Welcome!

EuPathDB Workshop 2010
Important Information

• It might rain, afternoon showers are common

• Air conditioned buildings may be cold, you may want a sweater

• You do not need your laptop for workshop sessions
  • Buses are most frequent 7AM-6:30PM

• Guides (Brian and Omar) will leave hotel tomorrow lead folks to the workshop room via University bus
Orbit Bus Route
Every 12 minutes
7AM - 6:30PM

~10 blocks
1.1 miles

3 blocks, .6 Km
Take the Orbit bus from the Arch to Tucker Hall at East Campus and Green Street
The Life Sciences building is on Greene street, up the hill on the Right Rm C128
General Schedule

• Workshop all day Mon - Tues 9-5:00
• Workshop till noon on Wednesday
• Schedule in your packets
• Dinner at the Coverdell Center on Monday (Informal!)
• Breakfast use coupons at Holiday Inn, Coffee and lunch provided (M-W). Tuesday Dinner is on your own.
Instructors & Friendly Faces

- Betsy Wang
- Jessie Kissinger
- Brian Brunk
- Omar Harb
- Mark Heiges
- Eileen Kraemer
Now you know us, Who are you?

• Please tell us your name and organization, country
• Please tell us a bit about what you do/study (3-5 sentences)
• Please tell us what you hope to learn at the workshop
• Please tell us how you learned about the workshop
Crash Course in Omics Terminology, Concepts & Data Types

Jessica Kissinger
June 6, 2010
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

• Must begin 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
Donor Site
Generated by http://www.bio.cam.ac.uk/seqlogo/logo

GTAAGTT
Computed using alpro and makeolog (Schneider & Stephens, 1999)

Acceptor site
Generated by http://www.bio.cam.ac.uk/seqlogo/logo

TAG
Computed using alpro and makeolog (Schneider & Stephens, 1999)
Different prediction methods often generate different results

Prediction 1

Prediction 2
# DNA sequence Data

<table>
<thead>
<tr>
<th>Data</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Genome Sequence</td>
<td>Sanger</td>
</tr>
<tr>
<td>Genome Survey Sequence</td>
<td>- WGS, PGS</td>
</tr>
<tr>
<td>BAC, Cosmid End Sequence</td>
<td>- ABI</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Next-Generation</td>
</tr>
<tr>
<td>SNP, CNV</td>
<td>- 454</td>
</tr>
<tr>
<td></td>
<td>- SOLiD</td>
</tr>
<tr>
<td></td>
<td>- Illumina</td>
</tr>
<tr>
<td></td>
<td>- Flx</td>
</tr>
<tr>
<td></td>
<td>Affymetrix Chip</td>
</tr>
</tbody>
</table>
Genome assembly

- 5X random genome shotgun
- Library insert size
- Mated end pairs
- Contigs
- Scaffolds
Pairs Give Order & Orientation

Scaffold

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

2-pair

Mean & Std.Dev. is known

End Reads (Mates)

Primer

550bp

SEQUENCE
Anatomy of a WGS Assembly
Six Frame Translation

ORF-finding

ORFs ≠ Genes
ATGCAGAAAAACGTTGCTCCTGCTCGAGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC
ACTTGGACACACAGATTTCACAAACATTCTTCTGTGTCGCAAAAAACGACCGCCCAAGAAGCCAGATTGCAACACGCGCTTCTGCTGGGTCT
CTCCAGGAAAATTGCAAAGACGGGCAGTCTGGACTTCCCTTCTCCATCTGCGCCAGAGATTTCAACGCCGTTGTCATGGGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCCTCTTCCCTCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGCAGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGAGGAGCGGGACTGTACGAGGCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGAGTTTCCGTGCGACGTTTTCTTCCCTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCGTTCCCTTTTGTCCGGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTGACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGAGCAACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGCAGCCATGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCATGTTCACTTTTAGAGCCATGAATTCCTTTGACTCATTTGCTCATGCGCCAGCATATTATTAACAAATGGAAGGACAATGGATACCGAACGG

>Translation Frame 1
MQKPVCLVVAMTPKRGIGINNGLPWHPLTTDFKHFRTVKTTPPEEASRLNGWLPRFKAKTGDSGLPSVPSVVRKRNAVMGRKTFHIESMPRFRPLVDRLNI
VVSSSLKEEDIAAEKPAEQQGQRVRCASLPAALSLLEDVYKDSVDQIFV
VGGAGLYEAALSLGVAHLYITRVAFEPFCDFPPFAPFPGDILSNKSTAAQAAPAEVSFVFPFCELGREKDNEATYRPIFISKTFSDNGVPYDFVLEKRRKTDDAATAEPSNAMSSLTSTRETTVPVHGLQAPSSAAIAEPLVACLMDDE
DRKKREQKELIRAVPHVHRFGHEEFQYLDLIADIINNGRTMDDRT
**Terminology**

- **transcriptional start**
- **ATG**
- **stop codon**
- **polyA**
- **5' UTR**
- **exon**
- **intron**
- **3' UTR**

**CDS:**
(coding sequence nt)

**protein:**
(aa)

**transcript:**
(CDS + UTRs, if avail.)

**genomic:**
(includes introns)
The mouse and human genome have large syntenic blocks of genes in common.

Figure 13-15
*Introduction to Genetic Analysis, Ninth Edition*
© 2008 W. H. Freeman and Company
Synteny among Plasmodia
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
RNA sequence/Expression Data

Data
• cDNA
• Expressed Sequence Tags (EST)
• RNA-Seq
• Microarray
• Ditag (SAGE-tags)
• Small RNA’s (various types)

Technology
• Sanger
• Next-gen
• Microarray-slides
• Microarray-chips
• SAGE
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.

• Always... has a time and space component
Microarrays

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

Experiments are almost always Competitions between conditions or stages
Ratios of experimental to control expression are often expressed as colors rather than numbers.
A Dendrogram of clustered expression profiles
Clustered Microarray Data
Genes with Similar Expression Profiles are Grouped together
Other RNA expression

- 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)
- Primary purpose is to assay relative levels of gene expression
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
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 Peptides

Peptides

Measurement

Liquid chromatography

Collision Induced Dissociation (CID)

Isolation

Fragmentation
Tandem MS protein data

(protonated mass 1410.6)

<table>
<thead>
<tr>
<th>Mass⁺</th>
<th>b-ions</th>
<th>y-ions</th>
<th>Mass⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>81.1</td>
<td>S</td>
<td>PAFDSIMAETLK</td>
<td>1323.6</td>
</tr>
<tr>
<td>185.2</td>
<td>SP</td>
<td>AFDSIMAETLK</td>
<td>1226.4</td>
</tr>
<tr>
<td>256.3</td>
<td>SPA</td>
<td>FDSIMAETLK</td>
<td>1155.4</td>
</tr>
<tr>
<td>403.5</td>
<td>SPAF</td>
<td>DSIMAETLK</td>
<td>1008.2</td>
</tr>
<tr>
<td>518.5</td>
<td>SPAFD</td>
<td>SIMAETLK</td>
<td>893.1</td>
</tr>
<tr>
<td>605.6</td>
<td>SPAFDS</td>
<td>IMAETLK</td>
<td>806.0</td>
</tr>
<tr>
<td>718.8</td>
<td>SPAFDSI</td>
<td>MAETLK</td>
<td>692.3</td>
</tr>
<tr>
<td>850.0</td>
<td>SPAFDSIM</td>
<td>AETLK</td>
<td>561.7</td>
</tr>
<tr>
<td>921.1</td>
<td>SPAFDSIMA</td>
<td>ETLK</td>
<td>490.6</td>
</tr>
<tr>
<td>1050.2</td>
<td>SPAFDSIMAE</td>
<td>TLK</td>
<td>361.5</td>
</tr>
<tr>
<td>1151.3</td>
<td>SPAFDSIMAET</td>
<td>LK</td>
<td>260.4</td>
</tr>
<tr>
<td>1264.4</td>
<td>SPAFDSIMAETL</td>
<td>K</td>
<td>147.2</td>
</tr>
</tbody>
</table>

$m/z$
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
# Population data

<table>
<thead>
<tr>
<th>Data</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Nucleotide</td>
<td>Technology</td>
</tr>
<tr>
<td>Polymorphisms, SNPs</td>
<td></td>
</tr>
<tr>
<td>Alleles</td>
<td>Technology</td>
</tr>
<tr>
<td>Allele frequency</td>
<td></td>
</tr>
<tr>
<td>Haplotypes</td>
<td></td>
</tr>
</tbody>
</table>
Homologous chromosomes (in a diploid)

- Allele for purple flowers
- Locus for flower-color gene
- Homologous pair of chromosomes
- Allele for white flowers
Loci, alleles and SNPs

SNP = Single Nucleotide Polymorphism
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)
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

**Data**
- Species, Strain,
- Isolate
- Location, Date
- SNP
- Sequence
- Allele

**Technology**
- PCR-RFLP
- Sequencing
- SNP chip
- GPS