Beyond Single Cells:

socio-cellular biology & developmental mechanisms

<u>Review section</u> for introduction to the social (cooperative) interactions between cells, the mechanisms involved, and how they lead to complex communities, embryonic development, and (perhaps) self-aware neural networks.

by

Michael W. Klymkowsky

Molecular, Cellular & Developmental Biology, University of Colorado Boulder

Preface

Life, as we know it, is based on cells. While the structural characteristic of cells can vary dramatically the overwhelming body of evidence indicates that cells are homologous. What does that mean? It means that all cells appear to share a common ancestor, known as the last universal common ancestor or LUCA.[1] The result is that while there are many (many) different types of organisms, there is only one type of life on Earth, characterized by shared molecular and cellular mechanisms.

While it is common to think of unicellular organisms as independent individuals, there is a growing realization that there are complex interactions between unicellular organisms, as well as within and between multicellular organisms.[2] The theme of this book (and the course it is meant to accompany) is that whether within an multicellular organism or microbial community, cells interact to produce and respond to complex "social" behaviors. Our focus is will be community effects displayed by bacteria and archaea, including the "self-sacrifice" of one cell for the benefit of others,[3] to the developmental and functional complexities that underlie the embryonic development and adult behavior of multicellular animals (metazoans). We will consider both transient metazoans, such as slime molds, and more familiar "constitutive" metazoans (animals). Interactions between organisms at the cellular level are involved in many diseases and play important roles in normal development.[4] We will consider how basic cellular processes have been adapted to produce these "communal" behaviors.

A surprising observation that has emerged from the molecular level study of different organisms is how apparently distinct social and developmental processes are based on a common set of conserved molecular and cellular mechanisms. What makes our approach rather different from a conventional developmental biology textbook is a focus on these common processes and how they were identified, typically through studies of uniquely experimentally accessible organisms. We will then follow their role(s) in "higher" systems, including humans. We will consider a number of organisms. Our goal is not to describe these various systems in detail, but to help you recognize how common processes produce species specific behaviors, including those involved in embryo formation. We are not the first to take this approach; it was a major theme of John Gerhart and Marc Kirschner's *Cells, Embryos, and Evolution* (Blackwell Science, 1997). On a personal note, I was introduced to the use of *Xenopus laevis* (clawed frog) embryos as an experimental system in John's lab; the generosity of John and his lab members is very much appreciated.

---

[1] All acronyms used can be found in appendix I at the end of the book.

[2] West et al., 2021. Ten Recent Insights for Our Understanding of Cooperation

[3] Lewis., 2000. Programmed Death in Bacteria & Vostinar e al., 2019. Suicidal selection: Programmed cell death can evolve in unicellular organisms due solely to kin selection

[4] Agrawal & Broderick.Inside help from the microbiome eLife 2023

Since we build on common molecular and cellular principles, we begin with a brief review (Chapter 1) of our pedagogical approach and foundational cellular and molecular processes, most of which you are likely to already be familiar with. These various biological processes are found over and over in various system, and speak to the evolutionary relationship of organisms. Much, but not all, of this background material is presented in the open-source, that is free, introductory biology text biofundamentals.[5] Following this review we move on to consider social interactions, first within unicellular systems and then in the development of multicellular systems. Each organism considered provides unique experimental opportunities that allow for identification of underlying molecular and cellular mechanisms. Our goal is to help you develop the working knowledge needed to analyze, understand and evaluate scientific claims related to sociocellular and developing systems in general, and human development in particular. We will consider recent advances in *in vitro* embryonic development, and the various methods involved. While extremely important, we will not consider the ethics or ethical implications of these studies; these are topics you may want to discuss in class. Our approach is science-focussed and science (at least according to this author) has little or nothing to say about what is right or wrong. Throughout our discussions, it is worth remembering that each system discussed has its own multi-million year evolutionary history and displays evidence of species specific adaptations - the result of evolutionary tinkering.[6] We will note and pass over many species-specific features.

Each section of the book includes questions to answer and ponder. Their purpose is to provide opportunities to practice generating plausible, rather than "correct", responses. We recognized that we live in the age of generative AI and that it can be helpful in getting started with writing and such, but at the end of the day you are responsible for being able to explain and evaluate topics covered on your own. You are encouraged to challenge statements made in class that make no sense to you, or that you find confusing; talking with others often helps. Try and identify what specific points you find confusing and why. In class, we rely on beSocratic [link] activities and social reading / commenting through nota bene [link] to fuel useful in class discussions. Working with these tools, and joining in and out of class discussions will help you test your progress toward an accurate working understanding of molecular and cellular mechanisms involved in sociocellular systems.

A note on footnotes: I have an inordinate fondness for footnotes. Typically, they contain references for further reading. Be careful to avoid getting lost in, or distracted by, them - the world is a labyrinth of treasures (and monsters).

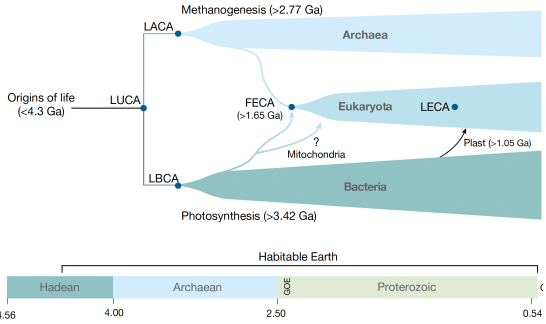
[5] The overall form of this text is shaped in large part by my experiences in building more coherent, concept-centric curricular in chemistry (Chemistry, Life, the Universe and Everything – CLUE and organic CLUE) and biofundamentals, both collaborative projects with Melanie Cooper. Further background information can be found here and here.

[6] Jacob, F. 1977. Evolution and Tinkering

Chapter 1: Working cell and molecular concepts & socratic black boxes

This course is designed for people who have completed introductory courses in cell and molecular biology. Since we are building on concepts presented earlier, we begin by reviewing the core molecular and cellular principles involved. We rely heavily on the socratic unpacking of "black boxes" to identify plausible and testable mechanisms that underlie various socio-cellular and developmental processes. This should also help you develop a critical scientific mindset.

Biological systems are the most complex "things" in the universe. Around 3.5 billion years ago, the available evidence indicates that lipid membrane-bounded, thermodynamically non-equilibrium chemical systems arose; they gave rise to the first cells, known as LUCA–the last universal common ancestor.[7] Through a process of DNA replication and controlled fragmentation (cell division), these systems produced more cells. This statement, a distillation of many observations, is known as the "Cell Theory".[8] Because DNA replication is not perfect and DNA is not stable and can react with radiation and various "mutagens" in its environment mutations occur. Mutations can influence the reproductive success of cells and the organisms that they form. The result is biological evolution. Evolutionary processes favor the specialization of distinct populations of organisms, species, with the end result, the diversification of life - the subject of Charles Darwin's revolutionary book "The Origin of Species".[9] Three distinct types of life, which share many common features inherited from LUCA, emerged: the bacteria, the archaea, and the eukaryotes. While most bacteria and archaea are single celled organisms, eukaryotes have evolved into a wide array of multicellular organisms, the plants, animals, and fungi.



It is common to see evolutionary processes portrayed as a struggle between individuals, leading to the "survival of the fittest" and, in the extreme, a "merciless war of all against all".[10] The reality is generally quite different; there are many examples of cooperation. Moreover, changes in the behavior of individuals (perhaps fostered by copying the behaviors of parents and others) within a population can lead to changes in the direction and strength of selection.[11]

---

[7] Javaux, E.J. (2019) Challenges in evidencing the earliest traces of life. Nature.

[8] Mazzarello, P. (1999) A unifying concept: the history of cell theory.

[9] Dennett, D. C. (1995). Darwin's dangerous idea. *The Sciences*, *35*(3), 34-40.

[10] Dugatkin, L.A. 2011. "The prince of evolution"

[11] Huber et al., 2009. The evolution of imitation: what do the capacities of non-human animals tell us about the mechanisms of imitation? Laland et al., 2015. The extended evolutionary synthesis: its structure, assumptions and predictions

The complexities of cells and their implications: Cellular complexity, and the complexity of the organisms that they form, arises from the number of distinct, but interacting molecular and cellular components involved. These components interact, often in multiple ways with multiple partners. The strengths of these interactions are based on molecular shapes and surface properties. These are properties that can be modified; in the case of proteins modifications include phosphorylation, acetylation, methylation, glycosylation, lipidation, and more; each can influence the specificity, strength, and effects of the interactions between molecules. Given the small size of cells, and the corresponding rather small (compared to conventional chemical systems) numbers of molecules of each type present – in particular the presence of one or two copies of each gene – biological systems display a significant stochastic component, often referred to as noisy behavior.[12] While a challenge, associated with avoiding catastrophic system failures, that is death, stochastic behaviors also offer opportunities - the ability of the system to produce a range of behaviors. Each cell currently alive has been "running" without interruption for ~3.5 billion years, in large part due to the effectiveness of various feedback systems that keep the cell functioning and alive.

The multiple interacting systems within each cell are influenced by a range of factors, these include environmental variables, some general such as temperature and O_2 concentration and others. Cellular systems can also be influenced by various social signals from (and to) and physical (adhesive) interactions with other cells, the primary focus of this book. An important factor is how a cell's "pre-existing state" influences the cell's response to environmental and internal changes. This pre-existing state includes the specific alleles of the genes they inherited, which genes are currently expressed at what levels, what proteins (and other molecules) are present, at what concentrations, together with the extent of their "post-synthesis" modification. Differences in the pre-existing states of cells of the same "type" can lead to dramatically different responses to the same "signal" or perturbation. It is worth noting that our understanding of the complexities of these systems is rarely complete.

While all "types" of organisms share common ancestors, populations (species) have diverged over the course of many millions of years. For example the common ancestor of mice and humans appears to have lived ~65 million years ago, but since that time, the populations that gave rise to modern mice and humans have diverged rather dramatically.[13]

Reviewing cellular and molecular basics: So that we can all be on the same page as far as possible, we will use aspects of various scientific papers to identify and review common molecular and cellular processes involved in socio-cellular interactions. To begin with we will consider molecular interactions and their stability and specificity (selectivity). We will then consider the structures of genes and proteins, and how they interact, together with the various feedback interactions that influenced patterns of gene expression and resulting biological behaviors. We will also consider the effects of mutations and noise (stochasticity) and how the structure of proteins often enables the construction of molecular chimeras with experimentally useful traits, such as fluorescent tags for visualizing proteins in living cells. We will review the basic processes of cell division and sexual

---

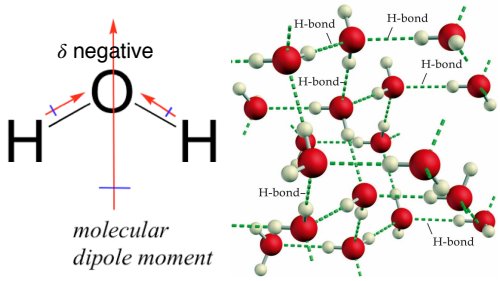
[12] A short-ish talk on stochastic processes: Coping with the Noisy Nature of Life in Teaching & other philosophical ruminations

[13] Mice, people, and dinosaurs JacksonLab website / Mouse models of human disease: An evolutionary perspective 2016

reproduction (meiosis and fertilization) and some molecular and cellular tools for experimentally manipulating and characterizing socio-cellular processes. Finally, we will consider the roles of model systems that have revealed common socio-cellular and developmental processes.

Molecular interactions: Molecules are composed of atoms, linked through covalent bonds. As you know from physics and chemistry[14], atoms consist of a tiny positively charged nucleus and a much larger negatively charged electron cloud. The result is that atoms display a small fluctuating electric field. As two atoms approach each other, they influence each other's electric fields. The result is that all atoms (and molecules) attract each other via electrostatic interactions known as London Dispersion Forces (LDFs). These are relatively weak forces, and can be readily overcome or overwhelmed when molecules or regions of molecules are consistently positively or negatively charged, as occurs in both nucleic acids and proteins which have acidic (groups that lose a proton, H+) or basic (groups that gain a proton) from the solvent, that is water. Interactions based on LDFs will be influenced by molecular shapes; they will be stronger when molecular surfaces are "complementary".

Covalent bonds, which occur when electrons are "shared" between atoms, are much stronger. Due to the intricacies of quantum mechanics, even though all atoms are neutral, the atoms of different elements differ in what is known as their electronegativity. The result is that when a covalent bond is made between atoms with sufficiently different electronegativities, for example oxygen (3.44) or nitrogen (3.04) and hydrogen (2.20) or carbon (2.55), the resulting bond is polarized, with a partially positively charged face associated with the atom with lower electronegativity and a partially negatively charged face associated with the atom of higher electronegativity (↓). Molecular interactions between such partially charged regions or between different regions of larger (macromolecules) are stronger than LDFs but weaker than those associated with acidic or basic features of a molecule; they play a critical role in determining the structure and physical properties of solutions via solvent-solvent, solvent-solute, and solute-solute interactions. Biological molecules are typically found dissolved (to various extents) in water H_2O (the solvent). Inspection (←) of the water molecule reveals that its two O-H bonds are both polarized. This bond polarization, together with its roughly tetrahedral geometry, means that a single water molecule can interact with four surrounding water molecules (or with other molecules with polarized bonds) through electrostatic interactions known as H-bonds. When a molecule that cannot make H-bonds with water molecules is inserted into water the result is a constraint on (ordering of) the orientation of water molecules - a thermodynamically unfavorable decrease in the entropy of the system.[15] Entropic effects are responsible for the separation of water and oil molecules (which cannot make H-bonds). While the system looks to be more ordered macroscopically, with separate oil and water layers, it is more disordered at the molecular level than it would be if the oil molecules were interspersed, with each oil molecule surrounded by a "cage" of water molecules. When oil and water separate, the water molecules are free to assume many more orientations with respect to



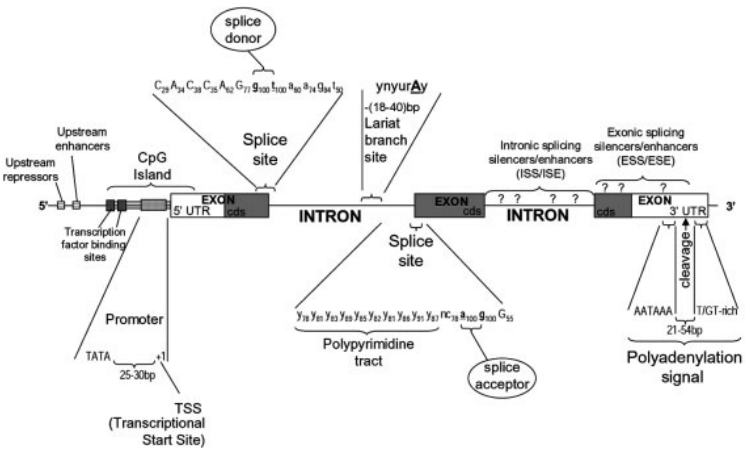
[14] Of course we recommend reading CLUE (Chemistry, Life, the Universe and Everything) if you want an accessible refresher.

[15] remember, thermodynamically favorable reactions are associated with an increase in entropy (ΔS)

one another. The "hydrophobic" effect is a key driver in determining a polypeptide's three-dimensional folding pattern. The presence or absence of polarized covalent bonds as well as acidic or basic (fully charged) groups in a molecule influences the three-dimensional structure of, and the interactions between, macromolecules - it underlies the folding of proteins, the ability of DNA-binding proteins (e.g. transcription factors) to recognize specific nucleotide sequences, the insertion of lipids and lipid-modified proteins into membranes, and many other processes. Understanding the atom bases of molecular interaction makes possible plausible predictions about a molecule's behavior (is it water or lipid soluble, does it form insoluble aggregates) and how specific mutations at specific position might alter its folding or interactions with other molecules.

Genes: A gene is a stretch of DNA that contains information accumulated through mutation and evolutionary mechanisms.[16] Prokaryotes (bacteria and archaea) typically contain a single copy of each gene. Their genomes (the totality of a cell's genetic information) are organized in a single circular DNA molecule, although there can be other circular DNAs (plasmids) present. Eukaryotic cells typically contain either one (haploid) or two (diploid) copies of each gene.[17] The eukaryotic genome is organized in a number of linear DNA molecules, with the number characteristic of each species. Such cells are said to be diploid. In the typical sexually reproducing species, one of the two copies of each gene is derived from one parent and one from the other; the exception being in species that use chromosome-based sex determination. When the two parental copies are the same, the same alleles, the organism is said to be homozygous for the gene or genetic locus; if they are different the organism is said to be heterozygous for that gene. An organism can be homozygous for some genes and heterozygous for others. Different alleles can be expressed differently and encode different versions of the gene's product(s).

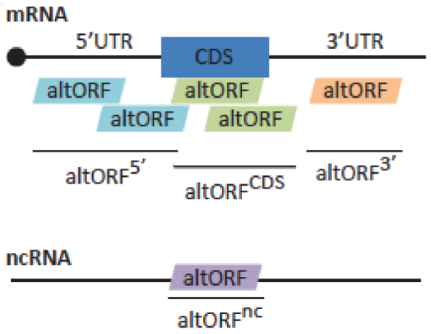
The typical (eukaryotic) gene has a distinct anatomy (↓). Prokaryotic genes are typically simpler and more complex; they lack intronic sequences and are often organized into "cistrons"–DNA sequences that encode multiple proteins, typically involved in a common metabolic or behavioral process. Understanding the anatomy of a gene is useful in that it allows us to make plausible predictions based on the effects of specific mutations, although it should go without saying that the actual effects of a particular mutation will always need to be established experimentally. In the case of both prokaryotic and eukaryotic genes, the gene is associated with nearby regulatory regions, referred to as promoters. Promoters are gene sequences that are recognized by specific transcription factor proteins; when bound, the transcription factor protein interacts with other proteins leading to the stable binding



[16] Exactly what is a gene can get complicated - see Portin & A. Wilkins (2017). The evolving definition of the term "gene".

[17] During the cell cycle, DNA replication will lead to a doubling of the number of copies of each particular gene. In some cases, DNA replication can occur without cell division,, leading an increase in copy number (or ploidy).

and activation of DNA-dependent, RNA polymerase and the synthesis of an RNA. This process is often termed gene expression. In eukaryotes, gene expression is complicated by the presence of multiple promoter and enhancer elements that can influence one another, influences mediated by various promoter and enhancer binding proteins. The result is the formation of a protein-DNA complex that can act to inhibit or enhance gene expression. While promoter regions are typically found adjacent to the main gene body, enhancers can be located far "up-stream or down-stream" of the gene body (the transcribed region); moreover, they can be in either orientation with respect to the gene body. The gene body contains the information to specify the RNA synthesized, and its subsequent processing, whether the RNA serves as a structural (e.g. ribosomal) RNAs, functional (e.g. tRNAs), regulatory "non-coding" RNAs (e.g. shRNAs), or as a messenger RNA (mRNA) that, in association with a ribosome, directs the synthesis of a polypeptide. In eukaryotes the gene body is typically divided into exons (coding regions) and introns (intervening regions). It is worth noting that many non-coding regions of the DNA direct RNA synthesis (→) and that some (many) RNAs can be both; recent studies have identified regions in what were thought to be non-coding RNAs that express micro-open reading frames, often using alternative start codons.[18]



modified from Samandi et al., 2017. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins

Part of the complexity in understanding the rolls and regulation of genes, particularly eukaryotic genes, is that their regulatory regions, particularly enhancers, can exist within coding regions–a gene can contain parts of other genes. For example, overlapping genes can be found in the two strands of a DNA molecule; one estimate is that ~25% of polypeptide encoding genes overlap.[19] A classic example is adult lactose tolerance, a trait found in human populations with a long history of raising domesticated animals from which milk, which is rich in the milk sugar lactose, can be harvested. In humans, the utilization of lactose depends upon expression of the *LCT* gene, which encodes the protein lactase. The *LCT* gene (and the lactase protein) is expressed in infants, who are dependent upon mother's milk for their nutritional needs. Tolerance of lactose in the diet in adults is associated with the continued expression of the *LCT* gene. Expression of *LCT* is turned off as children age due to the action of an negatively acting enhancer, located within an intron of the *MCM6* gene, some ~14 kilobase pairs (kbs) "upstream" of the main body of the *LCT* gene. Human adult lactose tolerance arises from a mutation that inactivates this negatively-acting enhancer.[20] The result is that *LCT* expression continues in adults. This mutation if widespread in populations in which adult lactose tolerance is common, apparently due to positive selection.[21]

Another complexity of eukaryotic genes is that they often have multiple distinct regulatory sequences (promoters and enhancers) and exons can be "spliced" (included or excluded in the final

---

[18] Dong et al. 2023. Small Open Reading Frame-Encoded Micro-Peptides: An Emerging Protein World. *International Journal of Molecular Sciences*, *24*, 10562 and Genes – way weirder than you thought (bioliteracy 2018)

[19] Nakayama et al., 2007. Overlapping of Genes in the Human Genome.

[20] Lactose digestion and the evolutionary genetics of lactose persistence ; The Evolution of Lactose Tolerance — HHMI BioInteractive Video

[21] World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection.

RNA) in multiple ways, leading to structurally and functionally distinct gene products, i.e. polypeptide. Which DNA regulatory elements are used and to what effect depends upon which regulatory factors, primarily DNA-binding, transcription-regulating proteins, as well as proteins and other molecules that may interact with them, are present in the cell, their concentrations (influenced by post-translational folding and degradation pathways) and various interactions and modifications that can control where they are located within the cell, e.g. are they concentrated in the nucleus or in the cytoplasm. DNA can also be modified by DNA methylation and the accessibility of specific DNA sequence regions can be modified by interactions with structural proteins, such as histones.

The origin(s) of genes: In the genome, genes are interspersed with non-coding, non-gene intra-genic regions. It was originally thought that only genes were used to direct the synthesis of RNA molecules, but recent studies reveal that most of the genome is transcribed. In some cases, the ability of these, typically short, transcripts (RNAs) to encode polypeptides has been identified based on their association with ribosomes, a technique known as Ribo-SEQ. Some of these RNAs use non-conventional start codons, that is, codons other than AUG.[22] It appears that originally "noisy" transcripts can be captured and conserved via natural selection based on their effects on reproductive success.[23] New genes have appeared *de novo*[24]*;* many of these *de novo* genes appear to have become essential rather quickly.[25] A number of putative *de novo* genes, not found in related primates, have been identified in humans.[26]

Once a gene exists it can be altered (the original DNA sequence changed) by mutation and meiotic recombination. Different versions of a gene present in a population are known as alleles. Organisms of the same species) often differ in the alleles they contain, but they generally have the same genes. A gene can disappear from a population or give rise to diverging copies of itself through the processes of genome dynamics. Simakov et al[27] have reported that homologous genes are often found together, but in different arrangements along chromosomes in various distantly related species, due to processes that result in genomic reshuffling.

---

[22] see Olexiouk et al., 2017*. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling and Leong et al., 2022. Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures

[23] Andrews & Rathnagel. 2014. Emerging evidence for functional peptides encoded by short open reading frames

[24] see: Schlotter. 2015. Genes from scratch – the evolutionary fate of de novo genes Schmitz & Bornberf-Bauer. 2017. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. Van Oss & Carvunis. 2019. *De novo* gene birth

[25] see New genes in Drosophila quickly become essential and The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo.

[26] Wu et al., 2011. De novo origin of human protein-coding genes

[27] Simakov et al 2022.Deeply conserved synteny & the evolution of metazoan chromosomes. *Science advances*, *8*, eabi5884.

A number of studies, beginning with the classic Luria-Delbruck experiment,[28] indicate that mutation, recombination, deletion, gene duplication and loss, and the de novo appearance of genes occur stochastically based on the molecular nature of DNA, the various molecular mechanisms active in cells, and environmental effects due to chemicals and radiation, and not because the organism "wants or needs" specific mutations. At the same time, how an animal develops and behaves can impact the selective value of particular mutations, producing various types of feedback loops.

Changes in gene sequence (or gene number) can have phenotypic (visible, measurable) effects; they can range from the non-detectable to the lethal. If these phenotypic effects impact reproductive success they can be “selected”, evolutionarily. Alleles with negative effects will tend to decrease or disappear from the population, and those with positive effects will tend to increase. When an allele becomes universal (in a population), it is said to have become "fixed". But evolutionary processes can produce complex result. For example, a trait that is beneficial when uncommon in a population can become less beneficial or even deleterious when common, so "selective pressure" can change over time. Similarly, a allele that has a negative effect, may have a different effect in the context of different alleles of other genes. Because genes are linked together on chromosomes, selection for or against an allele of one gene can influence the frequency of alleles of neighboring genes. Stochastic processes, such as genetic drift, population bottlenecks, and founder effects can influence the frequency of specific alleles within a population. Particularly in small populations, as often occurs during speciation, beneficial alleles can be lost and deleterious alleles preserve by chance. These principles apply both to the cells within a multicellular organism (somatic selection) as well as organisms within a population.

Questions to answer and ponder:

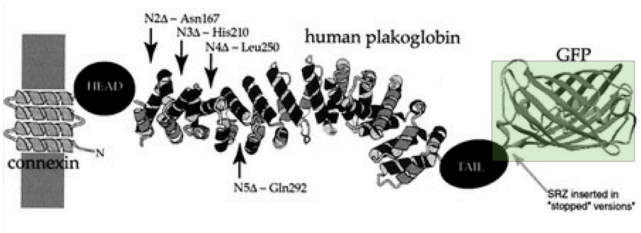
1. What methods and "rules of thumb" could you use to identify all of the genes in a newly described genome? How would you confirm your predictions?
2. Provide a plausible explanation (and sketch) for why an enhancer's orientation with respect to the gene body generally does not influence its activity.
3. Why isn't adult lactose tolerance universal in humans? How might a mutation lead to lactase expression in adults?
4. How might the expression of overlapping genes located on different strands of a DNA molecule influence each other's expression?
5. How might an animal's behavior influence selective pressures?

Implications of the modular structure of polypeptides and proteins: As in the case of genes, there is a functional anatomy of polypeptides and proteins. First, a polypeptide is synthesized in a linear manner, like the gene sequence that encodes it. As the N-terminus of the polypeptide emerges from the ribosome it starts to fold. The first step in this folding process has been described as a molten globule. Hydrophobic and hydrophilic residues (linked, of course, in a single chain) will fold so as to reduce or maximize, respectively, interactions with water. At this point various "secondary structures", primarily α-helices and β-sheets will form; these serve to maximize the H-bonding interactions associated with the chain's peptide bonds. In the third (tertiary) level of structural interactions, the folding polypeptide takes on a defined, albeit dynamic three dimensional structure. Finally, in proteins composed of multiple

[28] see. Luria & Delbruck 1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, *28*, 491 and Fusco et al. 2016. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nature communications*, *7*, 12760.

polypeptides, there is quaternary interaction that reflect how these polypeptide chains interact to form a functional protein. Students, instructors, and textbooks often confuse the terms polypeptide and protein.[29] The major class of directly encoded gene products are polypeptides. Proteins are defined in functional terms; they may be composed of one or multiple copies of one or multiple types of polypeptide encoded for by different genes, and can include non-polypeptide "co-factors".

The value of thinking in terms of polypeptide/protein structural "anatomy" is that it enable us to make plausible predictions as to the effects of mutations and to plan (with a reasonable hope of success) the construction of chimeric proteins. In the first case, we can predict how a particular type of mutation (amino acid substitution or alteration, e.g. deletion, addition, or frame-shift mutation) will influence polypeptide folding or protein function. Based on a large body of observational data, it is clear that polypeptides typically fold in discrete domains. This makes it possible to add a domain, by molecular engineering techniques, to the beginning or end of a polypeptide while retaining both the polypeptide's original functional activity and the activity of the added domain, whether as a targeting sequence (membrane anchor, nuclear localization or exclusion signal), a fluorescent domain, an antibody-recognized domain (an epitope), or a particular domain with a particular biological activity, such as a transcriptional activator or repressor domain.[30] Of course, as with pretty much every biological prediction, the complexity of these systems demands experimental confirmation of our predication. The rules of polypeptide folding and protein assembly are such that AI tools (such as AlphaFold) can be used to make plausible (and often accurate) predications of three-dimensional structure based on primary polypeptide sequence.[31]



On considering molecular interactions: Molecular interactions underlie most of what is going on in a cell. The extent to which a particular transcription factor protein is bound to a specific site in the genome, or to other proteins, whether it is targeted for activation by a protein kinase, or inactivation by a protein phosphotase or marked for degradation by the addition some other group (often ubiquitin), depends in intermolecular interactions. The formation of these interactions is dependent upon the surfaces of the interacting molecules and the degree to which they form H-bonds, complementary electrostatic interactions, van der Waals interaction, or hydrophobic pockets. The strength of these interactions is the sum of the enthalpic and entropic changes involved. In chemistry, such enthalpic and entropic effects are typically measured in bulk using purified components; something that is rarely possible or relevant to biological systems, particularly when the structures of the molecules involves can be influenced by allosteric effectors, various associated proteins, and molecular and protein

[29] bioliteracy: When is a gene product a protein when is it a polypeptide?

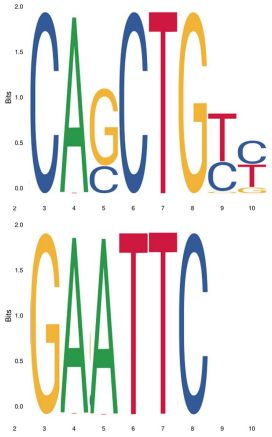
[30] The figure is from one of my scientific projects: Membrane-anchored plakoglobins have multiple mechanisms of action in Wnt signaling

[31] Pak et al., 2023. Using AlphaFold to predict the impact of single mutations on protein stability and function

chaperones, such as the high intracellular concentration of ATP found in cells[32], that act to stabilize protein structures.

A second, often overlooked, aspect of molecular interactions is the mechanism by which interactions are reversed. In essentially all cases, the disruption of an interaction is driven by kinetic energy delivered through thermal (temperature-dependent) collisions. The stronger the interaction, the more energy is required to break it. In probabilistic terms, since collisions are continuous but unpredictable (stochastic), the stronger the interaction between two molecules the less likely (per unit time) collisions with surrounding molecules will provide sufficient energy to break their interaction. The result is that exactly when a particular bond or interaction will break is unknowable, even when the "strength" of a bond or an interaction is known.

Consider how the stochastic nature of these processes influences cellular behavior. There are typically one or two copies of each gene. A particular gene may have a few sequences to which the proteins that regulate, either positively or negatively, its expression interact with high affinity. Typically the number of these proteins in a cell is low, measured in the thousands to millions of copies.[33] At the same time there are often millions to billions of base-pairs of DNA. While there are proteins, such a most restriction endonucleases, that bind with high specificity to specific DNA sequences (e.g. EcoRI - bottom panel ←), transcription factor proteins are different, their targeted binding sequences are more flexible (the top panel provides an example). Within their binding target, some sites are essential, other sites can vary. The result is that there are many potential binding sites in a genome and these sites can vary in terms of binding affinity/stability (resistance of the bound protein to being knock off by thermal motion). A particular transcription factor protein will interact with a number of binding sites with different functional properties, e.g. regulating different genes. Moreover, given the physical size of a protein, and the geometry of its DNA interactions, binding can be influenced by post-translational modifications, the presence of other proteins, and the concentrations of these molecules within the cell (or within the nucleus, for a eukaryote). As it concentration increases, the frequency at which a transcription factor protein will be found bound to lower affinity sites will increase. Since whether a gene is actively expressed or repressed can influence, in what may be a difficult to reverse way, the future state of the cell, the stochastic (noisy) nature of gene regulation can have complex effects (as we will see).[34]



Questions to answer and ponder:

1. What might make the effect of "turning on" a gene on the future behavior of a cell difficult to reverse?
2. How (in general) will gene expression change following the initial "turn on" of the expression of a gene for a particular transcription factor?

---

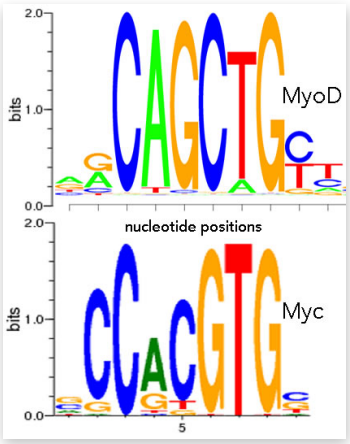
[32] Takaine et al., 2022. High and stable ATP levels prevent aberrant intracellular protein aggregation in yeast

[33] WHAT ARE THE COPY NUMBERS OF TRANSCRIPTION FACTORS?

[34] Making sense of noise.

Mutations, Alleles, and Morphs: One way to look at alleles is from a functional perspective. This was the approach taken by Herman J. Muller (1890-1967) working with the fruit fly *Drosophila melanogaster.* Classical geneticists exploited the small number (four) chromosomes and the highly polyploid nature of its salivary gland chromosomes.[35] They were able to generate, and visually confirm using a microscope, the location of chromosomal rearrangements, duplications, and regional deletions. Based on visible phenotypes in the context of gene deletions and duplications, he was able to place mutations into distinct groups or morphs.[36] In this scheme, a mutation that led to a complete loss of function (LoF) was termed amorphic; a hypomorph had reduced activity, while a hypermorph had increased activity. An antimorphic mutation was antagonistic to the function of the normal or wild type version of the gene while a neomorph displayed a new "function". This approach is over-simplified since a particular gene (gene product) generally has multiple functions and interaction partners and multiple genes are involved in any particular trait. A particular mutation may influence these functions differently. An particular allele could be hypomorphic for one trait and antimorphic for another. Nevertheless, it can be a useful context in which to consider the effects of mutations, particularly in terms of protein function.

How is it possible that a mutation can generate a new function? Consider the gene encoding the transcription factor MyoD, a protein that regulates the differentiation of skeletal muscle cells. Mutations in the *MyoD* gene are associated with embryonal rhabdomyosarcoma, a cancer of skeletal muscle. One neomorphic mutation changes the leucine normally present at position 122 to an arginine; this mutation alters the DNA sequences to which the MyoD protein binds.[37] The "wild type" MyoD protein binds to a consensus sequence (top →); the mutated protein binds better to sequences similar to those recognized by the transcription factor Myc (bottom →). Myc regulates genes associated with active cell division. The result is that a gene product (MyoD) that normally inhibits cell division and encourages the formation of muscle cells, now binds less well to these differentiation-associated genes and turns on genes that favor aberrant cell division – a key feature of cancer cells. The mutation is neomorphic because the mutated MyoD protein (known as MyoDAla$_{122}$→Arg) has a new function, activating cell division. It might also be considered hypomorphic or amorphic for its original function, inducing muscle differentiation.



Certain alleles have been associated with one or more specific phenotypic traits, often diseases or the probability of coming down with a specific disease or syndrome. If an allele-associated trait is visible when the locus is heterozygous for that allele, the allele is said to be dominant with regard to that trait. On the other hand, if the trait is not apparent when the locus is heterozygous, but is visible when the locus is homozygous for the allele, it is referred to as recessive. An amorphic allele can be dominant, a behavior known as haploinsufficiency, arising because one copy of the gene does not produce a sufficient amount of the gene product, or it can be recessive, if one functional copy of the gene is sufficient to produce the wild type phenotype. Of course things are often more complex. There are times when the trait observed when an organism is heterozygous for a

[35] Banding patterns in Drosophila melanogaster polytene chromosomes correlate with DNA-binding protein occupancy.

[36] Turning randomness into meaning at the molecular level using Muller's morphs.

[37] from Myc and MyoD and Deep Sequencing of MYC DNA-Binding Sites in Burkitt Lymphoma

genetic locus is different from the trait display by organisms homozygous for either allele. In addition, many alleles display what is known as incomplete penetrance and expressivity, a situation dependent upon genetic background effects and historic (evolutionary) events. It is worth remembering that every trait is influenced by and dependent upon many genes and the product(s) they encode, and that most genes, and their product(s) influence many traits. Finally, an allele can be dominant for one trait and recessive for another. Many students emerge from introductory genetics courses not realizing that the traits Mendel worked on were the result of extensive inbreeding. Most phenotypes are complex, and involve the contribution of many genes.

Some molecular genetic terminology: A missense mutation leads to the replacement of one amino acid by a different amino acid in a polypeptide. Mutations that do not change the encoded polypeptide's amino acid sequence but change the DNA sequence are known as synonymous mutations; such mutations produce what are known as single nucleotide polymorphisms (SNPs). While most SNPs do not produce a phenotype, there are cases where different tRNA genes are expressed. If a SNP changes a common to a rare codon, that is a codon for which the appropriate tRNA is missing or present at low concentration, it can produce a "codon-bias" effect.[38] If a codon is rare, the ribosome can "stall" along the RNA "waiting" for a rare tRNA to bind. This can lead to premature termination or a frame shift in the polypeptide. A non-sense mutation leads to stop codon replacing a codon encoding an amino acid; this can result in a truncated polypeptide. The effect of a non-sense mutation will depend upon where it occurs within a gene. Another type of mutation, known generically as indels for "insertion/deletion", leads to the insertion or deletion of one or more nucleotides from the gene sequence. Such mutations these can lead to a range of effects, from inserting or deleting amino acids to changing the reading frame. In genes with exons and introns, splice site mutations that disrupt the sequences involved in recognizing and removing introns from newly synthesized RNAs, again producing mRNA that encode altered and/or truncated polypeptides (or non-coding RNAs). Given that many gene products have roles in multiple processes, a single mutation can influence many traits. Such mutations are termed pleiotrophic.[39] Genetic and molecular studies tend to focus on genes with simple phenotypes.

Questions to answer and ponder:

1. A hypermorphic allele produces a dominant phenotype. Make and justify a phenotypic prediction for a hypermorphic/gene deleted animal.
2. A hypomorphic allele produces a dominant phenotype. Make and justify a phenotypic prediction for a hypomorphic/gene duplicate animal.
3. Purpose a model by which a spice site mutation generates an anti-morphic allele.

"Simple" asexual cell division: Ever since LUCA, each "new" cell has been derived from a pre-existing cell. Typically, this process involves cell growth followed by division resulting in two cells. These two cells are generally extremely similar, but they are not necessarily identical. Each contains a full copy of the original cell's genome, its DNA, but because DNA replication is not 100% accurate, the two

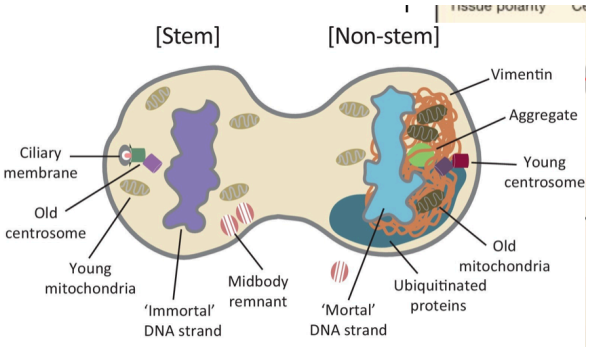
[38] see Plotkin & Kudla 2011 Synonymous but not the same: the causes and consequences of codon bias.

[39] Pleiotropy: One Gene Can Affect Multiple Traits

cells can differ due to replication-associated mutations. A more reproducible difference involves the DNA itself. Inside cells, DNA is subject to post-replication modifications, primarily methylation of C and A residues. During the course of DNA replication, the two strands that make up the molecule separate and two new strands are synthesized. The resulting replicated DNA molecules are composed of an old and a new strand. Because errors that occur during DNA replication involve the newly synthesized strand, the old (original) DNA strand can be used to make repairs, assuming that the cell's DNA error repair systems can distinguish old and newly synthesized DNA strands. One way to identify the old DNA strand is that it is methylated, while the newly synthesized strand is not. It takes time following replication before methylation of the new strand "catches up". By recognizing the old DNA strand and using it as the template to repair the newly synthesized strand, DNA mutation repair systems can bias repair to return the newly synthesized replicated DNA to its original sequence.

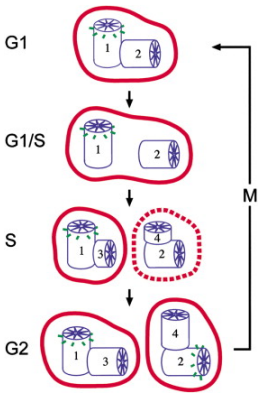
Another important implication is associated with the asymmetry of DNA methylation of newly synthesized DNA. Genes occur are present on both strands of the original DNA molecule and DNA methylation can influence gene expression (→). The methylation state of genes in the cells that arise after cell division will not the same, at least until DNA methylation of the newly synthesized strand "catches up" with the old methylated strand. The result is that there can be subtle (or not so subtle) differences in gene expression between the two cells. When the genes affected encode proteins that regulate gene expression, these differences can cascade and in some cases may be become effectively irreversible.

There is growing evidence for other intrinsic asymmetries that arise during cell division; some serve to sequester damaged cellular components, such as defective mitochondria and denatured and aggregated molecules and molecular complexes that may otherwise be difficult to eliminate using degradative enzymes (→).[40] The result can be cells that display quite different behaviors. Perhaps the most dramatic example is what is known as a stem cell. In a typical stem cell, after cell division, on cell can continue to divide essentially without limit while the other cell goes on to differentiate, with its ability to divide lost or significantly limited. To achieve an asymmetric distribution of damaged cellular components, asymmetrically localized cytoplasmic factors and environmental cues may be involved. As an example, you may have been introduced to centrosomes and centrioles and their roles in organizing microtubules, particularly in the mitotic animal cells. The centrioles come in pairs, typically oriented at a right angle to one another. The two can be distinguished by the presence of "distal" appendages on one (the mother or older centriole) and absent from the daughter centriole (←).[41] During the cell cycle, the two centrioles separate from one another, a new centrioles are assembled. One pair of centrioles will contain the mother, the other the daughter. Surrounded by the microtubule organizing centrosome, there will be a centriole pair at each pole of the



[40] Moore & Jessberger 2016. Creating Age Asymmetry: Consequences of Inheriting Damaged Goods in Mammalian Cells

[41] Image Stearns, T. (2001). Centrosome duplication: a centriolar pas de deux. *Cell*, *105*(4), 417-420.



mitotic spindle. When the cell divides, one cell will inherit the centriole pair that includes the original mother centriole, while the both centrioles in the other cell will both be newer. We now have a basis for a cellular asymmetry between the two resulting cells, and this asymmetry exists before the cell divided.

Sexual reproduction: Compared to simple cell division (mitosis and cytokinesis), the process of sexual reproduction is both mechanistically and evolutionarily more complex. Two morphologically different cells, generally but not always from two different organisms, have to find and fuse with one another. Such cooperation requires these cells to recognize one another as appropriate fusion partners. Only gametes of different "types" can fuse successfully. In animals, there are two dimorphic types of gametes. By convention, the larger and typically immobile of the two is produced by organisms known as female (♀). In females meiosis typically generates a single gamete, known as an oocyte, and in its mature form an egg, and three non-viable mini-cells, known as polar bodies. The smaller, and generally mobile type of gamete is produced by a male (♂) organism and is termed a sperm. Male meiosis produces four gametes. There are organisms that can produce both types of gametes (eggs and sperm) either at the same time, or at different points in their life-cycle, these organisms are termed hermaphrodites.[42] The difference in the size of the gametes (together with other factors) means that the two sexes can have discordant investments in reproduction. Typically, more, often much more, resources are required to build an egg than a sperm. This difference can become even more pronounced in terms of parental investment, a fact that underlies sexual selection, one of the key aspects of modern (Darwinian) evolutionary theory.[43]

Each gamete, egg or sperm, contains one and only one copy of each chromosome present in the original diploid cell. Historically, chromosomes were numbered based on their apparent size in histologically stained specimens (→). In humans, the largest chromosome, chromosome 1, contains ~250 million base pairs of DNA and over 2000 polypeptide-encoding genes, while the smallest, chromosome 22 contains ~52 million based pairs of DNA and ~500 polypeptide encoding genes.[44] Homologous chromosomes are also defined by the types and order of genes found along their length, known as synteny. Human chromosome #5 contains different genes than are found on chromosome #6. Moreover, the maternal (from the mother) version of each chromosome may contain different alleles compared to those found in the paternal (from the father) version. The maternally and paternally derived chromosomes are known as homologs. In the human, an organism's sex is normally determined by the 23rd chromosome pair. When XX, the organism is female, when XY the organism is male. To balance the "over-representation" of genes on the X chromosome in females, one or the other

---

[42] There are situation when the two cell involved in sexual reproduction are morphologically similar, in which case the "sexes" involved (and there can be many) are not termed male / female, but are different mating types. Only haploid cells of different mating types fuse to form a "new" diploid organism.

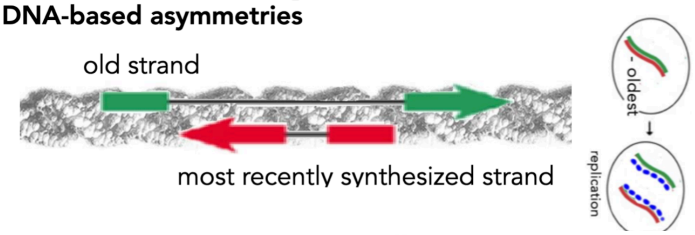
[43] How Darwin arrived at his theory of sexual selection & Mate choice and sexual selection: what have we learned since Darwin?

[44] We are only discussing polypeptide-encoding genes because it remains unclear whether (and which) other transcribed regions are genes, or physiologically significant.

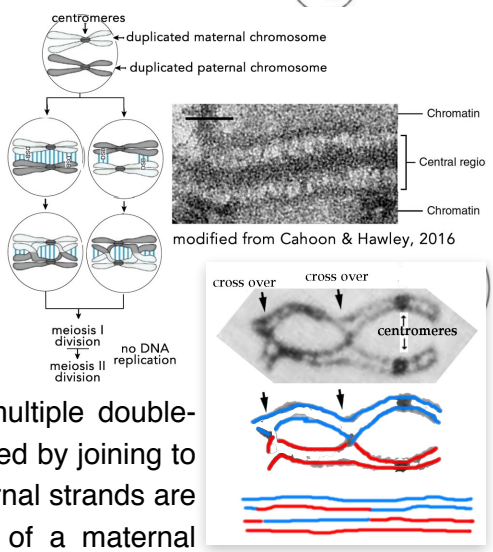
of these chromosomes is "inactivated" randomly during embryonic development. The result is that both males and females have one active X chromosome per cell, but females are a mosaic of clones composed of cells with either an active maternally- or an active paternally-derived X chromosome. All male cells have an active, maternally-derived X chromosome.

HERE

In asexual reproduction, whether of a cell or during the development of a multicellular organisms, the various versions (alleles) of the genes present are inherited together. Genetic variation arises from mutation, and is restricted to the cell/organism in which the mutation occurs. A population of cells/organisms, can be considered an assembly of distinct clones. The cooperative (social) process of sexual reproduction breaks this isolation. We will consider sexual production from a molecular and cellular perspective, but it is a complex subject that includes organismic morphology, behaviors, and social interactions - topics best considered elsewhere.



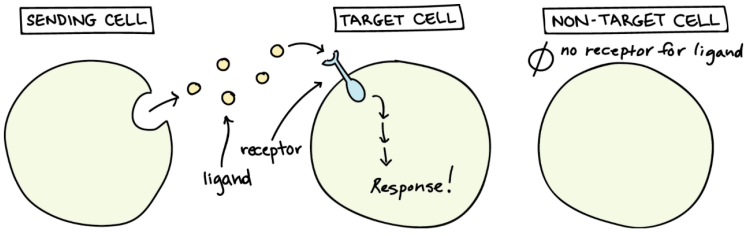
During asexual reproduction, replicated chromosomes are partitioned, one copy to each of the two cells produced. The cellular start of sexual reproduction involves meiosis, a modified form of mitosis – a single round of DNA replication is followed by two rounds of cell division, producing four haploid cells, two of which fuse (fertilization) to produce a new organism. During the first step of meiosis, following DNA replication, replicated maternal and paternal homologous chromosomes pair and their DNA sequences align closely. At this stage, the effects of large DNA deletions, inversions, or regional inversions can be visualized. Typically multiple double-strand breaks occur in the aligned strands; these breaks are healed by joining to the broken ends of another DNA strand. When maternal and paternal strands are involved, the result is known as recombination (→); a region of a maternal chromosome is swapped for a region of the paternal chromosome and visa versa. The effect of recombination is that alleles that were originally linked on the same chromosome are now associated with other alleles - alleles are no longer inherited together.



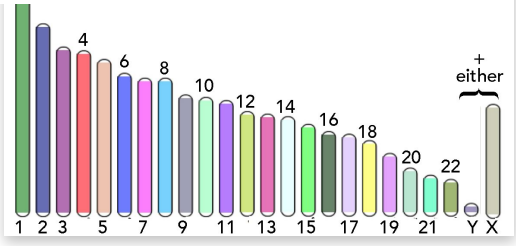
The second shuffling event in meiosis is the stochastic and independent segregation of replicated (and recombined) maternal and paternal chromosomes at the first meiotic division. When the chromosome pairs line up at the first division plane, the orientation of the maternal/paternal centromeres on each homolog chromosome pair point independently of one another. At this first meiotic division, the paternal and maternal centromeres remain attached, but separate from each other. The result is that two cells produced by the meiosis I division will inherit different combinations of maternal and paternal chromosomes. Independent segregation results in the further shuffling of alleles on different chromosomes. The final shuffling event occurs when two gametes, egg and sperm fuse to produce a new organism, genetically distinct from both parents. The fusion event, known as fertilization, is the most discontinuous event in the process of (sexually reproducing) life. Even so, fertilization does

not represent a true discontinuity, at least with respect to life – both sperm and egg are alive, as is the fertilized egg.[45] In a critical sense life (in the post-LUCA world) never begins – it continues and is transformed. That said, fertilization is the start of a new, genetically distinct organism. The fused cell that results from fertilization is known as a zygote. Through somatic (asexual) cell division the zygote (fertilized egg) will develop into an adult, composed of diploid cells. The cells of the adult that produce gametes are known as germ cells, and together are known as the organism's germ line. The rest of the adult is composed of somatic cells, cells that divide (if they divide) by mitosis. Meiosis is restricted to germ line cells and gamete formation.

Signaling systems, sigmoidal responses & threshold effects: Cells (and the organisms they form) are inherently responsive systems. They respond to the physical properties of their environment through changes in gene expression, protein stability and concentration, location, and activity, and resulting behaviors. A good example is explosion of algal endosymbionts by coral (cnidarians) when water temperatures rise, which leads to coral bleaching and reef disruption. Cells can respond in specific ways to many different types of molecular changes in their environments. Here, we are primarily concerned with social signals, signals that arise from other cells (or organisms) of the same type. Similar mechanisms are used to monitor and regulate variations in a cell or organism's internal states. We will consider these systems generically, and then in specific contexts. A typical signaling system consists of a signaling molecule, a receptor, and an effect on the system. The nature of the receptor will be determined by the nature of the signal and cell structure. The cell is surrounded by a lipid bilayer membrane. Typically a highly hydrophilic (water soluble) signal molecule will not easily pass through such a membrane. In most cases, the receptor for a water-soluble signaling molecule will be a membrane protein with three distinct structural domains - an extracellular, signal binding domain, a transmembrane (largely hydrophobic) domain, and an intracellular domain that changes in some way upon signal-receptor binding. There are also systems in which i) the signaling molecule can pass through the membrane without assistance or ii) can be endocytosed and then cross into the cytoplasm through secondary effects. The same signaling molecule can have different effects on different cells depending upon the exact types of receptors expressed and present, various co-receptors, and "down stream" factors. Activating one signaling system can influence the response of others, since downstream effects can interact. For example, activating signaling system A could lead to post-translational modifications that alter the ability of the receptor for signaling molecule B to bind or respond to the binding of B.

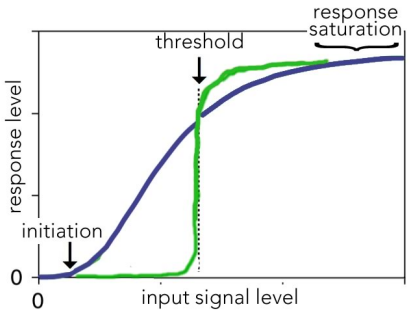


Not all cells can "hear" a particular chemical message. In order to detect a signal (that is, to be a target cell), a neighbor cell must have the right receptor for that signal. When a signaling molecule binds to its receptor, it alters the shape or activity of the receptor, triggering a change inside of the cell. Signaling molecules are often called ligands, a general term for molecules that bind specifically to other molecules (such as receptors). Image credit: Khan Academy.



[45] In fact, there are examples of cell fusion within organisms - as an example, during the development of skeletal muscle, muscle precursor cells fused to generate large multi-nuclear cells, known as myotubes.

A typical effect of a signaling molecule-receptor interaction is to activate or inactivate some process that leads to changes in enzymatic activities, intracellular localization of a protein (e.g. cytoplasmic to nuclear), and changes in gene expression (inhibiting or activating a repressor or activator of the expression of specific genes). Cells generally contains systems that serve to restore the cell to its pre-signal state. These systems are typically constitutively active; they may act to degrade or remove the signal molecule or reverse its effects on down-stream components. Such back-reactions explain why the response to a signal is typically a sigmoidal function of signal concentration (→). For example, when a signal induces the phosphorylation of a cytoplasmic effector protein, there is generally a constitutively active phosphotase that removes the phosphate group, returning the system to its pre-signal state. Only when the extent of the forward reaction overcomes this back reaction will the system begin to respond (the take-off point). The response will increase until it reaches in maximal level. When a system's take off and maximal response points are close together (in terms of signal concentration) the result is a "threshold effect". Threshold effects act as on-off switches and are widely used in developing and differentiating systems.



One point about cellular responses to signals (and threshold effects) is that the response can be either adaptive and temporary, or can lead to a change in the state of the cells, changes in which genes are expressed, which structures appear, and how the cell behaves and response to signals. Such changes, in part because they involve cascades of regulatory events, can be effectively irreversible or difficult to reverse. They can involve both positive and negative feedback interactions that alter the cell's regulatory state, and its future responses signals.

Modeling regulatory networks: In the best case, to understand the effects of a signal molecule on a cellular system, we would need to know many things, such as which genes are expressed, how long RNA and polypeptide synthesis and processing takes, which proteins are active, the concentration(s) proteins and other molecules, their distribution with the cell (nucleus, cytoplasmic, membrane-associated, etc, and their affinities for other molecules that influences how often they interact with other molecules and how long those interactions persists. Unfortunately, we never really have a full understanding of the system. There may be interacting components, perhaps present at low concentrations, that influence the behavior of the system. Moreover, the nature of interactions is stochastic, these are mistakenly termed random, but are actually noisy but lawful processes (like radioisotope decay) that, in biological systems, arise from the small numbers of molecules involved together with what is known a Brownian or thermal motion. If you have the time, here is a recent talk I gave on the subject [Coping with the Noisy Nature of Life].[46] It is, however, possible to make useful (albeit incomplete) models of biological systems. This typically involve sets of differential equations together with estimated physiologically relevant ranges of assumptions for the various parameters involved. These can then be used to project the behavior of the system at various signal concentration. The results can produce interesting behaviors.[47]

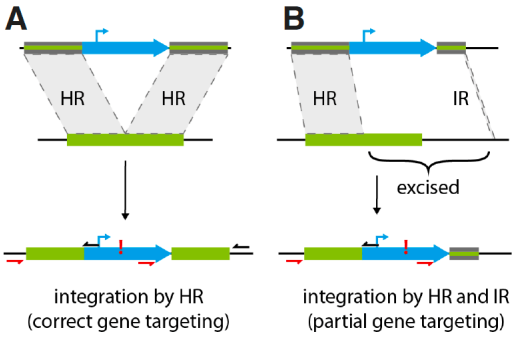
---

[46] see also Honegger & de Bivort, B. (2018). Stochasticity, individuality and behavior. *Current Biology* 28, R8-R12.

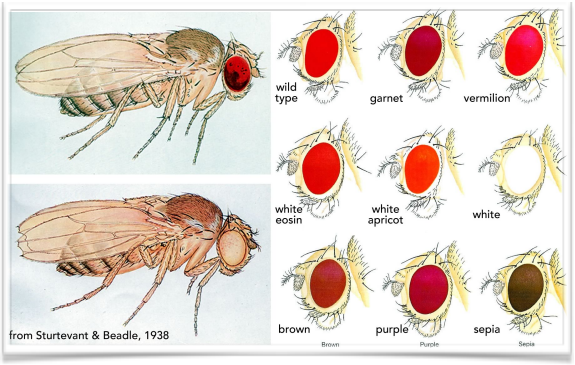
[47] Saka and Smith, Klymkowsky mechanistic behaviors.

Questions to answer and ponder:

1.



Forward (phenotype to gene) genetics: Initially, genetic studies were carried out through what is now known as a forward genetic screen. In such a screen mutations were generated by change (through exposure to radiation or mutagenic chemicals), mutated organisms were bred, and individuals that display specific phenotypic traits were isolated for further characterization. As an example, consider eye shape and color in the fruit fly *Drosophila melanogaster* (↓); these traits are easily accessible experimentally because a *Drosophila* embryo can develop into a fertile adult without an eye. It turns out, it is possible to identify mutant alleles in a number of genes that are involved in eye formation but leave other aspects of embryonic development apparently unaltered. If, however, a gene plays multiple roles in the developing organism, such a screen will not be able to identify it. It is for this reason that forward genetic screens for mutations that influence a particular process are never complete, that is, they do not identify every gene/gene product involved in a process. At the same time it is interesting to note that a number of genes, conserved between disparate species, have yet to have specific functional roles assigned to them.[48]



Backward (gene to phenotype) genetic screens: Modern technologies have given us alternatives to the classic genetic screen. Instead of inducing mutations "at random", it is possible to target specific genes for disruption, and then examine its effects on processes of interest. There are, however, two potential obstacles to such an analysis. The first is that all genes have to be identified and methods used to disrupt the gene need to act on that specific gene alone. This can be tricky, given the presence of overlapping genes, long non-coding RNAs and unrecognized regulatory enhancer elements, and small ORFs. The second issue is the sequential nature of developmental processes; a particular gene can be involved, at various times and in various places, in events that precede the process or behavior of interest. For example, events associated with the formation of the neural tube occur before the formation of the brain. If a gene is involved in neural tube closure, its role in later events can be difficult or impossible to discern using "simple methods".

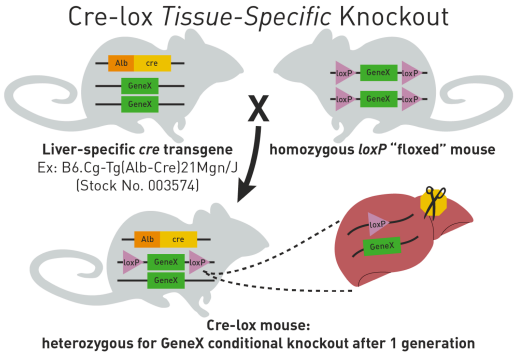
There are a number of commonly used approaches to experimentally manipulate gene expression within an organism. Homologous recombination (→[49]) exploits the cell's DNA repair

[48] Rocha, Jayaram et al., 2023. Functional unknomics: Systematic screening of conserved genes of unknown function

[49] Image from Strotbek et al, 2013. The moss Physcomitrella patens: methods and tools from cultivation to targeted analysis of gene function. *Int J. Dev Biol*, *57:* 553-564

systems; these systems use sequence identity to insert DNA fragments into specific genomic DNA sites or to delete regions of the genomic DNA. When the two homology regions are non-continuous, recombination will result in the deletion of intervening region producing a truncated or null allele. Insertion relies on double strand breaks in the DNA. Because these breaks are rare the system is inefficient. Its efficiency can be increased by methods that produce site specific DNA breaks, e.g. such as CRISPR-CAS9 (see below) to the region of interest. To determine if the genomic DNA has been altered, it is common to insert a "marker" of some kind between the two homology arms. The insertion of such a marker can have two effects, it can disrupt the targeted gene's open reading frame, producing a null mutation. It can also deliver a functional gene that, when inserted, is expressed and renders the cell resistant to some toxic molecule. When the toxic molecule is applied (typically in the context of cultured cells), only those cells that have undergone gene insertion, preferably by homologous recombination, survive. It is also possible to adapt this system to modify a gene's open reading frame, to add sequences that encode, for example, a fluorescent protein or antibody recognizable polypeptide (an "epitope tag") allowing for the visualization of the gene product in living or fixed cells. The specificity of insertion has to be confirmed experimentally, since "random", that is non-sequence specific insertions can occur (albeit at lower frequency). When totipotent embryonic stem cells (ESCs) are used, the altered cells can be used to make a mouse (or perhaps a human). We will consider that process in detail later on.

In "simple" homologous recombination, the gene is typically knocked out. If the targeted gene has multiple functions at multiple times and/or regions of an organism, the interpretation of the gene's functional role in a specific process can be hard or impossible to determine. To address this issue, we turn to what are known as conditional knock-out mutations. This involves a more complicated set of genetic modifications. Here I will describe one approach, know as CRE-LOX (→).[50] In the first step of this process, the gene to be targeted is modified, through homologous recombination, to include specific sequences, known as Lox or Flox sequences flanking the gene (or gene region) that to be deleted. These sequences are targets for a protein, the phage-derived Cre recombinase. If the two Lox sequences are in the same orientation, when the recombinase is expressed it can bind to them both and then catalyze a crossing-over event that results in the excision of the intervening region of the DNA. If the two Lox sequences are in opposite orientations, the recombination event results in the "flipping" of the intervening sequence. The effects depend on where the Lox sites were originally placed. When and where these gene recognization events take place is determined by the expression of the Cre recombinase. In this case, this generally involves generating a transgenic mouse in which the expression of Cre is under the control of an inducible (heat, drugs) or cell type specific promoter. In a wild type mouse, that is a mouse without Lox sequences, the expression of Cre has no effect, but a mouse with a Floxed allele, in those cells where transgene is expressed the Cre recombinase will



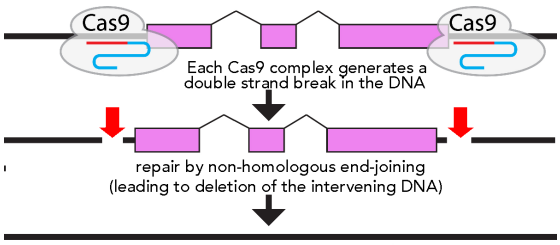
[50] for a review Kim et al., 2018. Mouse Cre-LoxP system: general principles to determine tissue-specific roles of target genes. *Laboratory animal research*, *34:* 147-159.

accumulate and catalyzed the reorganization (mutation) of the Floxed allele, leading to a mouse heterozygous for wild type and floxed alleles. To generate a homozygous floxed mouse, a second mating is necessary.

RNA interference: Another approach to making gene/gene product specific inhibitors involves the use of RNAs.[51] This approach was originally inspired by studies of the roundworm *C. elegans* (which we will return to) that revealed that there were genes that encoded RNAs that while not encoding a polypeptide were processed by various enzymes to form short RNAs that could recognize, that is bind to, and induce the degradation of (or block the translation of) target mRNAs. These microRNAs can be used to target the mRNAs of specific genes and minic the effects of null and hypomorphic (reduced expression) mutations. A caveat in the use of microRNAs is that their recognition sequences do not involve perfect sequence alignment with target mRNAs, which can lead to non-specific, off-target effects.

CRISPR-CAS9: There has been a revolution in genetic engineering technologies, CRISPR-CAS9 and related systems.[52] Like the impact of restriction enzymes (another bacterial anti-viral/foreign DNA defense system), CRISPR-CAS9 is a (bacterial) immunity system; serves to destroy incoming pathogen (viral) DNAs, while leaving the host cell's DNA unaltered. Cas9 is one of a family of enzyme that acts as an small RNA-directed double-stranded endonuclease. Sequence specificity is due to the associated "guide RNA" (gRNA) that determines where cleavage occurs along a DNA molecule. gRNAs are ~23 base pairs long, a sequence length that is long enough to (often) occur only once within the genome of an organism, even an organism with a multi-billion base pair genome such as humans. Versions of the Cas9 and related proteins have been engineered to catalyze base changes at the target site, or to act as transcriptional repressors or activators, rather than cutting the DNA strands.[53]

In response to DNA cleavage, the cell's DNA repair systems act to join the two ends of the cleaved DNA molecule back together again, but this joining is rarely accurate – base pairs can be lost or added, generating a mutated form of the original DNA sequence. If the gRNA target sequence is present in both alleles of a gene, both alleles can be mutated at the same time. One variation, to insure that a region is removed and the gene is disabled (e.g. generating a null mutation), is to use pairs of gRNAs (↓). The double strand break introduced by CRISP-CAS9 can also be exploited to insert DNA via homologous recombination into the cleavage site. The trick is to supply both active gRNA:Cas9 and DNA molecules with regions of homologous sequence flanking the targeted site, and containing the non-homologous sequences to be inserted between these gene homologous sequences. If the CRISPR-CAS9 system is introduced or activated early in the development of an organism all or most cells can be mutated, which can lead to multiple phenotypes.



[51] Tuschl. 2001. RNA Interference and Small Interfering RNAs

[52] In 2020, Emmanuelle Charpentier (b. 1968) and Jennifer Doudna (b. 1964) won the Nobel prize in Chemistry "for the development of a method for genome editing".

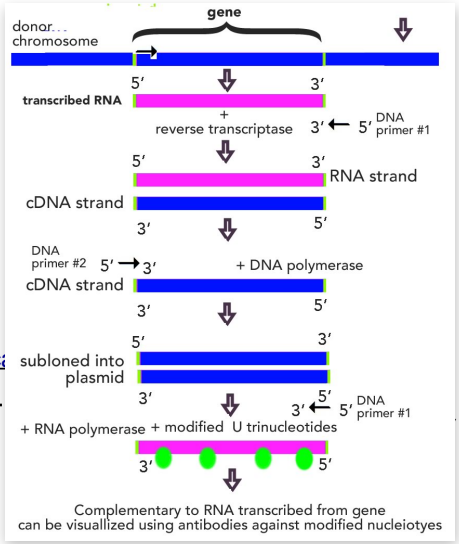
[53] The next generation of CRISPR–Cas technologies and applications

Alternatively, it is possible to activate the system only in specific cell types or at specific times of development (as described for the CRE-LOX system), allowing for finer experimental control.

Outcomes analysis tools: When we consider the effects of an experimental manipulation, we need tools to visualize molecular and cellular level effects, as well as looking at the organism as a whole. So how do we determined how an experimental manipulation has effected the system? How do we know where genes are expressed? There are a number of applicable methods that fall into two basic types - there are those that detect transcribed gene products (RNAs) and those that detect the polypeptides encoded. We consider them briefly here.

RT-PCR: A transformative technology, made feasible by the discovery of heat stable DNA-dependent, DNA polymerases, isolated from archaea that live in very high temperature environments (thermophiles and hyperthermophiles), polymerase chain reaction (PCR) has been a powerful technique for isolating and manipulating genes, as well as for visualizing gene expression and genome sequencing. In the context of gene expression analysis, a modified form of PCR is used, termed quantitative reverse transcriptase-PCR (RTPCR) is used to quantify the amount of a particular transcribed (expressed) RNA within a particular tissue or cell type. After isolating RNA from the sample "reverse transcriptase" and a RNA-specific DNA primer is used to make a complementary DNA copy of the target RNA, a cDNA.[54] The RNA-DNA strands are then separated by increasing the temperature of the system, and a second DNA primer acts together with a thermostable DNA-dependent, DNA polymerase to generate a copy of the cDNA, leading to a doubled stranded DNA molecule with primer sequences at each end. The amplification stage of the reaction involves cycles in which the two strands are separated by increasing temperature. Since the original two DNA primers are present in excess, when the temperature is reduced they bind back to the DNA strands, and initiate a new round of DNA-dependent, DNA synthesis. With each cycle the number of DNA strands doubles, so that there is exponential growth in the number of specific DNA molecules with each cycle. If the gene is not expressed, RNA is not present, and no amplified DNA will be synthesized. By using various tricks (beyond us here, but relatively simple to employ with the right equipment) the process can be made quantitative, so that it is possible to accurately compare the numbers of different types of RNA molecules (the products of a particular gene) present in the original sample, a measure of the level of gene expression, at least at the RNA level. Levels of encoded polypeptides can differ from the level of mRNAs. More recently, it has become possible to isolate and sequence the RNAs (or rather cDNAs derived from them) at the level of individual cells (discussed below).[55]

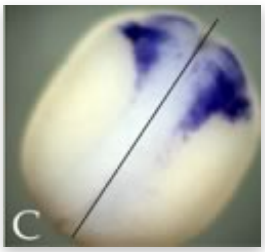
In situ hybridization: To visualize gene expression with spatial resolution, we use in situ hybridization. There are increasingly sensitive (and complicated) versions of the method, but they have the same molecular basis. When a gene is expressed, an RNA molecule complementary to one strand of the gene is synthesized;



[54] insert reference to reverse transcriptase.

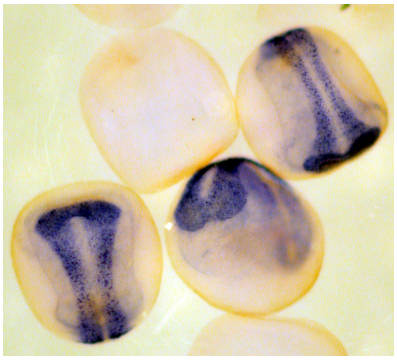
[55] A practical guide to single-cell RNA-sequencing for biomedical research and clinica

these “sense” RNAs accumulate - there level is controlled by their relative rates of synthesis and degradation. To identify cells that express the gene, we generate modified “anti-sense” RNA molecules, typically having subcloned the sequence into a plasmid that uses a phage promoter that drives synthesis from the anti-sense strand using modified forms of the UTP ribonucleotide (→). Once the fluorescent probe has been synthesized, the overall process is relatively simple. The tissue is chemically stabilized and permeabilized, so that RNA molecules can diffuse into and out of it. It is then incubated with either sense or anti-sense probes. Because of the complementary nature of nucleic acids, the anti-sense probe RNA will bind to any RNA transcripts that are present, while the sense probe is not expected to bind - it is used as a control. By controlling the hybridization temperature, we can minimize low affinity, non-specific interactions. The probe will be retained in cellular regions where the gene is expressed, and the RNA accumulates, and washed away from regions where the gene is not expressed. Antibodies, conjugated with various enzymes (typically alkaline phosphotase or horseradish peroxidase) and color-generating reactions can then be used to reveal probe RNA:mRNA complexes. As an example here (→) is a neurula stage *Xenopus laevis* embryo in which expression of Snai2/Slug in the neural crest has been visualized by in situ hybridization.[56] In situ hybridization can provide single cell resolution, distinguishing cells that do, from those that do not, express a particular gene. The specificity of the technique is influenced by the length of the probe and the hybridization temperatures used.



Single cell RNA Sequencing (scRNA SEQ): The advent of more efficient DNA sequencing methods, together with PCR-based amplification, has made it possible to isolate and sequence the RNA molecules within a single cell. Once sequenced, the number of reads (e.g. molecules) of each RNA (each gene product) can be counted to provide a catalogue of the genes expressed within a cell. While not complete because some RNAs are present at very low levels, and some are discarded during the course of sample preparation, scRNA SEQ provides a new level of cellular characterization; it reveals not just the genes expressed, but, in heterozygotes, whether one or both alleles are expressed. As an example, it can determine which of the two X chromosomes are active in a particular cell. An important insight has been that cells of the cell type vary in terms of gene expression, in part due to the stochastic nature of gene expression and down-stream effects of that variation. These variations can impact cell behavior and organismic phenotype.

Immunocytochemistry: RT-PCR, in situ hybridization, and scRNA SEQ report on RNA levels. In the case of mRNAs, RNA levels do not always correlate with polypeptide levels. One approach to avoid this disconnect is to use antibodies. Antibodies are proteins generated by the vertebrate immune system, they bind with high specificity to particular molecular targets. Antibodies act very much like anti-sense RNA in situ probes, binding to specific molecular (protein) targets. The example here (→) is a neurula stage *Xenopus laevis* (clawed frog) embryo stained for the Sox3 protein, a transcription factor involved in the specification of neural

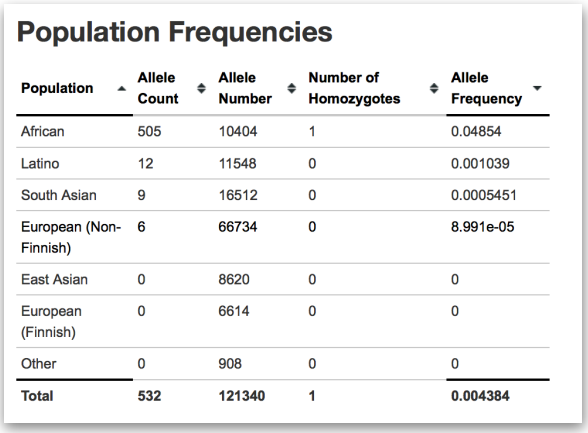


[56] from: An NF-κB and Slug Regulatory Loop Active in Early Vertebrate Mesoderm

cells in the nervous system. Immunocytochemistry depends on the presence of specific antibodies, which can be difficult to generate. What can be used instead are epitope-tags, short sequences of amino acids added (through genetic engineering approaches) to the gene encoding the protein. These can be recognized by an epitope-specific antibody. A full characterization of the proteins present in a cell or tissue, together with their interaction partners, relies on various physicochemical approaches. In one approach, antibodies are used to selectively precipitated that antibody's target, together with associated molecules. Other methods, beyond us here and for the moment, such as mass spectrometry help define proteins and protein interaction partners.[57]

Using web-based bioinformatic tools (BLAST and gnomAD): There are other web based tools to identify evolutionarily conserved regions in related gene products. Perhaps the most useful is BLAST. It enables you to take either a nucleotide or a polypeptide sequence and search for similar sequences in all sequenced genes (deposited in GenBack, a central repository). The program returns similar sequences in other organisms. The presence of such sequences is generally considered evidence for evolutionary relationships, horizontal gene transfer, or (more rarely) convergent evolution towards a similar function (think wings in insects and bats). The BLAST tool is also useful for designing gene specific DNA primers and for identifying parts of nucleic acid or polypeptide sequences that are conserved, that is, that vary the least from organism to organism – we might expect such regions to be particularly sensitive to mutational change. The absence of allelic (missense/non-sense) variants (in gnomAD) in such regions would argue for the action of positive selection.

In the modern (i.e. DNA sequencing) era, there is an ever increasing number of "libraries" of genomic and exomic (present in mature RNAs) sequence data. One of these, gnomAD includes more than 120,000 "normal" people from around the globe, and continues to increase in its population diversity.[58] To search the database, the user (you, for example), inputs a gene's official name, as listed in OMIM or GenBank. gnomAD then displays sequence data from unrelated individuals; this allows for the identification of alleles and mutations present in a range of human populations. Data from gnomAD enables us to make informed guesses as to the impact of various genetic differences on the activity of a gene product.[59] If, for example, a dominant allele has been linked to a disease and yet that allele is detected in the gnomAD database, we might suggest that either the allele is not the cause of the disease, or that the effects of the allele are influenced by variation (alleles) in other genes, leading to reduced penetrance and/or expressivity. If an allele is present in a heterozygous condition, but not a homozygous one, we can tentatively assume that negative selection is acting on the allele. If, on the other hand, alleles are present at different frequencies in different populations, that may be evidence for the



Population Frequencies

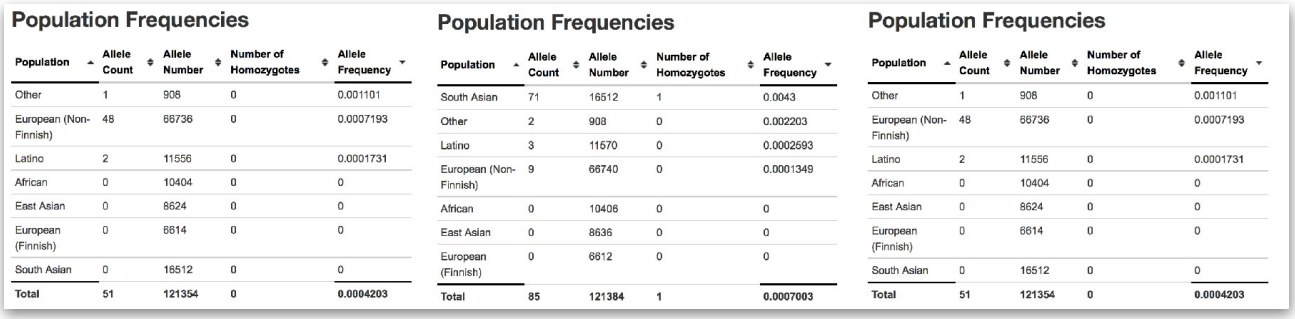
Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	505	10404	1	0.04854
Latino	12	11548	0	0.001039
South Asian	9	16512	0	0.0005451
European (Non-Finnish)	6	66734	0	8.991e-05
East Asian	0	8620	0	0
European (Finnish)	0	6614	0	0
Other	0	908	0	0
Total	**532**	**121340**	**1**	**0.004384**

[57] Here is an example of proteomic analysis: Region and cell-type resolved quantitative proteomic map of the human heart

[58] Genomics, Big Data, and Medicine Seminar Series – Daniel MacArthur

[59] The ExAC browser: displaying reference data information from over 60 000 exomes.

action of positive selection dependent on environmental factors. In addition, the frequency of alleles in different populations often reflects the effects of founder effects, bottlenecks, and drift. Take for example three alleles and the hemoglobin B (HBB) gene, p.Gly70Ser, p.Glu122Gln, and p.Gln40Ter (Ter=stop) (↓). We see that the Gly70Ser and Glu40Ter alleles are present primarily in non-Finnish Europeans,



Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Other	1	908	0	0.001101
European (Non-Finnish)	48	66736	0	0.0007193
Latino	2	11556	0	0.0001731
African	0	10404	0	0
East Asian	0	8624	0	0
European (Finnish)	0	6614	0	0
South Asian	0	16512	0	0
Total	**51**	**121354**	**0**	**0.0004203**

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
South Asian	71	16512	1	0.0043
Other	2	908	0	0.002203
Latino	3	11570	0	0.0002593
European (Non-Finnish)	9	66740	0	0.0001349
African	0	10406	0	0
East Asian	0	8636	0	0
European (Finnish)	0	6612	0	0
Total	**85**	**121384**	**1**	**0.0007003**

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Other	1	908	0	0.001101
European (Non-Finnish)	48	66736	0	0.0007193
Latino	2	11556	0	0.0001731
African	0	10404	0	0
East Asian	0	8624	0	0
European (Finnish)	0	6614	0	0
South Asian	0	16512	0	0
Total	**51**	**121354**	**0**	**0.0004203**

while the Glu122Gln allele is found in South Asians. It is not clear exactly what the effects of such missense mutations will be on the functions of the polypeptide – it could change folding, change interactions with other polypeptides and molecules, add or remove sites of post-translational modification, or change catalytic activity, if the polypeptide has such an activity. It is likely that the Glu40Ter mutation will produce a short, non-functional 39 amino acid polypeptide (compared to the 147 amino acid long wild type polypeptide). It is unlikely that the truncated protein is functional, but if it accumulates it could interfere with the function or molecular interactions of the full length polypeptide.

Questions to answer:

277. What things might you use BLAST for?


CHUCKIE 'D' SAYS:
EMBRACE
YOUR INNER FISH
Ray Troll, 2006