**Master Course**  
**Sequence Alignment**  

**Lecture 12**

**Genome Analysis – Comparative Genomics**

### Genome size

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>φX-174 virus</td>
<td>5,386</td>
</tr>
<tr>
<td>Epstein Bar Virus</td>
<td>172,282</td>
</tr>
<tr>
<td><em>Mycoplasma genitalium</em></td>
<td>580,000</td>
</tr>
<tr>
<td>Hemophilus Influenza</td>
<td>1.8 \times 10^8</td>
</tr>
<tr>
<td>Yeast (S. Cerevisiae)</td>
<td>12.1 \times 10^6</td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td><strong>3.2 \times 10^9</strong></td>
</tr>
<tr>
<td>Wheat</td>
<td>16 \times 10^9</td>
</tr>
<tr>
<td>Lilium longiflorum</td>
<td>90 \times 10^9</td>
</tr>
<tr>
<td>Salamander</td>
<td>100 \times 10^9</td>
</tr>
<tr>
<td>Amoeba dubia</td>
<td>670 \times 10^9</td>
</tr>
</tbody>
</table>

### The amount of genetic information in organisms

<table>
<thead>
<tr>
<th>Name</th>
<th>Genome size (Mb)</th>
<th># genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mycoplasma genitalium</em></td>
<td>0.5</td>
<td>470</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4.5</td>
<td>4400</td>
</tr>
<tr>
<td><em>Scherichia corvina</em></td>
<td>12</td>
<td>5500</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>120</td>
<td>18000</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>97</td>
<td>22,000</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>250</td>
<td>21,700</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>250</td>
<td>50,000</td>
</tr>
</tbody>
</table>

### Understanding genomes

**Yeast genome structure**
- \(-10Mbp\)
- 70% coding, 15% of intergenic region is regulatory

**Elements to identify**

<table>
<thead>
<tr>
<th>Genes</th>
<th>Regulatory elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNA</td>
<td>Systematic mutations upstream</td>
</tr>
<tr>
<td>de novo / other organisms</td>
<td>Clustering / common motif search</td>
</tr>
</tbody>
</table>

**Results**  
4800-6400 \(~80\) motifs

### Patterns of bullet holes

**Bullet holes in planes after a combat**
- Aim: protect vulnerable areas
- Abraham Wald was asked to analyse patterns of bullet holes

**Phylogenetic footprinting:**
- Airplane – genome
- Bullet hole – mutation
- Combat – natural selection

![Genome sequence](image-url)
Comparative genomics:
intro
More than one genome
in hand
no additional information
Look at strongly conserved regions

"Because evolution relentlessly tinkers with genome sequence and tests the results by natural selection, such [functional] elements should stand out by virtue of having a greater degree of conservation"

quantity of data -> quality of results

We start with...
Input: 4 Saccharomyces
- S. cerevisiae, S. paradoxus, S. mikatae, S. bayanus
- 5 to 20 mln years since the divergence of species
- Divergent enough to introduce noise where needed
- Related enough for orthologues to be easily detectable

Sequencing and comparison of yeast species to identify genes and regulatory elements, Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren & Eric S. Lander, Nature 423, 241 - 254 (15 May 2003)

Genome alignment

- Anchoring with ORFs
- Aligning region in between

50kb segment; arrow = direction of ORF, red = 1-1 match; blue = multiple match

Genome evolution - nucleotides

Nucleotide identity:

<table>
<thead>
<tr>
<th></th>
<th>S. paradoxus</th>
<th>S. mikatae</th>
<th>S. bayanus</th>
</tr>
</thead>
<tbody>
<tr>
<td>coding</td>
<td>90%</td>
<td>85%</td>
<td>80%</td>
</tr>
<tr>
<td>intergenic</td>
<td>80%</td>
<td>70%</td>
<td>60%</td>
</tr>
</tbody>
</table>

2x faster!

Each time S. cerevisiae is compared to another yeast species

Genome evolution - nucleotides

Measures of variation for all species (multiple alignment)

<table>
<thead>
<tr>
<th></th>
<th>identity</th>
<th>gap</th>
<th>frame shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>coding</td>
<td>60%</td>
<td>1.3%</td>
<td>0.14%</td>
</tr>
<tr>
<td>intergenic</td>
<td>30%</td>
<td>14%</td>
<td>10.2% +stop codons</td>
</tr>
<tr>
<td>difference</td>
<td>2x</td>
<td>10x</td>
<td>75x</td>
</tr>
</tbody>
</table>

Genes identification

Solution:
- Dummy method: long region without stop codon becomes putative ORF
- We reject putative ORF based on the gene region characteristics

Expected quality:
- High sensitivity of dummy method
- High specificity thanks to divergence
**Genes identification**

Results: gene catalogue revision
- From ~4000 named genes only 15 rejected. One mistake.
- 5538 genes (before: 6062)
- Different start/stop codons for 5% of genes
- ~60 new introns

**Regulatory elements identification**

Regulatory motifs
- Short 6-15bp
- Hard to identify
A small investigation of Gal4 motif shows conservation rates:

<table>
<thead>
<tr>
<th>Motif</th>
<th>Random</th>
<th>Gal4</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>3%</td>
<td>12.5%</td>
<td>4x</td>
</tr>
<tr>
<td>Coding</td>
<td>7%</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>½x</td>
<td>4x</td>
<td></td>
</tr>
</tbody>
</table>

**Detection of regulatory elements**

Make up random motif of form:  
\[CGA\text{x}x\text{x}TGG\]

Check validity with the test

Results

![Diagram showing discovered and known results](image)

**Summary**

Results
- Exact count of 5538 genes
- Better gene boundaries
- New introns identified
- New motifs found

Human
- Coding part of genome 2% (70% in yeast), regulatory motifs 3% of intergenic sequences (15% in yeast)
- repetitions

**An Example: ORF rejected by the test**

![Diagram showing ATG and TAA](image)

Not enough conservation...

**Genome evolution at large scale**

![Graph showing genome evolution](image)
Telomeres are specialized protein–DNA complexes that cap the ends of chromosomes. Like the plastic sleeves that stop shoelaces from unraveling, they protect the sequences that are needed for DNA to replicate when cells divide.

Telomeres are enormously complicated machines made up of specialized DNA, an enzyme called telomerase, and protein complexes that interact with DNA. They function to regulate the lifespan of a cell by shortening with each cell division until they become too small to serve their function and cause the cell to cease dividing.

When eukaryotic cells — those with distinct nuclei — first developed about 1.8 billion years ago, their chromosomes evolved to become linear in shape, rather than circular, as they were in prokaryotic cells, which lack nuclei. Many biologists have theorized that telomerase was born at the same time in order to protect the newly exposed ends of linear chromosomes.

Without proper telomere structure, chromosomes become unstable and cells die. Across evolution, telomere DNA is composed of tandemly repeated short sequences with one strand rich in G and T, e.g. 5’-d(TTAGGG)-3’ in vertebrates, and 5’-d(TTTTGGGG)-3’ in Oxytricha nova. This G-rich strand extends past the duplex portion of the telomere to form a single strand 3’ end. The telomere end binding protein from O. nova recognizes and binds this single strand DNA to form a unique capping complex. Human Telomer-DNA comprises 5000-15 000 bases bound by a large number of proteins. At each replication the telomere shrinks as 50-500 TTAGGG are lost, eventually leading to cell death. Tumor cells use telomerase to counter this process by elongating the telomere.

Telomerase is implicated in cancer — So-called G-quartett DNA structures are believed to inhibit telomerase.

Phylogenetic Footprinting

Phylogenetic footprinting is a method for the discovery of regulatory elements in a set of homologous regulatory regions, usually collected from multiple species.

- It does so by identifying the best conserved motifs in those homologous regions.
- The idea underlying phylogenetic footprinting is that selective pressure causes regulatory elements to evolve at a slower rate than the non-functional surrounding sequence. Therefore the best conserved motifs in a collection of homologous regulatory regions are excellent candidates as regulatory elements.
- The traditional method that has been used for phylogenetic footprinting is to construct a global multiple alignment of the homologous regulatory sequences and then identify well conserved aligned regions. However, this approach fails if the regulatory regions considered are too diverged to be accurately aligned.
Phylogenetic Footprinting

“What is important is conserved”

So you need divergent sequences

Phylogenetic shadowing

Additive divergence

“What is not important can be different”

so you need closely related sequences

Horizontal Gene Transfer

- Acquiring genetic material from other species
- Becomes part of genome
  - from now on transferred vertically

A few closely related species
- Chimp
- Baboon
- monkeys

An example of a position-specific weight matrix of a TF-binding motif adapted from the TRANSFAC database

The sequences that have been shown experimentally to bind to the human transcription factor GATA-1 have 14 positions, among which only positions 6–10 are fully conserved.

Abbreviations: R, G or A (purine); N, any; S, G or C (strong); D, G or A or T. Twelve sequences were used to build this matrix.


HGT in bacteria

- transduction
- conjugation
- transformation
**Detection of HGT**

- Composition bias
  - GC content
  - codon usage
- HGT in bacteria
  - *E. coli*: 13%
  - *M. genitalium*: 0

**SNP - Single Nucleotide Polymorphism**

- **Definition**
  - SNP and phenotype
- **Occurrence in genome**
  - Rarity of most SNPs (agrees with neutral molecular evolutionary theory)
  - SNPs in human population:
    - Inter-genic
    - Coding regions
    - Every 1400bp
    - Every 1430bp
  - High variance in genome!
- **Detection of SNPs:** Hybridization

**SNP - Phase inference**

- In the data from sequencing the genome the origin of SNP is scrambled
  - ...CTA_AAC_GGT...
  - ...CTA_ACG_GGT...
  - ...CTA_AAC_GGT...
  - ...CTA_ACG_GGT...

  Which SNPs are on the same chromosome (are in phase)?

**Linkage Disequilibrium, intro**

How hard is it to break a chromosome?

- An allele/trait SNP A and a are on the same position in genome (locus), thus on a single chromosome an individual can have either of them – but not both
  - $f_A$ - frequency of occurrences of trait A in population
  - $f_a = 1 - f_A$ are frequency occurrences of B and b

- Probabilities of occurrences of both traits on the same chromosome:
  - $f_{AB}$
  - $f_{Aa}$
  - $f_{Ab}$
  - $f_{ab}$

- LD and genomic recombination

**SNP - phase inference**

Determining the parent of origin for each SNP

- In this case: 00 00

Phase inference - the reason why many SNPs sequencing is done for child and two parents.

**Linkage Disequilibrium, calculation**

- When these alleles are not correlated we expect them to occur together by chance alone:
  - $f_{AB} = f_A f_B$
  - $f_{Ab} = f_A f_a$
  - $f_{ab} = f_a f_b$
  - $f_{aB} = f_a f_B$

- But if A and B are occurring together more often (dis-equilibrium state), we can write
  - $f_{AB} = f_A f_B + D$
  - $f_{Ab} = f_a f_B - D$
  - $f_{ab} = f_a f_b + D$
  - $f_{aB} = f_a f_B - D$

- where D is called the measure of disequilibrium.
- Of course from definitions above we have $D = f_{AB} - f_A f_B$
How can we use it?

- Phase inference tells us how SNPs are organized on chromosome
- Linkage disequilibrium measures the correlation between SNPs

Legend to preceding Figure

Fig. 1 Comparison of single-marker LD with haplotype-based LD.

a. LD between an arbitrary marker (at the 26th position, indicated with an asterisk) and every other marker in the data set using $D^2$.

b. Multiallelic $D^2$ is used to plot between the maximum-likelihood haplotype group assignment at the location of the 26th marker and that assignment at the location of every other marker in the data set.

c,d. Repeat of the comparison in a and b but with respect to a second marker (at the 61st position) in the map. Both pairs of graphs show the common feature that, when haplotypes rather than individual SNP alleles are considered to be the basic units of variation, the noise (presumably caused by marker history and properties of the specific statistic chosen) essentially disappears, resulting in a clear, monotonic and step-like breakdown of LD by recombination.

Legend to preceding Figure

Fig. 2 Block-like haplotype diversity at 5q31.

a. Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one.

b. Percentage of observed chromosomes that match one of the common patterns exactly.

c. Percentage of each of the common patterns among untransmitted chromosomes.

d. Rate of haplotype exchange between the blocks as estimated by the HMM. We excluded several markers at each end of the map as they provided evidence that the blocks did not continue but were not adequate to build a first or last block. In addition, four markers fell between blocks, which suggests that the recombinational clustering may not take place at a specific base-pair position, but rather in small regions.
Literature

- Gibson, Muse „A Primer of Genome Science”