Human Genome Analysis

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>Mycoplasma genitalium</td>
<td>0.5</td>
<td>470</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4.5</td>
<td>4400</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>12</td>
<td>5500</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>120</td>
<td>18000</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>97</td>
<td>22000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>3500</td>
<td>23457</td>
</tr>
<tr>
<td>Zea mays</td>
<td>2500</td>
<td>50000</td>
</tr>
</tbody>
</table>

SNPs and haplotypes

Passengers and their evolutionary vehicles

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
  
  \[
  \ldots CT^G AC^G CT^A A GT\ldots
  \]

  Possibility 1 Possibility 2

  chromosome \( \ldots CTGACCGT\ldots \) \( \ldots CTGACAGT\ldots \)

  chromosome \( \ldots CTTACAGT\ldots \) \( \ldots CTTACCGT\ldots \)

  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
  
  \( f_A \cdot \text{frequency of occurrences of trait } A \text{ in population} \)
  
  \( f_a \cdot f_b = 1 - f_a \cdot f_b \) are frequency occurrences of \( B \) and \( b \)

- Probabilities of occurrences of both traits on the same chromosome:
  
  \[
  \begin{array}{ccc}
  f_{AB} & A & B \\
  f_{Ab} & A & b \\
  f_{aB} & a & B \\
  f_{ab} & a & b \\
  \end{array}
  \]

  - LD and genomic recombination

---

**SNP – phase inference**

**Determining the parent of origin for each SNP**

\[
\ldots CT^G AC^G CT^A A GT\ldots
\]

\[
\ldots CT^G AC^G CT^A A GT\ldots
\]

In this case:

\[
\begin{array}{ccc}
GG & TA & \Phi \\
\end{array}
\]

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:
  
  \[
  \begin{align*}
  f_{AB} &= f_A f_B \\
  f_{Ab} &= f_A f_b \\
  f_{aB} &= f_a f_B \\
  f_{ab} &= f_a f_b \\
  \end{align*}
  \]

  - But if \( A \) and \( B \) are occurring together more often (disequilibrium state), we can write
    
    \[
    \begin{align*}
    f_{AB} &= f_A f_B + D \\
    f_{Ab} &= f_A f_b + D \\
    f_{aB} &= f_a f_B + D \\
    f_{ab} &= f_a f_b + D \\
    \end{align*}
    \]

  - 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

Haplotypes - vehicles for SNPs

- discrete haplotype blocks
- The haplotype blocks:
  - Up to 100kb
  - 5 or more SNPs
  - For example, this block shows just two distinct haplotypes accounting for 95% of the observed chromosomes

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAGAACCC</td>
<td>283 (83.2%) haplotype A</td>
</tr>
<tr>
<td>AATCGGGG</td>
<td>40 (11.8%) haplotype B</td>
</tr>
<tr>
<td>GATTAGCC</td>
<td>2 (0.6%)</td>
</tr>
<tr>
<td>GTGAGGG</td>
<td>2 (0.6%)</td>
</tr>
</tbody>
</table>

*Another 13 chromosomes (3.8%) were observed that matched haplotype A or B at all alleles except one, and might represent gene conversion or an undetected genotyping error.

Back to SNPs

Daly et al (2001), Figure 1

Haplotypes of a genome fragment

- Observed haplotypes with dotted lines wherever probability of switching to another line is > 2%
- Percent of explanation by haplotypes
- Contribution of specific haplotypes
Literature

- Gibson, Muse „A Primer of Genome Science“

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


Yeast Genome Analysis

- Anchoring with ORFs
- Aligning region in between

Genome alignment

...CGATGACTATTA...
...CGATGACTA- TA...
...C---GAGTATA...
...CGATGACTATTA...

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!

**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%</td>
</tr>
<tr>
<td>difference</td>
<td>2x</td>
<td>10x</td>
<td>75x</td>
</tr>
</tbody>
</table>

**Genome evolution at large scale**

*Telomeres - evolution's workshop*
- Clusters of ambiguity
- 7-52kb from each end
- Translocations between telomeres
- Rapid evolution
- Observed in Plasmodium falciparum (antigenic variation)
Whole-genome duplication (WGD)

- Poorly understood
- “Cataclysmic” genomic event
- Return to normal state
  - Mutations
  - Rearrangements
  - Gene loss

K. waltii as a proof of genome duplication

S. cerevisiae
- 5714 genes
- 16 chromosomes
- 7% Kw genes: no similarity to Sc

K. waltii
- 5230 genes
- 8 chromosomes

Yeast-genome duplication

- 8 -> 16 chromosomes
- Loss of 90% of duplicated genes
- Paired regions 90% of genome
- ~500 duplicated gene pairs

Evidence of whole genome duplication
**Gene correspondence**

Doubly Conserved Synteny blocks:
- conserved gene order
- genes less than 20kb apart
- double hits Kw – Sc
- (often) no duplicated genes

**DCS in numbers**

- 253 DCS blocks
- ~80% coverage of K. waltii genome
- Typical block
  - approx. 27 genes (max 81)
  - separated by ~3 genes
  - 1% of Kw matches >2 blocks in Sc

**DCS – overview of genomes**

**Centromere proof of genome duplication**

Note! No duplicated genes
**Evolutionary analysis: gene loss**

- **Genome sizes:**
  - Sc 13% larger than Kw
  - 12% paralogous genes

- **Pattern of loss:**
  - Many small deletions (~2 genes)
  - Balanced between regions
  - No chromosome loss
  - No large segmental deletions

**Evolutionary analysis: accelerated divergence**

- 500 paralogous pairs
- 80 pairs (17%): accelerated evolution
  - Sc genes which evolve 50% faster than Kw
Old and new functions of duplicated genes

- Only 1 gene accelerates (95% of cases)
  - one copy preserves the function
  - the other copy is free to diverge
- Functions derived from ancestral ones
  - silencing of Sir3 comes from origin-of-replication Orc1
  - Spatio-temporal differentiation

Knocking-out duplicated genes

- KO of paralogue
  - ancestral: lethal in 20%
  - derived: not lethal
- Derived copy:
  - not essential in a rich medium
  - (sometimes) lose essential aspects of its original function

Summary

- Genome duplication is followed by:
  - massive gene loss: 90% of new genes
  - gene specialization: only one of paralogues accelerates
- Tiny footprints of duplication: genome grows by 10%