGTATTGGCGCTGATAGCAGGTTC
CACGCATATCGAGATGATCAGCAATACGCTGGCGATGGGCTGGAACTCGCCGCTAACCACCGTACTGGCGTTACTGGCCTGTGGTTTTGCCTGCTTC
```

```
ATTGAGATGGGAAAAATTCCCTTTGATGTTGCTGAAGCAGAACAGGAATTACAGGAAGGCCCGCTGACCGAATATTCCGGTGCCGGGCTGGCGCTAG
CGAAATGGGGGCTGGGGCTGAAACAGGTCGTGATGGCATCACTGTTTGTGGCCCTGTTTCTGCCCTTTGGGCGCGCGCAAGAACTTTCTCTCGCCTG
CCTGCTGACTTCACTTGTCGTTACGCTGCTCAAGGTTTTGCTGATTTTTGTACTGGCCTCAATCGCAGAAAACACGCTGGCACGCGGGCGTTTTTTA
CTCATTCACCATGTGACCTGGCTTGGCTTCAGCCTTGCTGCGCTTGCATGGGTCTTCTGGTTAACCGGTCTGTAA
```

```
>hyfE
ATGACCGGTTCTATGATCGTAAATAATCTGGCGGGACTGATGATGCTGACATCGCTGTTTGTGATTAGCGTCAAAAGCTATCGCCTGTCATGCGGAT
TTTACGCCTGCCAGTCACTGGTGCTGGTGTCTATTTTCGCCACTCTCTCGTGCCTGTTCGCCGCAGAGCAACTGCTGATCTGGTCCGCCAGCGCCTT
TATCACCAAAGTGCTGCTGGTACCGTTAATCATGACTTACGCTGCACGAAATATTCCCCAGAACATCCCGGAAAAAGCGTTATTCGGTCCGGCAATG
ATGGCACTGCTCGCGGCGTTAATTGTCCTGCTTTGCGCATTTGTCGTTCAGCCCGTGAAGCTACCGATGGCTACCGGGCTGAAACCGGCGCTGGCGG
TAGCGTTAGGTCATTTTCTGCTTGGCCTGCTGTGCATTGTCAGCCAGCGCAATATCCTGCGGCAAATTTTTGGTTACTGCCTGATGGAAAACGGCTC
CCATCTGGTGCTGGCGCTTCTTGCCTGGCGAGCACCGGAACTGGTGGAAATAGGTATCGCTACCGACGCCATCTTCGCCGTCATTGTGATGGTGTTA
CTGGCAAGAAAAATATGGCGTACCCACGGCACGCTGGACGTGAACAACTTGACCGCGCTGAAGGGATAA
```

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 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 BLAST2seq from NCBI. This program will take two protein or nucleotide sequences and BLAST one (the "query") against the other (the "subject"), giving you a sequence alignment and additional information including an e-value (similar to a p-value, this is a statistical measure of how likely the similarities between two sequences is to have arisen purely by chance), 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* RclR protein sequence above with a homologous sequence from *Klebsiella pneumoniae* ("Expect" is the e-value, in this case 2 x 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    MDSLSHLLALLAPRCEVNLHCRFGGRWQAGHQQMRSGVVPWHVVLRGEGRLNV-GGQTHH  59
            MD+LS LL L AP+  ++ +C  G  WQ  H      V+ WH + +G  +L +  G+
Sbjct  1    MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHALTQGAAKLEMPTGEIFT  60

Query  60   LRAGDVVLLPHGSPHLMESLVEWGQVLPVAHRFNGTVTEMRAGPAEGALEMLCGEFYFGP  119
            LR G+VVLLP  S                AHR +    E           ++CG
Sbjct  61   LRPGNVVLLPQNS----------------AHRLSHVDNESTC--------IVCGTLRLQH  96

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

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

Query  239  VYHLTPQAWLSQLRMALAARLLRLERQTNLEVIAERCGFQSLASFSKRFKMRYGVTPGEW  298
            V   TP A L++LR+ +AA++   E    + VIAE  G+ S +SF K F   +G TPGE+
Sbjct  217  VSGTTPLAVLTKLRLQIAAQMFSRE-TLPVVVIAESVGYASESSFHKAFVREFGCTPGEY  275

Query  299  R  299
            R
Sbjct  276  R  276
```

Notice that in this format amino acid residues identical in both proteins (*conserved residues* or "identities") are shown with that amino acid letter in between the query and subject sequences and that chemically similar amino acids ("positives") 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 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 http://www.ebi.ac.uk/Tools/msa/. I often use MUSCLE for protein alignments, but Clustal Omega is also excellent. Use an alignment program appropriate for your particular samples. 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           -----------------MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
Methylobacterium      MAGPIRRRAGAPETAGADDPLSGLAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
Proteus               -----------------MDTLSQLLYLSQGQLQLDVFCQMKGHFSLPHVSSVEHETIFHL
Bordetella            -----------------MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEAAFHL
Pseudomonas           -----------------MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI
                                        * *. *  *   .   *  * : . :   *         :*

Escherichia           LTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDN---------------------
Methylobacterium      VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG-PPPSTPLPVAVRHA--HDLRF
Proteus               VLSGQCYVQIEKSAPIVLSEGTFLMLNRRQSHTLWSGERDIEP--PPFLHKNNGFLPVKY
Bordetella            LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRLPLRR
Pseudomonas           VMAGTCHCEFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ
                      :  * .  .   .  : :  * .::* :   * :

Escherichia           -------ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
Methylobacterium      KTTVGAEPDVELICGRLAFEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLAGIREEL
Proteus               TKSEDQTQHVDLLCGRMAYAKGSGLLLLNGFPDMVVA-NLVEMPGLTVLNLFSQLLREEA
Bordetella            NTAPEQQADVDLLCGRFSYDRGAGDLFARSLPGVLHV-PLA-H-HLPQLQPLIAMLRAEA
Pseudomonas           LNGPG--EALDMLCGSYRYHAGASLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA
                                 ::**      .   :   .:*  :    :        :      :: *

Escherichia           RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
Methylobacterium      NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRLLGQPLTARAVLAMVRDPVHPW
Proteus               INANQGAAAILNGLAQTLFAFALRVYGQKPDINSSWLALLAEPRLSRVFNSMLNEPQKGW
Bordetella            ASPLPGAAAVINALGQALLALALRAYGQREEVPANMLALAADSRIGPSVRAMIQDPGQAW
```

```
Pseudomonas           ESTRSGARSIIDALATALFALTLRAYLDRQPLGDGLFGLLGDARLGRALQVMLECPEQAW
                       .   *  :: . :  ::: : :*            : *  .   .  .  *:. * : *

Escherichia           TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGY
Methylobacterium      TLDALAATAAASRASLVRAFRAAAGVAPLEFLTDLRLGLAHHRLRTETVSLDRLAAEVGY
Proteus               TLDSLANVASMSRATFVRQFKATANTTPGEVLQSIRMLKALSLLQQNKYTLSDIAERVGY
Bordetella            TIETLGNKAAMSRATYARHFRSRAGMTVGEFLLRIRMMHASALLNHSQRSQRDIAEQVGY
Pseudomonas           TVERLAQQAAMSRASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAAETGY
                      *:: *.  *  ***: .: *   :  :   .*  :*:  *   :         :*  .**

Escherichia           ASESSFHKAFVREFGCTPGEYRERVRQLAP-----------
Methylobacterium      QSAAALSRAFLRKYGIRPGQARQAEAPPAG-----------
Proteus               QSEAAFSKAFKSVFNCRPGQWKKQQSKV-------------
Bordetella            QSEAAFGKAFREIMGQTPGQWRRLHRNARPVDTARRSDPKQ
Pseudomonas           QSEAAFSRNFRQAFGESPGRFRRQADASR------------
                       * ::: : *       **. :.
```

In Clustal format, the punctuation under each block indicates conserved positions: "*" indicates completely conserved residues, ":" indicates very similar residues, and "." 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:

```
Escherichia           -----------------MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
Methylobacterium      MAGPIRRRAGAPETAGADDPLSGLAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
Proteus               -----------------MDTLSQLLYLSQGQLQLDVFCQMKGHFSLPHVSSVEHETIFHL
Bordetella            -----------------MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEAAFHL
Pseudomonas           -----------------MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI

Escherichia           LTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDN---------------------
Methylobacterium      VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG-PPPSTPLPVAVRHA--HDLRF
Proteus               VLSGQCYVQIEKSAPIVLSEGTFLMLNRRQSHTLWSGERDIEP--PPFLHKNNGFLPVKY
Bordetella            LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRLPLRR
Pseudomonas           VMAGTCHCEFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ

Escherichia           -------ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
Methylobacterium      KTTVGAEPDVELICGRLAFEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLAGIREEL
Proteus               TKSEDQTQHVDLLCGRMAYAKGSGLLLLNGFPDMVVA-NLVEMPGLTVLNLFSQLLREEA
Bordetella            NTAPEQQADVDLLCGRFSYDRGAGDLFARSLPGVLHV-PLA-H-HLPQLQPLIAMLRAEA
Pseudomonas           LNGPG--EALDMLCGSYRYHAGASLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA

Escherichia           RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
Methylobacterium      NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRLLGQPLTARAVLAMVRDPVHPW
Proteus               INANQGAAAILNGLAQTLFAFALRVYGQKPDINSSWLALLAEPRLSRVFNSMLNEPQKGW
Bordetella            ASPLPGAAAVINALGQALLALALRAYGQREEVPANMLALAADSRIGPSVRAMIQDPGQAW
Pseudomonas           ESTRSGARSIIDALATALFALTLRAYLDRQPLGDGLFGLLGDARLGRALQVMLECPEQAW

Escherichia           TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGY
Methylobacterium      TLDALAATAAASRASLVRAFRAAAGVAPLEFLTDLRLGLAHHRLRTETVSLDRLAAEVGY
Proteus               TLDSLANVASMSRATFVRQFKATANTTPGEVLQSIRMLKALSLLQQNKYTLSDIAERVGY
Bordetella            TIETLGNKAAMSRATYARHFRSRAGMTVGEFLLRIRMMHASALLNHSQRSQRDIAEQVGY
Pseudomonas           TVERLAQQAAMSRASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAAETGY

Escherichia           ASESSFHKAFVREFGCTPGEYRERVRQLAP-----------
Methylobacterium      QSAAALSRAFLRKYGIRPGQARQAEAPPAG-----------
Proteus               QSEAAFSKAFKSVFNCRPGQWKKQQSKV-------------
Bordetella            QSEAAFGKAFREIMGQTPGQWRRLHRNARPVDTARRSDPKQ
Pseudomonas           QSEAAFSRNFRQAFGESPGRFRRQADASR------------
```

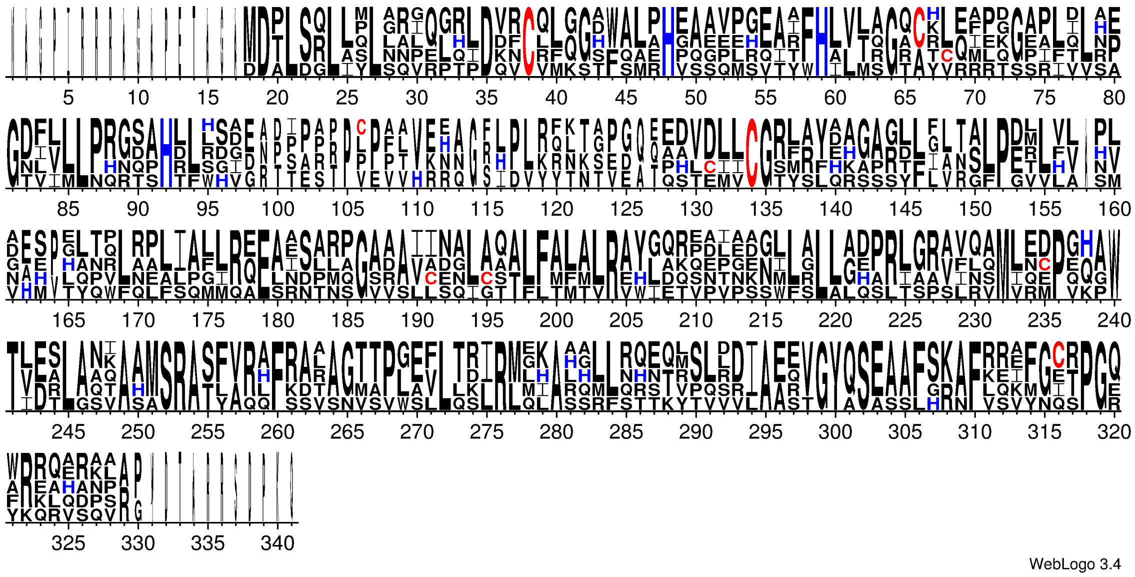
You can also present the same alignment in FASTA format, which is less human-readable, but more convenient for handling larger numbers of sequences:

```
>Escherichia
-----------------MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
LTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDN---------------------
-------ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGY
ASESSFHKAFVREFGCTPGEYRERVRQLAP-----------
```

```
>Methylobacterium
MAGPIRRRAGAPETAGADDPLSGLAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG-PPPSTPLPVAVRHA--HDLRF
KTTVGAEPDVELICGRLAFEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLAGIREEL
NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRLLGQPLTARAVLAMVRDPVHPW
TLDALAATAAASRASLVRAFRAAAGVAPLEFLTDLRLGLAHHRLRTETVSLDRLAAEVGY
QSAAALSRAFLRKYGIRPGQARQAEAPPAG-----------
>Proteus
-----------------MDTLSQLLYLSQGQLQLDVFCQMKGHFSLPHVSSVEHETIFHL
VLSGQCYVQIEKSAPIVLSEGTFLMLNRRQSHTLWSGERDIEP--PPFLHKNNGFLPVKY
TKSEDQTQHVDLLCGRMAYAKGSGLLLLNGFPDMVVA-NLVEMPGLTVLNLFSQLLREEA
INANQGAAAILNGLAQTLFAFALRVYGQKPDINSSWLALLAEPRLSRVFNSMLNEPQKGW
TLDSLANVASMSRATFVRQFKATANTTPGEVLQSIRMLKALSLLQQNKYTLSDIAERVGY
QSEAAFSKAFKSVFNCRPGQWKKQQSKV-------------
>Bordetella
-----------------MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEAAFHL
LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRLPLRR
NTAPEQQADVDLLCGRFSYDRGAGDLFARSLPGVLHV-PLA-H-HLPQLQPLIAMLRAEA
ASPLPGAAAVINALGQALLALALRAYGQREEVPANMLALAADSRIGPSVRAMIQDPGQAW
TIETLGNKAAMSRATYARHFRSRAGMTVGEFLLRIRMMHASALLNHSQRSQRDIAEQVGY
QSEAAFGKAFREIMGQTPGQWRRLHRNARPVDTARRSDPKQ
>Pseudomonas
-----------------MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI
VMAGTCHCEFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ
LNGPG--EALDMLCGSYRYHAGASLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA
ESTRSGARSIIDALATALFALTLRAYLDRQPLGDGLFGLLGDARLGRALQVMLECPEQAW
TVERLAQQAAMSRASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAAETGY
QSEAAFSRNFRQAFGESPGRFRRQADASR------------
```

In FASTA alignment format, each protein sequence is listed separately, with gaps indicated by "-". 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 (weblogo.threeplusone.com) is my tool of choice 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:



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 a 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
--------MDXLSXLLXLXRXQXRLDVRCQLXGXWALPHEAAVPGEAXFHLVXAGQCXLXXPXGAPLXLXXGDFXLLPRGSAHLLXSXE--XPX-
PXPXXVEXAG-LPLRXXTXPGX-EDVDLLCGRLXYXAGAXLLXLTXLPXXLVXPXXESXXLTXLRPLIALLRXEAXSARPGAAAXINALAQALFA
LALRAYXQRXXXXXXLLALLXDPRLGRXVQAMLEDPGXAWTXESLANXAAMSRASFVRAFRAXAGTTPXEXLTRIRMXKAXXLLRQEXXSLXDIA
EEVGYQSEAAFSKAFRRXFGCXPGQWRRQXRXXXX-----------
```

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 "-" 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 www.bioinformatics.org/sms/iupac.html.

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

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 1: 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.

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 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 (by van Leeuwenhoek, Hooke, and many others) 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 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 a more modern example of the same process of technological improvement leading to new kinds of observations.

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 are categorical or descriptive. Beware of assigning numerical values to categorical measurements and then treating them as quantitative.

THE GENETIC TOOLKIT

At a very simple level, molecular genetics techniques do one of two things: move new DNA into a cell or change the genes 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 6 - 9** we'll discuss different ways of moving new DNA into cells, as well as homologous recombination, a mechanism which can incorporate new DNA into a cell's genome.

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 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.)

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 mutations 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 cholera* 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 a 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 is often written as being "His$^-$", pronounced "hiss-minus".) Phenotypes can be "strong", "weak", or "leaky", terms that are not strictly defined, but generally express how easy they are 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, 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. A strain that grows slowly in the absence of a particular nutrient is a *bradytroph*, although this is a much less commonly used term.

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). 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.

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 attempt 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 do 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 cellular growth rate. 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.

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 www.russelllab.org/aas for a detailed resource on 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. *Silent mutations* are point mutations that, due to the *degeneracy* of the amino acid code (that is, 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.

Table 2.1. Types of point mutations

transition	purine (AG) to purine, pyrimidine (CT) to pyrimidine
transversion	purine to pyrimidine, 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

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.

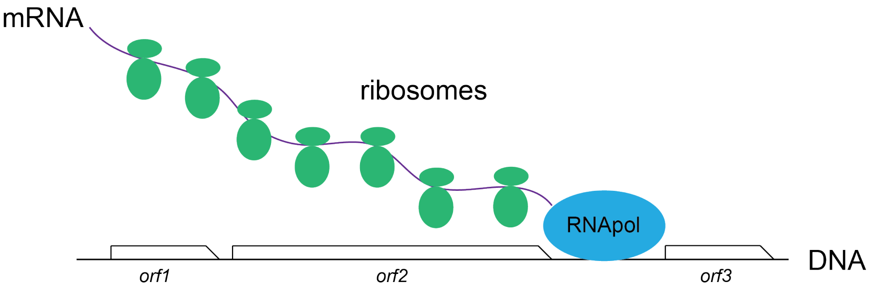
It is important to note that *null mutations*, in which the product of a mutated gene loses activity (also called *gene knockouts*), are always more common than *gain-of-function mutations*, where a new activity is generated, but all types of mutation can result in the addition of new functions under 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*.

Some mutations will be *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. You will not be able to isolate lethal mutants in the lab without specialized methods. Lethal mutations could include null mutations of *essential genes* or gain-of-function mutations creating toxic effects. Some mutations result in *conditional phenotypes*, in which phenotypes (often lethality) are only observed under some conditions. A common and useful example of this are *temperature-sensitive* mutants, which result in gene products that are destabilized and do not function at high temperatures.

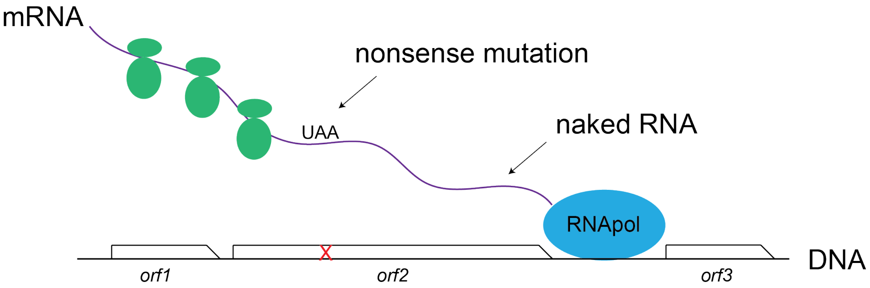
POLARITY

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. 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 most polar effects occur.

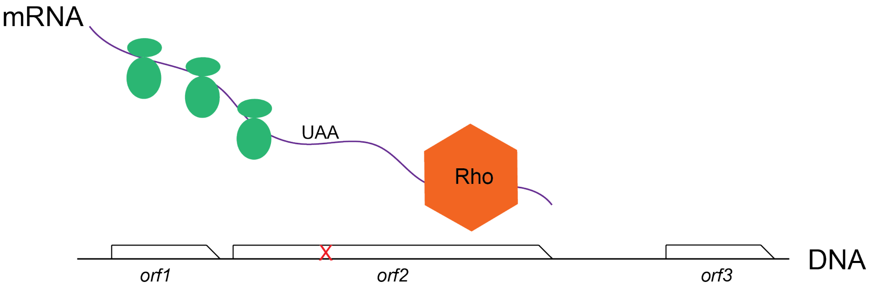
Normally, when bacterial RNA polymerase is transcribing an operon (shown in the figure below as *orf1* – *orf2* – *orf3*), the mRNA produced is coated in ribosomes actively translating that mRNA into protein. (Transcription and translation are linked in bacteria.)



However, 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):



In bacteria that contain a homolog of a protein called Rho (which most bacteria have), that protein recognizes such stretches of "naked" RNA, chases RNA polymerase, and when it catches up, it stops transcription, causing RNA polymerase to 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, but in this case, it can 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*.



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. *Random mutagenesis* creates mutations at random points 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.**) 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 knockouts of any given gene occur in about 1 in every 10^5 cells, although these numbers can vary widely depending on the region of the chromosome involved. The frequency at which mutations yielding a particular phenotype arise can be very informative. For example, if the phenotype you are looking for arises spontaneously in 1 in every 10^5 cells, you can reasonably conclude that it could be caused by a loss-of-function mutation in a single gene.

Table 2.2 lists the rough frequency to be expected for different kinds of spontaneous mutations. We will discuss recombination, plasmids, and transposons in future lectures.

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

spontaneous knockout of gene function: 10^{-5}
any particular 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 tandem 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: $10^{-1} – 10^{-2}$

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 $10^{-5} \times 10^{-5} = 10^{-10}$. We will discuss ways of increasing the rate of random mutations 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 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:

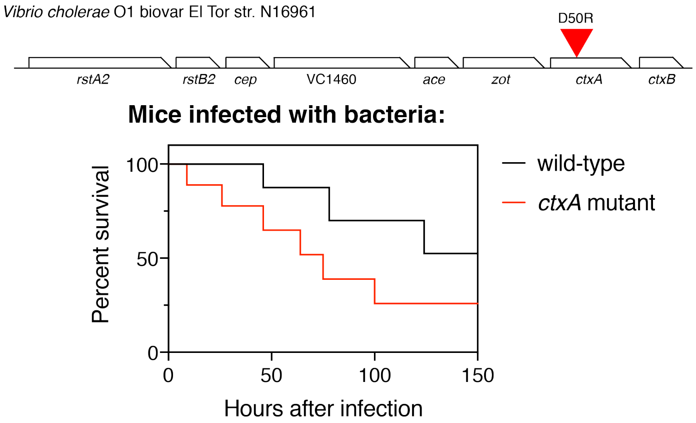
- incorporate all of the available data
- propose a mechanism that explains the behavior of the system
- make testable predictions about the system being studied

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

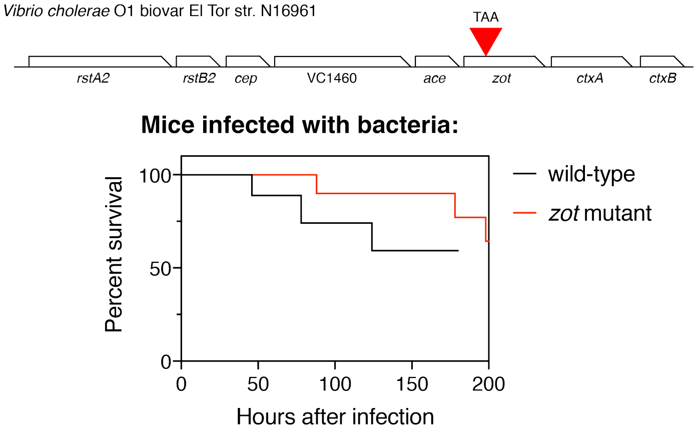
The following figure shows an operon from the cholera pathogen *Vibrio cholerae*. You have isolated a strain with the indicated *ctxA* missense mutation, and compared the survival of mice infected with this mutant and the wild-type strain.



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:



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 disease is caused by bacteria," is a valid hypothesis, since it can be disproved by observing even one case of disease which is not caused by bacteria.

When I ask you to propose a hypothesis:

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

These are obviously closely related concepts. 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.

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.

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 contains *gusA*) are white. You spread MG1655 on X-Gluc plates, and are able to isolate spontaneous blue colony mutants.

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

Problem #2

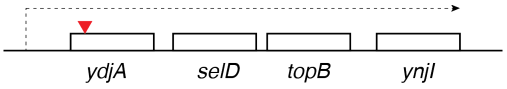
Sinorhizobium meliloti is a plant symbiont that forms nitrogen-fixing nodules on the roots of alfalfa plants. When studying its metabolism, you find the results in the table below. "*Minimal media*" are growth media that contain only the compounds that a given organism absolutely requires for growth. In this case, it indicates media with no amino acids added. MetE and MetH are two different *isozymes*, non-homologous enzymes that catalyze the same reaction, in this case, the last step of methionine synthesis.

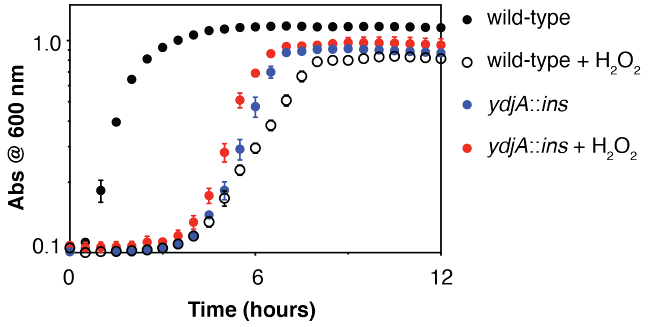
Strain	Ability to form nodules	Ability to grow on minimal media
wild-type	yes	yes
Δ*metE*	yes	yes
Δ*metH*	no	yes
Δ*metE* Δ*metH*	no	no

- 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 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 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 <u>good</u> 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 <u>one</u> 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 <u>change</u>, the dependent variable(s) is what you <u>measure</u>.

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

- <u>Will it answer the question?</u> 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?
- <u>Is it possible?</u> 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?
- <u>Is it elegant?</u> 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 the results you need.

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 expected to result in such a change. Negative controls are particularly useful for ensuring that no contamination or other problems are interfering with measurements to give *false positive results*. Positive controls demonstrate that the measurement system is capable of observing the expected changes, and rule out the possibility of *false negative results*.

When I ask you to design an experiment in 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

In the previous class, we discussed using mutants to understand biological phenomena. In this class, 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.

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 acridine orange causes frameshifts. 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. (From a practical standpoint, always be careful using mutagens in the lab. They will mutate your DNA just as efficiently as they mutate bacterial DNA!)

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 somewhat easier with the advent of relatively inexpensive genome sequencing technology.

A common and practical way to make random gene-inactivating null mutations is the use of *transposons* or *insertion elements*. Transposons are parasitic DNA fragments that are able to "hop" or insert themselves into a 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. Transposon mutants, like other insertions, nearly always destroy the function of the gene they integrate into, and are highly polar, which limits their usefulness for some kinds of mutagenesis experiments.

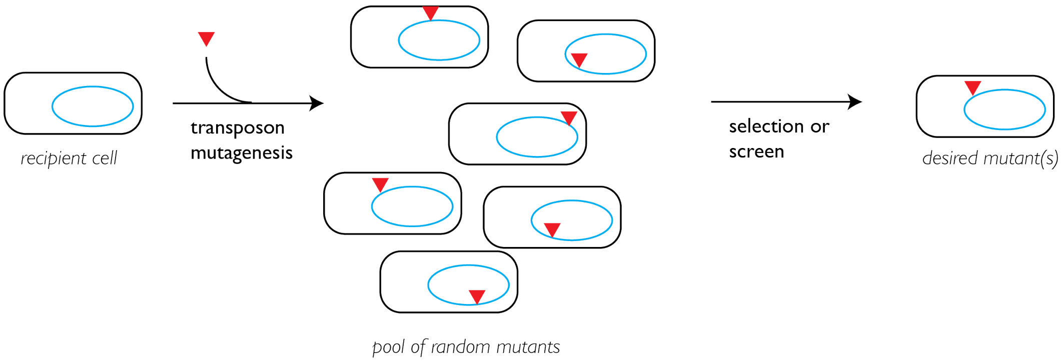


Figure 3.1: Transposon mutagenesis generates a library (or pool) of strains, each containing one randomly located transposon insertion. This library can then be screened or selected to identify insertions that cause phenotypes of interest.

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, and have been modified to carry antibiotic resistance genes. This allows you to treat a population of bacteria with the transposon and select for only those cells that have successfully integrated it into their chromosome on media containing the antibiotic. Each individual cell will have only one insertion, but since they occur at random positions, pooling together many cells can give a mixture 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 (see the next section for more details on screens and selections).

A transposon library should usually contain 20,000 or more independent mutants to ensure that at least one insertion is present in every non-essential gene. This is called a "saturated" library, and 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. Since transposons have a known sequence, it is relatively easy to identify exactly where in the chromosome it has integrated by PCR and Sanger sequencing.

Tn-seq (*transposon sequencing*, variations of which are also called INSeq, TraDIS, or HITS) is a relatively new and 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.

DISCUSSION PROBLEM SET #7: LIMITATIONS OF TRANSPOSONS

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 would be useful industrially. However, despite your best efforts (see next section) you are unable to isolate any such strains.

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

FINDING INTERESTING MUTANTS

All kinds of mutations occur spontaneously, but not every mutation is interesting. We use bacterial genetics to ask questions about specific phenomena, 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*.

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 <u>individually</u> 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.

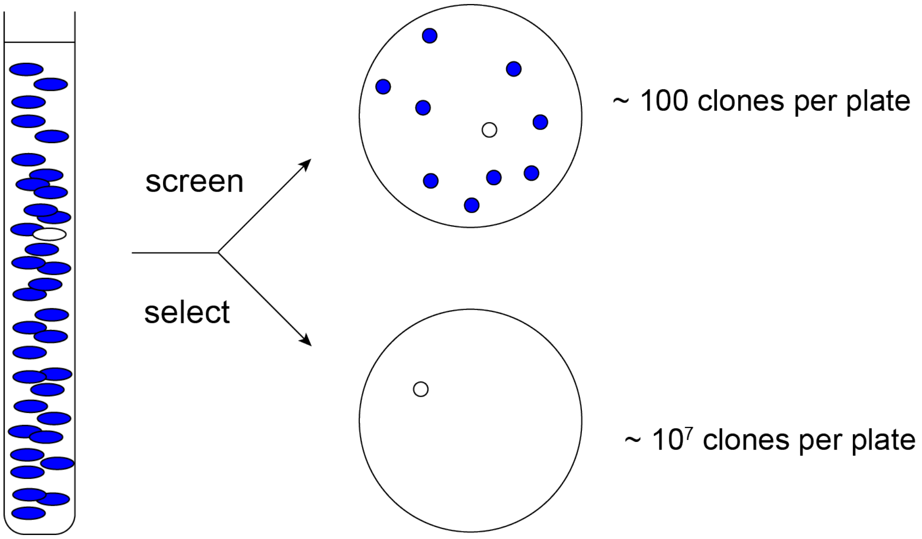


Figure 3.2: Selections allow you to identify much rarer mutants (white cells or colonies in this diagram) than is possible with a screen. A typical bacterial culture contains 10^8 – 10^9 cells per milliliter, and it is possible to spread about 100 microliters of culture on the surface of an agar plate.

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. If you have conditions where the wild-type grows, but the mutants you're interested in grow <u>faster</u>, then the mutants will slowly become a larger and larger proportion of the population, thereby "enriching" the population in interesting mutants. You typically follow up an enrichment (or several cycles of enrichment) with a screen to identify individual mutant strains. This can greatly reduce the number of colonies that you need to screen to find mutants of interest.

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.

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 look at an example experiment, in which we will use a screening approach to identify mutations in genes involved in sporulation in the opportunistic pathogen *Clostridium difficile*. Spores are easily visualized through the microscope:

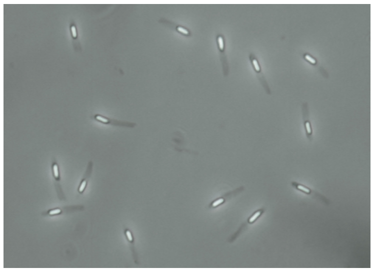


Figure 3.3: Microscopic image of bacterial spores (Wikipedia).

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile.*

Experimental Design:

1) Generate a saturating transposon library of *C. difficile* mutants.

2) Screen mutants microscopically for defects in spore morphology.

3) Sequence the genomes of mutants with spore defects to identify the sites of mutation.

Independent Variable: position of individual transposon insertions

Dependent Variables:

1) spore numbers (a quantitative measurement)

2) spore appearance (a qualitative measurement)

Negative Controls: (eliminate false positive results)

1) Confirm spore production and appearance in the wild-type

Positive Controls: (eliminate false negative results)

1) Confirm that the mutagenesis was successful (most transposons confer antibiotic resistance, which is easy to test for).

2) If possible, test a known spore-deficient strain.

Potential Outcomes:

1) Mutations that interfere with spore formation are isolated. This supports the original hypothesis, and will allow us to identify the genes involved.

2) No mutations that interfere with spore formation are isolated. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from not having screened enough mutants to find (by chance) any with visible differences in spore formation.

BRUTE FORCE AND ELEGANCE IN EXPERIMENTAL DESIGN

"Elegance" is an elusive and 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 microscopic examination of tens of thousands of mutants. Can we redesign our experiment to be more elegant?

Here's another possibility (with asterisks indicating steps that have changed):

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile.*

Experimental Design:

1) Generate a saturating transposon library of *C. difficile* mutants.

*2) Screen individual mutants for the ability to grow after ethanol treatment.

*3) Examine mutants that did not survive ethanol treatment microscopically for defects in spore morphology.

4) Sequence the genomes of mutants with spore defects to identify the sites of mutation.

Independent Variable: position of individual transposon insertions

Dependent Variables:

*1) growth after ethanol treatment (a qualitative measurement)

2) spore numbers of ethanol-sensitive strains (a quantitative measurement)

3) spore appearance of ethanol-sensitive strains (a qualitative measurement)

Negative Controls: (eliminate false positive results)

1) Confirm ethanol resistance, spore production, and appearance in the wild-type

*2) If possible, test a known spore-deficient strain.

Positive Controls: (eliminate false negative results)

1) Confirm that the mutagenesis was successful (most transposons confer antibiotic resistance, which is easy to test for).

*2) Confirm that the mutants do grow <u>before</u> ethanol treatment.

*3) Confirm that your ethanol treatment successfully kills vegetative cells of *C. difficile*.

Potential Outcomes:

1) Mutations that interfere with ethanol resistance and spore formation are isolated. This supports the original hypothesis, and will allow us to identify the genes involved. (It will also identify mutants that are ethanol-resistant, but <u>don't</u> have visible spore defects.)

2) No mutations that interfere with spore formation are isolated. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from not having screened enough mutants to find (by chance) any with visible differences in spore formation.

This is still a screen, but it is a much less labor-intensive one, since growth and ethanol treatment of bacteria can be carried out in 96- or 384-well plates and easily scored by measuring absorbance. Only mutants that have demonstrated defects in ethanol resistance will be subjected to time-consuming microscopic examination. It will, however, still be a lot of work screening tens of thousands of individual mutants.

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

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile.*

Experimental Design:

1) Generate a saturating transposon library of *C. difficile* mutants.

*2) Use Tn-seq to identify all of the insertions in the pooled library.

*3) Grow the entire pooled library, then add ethanol to kill vegetative cells.

*4) Regrow the survivors, and use Tn-seq to identify all of the insertions in those surviving cells.

Independent Variable: * ethanol treatment (before and after)

Dependent Variables: * the frequency of each transposon insertion in each pool (a quantitative measurement)

Negative Controls: (eliminate false positive results)

*1) "Before" data will not include insertions in any known essential genes.

*2) Confirm that your ethanol treatment has successfully killed all vegetative cells in your sample.

Positive Controls: (eliminate false negative results)

*1) Confirm that the mutagenesis was successful (the "before treatment" pool contains at least one transposon insertion in every non-essential gene).

Potential Outcomes:

1) You identify transposon insertions that are present in the "before treatment" pool and not present in the "after treatment" pool, and therefore prevent survival of ethanol treatment.

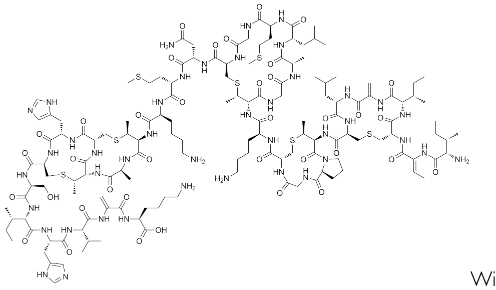
2) The distribution of transposon insertions is the same before and after ethanol treatment. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from your library being too small and not containing insertions in the genes required.

This experimental design will, in a single step, give you a list of genes in which transposon insertions result in sensitivity to ethanol, which you can plausibly hypothesize will be defective in their ability to form spores. It's still a screen, and you will need to do a secondary experiment to isolate individual mutants and examine them microscopically, but this design elegantly identifies all of the genes in *C. difficile* that are required for ethanol resistance.

DISCUSSION PROBLEM SET #8: SCREENS AND SELECTIONS

<u>Problem #1</u>

Nisin is an antimicrobial bacteriocin produced by some strains of *Lactococcus lactis* which efficiently inhibits the growth of a wide range of Gram-negative and Gram-positive bacteria. It is commonly used to prevent bacterial growth 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 *Listeria monocytogenes.*

1) Design a mutant hunt to identify genes involved in nisin resistance using a <u>screen</u>.

2) Design a mutant hunt to identify genes involved in nisin resistance using a <u>selection</u>.

For each experimental design, state:

- your hypothesis
- the method of mutagenesis you will use (and why)
- the independent and dependent variables (*what will you change, and what will you measure?*)
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

<u>Problem #2</u>

Salmonella enterica can grow on ethanolamine, a common carbon and nitrogen source present in the mammalian gut, especially during inflammation. You hypothesize that *S. enterica* has genes encoding a pathway specifically required for growth on ethanolamine.

Design a mutant hunt that would allow you to identify any such genes, and 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
- potential outcomes of your experiment, and how you will interpret them

<u>Problem #3</u>

Anaplasma phagocytophilum is an emerging tick-borne pathogen that survives and replicates in human neutrophils, causing a severe, life-threatening fever.

Design a mutant hunt that would allow you to identify genes required for virulence in *A. phagocytophilum*, and 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
- potential outcomes of your experiment, and how you will interpret them

REVERTANTS AND SUPPRESSORS

If a mutation causes a slow growth phenotype, you will sometimes observe *secondary mutations* in that strain that restore wild-type-like growth. These arise because in the process of observing the poor growth phenotype, you are also selecting for any mutant that <u>is</u> able to grow well under those conditions. Such a mutant is called a *revertant*, because the phenotype has reverted back to "wild-type". In some cases, this may actually be due to a mutation that directly changes the original mutated gene back to the wild-type sequence. This is far more likely with point mutations than with other kinds of mutations. This kind of revertant and other mutations in the same gene as the original mutation are referred to as *intragenic suppressors*.

However, it is often more interesting to identify *intergenic suppressor* mutations, which are mutations in <u>other</u> 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 restores growth, you have very strong evidence that those two genes are involved in the same biological process.

Multicopy suppressors are genes that, when present in more copies than in the wild-type (see **Lecture 6** on plasmids), suppress the phenotype of a mutation in a different gene.

It is worth noting that revertants may have a wild-type phenotype, but nearly always have a mutant genotype. 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 heat shock regulator *rpoH* 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 unregulated increase in protein-stabilizing chaperones, but if you didn't know that, you might make the wrong conclusions about the function of *rpoH*.

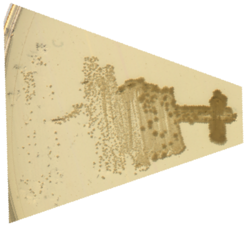


Figure 3.3: The appearance of spontaneous revertants. Note how faster-growing colonies containing suppressor mutations are arising out of a streak of slower-growing parent cells (which are themselves mutants that do not grow especially well under these conditions).

In a related phenomenon, there are also mutations that have no phenotype on their own, but have measurable 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.

DISCUSSION PROBLEM SET #9: SUPPRESSORS AND REVERTANTS

Exiguobacterium acetylicum is a commensal bacterium that colonizes the gut of some fish species, influencing their health. It is difficult to manipulate genetically, and you are unable to generate transposon insertion mutants in this species, but you are able to use chemical mutagenesis and an enrichment and screening process to identify mutants with reduced motility.

Several of the non-motile mutants you isolate have mutations in the gene of unknown function *ea2862*. You think that identifying suppressors of these mutations might help you figure out the function of *ea2862*.

Design a mutant hunt that would allow you to identify *ea2862* suppressors, and 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
- potential outcomes of your experiment, and how you will interpret them

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. 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.

Each approach to a problem has different advantages and disadvantages and gives you different kinds of results, so combining multiple approaches is the most rigorous way to test a hypothesis. When designing experiments for this class, I will often ask you to describe more than one distinct experiment to answer any given question. 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"?

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. 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.

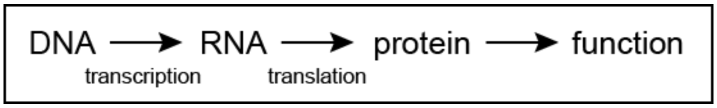
LECTURE 4: PRINCIPLES OF REGULATION

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 flow of information from DNA to protein function (the "Central Dogma"), 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 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.

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.

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 often called *derepression*.

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* in the mutant.

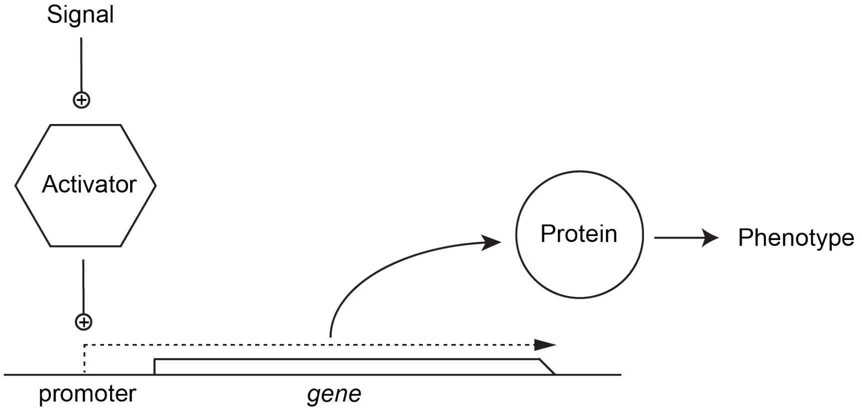


Figure 4.1. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional activator.

The following table illustrates how null mutations in different components of the circuit shown in Figure 6.1 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	+
activator-	absent	-
activator-	present	-
gene-	absent	-
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 expression 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, which can have very different effects depending on the genes in question.

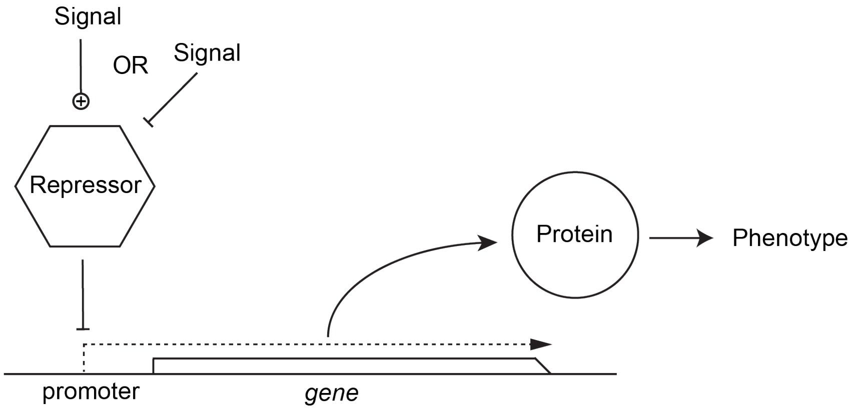


Figure 4.2. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional repressor.

The following table illustrates how null mutations in different components of the circuit shown in Figure 6.2 would be expected to change the observable phenotype if the repressor responds to a signal that activates its repressive functions:

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

If the repressor responds to a signal that inactivates 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	+
repressor-	absent	+
repressor-	present	+
gene-	absent	-
gene-	present	-

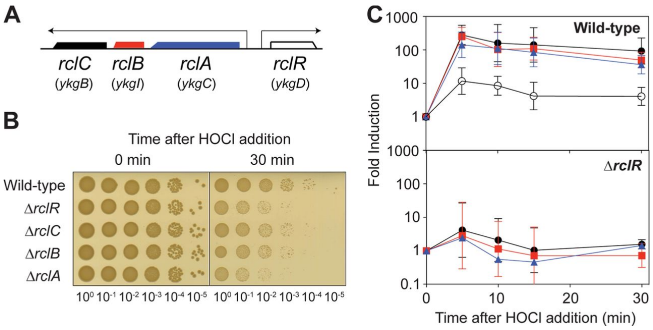
Operator sequences are sites in DNA (usually within the promoters of genes) where proteins or other regulators bind to control gene expression. Mutations in operators can change how well those regulators bind, leading to a variety of phenotypes depending on the nature of the operator mutation and the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

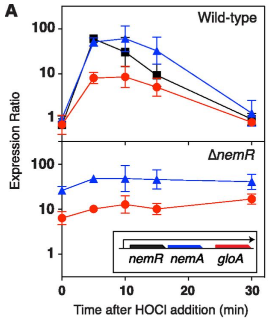
Global regulators control many genes or gene products throughout a cell's genome. They may be positive regulators of some of those genes and negative regulators of others, and mutations of global regulators often have very complex pleiotropic phenotypes. *Local regulators* control only a small set of 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.

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. This is particularly true for genes with multiple regulators (which is probably most of them, in bacteria).

DISCUSSION PROBLEM SET #10: MUTATIONS IN REGULATORS

Transcription of the *nemR-nemA-gloA* and *rclABC* operons of *E. coli* are induced by exposure to hypochlorous acid (HOCl), and are both required for *E. coli* to efficiently survive exposure to HOCl. Based on sequence homology, NemR and RclR are probably DNA-binding proteins. (Note that the gene expression data is color-coded by gene for each operon individually in these figures.)





Is RclR a positive or a negative regulator? What about NemR? Why?

Would you expect a Δ*nemR* mutant to be more sensitive to killing by HOCl? Why or why not?

SURVEY OF REGULATORY MECHANISMS

In the next part of this chapter, I will briefly describe some of the different types of regulation that are known 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. My goal here is to give you a broad sense of how complex regulation can be. 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 both these factors and by the *stability* of that mRNA.

Regulating transcription initiation is 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. 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.

1. Transcription initiation. Initiation of transcription takes place at a *promoter* where RNA polymerase binds to the DNA. 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}, 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.

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 involved in the construction of complex molecular machines (*e.g.* flagella). They recognize consensus sequences different from those found in promoters transcribed by RNA polymerase containing the housekeeping sigma. The concentration and activity of alternative sigma factors are tightly controlled, often using multiple mechanisms of transcriptional and post-transcriptional regulation. Different bacterial species may contain anywhere from one to dozens of sigma factors, depending on the complexity of their environment and developmental pathways.

Other features of promoters can also influence the efficiency of transcription initiation. *UP elements* are AT-rich 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) 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

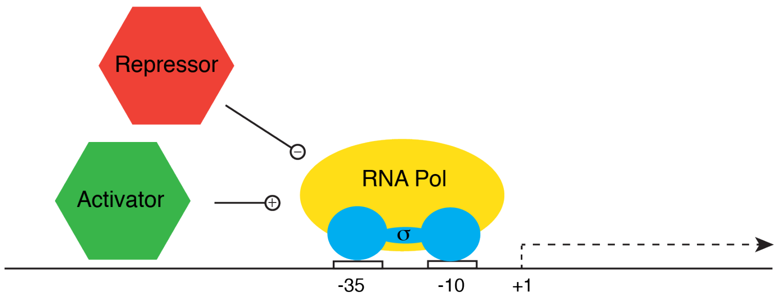


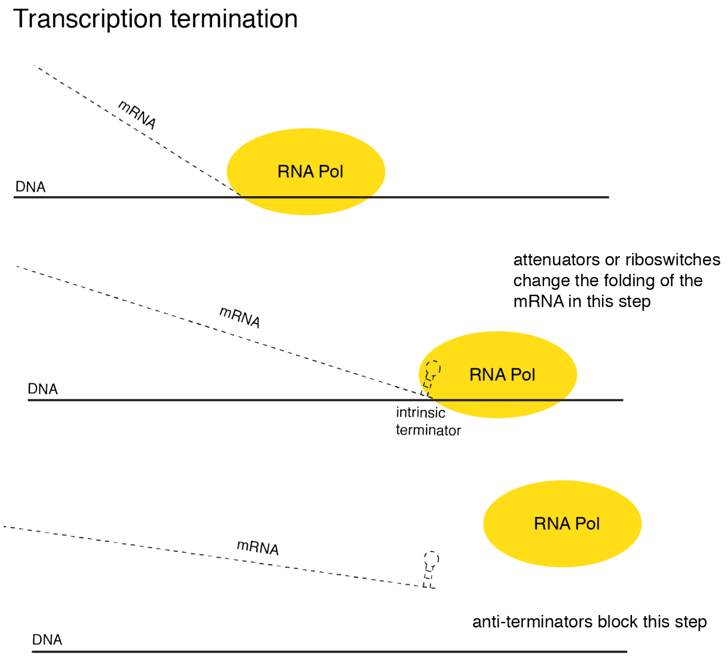
Figure 4.3. Regulators that can affect transcription initiation.

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 occluding 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).

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 (*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 recognizes single-stranded RNA with no ribosomes attached and chases RNA polymerase to terminate transcription. (As mentioned in **Lecture 2**, this is why nonsense mutations are polar: they result in long stretches of untranslated RNA in mRNAs.) *Rho-independent termination* also occurs (for about half of transcripts in *E. coli*). 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 several uracil residues.


Transcription termination

attenuators or riboswitches change the folding of the mRNA in this step

anti-terminators block this step

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

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. Attenuation is a mechanism by 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 biosynthesis operon (briefly mentioned in the reading for **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. *Riboswitches* are RNA structures, usually found in the 5' *untranslated region* (UTR) of an mRNA, which bind specific metabolites (e.g. amino acids or vitamins) and form structures that can include terminators or anti-terminator loops.

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. 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 (specifically, RNAse III). *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 regulate more than one mRNA.

They nearly always require the RNA-binding protein Hfq for activity. An *hfq* mutant is therefore defective in all sRNA-mediated regulation, which can be useful for determining whether sRNAs are involved in a regulatory phenotype.

sRNA regulation of mRNA degradation

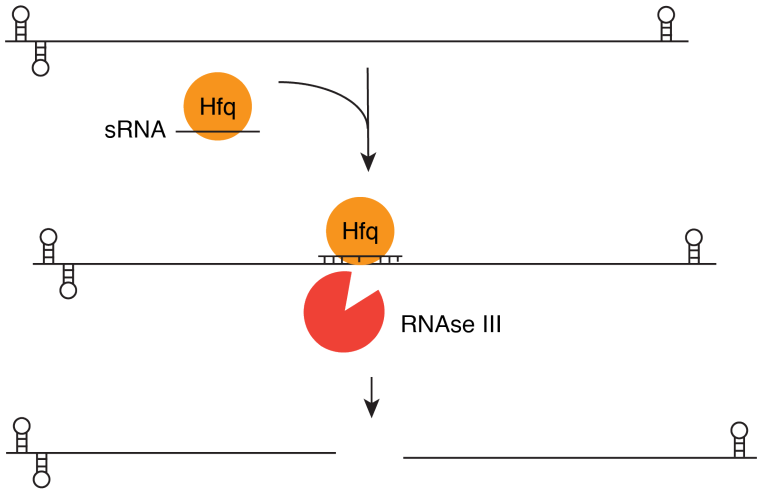


Figure 4.5. sRNA and Hfq-mediated mRNA degradation, by forming a double stranded RNA targeted by RNAse III.

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 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 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, but can rapidly become prohibitively expensive if a lot of different samples need to be analyzed.

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 occasionally, chromosomal insertions) in which the promoters of genes of interest are cloned 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 luciferase, which produces light.

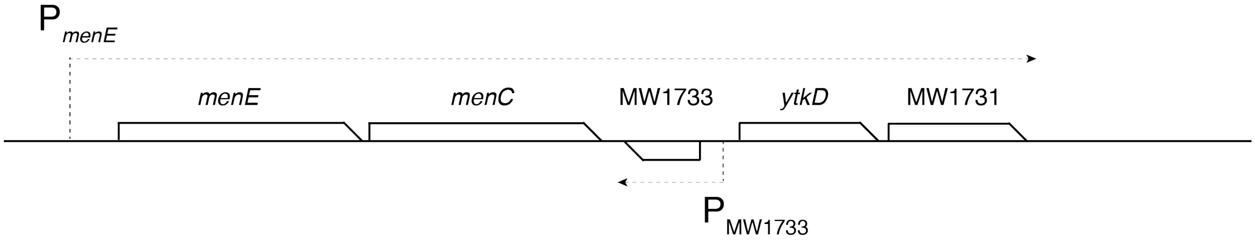
The advantages of using transcriptional reporters are that they tend to be the simplest, cheapest way to measure expression from a given promoter. There are several disadvantages, though. First is that they are intrinsically non-physiological, since they are cloned promoters driving non-physiological products from multi-copy plasmids (see **Lecture 6** for more on plasmids). Secondly, high production of reporter gene products may be toxic (fluorescent proteins) or require large amounts of energy (luciferase). Thirdly, cellular growth conditions can affect reporters in ways that they would not affect the actual gene product. Both GFP and luciferase require aerobic conditions, 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.

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 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.

DISCUSSION PROBLEM SET #11: TRANSCRIPTIONAL REGULATION

Problem #1

An unusual locus called a "non-contiguous operon" has recently been identified 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 MW1733 mRNA is expressed, the amount of *menE-menC-ytkD*-MW1731 mRNA detectable in the cell 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 model and testable hypothesis to explain the regulation of the *menE-menC-ytkD*-MW1731 operon by MW1733.

2) Propose a genetic experiment that will test your hypothesis. All standard genetic tools are able 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

Problem #2

While studying the actinomycete *Streptomyces griseus*, you identify a mutant that does not produce spores or antibiotics. Sequencing reveals that the mutation is a premature stop codon in a sigma factor homolog.

GmbH

Propose a series of experiments to determine which gene or genes in *S. griseus* are regulated by this sigma factor and which ones are required for spore and/or antibiotic production. For each experiment, 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 for mRNA. Regulation of translation allows the cell to maintain a pool of mRNA that it does not need to transcribe before producing protein, 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 ACCUCCUUA, and therefore the consensus sequence for S.D. sites in *E. coli* is AGGAGGU, which base pairs with the underlined 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 genes encoded by the same RNA can be translated at different rates.

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 but unfolds at higher temperatures (found, for example in the virulence-associated *prfA* transcript from *Listeria monocytogenes*).

Translation initiation

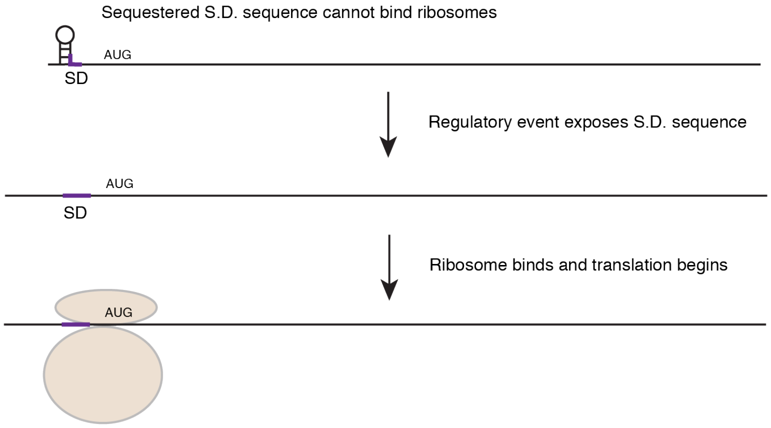

Sequestered S.D. sequence cannot bind ribosomes
AUG
SD
Regulatory event exposes S.D. sequence
AUG
SD
Ribosome binds and translation begins
AUG

Figure 4.6. 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.

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 well translated (and is likely to be prone to termination or degradation, as described above). There is also some 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, inaccessible to 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" found in mycobacteria. Each protease has different specific substrates, although they often overlap extensively. Proteases recognize specific signal sequences (*degrons*) in their target proteins, and the presence 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.

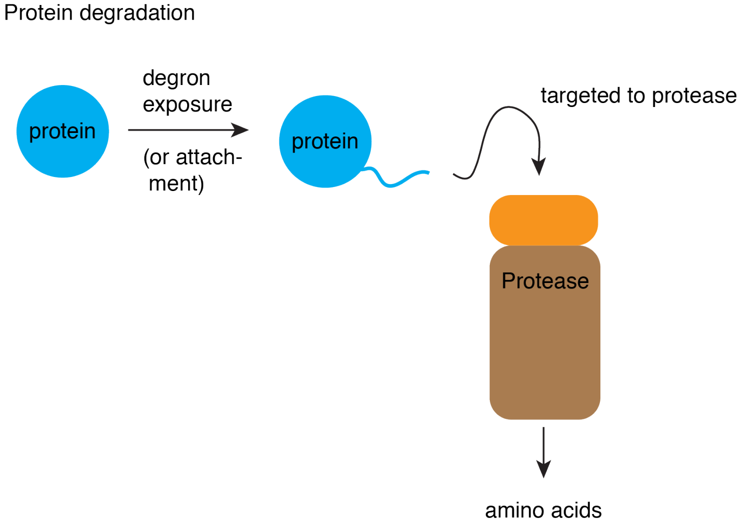


Figure 4.7. Targeting and degradation of proteins by protease complexes.

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.

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	Genome-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. To quantify a protein by immunoblotting, 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 very challenging.

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, and the FLAG-tag, an entirely artificial antigen with the amino acid sequence DYKDDDDK. Adding such a 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 easier 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 **Lecture 6** for more on plasmids).

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 detect low-abundance proteins.

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

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, with the additional complication that any translational signals (pause sites, rare codons, *etc.*) found within the coding sequence of the gene will not be present in the fusion.

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. This is a powerful technique that has a lot of potential uses, but is currently expensive and requires considerable technical expertise.

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 approach 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.

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 much more difficult to measure 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, it can be quite wasteful, since producing inactive proteins requires 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.

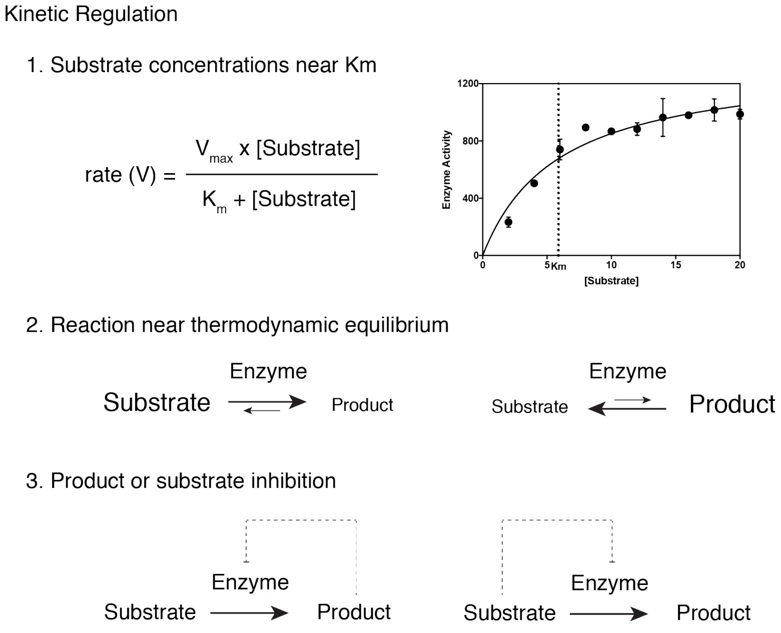

Kinetic Regulation
1. Substrate concentrations near Km
rate (V) = Vmax x [Substrate] / Km + [Substrate]
Enzyme Activity
1200
800
400
0
0
5
Km
10
15
20
[Substrate]
2. Reaction near thermodynamic equilibrium
Enzyme
Substrate
Product
Enzyme
Substrate
Product
3. Product or substrate inhibition
Enzyme
Substrate
Product
Enzyme
Substrate
Product

Figure 4.8. A variety of ways enzymes can be regulated by their biochemical properties.

The nature of enzyme kinetics means that the activity of many 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.

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 means of competition for binding in the active site of the enzyme.

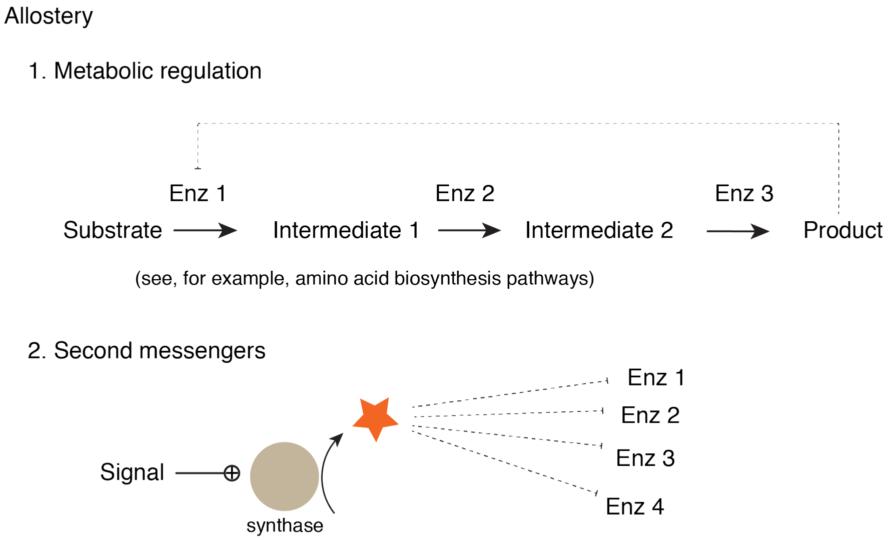


Figure 4.9. Examples of allosteric regulation of protein activity by small molecules.

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.

There are many examples of allosteric regulation by *second messengers*, which are small molecules produced under certain conditions which affect the activity of proteins throughout the cell. Many of these 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.

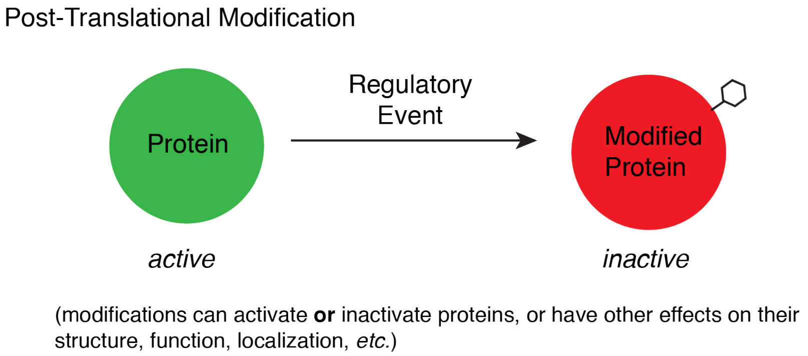

Post-Translational Modification
Protein
Regulatory Event
Modified Protein
active
inactive
(modifications can activate **or** inactivate proteins, or have other effects on their structure, function, localization, *etc.*)

Figure 4.10. Protein activity can be changed by covalent modification of the protein. These modifications are typically the result of the activity of other enzymes, which are typically regulated in turn by one or more of the mechanisms discussed in this chapter.

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

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.

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. 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. 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 sulfinic ($-SO_2H$) or sulfonic ($-SO_3H$) acid, and the *Bacillus subtilis* transcription factor PerR responds to peroxide stress via the irreversible oxidation of a histidine residue. Presumably, the resulting 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.

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 very often differ for every 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 allow much simpler normalization for protein abundance.

Enzyme activity assays are the most direct way to assess whether an enzyme 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 PhoA measurements. 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 the pathway, making it impossible to measure the synthesis of product *in vivo*. If the product of an enzyme is chemically unstable, it will also be difficult to measure *in vitro*. We recently encountered this problem in my lab with an enzyme that catalyzes the reduction of Cu^{2+} to Cu^{1+}. Cu^{1+} is oxidized back to Cu^{2+} by atmospheric O_2, which made the rate of Cu^{1+} accumulation very difficult to track.

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. *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.

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 overexpressed without toxicity? Is the protein soluble? Membrane proteins are never soluble, so there are different detergents 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? If not, how can you separate the protein from other cellular proteins without a tag?

Allosteric and kinetic regulation is typically easiest to measure for proteins that can be purified and assayed *in vitro*.

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 easily 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* (<u>ch</u>romatin <u>i</u>mmuno<u>p</u>recipitation <u>seq</u>uencing) 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.

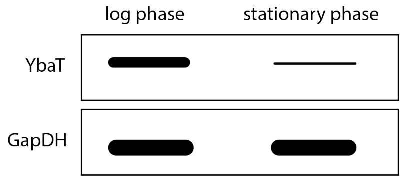
Protein-protein interactions can be measured both *in vivo* and *in vitro*, although *in vitro* techniques with purified proteins are much more likely to give quantitative measurements of binding affinity. *Two-hybrid assays* are clever *in vivo* screens (or sometimes selections) that link protein-protein 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.

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. As protein levels and activity change, this flux will shift, reflecting changes in cellular metabolism.

DISCUSSION PROBLEM SET #12: POST-TRANSCRIPTIONAL REGULATION

<u>Problem #1</u>

While studying the metabolism of bacteria in the stationary phase of growth, you identify a protein (YbaT) that is much more abundant in stationary phase than during logarithmic growth. ("GapDH" is glyceraldehyde dehydrogenase, a constitutively expressed glycolytic enzyme.)



Mutants lacking *ybaT* die after 4 days of incubation at 37°C, while > 90% of wild-type cells survive.

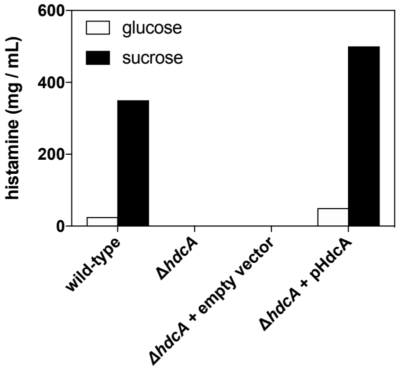
1) What measurements could you make to determine whether this regulation is transcriptional or post-transcriptional?

2) If those measurements indicate that YbaT is post-transcriptionally regulated, propose a genetic experiment to identify factors required for YbaT regulation. Sate:

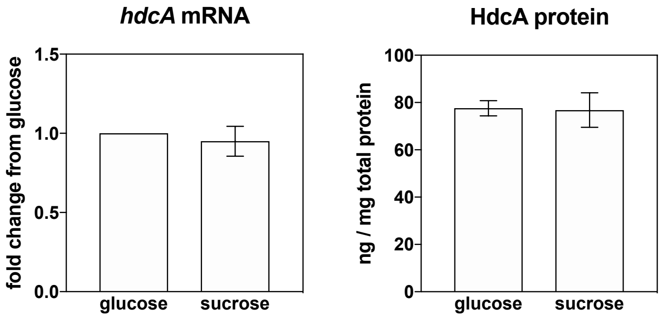
   - the method of mutagenesis you will use (and why)
   - are you using a screen, a selection, or an enrichment to identify relevant mutants?
   - the independent and dependent variables
   - both positive and negative controls
   - potential outcomes of your experiment, and how you will interpret them

Problem #2

You isolate a *Lactobacillus* strain from the microbiome of a mouse and discover that it synthesizes the anti-inflammatory compound histamine when grown in media containing sucrose, but not in media containing glucose. This is dependent on the presence of the *hdcA* gene, which encodes histidine decarboxylase, an enzyme that converts the amino acid histidine to histamine. (Note that in the figure below, "pHdcA" indicates a plasmid or "vector" encoding the *hdcA* gene. We will discuss plasmids in more detail in **Lecture 6**.)



You would like to understand how histamine synthesis is regulated. You do qRT-PCR to measure *hdcA* mRNA levels and quantitative Western blots to measure HdcA protein levels, with the following results:



- Based on these data and your knowledge of regulation, propose two possible models that might explain the regulation of histamine synthesis in response to sucrose.

- Describe an experiment or series of experiments that would allow you to distinguish between the two hypotheses you described above. 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 systems 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: 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 in the previous 4 days 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 <u>question</u> 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

Ding *et al.* (2019) "Induction of *Rhodobacter capsulatus* Gene Transfer Agent Gene Expression Is a Bistable Stochastic Process Repressed by an Extracellular Calcium-Binding RTX Protein Homologue." J Bacteriol 201(23):e00430-19.

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, although since this is a recent paper in a non-open access journal, you will probably need to be logged into a UAB network to access the full-length document.

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 6: PLASMIDS

INTRODUCTION

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

SCIENTIFIC PROCESS 5: CORRELATION AND CAUSATION

It is extremely important to distinguish 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.

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 class. Falkow's Postulates (from Falkow 1988 Rev Infect Dis 10 supp 2: S274-6) 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 #13: CORRELATION AND CAUSATION

Problem #1

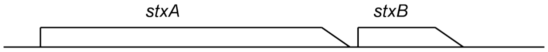
People with inflammatory diseases of the gut have different proportions of bacteria in their gut microbiomes than do healthy people. This typically includes higher populations of *E. coli* and lower populations of *Faecalibacterium* species.

Propose an experiment to determine whether inflammation causes changes in bacterial populations 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

Problem #2

The lysogenic ϕStx phage of *Shigella dysenteriae* carries the *stxAB* operon:



Nonsense mutations isolated in either *stxA* or *stxB* prevent *S. dysenteriae* from causing disease, but expressing a wild-type copy of *stxA* in an *stxA* mutant does not restore virulence.

Is *stxA* a virulence factor?

Propose a model to explain these results, and an experiment (based on Falkow's postulates) to test your model. 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. **Chapter 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 7**.)

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 7**), 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. They 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 the "plasmid", and it has a copy number of one, it is likely to be considered a chromosome. 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 possible to select for their presence in transformed cells. See below for a reasonably comprehensive list of plasmid features and 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.

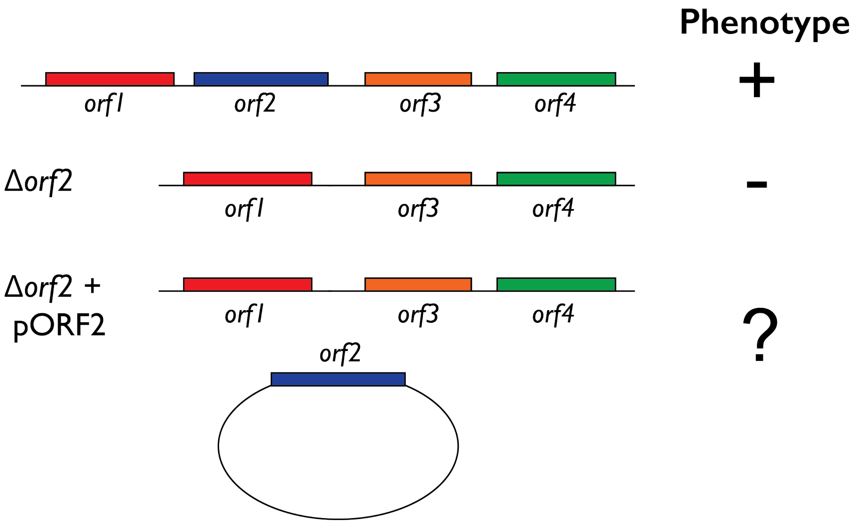


Figure 6.1. Illustration of complementation analysis testing whether deletion of *orf2* is responsible for the "-" phenotype.

As a real-world example (drawn from my dissertation research), lets consider the *bluB* gene of the photosynthetic bacterium *Rhodospirillum rubrum*. I constructed a Δ*bluB* mutant and observed that it grew poorly in the absence of the vitamin B_{12} precursor dimethylbenzimidazole (DMB). 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 Δ*bluB* mutant (site-directed mutagenesis of *R. ruburm* is not trivial!), so there was certainly a chance that other mutations would have arisen.

The following figure (from Gray & Escalante 2007 PNAS 104:2921), shows how I was able to demonstrate this using a plasmid encoding the *bluB* gene (here indicated as "p*bluB*$^+$"):

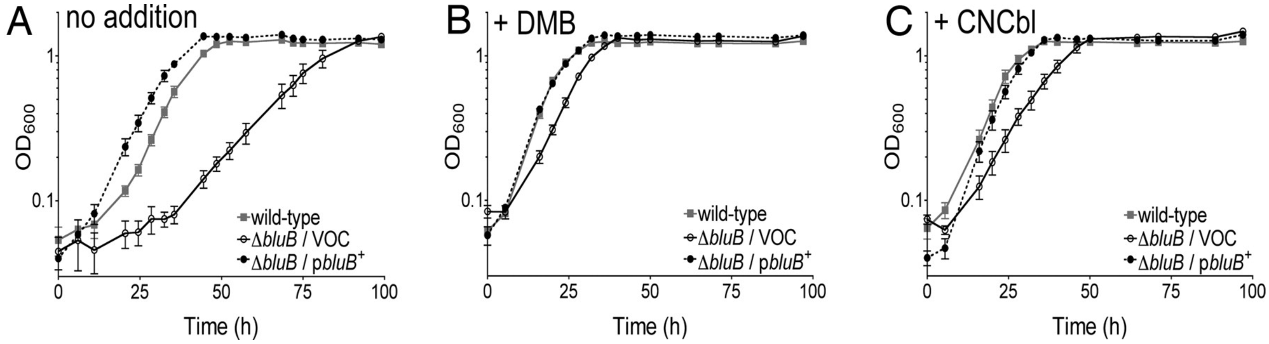


Figure 6.2. BluB is necessary for DMB synthesis. *R. rubrum* wild-type and Δ*bluB* mutant cultures containing the indicated plasmids were grown photosynthetically in the presence of DMB or cyanocobalamin (CNCbl), as indicated.

"CNCbl" is cyanocobalamin, the chemical name for vitamin B_{12}. Observe that the wild-type (grey lines) grows well under all three conditions, but the Δ*bluB* mutant containing an empty plasmid (*vector-only control* or *VOC*) has a significant growth defect in the absence of DMB or B_{12}. Critically, complementing the mutant with the p*bluB*$^+$ plasmid (dotted lines) restored growth in the absence of DMB or B_{12} (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.

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 by far 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 p*bluB*$^+$ 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. Its expression can be tightly controlled, depending on the promoter present in the plasmid (see **Lecture 4** for more about promoters and gene expression), so you can tune the amount of its encoded RNA or protein that is produced. Protein products can be easily fused to GFP or other proteins for detection or purification. Plasmids can easily be mutated, either *in vivo* or *in vitro*, to rapidly test the effect of specific mutations on gene activity. Randomly mutating a plasmid (for example, by propagating it in a *mutator strain* that lacks DNA repair genes) allows *localized mutagenesis* of a single gene, rather than the entire genome. *Site-directed mutagenesis* is much easier on a plasmid than in the chromosome, and allows very precise experimental designs.

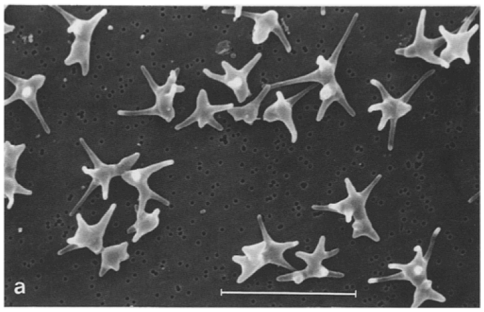
Another common and very useful technique is the construction of *plasmid libraries*, which are pools of plasmids containing many different cloned inserts. A *genomic library* contains random fragments of the entire genome of an organism, and a *cDNA library* contains DNA reverse transcribed from an mRNA 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-factor 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.

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 Plasmid Guide is particularly valuable.

DISCUSSION PROBLEM SET #14: USING PLASMIDS IN GENETIC EXPERIMENTS

Problem #1

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



MicroBestiary

A. adetum is a Gram-negative alpha proteobacterium, and you are able to generate random mutant libraries and introduce expression plasmids into this species.

Design a genetic experiment to determine what genes are required for cell shape determination in *A. adetum*.

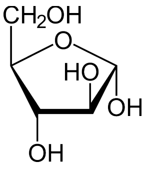
State:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?
- the independent and dependent variables

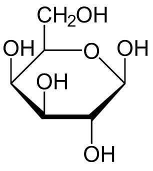
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

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.



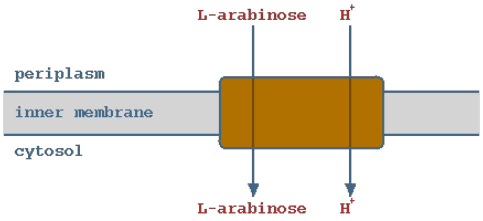
arabinose



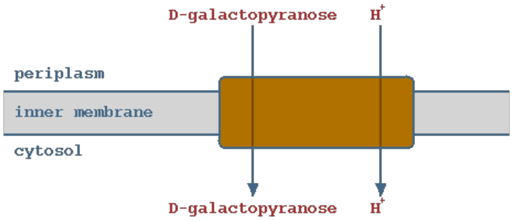
galactose

Surprisingly, the transporters for importing arabinose and galactose (AraE and GalP, respectively) in *E. coli* are 65% identical to each other at the amino acid level. AraE cannot transport galactose and GalP cannot transport arabinose.

AraE



GalP



Design a genetic experiment using plasmids to identify amino acids involved in substrate specificity in AraE and/or GalP. (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
- potential outcomes of your experiment, and how you will interpret them

Problem #3

Soil bacteria are suspected to be the source of most antibiotic resistance genes, but >99% of them cannot be cultured in the laboratory. Metagenomic sequencing of DNA extracted from soil can identify the presence of known antibiotic resistance genes in a sample, but cannot identify new genes.

Smithsonian Magazine

Design an experiment using metagenomic plasmid libraries to identify genes from uncultured soil organisms that encode antibiotic resistance. State:

- the independent and dependent variables
- both positive and negative controls
- 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 "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 Bolivar and Rodriguez, who were postdocs in Herbert Boyer's lab. However, this is not at all a universal rule. The "UC" in pUC18 stands for "University of California", the "ET" in pET-21b (one of a very large family of "pET vectors") stands for "expression by T7 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.

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:

pMGSA1
pMGSA2

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 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.

origin of replication (*ori* or *oriC*): Every plasmid will have 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) 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, and sometimes have separate origins of replication for each species. 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*, which will allow replication at a *permissive temperature* (often 30°C) but not at a *restrictive temperature* (often 42°C).

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. Two plasmids with the same origin will be *incompatible*, so origins of replication are sometimes called *compatibility groups*. A single plasmid cannot contain two origins of replication that function in the same organism.

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, 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

(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 bacteria.

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 depends on the inhibition of Tc^{R} cells by fusaric acid. The following paper by Maloy and Nunn (1981) J Bacteriol 145(2):1110-2 expands on this method and describes media for using it in *E. coli*: http://jb.asm.org/content/145/2/1110.full.pdf, should that ever happen to be useful to you.

Multiple Cloning Site (MCS): A small region of the plasmid with several closely spaced restriction sites to allow simplified insertion of cloned genes (see **Lecture 7**).

promoter: A DNA sequence which allows expression of genes on the plasmid. While every gene on the plasmid must have a promoter to be expressed, in most cloning vectors there is a separate 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 have their level of expression tuned 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-thiogalactopyranoside) and promoters controlled by other sugars, like arabinose or xylose. Many *overexpression vectors* used to produce 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. Some plasmids include an RBS, and some do not. In the latter case, you must include an RBS in gene sequences you clone in order for them to be expressed.

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 6xHistidine or GST tags), changing the physical properties of the protein (fusion with maltose binding protein increases the solubility of a protein, and fusion with a *signal sequence* can target a protein for secretion out of the cell), 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 the FLAG tag, which is 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 you do not include a stop codon in between your cloned sequence and the fusion protein (if it is a C-terminal protein fusion).

origin of transfer (*oriT*): A DNA sequence allowing the plasmid to be *mobilized* 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.

f1 origin: Many older plasmids will contain an origin of replication derived from the filamentous phage f1, 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 f1. This was useful when DNA sequencing technologies required large amounts of single-stranded DNA, but is now largely obsolete.

***cos* sites**: Like the f1 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.

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, but the map is likely to be more useful for most purposes. The following figure illustrates what some plasmid maps might look like:

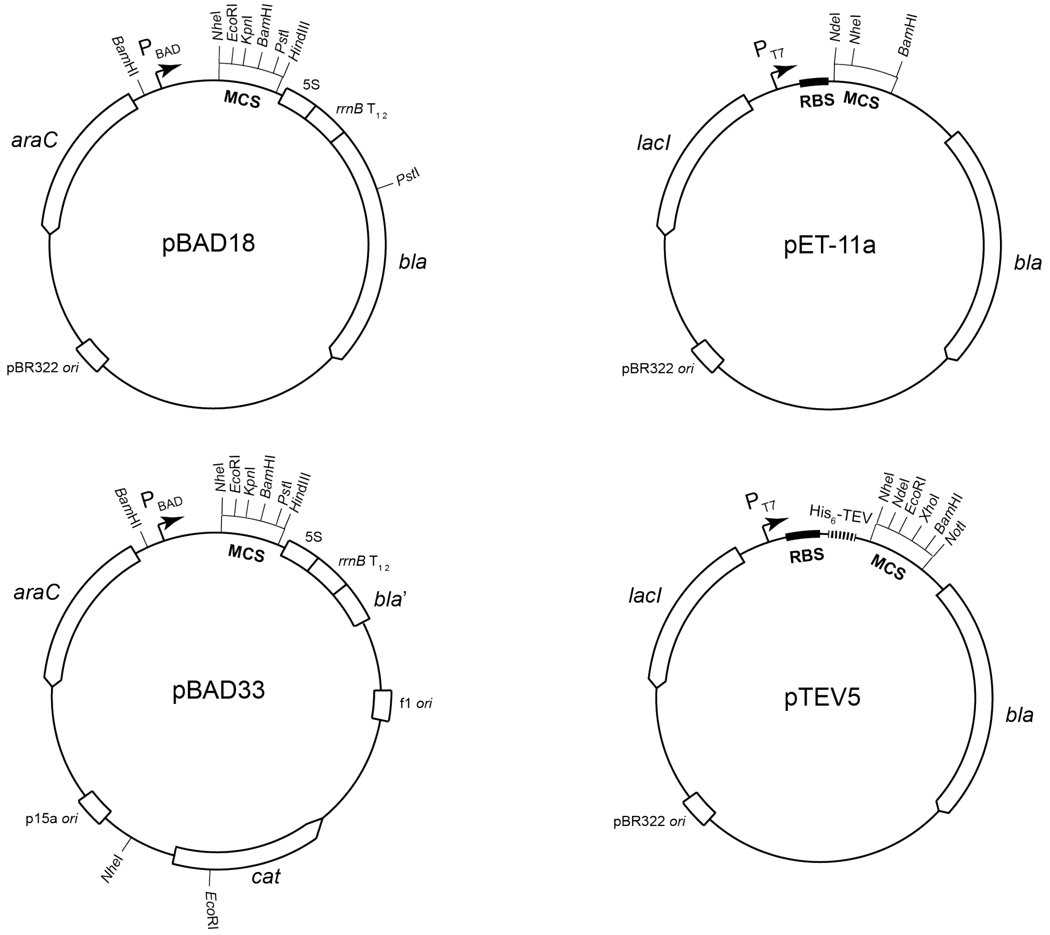


Figure 6.3. Sample plasmid maps.

- pBAD18 and pBAD33, from Guzman *et al.* (1995) J Bacteriol 177(14):4121, 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 regulator of 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 *rrnB* T$_{1\ 2}$" transcriptional terminator immediately after the MCS, to prevent read-through transcription.

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 a protein that protects the cell against chloramphenicol, 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 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 the 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 overexpression strain BL21[DE3]).

- pTEV5, from Rocco *et al.* (2008) Plasmid 59(3): 231-239, 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.

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 **Chapter 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 an online tool that searches DNA sequences for restriction sites.)

LECTURE 7: PRINCIPLES OF GENETIC ENGINEERING

INTRODUCTION

Constructing a DNA molecule with sequences from two or more different organisms is called *recombinant DNA technology*, 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 key tool for molecular genetics of bacteria. This chapter is a discussion of the principles and techniques used to engineer plasmids.

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 on 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 and will 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 almost all DNA molecular construction is done in *E. coli*, where the most highly developed tools are available. The resulting DNA products can then be transferred to other species of interest.

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**).

	In vitro	*In vivo*
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 relatively inexpensive to sequence DNA up to and beyond the length of an organism's genome.

For routine sequencing of short sections of DNA molecules (< 1000 bp), *Sanger sequencing* is the most common and cheapest method. A variety of so-called "*next-generation sequencing*" (NGS) and emerging "third generation" sequencing technologies exist that allow us to sequence whole genomes and complex mixtures of DNA from many organisms (*metagenomes*), usually by computationally aligning millions of very short sequence 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.

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*.

Write

It is possible to chemically synthesize DNA molecules with a desired sequence *in vitro*, but current methods typically only allow accurate synthesis of DNA chains up to about 100 nucleotides long. These are called *oligonucleotides* (or "oligos"), and are relatively inexpensive (as little as 20 cents per nucleotide, if ordered in bulk).

Since even a single gene is usually hundreds or thousands of nucleotides long, it is not typically practical to make large and complex DNA molecules from scratch (although see "Gene Synthesis" below, for a common, commercially available, and not too exorbitantly expensive way to do so, when needed). There is no equivalent technique for generating entirely new DNA sequences from scratch *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.

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 for us. 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*. 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 and multiple cycles of annealing and extension are repeated, the result is exponential copying of the sequence between the two primers. 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 sequences, but some protocols use nucleases with less specificity for particular purposes.

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. Purified restriction enzymes are used in many protocols to cut DNA molecules into defined fragments. The names of restriction enzymes are based on the species they were originally isolated from. *Eco*RI and *Eco*RV are the first and fifth restriction enzymes isolated from *E. coli* strain R, for example, and *Hind*III 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.

DNA fragment 1

CATATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAA**GGATCC**

GTATACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATT**CCTAGG**

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

```
...GAAGGAGATATACATATGGCTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCGGCTGCTAACAAA...
...CTTCCTCTATATGTATACCGATCGTACTGACCACCTGTCGTTTACCCAGCGCCTAGGCCGACGATTGTTT...
```

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

```
TATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAAG
  ACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATTCCTAG
```

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

```
...GAAGGAGATATACA     TATGGCTAGCATGACTGGTGGACAGCAAATGGGTCGCG      GATCCGGCTGCTAACAAA...
...CTTCCTCTATATGTAT     ACCGATCGTACTGACCACCTGTCGTTTACCCAGCGCCTAG      GCCGACGATTGTTT...
```

The most recent major addition to the molecular biology tool kit is a technology for cutting DNA *in vivo*. *CRISPR* (which stands for clustered regularly interspaced short palindromic repeats, referring to the 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 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 at https://www.genscript.com/gsfiles/techfiles/CRISPR_handbook.pdf, 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 is able to protect 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. 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 useful in some cloning and mutagenesis procedures.

CRISPR targets longer, 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-repressing proteins.

📋 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 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 a double-stranded DNA molecule. If a molecular biology protocol results in a DNA molecule with a single nick, this will be pasted together when that molecule replicates *in vivo*. Bacteria do not typically ligate double strand breaks *in vivo*, although this does happen in eukaryotic cells (where it is called "non-homologous end-joining").

Generating recombinant DNA *in vitro* with purified ligase is more versatile. The most common enzyme used is the ligase from the bacteriophage T4. At high enzyme concentrations, T4 ligase will join blunt-ended linear DNA fragments into linear or circular products. 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. 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 *Nde*I and *Bam*HI

```
TATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAAG
  ACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATTCCTAG
```

Vector, digested with *Nde*I and *Bam*HI

```
...GAAGGAGATATACA          GATCCGGCTGCTAACAAA...
...CTTCCTCTATATGTAT            GCCGACGATTGTTT...
```

Ligated product

```
...GAAGGAGATATACATATGTTTAAAAAATCTGTTTTATTTGCAA...CACCAATGCAGATGATAAAATAATTCTGATAAGGATCCGGCTGCTAACAAA...
...CTTCCTCTATATGTATACAAATTTTTTAGACAAAATAAACGTT...GTGGTTACGTCTACTATTTTATTAAGACTATTCCTAGGCCGACGATTGTTT...
```

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!

EXAMPLES OF COMMON MOLECULAR BIOLOGY PROTOCOLS

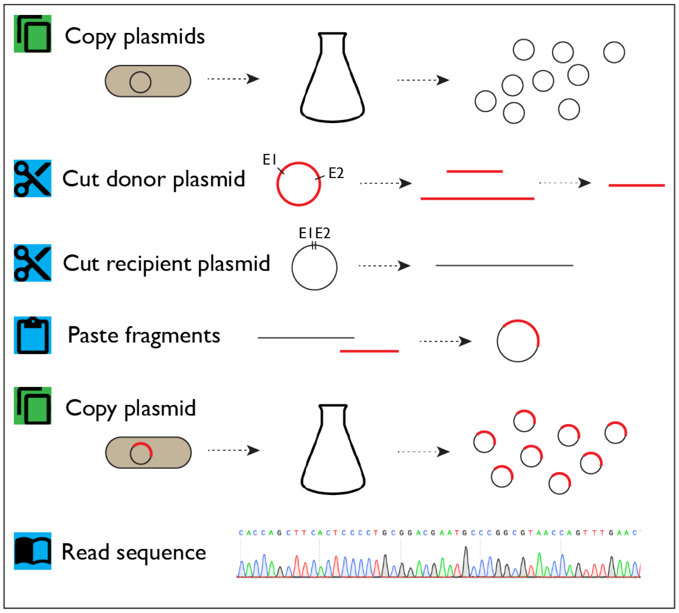
In this section, I will break down a series of protocols into their component steps, both in outline and graphical form. I will proceed from fairly simple procedures to more complex ones. Notice, however, that many of the steps are the same for all or almost all of the protocols. For example, essentially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

SUBCLONING

Subcloning is a protocol in which a DNA fragment from one plasmid is moved into another plasmid. Most plasmids contain arrays of defined restriction enzyme recognition sites called multiple cloning sites to make this kind of procedure straightforward.



Protocol:

1. *Copy – in vivo*

- grow cells containing donor and recipient plasmids to make large amounts of each

2. *Cut – in vitro*

- digest the donor plasmid with restriction enzymes that cut on either side of the gene, ideally two different enzymes that each create a different sticky end
- optionally, separate the resulting fragments on a gel and purify the fragment containing the gene you wish to subclone

3. *Cut – in vitro*

- digest the recipient plasmid with the same restriction enzyme(s)
- optionally, treat with a phosphatase (*e.g.* shrimp alkaline phosphatase) to remove the 5' phosphate from the DNA; this prevents ligase from rejoining the fragments of the recipient plasmid to each other

4. *Paste – in vitro*

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

5. *Copy – in vivo*

- transform the ligation mixture into a fresh bacterial strain and grow the culture under conditions that select for the desired plasmid to make a large amount of recombinant plasmid product

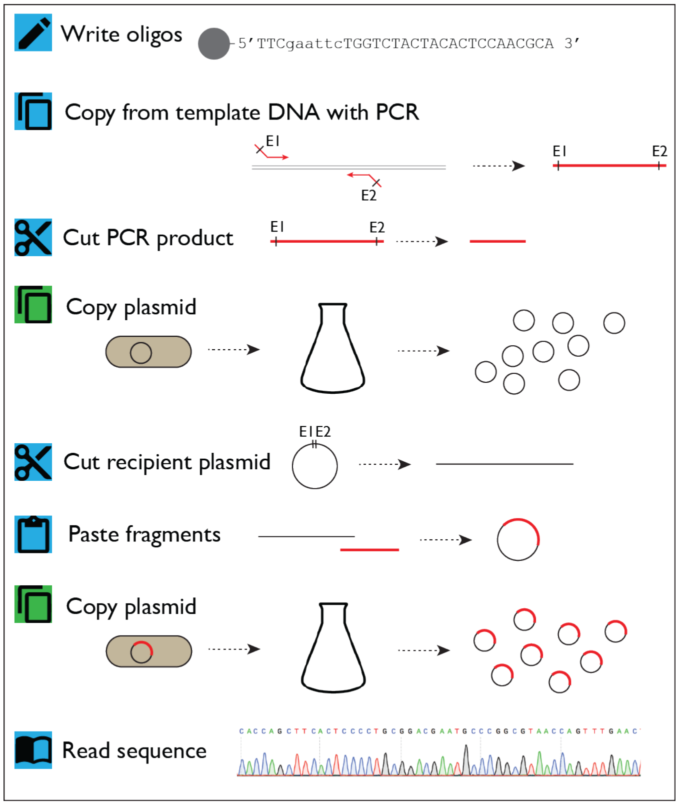
6. *Read – in vitro*

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

Exactly the same procedure can be done with completely or partially restriction-digested genomic DNA instead of a donor plasmid, 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.

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. Because PCR primer sequences 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.


Write oligos
5'TTCgaattcTGGTCTACTACACTCCAACGCA 3'
Copy from template DNA with PCR
E1
E2
E1
E2
Cut PCR product
E1
E2
Copy plasmid
E1E2
Cut recipient plasmid
Paste fragments
Copy plasmid
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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 DNA of interest from a template (for example, genomic DNA) using the primers designed in step 1

3. *Cut – in vitro*

- digest the PCR-amplified 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*

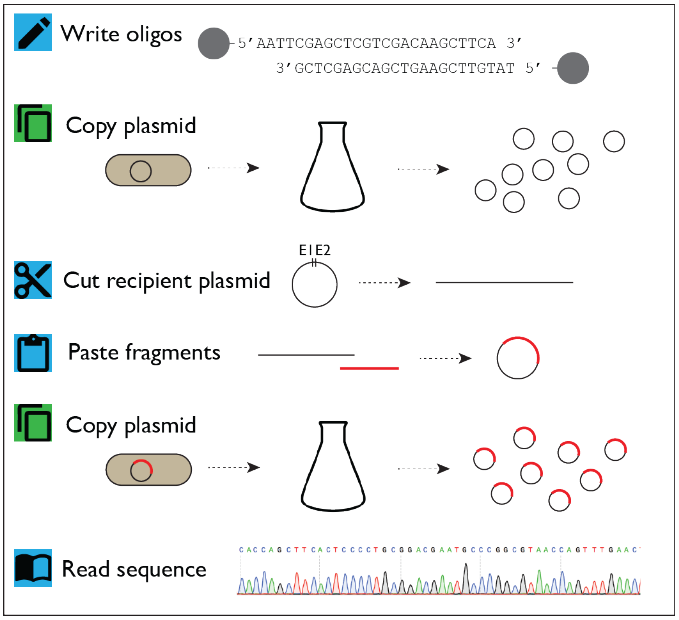
- transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

8. *Read – in vitro*

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

CLONING SMALL FRAGMENTS

It is often useful to clone very short DNA sequences. Many older plasmids have simple multiple cloning sites with only a few restriction sites, and you might want to add more. You might have a gene in a plasmid that you would like to add a promoter to, or a short amino acid tag for protein purification. In cases like this, you can *write* the sequence to be cloned directly. The main complication for this kind of cloning is screening for small inserts, which are often too small to be seen easily on a gel. It's therefore often a good idea to include a unique restriction site within the insert that will allow you to rapidly distinguish between your desired product and the original plasmid without having to sequence every possible candidate.


Write oligos
5′ AATTCGAGCTCGTCGACAAGCTTCA 3′
3′ GCTCGAGCAGCTGAAGCTTGTAT 5′
Copy plasmid
E1E2
Cut recipient plasmid
Paste fragments
Copy plasmid
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

Protocol:

1. *Write – in vitro*

- design PCR primers that contain your sequence of interest and are complementary to each other, with sticky ends for cloning
- anneal the primers to each other by mixing them, heating to 95°C, then cooling slowly to room temperature
- optionally, use T4 polynucleotide kinase to phosphorylate the 5' end of the annealed DNA fragment (oligos are not normally synthesized with a 5' phosphate, and that phosphate is necessary for ligase activity)

2. *Copy – in vivo*

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

3. *Cut – in vitro*

- digest the recipient plasmid with restriction enzyme(s) that match the sticky ends you designed into your primers
- treat with a phosphatase to remove the 5' phosphate from the recipient DNA (if you have phosphorylated the insert)

4. *Paste – in vitro*

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

5. *Copy – in vivo*

- transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

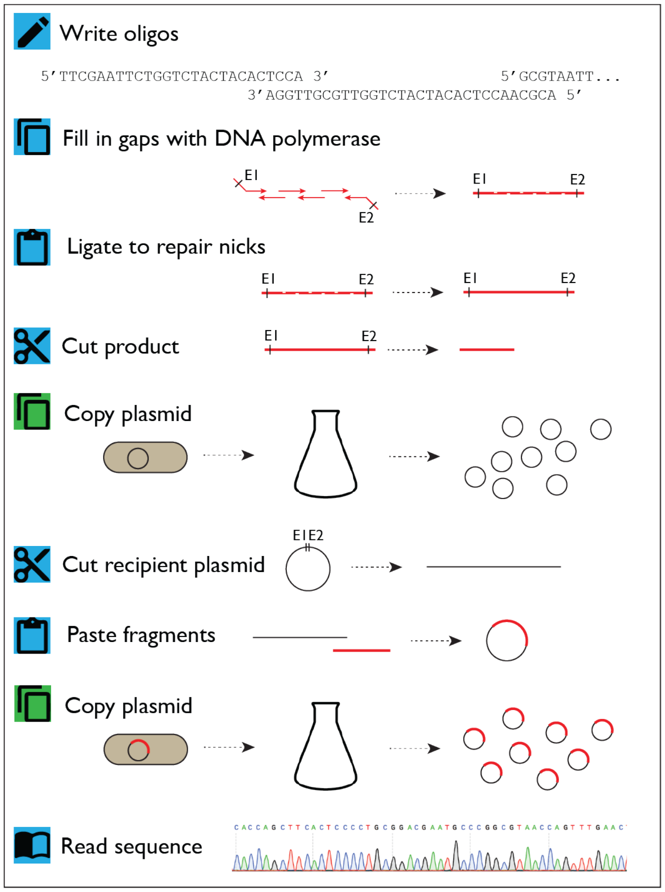
6. *Read – in vitro*

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

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". The resulting DNA can then be cloned as usual. Many companies will synthesize DNA for you in this way fairly inexpensively (about 35 or 40 cents per base pair). You could also do it yourself, although it requires careful primer design. In practice, synthesizing sequences longer than a single gene is usually not worth it, but the Craig Venter Institute has used this method to (very expensively!) synthesize an entire bacterial genome.

One common reason to have a gene synthesized rather than cloning it directly from genomic DNA is 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 effect 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.



<u>Protocol:</u>

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*

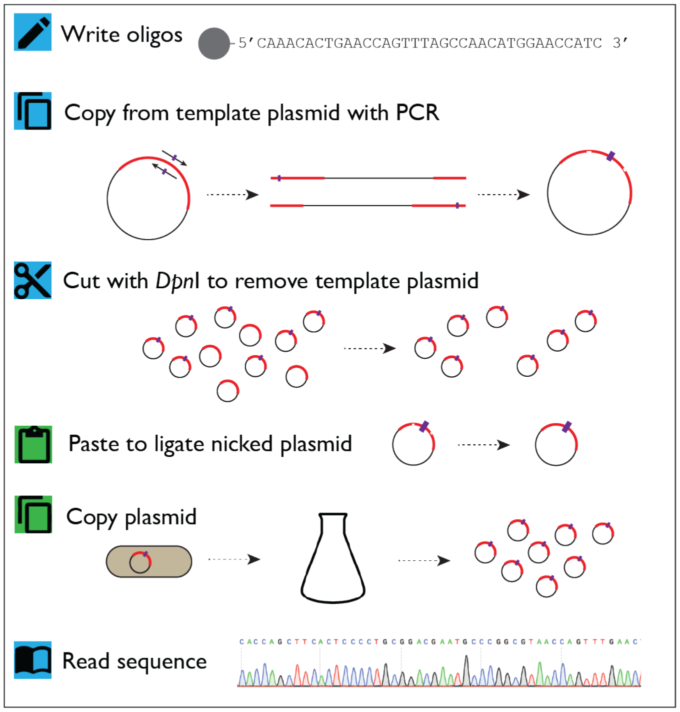
- transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

9. *Read – in vitro*

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

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.


Write oligos
5' CAAACACTGAACCAGTTTAGCCAACATGGAACCATC 3'
Copy from template plasmid with PCR
Cut with *DpnI* to remove template plasmid
Paste to ligate nicked plasmid
Copy plasmid
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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 at http://www.bioinformatics.org/primerx/ 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 *Dpn*I, which cuts methylated DNA at the sequence 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

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).

DISCUSSION PROBLEM SET #15: UNDERSTANDING NEW PROTOCOLS

There are a lot of different techniques available in the literature that people have used to construct particular kinds of recombinant DNA molecules. It is useful to be able to break them down into their component steps to make sure that you understand how a particular protocol is done, and how it can be used.

Here, for example, are links describing two protocols commonly used to construct plasmids with complex inserts:

SOEing PCR: https://www.future-science.com/doi/full/10.2144/000114017

Gibson Assembly: www.addgene.org/protocols/gibson-assembly/

Based on these links (and any other resources you can find), break down each of these procedures into a series of "read", "write", "copy", "cut", and "paste" steps, in the same way that protocols were broken down in the previous section. Make sure to include all necessary steps!

DISCUSSION PROBLEM SET #16: DESIGNING CONSTRUCTS FOR GENETIC EXPERIMENTS

Problem #1

Describe a detailed protocol for generating a plasmid which will allow inducible expression of a protein fusion between the MreB cytoskeletal protein of *E. coli* and the red fluorescent protein mCherry.

As raw materials, you have wild-type *E. coli* MG1655 genomic DNA and the plasmids linked below (and whatever standard genetic tools you care to use):

pBAD30: www.addgene.org/vector-database/1847/

pmCherry: www.addgene.org/vector-database/6597/

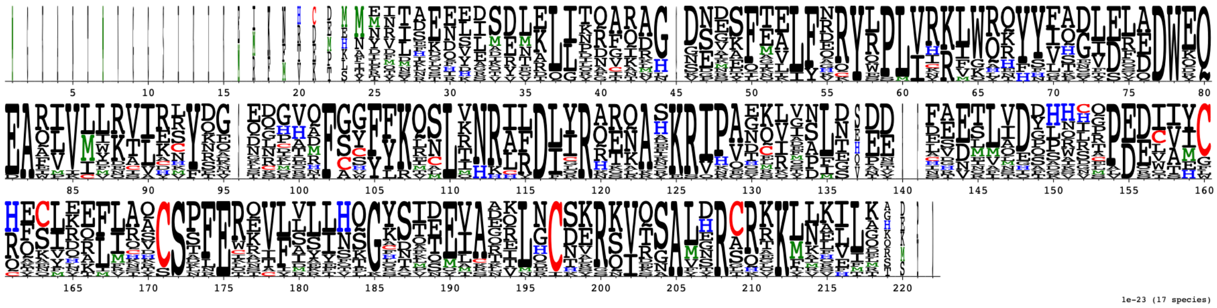
You can find the sequence of the mCherry gene on GenBank: www.ncbi.nlm.nih.gov/nuccore/MH070102.1

Be sure to include all of the necessary steps and draw a map of the resulting plasmid product.

What do you expect to observe when you express the MreB-mCherry fusion protein you have constructed in *E. coli*?

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, 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 number of useful *Lactobacillus-E. coli* shuttle vectors available. For the purposes of this experiment, use pTRKH2: www.addgene.org/71312/.)

- the independent and dependent variables

- both positive and negative controls
- 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 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, distantly related organisms (*horizontal gene transfer*).

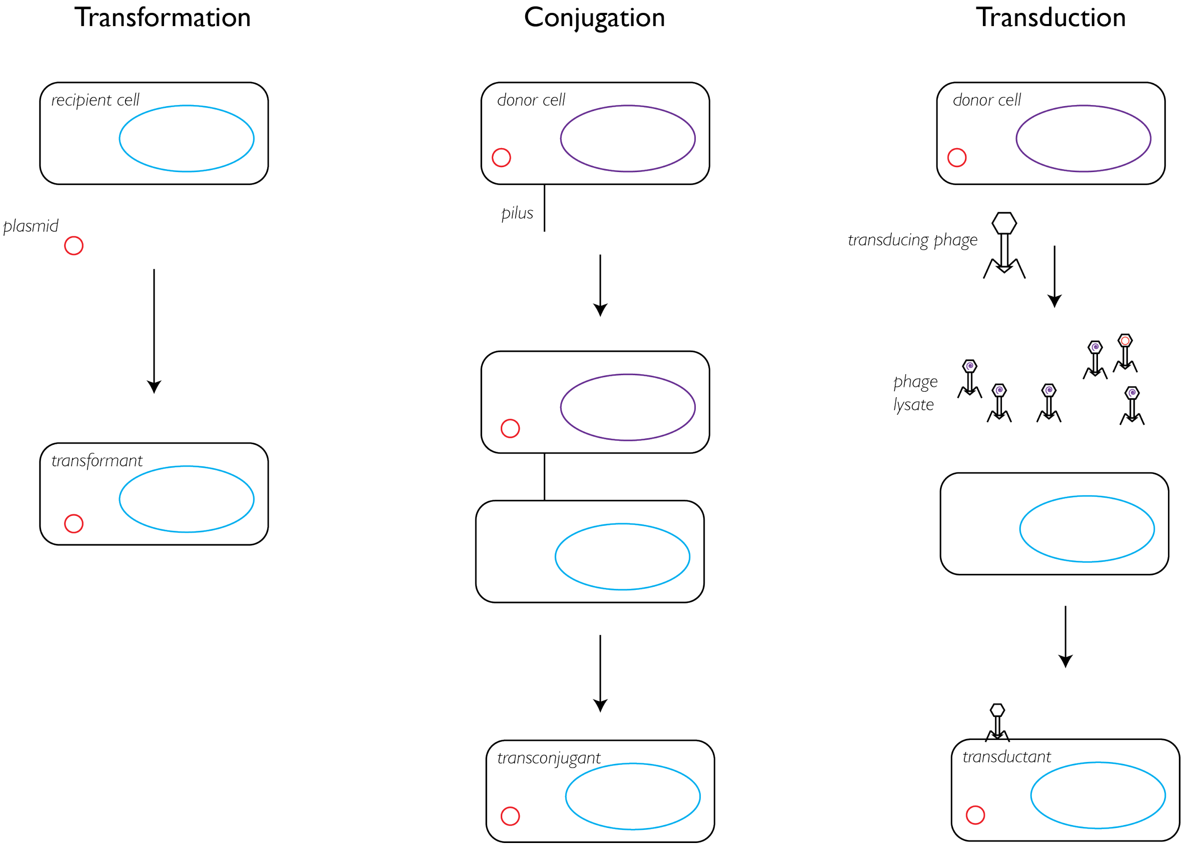


Figure 7.1. Moving a plasmid (red circle) into a recipient cell (blue chromosome) by three different methods. Note that transduction also results in phage particles containing fragments of chromosomal DNA from the donor cell (purple chromosome), which may also be transferred into the recipient cell. See the "Common Protocols" section of this chapter for more details.

Transformation is a process in which 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 can transform 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.) 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. 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. (Confusingly, for eukaryotic cells, direct uptake of DNA is called "transfection" instead of transformation.)

Conjugation is a process in which bacterial cells form tubular structures (called *pili*, which is the plural of *pilus*) on their surfaces and transfer DNA 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 more-challenging species, 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. 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. This is contrasted with *specialized transducing phage*, which are less useful and only 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 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 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*, 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.

DISCUSSION PROBLEM SET #17: GENE TRANSFER

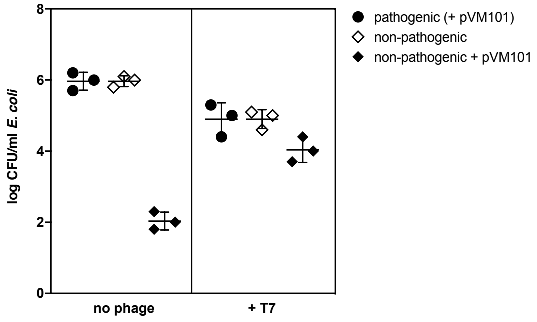
Problem #1

While studying antibiotic resistance in *S. aureus*, you discover that mixing an erythromycin-resistant strain with a chloramphenicol-resistant strain in media without antibiotics results in the appearance of strains resistant to both antibiotics.

Propose an experiment to determine the mechanism by which this genetic exchange occurs.

Problem #2

During intestinal infections, the number of phage particles in the gut increase dramatically. Pathogenic strains of *E. coli* often contain plasmids which encode toxins, siderophores, and other virulence factors. One of these, called pVM101, is more than 150 kbp in size. You infect mice with a combination of pathogenic (pVM101-containing) and non-pathogenic *E. coli* and measure the rate of transfer of pVM101 to the non-pathogenic strain during the infection. You find that adding the *E. coli*-infecting phage T7 greatly increases this transfer.



T7 is an obligately lytic phage with a 40 kbp genome. Propose a model to explain your observations, and an experiment to test that model.

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 smaller fragments of chromosomal DNA, but generally only between fairly closely related strains. It is important to remember that plasmids can be moved from one cell to another by transduction, it is just 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.
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. 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*.

The main physiological function of homologous recombination in cells is in DNA damage repair, and the complex details of its mechanism are beyond the scope of this course. However, it is important to have a general sense of how it works, since many genetic engineering procedures depend on it.

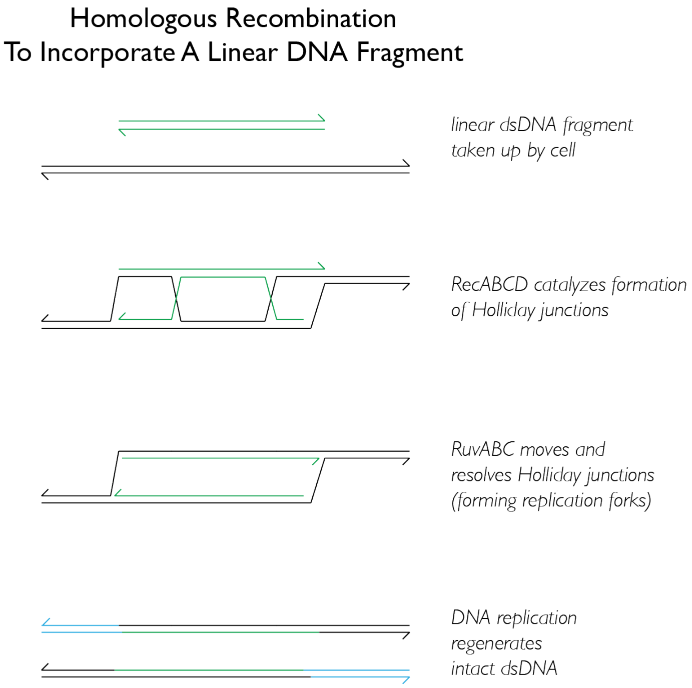


Figure 7.2. A very simplified diagram illustrating incorporation of a linear DNA fragment into a bacterial chromosome by homologous recombination.

As shown in Figure 5.2, 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. 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 of the *rec* or *ruv* genes are unable to carry out homologous recombination and are extremely sensitive to DNA damaging chemicals and radiation.

Lysogenic bacteriophage encode their own recombinases, which they use to integrate themselves into the host chromosome. These often require much shorter regions of homology than RecABCD to stimulate recombination, which has made them useful tools for molecular genetics, as we will discuss below.

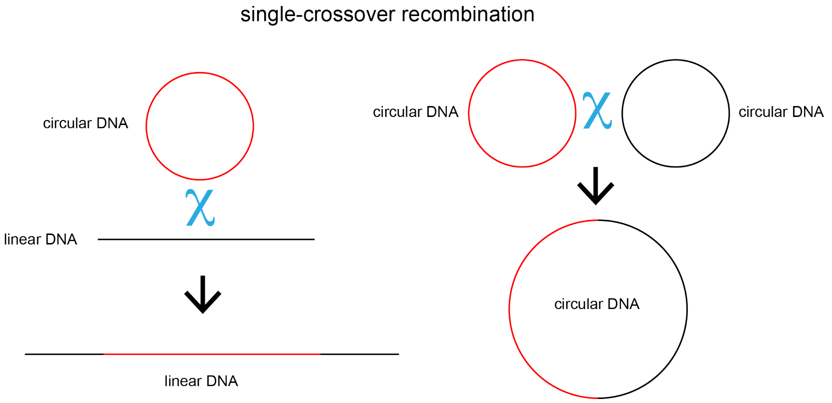
USING HOMOLOGOUS RECOMBINATION FOR GENETIC ENGINEERING

Swap

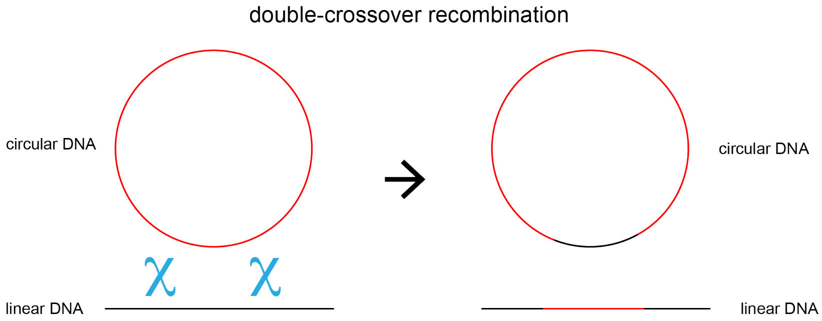
As mentioned in the last chapter, one of the six fundamental procedures in molecular biology is recombination, which I abbreviated as "swap", since it results in swapping or exchanging sequences from one DNA molecule to another. In the next section of this chapter, I will describe a variety of genetic engineering protocols that depend on recombination, to illustrate what is possible. Swap steps always happen *in vivo*. The 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 itself.

It is important to distinguish between *single-crossover* and *double-crossover* recombination events, which are distinguished by requiring either one or two independent homologous recombination events. Single-crossover occurs much more frequently, and combines two DNA molecules into a single product, but 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 almost exclusively rely on double-crossover recombination, which generates 2 products, but no DNA strand breaks in the circular molecule.



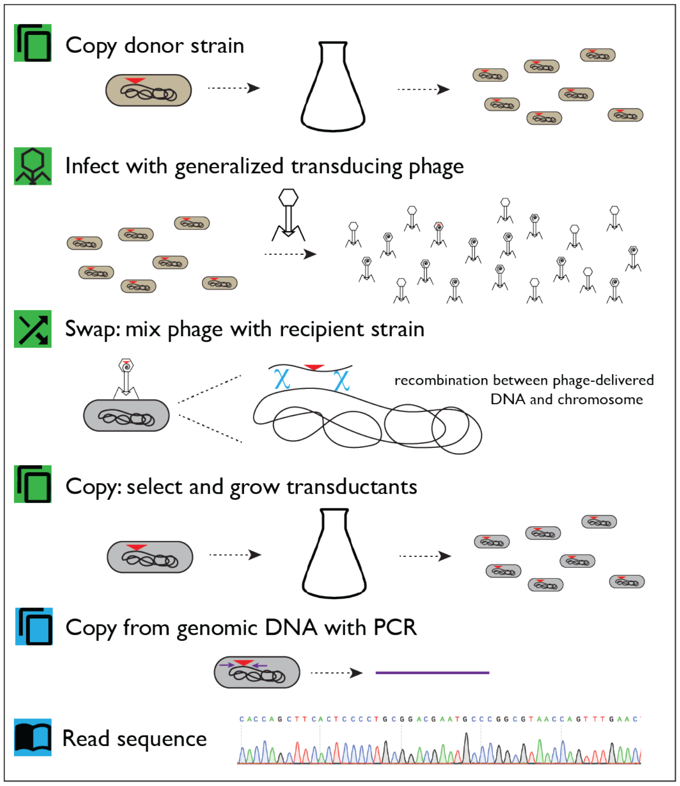
In either case, but especially when demanding a double-crossover product, recombination is a rare event, so having a strong selection for strains that contain the desired final product is essential.

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 many of them also require some amount of plasmid engineering, which can be done by the methods described in the previous chapter.

TRANSDUCTION

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.


Copy donor strain
Infect with generalized transducing phage
Swap: mix phage with recipient strain
χ
χ
recombination between phage-delivered DNA and chromosome
Copy: select and grow transductants
Copy from genomic DNA with PCR
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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 generalized transducing phage and harvest phage particles, some of which will contain DNA from the donor cell chromosome

3. *Swap – in vivo*

- add phage containing selectable marker 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; this is commonly done by chelating away calcium, which many transducing phage require for attachment

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*

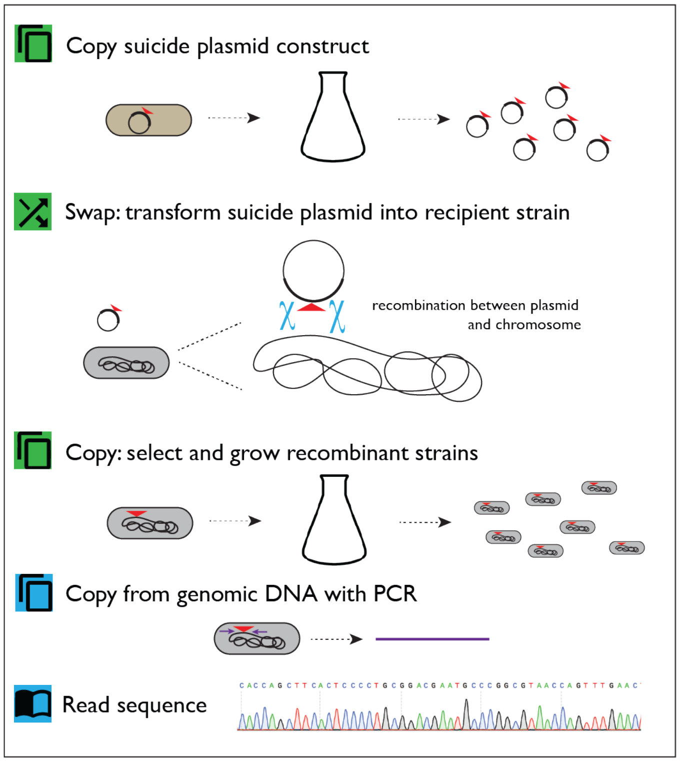
- sequence the PCR product to confirm that the selected transductant has the desired sequence derived from the donor strain

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, 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 is commonly 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*. (This can also be used to calculate the distance between two mutations on the chromosome, if you don't already know that information. This is called *linkage mapping* and has been made almost entirely obsolete by inexpensive genome sequencing.)

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 "colony PCR", and works because the 95-98°C melting step of the PCR cycle lyses 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. RecA usually requires very 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.


Copy suicide plasmid construct
Swap: transform suicide plasmid into recipient strain
recombination between plasmid and chromosome
Copy: select and grow recombinant strains
Copy from genomic DNA with PCR
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

<u>Protocol:</u>

1. *Copy – in vivo*

- grow cells containing suicide plasmid 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*

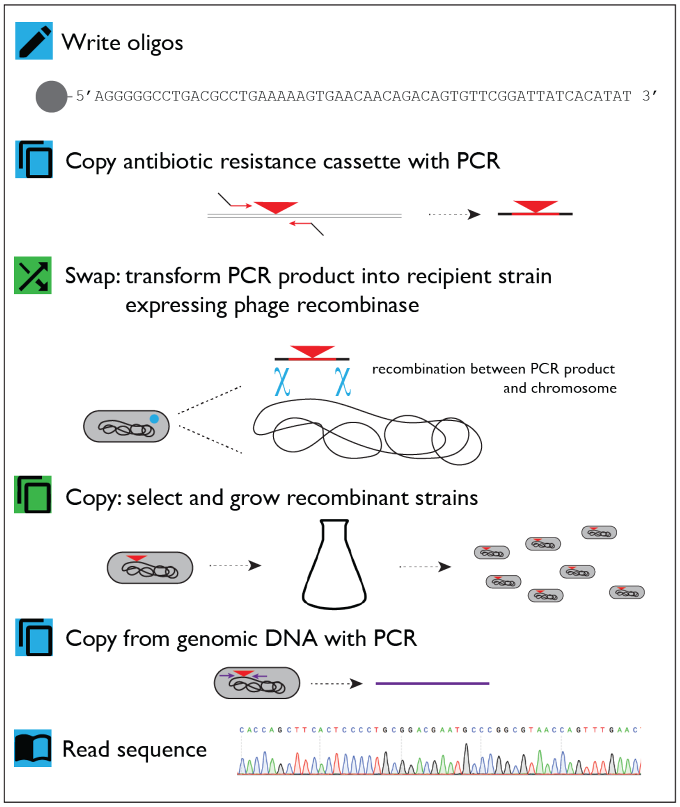
- sequence the PCR product to confirm that the selected strain has the desired mutation

Since you will be selecting for the antibiotic resistance encoded by the cassette you want to insert into the chromosome, using a suicide vector is essential. Otherwise, all of the Ab^R colonies you obtain will simply be plasmid transformants, not chromosomal mutants.

Note that this is the only protocol in this chapter which can use single-crossover recombination, since the suicide vector is a circular DNA molecule. In this case, the "swap" step might involve one recombination step to integrate the plasmid into the chromosome, followed by a second single-crossover recombination step to "loop" the integrated plasmid out, which will (about 50% of the time) result in the chromosome containing the allele that was originally in the vector.

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 sequence homologous to the host chromosome. The PCR products used for recombineering almost always contain an antibiotic resistance gene to allow selection of recombinants.


Write oligos
5' AGGGGGCCTGACGCCTGAAAAAGTGAACAACAGACAGTGTTCGGATTATCACATAT 3'
Copy antibiotic resistance cassette with PCR
Swap: transform PCR product into recipient strain expressing phage recombinase
recombination between PCR product and chromosome
Copy: select and grow recombinant strains
Copy from genomic DNA with PCR
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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*

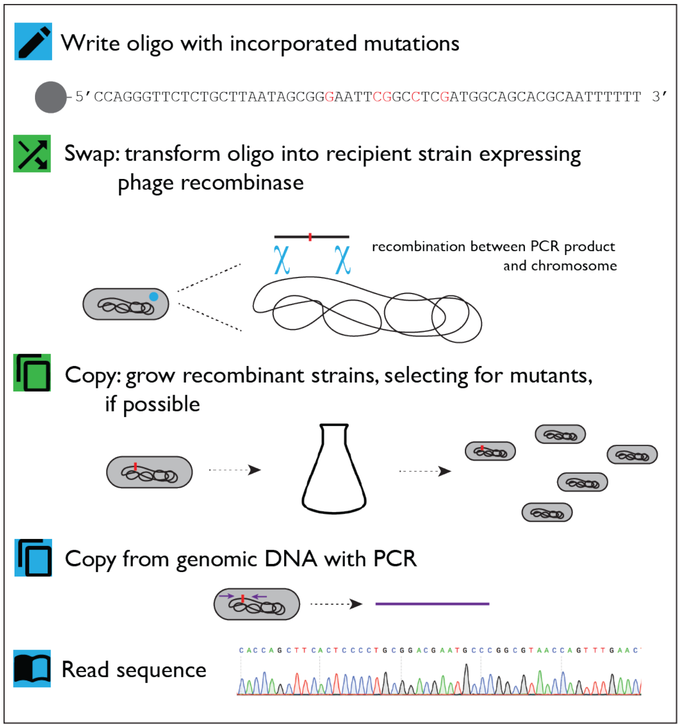
- sequence the PCR product to confirm that the selected strain has the desired mutation

The phage recombinase (very commonly λ Red, 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 **Chapter 7**. It is not generally healthy for bacteria to constitutively express recombinases, which can lead to unwanted chromosome rearrangements. Recombinase expression plasmids for recombineering often have temperature-sensitive origins of replication to make curing the plasmid easy after the desired chromosomal mutation(s) have been made.

OLIGO-DIRECTED RECOMBINEERING

The phage recombinases used for recombineering also work with single-stranded DNA templates. It is only possible 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.

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.


Write oligo with incorporated mutations
5' CCAGGGTTCTCTGCTTAATAGCGGGAATTCGGCCTCGATGGCAGCACGCAATTTTTT 3'
Swap: transform oligo into recipient strain expressing phage recombinase
χ
χ
recombination between PCR product and chromosome
Copy: grow recombinant strains, selecting for mutants, if possible
Copy from genomic DNA with PCR
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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

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). Making 3 or 4 silent mutations directly adjacent to the mutation of interest can improve efficiency greatly, as can synthesizing the oligo with more-stable phosphorothioate linkages at the 5' or 3' ends. 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 addition to the molecular genetics toolkit is CRISPR, which stands for clustered regularly interspaced palindromic repeats. 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 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 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.

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 proteins to visualize where particular DNA elements are found in cells, and has also been used to either activate or (more commonly) repress gene expression (resulting in a *gene knock-down* instead of a gene knockout).

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. This has stimulated an active debate in the scientific community about the ethics of genetic engineering in higher organisms.

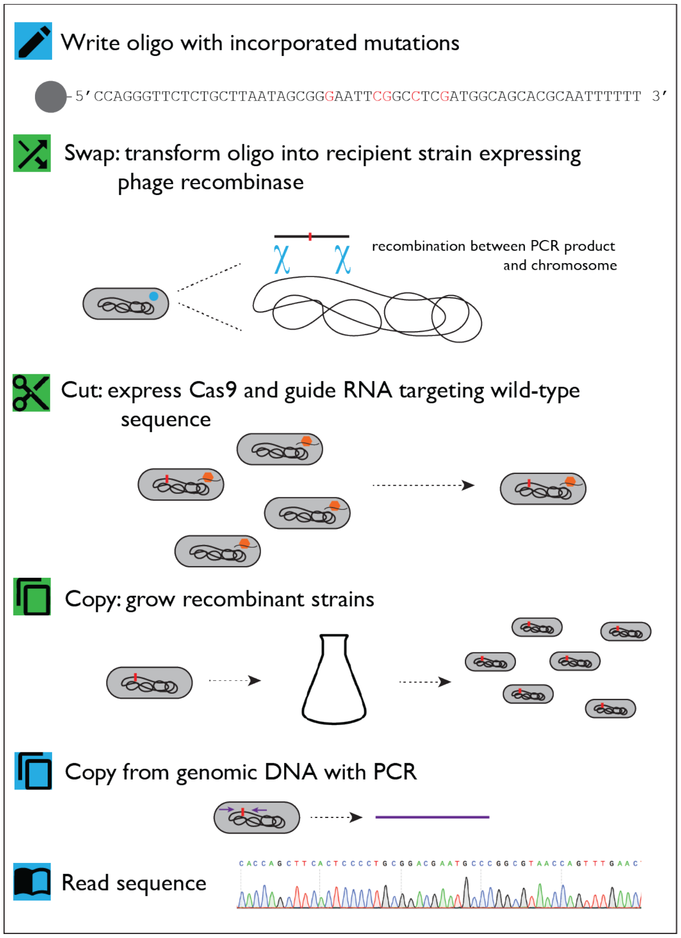
CRISPR-ASSISTED RECOMBINEERING

The most common use of CRISPR in bacterial genetics is in combination with recombineering. 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, essentially creating a selection for the desired mutant. 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 required compared to the previous protocol: constructing a plasmid that will express the guide RNA to target Cas9. Any of the cloning methods from the previous chapter can be used, but since guide RNAs are very small (about 20 nt), "Cloning Small Fragments" is the most common approach.

As with recombinase expression plasmids, it is important that the plasmid(s) expressing Cas9 and the guide RNA be easily curable. This is often accomplished with temperature-sensitive origins of replication.

It is well worth searching the literature (and the plasmid repository at addgene.org) to see if anyone has developed a CRISPR-based mutagenesis system for your organism of interest, although since it is so 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.


Write oligo with incorporated mutations
5′ CCAGGGTTCTCTGCTTAATAGCGGGAATTCGGCCTCGATGGCAGCACGCAATTTTTT 3′
Swap: transform oligo into recipient strain expressing phage recombinase
recombination between PCR product and chromosome
Cut: express Cas9 and guide RNA targeting wild-type sequence
Copy: grow recombinant strains
Copy from genomic DNA with PCR
Read sequence
CACCAGCTTCACTCCCCTGCGGACGAATGCCCGGCGTAACCAGTTTGAAC

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

SINGLE-COPY INSERTION ELEMENTS

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. 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 insert 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 are a very clean method to use for complementation experiments.

For phage-based systems, it is important that the inserted sequence not include the genes required for production of live phage particles, since cultures with active infections behave very differently from uninfected cells. The resulting irreversible insertion is called a *defective prophage* or *stable lysogen*. Perhaps 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 is a widely-used system for making single-copy insertions. 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.

DISCUSSION PROBLEM SET #18: EXPERIMENTAL DESIGN WITH RECOMBINATION

Problem #1

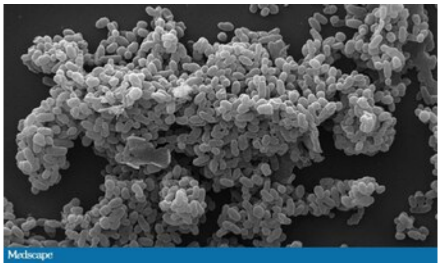
For many model organisms, *knockout collections* have been generated that consist of thousands of individual mutants, generally one in each non-essential gene for that organism. These may be generated by isolating individual transposon insertion mutants (like the Nebraska Transposon Mutant Library for *Staphylococcus aureus* USA300_FPR3757; www.unmc.edu/pathology/csr/research/library.html) or by recombineering (like the Keio collection for *E. coli* BW25113; www.ncbi.nlm.nih.gov/pmc/articles/PMC1681482/), but regardless, are a tremendously useful resource.

While studying responses to starvation stress in *E. coli*, you find that mutations in *rpoS* and *rpoN* both reduce the ability of *E. coli* to grow in minimal media. These genes encode alternative sigma factors, and you hypothesize that they each drive the expression of different sets of genes needed under those 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*.

Describe a detailed protocol to build an *E. coli rpoS rpoN* mutant. You have access to the Keio collection and all standard genetic tools.

Problem #2

Akkermansia muciniphilia is a mucus-degrading bacterium found in the mammalian intestine that is associated with reductions in obesity. Human derived strains of *A. muciniphilia* do not colonize mice efficiently, and *vice versa*. You use UV mutagenesis to mutagenize a human strain and select for mutants that colonize mice well.



MedScape

All of the mutants you isolate have multiple point mutations throughout their chromosomes, but you isolate several strains with a particular C to T point mutation in *waaL*, a gene involved in lipopolysaccharide biosynthesis. You hypothesize that this mutation changes the surface properties of *A. muciniphilia*, conferring host-specific colonization.

Propose an experiment using recombination to test this hypothesis. State:

- a detailed description of how you will construct the necessary strains
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

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-Nafría (2019) "*In vivo* DNA assembly using common laboratory bacteria: A re-emerging tool to simplify molecular cloning." J Biol Chem 294(42):15271-15281.

SUMMARY OF EXPERIMENTAL DESIGN PRINCIPLES

This section 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 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

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
abstract	a short summary of a paper
alignment	a visual representation of homology between DNA, RNA, or protein sequences
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 sigma factors	sigma factors responsible for recognizing promoters other than those recognized by the housekeeping sigma factor; often involved in stress response or development
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 cassette	a gene encoding a product that confers antibiotic resistance, along with all of the additional sequences needed to ensure its expression
attachment site	(also "*att* site") the specific DNA sequence at which lysogenic bacteriophage (and some transposons) insert themselves into their host chromosomes
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
bacteriophage	a virus that infects bacteria
biochemistry	the study of the physical properties of biological molecules
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
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
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)
chimeric protein	see "protein fusion"
ChIP-seq	(also "chromatin immunoprecipitation sequencing") an *in vivo* 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 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
compatibility group	see "origin of replication"
competent cells	bacteria capable of taking up DNA directly from their environment (transformation)
complementation analysis	an experimental design that establishes genetic causation by removing and replacing individual genes
conditional phenotype	a phenotype that is only observed under specific growth conditions
conjugation	DNA transfer between cells *via* pili; requires *tra* 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
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
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
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
data	high-quality, carefully recorded observations
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
divergently transcribed	genes adjacent to each other on the chromosome, but transcribed in opposite directions
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"
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
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
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
enrichment	a procedure that increases the proportion of mutants of interest in a population
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
exonuclease	a nuclease that degrades a DNA or RNA molecule from one end
experiment	a test of the effects of a specific manipulation on a system
f1 origin	a site that allows a plasmid to be packaged as concatenated single-stranded DNA when the host bacterium is infected with bacteriophage f1
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?
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
frameshift mutation	insertion or deletion of 1 or 2 nucleotides (or any number not divisible by 3)
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
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
genes of unknown function	genes with no currently known role in the cell
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
global regulator	a regulator that controls many genes or gene products from around the genome
guide RNA	a short sequence that serves to direct Cas9 nuclease to a specific target site
hairpin	a DNA or RNA structure that is folded into a small, stable loop
Holliday junction	the crossover point between two homologous DNA sequences that is the essential intermediate in homologous recombination
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

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
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
initiating nucleotide	the first nucleotide of a transcribed RNA
insertion	the addition of extra DNA sequence into the chromosome
intergenic suppressor	a mutation in a different gene that reverses the phenotype of a mutation
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
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
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
lysogen	a bacterial cell containing a prophage
lysogenic phage	a bacteriophage able to integrate itself into the chromosome of a host cell
mass spectrometry	a powerful technique for determining the molecular weight of molecules
material transfer agreement	paperwork necessary to transfer research materials from one university to another
megabase pair	1,000,000 base pairs
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
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}
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
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
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 overexpressed
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
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
N-acylation	covalent addition of acyl groups to lysine residues in proteins
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
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
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
null mutation	a mutation that inactivates a gene product
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 regulatory protein binds
operator sequence	the DNA sequence to which a regulator binds
operon	several genes encoded on the same mRNA
origin of replication	(also "*ori*" or "*oriC*") the site which determines the ability of a plasmid to replicate within a cell, it's copy number, and host range
origin of transfer	(also "*oriT*") a DNA sequence allowing a plasmid to be mobilized by conjugation
orthologs	(also "orthologous genes" or "orthologous proteins") homologs in different genomes
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
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
peptide	a short protein
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
permissive temperature	for temperature sensitive mutants, the temperature at which the gene functions
phage recombinase	a highly efficient recombinase derived from a lysogenic bacteriophage
phagemid	a plasmid with an f1 origin
phenomenon	a measurable event in objective reality
phenotype	the measurable physical properties of an organism

phosphatase	an enzyme that removes phosphate groups from a substrate
phosphorylation	covalent addition of a phosphate group to a molecule
phylogenetic tree	a visual representation of evolutionary relationships
phylogeny	the evolutionary relationship between organisms or genes, inferred from homology
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)
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
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
polymerase chain reaction	(also "PCR") a very common method that uses DNA polymerase to amplify large amounts of a specific DNA molecule *in vitro*
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 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
prestige journal	a journal which only publishes "high-impact" science; *Nature, Science, Cell, 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
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
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 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 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
prototroph	a strain that does not require a particular nutrient (compare to auxotroph and bradytroph)
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
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
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
rare codons	codons that are not translated efficiently in an organism due to low numbers of tRNAs for that codon
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
replica printing	using a sterile piece of velvet as a printing block to transfer colonies to several different plates; useful for screens
reproducibility	a desirable property of experiments: they give the same result each time
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 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
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
Sanger sequencing	a common and inexpensive way of sequencing several hundred to 1000 bp of DNA
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 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
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
Shine-Dalgarno sequence	see “ribosome binding site”
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 sequence	an N-terminal protein sequence that is recognized by cellular export machinery and directs the cell to secrete the protein
silent mutation	a mutation of an amino acid encoding codon to a different codon encoding the same amino acid
single-crossover recombination	a recombination event that requires only one homologous recombination event, such as integrating a circular plasmid into a circular chromosome
single nucleotide polymorphism	see “point mutation”
site-directed mutagenesis	(also “targeted mutagenesis”) constructing a specific mutation at a specific site in a DNA

	molecules
site-directed mutagenesis	precise construction of specific mutations at specific sites in a DNA molecule
society journal	a journal published by a scientific professional society
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
spontaneous mutagenesis	random mutations resulting from natural mistakes made by DNA polymerase during replication
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
sticky end	a staggered double strand break produced by some restriction enzymes
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
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)
synthetic lethality	two genes which can be knocked out individually, but not simultaneously
temperature-sensitive origin of replication	an origin of replication that only functions at low temperature, typical of some suicide vectors
temperature-sensitive mutant	a mutant that grows at low temperature, but not at high temperature; typically due to mutations that destabilize essential proteins
terminator	a sequence which stops transcription
testable	see "falsifiable"
tra functions	genes encoding the machinery that allows transfer of plasmids with an appropriate *oriT* 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
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
treatment	see "independent variable"
two-hybrid assay	a screening method that uses protein fusions to identify protein-protein interactions *in vivo*
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
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