**What is Genomics?**

- Study of complete set of genes
- Global vs local
- Genome Projects ([Table](#))
- Structural vs. Functional genomics

**Steps in Whole Genome Mapping:**

- High resolution genetic mapping
- Physical mapping
- DNA sequencing

**Applications of the Complete DNA Sequence**

- Functional Genomics
- Bioinformatics

---

**High Resolution Genetic Mapping**

- used to place molecularly defined differences on linkage maps or cytogenetic maps
- provide molecular landmarks for building higher resolution physical and sequencing maps
- builds on **low-resolution genetic maps**

**DNA polymorphisms:** molecularly defined differences between individuals

**Mapping Techniques Used to Determine Position of DNA Marker on a Chromosome:**

1) **Meiotic Recombination Maps**

- based on analyzing recombinant frequency in dihybrid and multihybrid crosses
- done more easily in *Saccharomyces cerevisiae, Ceanorhabditis elegans, Drosophila melanogaster* etc.
- due to ease of controlled experimental crosses
- Not as easy for humans since crosses are harder to obtain and progeny sizes are too small
- measurements of loci with known phenotypic effect showed intervals in between genes which contain vast amounts of DNA
- To fill in these gaps, various kinds of **polymorphic DNA markers** need to be exploited

Examples of polymorphic DNA markers (RFLPs, SSLPs, and RAPDs):
a) **RFLPs**
   - Restriction Fragment Length Polymorphism
   - Restriction enzyme recognition sites present in some strains but not in others (presence/absence)

b) **SSLPs**
   - Simple Sequence Length Polymorphisms
   - Advantages over RFLPs include larger number of alleles as well as higher levels of heterozygosity

**Types of SSLPs:**

i) **Minisatellite markers**
   - based on VNTRs: Variable Number of Tandem Repeats
   - VNTRs are 1 - 5 kb sequences of repeating units that are 15 - 100 nucleotides long
   - When genome cut by restriction enzymes, and there are no recognition sites located within VNTRs, Southern blot analysis reveals a large number of different sized fragments bound by VNTR probe.
   - Due to high variability in number of tandem repeats from person to person, set of fragments revealed is highly individualistic.
   - Often called DNA fingerprints

ii) **Microsatellite markers**
   - based on variable number of dinucleotides repeated in tandem
   - CA with complementary GT most common
   - Use PCR to make probes for DNA flanking marker. First digest the DNA with a restriction enzyme such as Alu1. Clone the fragments into a sequencing vector and then identify those containing the CA/GT repeats with a CA/GT probe. Sequence these vectors and create PCR primer pairs. These primers are designed to recognize single copy DNA sequences flanking the marker. Use these primers with genomic DNA in PCR amplification. Gel electrophoresis can be used to determine size differences.

c) **RAPDs**
   - Randomly Amplified Polymorphic DNA
   - based on random PCR amplification
   - A single PCR primer is designed at random and used to amplify different regions of the genome.

Also **SNPs should be mentioned.**

- SNPs are Single Nucleotide Polymorphisms
- Many due to neutral variation, such as third codon position
- There is one SNP difference every 1000 base pairs of human DNA. Since the human genome is about 3 billion base pairs long there are 3 million differences between any two of our genomes. A very rich source of variation!

2) **Cytogenetic Maps**
   - produced by relating locations of DNA markers to cytogenetic landmarks such as chromosome bands and puffs
Ways to do this:

a) **In Situ Hybridization Mapping**
   - if a cloned DNA sequence is available for area of interest, label it as a probe and use it to hybridize to chromosomes in situ
   - individual chromosomes are recognizable through morphology differences such as size, banding pattern, centromere location
   - map the probe sequence to approximate position on chromosome
   - labels used for probes: **FISH** - Fluorescent In Situ Hybridization

b) **Rearrangement Breakpoint Mapping**
   - based on using DNA breakpoints as molecular landmarks
   - when cloned DNA spanning a breakpoint has been identified, breakpoints are easily detected on Southern blots as two bands of hybridization, while in normal chromosomes you would only see one band

c) **Radiation Hybrid Mapping**
   - does not require marker heterozygosity
   - irradiate cells with Xray fragments to break up the chromosomes (human)
   - insert these fragments into rodent cells
   - fragmented human chromosomes fuse to chromosomes of rodent cells
   - create a series of clones each containing a different random assortment of fragments of human chromosomes
   - isolate and denature DNA from each cell line
   - introduce a labeled human DNA probe to identify positions of human DNA homologous to probe
   - analyze the data from many probe hybridizations to determine co-retention of DNA markers
   - co-retention of different human markers allows high-resolution mapping of the chromosomal loci of the DNA markers

---

**Physical Mapping**

- Physical mapping is an intermediate step in sequencing the entire genome
  (genetic map > physical map > sequence map)
- A complete physical map of the genome includes:
  - maps for each chromosome in the haploid chromosome set
  - for each chromosome, continuous overlapping cloned genomic DNA segments extending from one telomere of the chromosome to the other

**Vector** - plasmid or phage chromosome used to carry cloned DNA segment (or insert)  Chapter 8, MGA
- Main types used: YAC (Yeast Artificial chromosomes) or Cosmids
  - BAC (Bacterial Artificial Chromosomes)
  - PAC (Phage P-1 based Artificial Chromosomes)

**Contig** - set of ordered overlapping clones that constitute a a chromosomal region or a genome

**Techniques for Identifying Clone Overlaps:**
1) **Ordering by Clone Fingerprints**

- genomic insert (clone) carried by vector has a unique sequence that can be used to generate a **DNA fingerprint**
- multiple restriction enzyme digestion generates a set bands, unique in number and position, representing the fingerprint for a particular clone
- patterns of bands from multiple clones are read by computer and aligned to determine the degree of overlap between inserted DNA segments
- the proportion of bands shared between two clones indicates whether there is true overlap (usually 20-25%)
- important technique in developing physical maps for *C. elegans*, mouse and human genomes

2) **Ordering by STSs**

- Sequence-Tagged Sites are short unique sequences that can be amplified using defined PCR primers
- derive from sequenced regions of the genome so can be used as landmarks for clone classification in creating physical map
- clones that share STSs must overlap; the more STSs they share, the more they overlap
- resulting physical map is a **STS content map**
- combination of fingerprinting and STS content mapping has resulted in complete and near-complete physical maps for many organisms, such as *C. elegans*

**Simplifying Physical Mapping by Subdividing the Genome:**

- many biological and technical challenges in creating physical maps that are true reflections of the genome
- in human genome, there are megabase-size regions that are duplicate copies on two different chromosomes (biological challenge)
- some regions of the genome do not clone efficiently in standard vectors, leaving gaps in the physical map (technical challenge)
- subdividing the genome into smaller working entities can circumvent some biol and tech challenges as the number of clones required to complete the physical map in a given subregion is much less

1) **Chromosome Specific Libraries**

- separate actual DNA molecules of the genome into those contained within specific chromosomes
- library serves as source of clones for fingerprint or STS content physical mapping
- individual chromosomes are identified

**Techniques:**

a) **Pulse Field Gel Electrophoresis (PGFE)**  *Chapter 2, MGA*

- modification of standard gel electrophoresis that adjusts conditions to permit separation of large DNA molecules
- can isolate individual chromosomes if they are small (eg. yeast chromosome)
- for large chromosomes, can isolate chromosome fragments by "cutting" with "rare cutter" enzymes (eg. *NotI* enzyme cuts every 64,000 bp)

b) **Fluorescence-Activated Chromosome Sorting (FACS) or Flow Sorting**

- cells disrupted to liberate whole metaphase chromosomes into liquid suspension
- chromosomes stained with 2 dyes, one binding to AT-rich regions, the other to GC-rich regions
- each chromosome has a unique ratio of AT-rich to GC-rich regions, used to distinguish between different chromosomes
2) **Ordering by FISH (Fluorescent In Situ Hybridization)**

- technique used to confirm physical map order of cloned DNAs
- many clones in situ hybridize to the same landmark regions (using human chromosome banding techniques), thus, FISH puts clones into one of a number of cytogenetic regions within a given chromosome
- clones then evaluated by fingerprint analysis or STS content mapping to produce physical map
- provides independent way to corroborate results with the physical maps produces from fingerprint or STS content mapping
  
  Cytogenetic map of BAC and PAC clones localized by FISH mapping in human genome

---

**DNA Sequencing**

- Four bases include A, C, T, and G
- Human genome equals 3 x 10^9 base pairs and includes an X and Y chromosome as well as 22 autosomes
- All current sequencing techniques are clone based
- First make a clone or subclone library and then sequence all or part of inserts of individual clones in the library. From these sequences form a consensus sequence

**There are Two Ways to Assemble a Consensus Sequence:**

1) **Ordered Clone Sequencing**

- produce physical map of genome
- ordered subset of minimally overlapping clones selected for sequencing
- consensus sequence for each clone
- assemble in order on physical map

2) **Whole Genome Shotgun Sequencing**

- obtain sequence reads from randomly selected clones from whole genome library
- no information on where clones map in genome
- homologous sequence allows assembly of sequences into consensus sequence over whole genome

**Sequencing Strategies in Bacteria:**

- bacterial DNA is single copy and only a few megabase pairs in size
- due to simple system, whole genome shotgun assembly can be applied
- gaps in consensus sequence can be filled in by primer walking
  
  **Primer Walking** - use of a primer based on a sequenced area of a genome to sequence into a flanking unsequenced area
- shotgun sequencing does not work well in eukaryotic systems since it is not composed entirely of single copy DNA and may contain repetitive genome sequences

**Repeated Genome Sequences:**

- repeated genome sequences are identical sequence strings present many times in the genome
- problematic in eukaryotic systems
1) **Tandem Repeat Arrays**

- **tandem repeats** are sequences in multiple copies adjacent to one another, variable in size and number of repetitions
  (Recall: VNTR-minisatellites, microsatellites).
  
  a) **Tandemly repeated genes**  
  b) **Non-coding tandem repeats** - telomeres and heterochromatin

2) **Mobile Genetic Elements: Dispersed Repeats**  
   (Summary Table)

- dispersed in genome and move to new locations via **transposition**
  
  a) **Transposons**
  b) **Retrotransposons**
  c) **LINE** (long interspersed elements)
  d) **SINE** (short interspersed elements)

**Tackling Genomes with Repetitive Sequences**

1) **Assembling a Sequence from Ordered Clones**

- straightforward assembly of many of the dispersed repeats since they are present only once in the individual clone
  
  - **Minimum Tiling Path** is a subset of clones with clear but minimal overlap (ie. minimum # of clones that represent entire genome)
  - relies on physical map to order and orient the clone sequences

2) **Whole Genome Shotgun Assembly**

- connects the single-copy sequences on either side of the repetitive element but ignores the sequence of the repetitive element itself.
  
  - sequenced clones aligned by their homologous sequence overlaps into **contigs** (but in no particular order)
  - **paired-end sequences** (sequences corresponding to either end of cloned insert) are used to span gaps between **contigs** and place them in correct genomic order and orientation
  - **scaffolds** - ordered set of contigs in which there are unsequenced gaps connected by paired-end sequence reads

For a visual comparison of these methods see: Figure 9-29

---

**Functional Genomics**

- functional genomics includes study of expression and interaction of gene products on a **global** level, that is, using genomic approaches to study some aspect of all gene products simultaneously
- how molecules cooperate and interact to effect all the processes and phenotypes that make up a biological system
- genome refers to "gene" plus "ome", or the global data set for "all genes"
- various other 'ome's are being worked on: transcriptome, proteome, interactome and phenome
Studying the Transcriptome and Interactome Using DNA Chips:

**DNA chips:** chips the size of a microscope cover slip which contain samples of DNA laid out in series

- automation and miniaturization of assay methods
- contain samples of DNA laid out as a series of microscopic spots bound to a glass "chip"
- can contain all genes of complex genome
- can assay all gene products in a single experiment
- method alternative to mutational analysis; rather than amassing mutations to disrupt a particular process, chip technology detects the specific mRNAs expressed in that process
- can also be used to detect protein-DNA interactions

Constructing DNA Chips:

- microscopic droplets of DNA added to slide via a robotic machine (thousands of samples can be applied to one chip)
- DNA dried and treated to bind to glass

1) **One protocol detecting which genes are active at a particular stage of development in a cell:**

- array of known cDNAs from different genes are applied to chip
- chip exposed to *fluorescently labelled probe*, such as, RNA extracted from particular cell at particular stage of development
- binding of probe molecules to homologous DNA spots monitored automatically by laser beam-illuminated microscope
- detect spots on chip where probe binds to determine which genes are active at the particular stage of interest
- [Animation](#)

2) **Another protocol for building oligonucleotides for detection of active genes:**

- array of *oligonucleotides* are chemically synthesized on chip, one nucleotide at a time
- chip covered with protecting groups that prevent DNA deposition
- mask placed on chip containing holes where sites of deposition are to occur
- shine a laser beam on holes where synthesis will begin, this will knock off protecting groups
- bathe chip in first nucleotide to be added (containing protective group to avoid adding dimers)
- sequential additions of laser beams, appropriate masks and bathing in nucleotides allow for construction of oligonucleotide
- once this is done, these chips are ready to bind to fluorescent probes isolated at some developmental stage of interest
- chip is analyzed similar to method above
- [Animation](#)

Studying the Interactome with the Yeast Two-Hybrid System:

- investigates interactions between proteins [Figure 9-40](#)
- uses the yeast GAL4 transcription activator
- GAL4 has two domains: a DNA binding domain which binds to site of transcriptional activation and an
activation domain which is responsible for activating transcription, but cannot do so without the DNA binding domain
- gene for GAL4 is divided between plasmids
- gene for protein of interest is spliced next to DNA binding domain of GAL4 = bait
- the other protein gene is spliced next to the rest of the GAL4 gene on other plasmid = target
- both plasmids introduced into cell or cells of organism together and observed for activation via a reporter gene (gene for an easily detected protein)

Bioinformatics

- deciphering meaning from the raw 4-letter DNA sequence by using computational analysis to predict mRNA and polypeptide sequences.

Problems with deciphering information content of DNA:

1. Do not know all of the specific DNA sequences that encode the thousands of docking sites for DNA or RNA-binding regulatory proteins.

2. A given DNA sequence can encode for different things depending on its location within the DNA. If located in coding region, the sequence would code for amino acid, if located in non-coding region, the sequence would act as binding site for regulatory protein.

3. Two or more different sequences can serve the same function.

Using Bioinformatics to Determine an Organisms Proteome

- proteome (complete set of polypeptides encoded by a genome)

Bioinformatics uses several independent sets of information to do this:

1. cDNA sequences (complimentaryDNAs are DNA copies of mRNAs) cDNAs are aligned with genomic DNA to determine the position of introns and exons.

2. Docking site sequences marking the start and end points for the events in information transfer (transcription, pre-mRNA splicing, translation).

3. Sequences of related polypeptides. Common statistical tool for aligning proteins is BLAST (Basic Local Alignment Search Tool)

4. Codon bias - species-specific usage preferences for some codons over other encoding for the same amino acid. Presence of the preferred codon in predicted mRNA sequence supports the accuracy of the prediction.

Predictions of mRNA and polypeptide structure from genomic DNA sequence depend on an integration of information from cDNA sequence, docking site predictions, polypeptide similarities, and codon bias. Summary Figure