BioSci D145 Lecture #2

• Bruce Blumberg (blumberg@uci.edu)
  - 4103 Nat Sci 2 - office hours Tu, Th 3:30-5:00 (or by appointment)
  - Phone 824-8573

• TA - Riann Egusquiza (regusqui@uci.edu)
  - 4351 Nat Sci 2 - office hours M 1-3 (this week, Wednesday 1:30-3:30)
  - Phone 824-6873

• check e-mail and noteboard daily for announcements, etc..
  - Please use the course noteboard for discussions of the material

• Updated lectures will be posted on web pages after lecture

• Don’t forget to discuss term paper topics with me
How about term paper topics?

- Example term papers are posted on web site

  - Specific aims
  - Background and significance
  - Research plan
  - References

  \{ ~2 pages
  \}
  \{ ~3 pages
  \}
  No limit (but be reasonable)

- Please e-mail me (or stop by and discuss) your topic as soon as possible

- Remember that your goal is to propose a study of something you find interesting that has not already been done
  - Exercise your imagination
  - Indulge your intellectual curiosity
  - Expand your BioSci 199 research interests.
Organization and Structure of Genomes (contd)

- Gene content is proportional to single copy DNA
  - Amount of non-repetitive DNA has a maximum, total genome size does not
  - What is all the extra DNA, i.e., what is it good for?
    - Repetitive DNA
    - Telomeres
    - Centromeres
    - Transposons
    - Junk of all sorts

- Where did all this junk come from and why is it still around?
  - DNA replication is very accurate
  - Selective advantage?
  - OR
Organization and Structure of Genomes (contd)

- What is this highly repetitive DNA?

- Selfish DNA?
  - Parasitic sequences that exist solely to replicate themselves?

- Or evolutionary relics?
  - Produced by recombination, duplication, unequal crossing over

- Probably both
  - Transposons exemplify “selfish DNA”
    - Akin to viruses?
    - Crossing over and other forms of recombination lead to large scale duplications

- BUT, note that the ENCODE (encyclopedia of DNA elements) considers almost 100% of genome to be functional.
Transcription of Prokaryotic vs Eukaryotic genomes (stopped here)

- **Prokaryotic genes are expressed in linear order on chromosome**
  - mRNA corresponds directly to gDNA

- **Most eukaryotic genes are interrupted by non-coding sequences**
  - Introns (Gilbert 1978)
  - These are spliced out after transcription and prior to transport out of nucleus
  - Post-transcriptional processing in an important feature of eukaryotic gene regulation

- **Why do eukaryotes have introns, i.e., what are they good for?**
  - Main function may be to generate protein diversity
  - Harbor regulatory sequences
Introns and splicing

- Alternative splicing can generate protein diversity
  - Many forms of alternative splicing seen
  - Some genes have numerous alternatively spliced forms
    - Dozens are not uncommon, e.g., cytochrome P450s
Introns and splicing

- Alternative splicing can generate protein diversity (contd)
  - Others show sexual dimorphisms
    - Sex-determining genes
    - Classic chicken/egg paradox
      - how do you determine sex if sex determines which splicing occurs and spliced form determines sex?
Origins of intron/exon organization

- Introns and exons tend to be short but can vary considerably
  - “Higher” organisms tend to have longer lengths in both
  - First introns tend to be much larger than others - WHY?

- Often contain regulatory elements
  - Enhancers
  - Alternative promoters
  - etc
Origins of intron/exon organization

- Exon number tends to increase with increasing organismal complexity
  - Possible reasons?
    - Longer time to accumulate introns?
    - Genomes are more recombinogenic due to repeated sequences?
    - Selection for increased protein complexity
  - Gene number does not correlate with complexity
  - therefore, it must come from somewhere

**Figure 2.23** Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totalled in the leftmost column.)
Origins of intron/exon organization

- When did introns arise
  - Introns early - Walter Gilbert
    - There from the beginning, lost in bacteria and many simpler organisms
  - Introns late - Cavalier-Smith, Ford Doolittle, Russell Doolittle
    - Introns acquired over time as a result of transposable elements, aberrant splicing, etc
    - If introns benefit protein evolution - why would they be lost?

  - Which is it?
    - Introns late (at the moment)

- What is common factor among animals that share intron locations?

  All deuterostomes (echinoderms, chordates, hemichordates, xenoturbellids - diverged about 580 x 10^6 years ago
Evolution of gene clusters

- Many genes occur as multigene families (e.g., actin, tubulin, globins, Hox)
  - Inference is that they evolved from a common ancestor
  - Families can be
    - clustered - nearby on chromosomes (α-globins, HoxA) 
    - Dispersed - on various chromosomes (actin, tubulin) 
    - Both - related clusters on different chromosomes (α,β-globins, HoxA,B,C,D) 
  - Members of clusters may show stage or tissue-specific expression 
    - Implies means for coregulation as well as individual regulation

![Figure 4.1](image-url) Each of the α-like and β-like globin gene families is organized into a single cluster that includes functional genes and pseudogenes (ψ). The organization of the clusters in higher primates is conserved. All of the active genes are transcribed from left to right.

<table>
<thead>
<tr>
<th>Stage of development</th>
<th>Hemoglobin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryonic (&lt;8 weeks)</td>
<td>ζ2ε2, ζ2γ2, α2ε2</td>
</tr>
<tr>
<td>Fetal (3–9 months)</td>
<td>α2γ2, α2δ2, α2β2</td>
</tr>
<tr>
<td>Adult (from birth)</td>
<td>α2γ2, α2δ2, α2β2</td>
</tr>
</tbody>
</table>
Evolution of gene clusters (contd)

- multigene families (contd)
  - Gene number tends to increase with evolutionary complexity
  - Globin genes increase in number from primitive fish to humans
  - Clusters evolve by duplication and divergence

**Figure 4.2** Clusters of $\beta$-globin genes and pseudogenes are found in vertebrates. Seven mouse genes include two early embryonic, one late embryonic, two adult genes, and two pseudogenes. Rabbit and chick each have four genes.
Evolution of gene clusters (contd)

- History of gene families can be traced by comparing sequences
  - Molecular clock model holds that rate of change within a group is relatively constant
  - Not totally accurate - check rat genome sequence paper
  - Distance between related sequences combined with clock leads to inference about when duplication took place

![Figure 4.4](image1.png)

**Figure 4.4** Divergence of DNA sequences depends on evolutionary separation. Each point on the graph represents a pairwise comparison.

![Figure 4.5](image2.png)

**Figure 4.5** Replacement site divergences between pairs of β-globin genes allow the history of the human cluster to be reconstructed. This tree accounts for the separation of classes of globin genes. Duplications of individual genes are of unknown origin. The time of the α-ζ divergence is not known.
Types and origin of repetitive elements

• DNA sequences are not random
  - genes, restriction sites, methylation sites
• Repeated sequences are not random either
  - Some occur as tandemly repeated sequences
  - Usually generated by unequal crossing over during meiosis
  - These resolve in ultracentrifuge into satellite bands because GC content differs from majority of DNA
  - This “satellite” DNA is highly variable
    • Between species
    • And among individuals within a population
    • Can be useful for mapping genotyping, etc
  - Much highly repetitive DNA is in heterochromatin (highly condensed regions)
    • Centromeres are one such place
Types and origin of repetitive elements (contd)

- Dispersed tandem repeats are “minisatellites” 14-500 bp in length
  - First forensic DNA typing used satellite DNA - Sir Alec Jeffreys
  - Minisatellite DNA is highly variable and perfect for fingerprinting

Figure 4.22: Alleles may differ in the number of repeats at a minisatellite locus, so that cleavage on either side generates restriction fragments that differ in length. By using a minisatellite with alleles that differ between parents, the pattern of inheritance can be followed.
### Types and origin of repetitive elements - dispersed repeated sequences

<table>
<thead>
<tr>
<th>Type</th>
<th>Structural Features</th>
<th>Mechanism of Movement</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DNA-MEDIATED TRANPOSITION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterial insertion sequences</td>
<td>≈50-bp inverted repeats flanking region encoding transposase and, in some, resolvase</td>
<td>Excision or copying of DNA and its insertion at target site</td>
<td>IS1, IS10</td>
</tr>
<tr>
<td>(IS elements)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterial transposons</td>
<td>Central antibiotic-resistance gene flanked by IS elements</td>
<td>Copying of DNA and its insertion at target site</td>
<td>Tn9</td>
</tr>
<tr>
<td>Eukaryotic transposons</td>
<td>Inverted repeats flanking coding region with introns</td>
<td>Excision of DNA and its insertion at target site</td>
<td>P element (<em>Drosophila</em>) Ac and Ds elements (corn)</td>
</tr>
<tr>
<td><strong>RNA-MEDIATED TRANPOSITION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viral retrotransposons</td>
<td>≈250- to 600-bp direct terminal repeats (LTRs) flanking region encoding reverse transcriptase, integrase, and retroviral-like Gag protein</td>
<td>Transcription into RNA from promoter in left LTR by RNA polymerase II followed by reverse transcription and insertion at target site</td>
<td>Ty elements (yeast) Copia elements (<em>Drosophila</em>)</td>
</tr>
<tr>
<td>Nonviral retrotransposons</td>
<td>Of variable length with a 3’ A/T-rich region; full-length copy encodes a reverse transcriptase</td>
<td>Transcription into RNA from internal promoter; folding of transcript to provide primer for reverse transcription followed by insertion at target site</td>
<td>F and G elements (<em>Drosophila</em>) LINE and SINE elements (mammals) <em>Alu</em> sequences (humans)</td>
</tr>
</tbody>
</table>
Types and origin of repetitive elements - dispersed repeated sequences

- **Main point is to understand how such elements can affect evolution of genes and genomes**
  - Gene transduction has long been known in bacteria (transposons, P1, etc)
  - LINE (long interspersed nuclear elements) can mediate movement of exons between genes
    - Pick up exons due to weak polyadenylation signals
    - The new exon becomes part of LINE by reverse transcription and is inserted into a new gene along with LINE
      - Voila - gene has a new exon
    - Experiments in cell culture proved this model and suggested it is unexpectedly efficient
    - Likely to be a very important mechanism for generating new genes
Genome Structure

- The big picture
  - Chromosomes consist of coding (euchromatin) and noncoding (heterochromatin) regions
    - Various physical methods can distinguish these regions
      - Staining
      - Buoyant density
      - Restriction digestion
    - Heterochromatin is primarily tandemly repeated sequences
    - Euchromatin is everything else
      - Genes including promoters, introns, exons
      - LINES, SINES micro and minisatellite DNA
    - Patterns of euchromatin and heterochromatin can be useful for constructing genetic maps
      - Heterochromatin is trouble for large scale physical mapping and sequencing
        - May be hard to cross gaps
Genome evolution

- Genomes evolve increasing complexity in various ways
  - Whole genome duplications
    - Particularly important in plants
  - Recombination and duplication mediated by SINEs, LINEs, etc.
    - Expands repeats, exon shuffling, creates new genes
  - Meiotic crossing over
    - Expands repeats, duplicates genes
  - Segmental duplication - frequent in genetic diseases
    - Interchromosomal - duplications among non-homologous chromosomes
    - Intrachromosomal - within or across homologous chromosomes
Genome evolution (contd)

- Several papers discuss details of genome evolution as studied in closely related species
  - Dietrich et al. (2004) Science 304, 304-7
  - S. cerevisiae vs two other species of yeast
    - Saw genome duplications and
    - evolution or loss of one duplicated member but never both

![Genome evolution diagram](image)
Mapping Genomes

• Why map genomes?
  - Locate genes causing mutations or diseases
    • Figure out where identified genes are
  - Prepare to sequence
  - Discern evolutionary relationships

• How do we go about mapping whole genomes?
  - restriction endonuclease digestion
    • Impossible for all but the tiniest genomes
    • Requires ability to precisely resolve very large fragments of DNA
  - Must be able to separate chromosomes or huge fragments thereof
    • Then map various types of markers onto these fragments
    • STS, ESTs, RFLPs
  - Modern approach
    • Construct large insert genomic libraries
      - Map relationship to each other
      - map markers onto large insert library members
    • Map to chromosomes
Mapping Genomes - comparison of maps

Physical map

Cytological map

Genetic map

Transcripts Insertions

0 10 20 30 40 50 60 cM

yw sev fog

genes

A PRIMER OF GENOME SCIENCE 3e, Figure 1.3

Construction of genomic libraries

- What do we commonly use genomic libraries for?
  - Genome sequencing (most approaches use genomic libraries)
  - gene cloning prior to targeted disruption or promoter analysis
  - positional cloning
    - genetic mapping
      - Radiation hybrid, STS (sequence tagged sites), ESTs, RFLPs
    - chromosome walking
    - gene identification from large insert clones
    - disease locus isolation and characterization

- Considerations before making a genomic library
  - what will you use it for
    - what size inserts are required?
  - Are high quality validated libraries available?
    - Caveat emptor
      - Research Genetics X. tropicalis BAC library is really Xenopus laevis
    - apply stringent standards, your time is valuable
Genomic libraries (contd.)

- Considerations before making a genomic library (contd)
  - availability of equipment?
    - PFGE
    - laboratory automation
    - if not available locally it may be better to use a commercial library or contract out the construction

Fig. 1.1. **Voltage clamping by the CHEF Mapper system.** A. Relative electrode potentials when the + 60° field vector is activated. B. Relative electrode potentials when the - 60° field vector is activated.

Fig. 1.2. **Voltage clamping by the CHEF Mapper system in the FIGE mode.** A. Relative electrode potentials when the 0° field vector is activated. B. Relative electrode potentials when the 180° field vector is activated.

---

Fig. 9.1. Two gels, one in 0.5x TBE and the other in 1.0x TAE, were run to show the difference in mobility of DNA in the two buffers. S. cerevisiae chromosomes were separated on a 1% gel with a 60 second pulse for 15 hours, followed by a 90 second pulse for 9 hours, at 6 V/cm. Notice the increased migration of the DNA molecules in the TAE gel when compared with the TBE gel.
Genomic libraries (contd.)

- Goals for a genomic library
  - **Faithful representation of genome**
    - clonability and stability of fragments essential
    - >5 fold coverage i.e., library should have a complexity of five times the genome size for a 99% probability of a clone being present.
  - easy to screen
    - plaques much easier to deal with colonies UNLESS you are dealing with libraries spotted in high density on filter supports
  - easy to produce quantities of DNA for further analysis
Construction of genomic libraries

- Prepare HMW DNA
  - bacteriophage λ, cosmids or fosmids
    - partial digest with frequent (4) cutter followed by sucrose gradient fractionation or gel electrophoresis
      - Sau3A (^GATC) most frequently used, compatible with BamHI (G^GATCC)
    - why can’t we use rare cutters?
      - Unequal representation of restriction sites in genome
  - Ligate to phage or cosmid arms then package in vitro
    - Stratagene >>> better than competition
  - Vectors that accept larger inserts
    - prepare DNA by enzyme digestion in agarose blocks
      - why?
        - So DNA does not get mechanically sheared
  - Partial digest with frequent cutter
  - Separate size range of interest by PFGE (pulsed field gel electrophoresis)
  - ligate to vector and transform by electroporation
Construction of genomic libraries (contd) stopped here

- What is the potential flaw for all these methods?
  - Unequal representation of restriction sites, even 4 cutters in genome
  - large regions may exist devoid of any restriction sites
    - tend not to be in genes

- Solution?
  - Shear DNA or cut with several 4 cutters, then methylate and attach linkers for cloning
  - benefits
    - should get accurate representation of genome
    - can select restriction sites for particular vector (i.e., not limited to BamHI)
  - pitfalls
    - quality of methylases
    - more steps
    - potential for artefactual ligation of fragments
      - molar excess of linkers
Construction of genomic libraries (contd)

- What sorts of vectors are useful for genomic libraries?
  - Plasmids?
  - Bacteriophages?
  - Others?

- Standard plasmids nearly useless
- Bacteriophage lambda once most useful and popular
  - Size limited to 20 kb
- Lambda variants allow larger inserts - 40 kb
  - Cosmids
  - Fosmids
- Bacteriophage P1 - 90 kb
- YACs - yeast artificial chromosomes - megabases
- New vectors BACs and PACs - 300 kb