30. Human genomics and polymorphic markers
Reading 357-359, 360-361, 366-370

- Montgomery Slatkin, 4155 VLSB. Office hours Wednesdays 1:30-3:30. Email slatkin@berkeley.edu.
- The final will be held Friday, Dec. 19th, from 8-11 AM in Pauley. It will cover material only in this part of the course. The final will be based on the notes, including assigned problems in the book and all problems in the notes, as well as other material in the notes. Material in the book not covered in the notes will not be on the final.
- This part of the course will deal with gene mapping, population genetics and molecular evolution, with particular emphasis on recent developments in studies of humans.
- The notes will be posted the day before each lecture, if possible. They will be updated later if there are corrections or changes in content, but not if some material is not presented until a later lecture.
- I will not cover the technology of DNA sequencing. Sequencing via cloning is outlined in the text. Chromosomes are broken into overlapping pieces, each of which is cloned. The ends of the clones are Sanger sequenced and the sequences are assembled by using the overlapping regions of fragments. **10X coverage** achieves an sequencing error rate <1/10,000.
- The goal of genetic mapping is to locate a gene with alleles that cause differences in phenotype relative to markers whose position in the genome is known.
- The starting point is the creation of a **genetic map** (or **linkage map)** which describes the recombination frequencies relative to one another and the order of a set of loci, which are called **marker loci**. The units are **map units** or **centiMorgans**. The map distance between very closely linked loci is the recombination frequency (RF). The map distance between loci farther apart is greater than the RF because the map distance includes double crossovers.
- Loci used in classical genetic studies determine the state of easily detectible, discrete phenotypic differences, e. g. round and wrinkled peas, white or red eyes, presence or absence of symptoms of cystic fibrosis. Traits like these are called **visible polymorphisms**. How many visible polymorphisms are known for a species depends on what nature provides and how much effort has been made to find them.
- Modern genetic studies rely on **molecular polymorphisms**, which are differences in DNA sequence. They almost never produce visible effects but are detectable using various biochemical methods.
- There are three kinds of molecular polymorphisms commonly used, restriction site polymorphisms (**RFLPs**), single nucleotide polymorphisms (**SNPs**) and simple sequence repeats (**SSRs**, also called microsatellites).
- RFLPs are rarely used now but are of interest for understanding the recent history of genetics.

- Alleles at **SSRs** are distinguished by the number of repeats of a 2-5 base motif, typically 4-50 times. Almost all SSRs are not in coding sequences and appear to have no effect of phenotype, although SSRs that cause diseases such as Fragile X and Huntington's are notable exceptions. They are convenient for construction high-density linkage maps in vertebrates because they are very abundant, roughly every 6 kb in the human genome, they have many alleles. SSRs are easy to find in a genome by hybridization to a repeat motif. Furthermore, once unique primer sequences are found, it is easy to screen large number of individuals. The human linkage map is based on more than 20,000 SSRs.
- SNPs are far more abundant in the human genome than SSRs. There is an average of one SNP per kb. SNPs are diallelic.
- Linkage maps in humans are made by screening **parent-offspring trios** for molecular markers and locating all recombination events. A widely used set of reference families, the CEPH families (http://www.cephb.fr/en/cephdb/).
- A **physical map** describes the locations of loci relative to physical properties of chromosomes, either visible landmarks on chromosomes or, if the genome sequence is available, the number of base pairs separating two loci.
- The physical and genetic maps have the **same order of loci**. On the scale of whole chromosomes, the distance on a physical map is approximately proportional to the distance on the linkage map. In humans, **1 mb≈1 cM**, which follows from the fact that the total number of bases is $3x10^9$ and the total map length is 30 Morgans. In mice, 1mb≈2cM.
- In humans, data from parent-offspring trios shows that recombination rates vary within a chromosome by a factor of 5-10. Recombination rates in females tend to be higher than in males by as much as a factor of 5, but in a few regions male rates are higher.
- At a fine scale, local recombination rates vary tremendously. In humans there are recombinational **hotspots**. Jeffreys et al (2001) measured recombination in males by sperm typing. demonstrated that local rates differ by a factor of more 1000 within a 216 kb region on chromosome 6.
- The **karyotype** is the visual description of complete set of chromosomes. Physical landmarks are seen when human chromosomes are stained with Giemsa dye, which distinguishes regions of relatively high G+C content as light bands and low G+C content as dark bands.
- Using hybridization methods such as **FISH**, a locus that has been cloned can be mapped to a chromosome band. Hybridization methods has several advantages over linkage mapping. (1) Hybridization does not require polymorphism in the cloned locus. (2) Hybridization does not require polymorphic linked markers. (3) Only a chromosomal preparation from a single individual is needed. (4) The relationship of the locus to other chromosomal features such as inversions or deletions may be apparent. However, the resolving power of FISH protocol is about 4-8 mb only.

- Sequence similarity in different species shows that chromosomal segments are rearranged but there are recognizable regions of **synteny** in which gene order is conserved, even between humans and mice.
- The complete sequence of a human genome has been available since 2001. The original (reference) sequence was a mosaic of sequences from a few anonymous people. Several individuals have now been sequenced and their sequences are publicly available. The reference sequence revealed several things.
- (1) There are between 30,000 and 40,000 genes. The exact number is not known because of the difficulty of recognizing small (<25bp) RNA genes and small coding genes. Longer coding sequences are recognized by **open reading frames (ORFs)**, which are segments with no stop codon. There are 64 codons, only 3 or 4.7% of which are stop codons. If you randomly assemble nucleotides, the average length of an ORF would be 1/0.047≈21 codons. In a long randomly assembled sequence ORFs of 30-60 codons could easily arise by chance in long stretches of random nucleotides.
- (2) There are more than **2000 RNA** genes, including tRNAs, rRNAs, snoRNAs, and smRNAs.
- (3) There is **extensive sequence similarity** of human protein-coding genes to gene in other species, 99% to chimpanzees, 61% to flies, 43% to worms and 46% to yeast. Genes with such similar sequences are **homologous**, meaning that the descended with modification from a common ancestral gene.
- (4) There are ~2000 **multigene families**, which are made up of genes that arose by past duplication events. Some families contain hundreds of genes (Table 10.2). Members of multigene families are **paralogous**. Genes in closely related species can be recognized as **orthologous**, meaning they are descended from a gene in the common ancestor of those two species. For example, the β-globin in humans and in chimps are orthologous. The two α-globin genes in humans and the β-globin gene in humans are parologous.
- (5) Roughly **50%** of the human genome is **repeat sequence**, including transposons (45%), processed pseudogenes, SSRs (3%), segmental duplications, and blocks of repeated sequences in centromeres and telomeres. Three type transposons make up a substantial fraction of the human genome, LINEs (~850,000 copies, 6-8 kb each, 21% of the genome), SINEs (1,500,000 copies, 100-300 bp each, 13%), and retroposons (450,000 copies, 1.5-11 kb each, 8%).
- (6) Gene density and gene size differ greatly. Average density if 30,000/3,000 mb or 10 genes per mb. The highest density is on Chr. 6, where a 700 kb region contains 60 genes. Gene rich regions tend to have higher than average G+C content. There are also **gene deserts**, regions >1 mb containing no genes. The largest gene desert is 4.1 mb. **Big genes** have coding sequences that span at least 500 kb. Dystrophin spans 2.3 mb.

Kong A. et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241-247
http://www.nature.com/ng/journal/v31/n3/full/ng917.html

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics 29:217-222 http://www.nature.com/ng/journal/v29/n2/abs/ng1001-217.html

Problems: Ch. 10—11b, 12-14, 23, 28, 29

Additional problems

30.1 What features do loci have to share for them to be usable for genetic mapping?

Ans. They must be easily genoptyped in large numbers of individuals. They must Mendelize. It is very useful if they are co-dominant.

30.2 If there were no recombinational hotspots in the region that Jeffreys et al. tested, what would you expect the results to look like?

Ans. Recombination events would be scattered uniformly throughout the region.

30.3 What is necessary before you can use the FISH protocol for gene mapping?

Ans. You have to have a clone of the gene you want to map.