**Crash Course in Omics Terminology and Concepts**

Jessica Kissinger  
June 4, 2007

---

**Genome assembly**

- 5X random genome shotgun
- Library insert size
- Mated end pairs
- Contigs
- Scaffolds

---

**Shotgun DNA Sequencing (Technology)**

- DNA target sample
- SHEAR
- SIZE SELECT
- End Reads (Mates)
- LIGATE & CLONE
- Vector

---

**Whole Genome Shotgun Sequencing**

- Collect 10x sequence in a 1-to-1 ratio of two types of read pairs:
  - Short
  - Long
  - ~700,000,000 reads for Human.
- Collect another 20X in clone coverage of 50Kbp end sequence pairs:
  - ~1,200,000 pairs for Human.
- Early simulations showed that if repeats were considered black boxes, one could still cover 99.7% of the genome unambiguously.

---

**Clone-by-Clone Genome Sequencing**

- Physical Mapping
- Minimum Tiling Set
- Shotgun Assembly
- Target
- (~3,000 BACs for Human)
- 2 separate processes
- Physical map of clones
- sequencing libraries must be made for every clone
+ assembly problem ‘easy’ and well understood
**Pairs Give Order & Orientation**

Assembly without pairs results in contigs whose order and orientation are not known. Pairs, especially groups of corroborating ones, link the contigs into scaffolds where the size of gaps is well characterized.

**Anatomy of a WGS Assembly**

**An Assembly of reads**

**Reads and their “trace” files**

**Six Frame Translation**
Terminology

Different prediction methods often generate different results.
**P. falciparum Chromosome 2 - past and present**

**Chromosomal Synteny between organisms**

**C. parvum lacks HXGPRT**

Other Apicomplexans

\[
\begin{align*}
\text{ATP} & \rightarrow \text{AMP} \rightarrow \text{IMP} \rightarrow \text{XMP} \rightarrow \text{GMP} \rightarrow \text{GTP} \\
\text{Apt} & \rightarrow \text{Ado} \\
\text{HXGPRT} & \\
\text{C. parvum} & \\
\text{C. parvum} & \\
\text{Cryptosporidium} & \\
\end{align*}
\]

**C. p. Thymidine Kinase represents a lateral transfer from a bacterium**

Striepen et al, 2004 PNAS 101(9): 3154-3159

**Comparative genomics of pyrimidine biosynthesis in Apicomplexa**

<table>
<thead>
<tr>
<th>Gene/Enzyme</th>
<th>C. parvum</th>
<th>T. gondii</th>
<th>T. annulata</th>
<th>P. falciparum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu + CO₂</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Aspartate carbamoyltransferase</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Dihydroorotase</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Dihydroorotate dehydrogenase</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Oxaloacetate phosphatase-aminotransferase</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>UMP</td>
<td>Absent</td>
<td>Present</td>
<td>Present</td>
<td>Present</td>
</tr>
</tbody>
</table>

Striepen et al, 2004 PNAS 101(9): 3154-3159

**C. parvum genome encodes two unique salvage enzymes**

Striepen et al, 2004 PNAS 101(9): 3154-3159
Evolutionary relationships

- Homology - related by evolutionary descent not equivalent to similarity
- Orthology - same gene in different organisms, e.g. alpha hemoglobin in humans and chimps
- Paralogy - genes within an organism related by gene duplication, e.g. alpha and beta hemoglobin in humans
- Xenology - genes related by gene transfer

Functional Genomics

Always... has a time and space component

Expression Profiles

- The pattern of expression of one or more genes over time or a set of experimental conditions, e.g. during development or a drug treatment or in a genetic mutant such as a gene knock-out.

Microarrays

- cDNA microarrays
- "GeneChip" in situ synthesized oligonucleotide arrays
- Oligomer (~70mer) arrays

Experiments are almost always Competitions between conditions or stages

cDNA array grid

Sequence of one gene

Microarray

Microscope slide

cDNA Microarrays

Robotic microarrayer
Chip Oligo Array Hybridization

Probe labeling

mRNA

Oligo dT

Reverse transcriptase

Labeled cDNA

General Scanning ScanArray 3000

Scans, Red, Green and Merged
Ratios of experimental to control expression are often expressed as colors rather than numbers.

A Dendrogram of clustered expression profiles

Table 4.1 Summary table of data for one pooled DNA microarray.
The fluorescence intensity for each color was determined, and the ratio of red signal to green signal was calculated for each of the 3,200 genes on the full microarray. Also contained in the data is the location for each spot. Where a microarray is hybridized, it is important to know where each gene is located on it.

<table>
<thead>
<tr>
<th>Block</th>
<th>Carbons</th>
<th>Rear</th>
<th>Gene Name</th>
<th>Red</th>
<th>Green</th>
<th>Red/Green Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>S1s</td>
<td></td>
<td>2.345</td>
<td>2.457</td>
<td>0.955</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>S2c</td>
<td></td>
<td>5.539</td>
<td>2.158</td>
<td>2.663</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>S3c</td>
<td></td>
<td>4.493</td>
<td>1.450</td>
<td>3.099</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>S4c</td>
<td></td>
<td>5.880</td>
<td>3.509</td>
<td>1.700</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>S5c</td>
<td></td>
<td>1.284</td>
<td>1.258</td>
<td>0.939</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>S6c</td>
<td></td>
<td>1.821</td>
<td>2.514</td>
<td>0.723</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>S7c</td>
<td></td>
<td>2.787</td>
<td>1.362</td>
<td>2.044</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>S8c</td>
<td></td>
<td>2.002</td>
<td>5.202</td>
<td>0.383</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>S9c</td>
<td></td>
<td>3.925</td>
<td>1.219</td>
<td>3.217</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>S10c</td>
<td></td>
<td>2.746</td>
<td>1.836</td>
<td>1.490</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>S11c</td>
<td></td>
<td>2.310</td>
<td>4.245</td>
<td>0.543</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>S12c</td>
<td></td>
<td>1.896</td>
<td>2.396</td>
<td>0.791</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>S13c</td>
<td></td>
<td>1.023</td>
<td>3.284</td>
<td>0.311</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>S14c</td>
<td></td>
<td>1.688</td>
<td>2.388</td>
<td>0.714</td>
</tr>
</tbody>
</table>

Table 4.4 Summary of the hierarchical clustering algorithm applied to the 12 genes in Table 4.2.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Object 1</th>
<th>Object 2</th>
<th>Correlation</th>
<th>Nearest Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J</td>
<td>S1</td>
<td>1.195</td>
<td>[S12]</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>S2</td>
<td>0.764</td>
<td>[S2]</td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>S3</td>
<td>0.958</td>
<td>[S12]</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>S4</td>
<td>0.958</td>
<td>[S12]</td>
</tr>
<tr>
<td>5</td>
<td>S1</td>
<td>S5</td>
<td>0.84</td>
<td>[S12]</td>
</tr>
<tr>
<td>6</td>
<td>S2</td>
<td>S6</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>7</td>
<td>S3</td>
<td>S7</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>8</td>
<td>S4</td>
<td>S8</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>9</td>
<td>S5</td>
<td>S9</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>10</td>
<td>S6</td>
<td>S10</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>11</td>
<td>S7</td>
<td>S11</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>12</td>
<td>S8</td>
<td>S12</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>13</td>
<td>S9</td>
<td>S13</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>14</td>
<td>S10</td>
<td>S14</td>
<td>0.949</td>
<td>[S12]</td>
</tr>
<tr>
<td>15</td>
<td>S11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>S12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>S13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>S14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2 Correlation coefficients between each pair of genes, based on log transformed gene expression data in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84</td>
<td>0.94</td>
<td>-0.98</td>
<td>-0.38</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
<td>-0.48</td>
</tr>
</tbody>
</table>
Protein Expression

Typical 2D gel

High throughput mass spectrometry

- Direct identification of proteins from biological sample.
- Capillary liquid chromatography apparatus (LC) coupled with...
- Electrospray tandem mass spectroscopy (MS/MS)
- "Sequest" software links mass spectra with genomic sequence database.

How Tandem MS Works

Complex mixture

Protein

Peptides

+ Ionized

Measurement

Isolation

Fragmentation

Collision Inducted Dissociation (CID)

Liquid chromatography

Tandem MS protein data

Individual mass spectra

P A P K

P A P K

P A P K

P A P K

Combined mass spectrum

P A P K
**Sequest Database Search**

- Mass Spectrometer
- Protein Database
- Nucleic Acid Database
- EST Database
- Tandem Mass Spectrum
- Theoretical Mass Spectrum
- Correlation Analysis
- Ranked Score of Matched Peptides

**Note:** ORFs in addition to predicted Genes must be searched

**A protein network**

Networks have edges and nodes

**Protein Networks, often established by Y2H studies**

**Ontologies**

- Terms to link equivalent concepts that have different names
- Gene Ontologies
- Sequence Ontologies
- BioMedical Ontologies
- Cell component Ontologies
- http://www.bioontology.org/

**Community Databases**

- PFAM - http://pfam.janelia.org/
- Interpro - http://www.ebi.ac.uk/interpro/
- Cyc - http://metacyc.org/