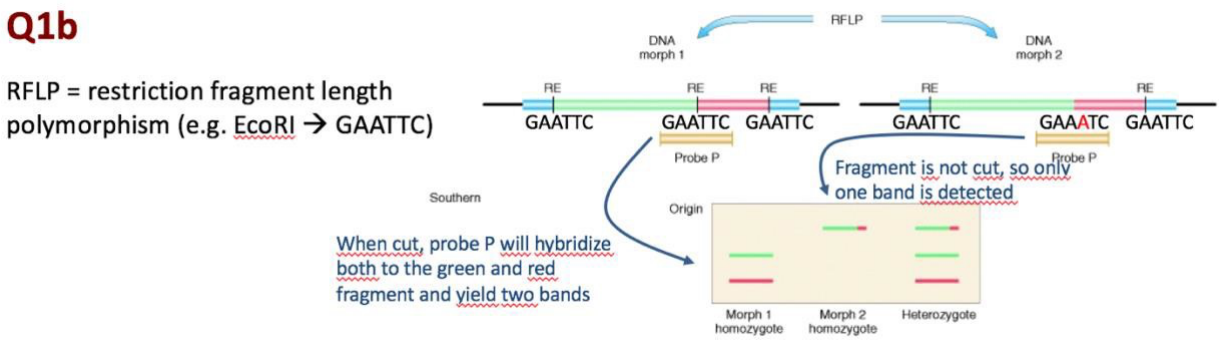
Questions and Answers Lecture 1: Introduction to Genomics

1) a. What was David Botstein's contribution to the Human Genome project, b. explain the underlying principles, and why was it crucial for the project to succeed?

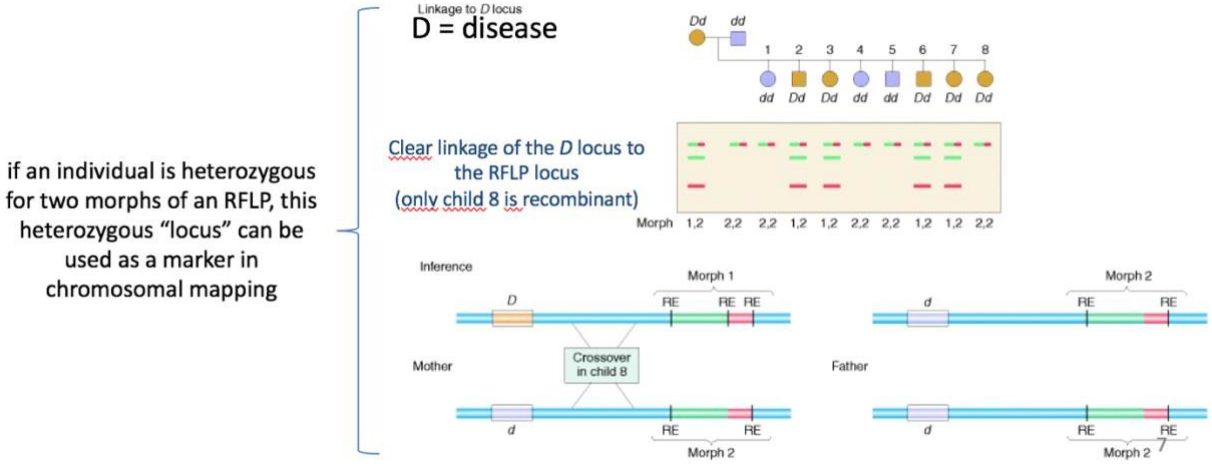
The Human Genome Project (HGP) was an international scientific research project with the goal of determining the sequence of nucleotide base pairs that make up human DNA, and of identifying and mapping all of the genes of the human genome from both a physical and a functional standpoint.

Botstein's contribution – Genetic linkage map using restriction fragment length polymorphisms (RFLP), used for first molecular genotyping assays

Principle:



By performing this on many individuals, using different RFLPs, we can start to examine which RFLPs are linked and thus genetically close (this principle is shown using a "disease" phenotype as a more intuitive read-out, but in practice, this "disease marker" is replaced by another RFLP.



Doing this systematically for many RFLPs allows one to start constructing a genetic map of the genome, which shows the linear order of the genes along a chromosome with distance proportional to the frequency of recombination (i.e. the closer two genes are, the greater the probability that they will crossover together). Using this genetic map provided a first guide (like road signs) regarding which sequence fragments are close to one another (which chromosome, which arm etc.), thus allowing the construction of a physical genome sequence map.

2) Define 1 cM. What's the physical equivalent in humans? Is it the same in Drosophila for example?

1 cM – Centimorgan - 1 genetic map unit = 1% chance (=1 out of 100 recombinations) that a marker at one genetic locus on a chromosome will be separated from a marker at a second

locus due to crossing over in a single generation.
About 1Mb in humans *vs 500* kb in *Drosophila melanogaster*.

3) A. What was the major difference in sequencing approach between the "public" and "private" Initiatives? Explain the underlying principles and why there was common belief that the private effort would fail.
Chromosome walking *vs*. Shotgun sequencing.
Chromosome walking is a technique with which an unknown region of a chromosome can be explored. It is generally used to isolate a locus of interest for which no probe is available but that is known to be linked to a gene which has been identified and cloned. A fragment containing a known gene is selected and used as a probe to identify other overlapping fragments which contain the same gene. The nucleotide sequences of these fragments can then be characterized. This process continues for the length of the chromosome.

Shotgun sequencing is a method used for sequencing DNA strands. DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain reads. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing.
The downside it presented was the challenging assembly of fragments (solved by TIGR, at least for less complex organisms).

B. Sanger sequencing revolutionized genome biology. Explain briefly how it works and which "upgrades" it required by Smith and Hunkapillar to make it useful for sequencing the human genome.
* Sanger sequencing (chain termination sequencing, 1975) is a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. It uses radioactivity to visualize bands (4 reactions, each involving one kind of ddNTP).
* Smith and Hunkapillar introduced fluorescent dyes and the use of a capillary column, resulting in only one and automated reaction.

4) When Bill Clinton announced in June 2000 the availability of the human genome sequence, was his statement correct? Please clarify.
No, published was merely a draft, no single chromosome was accomplished at that time. Chromosome 1 only fully sequenced in 2006.

5) A. Explain the principle of Illumina sequencing. B. Name other sequencing approaches and describe the principle underlying each in 2-3 sentences.
Samples of DNA were normally sheared into a random library of 100-300 base-pair long fragments. After fragmentation, the ends of the obtained DNA- fragments are repaired and an A- overhang is added at the 3'-end of each strand. Afterwards, adaptors which are necessary for amplification and sequencing are ligated to both ends of the DNA-fragments. These fragments are then size selected and purified.

NEW: Because such ligation is inherently inefficient, random shearing is now routinely replaced by transposase (Tn5)-based fragmentation whereby the Tn5 carries P5 / P7-compatible adapters which enable a short PCR (amplification) with P5 / P7 primers such that now you obtain also 300 bp DNA fragments with P5 / P7 adapters at both sides….

The Cluster Generation is performed on the Illumina cBot: Single DNA-fragments are attached to the flow cell by hybridizing to oligos on its surface that are complementary to

the ligated adaptors. The DNA-molecules are then amplified by bridge amplification which results in hundreds of millions of unique clusters. Finally, the reverse strands are cleaved and washed away and the sequencing primer is hybridized to the DNA-templates.
During sequencing the huge number of generated clusters are sequenced simultaneously. The DNA-templates are copied base by base using the four nucleotides (ACGT) which are fluorescently-labelled and reversibly terminated. After each synthesis step, the clusters are excited by a laser which causes fluorescence of the last incorporated base. Next, the fluorescence label and the blocking group are removed allowing the addition of the next base. The fluorescence signal after each incorporation step is captured by a built-in camera, producing images of the flow cell.
(please note that it would be helpful to support this description with illustrative figures)

- Pacific Biosciences: single molecule real time sequencing (SMRT) based on a fluorescent pulse but unlike reversible terminators, real-time nucleotides do not halt the process of DNA synthesis.
- Nanopore sequencing – DNA sequencing based on threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side to the other side of the membrane, which in turn affects the current created by the ion flow. Each base blocks the flow to a different degree, specifically altering the current.

6) *A. How does genome size relate to complexity? B. And what about the gene number?*
No correlation between complexity and number of genes.

7) *What's the difference between homologs, orthologs, and paralogs?*
Homolog (qualitative term -> % identity, not % homology) - Overall
A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship between genes separated by the event of genetic duplication (see paralog).
Ortholog (between species): Genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identifiction of orthologs is crtitical for reliable prediction of gene function in newly sequenced genomes.
Paralogues: (within genome): Genes related by duplication often (bot not always) within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

8) What is meant when it is stated that the human FOXP2 has undergone accelerated evolution?
Gene mutation in people with facial motor control and mental processing of language. Evolutionary stable protein underwent accelerated changes and is believed to have contributed to human-specific traits. Proposal that aa composition of human FOXP2 has undergone accelerated evolution, with change occurring around the time of language emergence in human. However, this is still being questioned and recent evidence using a higher number of species revealed that these changes may not be sapiens-specific but already be present in other hominid species. This does not exclude the important role of FOXP2 in shaping language abilities, but it suggests that it is a more complex picture than was so far appreciated.

9) *What are syntenic genomic regions?*

Co-localization of genes on chromosomes of different species.
Very conserved regions can reflect selection for functional relationship between genes in a region.

10) Can humans and chimps interbreed?
No, humans have 23 chromosome pairs while apes have 24, so there will be pairing issues during meiosis, likely rendering any offspring infertile…..

11) Provide a rough proportional breakdown of the human genome in constrained, repetitive, and coding sequence.
~ 5% constrained, ~ 42% repetitive, ~ 2% coding

12) What is the difference between LINEs, SINEs, and DNA transposons, and which constitute the largest genome portion? And what's the difference between an autonomous and non-autonomous repetitive element?
LINE elements have shed the viral body and act as selfish genes (automomus) by simply staying in the host, whereas SINEs (non-autonomus) depend on LINEs to be transcribed. LINEs – largest portion of the genome (~ 21%) vs. SINEs (~ 13% and DNA transposons (~ 3%). LINEs code for reverse transcriptase and go through an intermediate RNA phase.
DNA transposons, on the other hand, code for transposase (or related transposase) and insert double-stranded DNA into host genome.

13) Which kind of repetitive elements enlarges the genome? Explain the mechanism.
LINEs and SINEs insert themselves into the host genome through RNA pol II of the host. The ORFs code for reverse transcriptase with high specificity for L1 mRNA as well as for endonuclease, allowing integration in the genome.

14) A. What's the difference between a micro- and mini-satellite? B. Describe how they may originate.
Satellites: section of repeated DNA stretches (e.g. GGGCAGG).
Micro-satellite: 30-300 bp locus size, ~ 200,000 loci in population, 2-5 bp repeat unit, might originate from an event of the DNA polymerize re-annealing out of register during replication, after completion of replication DNA repair mechanism, a new allele is created.
Mini-satellite: 1-20 kp locus size, ~ 30,000 loci in population, 10-100 bp repeat unit, might originate from a misalignment of repeat containing sequence and an unequal crossing-over at the end of which two resulting alleles have differently sized mini-satellites.

15) Explain the mechanism of DNA fingerprinting and a concrete application.
DNA fingerprinting is the process of determining an individual's DNA characteristics, called a DNA profile, that is very likely to be different in unrelated individuals, thereby being as unique to individuals as are fingerprints. DNA profiling should not be confused with full genome sequencing.
PCR-based DNA profiling: Determine sequences flanking microsatellites -> Amplify alleles by PCR -> Analyse PCR products by gel electrophoresis and staining (now: increased sensitivity and automation).

DNA profiling is most commonly used as a forensic technique in criminal investigations to identify an unidentified person or whose identity needs to be confirmed, or to place a person at a crime scene or to eliminate a person from consideration.
DNA profiling has also been used to help clarify paternity, in immigration disputes, in parentage testing and in genealogical research or medical research. DNA fingerprinting has also been used in the study of animal and floral populations and in the fields of zoology,

botany, and agriculture.