Lectures 30 and 31 – Genome analysis

I. Genome analysis
   A. two general areas
      1. structural
      2. functional
   B. genome projects – a status report
      1. 1st sequenced: several viral genomes
      2. mitochondria and chloroplasts
      3. many bacteria
      4. many complete nuclear genomes have been sequenced

<table>
<thead>
<tr>
<th>species</th>
<th>genome size (Mb)</th>
<th>date completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli (bacteria)</td>
<td>4.64</td>
<td>1997</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (yeast)</td>
<td>12.07</td>
<td>1996</td>
</tr>
<tr>
<td>Caenorhabditis elegans (nematode)</td>
<td>100</td>
<td>1998</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>180</td>
<td>2000</td>
</tr>
<tr>
<td>Arabidopsis thaliana (mustard weed)</td>
<td>125</td>
<td>2000</td>
</tr>
<tr>
<td>Fugu rubripes (pufferfish)</td>
<td>380</td>
<td>2002</td>
</tr>
<tr>
<td>Oryza sativa (rice, two strains)</td>
<td>466</td>
<td>2002 (draft)</td>
</tr>
<tr>
<td>Mus musculus (mouse)</td>
<td>3200</td>
<td>2002*</td>
</tr>
<tr>
<td>Homo sapiens (humans)</td>
<td>3200</td>
<td>2002* (draft)</td>
</tr>
</tbody>
</table>

C. why perform whole genome analyses?
   1. genomic DNA is blueprint for making organism
   2. comparison – shows important sequences
   3. human genomics – clone genes associated with genetic disorders
   4. many projects generate ordered clones that cover genome

D. general strategy
   1. position genes and molecular markers along chromosome map
   2. position individual cloned DNA fragments relative to one another
   3. sequence DNA
   4. assemble sequence

II. Clone genome into workable sized pieces
   A. goal is ordered set of overlapping clones that together cover the entire genome
      1. first clone large pieces into YACs (or BACs, PACs)
      2. order the clones into contigs
         a. by FISH (fluorescence in situ hybridization)
b. by sequence tagged sites (STSs)
   - unique sites on chromosome, often specific PCR products
   - if from cDNAs, they’re called expressed sequence tags (ESTs)

c. by other means
d. generate smaller (often cosmid) clones, and order them

B. sequence the genome
   1. sequence individual cosmids
      - assemble the sequence
   2. alternative – sequence random clones

III. Physical chromosomal maps
   A. Can map polymorphisms (sequence differences between individuals in population)
      1. single nucleotide polymorphisms (SNPs)
         - May alter restriction site. Produces restriction fragment length polymorphism (RFLP)

         1. \( \text{EcoRI} \)
            \[
            \begin{array}{ccc}
            & 1.2kb & 5.6kb & 3.1kb \\
            \text{YAC1} & - & - & + \\
            \text{YAC2} & - & + & - \\
            \text{YAC3} & - & - & - \\
            \text{YAC4} & + & - & - \\
            \text{YAC5} & - & - & - \\
            \end{array}
            \]

         2. \( \text{EcoRI} \)
            \[
            \begin{array}{ccc}
            & 6.8kb & 3.1kb \\
            \text{YAC1} & - & - & - \\
            \text{YAC2} & + & - & - \\
            \text{YAC3} & - & - & - \\
            \end{array}
            \]

2. eg. of use of RFLP
3. simple sequence repeats (SSRs)

B. Can map genetic markers (genes which when mutant produce phenotype) relative to physical markers

IV. The Human Genome
A. Genes
   1. Number of genes
   2. Types of genes
      a. noncoding RNAs
         i.
         ii.
         iii.
         iv.
         v.
      b. protein - increased complexity vs. simpler eukaryotes
         i. somewhat more genes
         ii. arrangement of domains
         iii. more paralogs
            - paralog
         iv. alternative splicing
         v. chemical modifications

B. Repetitive DNA (>50% of human DNA)
   1. transposon derived (~45%)
   2. pseudogenes
   3. SSRs (3%)
4. segmental duplications of 10 - 300kb pieces (5%)
5. repeats at centromeres, telomeres and other

C. Types of gene organization
1. gene families - some genes are present in many copies
   a. different copies may provide same or different function
   b. olfactory receptor genes - 1000 in humans
      - evolved in two ways:

2. gene rich regions
   a. major histocompatibility complex (MHC) on chromosome 6, 60 genes within 700kb
3. gene deserts - 82 currently recognized
   a.  
   b. in total span 144 Mb or 3% of genome
   c. why?
      i.
      ii.
      iii.

D. Combinatorial strategies to increase diversity
1. at DNA level

2. at RNA level

E. Use of genome info: eg – cloning of cystic fibrosis gene
1. mapped gene to small interval between met (proto-oncogene) and D788 (polymorphism)
2. used chromosome walk to isolate clones
3. sequenced clones to identify candidate genes
4. identify gene:
   a. looked for gene to be present in other species
   b. ought to look like a gene
   c. gene should be transcribed
d. some sweat gland symptoms

e. found mutations in affected individuals in one gene

f. wild-type gene rescued cells from affected individuals

F. now, if clones available, get them; if genome sequenced, screen for candidates

V. Functional genomics - determine function of all genes

A. Mutation
   1. identify all ORFs in sequence
   2. systematically knock out all predicted genes
   3. assess phenotypes produced
   4. limitations:
      - many are lethal
      - many produce subtle mutant phenotypes

B. look for protein interactions
   1. Why?
      a.

      b.

   2. How?
      a. identify suppressor mutations

      b. protein interaction by yeast two hybrid
         i. ~genetic method that can identify interactions
ii. interaction between “bait” and target brings GAL4 DNA binding and activation domains together
iii. gives transcription of lacZ
iv. used to test candidate interactions or to screen for unknown protein interactions

C. global mRNA regulation assayed by DNA chips
   1. uses:
      a. can identify all mRNAs expressed at certain developmental stage
      b. can look at total mRNA expressions under certain conditions
      c. can see how mutations affects total mRNA expression
   2. how:
      a. put DNA from every gene on chip
      b. fluorescently label cDNA probe
      c. hybridize to DNA on chip