An introduction to genetics and molecular biology

Cavan Reilly

September 4, 2019

Table of contents

Biology

Biology is the study of life, however it is surprisingly difficult to define life.

A living thing is usually defined descriptively by what it does: metabolism, reproduction and growth are some of the crucial activities of a living thing.

There are many subdisciplines in biology: ecology, microbiology, molecular biology, immunology, virology, genetics...and many that overlap with other areas of inquiry: biochemistry, bioinformatics, biophysics...

Life

Features that are common to all forms of life include some sort of cell membrane that keeps the contents of the cell secluded from its environment and the presence of deoxyribonucleic acid (DNA).

The 3 basic groups of organisms found on Earth are classified as archea, prokaryotes and eukaryotes.

Archea and prokaryotes are unicellular organisms while some eukaryotes are multicellular.

Not much is known about archea however they have been found in the human gut and many other habitats.

Most forms of life are unicellular: the majority of the biomass (i.e. living matter) of the Earth is composed of bacteria.

Eukaryotes

Bacteria are everywhere, from hot springs to the inhospitable environment of the human gut.

Eukaryotes have a nucleus that contains the DNA of the cell whereas prokaryotes do not.

They also have a number of organelles that the other 2 groups do not and eukaryotes use different methods for cell division.

There are 4 divisions of Eukaryotes: protists, fungi, plants and animals.

DNA

We refer to the DNA of an organism as its *genome*.

DNA is a molecule that has a sequence of sugar molecules (and a phosphate group) with 1 of 4 distinct *bases*: adenine (A), cytosine (C), guanine (G) and thymine (T).

We think of DNA sequences in terms of sequences of bases.

A base bound to a sugar and phosphate group is called a *nucleotide*.

Each base has a complementary base that it will naturally form a hydrogen bond with: the complement pairs are A-T and G-C.

Chromosomes

In its naturally occurring form, a single stranded DNA molecule binds to another single stranded DNA molecule so that each base on one strand binds to its complement.

This pair of molecules tends to fold into a double helix.

The human genome has approximately 3 billion basepairs and about 99.9% of these are identical among humans.

In healthy humans, DNA exists as 23 pairs of distinct molecules known as *chromosomes*. Other organisms have different numbers of chromosomes and some even have more than just a pair of each.

One pair, the sex chromosomes, determine the gender of the person.

Gene expression

In its naturally occurring state DNA is bound to a variety of molecules, and the manner in which these other molecules bind to DNA impacts the functionality of DNA.

Which molecules are bound to a DNA molecule varies over time and in response to stimuli.

Cells use DNA as a template for making proteins.

Proteins are molecules generated by cells that allow the cell to fulfill a function.

Gene expression (cont.)

Ribonucleic acid (RNA) is much like DNA except uracil takes the place of thymine, it doesn't tend to form bonds with its complement and it takes many different shapes.

RNA has many functions in a cell, but it is "most famous" for transporting information from DNA to the sites where proteins are made (i.e. ribosomes).

Ribosomes are organelles composed of RNA molecules.

MicroRNAs are short RNA molecules that can bind to other nucleic acids and alter their function.

Gene expression (cont.)

A subsequence of the genome that is used to make a specific protein is called a *gene coding region*, or sometimes just a *gene*.

The term *gene* is used somewhat loosely: we might mean the subsequence of bases at the gene coding region or the location in the genome of the subsequence.

The term *locus* is used to refer to a location in the genome (there may or may not be a gene at a locus).

Although the number of genes isn't fully known, there are about 21,000-23,000 human genes.

Gene expression (cont.)

Most human genes are not composed of contiguous segments of DNA.

Just prior to the start of a gene are certain sequence features that are nearly common across all genes that regulate how the gene is used.

After this regulatory region there is a section called the 5' (read: 5 prime) untranslated region (UTR), then the first exon.

After this exon one encounters the first intron which is a section of the gene that is not used to make protein.

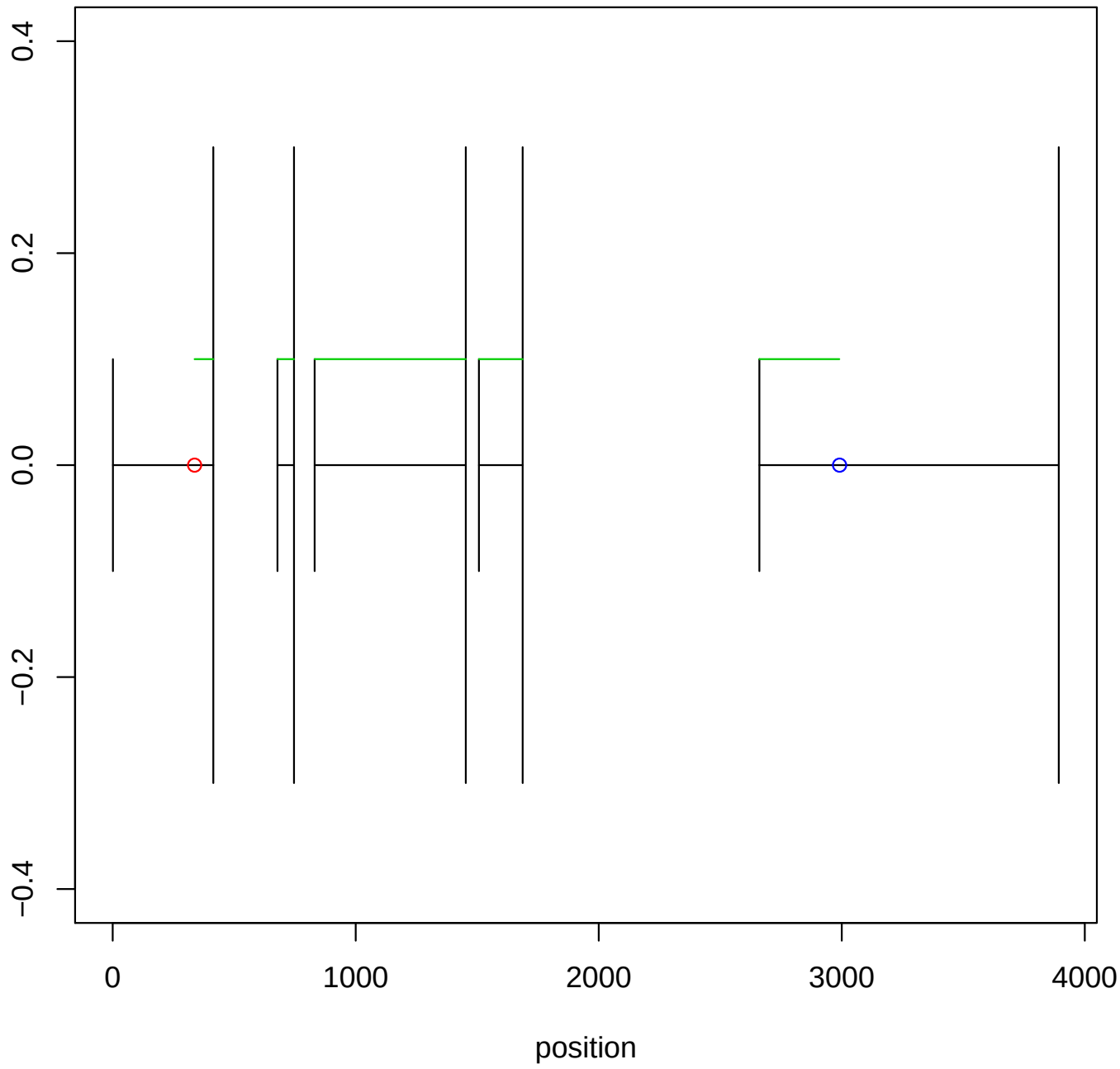
Gene expression (cont.)

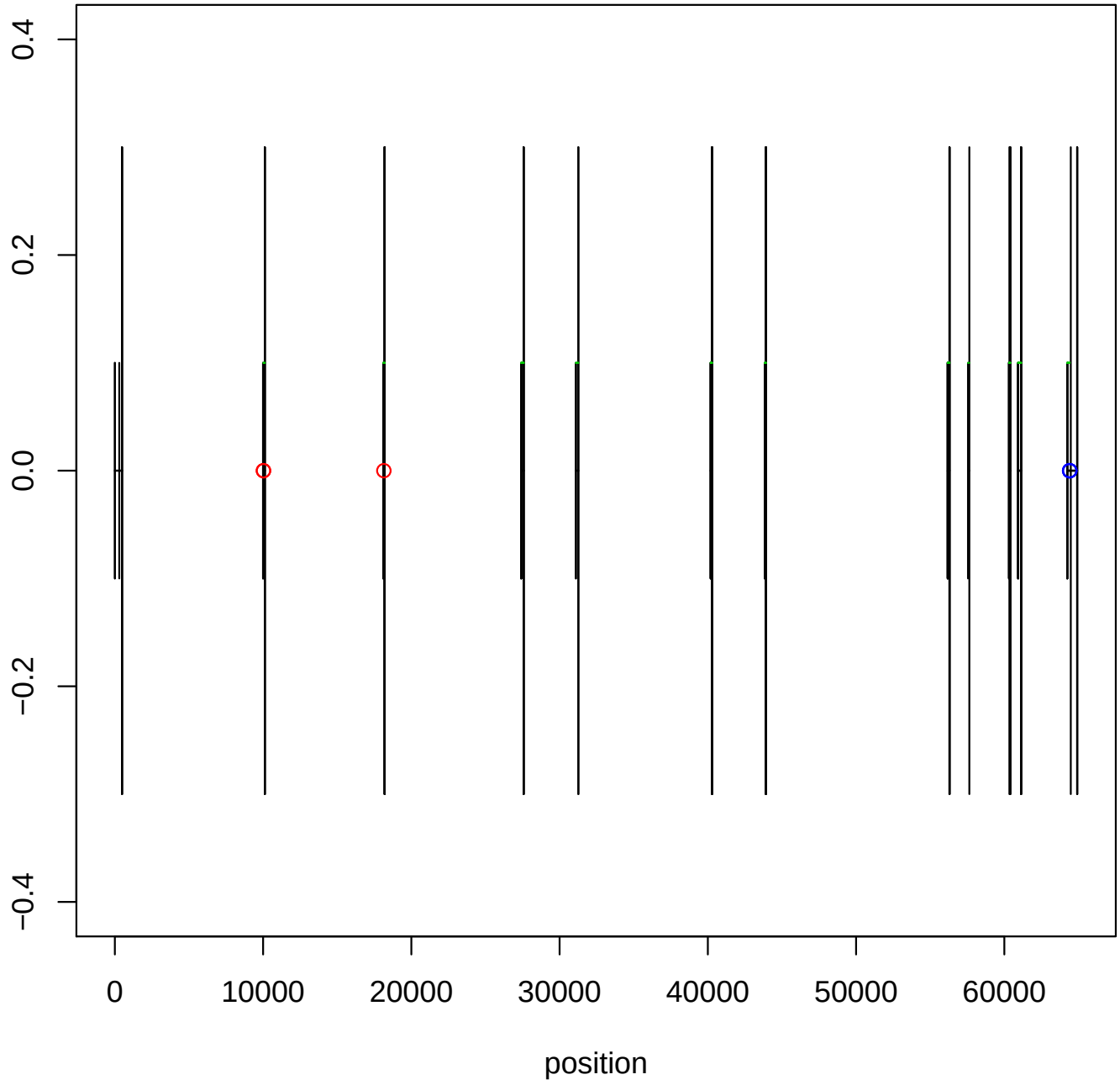
There are typically 5-10 exons in human genes and the introns can be thousands of bases in length.

Then the 3' UTR is found after the last exon.

Over 90% of human genes undergo alternative splicing: there are multiple mRNA molecules that can be made from the gene.

This is possible because the gene can use just some of the exons, or use different 5' and 3' UTRs, and some genes even have multiple start and stop locations.


0.4
0.2
0.0
−0.2
−0.4
0
1000
2000
3000
4000
position


0.4
0.2
0.0
−0.2
−0.4
0
10000
20000
30000
40000
50000
60000
position

Gene expression (cont.)

Transcription is the process by which alterations in the collection of proteins bound to a DNA molecule lead to copying a gene coding region to create an RNA molecule.

This RNA molecule has information about the particular sequence of bases at this gene coding region.

After copying the entire gene, the introns are removed (but not the UTRs) and structures are added to the ends to preserve the molecule: at that point we call the molecule a messenger RNA (mRNA).

Gene expression (cont.)

Translation is the process by which a cell uses the information in the mRNA molecule to create a protein.

Proteins are composed of a sequence of amino acids: there are 20 naturally occurring amino acids.

The *genetic code* is the basis for the system for converting DNA sequences into sequences of amino acids.

A *codon* is a collection of 3 contiguous nucleotides: each codon in an mRNA maps to 1 amino acid according to a fixed set of rules.

Gene expression is a term that is used to describe the entire process of translation and transcription of a gene.

Mendelian genetics

Since there is some variation among DNA sequences in humans, there are genes that have different subsequences among people.

Genes (or loci) that exhibit differences among members of a species are called *polymorphic*.

The possible subsequences at a gene (or a locus) are called *alleles*.

Since all humans have 2 copies of each chromosome, every gene has 2 alleles. One is inherited from the mother (the maternal allele) and the other from the father (the paternal allele).

Mendelian genetics (cont.)

The alleles that a subject has at a locus (or set of loci) is called that subject's *genotype*.

If the 2 alleles at a locus are distinct we say that the organism is *heterozygous* at that locus otherwise the organism is *homozygous*.

An observed feature of an organism is referred to as a *phenotype*.

Mendelian genetics (cont.)

If a trait is at least partially genetically transmitted, the extent to which it is hereditary is called its *heritability*.

There are methods for estimating this just using observable features of organisms-in particular you don't need to know which genes are responsible for the trait.

Note: Heritability is used in a technical sense that only applies to continuous traits.

The manner in which a genotype impacts a hereditary phenotype depends on the trait.

Mendelian genetics (cont.)

In Mendelian genetics, individuals with homozygous genotypes display the phenotype corresponding to the common allele, but heterozygous genotypes can display either phenotype.

If an organism has a heterozygous genotype and displays a certain trait, then that trait is said to be dominant and the other trait is recessive.

Traits that are controlled by one gene in a recessive or dominant fashion are called *Mendelian traits* otherwise they are called *complex traits*.

Cystic fibrosis is a well documented Mendelian trait: currently 3,000 loci have been implicated in Mendelian traits.

Haplotype

When we consider 2 loci simultaneously, if we know which pairs of alleles are on the same chromosome then we know the *haplotype*.

For example, consider 2 genes (denoted by a letter) where each gene has 2 alleles (which we distinguish by case).

If someone has genotype AA and Bb, then we know that 1 chromosome must have alleles AB and the other chromosome must have alleles Ab-in this case we know the haplotype.

Haplotype (cont.)

In contrast, if someone has genotype Aa and Bb then 2 configurations are possible: one chromosome has alleles AB while the other has ab or one chromosome has alleles Ab while the other has alleles aB-here the haplotype is unknown.

Conventional assays for genotyping a subject (i.e. determining the genotype for many loci simultaneously) do not give information about haplotypes, they only provide information about genotypes for each locus separately.

This presents a fundamental challenge for determining which genes are involved in disease processes.

Meiosis

One important function of some cells is that they go through the *cell cycle* and replicate-a process called mitosis.

During the process of replication, the cell makes copies of its chromosomes and then passes the copied chromosomes on to the next generation-this can lead to errors, or *mutations*.

Such mutations are thought to be an important source of genetic diversity.

However there are much stronger forces that lead to genetic diversity.

Meiosis (cont.)

The generation of diversity in offspring is essential for survival of a species.

Such diversity allows a species to be able to adapt to changes in its environment since some individuals might be able to survive due to mutations.

This phenomenon is at the heart of natural selection.

This is illustrated by organisms that can reproduce sexually and asexually: when there are sufficient resources the organisms undergo sexual reproduction (e.g. Cryptosporidium).

Meiosis (cont.)

Meiosis is the process that leads to the formation of sex cells (i.e. sperm and egg), also known as *gametes*.

Meiosis starts in a manner similar to mitosis: both copies of all of the chromosomes are replicated.

After this is a stage called prophase I in which the (replicated) pairs from both parental sources line up and separate, going into 2 daughter cells.

After this each of these daughter cells has 2 copies of each chromosome.

Meiosis (cont.)

In the next step of meiosis the daughter cells split again to create a total of 4 cells where each of these cells only has one of each chromosome (rather than a pair).

When the daughter cells divide, each chromosome from a pair is equally likely to be transmitted to the resulting cells.

Thus each gamete has a mix of chromosomes-some come from the maternal source and some from the paternal source.

When 2 gametes fuse the resulting cell will thereby have a pair of each of the chromosomes and the resulting organism will have a unique genome.

This process leads to substantial genetic diversity.

Recombination

While the random assortment of chromosomes to gametes generates new genomes, there is another mechanism that increases genetic diversity: *recombination*.

During meiosis (during prophase I), before the pairs of chromosomes split up and go to different cells, they sometimes will swap sections of chromosomes between pairs to create entirely new chromosomes.

This process is called recombination.

Recombination (cont.)

As an example, suppose we consider the fate of some chromosome after a recombination. If we use letters to denote loci, case to distinguish alleles and suppose that the maternal and paternal alleles are all distinct then:

Before recombination:

chromosome 2 (maternal):	`ABCDEFG`
chromosome 2 (paternal):	`abcdefg`

After recombination:

chromosome 2 (one copy):	`ABCdefG`
chromosome 2 (the other copy):	`abcDEFg`

This process creates even further genetic diversity.

Recombination rate

The *recombination rate* between 2 loci is the probability of the 2 loci having alleles from different parental sources.

If 2 loci are very close then it is unlikely that a recombination will separate them, consequently the recombination rate is low.

Conversely, if 2 loci are on distinct chromosomes then the probability that the alleles come from different parental sources is $\frac{1}{2}$ since chromosomes pass to gametes independently of one another.

Hence for all pairs of loci, the recombination rate is in the interval $\left(0, \frac{1}{2}\right]$.

Markers

A *marker* is some known feature of a genome.

For example, the center of a chromosome can be observed under a microscope (during certain stages of the cell cycle) hence this is a marker.

A *polymorphic marker* is a marker that varies among members of a species-all (normal) chromosomes have a center so the center is not a polymorphic marker.

A *single nucleotide polymorphism* (SNP) is a single nucleotide that is known to vary among individuals.

While older studies used other types of markers, SNPs are the markers of choice today.

SNPs

Millions of SNPs have been identified in the human genome and low cost assays are available for genotyping subjects at these SNPs.

There are also known sites which are called *indels*.

These are locations where some people have additional nucleotides compared to others (i.e. insertions),

or some people are missing nucleotides compared to others (i.e. deletions).

So the actual length of the genome in terms of nucleotides varies among members of a species.

Genetic linkage

Linkage analysis is a set of techniques whose goal is to estimate the recombination fraction between loci where (at least) 1 locus alters risk for disease.

If we know the recombination rate between a marker (whose location on the genome is known) and a locus that alters disease risk is near zero, then we have found the location of a locus that alters disease risk (i.e. we have found the "disease gene").

By using many markers that cover the genome we can therefore go fishing for the location of the disease gene-*genome scans*.

Genetic linkage (cont.)

By knowing which genes have alleles that alter risk for some disease we can gain insight into the disease and perhaps develop better treatments for patients with the disease.

We can also provide information for pregnant mothers regarding the probability that their offspring will have a particular disease.

Such genetic counciling is very important for some human subpopulations as debilitating recessive traits circulate in some subpopulations.

A simple example

To understand the difficulties of linkage analysis, we will consider a simple example.

Let's assume that there is a gene with certain alleles that alter the risk for some disease.

Suppose further that we are willing to assume that this disease is a dominant trait.

linkage example (cont.)

Consider a mating where one parent is affected by this trait and the other is not (and some offspring are affected and some are not).

Note that by assuming the disease has a genetic source and acts in a dominant fashion we know the genotype for the disease gene:

all affecteds (i.e. those with the disease) are heterozygous with a copy of the disease allele

all unaffected subjects are "homozygous" for the "normal" allele (there could be multiple normal alleles).

linkage example (cont.)

Suppose we genotype both parents at some marker and discover that both parents are homozygous, but with different alleles (say the diseased parent is A_1, A_1 and the other is A_2, A_2).

Use D to represent the disease allele and N to represent the normal allele at the disease locus.

This implies that all offspring will be heterozygous at this marker allele (with genotype A_1, A_2).

Hence an affected offspring will be heterozygous at the marker locus and the disease locus, i.e. this offspring will be *doubly heterozygous*.

linkage example (cont.)

Suppose these parents have an affected offspring. Note that we will also know the haplotype here which we denote $D, A_1|N, A_2$.

Now suppose the affected offspring mated with someone who is not affected by this disorder and was homozygous at the marker locus.

If this mating resulted in an affected offspring then by examining the alleles at the marker locus we are able to determine if a recombination occurred between the marker locus and the disease locus.

linkage example (cont.)

Suppose the unaffected parent in the second mating had genotype A_1, A_1.

Then the marker genotype of the affected offspring will be either A_1, A_2 or A_1, A_1.

In the former case, the affected parent transmitted the disease allele and the allele A_2 hence there was a recombination.

In the latter case, the affected parent transmitted the disease allele and the allele A_1 hence there was not a recombination.

linkage example (cont.)

Now suppose the unaffected parent in the second mating had genotype A_1, A_2.

Then the genotype of the affected offspring will be either A_1, A_1 (no recombination), A_1, A_2 (can't tell if recombination occurred) or A_2, A_2 (a recombination).

If we had many matings where we could deduce if a recombination occurred then we could use the sample proportion of recombinations to estimate the recombination fraction.

If we found a marker that was close to the disease locus (in terms of the recombination fraction) then we would know the location of the disease gene.

linkage example (cont.)

We could then use tools available online to find if there are any genes known to be at that location.

For example, let's look up rs1501299-the link to follow is the one to the NCBI website- this takes you to dbSNP which has information about this SNP.

Typically there will be many genes (hundreds) that are "close" to most markers when we use data from related subjects.

This is because related subjects typically share large chromosomal regions so that the ability to localize a disease gene is poor.

LOD scores

Rather than estimate the recombination fraction, geneticists test the null hypothesis that the recombination fraction is $\frac{1}{2}$ against the alternative that it is less than $\frac{1}{2}$.

If the null is rejected then the 2 loci are said to be *linked*.

Geneticists have traditionally conducted this test by computing the *LOD score*.

LOD scores (cont.)

The LOD score is the base 10 log of the test statistic for testing the null hypothesis of that the recombination fraction is $\frac{1}{2}$, and the null is rejected if the LOD score exceeds 3.

This turns out to be equivalent to rejecting the null hypothesis if the p-value is less than 0.0002.

This seemingly very "conservative" procedure makes sense given that one is typically testing many markers at once.

In fact given the current practice of examining million of markers, this is probably not conservative enough-currently 5.0×10^{-8} is widely used for genome wide association studies.

Estimating the recombination fraction given a family structure (i.e. a pedigree), disease status and a set of markers is incredibly computationally challenging.

Genetic associations

So far, we've focused on related subjects (families), but in fact all humans are related at some level.

This fact manifests itself in the genome via a phenomenon called *linkage disequilibrium* (LD).

Linkage disequilibrium exists when 2 (or more) alleles tend to be present in the same haplotype more than we expect by chance.

It arises due to recombination: as sequences of meioses go over the generations, ancestral chromosomes are continually cut up and reassembled due to recombination.

Genetic associations (cont.)

Within a family we noted that recombination will tend to produce shared large pieces of chromosomes-this is a problem for linkage analysis since 10-20 Mb (i.e. megabase) sections will be found.

When we look at unrelated subjects, much shorter segments of chromosomes will be retained since many more generations of meioses will separate these 2 unrelated individuals.

This phenomenon used to be exploited for the purpose of "fine mapping": after linkage is detected using family data, unrelated cases and controls would be used to test for an association between the presence of a particular allele and case status.

Genetic associations (cont.)

Such a test could just be conducted using Pearson's χ^2 for contingency tables: case versus controls versus presense of a certain allele or not.

Note that in linkage analysis, if a marker is linked to a disease allele then subjects in that family with the disease would tend to have the same allele at the marker, but different families could have different marker alleles that are transmitted with the disease allele.

In contrast, in association studies, we expect to find the same allele at the marker locus that is associated with the disease allele.

Admixture and confounding

Unfortunately, genetic associations can arise for reasons other than linkage between the 2 loci in question.

This can arise due to the existence of subpopulations that are genetically similar in a heterogeneous population (admixture).

For example, African-Americans are at greater risk for sickle cell anemia and have alleles at marker loci that have different frequencies than the rest of the US population.

Admixture and confounding (cont.)

Hence if one tests for an association between having sickle cell anemia and the presence of a certain allele at some marker, one will tend to find such an association due to the confounding effects of ethnicity.

This has led to several different strategies for attempting to correct for this source of bias.

Family based controls

The idea behind family based controls was originally developed for the *trio* study design in which one has access to a collection of parents and their affected offspring.

In this context, the family based control genotype is the pair of alleles that were not transmitted to the affected offspring.

As an example, suppose the alleles at some marker locus for one parent are A_1, A_2 and for the other parent they are A_3, A_4.

If the genotype of the affected offspring is A_1, A_3 then that is the case genotype and the control genotype is just A_2, A_4.

Family based controls (cont.)

One can then make the following table and use a test statistic called McNemar's χ^2 test to test for an association.

	control A_1	control $\overline{A}_1$	total
case A_1	a	b	w
case $\overline{A}_1$	c	d	x
total	y	z	n

This is known as the transmission disequilibrium test (TDT) and it has been extended to general pedigrees and quantitative traits.

The family based association test (FBAT) is a generalization of the TDT and many of its extensions.