Physical mapping of DNA

In testis, spermatogonia divide mitotically continuously. More precisely, there are spermatogonia stem cells that divide asymmetrically, one offspring being stem cell again (self-renewal), one proceeding though several more mitoses to form primary spermatocytes, which enter meiosis (in puberty). Meiosis is symmetrical, creating two secondary spermatocytes in meiosis I, which in turn create each two haploid spermatids. The spermatids differentiate into spermatozoa during process called spermiogenesis (making flagellum, acrosome, condensing DNA in nucleus by replacing histones by protamines).

In contrast to linkage mapping, which was first performed (by TH Morgan in fruitflies) using only the polymorphic phenotype, without definitive knowledge about biochemical basis of heredity, physical mapping usually requires DNA, only in hybridoma techniques also the gene product (RNA or protein) can be used. Many of the methods listed below are now obsolete or little used in human genetics, since we have the reference human genome since 2001. The genome sequence is in principle the most detailed physical map. Therefore only examples are listed, this is not exhausting list.

Low resolution techniques help assign genes to chromosomes

hybridoma technique

human cells are fused with hamster cells (using polyethylene glycol), the resulting hybrids are unstable at the beginning, losing some chromosomes. If you grow more different hybrid cell clones, each clone has a different set of chromosomes. Therefore if you have a human specific enzyme or other protein you can assay in a set of human-hamster hybrids, you can compare presence of the assayed protein with presence of different chromosomes – if you find a match, you assign a gene producing that protein to a particular chromosome. Of course, you cannot be more precise.

FISH

Fluorescent in situ hybridization can place a gene to a particular chromosomal locus with cytogenetic precision of a chromosomal band – that is very roughly down to 10 Mbp. You need a good piece of the gene DNA (typically roughly 100 kbp or more). This DNA is converted to labeled probe (usually by some kind of in vitro replication with labeled nucleotides). Lymphocytes are used to prepare mitosis spreads (like for karyotyping), but instead of Giemsa staining, the chromosome DNA is denatured, the probe is also denatured and applied to the chromosomes and renatured. DNA hybrid is formed and visualized with fluorescence. Single copy gene should present with four signals, two on each homologous chromosome.

High resolution techniques pave the way for sequencing

Because no sequencing method nowadays surpasses 10 kbp, and typically less than 1000 bp readable sequence is obtained from a single DNA fragment, it is impossible to sequence a whole chromosome in one read, chromosomes being in the order of 100 Mbp. Therefore the chromosome (or genome as a whole) has to be first fragmented, the fragments sequenced and then the fragments assembled together to derive chromosome sequence. The assembly of DNA fragments that can be somehow manipulated (e.g. sequenced) is called a DNA library. One typical way is ligating the fragments into plasmids that can be propagated in bacteria, other way is ligation to short oligonucleotides. To enable assembly of the chromosome from fragments, the fragments have to be overlapping. Therefore it is impossible to use a single haploid cell, multiple DNAs for each chromosome are needed, each fragmented in a different way (randomly, if possible). This can be achieved by various endonucleases or mechanical shearing (e.g. by ultrasound). There is one big problem for sequence assembly – repetitive sequences that form more than 50% of our genome makes it quite hard (also the shear amount of necessary fragments – 3 Gbp is the haploid genome; that means 3 milion fragments not counting the overlaps, typically you need at least 10-30times more to account for overlaps and uneven coverage of different loci. Therefore, in the publicly funded human genome project, a strategy was used to first make a DNA library with extra-large fragments – in the order of 100 kbp, therefore you need “only” 300 000 clones for 10x average coverage. Such libraries are typically propagated as special large bacterial plasmids called “bacterial artificial chromosomes” (BACs). They were derived from F-plasmids, but the genes for sex pili formation were excised. These BACs were than searched for overlaps between them to form a set of contiguous overlapping fragments (so called “contig”). From the contig, a “minimal tiling path” of individual BACs was selected for sequencing. Regarding our topic of physical mapping, the contig is a high-resolution physical map of the chromosome. Methods for searching for the overlaps are various, based on hybridization, PCR of sequenced ends or comparing pattern of restriction endonuclease digest, but these are well beyond the scope of this text. Anyway, all these methods are very laborious and in principle it was an important factor that caused a single human genome cost of billions USD. Consequently, private company Celera Genomics was first to develop methods that can assemble the genome without ordering the fragments into contigs before sequencing. See next section.

Sequencing strategies that enable de novo genome assembly

As described in previous section, assembly of the chromosome (and whole genome) sequence from overlapping fragments encounters a barrier of repetitive sequences. Strategies that are working around this are usually based on paired reads. The simplest is paired end reads. In this procedure, DNA library is generated using fragments that are too long to be sequenced completely, e.g. 2 kb for readings of 250 bp. So only both ends are sequenced, not the middle of the fragment, but the information about the ends belonging to each other, and the size of the missing middle piece is retained. In this way, sequence can be assembled around the repetitive sequences (in our example repeats not much longer than 1.5 kb). To bridge longer repeats, the fragments may be too long for proper handling during the sequencing preparation, so the fragments are first prepared as long, but then the ends are connected and the middle is cut off to reduce the size. These constructs are called mate-pairs. Even with these measures, it is impossible to sequence the largest tandem repeats, found in centromeres, telomeres and subtelomeric regions and short arms of acrocentric chromosomes.