Welcome!

EuPathDB Workshop 2015
This is our 10th Anniversary!!!

EuPathDB
Eukaryotic Pathogen Database Resources
http://EuPathDB.org

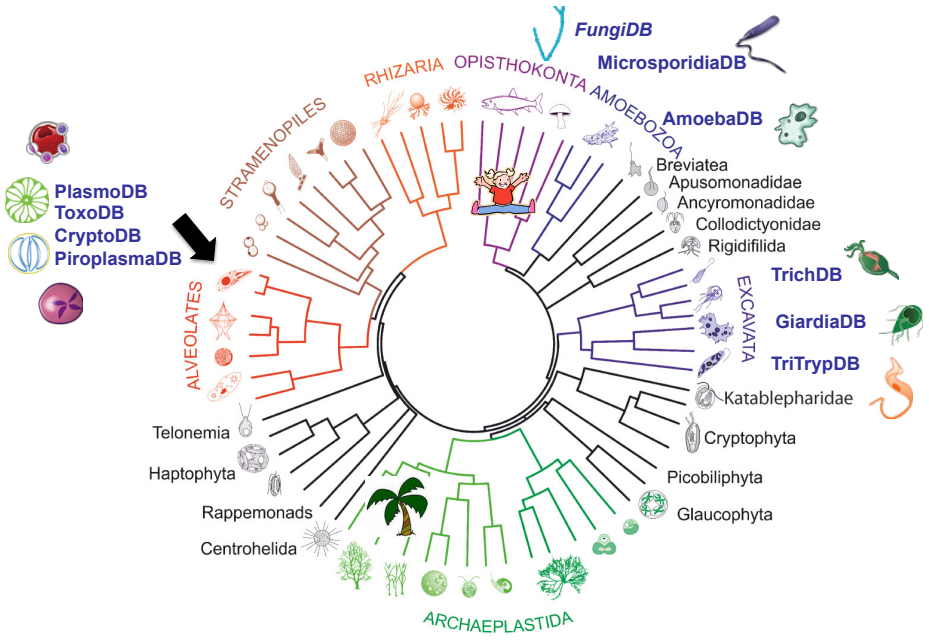
Instructors & Friendly Faces

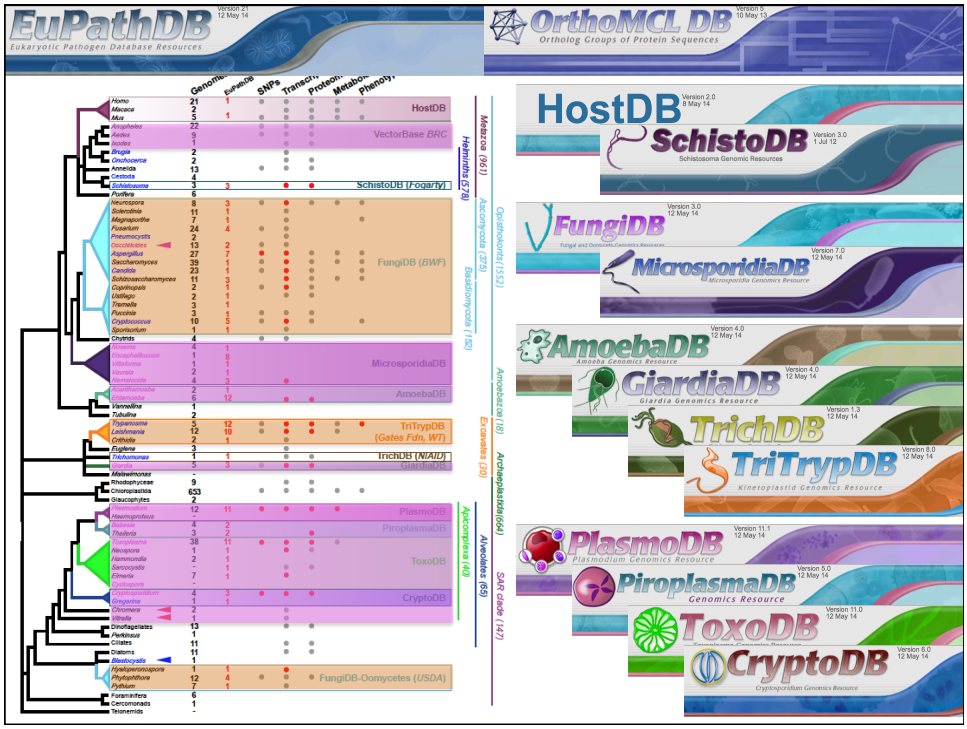
- Betsy Wenthe 
- Brian Brunk 
- Susanne Warrenfeltz 
- Jessie Kissinger 
- Omar Harb 
- David Roos 

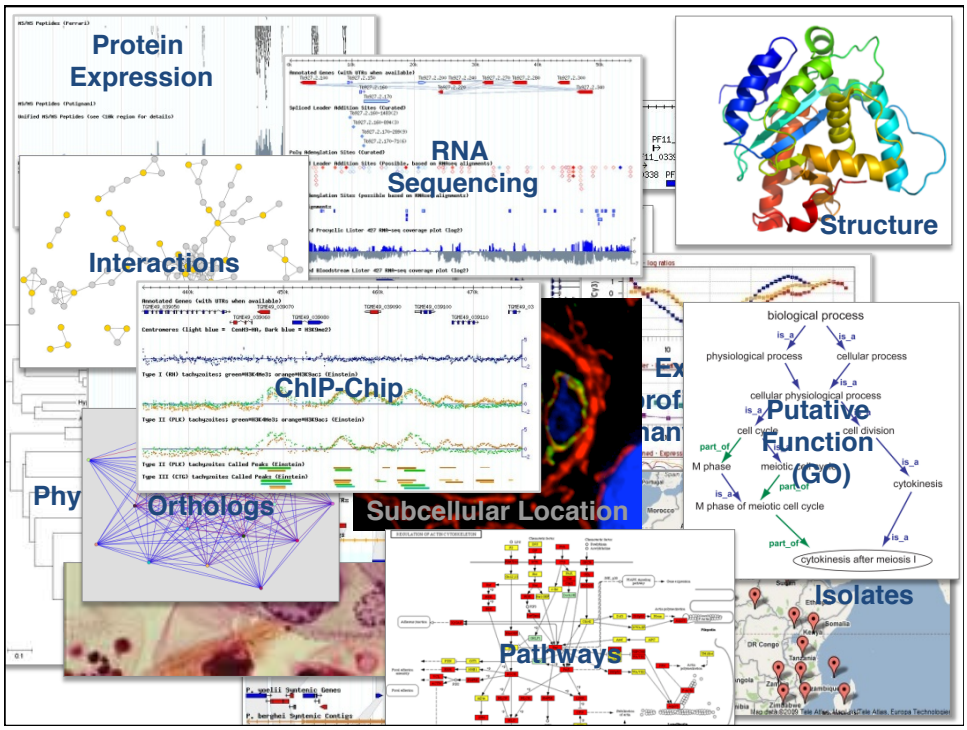
Crash Course in Omics Terminology, Concepts & Data Types

Jessie Kissinger

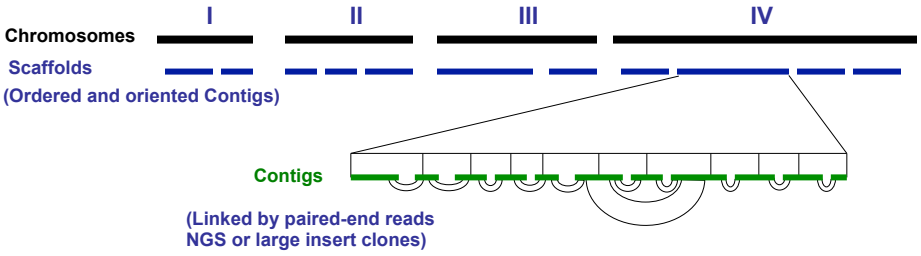
May 31, 2015




EuPathDB
Eukaryotic Pathogen Database Resources
OrthoMCL DB
Ortholog Groups of Protein Sequences
HostDB
SchistoDB
Schistosoma Genomic Resources
FungiDB
MicrosporidiaDB
Microsporidia Genomics Resource
AmoebaDB
Amoeba Genomics Resource
GiardiaDB
Giardia Genomics Resource
TrichDB
TriTrypDB
Kinetoplastid Genomics Resource
PlasmoDB
Plasmodium Genomics Resource
PiroplasmaDB
Genomics Resource
ToxoDB
CryptoDB
Cryptosporidium Genomics Resource
HostDB
VectorBase BRC
SchistoDB (Fogarty)
FungiDB (BWF)
MicrosporidiaDB
AmoebaDB
TriTrypDB (Gates Fdn, WT)
TrichDB (NIAID)
PlasmoDB
PiroplasmaDB
ToxoDB
CryptoDB
FungiDB-Oomycetes (USDA)
Metazoa (961)
Helminths (578)
Ascomycota (279)
Basidiomycota (132)
Opisthokonts (1552)
Amoebozoa (18)
Excavates (30)
Archaeplastida (64)
Alveolates (69)
SAR clade (147)
Apicomplexa (49)

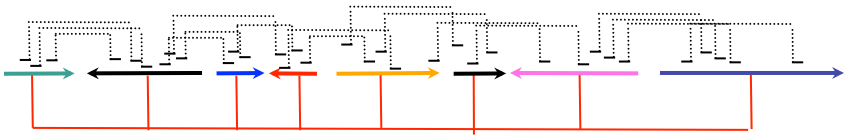

Protein Expression
RNA Sequencing
Structure
Interactions
ChIP-Chip
Orthologs
Subcellular Location
Pathways
Isolates
Putative Function (GO)
biological process
physiological process
cellular process
cellular physiological process
cell cycle
cell division
M phase
meiotic cell cycle
cytokinesis
M phase of meiotic cell cycle
cytokinesis after meiosis I

30,000 ft View – Genome Assembly



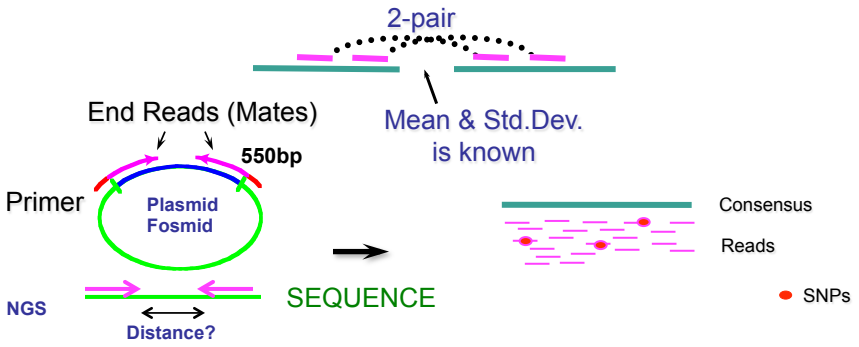
5X genome sequence means that sequences equivalent to 5X the genome size were generate e.g. Genome size = 10 Mbp, then 50Mbp of random sequences were generated

Pairs Give Order & Orientation

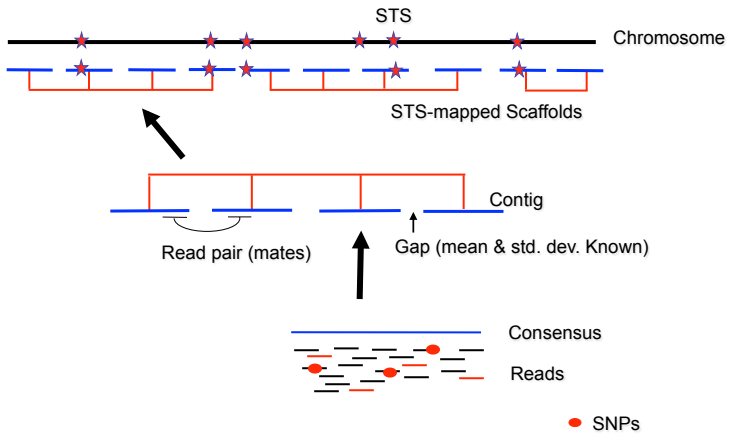


Scaffold

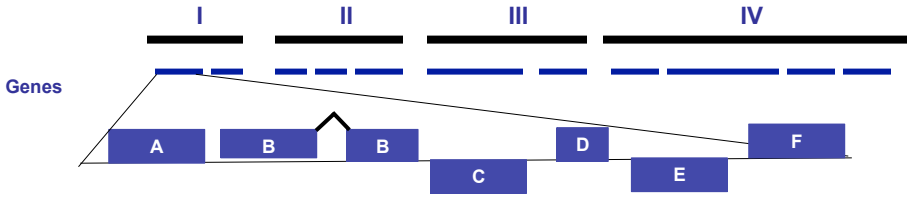
Gaps in scaffolds are traditionally indicated by 100 "N's"

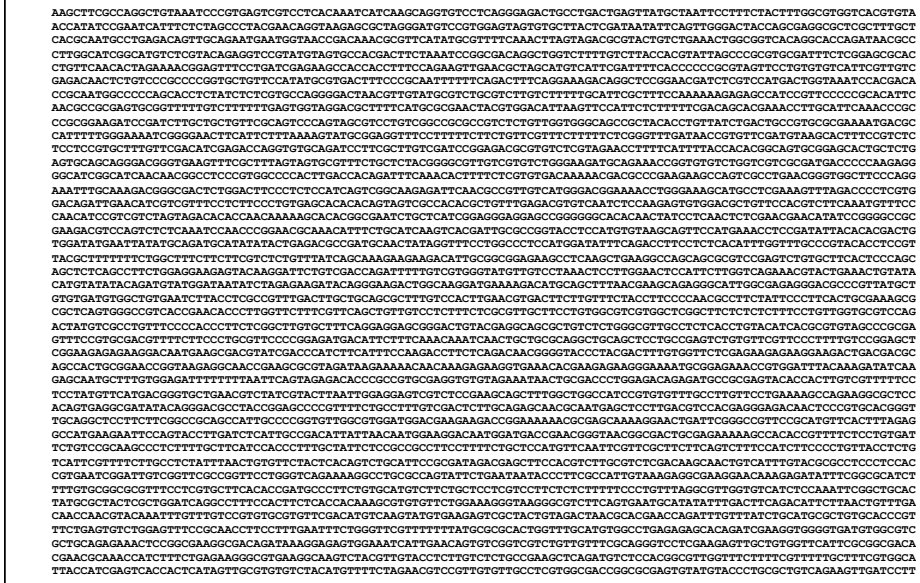


Anatomy of a WGS Assembly

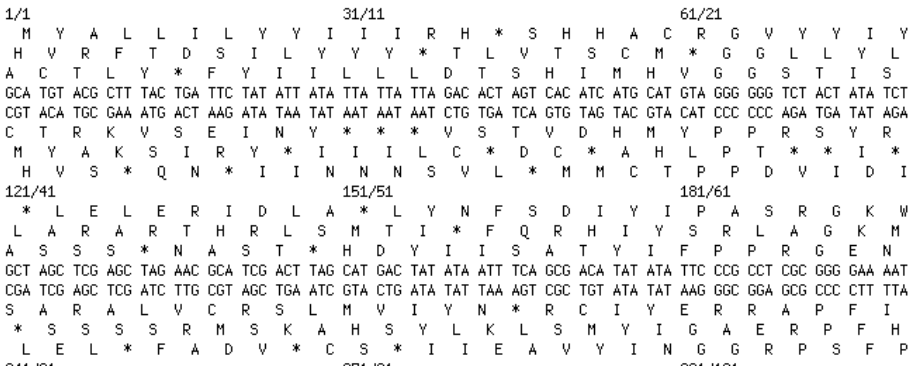


30,000 ft View - Annotation





Six Frame Translation
ORF-finding

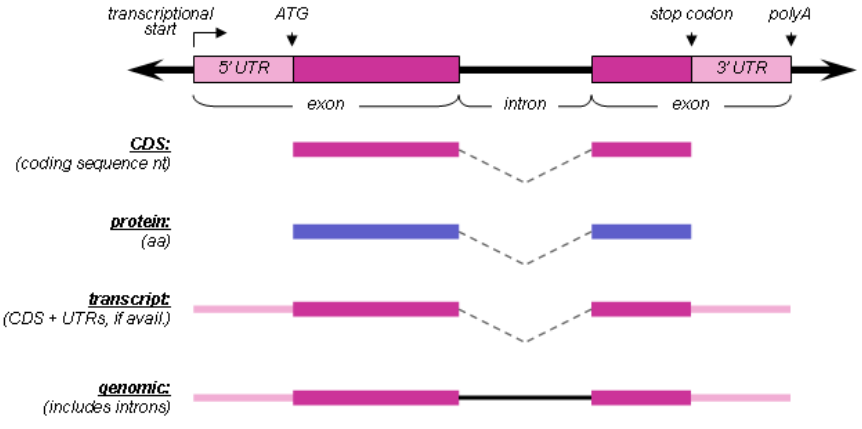


ORFs ≠ Genes

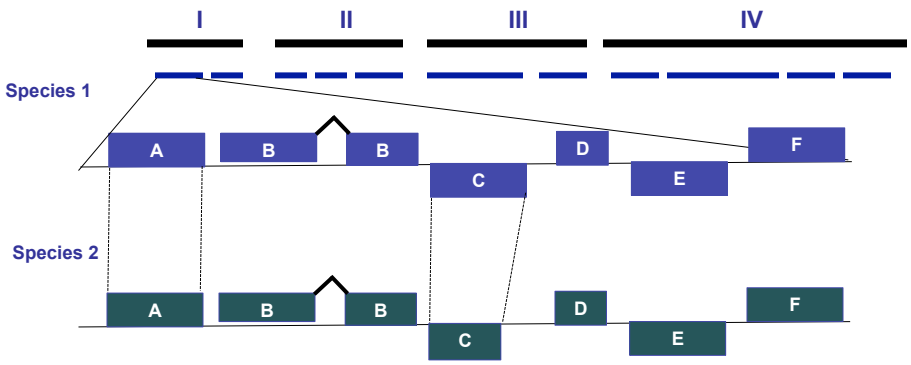
```
ATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC
ACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCGAAGAAGCCAGTCGCCTGAACGGGTGGCT
TCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATG
GGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGACAGATTGAACATCGTCGTTTCCTCTTCCC
TCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGAGGAGCGGACTGTACGAG
GCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGAGTTTCCGTGCGACGTTTTCTTCC
CTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCGT
TCCCTTTTGTCCGGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCA
GACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGAGCA
ACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGCAGCCAT
TGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCAT
GTTCACTTTAGAGGCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATG
ACCGAACGG
```

```
>Translation Frame 1
          MQKPVCLVVAMTPKRGIGINNGLPWPHLTTDFKHFSRVTKTTPEEASRLN
          GWLPRKFAKTGDSGLPSPSVGKRFNAVVMGRKTWESMPRKFRPLVDRLNI
          VVSSSLKEEDIAAEKPQAEGQQRVRVCASLPAALSLLEEEYKDSVDQIFV
          VGGAGLYEAALSLGVASHLYITRVAREFPCDVFFPAFPGDDILSNKSTAA
          QAAAPAESVFVPFCPELGREKDNEATYRPIFISKTFSDNGVPYDFVVLEK
          RRKTDDAATAEPSNAMSSLTSTRETTPVHGLQAPSSAAAIAPVLAWMDEE
          DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADIINNGRTMDDRT
```

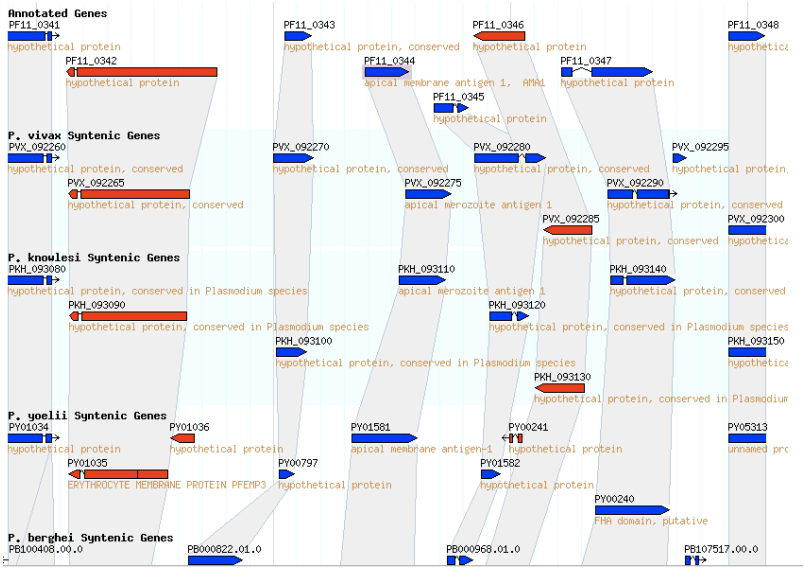
Terminology



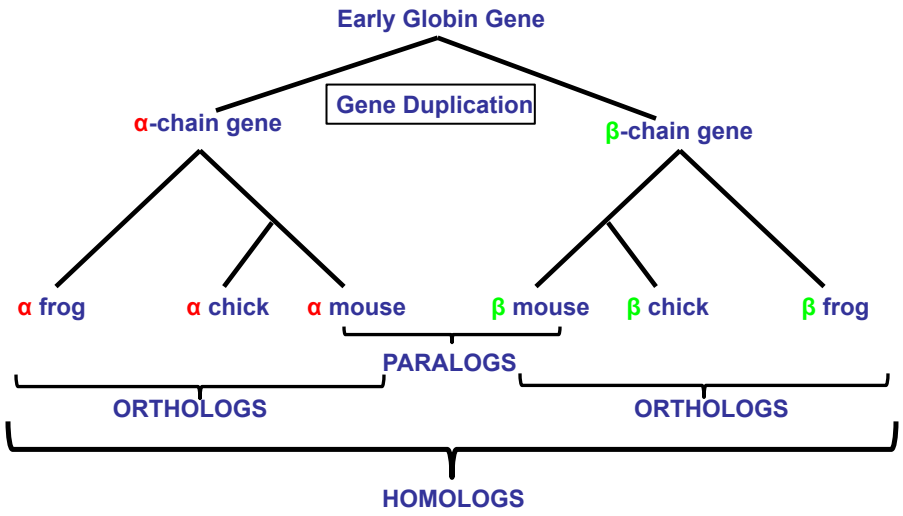
30,000 ft View - Synteny



Synteny among Plasmodia



Homology



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:

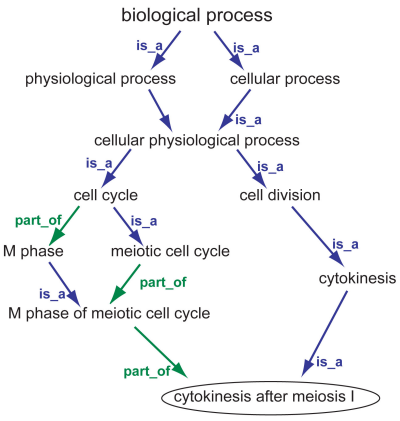
D. melanogaster gene CG3340 annotated as: "*Kruppel*" and *P. falciparum* gene PF3D7_1209300 annotated a "putative KROX1"

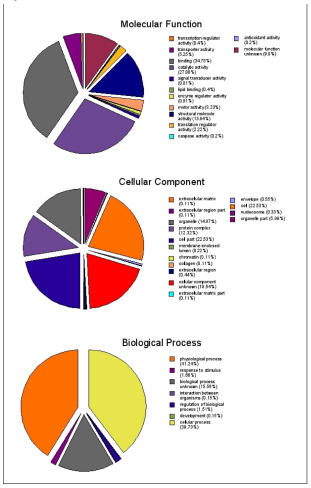
Can both be annotated with GO term:

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





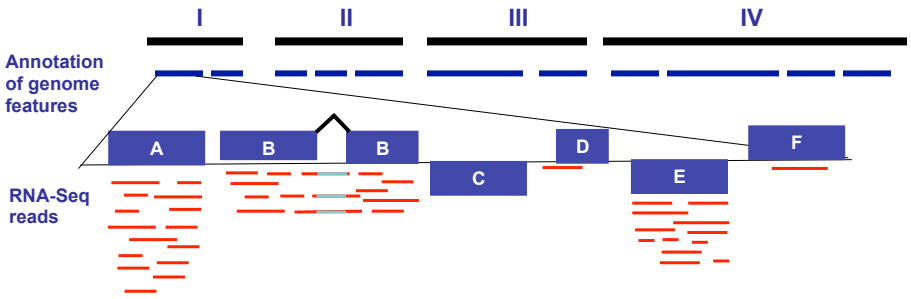
Expression Profiles (RNA and Protein)

- 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

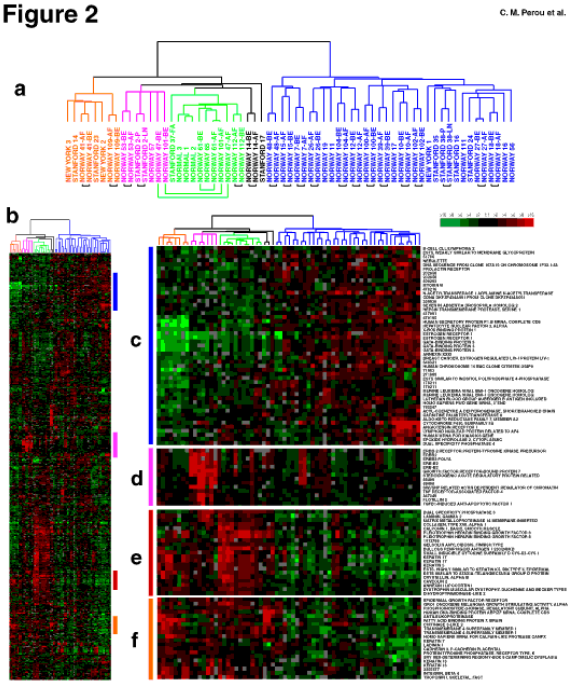
- RNA-Seq (NGS)
  - Little sequence bias
  - Quantitative
  - Usually are strand-specific
  - Can be used to identify UTR's and exon splice junctions
- Expressed Sequence Tags, ESTs
  - Usually represent partial cDNA
  - Often clustered
  - Come from libraries that may, or may not be normalized
  - Often used to identify genes in genomes and locations of introns
- SAGE tags
  - Serial Analysis of Gene Expression

30,000 ft View – RNA-Seq



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

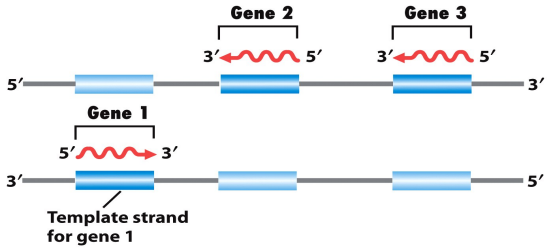


Figure 8-3
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

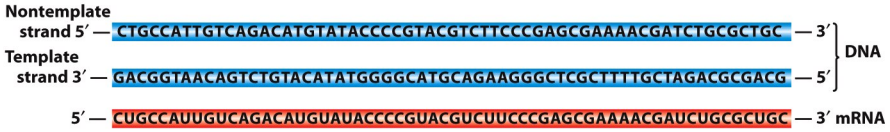
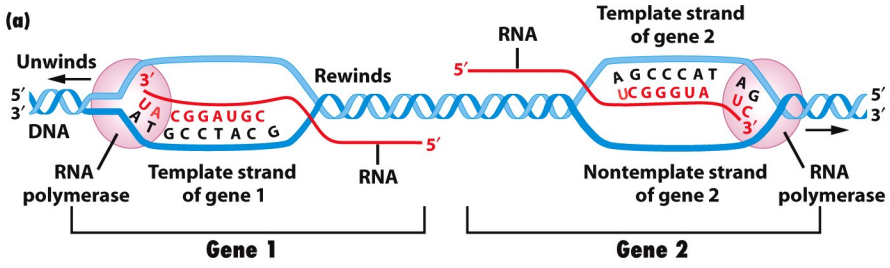


Figure 8-6
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

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



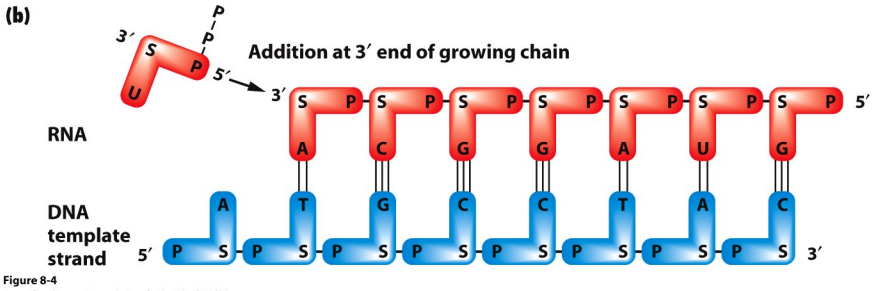


Figure 8-4
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

Complex patterns of eukaryotic mRNA splicing: What is a Gene?

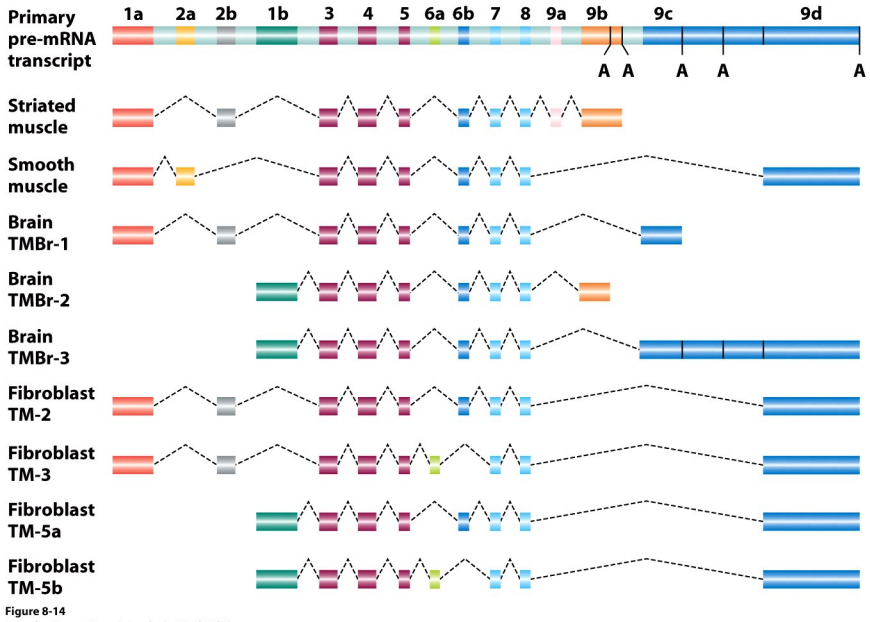
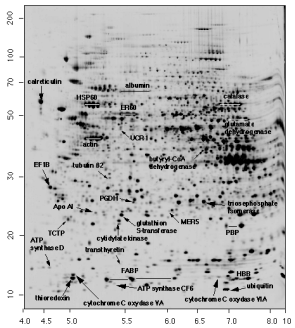


Figure 8-14
Introduction to Genetic Analysis, Ninth Edition

Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans

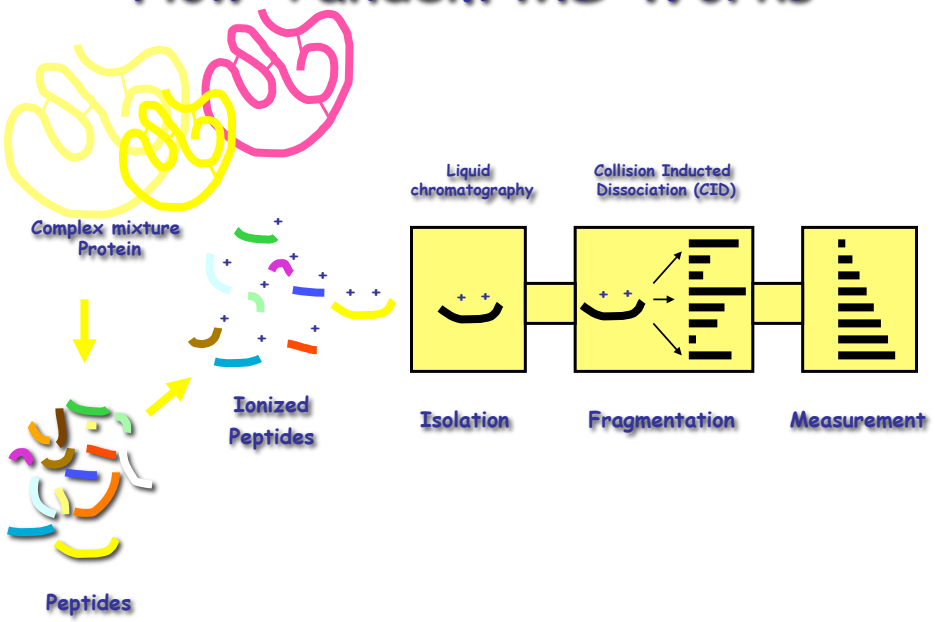


Technology

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

Typical 2 D gel

How Tandem MS Works



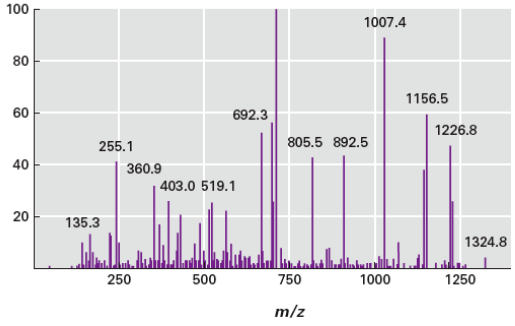
Tandem MS protein data

a)

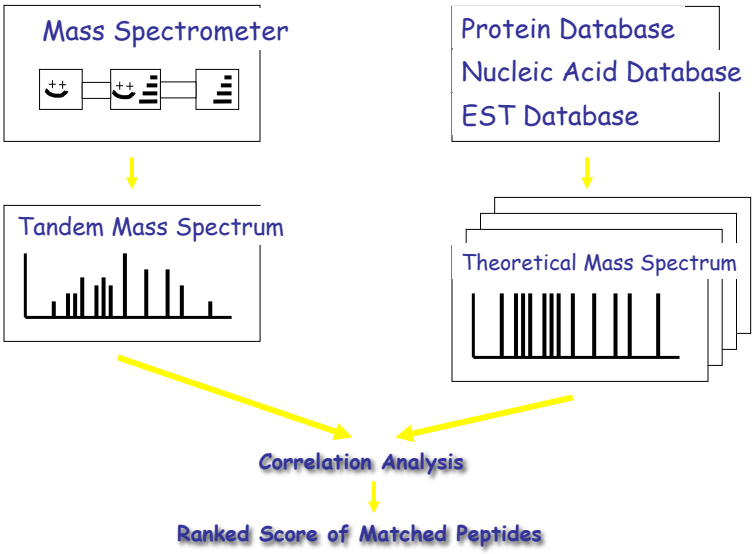
S–P–A–F–D–S–I–M–A–E–T–L–K
(protonated mass 1410.6)

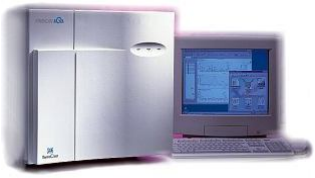
Mass$^+$	b-ions	y-ions	Mass$^+$
81.1	S	PAFDSIMAETLK	1323.6
185.2	SP	AFDSIMAETLK	1226.4
256.3	SPA	FDSIMAETLK	1155.4
403.5	SPAF	DSIMAETLK	1008.2
518.5	SPAFD	SIMAETLK	893.1
605.6	SPAFDS	IMAETLK	806.0
718.8	SPAFDSI	MAETLK	692.3
850.0	SPAFDSIM	AETLK	561.7
921.1	SPAFDSIMA	ETLK	490.6
1050.2	SPAFDSIMAE	TLK	361.5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

b)



Sequest Database Search

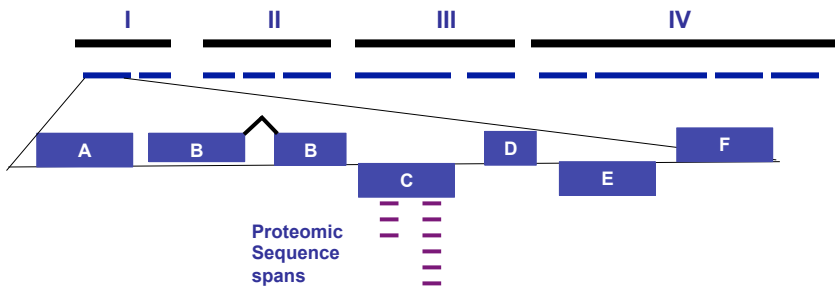




Peptide database

Note: ORFs in addition to predicted Genes must be searched

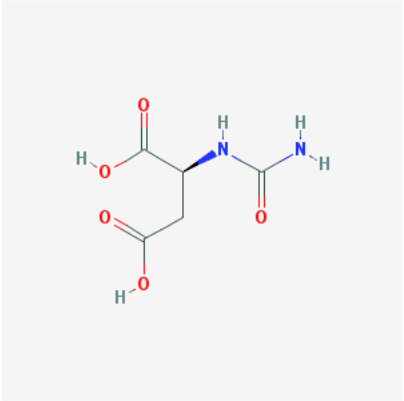
30,000 ft View - Proteomics



Overview

PubChem Compound ID:	CID:93072
PubChem Substance ID(s):	3727
Synonyms:	N-Carbamoyl-L-aspartate
Molecular Weight:	176.12742
Molecular Formula:	$C_5H_8N_2O_5$

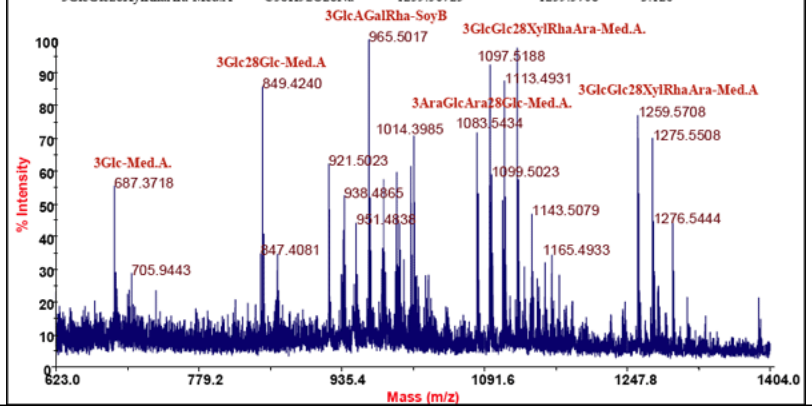
2D Structure



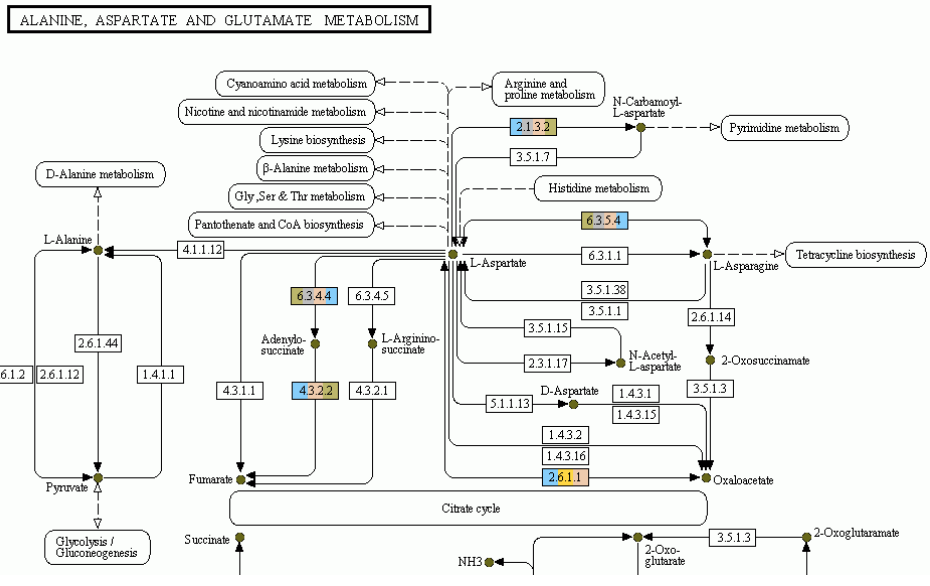
Mass Spectrometry can be used to measure metabolic and other chemical compounds

Complex mixtures can be analyzed and interpreted

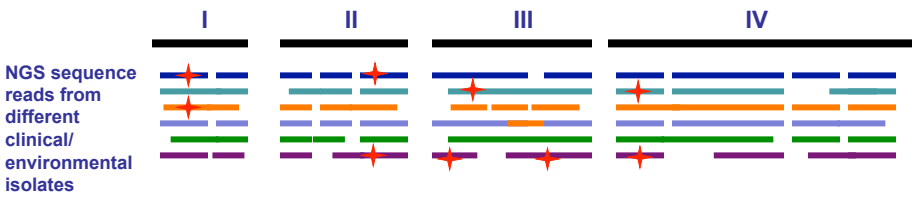
Saponin	Emperical Formula	Theoretical Accurate M/Z	Experimental Accurate M/Z	Error (ppm)
3Glc-Med.A.	C36H56O11Na	687.37201	687.3718	-0.338
3Glc28Glc-Med.A.	C42H66O16Na	849.42483	849.4240	-0.945
3GlcAGalRha-SoyB	C48H78O18Na	965.50856	965.5017	-7.172
3AraGlcAra28Glc-Med.A.	C52H84O22Na	1083.53517	1083.5454	7.715
3GlcGlc28XylRhaAra-Med.A.	C52H82O23Na	1097.51443	1097.5188	3.922
3GlcGlc28XylRhaAra-Med.A	C58H92O28Na	1259.56725	1259.5708	3.126



Metabolites can be linked to metabolic pathways and enzymes



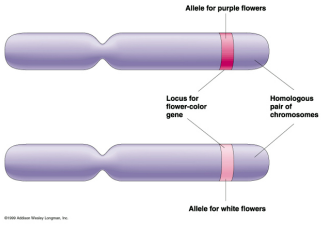
30,000 ft View- NGS SNPs

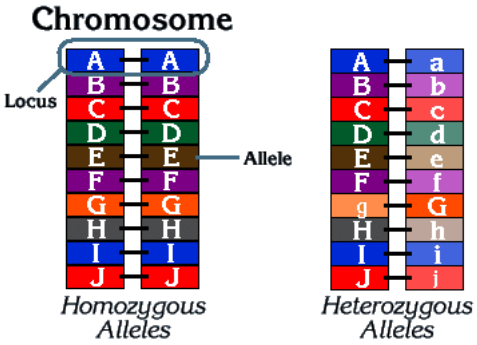


Alleles and Phenotype

- 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 chromsomes (in a diploid)





A AAGCCTCATC

a ACGCCTCATC

SNP =Single Nucleotide Polymorphism

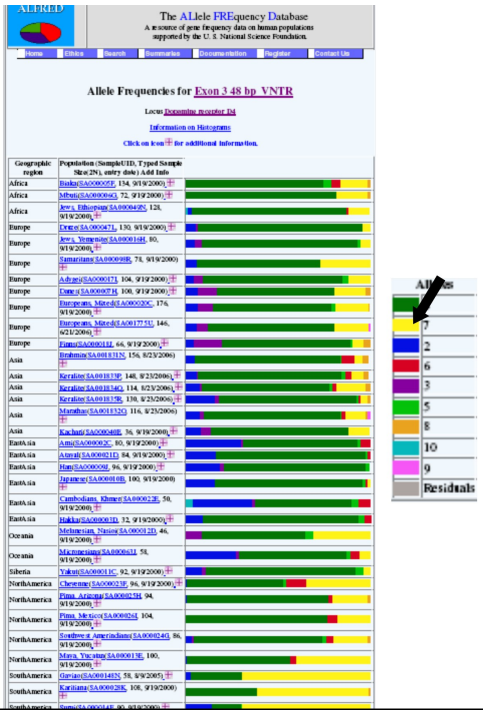
Population data

Data

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

Technology

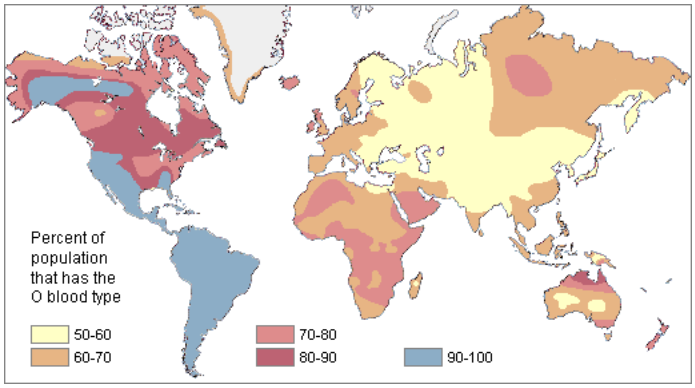
- Chip-Seq
- NGS



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

Data

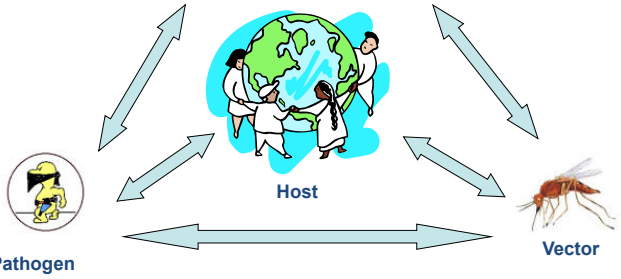
- Species, Strain,
- Isolate
- Location, Date
- SNP
- Sequence
- Allele
- phenotpe

Technology

- PCR-RFLP
- Microsatellites
- Sequencing
- SNP chip
- GPS

Infectious Disease Paradigm

Experimental systems



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

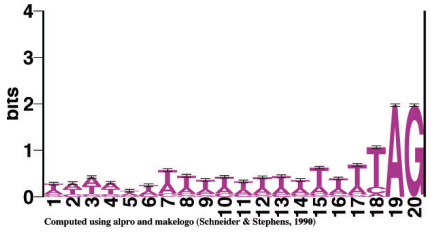
- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats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

How to find an intron

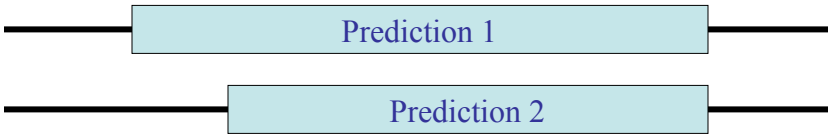
- Usually begins with GT and end with AG
- Must be longer than 19 nucleotides
- Must contain a branchpoint "A"
- Donor GT often followed by a sequence pattern. This pattern is species-specific
- Acceptor AG often proceeded by pyrimidine stretch
- Has a mean length of "X" as is observed in this species

Donor Site
Generated by http://www.bio.cam.ac.uk/seqlogo/log

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



Different prediction methods often generate different results

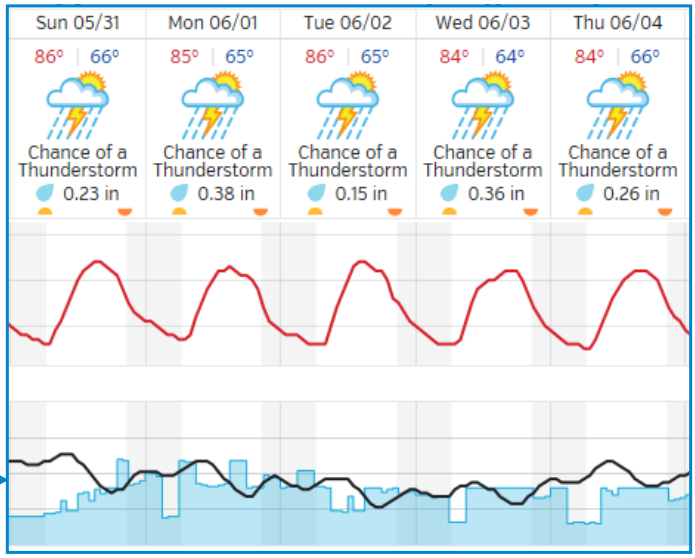


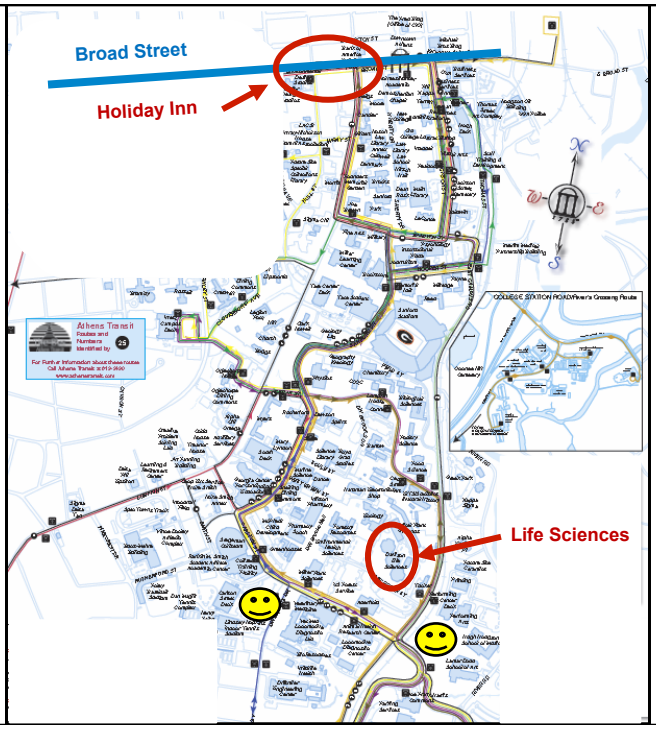
We provide lots evidence so that you can decide or design an experiment to confirm!

Welcome!!!! Logistics

EuPath Workshop 2015

Expect thunderstorms which can be dramatic.
Air conditioned buildings may be cold.





General Schedule

- Mon – 8:30 – 6:00
- Tue & Wed 9:00 – 6:00
- Breakfast use coupons at Holiday Inn,
- Lunches, coffee and snacks provided
- Dinner at the Coverdell Center on Monday - YUM
- Tuesday and Wednesday dinner is on your own.
- Tuesday or Wednesday hike at Bot Gardens?

How to get to the Workshop Room

- Hotel Lobby 8:00 Monday for RIDE to workshop
  - Guides (Haiming and Betsy) will leave hotel lobby tomorrow morning
  - Tuesday and Wed = 8:20 in lobby
- Hotel Lobby 7:45 am for walking
  - Brian walking 7:45 from hotel lobby and it takes about 25 min
  - Tuesday and Wed = 8:15 in lobby


EuPathDB
Eukaryotic Pathogen Database Resources
A EuPathDB Project
Gene ID: cgd7_230
Gene Text Search: synth*
About EuPathDB | Help | Contact Us | Login | Register
Home
New Search
My Strategies
My Basket (0)
Tools
Data Summary
Downloads
Community
Data Summary
News
3 May 2011 AmoebaDB 1.4 Released
3 May 2011 New Features in GiardiaDB
3 May 2011 MicrosporidiaDB 1.4 Released
3 May 2011 PlasmoDB 7.2 Released
3 May 2011 ToxoDB 6.4 Released
3 May 2011 New Features in TrichDB
3 May 2011 TriTrypDB 6.4 Released
All EuPathDB News
Community Resources
Web Tutorials
Information and Help
EuPathDB Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Diseases is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens (Babesia, Cryptosporidium, Encephalitozoon, Entamoeba, Enterocytozoon, Giardia, Leishmania, Neospora, Plasmodium, Theileria, Toxoplasma, Trichomonas and Trypanosoma).
Brian Jessie Eileen Omar Mark Susanne Betsy Cristina
Identify People by:
Expand All | Collapse All
Name
Gender
Education
Nationality
Disposition
Sunny
Not so sunny
Identify Other Data Types:
Expand All | Collapse All
Lion
Tigers
Bears
Oh My
EuPathDB 2.10 May 3, 2011
©2011 The EuPathDB Project Team
EuPathDB
Please Contact Us with any questions or comments

Questions?

Now that you know us, We would like to know you

- Please tell us your name and a bit about your research.
- Also, if there is something specific that you came here to learn, please state it so that we can cover it if possible