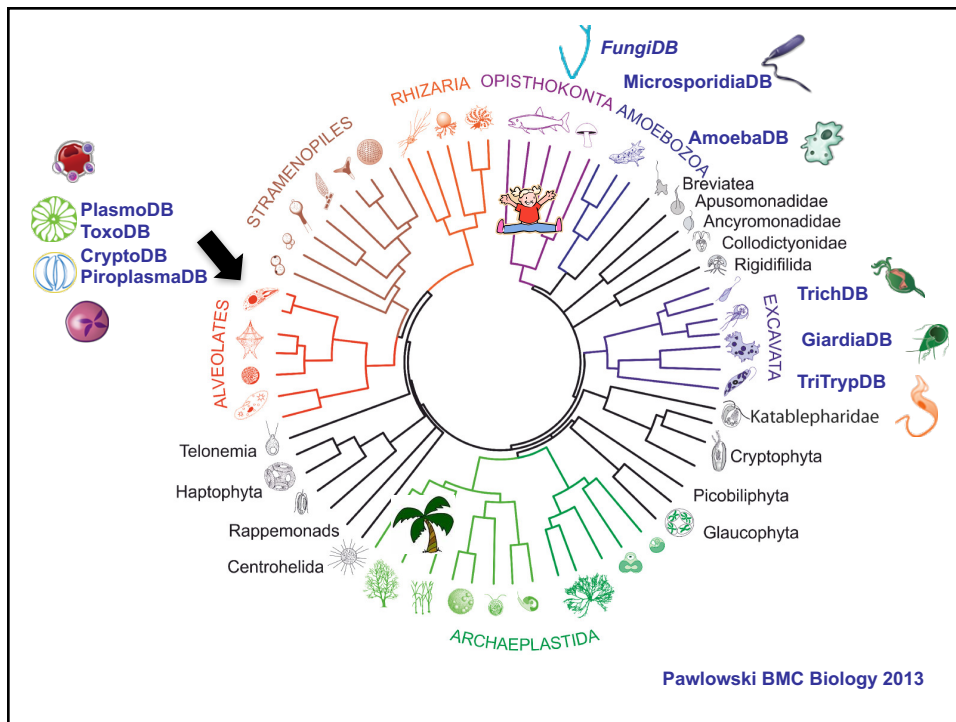
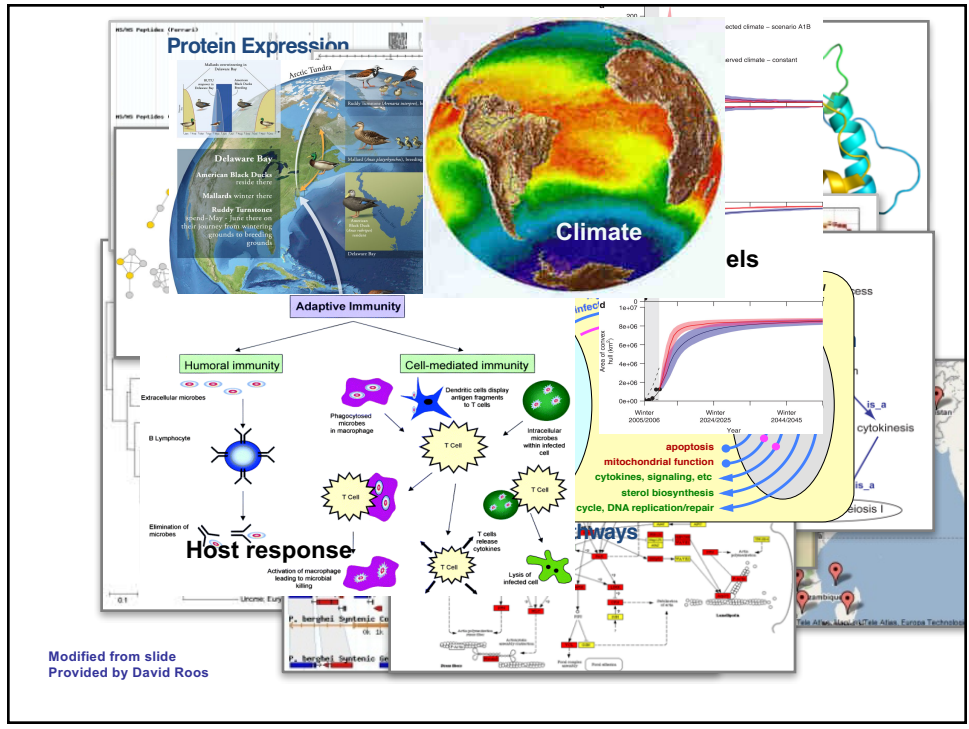
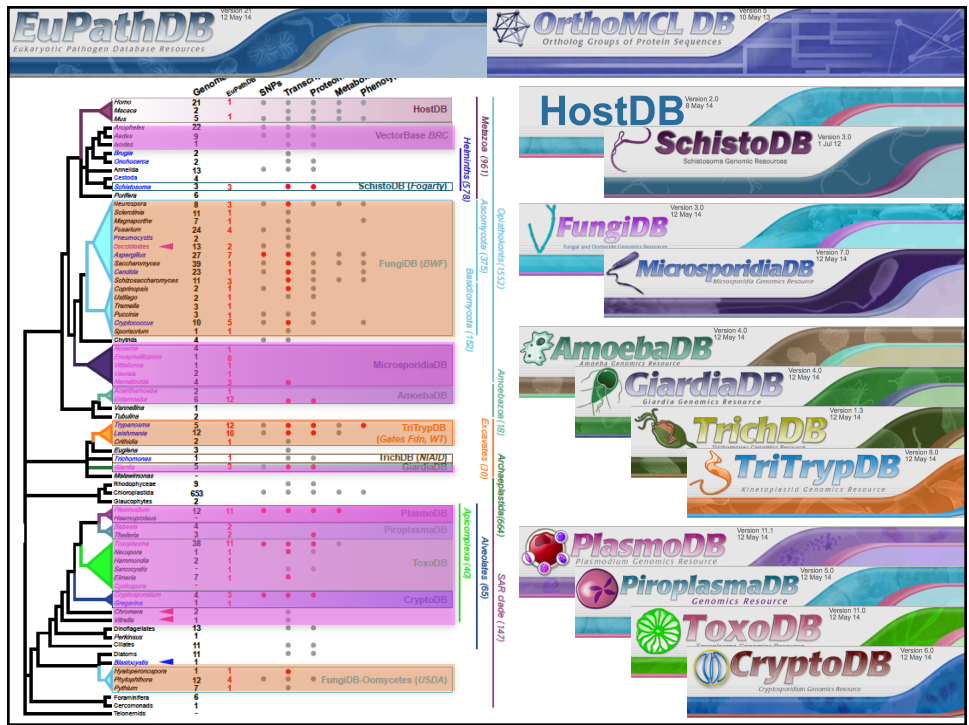


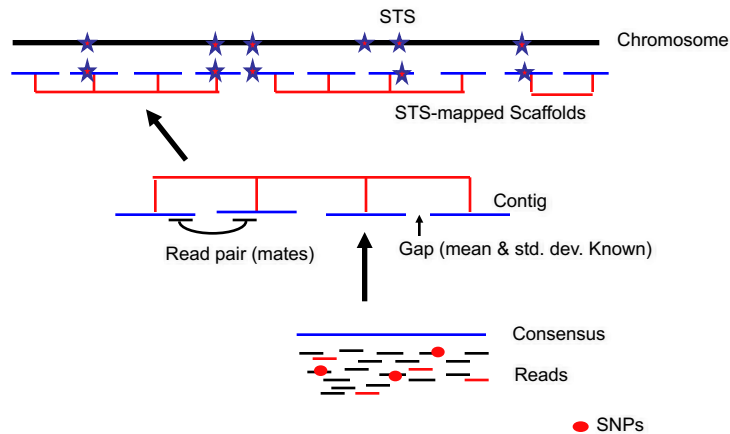
Crash Course in Omics Terminology, Concepts & Data Types

Jessie Kissinger
August 8th, 2016

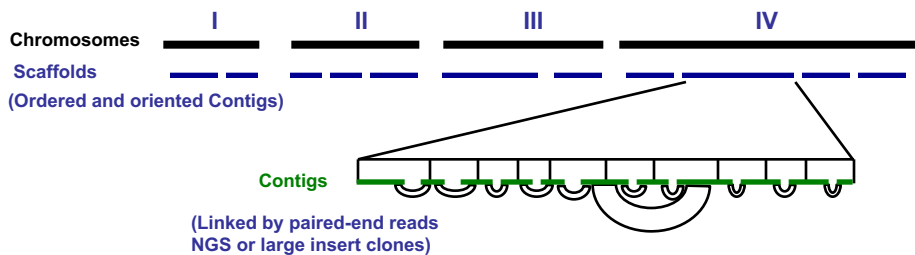




Anatomy of a WGS Assembly

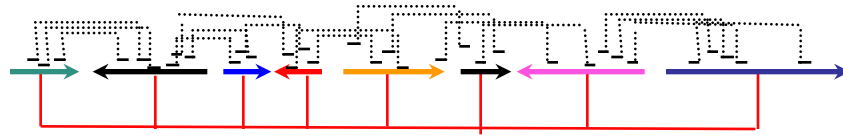


30,000 ft View - Genome Assembly

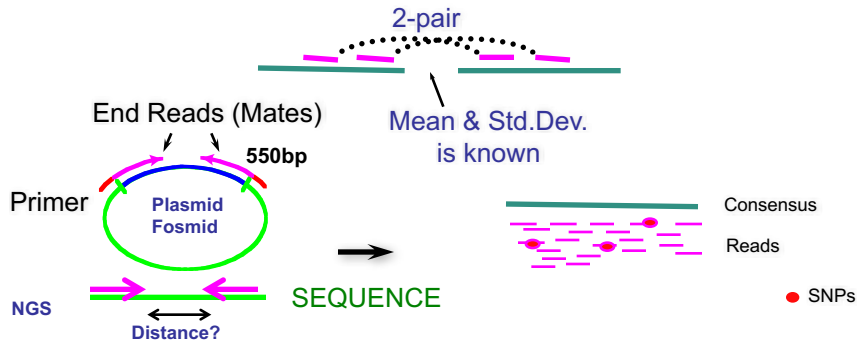


5X genome sequence means that sequences equivalent to 5X the genome size were generated e.g. Genome size = 10 Mbp, then 50Mbp of random sequences were generated

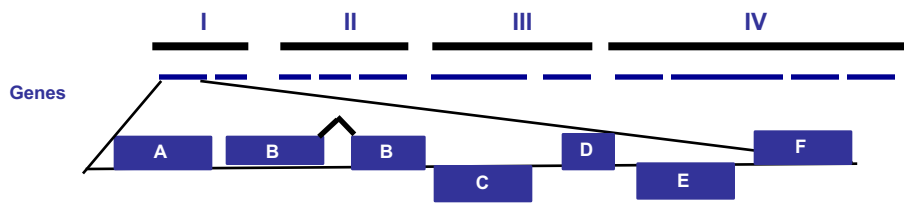
Pairs Give Order & Orientation



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



30,000 ft View - Annotation



AAAGCTTCGCCAGGCTGTAATCCCGTGAAGTCCCTCAAAATCATCAAGAGGTCCTCAGGGAGACTGCCTGACTGAGTTATGCTAATTCCTTTCTACTTTGGCGTGGTCCAGGTGA
ACCATATCCGAATCATTTCTAGCCCTACGACAGGTAAGAGCGCTAGGAGTCCGTTGGAGTAGTGTCTTAACGATAATATTCAGTTGGGACTACAGCGAGGCGCTCGTTCCT
CAGGCAATTCCTGAGACAGTGGCAGATGAAATGTAACCGACAAACGGTTCATATGCGTTTTCAAAATAGTAGAGCGGTACTGCTGAAACTGGCGGTACAGGCGACAGATAACGCC
CTTGGCATGGCAGTCTGTACAGAGTCCGTATGATGTCCAGACTTCATAATTCGGGAGAGGCTGGTCTTTGGTTACACAGATTAAGCGCGGTGGGATTCTCGGAGCGAC
CTTTCACACTGAAAGAGGGATTTCTGTGAGAGGCGCACCTTCCAGAAATGAAAGCGCTGAGCTATGATTTCCACCCCGCGGTGCTGTGTGATTCGTTGTGCTG
GAGACAACTCTGTCCCGCCCGGTGTGTCCATAATGCGGTACTTCCCGCAATTTTTTCAGACTTTCAGAAAGACAGGCTCCGGAACGATCTCGTCCATGACGTGTAATCCAGCA
CCGCAATGGCCCGACCTTATCTCTCGTCCAGGGGACTAAGCTGTATGCGCTGCGTCTTTGCTTTTTGCTATTCCTTTCAAAAGAGAGGACCTCCGTTCCCGCCGACATTC
AACGCCCGGATCGGGTTTTGCTTTTTGAGTGGTAGGAGCTTTTCATGCGCGAATACGTTGGAATTAAGTTCCATCTCTTTTTGACAGCAGCAAAACCTGCAATCAAAACCGG
CCCGAGGAATCCGACTTTGCTGCTGTCAGCTCCAGTAGCGTCTGTGCGCCGCGCTCTCTGTTGGTGGGCGCGCTACACCTGTTATCTGACTGCGCGTGGCAAAATGAGCG
CATTTTTGGGAAATCGGGAACTTCATTTTAAAGTATGGCGAGTTCCTTTTTCTGTCTGCTTTTTCTTTCTGGGTTGATAACCGGTTTCGATGAAAGCACTTCGCGTCT
TCTCCGTCCTTTTCAGACTCGAGCCAGGTCTCGAGTCTTCCTGCTGTGATTCGGAGACCGTCTGCTGTAGACTTTTTCAATTTACACAGGCACTGGGACACTGCTG
AGTCGACAGCGAGCGGTGAGTTTCGCTTACGCTAGCGTTCCTGCTTACGGGCGTGTCTGTGTGAGAGTGCAGAAACCGTGTCTGCTGTCCGATGACCCCAAGAG
GGCATCGGCATCAACACCGCTCCCGTGGCCCACTGACCAACAGATTTCAACACTTTTTCTGTGTGCAAAAACAGCCCGCAAGAGACGCTGCGTGAACGGGTGCTCCAG
AAATTTGAAAAGCGGGGACTCTGGACTCCCTCTCATAGTCGGCAAGAGATTAACCGCGTGTCTATGGGACGGAAAACCTGGAAAAGCATGCTGAAAAGTTTAGACCGTGTG
GACAGATGAACATCGCTGTTCCCTCTCCCTGTAGACACACAGTAGTGCCACACCGTGTGAGAGGTGCAATTCGAAGAGTGGACGCTGTCCACGCTCTCAAAATGTTTC
CACATCGCGTGTCTAGTAGACACCAACAAAAGACACAGCGGAACTGTCTCATCGAGGGAGGAGCCGGGGGCACACACTATCCCACTCTCGAACGAACATTCGGCGCGC
GAGACCTCGCTCTCAATCCAAACCGGACGCAACACTTCGACATGACGATTGGCGCGTACCTCCATGTTGTAAGCAGTCCATGAAACTCCGATATTAACAGACAGCT
TGGATATGATATATGAGATGCAATATATCGAGACCGGATGCACTATAGGTTTTCTGGCGCTCCAGGATATTCAGACTCTCTCCACATTTGCTTCCCGTACCTCCGCT
TAGCCTTTTTCTGCTCTTCTCTCTGTTATCAGCAAGAAGAGACATTGCGCGGAGAACCTCAAGCTGAAGGCGAGCGCGTCCGAGTCTGTACTCCAGC
AGCTCTCAGCCTTCGAGGAGAGTACAGGATTCTGTGACCAAGATTTTGTCTGGTATGTTGTCTAACTCTCTGGAACCTCAATTTGGTCAGAAAGCTGAAACTGTATA
CATGTATACAGATGTATGGATAATCTAGAGAAGATACAGGAAAGCTGGCAAGGATGAAAAGCATGCACTTTAACGAAAGAGGGCATTGGCAGAGGAGCGCCGTTATGCT
GTGTATGTGGCTGTGAATACCTCCCGGTTGACTGTCTGACGCGTTTTGCTCCACTGAACTGACTTGTGTTTACTCTCCCAAGCGCTTATTCCCTCACTCGCAAGCGC
CGCTCAGTGGCTCCAGCAACCCCTGGTCTTTCTGTCAGCTGTGTCTCTTTCTGCGCTGCTCTGTTGGCGTGGGCTGGGCTGGCTCTCTCTCTTTCTGCTGGTCCAG
ACTATGTCGCTGTTTTCCCGACCCCTCTCGGCTGTGTCTTCAGGAGGAGCGGAGCTGTAGAGAGCGAGCGTCTCTGGCGTCTGCTCTCCTCTGATGATCACCGGTGTAGCCGGA
GTTCCTGTAGACTTTTTCTCCCTCGTCCCGGAGATGACTTTTTCAACAACTAACTCTCGAGAGGCTGAGCTCCCTCGGAGTCTGTGTCTCTCTGCTGCTGCTGCTG
CGGAGAGAGAGGACCAATGAGCGACTATGACCCACTCTTCCAAAGACTCTCAGACAAAGCGGTACCTCAGACTTTGTGGTCTCGAGAGAGGAGAGACTGAGCAGCC
AGCACTCGGAGACCGGTAAAGCGCAACGAAGCGCTGATAGAAAACAAACAAAGAGAGGTAAGAACAGAGAGGAGAAAATCGGAGAAAACCGTGATTACAGAAATACAA
GCCAATCGGAGATTTTTTAAATTCAGTAGAGACCCCGCGGTGGAGGTGTAGAAAATAACTGCGACCTGAGAGACAGAGATGCGCGAGTACACCACTTTTTCT
TCCTATGTCATGACGGTGTGAAGCTCTCTGACTTAAATGGAGGAGTGTCTCCGAGGAGCTTTGGCTGGCCATCCGCTGTGTTTTGCTGTCTGAAAAGCAGAGGGCGCTC
ACAGTGGCGATATACAGGAGCGCTACCGGAGCGCGTTTTCTGCTTGTGACTCTTGCAGAGCAACGATGAGCTCCTTACGCTCCAGGAGGAGCAACTCCCGTGCAGCGGT
TGCAGGCTCTCTTGGCGAGCAATGGCGGTTGGCTGGCTGGAGAGAGAGCGTGAAGAAAACCGGAAAGAACTGATGGGCGGTGGCGGAGCTGCAATGTTACTTAGAG
GCCATGAGATATCCAGTACTTGAATCTATGCCAGATATTAAACGAAGAGACAATGGAGCCGACCGGTAAGCGCACTGCAGAAAAGCCACACCGTTTTCTCTGTGAT
TCTGCGGAAAGCCCTTTGCTTCACTCACCTTGTCTATTCGCGCGCTCTCTTTTTCTGTCTCAATGTTCAATGTTGCTCTCTCACTCTTCCACTCTCCCGTTCACCTG
TCAATGCTTTTTCTGCTTATTAACTGTGTTACTACAGTCTGATTCGCGGATGAGCAGCTCACGCTGCTGCTCGCAAGCAACTGTCATTGTAGCGCGCTCCCTCCAC
CGTGAATCGGATTTGCGTTCGCGGTTCTGGGTCAGAAAAGCGCTCGCCAGTATTCTGAATAATACCTTCGCAATGTAAGAGGCGAGGAACAAAGAGATATTTCGGCGCACT
TTTTGCGCGCGCTTTCTGCTGCTTCCACAGATGCCCTCTGCTGCATGCTTCTGCTCTCGTCTCTCTCTTTTTCCCTGTTTGGGCTTGGTGTACTCCAAAATCGGCTGCAC
TATGCTACTCGTGGATCAGGCTTTCACCTCTCACCAAAAGCTGTGTCTGAAAGGTAAGGCGCTTCACTGAAATGATATTTGACTTCAGACTCTTCAACTGTTTGA
CAACCAGCTCAAAATTTTTGCTCGGTGGCTGTGCAATGCTAGAGAGTCTACTGTGACATCAAGAGAACAGATTTTTTAACTGCAAGGCGGTGGCAGCT
TTCTGAGTCTGAGATTTCGCAACCTCCTTTGAAATTTCTGGTGTGTTTTTATGCGCGCACTGTGTTGATGTGCGCTGAGAGAGACAGATGAAAGGCGGTGATGTGGCGCT
GCTGAGAGAACTCCGCGGAGAGGCGACAGATAAGAGAGATGGAATCATTGAACAGTGTGCGTGTGTTTTGCGAGGCTCCTGAAAGTGTGTTGTTCTTCCGGCGGACA
CGAACCGCAACCTCTTTCGAAAGGCGTGAAGGAGTCTACTGTTGACCTCTGCTCTGCGGAAAGCTCAGATGCTCCACCGCGGTGGTTTTCTTCTGTTTTGCTTTCGCGGCA
TTACCATCGAGTCAACCTATAGTTGCGTGTCTACATGTTTTCTAGAAGTCCGTTGTGTTCCCTGTTGGGACCGCGGAGTGTATGACTCGCGTGCAGAAATGATCTT

Six Frame Translation ORF-finding

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

ORFs ≠ Genes

```

ATGCAGAAACCGGTGTGTCTGGTCTGGGATGACCCCAAGAGGGGCATCGGCATCAACAACGGCTCCCGTGGCCCC
ACTTGACCACAGATTCAAACACTTTCTCGTGTGACAAAACGAGCCCGAAGAACCGAGTCCGCTGAACGGTGGCT
TCCAGGAAATTTGAAAGACGGCGACTCTGGACTCCCTCTCATCAGTCGGCAAGAGATCAACGGCGTGTCAATG
GGACGAAACCTGGGAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCTTCTCTCC
TCAAAGAAAGACATTGGCGGAGAGCCCTAAGCTGAAGCCAGCAGCGCTCGAGTCTGTGCTTCACTCCACG
AGCTCTAGCCTTCTGGAGGAGATCAAGGATTCTGTCCAGCAGATTTTGTCTGGAGAGCGGACTGTACGAG
GCAAGCCGTCTCTGGCGCTTCCCTCTCACTGTACATCAAGCGGTGAGCCCGAGATTTCCGGCAGCTTTCTTCC
CTCCCTTCCCGGAGATGACATCTTTCAAACAATCAACTGCTGCCAGGCTCAGCTCTCCGAGTCTGTGTCTGT
TCCCTTTTCCGGGCTCGGAAGAGAAAGCAATGAAGGAGCATCGACCATCTTCAATTTCAAGACCTTCA
GACAACGGGTACCTACGACTTGTGTTCTGAGAAGAGAAAGACTGACAGCCAGCCACTGGCGAACCGACA
ACGCAATGAGTCCCTTACGCTCCACGAGGGAGACAACCTCCGTCACGGGTTGACGGCTCTTCTTCCGCCGACGCAT
TGCCCCGTGTTGGCCTGGATGGACGAAGACCCGAAAAACCGAGCAAAAGGAACTGATTCGGGCCGTCCCGCAT
GTACACTTAGAGCCATGAAGATTCAGTACCTTGATCTCATTCGACATTATTACAAATGGAAGGCAATGGATG
ACCGAAGG

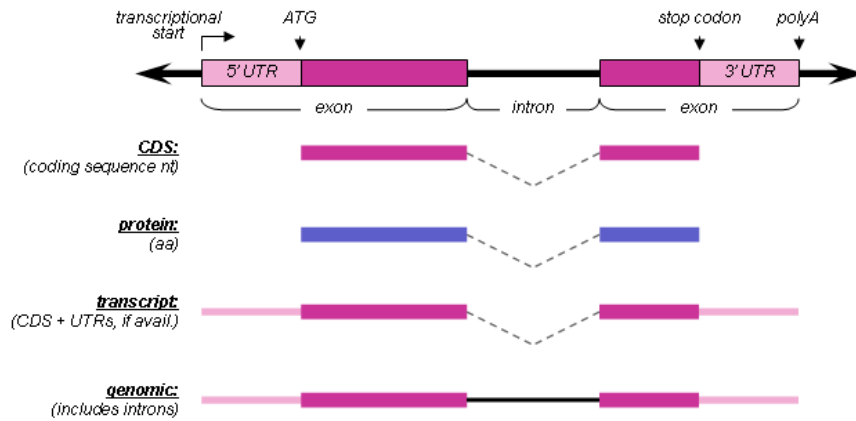
```

```

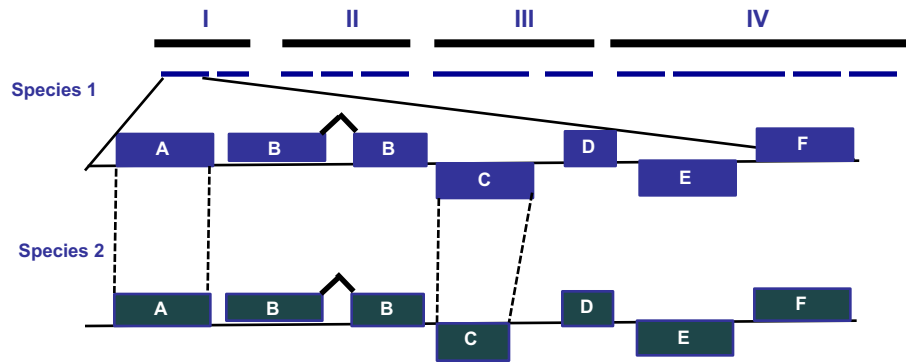
>Translation Frame 1
MQKPVCLVVAMTPKRGIGINNLFPWHLTTDFKHFSRVTKTPEEASRLN
GWLPRKFAKTGDSGLPSPVGRFNNAVVMGRKTWESMPRKFRPLVDRINI
VVSSSLKEEDIAAEKPPQAEQQRVVCSLPAALSLEEEYKDSVDQIFV
VGGAGLYEAALSLGVASHLYITRVAREFPDVFPAFPGGDILSNKSTAA
QAAAPAESVVFPCPELGREKDNEATYRPIFISKTFSDNGVPYDFVLEK
RRKTDAAATAEFSNAMSSLTSTRETTFVHGLQAPSSAAAIAPLAWMDEE
DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADI INNGRTMDDRT

```

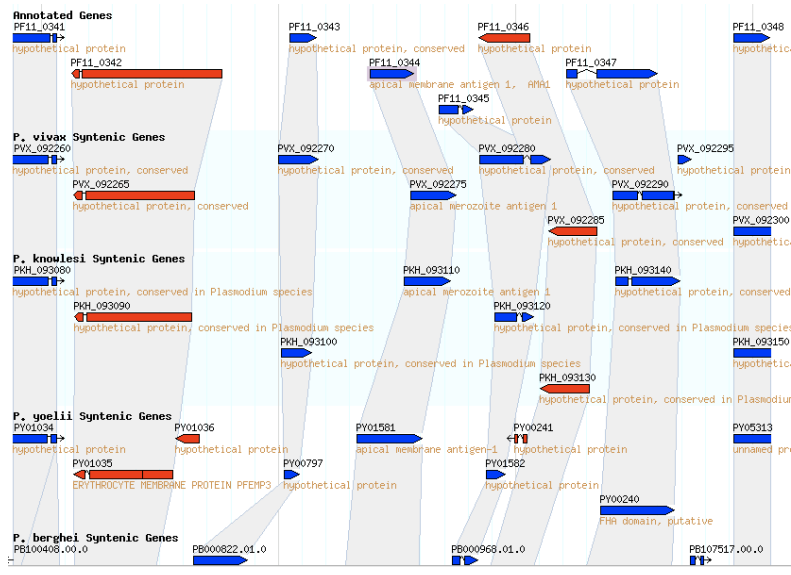
Terminology



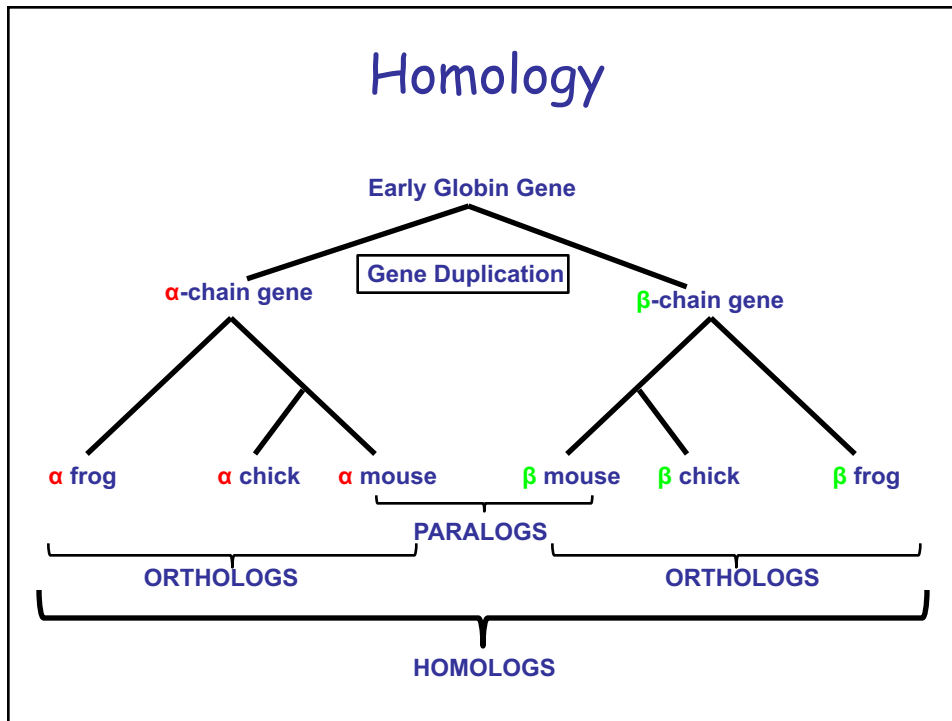
30,000 ft View - Synteny



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.

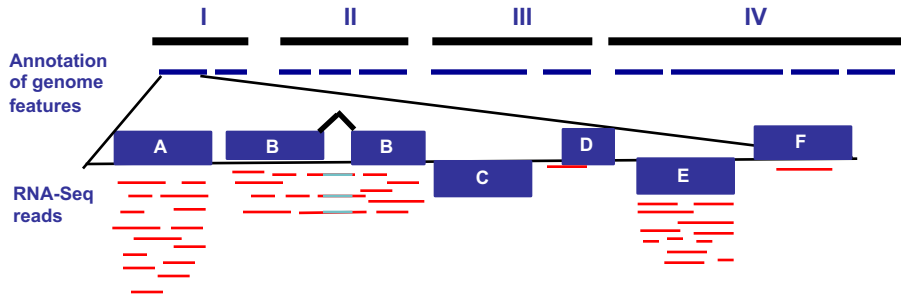
Expression Profiles (RNA and Protein)

- 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

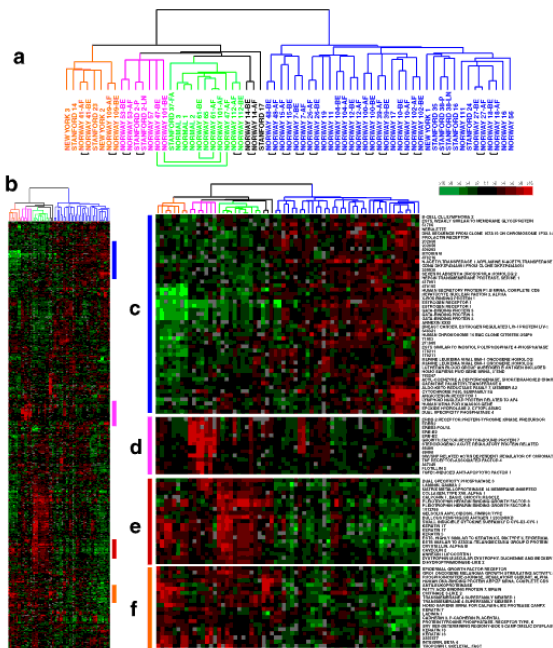


FPKM = Fragments per kilobase of exon per million fragments mapped

**Clustered
Microarray
Data
Genes with
Similar
Expression
Profiles are
Grouped
together**

Figure 2

C. M. Perou et al.



Genes can be located on either DNA strand
 Convention - Gene location = non-template strand, i.e.
 same as the mRNA

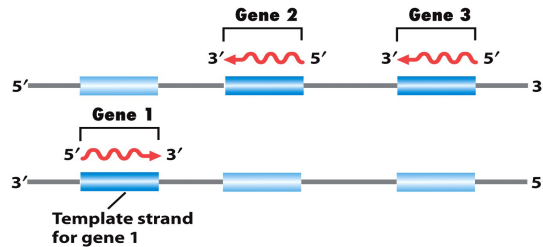


Figure 8-3
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

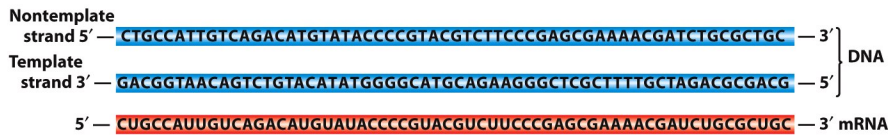


Figure 8-6
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

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

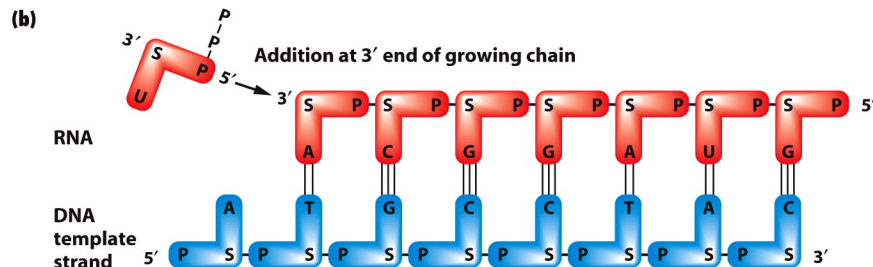
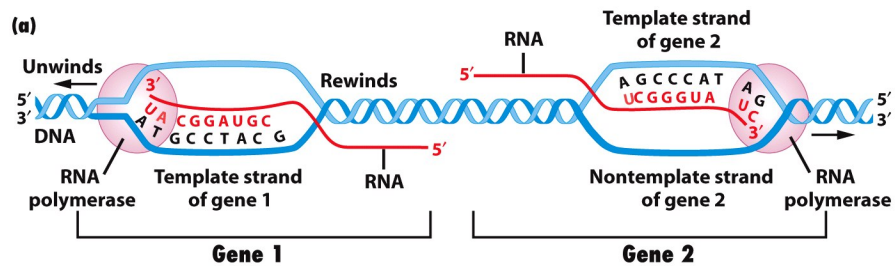


Figure 8-4
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

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

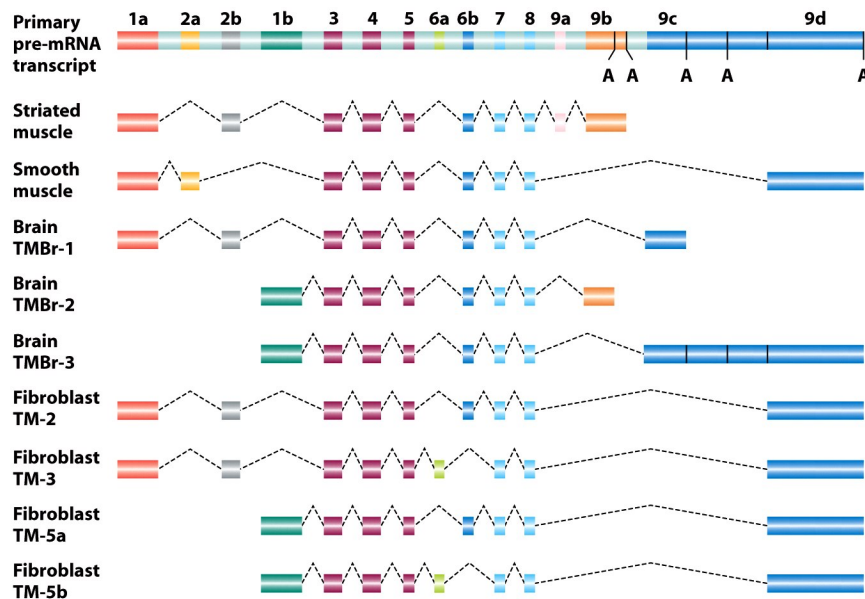


Figure 8-14
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

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 preceded by pyrimidine stretch
- Has a mean length of “X” as is observed in this species

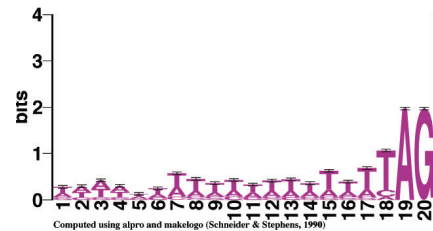
Donor Site

Generated by <http://www.bio.cam.ac.uk/seqlogo/logo>

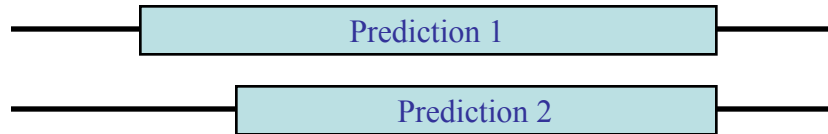


Acceptor site

Generated by <http://www.bio.cam.ac.uk/seqlogo/logo>



Different prediction methods often generate different results



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

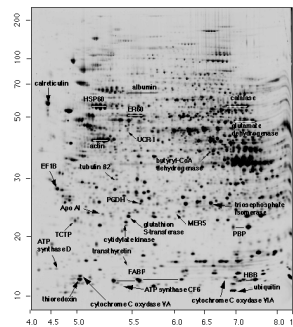
Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans

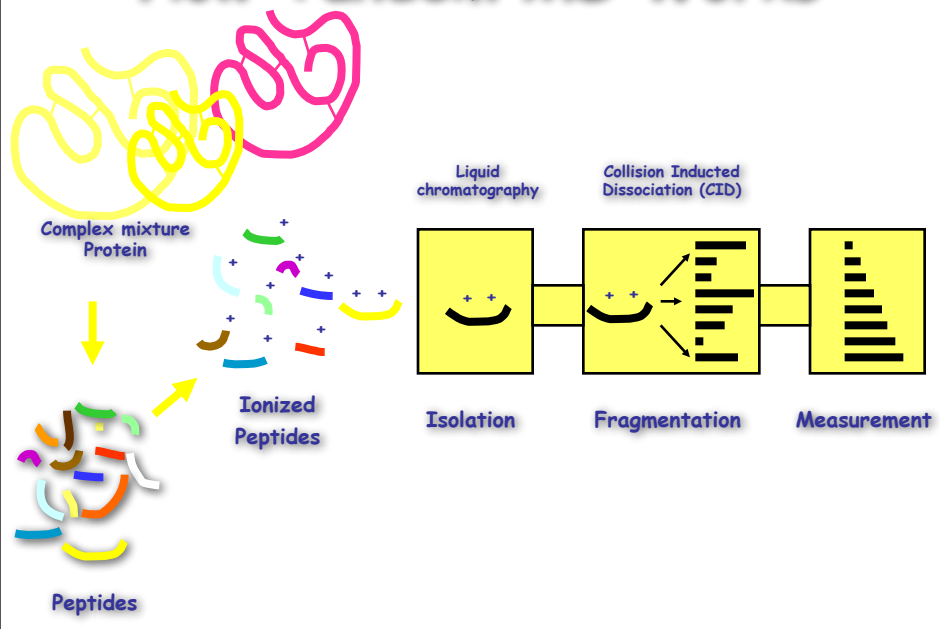
Technology

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



Typical 2 D gel

How Tandem MS Works



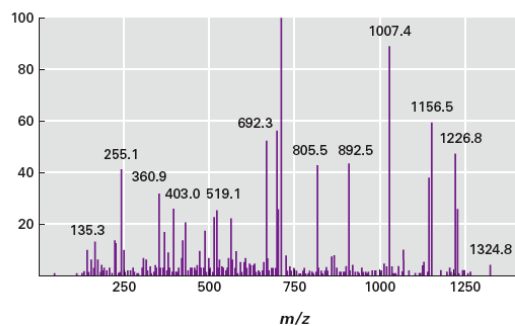
Tandem MS protein data

a)

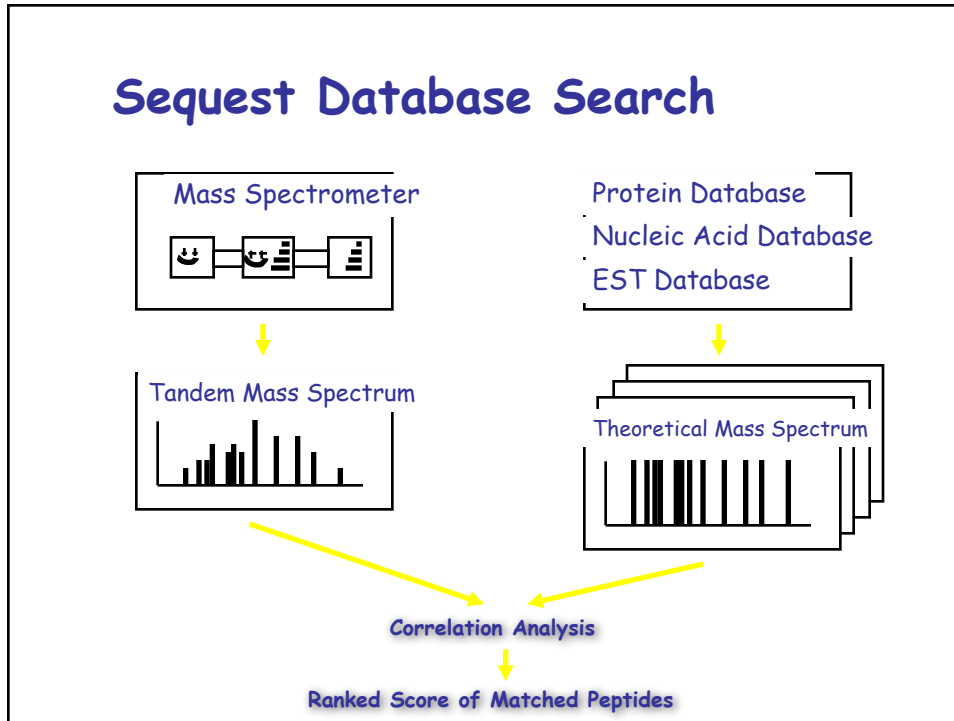
S-P-A-F-D-S-I-M-A-E-T-L-K
(protonated mass 1410.6)

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	SPAFDSIMAE	LK	260.4
1264.4	SPAFDSIMAETL	K	147.2

b)



Sequest Database Search



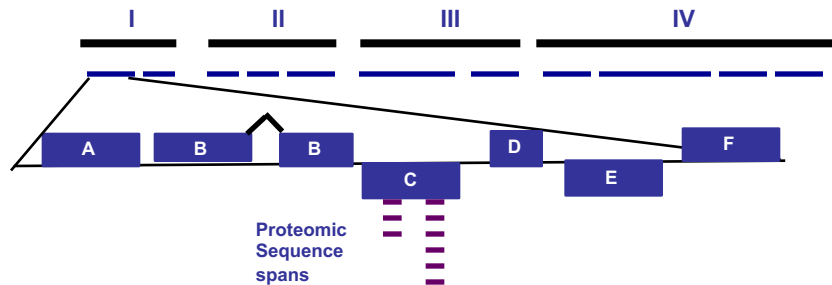
Peptide database

```

ENNPCKLQYDNTNVTHGFGQEQPCETDIVERFSDTEGAQCDDKKIKDNSEGACAPYRRL
HVCVRNLENINDYSKINNKHNLVCLAAKYEGETGRYPQHETNPDKSOLCTVLA
RSPADIGDIIRGKDLRYGGNTKEKKKKKLEENLKTIFGHIYDELKNGKTNGEELQKRY
RGDKDNDFYQLREDWWDANRETWVKAITCNAGSYQSQPTCGRGEIPYVTLKSCQCIAGE
VPTYFDYVQYLRWFEEWAEDFCRKKKKIPNVKTNCRQVQRGKEYCDRDGYNCGTIR
KQYIYRLDTCCKSLACKTFAEWIDNOEQDFKQKQYQNEISGGGGRQRKRSTHSTKE
YEGYKHFNEELRNEGKDVRSFLQLLSKEKICKERIQVGEETANYGNFENESNTFSHTEY
CDRCPLCGVDCSSDNCRKKPKSCDEQITDKEYPPENTTKIPKLTAEKRTGILKKYKFP
CKNSDGNNGGQIKKWECHYEKNDKDDGNDINNCIQGDWTKSKNVYYPISYYSFFYGSII
DMLNESIEWRERLKCINDAKLKRKRGCKNPECYKRWVEKKKDEWDKIEFFPKQKDL
LKDIAAGMDAGELLEFYENIFLEDMKNANGPKVIEKFEILGKENEVQDPLKTKKTID
DFLEKELNEAKNVEKNPDNECPKQKAPGDGAAPSDPPREDITHHDEHSSDEDEEEEEE
EEQQPPAEGTEQGEKSEKVEVVEQOETPOKDEKTVPTTTPTVDVCDTVKALADTGS
NAACSLKYVTGRMWCIAPSMISGKDCVPPPTCEICLYYLKLLDPTOKLLEA
FIKTAQETPLLDIDNENETFLDLPCTLNGEILPEDFRQNEIFQDHD
LFLGRYIGRDLNHRITIVYQVETLFGKIDKQZETGITSKINHEKALQEL
GGKKTLETETYNYSNDRNHLTGTLNEFASRPSFLRWMTWGDQFCRERITQLQLKER
CMWESRNGDRGKDDKKECTEACTYKEWLTWQDNYKKQNRYTEVKGTSYKEDSDVK
ESKYAHGYLRKILKNIICTSGTDIAVCNCEMGSTTDSSNNNDNIPESLKYPIEIEEGCT
CKDPSPEVIEPKVPEPKVLPKPKLPKRQPKERDFPTPALKNAMLSSTIMWSIGGFA
TFTFYLKTKKSTIDLLRVINIPKSDYDIPTKLSPNRYIPYTSKGYRGRYIYLEGDSG
TDSGYTDHSDITSSSESEYEELDINDIYAPRAPKYKTLIEVVLEPSGNNTASGNNTPS
DTQNDIQNDGIPSSKITDNEWNTLKDEFISQYLOSEQPNDVPNDYSSGDIPLNTQPNPLY
FDNPDEKPFITSIHDRDLYSGEYSYVNMVNTNDIPISGKNGTYSGLDINDSLNSNN
          
```

Note: ORFs in addition to predicted Genes must be searched

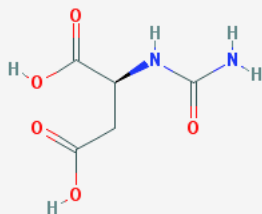
30,000 ft View - Proteomics



Overview

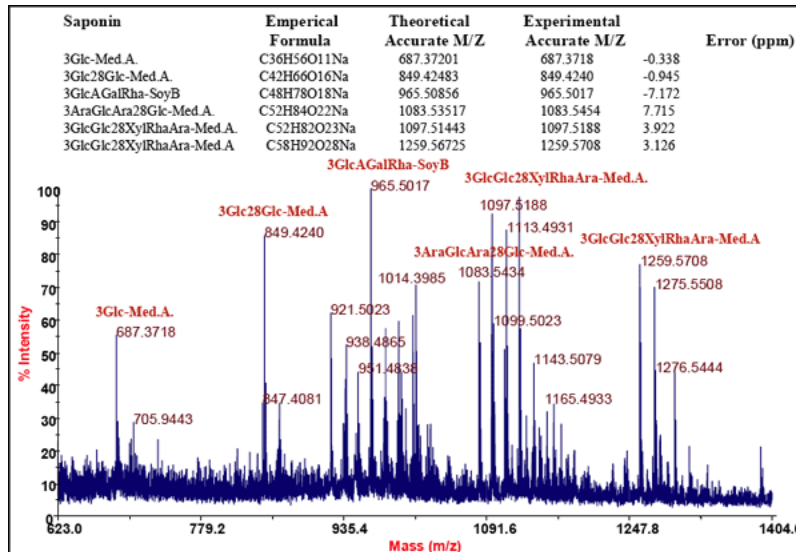
PubChem Compound ID: [CID:93072](#)
PubChem Substance ID(s): 3727
Synonyms: N-Carbamoyl-L-aspartate
Molecular Weight: 176.12742
Molecular Formula: $C_5H_8N_2O_5$

2D Structure

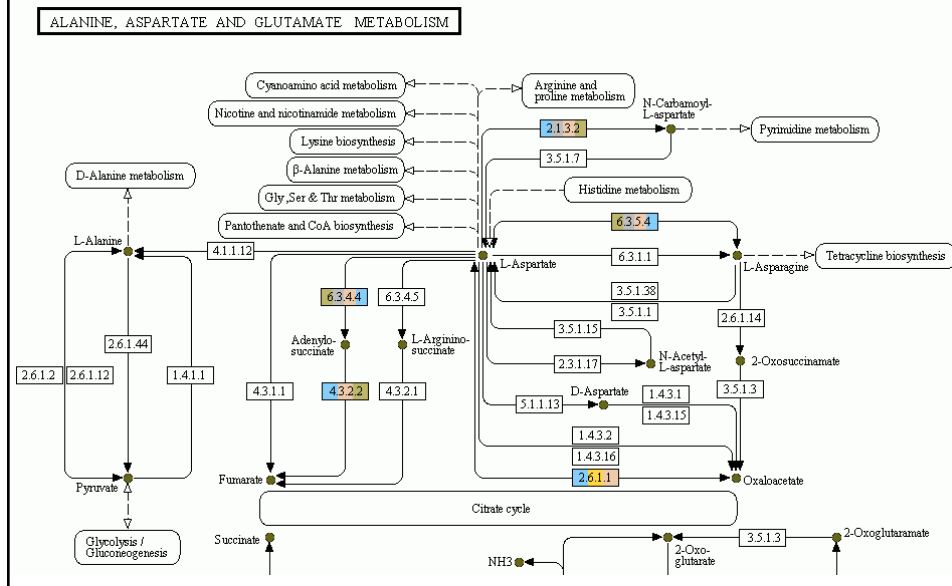


Mass Spectrometry can be used to measure metabolic and other chemical compounds

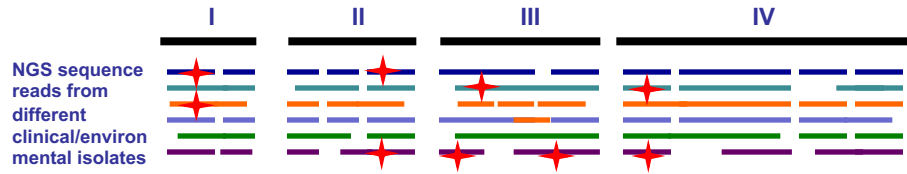
Complex mixtures can be analyzed and interpreted



Metabolites can be linked to metabolic pathways and enzymes



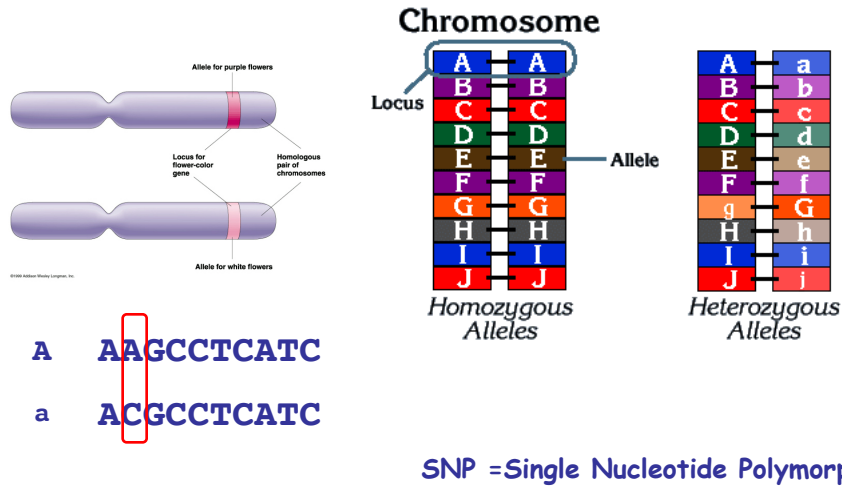
30,000 ft View- NGS SNPs



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 chromosomes (in a diploid)



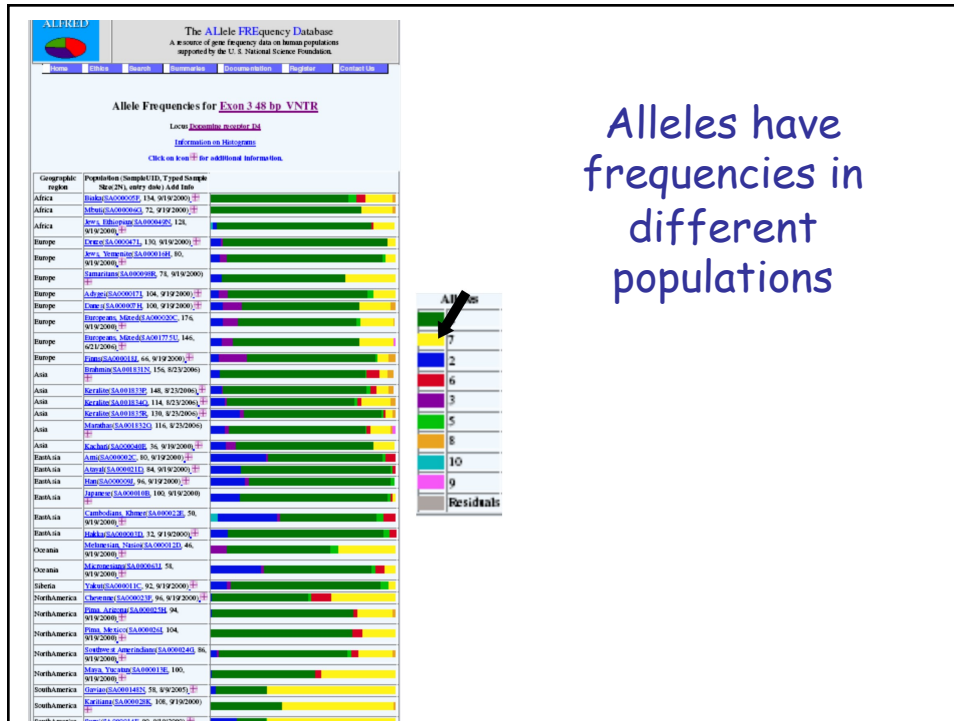
Population data

Data

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

Technology

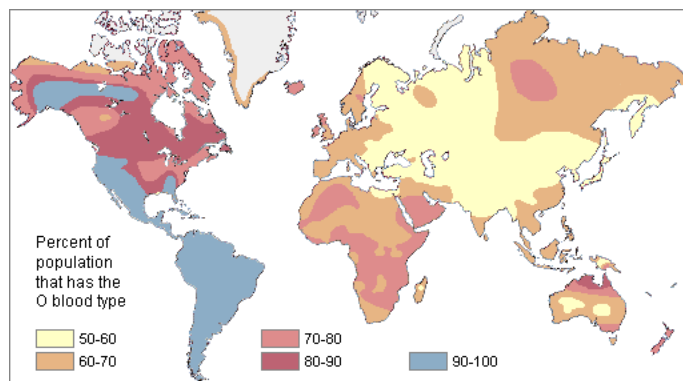
- Chip-Seq
- NGS



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

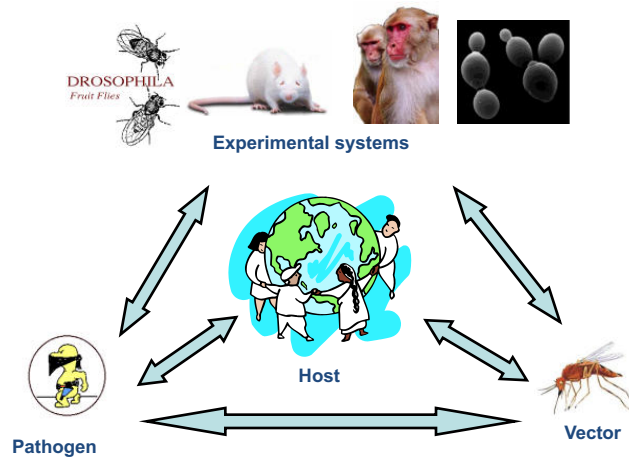
Data

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

Technology

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

Infectious Disease Paradigm



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

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 PDF