Welcome!


Instructors & Friendly Faces

- Betsy Wenthe
- Brian Brunk
- Susanne Warrenfeltz
- Jessie Kissinger
- Omar Harb
- Eve Basenko
Crash Course in Omics Terminology, Concepts & Data Types

Jessie Kissinger
June 12, 2016
5X genome sequence means that sequences equivalent to 5X the genome size were generated, e.g., Genome size = 10 Mbp, then 50 Mbp of random sequences were generated.

Pairs Give Order & Orientation

Scaffold

Gaps in scaffolds are traditionally indicated by 100 “N’s”

End Reads (Mates)

Mean & Std. Dev. is known

2-pair

Primer

NGS

Distance?

550bp

SEQNECE

Consensus Reads

SNPs

Sequence?
Anatomy of a WGS Assembly

30,000 ft View - Annotation
Six Frame Translation ORF-finding

ORFs ≠ Genes
>Translation Frame 1
MQKPVCLVYVAMTKRGGCNGGLWGFLTTDFKHFSRVKTTPEEASRLNGWLPRKFAKTGDSGLPSPSVGKRFNAVVMGRKTWESMPRKFRPLVDRLN
VVSSSLKEDEIAAEKPQAEGQQRVRCASLPAALSLLEEEYKDSVDQIFVVGGAGLYEAALSLGVASHLYITRVAREFPDVFFPAFPGDDILSNKSTAAQAAAPAESVFVFHCPELGRASHNATRPFRFISKVPSDVYDPYVEK
RSTCDAATAEFSNAMESSTATTTFVVLQAPSAJAFFLMD
DRKREQQSKAHPVHHRFGEFQYLKLIDIINNKSTMIRT

Terminology
30,000 ft View - Synteny

Species 1

Species 2

Synteny among Plasmodia
Synteny shows relationships in positioning: Ontologies show relationships in meaning

- The Gene Ontology – GO provides terms to link genes with similar functions and/or locations in the cell.
- An ontology was needed because the cultural traditions in different organisms led to different gene naming schemes that made it difficult to identify orthologous genes with the same function.
For Example:

*D. melanogaster* gene CG3340 annotated as: “Kruppel” and *P. falciparum* gene PF3D7_1209300 annotated a “putative KROX1”

Can both be annotated with GO term: **GO:0003705** (RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity)

Both proteins, functionally, are **Zinc Fingers** despite their different names

Note that the Gene Ontologies themselves contain only information about terms in the ontology and their relationships to other terms
Expression Profiles (RNA and Protein)

- 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.
- Always... has a time and location component

RNA expression

- RNA-Seq (NGS)
  - Little sequence bias
  - Quantitative
  - Usually are strand-specific
  - Can be used to identify UTR’s and exon splice junctions

- Expressed Sequence Tags, ESTs
  - Usually represent partial cDNA
  - Often clustered
  - Come from libraries that may, or may not be normalized
  - Often used to identify genes in genomes and locations of introns

- SAGE tags
  - Serial Analysis of Gene Expression
30,000 ft View - RNA-Seq

Annotation of genome features

RNA-Seq reads

FPKM = Fragments per kilobase of exon per million fragments mapped

Clustered Microarray Data
Genes with Similar Expression Profiles are Grouped together
Genes can be located on either DNA strand
Convection - Gene location = non-template strand, i.e. same as the mRNA

Overview of transcription: Either strand can serve as a template for a gene

Figure 8.4
Introduction to Genetic Analysis, Ninth Edition
© 2008 W.H. Freeman and Company
Complex patterns of eukaryotic mRNA splicing: What is a Gene?

Protein Expression/Sequence

Data
- MW- Isoelectric point
- MW
- Sequence/spans

Technology
- 2D gel electrophoresis
- Mass spectrometry
- Tandem MS (MS-MS, LC MS-MS etc)

Typical 2 D gel
How Tandem MS Works

How Tandem MS Works

Complex mixture
Protein

Ionized
Peptides

Peptides

Ionized
Peptides

Complex mixture
Protein

Measurement

Collision Induced Dissociation (CID)

Liquid chromatography

Isolation

Fragmentation

Tandem MS protein data

(protonated mass 1410.6)

Mass* b-lons y-lons Mass*
81.1 S PAFDIMAETLX 1323.6
185.2 SP AFDSIMAETLX 1226.4
256.3 SPAFD SIMAETLX 1155.4
403.5 SPAF SIMAETLX 1098.2
518.5 SPAFDS SIMAETLX 992.1
605.6 SPAFDSI MAETLX 886.0
718.8 SPAFDSI MAETLX 692.3
950.0 SPAFDSIM AETLX 567.1
921.1 SPAFDSIMAE TLK 490.8
1050.2 SPAFDSIMAE TLK 361.5
1151.3 SPAFDSIMAE TLK 296.4
1264.4 SPAFDSIMAE TLK 147.2

b) 250 500 750 1000 1250

20 40 60 80 100

255.1 692.3 905.5 925.5 1077.4
1334.8

360.9 400.0 519.1 1156.5 1226.8

135.3

m/z
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

Peptide database

Note: ORFs in addition to predicted Genes must be searched
Mass Spectrometry can be used to measure metabolic and other chemical compounds.
Complex mixtures can be analyzed and interpreted

<table>
<thead>
<tr>
<th>Saponin</th>
<th>Empirical Formula</th>
<th>Theoretical M/Z</th>
<th>Experimental M/Z</th>
<th>Error (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Glc:3Ac:2M.O.A</td>
<td>C68H100O11Na</td>
<td>587.5717</td>
<td>587.3721</td>
<td>-8.651</td>
</tr>
<tr>
<td>3Glc:3Ac:2M.O.A</td>
<td>C70H100O11Na</td>
<td>587.5717</td>
<td>587.3721</td>
<td>-8.651</td>
</tr>
<tr>
<td>3Glc:3Ac:2M.O.A</td>
<td>C70H100O11Na</td>
<td>587.5717</td>
<td>587.3721</td>
<td>-8.651</td>
</tr>
<tr>
<td>3Glc:3Ac:2M.O.A</td>
<td>C70H100O11Na</td>
<td>587.5717</td>
<td>587.3721</td>
<td>-8.651</td>
</tr>
</tbody>
</table>

Metabolites can be linked to metabolic pathways and enzymes

**ALANINE, ASPARATE AND GLUTAMATE METABOLISM**

[Detailed metabolic pathway diagram]

18
30,000 ft View - NGS SNPs

NGS sequence reads from different clinical/environmental isolates

Alleles and Phenotype

- Some phenotypes are caused by a single locus in the genome and a single allele at that locus (e.g. some flower colors, or *Drosophila* eye color)
- Other phenotypes (Type-I diabetes, heart disease) are multi-locus or “complex” (i.e. many genes are involved, each potentially with many alleles)
**Homologous chromosomes (in a diploid)**

**A** AAGCCTCATC

**a** ACGCCTCATC

**SNP** = Single Nucleotide Polymorphism

**Population data**

**Data**
- Single Nucleotide Polymorphisms, SNPs
- Alleles
- Allele frequency
- Haplotypes

**Technology**
- Chip-Seq
- NGS
Alleles have frequencies in different populations.

Populations and alleles have geographic boundaries.

A parasite isolate comes from a particular population, a particular location and will have a specific haplotype (e.g. representation of alleles) often characterized via SNPs.
## Parasite Isolates

<table>
<thead>
<tr>
<th>Data</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species, Strain,</td>
<td>PCR-RFLP</td>
</tr>
<tr>
<td>Isolate</td>
<td>Microsatellites</td>
</tr>
<tr>
<td>Location, Date</td>
<td>Sequencing</td>
</tr>
<tr>
<td>SNP</td>
<td>SNP chip</td>
</tr>
<tr>
<td>Sequence</td>
<td>GPS</td>
</tr>
<tr>
<td>Allele</td>
<td></td>
</tr>
<tr>
<td>phenotype</td>
<td></td>
</tr>
</tbody>
</table>

## Infectious Disease Paradigm

Experimental systems

![Experimental systems diagram](image-url)
Metadata - The next Frontier

- Data about the data are critical
- What makes a data set valuable? (The reason it was generated...but often this is missing)
- How can you find the data set you need? Pull down Menu? A search of data set properties?
  - Data generator
  - Clinical outcome
  - Geographic location
  - Phenotype

Bioinformatics uses algorithms

- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats (like the experimental conditions, often call parameters in bioinformatics.)
How to find an intron

- Usually begins with GT and end with AG
- Must be longer than 19 nucleotides
- Must contain a branchpoint “A”
- Donor GT often followed by a sequence pattern. This pattern is species-specific
- Acceptor AG often proceeded by pyrimidine stretch
- Has a mean length of “X” as is observed in this species
Different prediction methods often generate different results

Prediction 1

Prediction 2

We provide lots evidence so that you can decide or design an experiment to confirm!

The End

• If you have questions, I and the other instructors will be around all week and we are happy to talk to you.
• These slides are posted on the course Schedule