Crash Course in Omics
Terminology, Concepts & Data Types

13th Annual EuPathDB Workshop
Jessie Kissinger
June 17th, 2018
Sequence Technologies/platforms

• Hybridization (Southern, Northern, Microarray, chip Arrays (Affymetrix), now.. “bait + Seq”

• Sanger sequencing, 454, Solexa/Illumina, Solid, Ion torrent, PacBio, Nanopore, 10X Genomics Chromium

DNA sequencing technologies: 2006-2016 Elaine R Mardis
NATURE PROTOCOLS | VOL.12 NO.2 | 2017 | 213
30,000 ft View - Genome Assembly

5X genome sequence means that sequences equivalent to 5X the genome size were generated e.g. Genome size = 10 Mbp, then 50Mbp of random sequences were generated
Anatomy of a WGS Assembly

Chromosome

STS-mapped Scaffolds

Contig

Read pair (mates)

Gap (mean & std. dev. Known)

Consensus

Reads

SNPs
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

550bp

Primer

Plasmid Fosmid

Consensus Reads

SNPs

NGS

Distance?

SEQUENCE
30,000 ft View - Annotation

Genes

I
II
III
IV

A
B
B
C
D
E
F
### Six Frame Translation

**ORF-finding**

<table>
<thead>
<tr>
<th>1/1</th>
<th>31/11</th>
<th>61/21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MY ALL I I L Y I I I I R H * S H H A C R G V Y Y I Y</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H V R F T D S I L Y Y Y * T L V T S C M * G G L L Y L</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A C T L Y * F Y I I I I L L D T S H I M H V G G S T I S</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GCA TGT ACG CTT TAC TGA TTC ATT GTA TTA TTA TTA GAC ACT AGT CAC ATC ATG CAT GTA GGG GGG TCT ACT ATA TCT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CGT ACA TGC GAA ATG ACT AAG ATA TAA TAT AAT AAT AAT CTG TGA TCA GTG TAG TAC GCA GTG CAT CCC CCC AGA TGA TAT AGA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C T R K V S E I N Y * * * V S T V D H M Y P P R S Y R</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>**M Y A K S I R Y * I I I I C * D C * A H L P T * * I ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H V S * Q N * I I N N N S V L * M M C T P P D V I D I</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>**121/41</td>
<td>151/51</td>
<td>181/61**</td>
</tr>
<tr>
<td><strong>LE L E R I D L A L Y N F S D I Y I P A S R G K W</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L A R A R T H R L S M T I F Q R H I Y S R L A G K M</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A S S S S N A S T * H D Y I I I S A T Y I I F P P R G E N</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GCT AGC TGC AGC TAG AAC GCA TCG ACT TAG CAT GAC TAT ATA ATT TCA GCC ACA TAT ATA TCT CCG CTC GGC GGG GAA AAT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CG A TGC AGC TGC ATC TTG CTG AGC TGA ATC GTA CTG ATA ATT TAA AGT CGC TGT ATA TAT AAG GGC GGA GGC CCC CTT TTA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S A R A L V C R S L M V I Y N * R C I Y E R R A P F I</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>S S S S R M S K A H S Y L K L S M Y I G A E R P F H</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L E L * F A D V * C S * I I E A V Y I N G G R P S F F</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ORFs ≠ Genes**
**Terminology**

- **transcriptional start**
- **ATG**
- **stop codon**
- **polyA**

- **5' UTR**
- **3' UTR**
- **exon**
- **intron**

**CDS:**
(coding sequence nt)

**protein:**
(aa)

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

**genomic:**
(includes introns)
Synteny among Plasmodia
Homology

Early Globin Gene

Gene Duplication

α-chain gene
α-frog
α-chick
α-mouse

β-chain gene
β-mouse
β-chick
β-frog

PARALOGS

ORTHOLOGS

HOMOLOGS
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 as “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.

Convention - 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?

Figure 8-14

Introduction to Genetic Analysis, Ninth Edition
© 2008 W.H. Freeman and Company
CRISPR-CAS

How CRISPR works

1. The Cas9 protein forms a complex with guide RNA in a cell
2. This complex attaches to a matching genomic DNA sequence adjacent to a spacer (yellow segment)
3. The Cas9-RNA complex cuts the double strands of the DNA
4. Programmed DNA may be inserted at the cut

- Need to provide both the enzyme and the guide RNA to the cell
- Need to design the guide RNA to the gene of interest, ideally at multiple target locations per gene

Ball et al., MRS Bulletin November 2016
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 2D gel
How Tandem MS Works

Complex mixture
Protein

Ionized Peptides

Peptides

Liquid chromatography

Collision Inducted Dissociation (CID)

Isolation

Fragmentation

Measurement
### Tandem MS protein data

#### a)

<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>

#### b)

![Mass spectrum graph](mass_spectrum.png)
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
30,000 ft View - Proteomics
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 Accurate M/Z</th>
<th>Experimental Accurate M/Z</th>
<th>Error (ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Glc-Med.A.</td>
<td>C36H56O11Na</td>
<td>687.37201</td>
<td>687.3718</td>
<td>-0.338</td>
</tr>
<tr>
<td>3Glc28Glc-Med.A.</td>
<td>C42H66O16Na</td>
<td>849.42483</td>
<td>849.4240</td>
<td>-0.945</td>
</tr>
<tr>
<td>3GlcAGalRha-SoyB</td>
<td>C48H78O18Na</td>
<td>965.50856</td>
<td>965.5017</td>
<td>-7.172</td>
</tr>
</tbody>
</table>

![Mass spectrum graph showing peaks at various masses with annotations for different saponins.](image)
Metabolites can be linked to metabolic pathways and enzymes
Figure 1. Overview of existing pathway analysis methods using gene expression data as an example.

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375
Gene & Pathway Enrichment

Gene list:
Up/Down-regulated based on some experiment, e.g. RNA-Seq

Background-Pathway information: All genes known to be involved in some process, e.g. glycolysis or cell signaling. ALL pathways are examined

Input Gene list

<table>
<thead>
<tr>
<th>Input Gene list</th>
<th>Background genes by GO or Pathway</th>
</tr>
</thead>
</table>

Result: GO:ID or Pathway ID that is enriched

Statistics: Are more genes observed than expected (P-value)
Multiple hypothesis testing (Bonferroni, Benjamini-Hochberg)

Atul Butte Review: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375
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)

AAGCCTCATC
ACGCCTCATC

SNP = Single Nucleotide Polymorphism
30,000 ft View - NGS SNPs

NGS sequence reads from different clinical/environmental isolates

$\star$ = SNP
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.
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
Donor Site
Generated by http://www.bio.cam.ac.uk/seqlogo/logo

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

Computed using alpro and makelogo (Schneider & Stephens, 1990)
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!
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)
• Introducing the “data set”
• 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
Data sharing standards
The End

• If you have questions, I and the other instructors will be around and we are happy to talk to you.
• These slides are available to you as a PDF on the workshop web site.