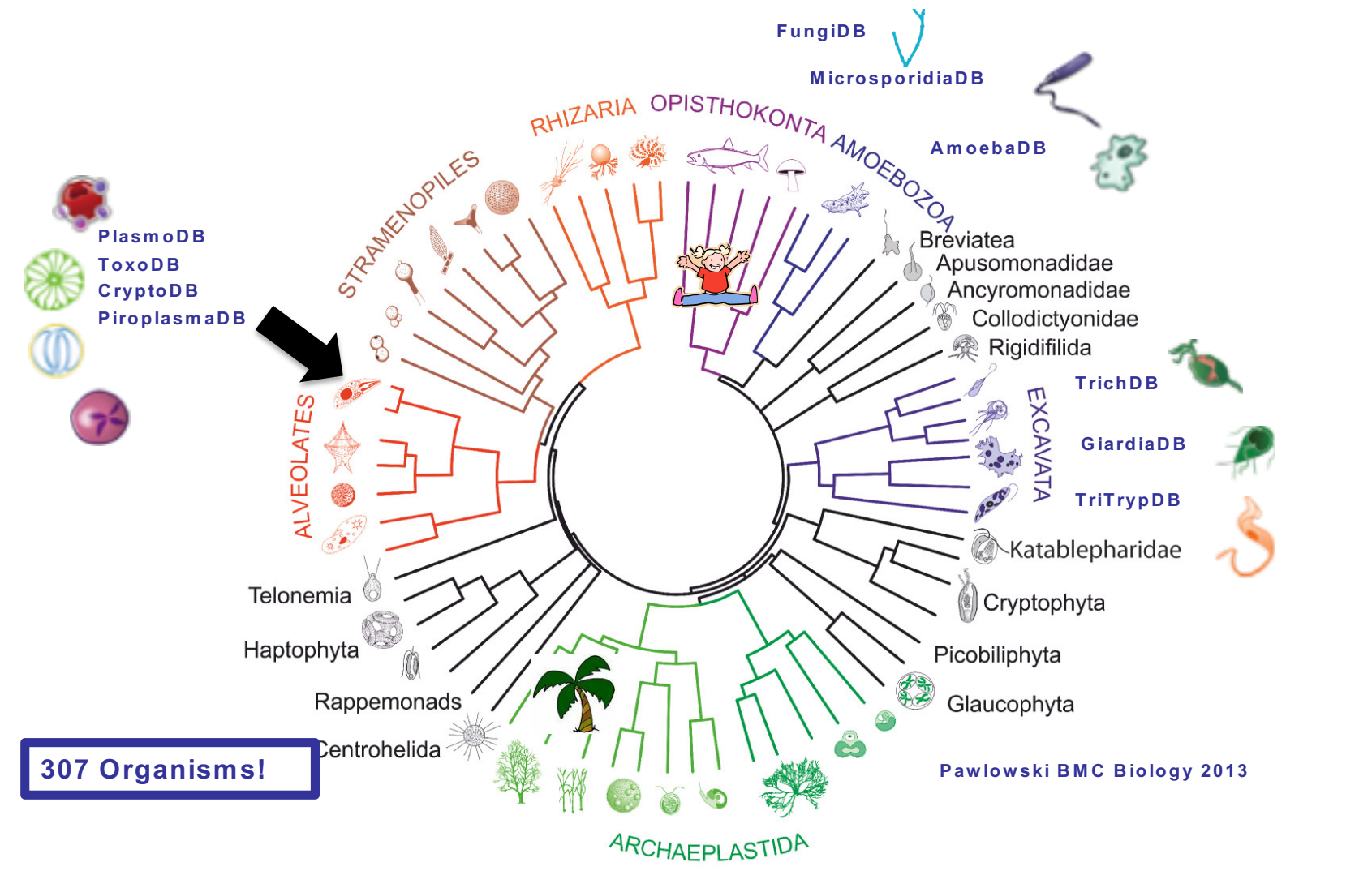


# Crash Course in Omics Terminology, Concepts & Data Types

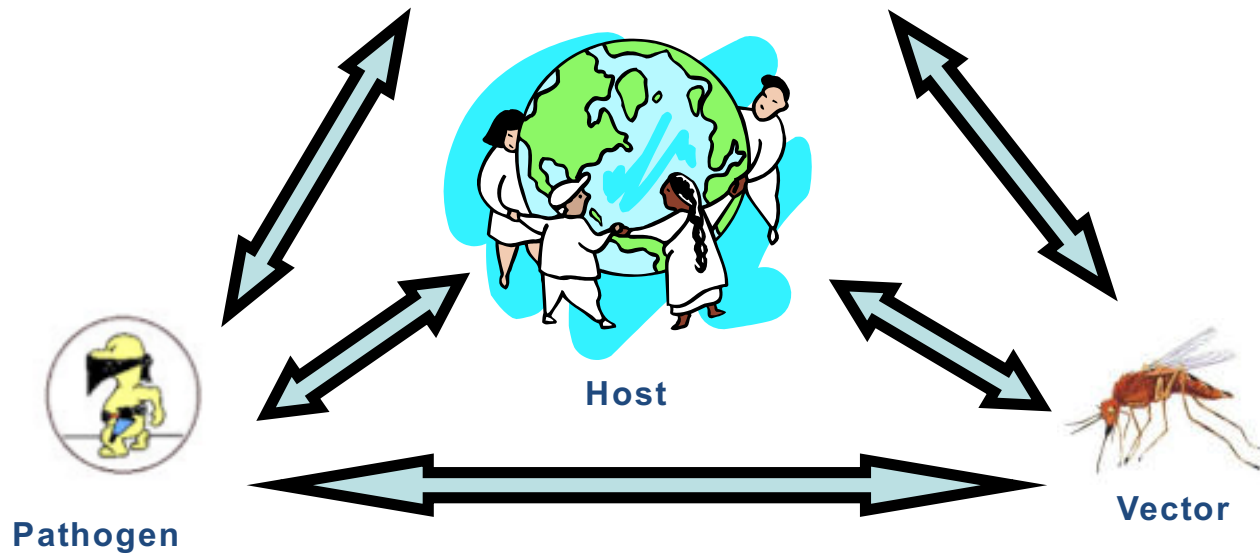
13<sup>th</sup> Annual EuPathDB Workshop

Jessie Kissinger

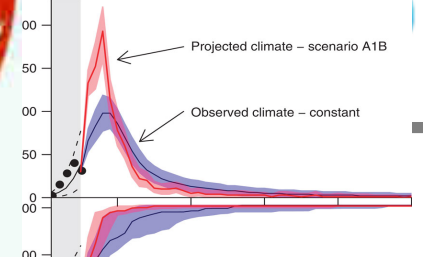
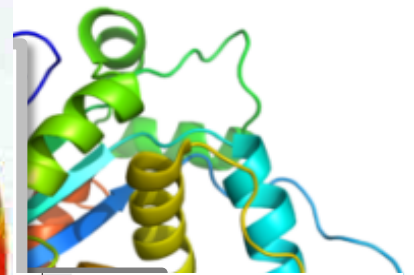
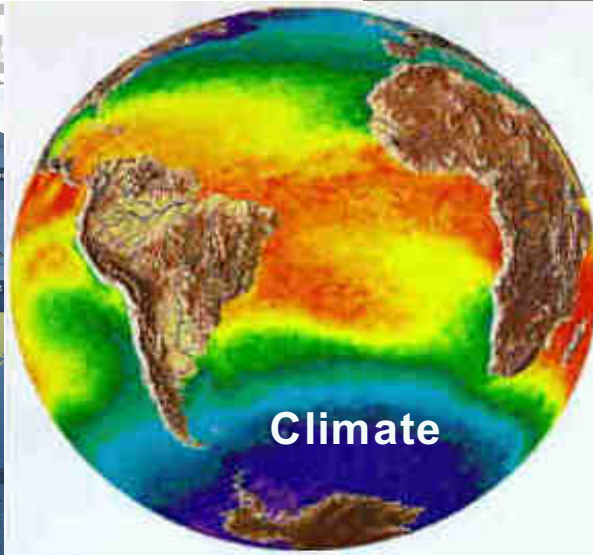
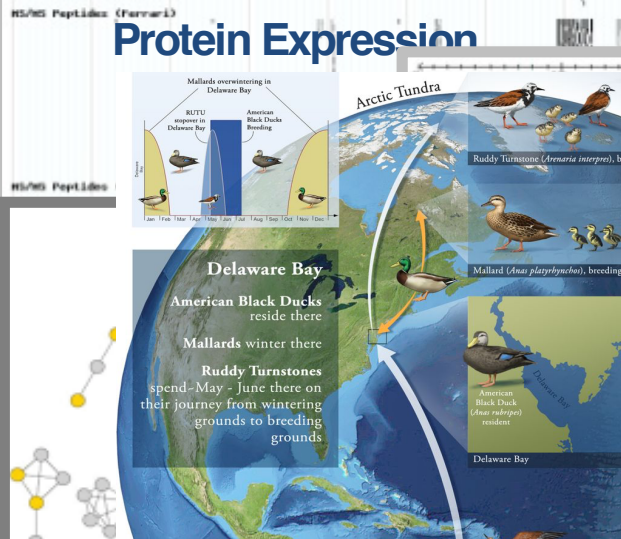
June 17th, 2018



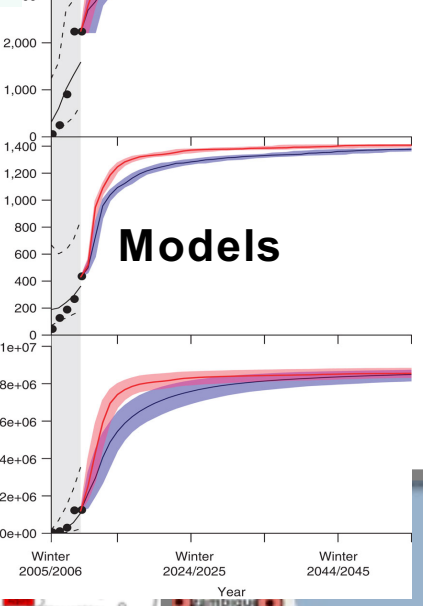
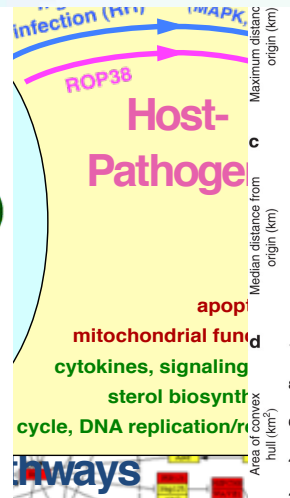
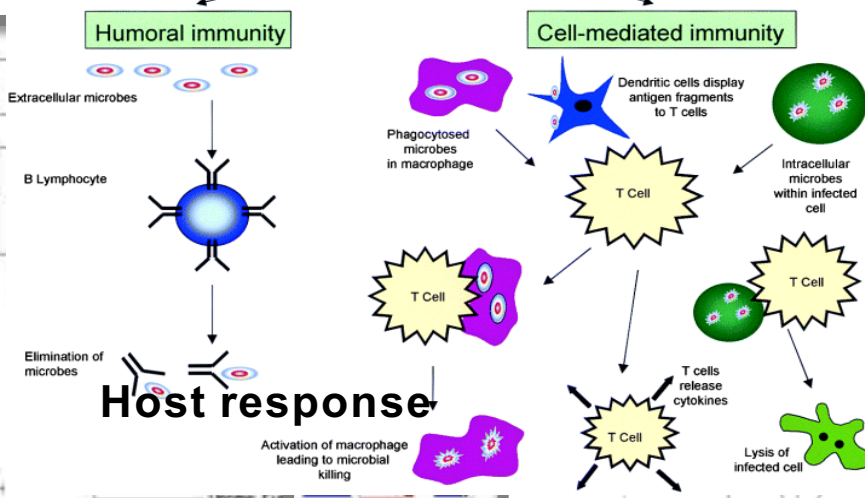
# Infectious Disease Paradigm



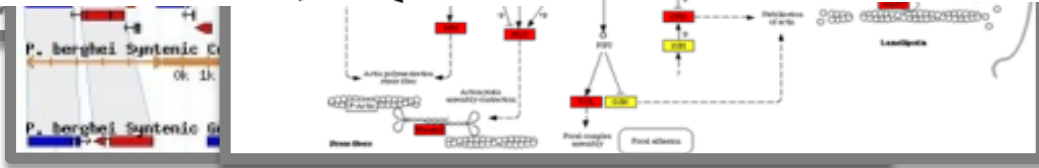
# Protein Expression



## Adaptive Immunity



Modified from slide  
Provided by David Roos

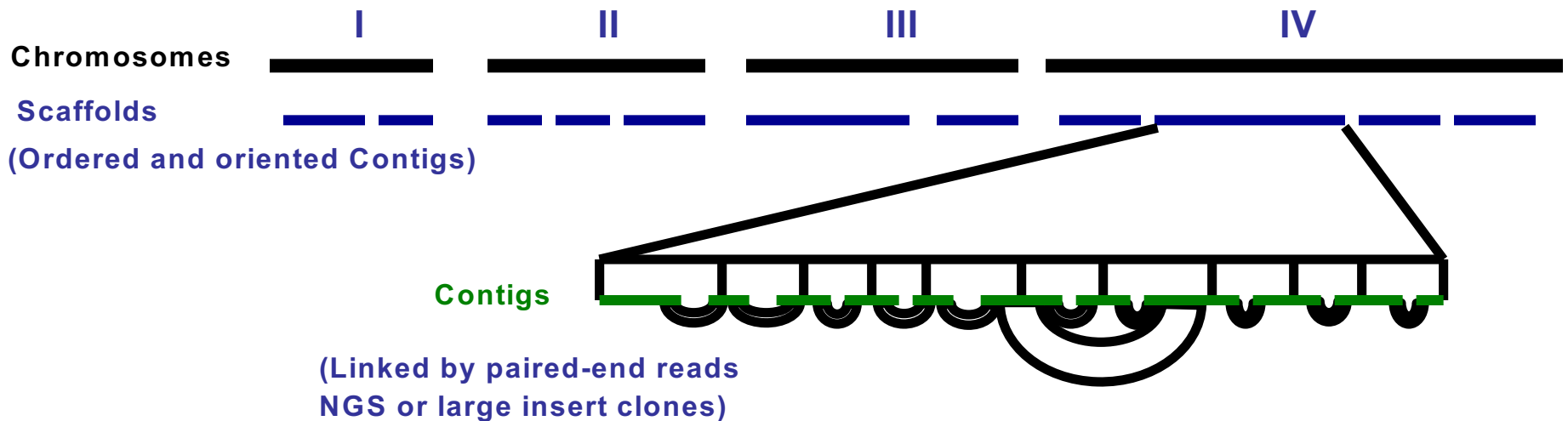


Tele Atlas, MapGIS, Tele Atlas, Europa Technolo...

# Sequence Technologies/platforms

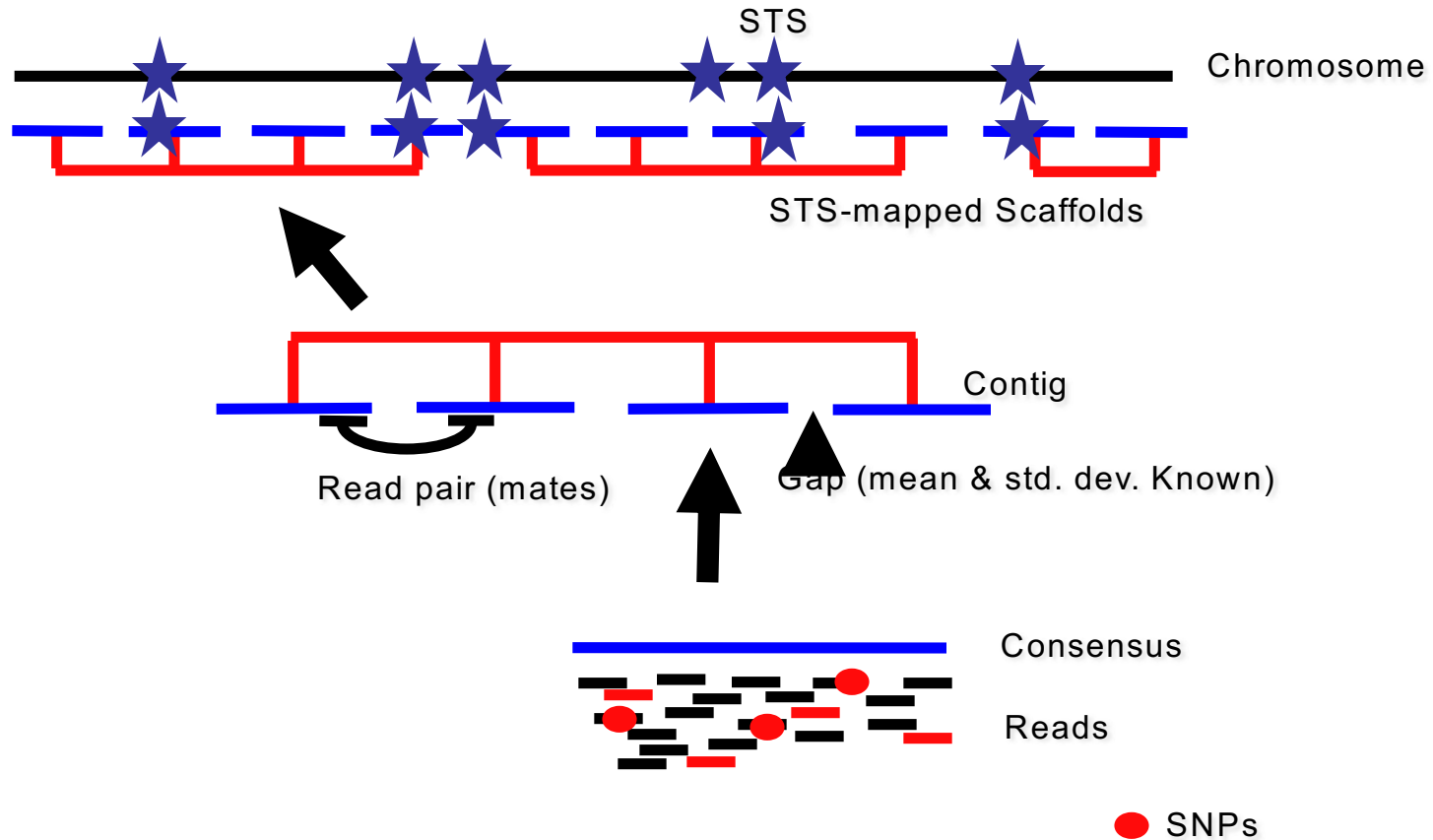
- Hybridization (Southern, Northern, Microarray, chip Arrays (Affymetrix), now.. "bait + Seq"
- Sanger sequencing, 454, Solexa/Illumina, Solid, Ion torrent, PacBio, Nanopore, 10X Genomics Chromium

# 30,000 ft View - Genome Assembly

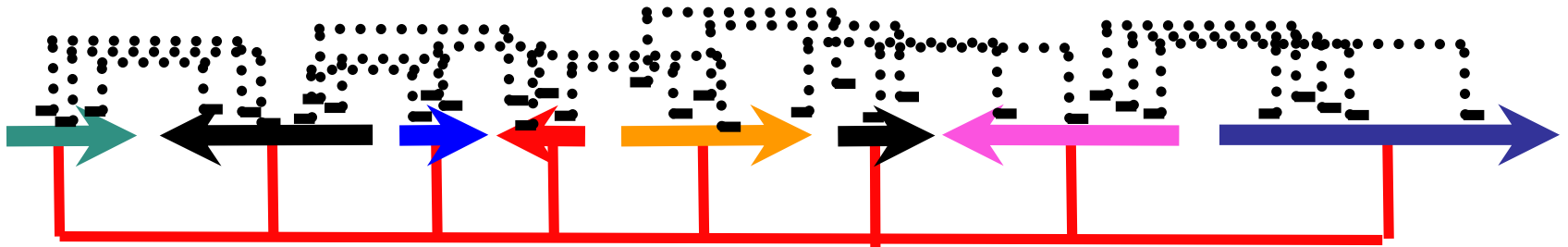


5X genome sequence means that sequences equivalent to 5X the genome size were generated e.g. Genome size = 10 Mbp, then 50Mbp of random sequences were generated

# Anatomy of a WGS Assembly



# Pairs Give Order & Orientation



Scaffold

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



End Reads (Mates)

Mean & Std.Dev.  
is known

550bp

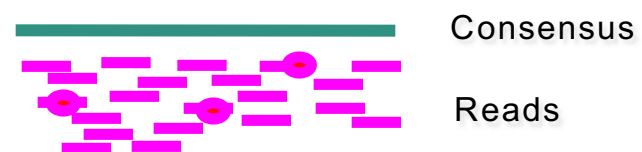
Primer

Plasmid  
Fosmid

NGS

Distance?

SEQUENCE



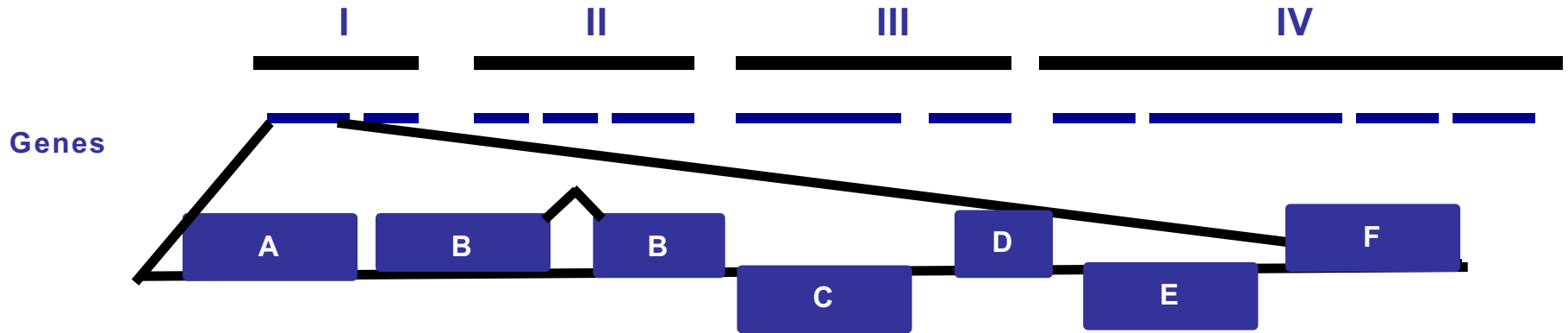
Consensus

Reads

● SNPs



# 30,000 ft View - Annotation



AAGCTTCGCCAGGCTGTAATCCCGTGAGTCGTCTCAACAAATCATCAAGCAGGTGTCTCAGGGAGACTGCCTGACTGAGTTATGCTAATTCCTTTCTACTTTGGCGTGGTCACGTGTA  
ACCATATCCGAATCATTTCTCTAGCCCTACGAACAGGTAAGAGCGCTAGGGATGTCGGTGGAGTAGTGTGCTTACTCGATAATATTCAGTTGGGACTACCAGCGAGGCGCTCGCTTTGCT  
CACGCAATGCCTGAGACAGTTGCAGAATGAATGGTAACCGACAAACGCGTTCATATGCGTTTTCAAACCTTAGTAGACGCGTACTGTCTGAAACTGGCGGTACAGGCACCAGATAACGCC  
CTTGGCATCGGCATGTCTCGTACAGAGGTCCGTATGTAGTGCCACGACTTCTAAATCCGGCGACAGGCTGGTCTTTTGTCTTACCACGTATTAGCCCGGTGCGGATTTCTCGGAGCGCAC  
CTGTTCAACACTAGAAAACGGAGTTTCTGATCGAGAAGCCACCACCTTTCCAGAAAGTTGAACGCTAGCATGTCAATTCGATTTTACCCCCCGGTAGTTCCTGTGTGTCAATTCGTGTGTC  
GAGACAACTCTGTCCCGCCCCGGTGTGTCCATATGCGTGACTTTCCCGCAATTTTTTTCAGACTTTTCAGGAAAGACAGGCTCCGGAACGATCTCGTCCATGACTGGTAAATCCACGACA  
CCGCAATGGCCCCCAGCACCTCTATCTCTCGTGCCAGGGGACTAACGTTGTATGCGTCTGCGTCTTGTCTTTTTGCATTTCGCTTTCCAAAAAAGAGAGCCATCCGTTCCCGCCGACATTC  
AACGCCGAGTGGCGTTTTTGTCTTTTTGAGTGGTAGGACGCTTTTCATGCGCGAACTACGTGGACATTAAGTTCATTCTCTTTTCGACAGCACGAAACCTTGCAATCAAACCCGC  
CCGCGAAGATCCGATCTTGTCTGCTGTTCGAGTCCCAGTAGCGTCTGTGCGGCCGCGCTCTGTGTGGTGGGCGAGCCGCTACACCTGTTATCTGACTGCCGTGCGGAAAATGACGC  
CATTTTTGGGAAAATCGGGGAACCTCATTCTTTAAAAAGTATGCGGAGGTTTTCTTTTTCTTCTGTTCTGTTTTCTTTTTCTCGGGTTTGATAACCGTGTTCGATGTAAGCACTTCCGCTC  
TCCTCCGTGCTTTGTTTCGACATCGAGACCAGGTGTGAGATCCTTCGCTTGTGATCCGGAGACGCGTGTCTCGTAGAACCTTTTCATTTTACCACACGGCAGTGGGAGCACTGCTCTG  
AGTGCAGCAGGGACGGGTGAAGTTTCGCTTTAGTAGTGCCTTCTGCTCTACGGGGCGTTGTGCTGTCTGGGAAGATGCAGAAACCGGTGTGTCTGGTGTGCGGATGACCCCCAAGAGG  
GGCATCGGCATCAACAACGGCCTCCCGTGGCCCCACTTGACCACAGATTTCAAACACTTTTTCTGTTGTGACAAAAACGACGCCGGAAGAAGCCAGTGCCTGAACGGGTGGCTTCCAGG  
AAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTC AACGCCGTTGTATGGGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTG  
GACAGATTAACATCGTTCGTTCCCTCTTCCCTGTGAGCACACAGTAGTGCACACGCTGTTTGGAGCGTCAATCTCCAAGAGTGTGGACGCTGTTCCACGTTCAAATGTTTCC  
CAACATCCGTCGTCAGTAGACACACCAACAAAAAGCACACGGCGAATCTGCTCATCGGAGGGAGGAGCCGGGGGCACACAACATCCTCAACTCTCGAACGAACATATCCGGGGCCG  
GAAGACGTCAGTCTCTCAAATCCAACCCGGAACGCAAAACATTTCTGCATCAAGTCACGATTGCGCCGGTACCTCCATGTGTAAGCAGTTCCATGAAACCTCCGATATTACACAGCACTG  
TGGATATGAATTATATGCAGATGCATATATACTGAGACGCCGATGCAACTATAGTTTCTGCGCCCTCCATGGATATTTTCAGACCTTCTCTCACATTTGGTTTGGCCGTACACCTCCGT  
TAGCTTTTTTTCTGGCTTTCTTCTTCGTCCTGTATTATCAGCAAAGAAGAAGACATTGCGGGGAGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCTCCGAGTCTGTGCTTCACTCCAGC  
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTTCTGTGACACAGATTTTTTGTGCTGGGTATGTTGTCTTAAACTCCTTGGAACTCCATTTCTGGTCCAGAAACGTAAGTAACTGTATA  
CATGTATATACAGATGTATGGATAATATCTAGAGAAGATACAGGGAAGACTGGCAAGGATGAAAGACATGCAGCTTTAACGAAGCAGAGGGCATTTGGCGAGAGGGACCCCGTTATGCT  
GTGTAGTGGCTGTGAATTTTACTTCGCGTTTTGACTTGTCTGAGCGCTTTGCTCAGCTTGTGTTTCTACCTTCCCCAACGCCCTTCTATTTCCCTTCACTCGGAAAGCG  
CGCTCAGTGGGCGCTCACCGAACCCCTTGGTTTTCTGTTTCAGCTGTTGTCTCTTTTTCTCGCTTGTGTTTCTGTTGGCGTGTGCTCGGCTTCTCTCTTTTTCTGTTGGTGGCTCCAG  
ACTATGTGCGCTGTTTCCCCACCCTTCTCGGCTTGTGCTTTCAGGAGGAGCGGGACTGTACGAGGCGAGCGTGTCTCTGGGCGTTCGCTCTCACCTGTACATCACGCGTGTAGCCCGGA  
GTTTCCGTGCGACGTTTTCTTCCCTGCGTTCCCCGGAGATGACATTTCTTCAAACAATCAACGTCTGCGCAGGCTGCAGCTCCTGCGGAGTCTGTGTTGCTTCCCTTTTGTCCGGAGCT  
CGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCAATTTCCAAGACCTTCTCAGACAACGGGGTACCTACGACTTTGTGGTTCGAGAAGAGAAGGAAGACTGACGACGC  
AGCCACTGCGGAACCGGTAAGAGGCAACCGAAGCGCGTAGATAAGAAAAACAACAAAGAGAAGGTGAAAACAGAGAAGGGAAGAAATGCGGAGAAACCGTGGATTTACAAAGATATCAA  
GAGCAATGCTTTGTGGAGATTTTTTTAATTCAGTAGAGACACCCGCGTGCAGGTTGTGTAGAAATAACTGCGACCCCTGGAGACAGAGATCCGCGGATACACCCTTGTGCTTTTCC  
TCCTATGTTTATGACGGGTGCTGAACGCTATCGTACTTAATTTGGAGGAGTGTCTCCGAAGCAGCTTTGGCTGGCCATCCGTGTGTTTGCCTTTGCTGAAAAGCCAGAAGGCGCTCC  
ACAGTGGAGCGATATACAGGGACGCTACCGGAGCCCCGTTTTCTGCTTTTGTGACTCTTGCAGAGCAACGCAATGAGCTCCTTGCAGTCCACGAGGGAGACAACCTCCCGTGCACGGGT  
TGCAGGCTCCTTCTCGGCCGACGCAATTGCCCCGGTGTGGCGTGGATGGACGAAGAAGACCGGAAAAACCGGAGCAAAAGGAACGATTGGGCGGTTCCGATGTTCACTTTAGAG  
GCCATGAAGAATCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATGACCGAACGGGTAACGGCGACTGCGAGAAAAGCCACACCGTTTTCTCTGTGAT  
TCTGTCCGCAAGCCCTCTTTTGTTCATCCACCTTTGCTATTCTCCGCGCCCTTCTTTTTCTGCTCCATGTTCAATTCGTTTCGCTTCTTTCAGTCTTTCCATCTTCCCTGTTACCTCTG  
TCATTCGTTTTCTTGCCTCTATTTAACTGTGTTCTACTCACAGTCTGCATTCGCGGATAGACGAGCTTCCACGCTTTCGCTCTCGACAAGCAACTGTCATTTGTACGCGCTCCCTCCAC  
CGTGAATCGGATTTGCGTTCCGGGTTCTGGGTGAGAAAAGGCTGCGCCAGTATTCTGAATAATACCTTTCGCAATTGTAAGAGGGCAAGGAACAAGAGATATTTCCGGCGCATCT  
TTTTGTGCGGCGCTTTCTCGTGTTCACACCGATGCCCTTCTGTGCATGTCTTCTGCTTCTCTCTTTTTTCCCTGTTTAGGCGTTGGTGTATCTCCAAATTCGGCTGCAC  
TATGCGCTACTCGTGGATCAGGCTTTCCACTTCTCACCAAAAGCGTGTGTTCTGGAAAGGGTAAGGGCGCTTTCAGTGAATGCATATATTTGACTTCAGACATTTCTTAACTGTTTGA  
CAACCAACGTACAAATTTGTTTGTCCGTGTGCGTGTTCGACATGTCAAGTATGTGAAGAGTGCCTACTGTAGACTAACGCACGAACCAGATTTGTTTATCTGCATGCGCTGTGCACCCGT  
TTCGAGTGTCTGGAGTTTCCGCAACCTTCTTTGAATTTCTGGTTCGTTTTTTATGCGCGCACTGGTTTTGCATGTGGCTGAGAGAGCACAGATCGAAGGTGGGGTGTATGGCGTC  
GCTGCAGAGAACTCCGGCGAAGGCAGACATAAAGGAGAGTGGAAATCATTGAACAGTGTGGTTCGTTGTTTTCGACAGGTTCCCGAAGAGTTGCTGTGTTTTCATTCCGGCGGACA  
CGAACGCAAACCATCTTTCTGAGAAGGGCGTGAAGGCAAGTCTACGTTGTACCTCTTGTCTCTGCCGAAGCTCAGATGTCTCCACGGCGTTGGTTTCTTTTTGCTTTTTCGTTGGCA  
TTACCATCGAGTACCACCTCATAGTTGCGTGTGTCTACATGTTTTCTAGAACGTCGGTGTGTTGCTCGTGGCGACCGGCGGAGTGTATGTACCCTGCGCTGTGAGAAGTTGATCCTT

# Six Frame Translation ORF-finding

```
1/1                               31/11                               61/21
M Y A L L I L Y Y I 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 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 GCG ACA TAT ATA TTC CCG CCT CGC GGG GAA AAT
CGA TCG AGC TCG ATC TTG CGT AGC TGA ATC GTA CTG ATA TAT TAA AGT CGC TGT ATA TAT AAG GGC GGA GCG 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
```

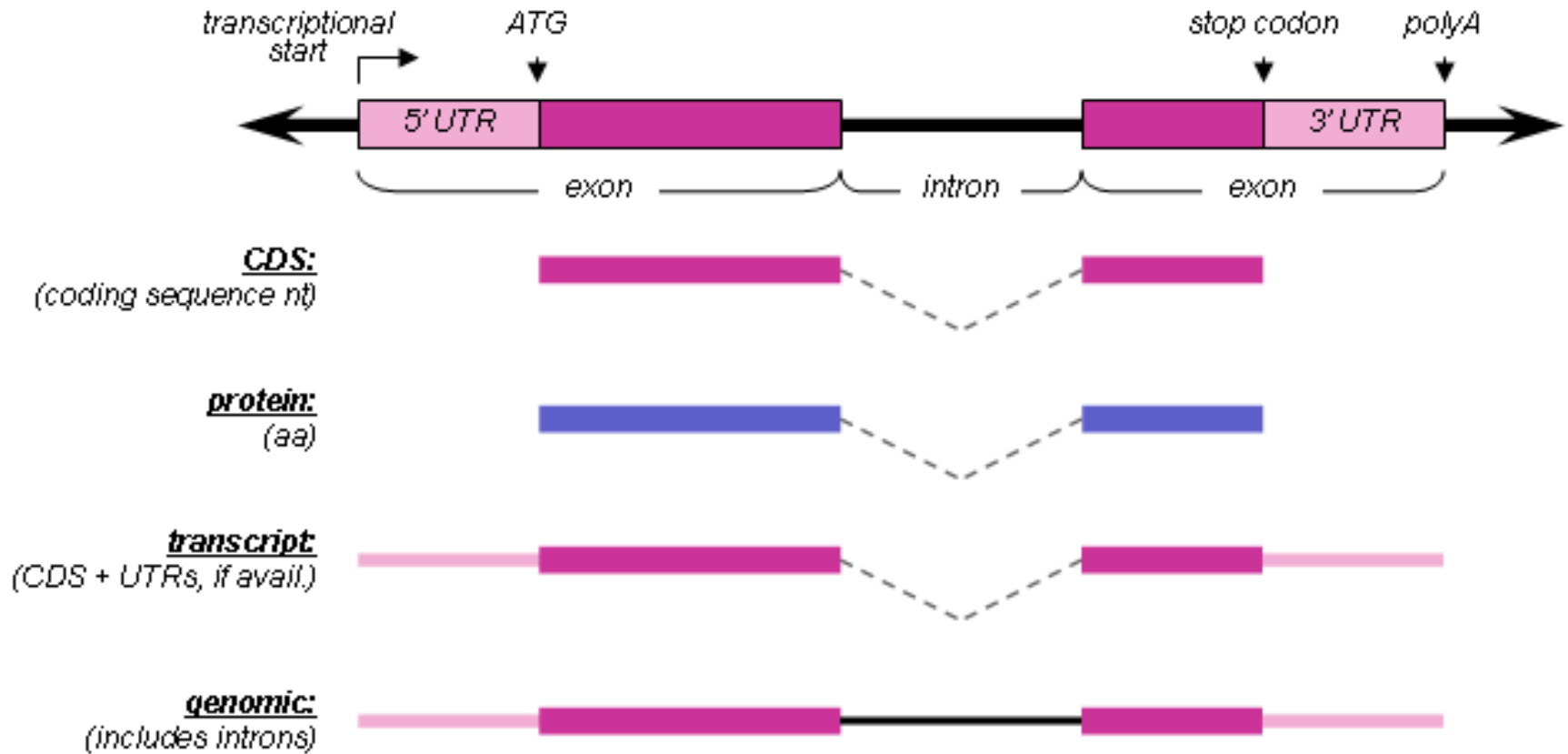
**ORFs ≠ Genes**

ATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC  
ACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAACGACGCCCGAAGAAGCCAGTCGCCCTGAACGGGTGGCT  
TCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTATG  
GGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCTCTTCCC  
TCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC  
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTGACAGATTTTTGTGCGTGGGAGGAGCGGACTGTACGAG  
GCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGGAGTTCCGTGCGACGTTTTCTTCC  
CTGCGTTCCCCGGAGATGACATCTTTCAAACAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTTCTG  
TCCCTTTTGTCCGAGCTCGGAAGAGAGAAGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCA  
GACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGACGCCACTGCGGAACCGAGCA  
ACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACCTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGAGCCAT  
TGCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAACGCGAGCAAAGGAACTGATTCCGGCCGTTCCGCAT  
GTTCACTTTAGAGGCCATGAAGAATCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATG  
ACCGAACGG

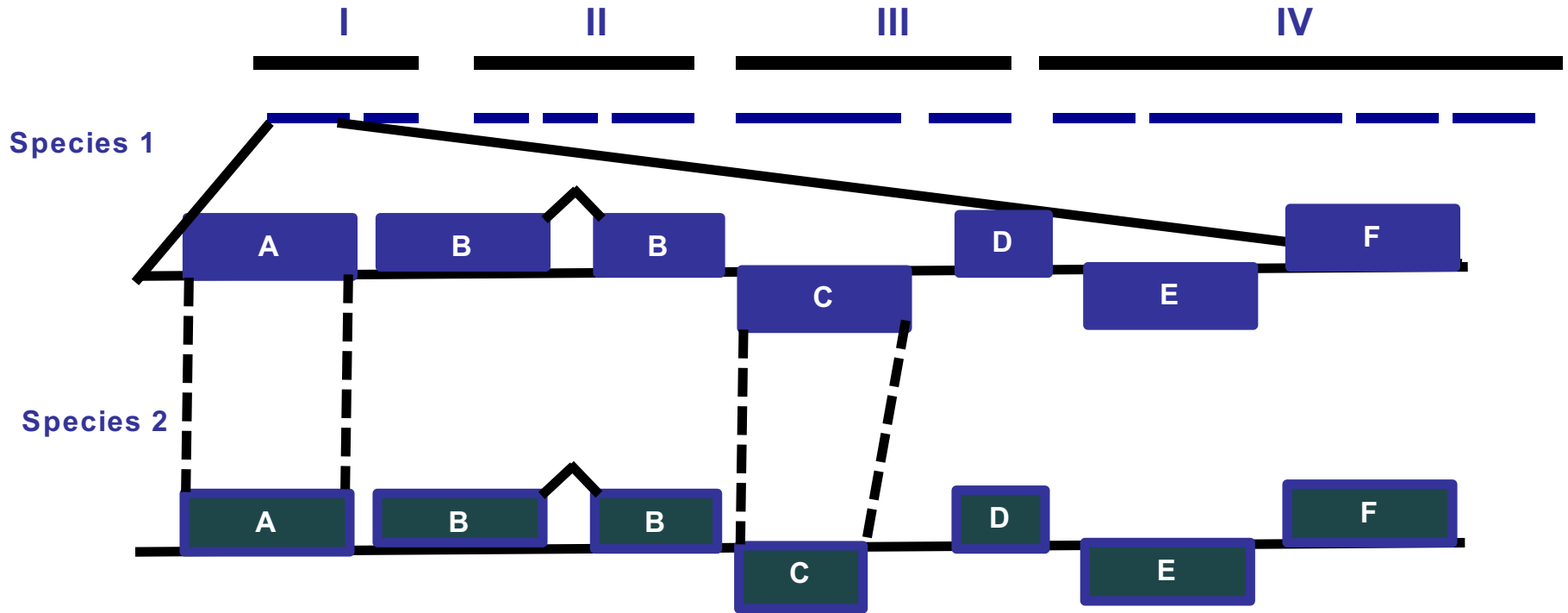
>Translation Frame 1

MQKPVCLVVAMTPKRGIGINNGLPWPHLTTDFKHFSRVTKTTPEEASRLN  
GWLPRKFAKTGDSGLPSVSGKRFNAVVMGRKTWESMPRKFRPLVDRJNI  
VVSSSLKEEDIAAEKPOAEGQQRVRCASLPAALSLEEEYKDSVDQIFV  
VGGAGLYEAALSLGVASHLYITRVAREFPDVFVPAFPGDDILSNKSTAA  
QAAAPAESVFPFCPELGREKDNEATYRPIFISKTFSDNQVDFVLEK  
RRKTDDAATAEPSNAMSSLTSTRETTVHGLQAPSSAAAIAPVLAWMDEE  
DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADI INNGRTMDDRT

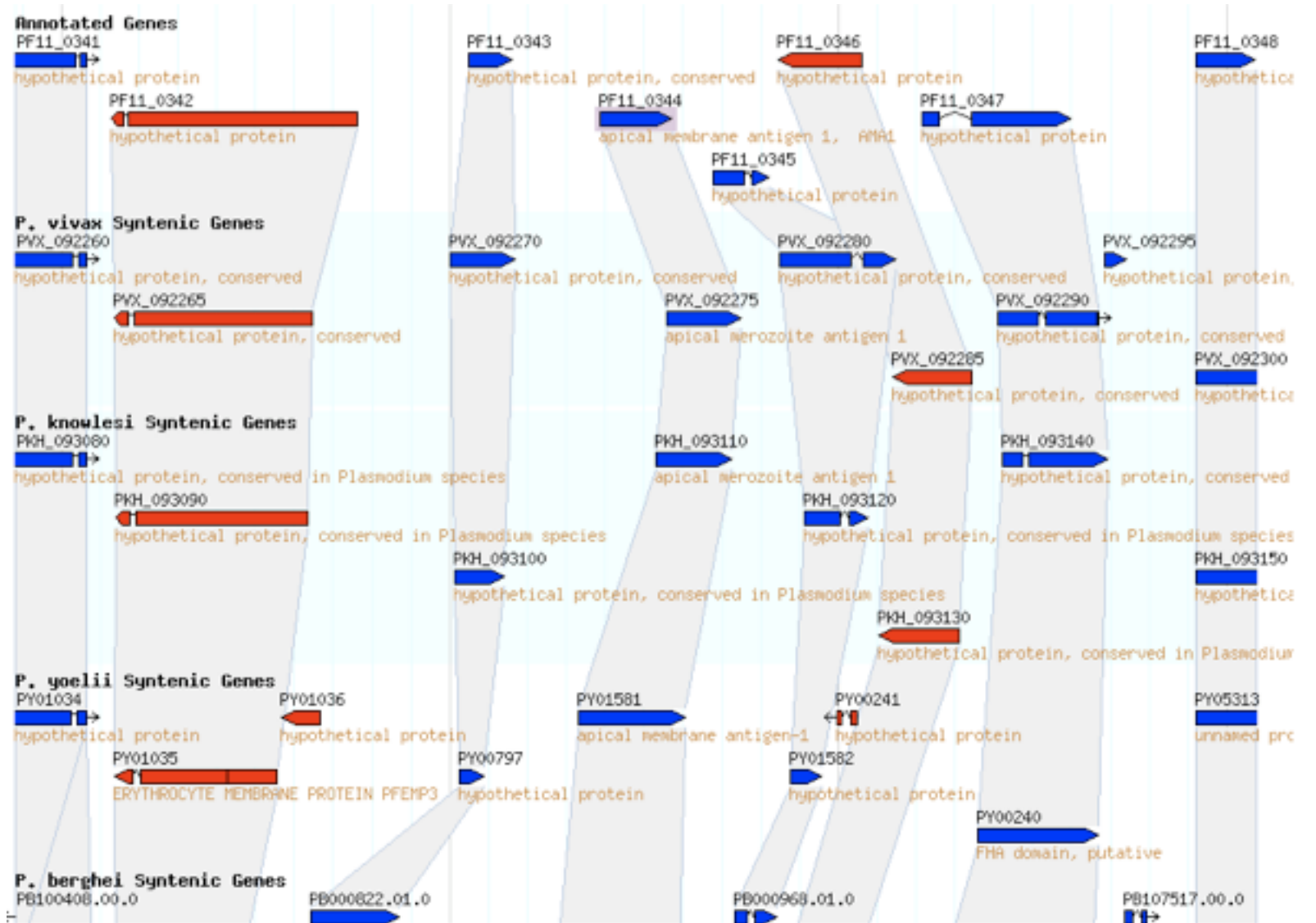
# Terminology



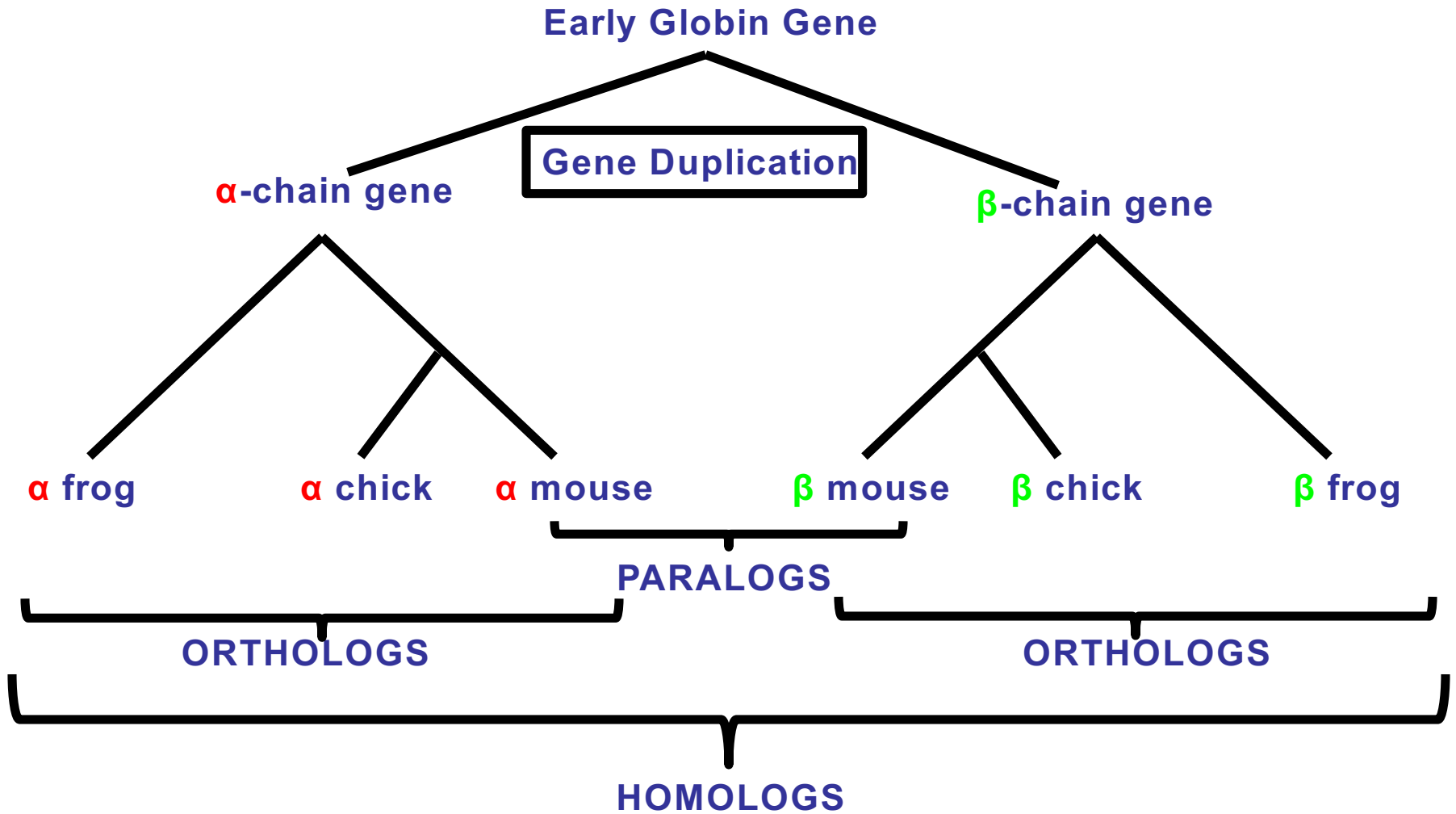
# 30,000 ft View - Synteny



# 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.

# For Example:

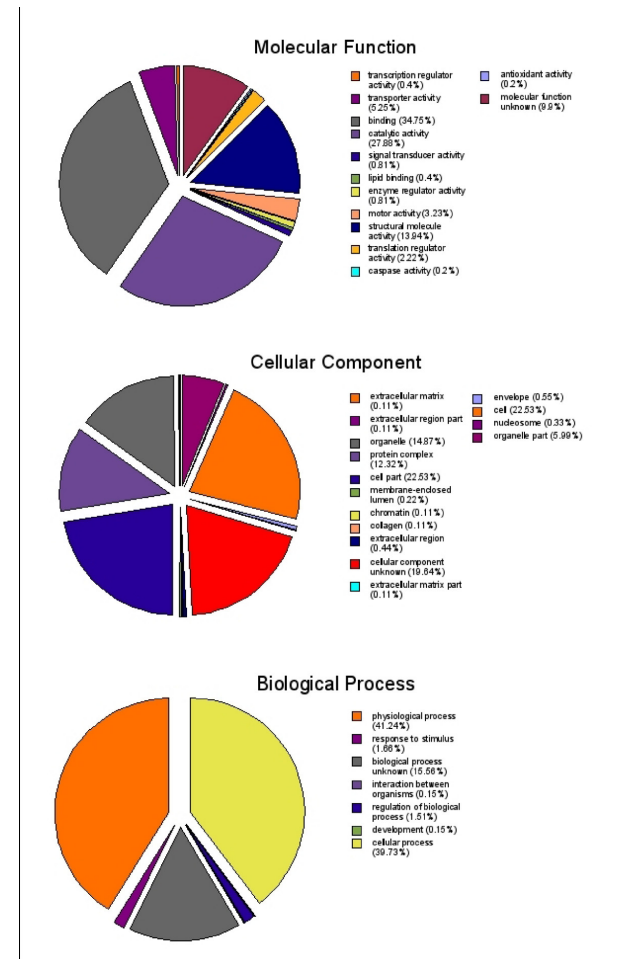
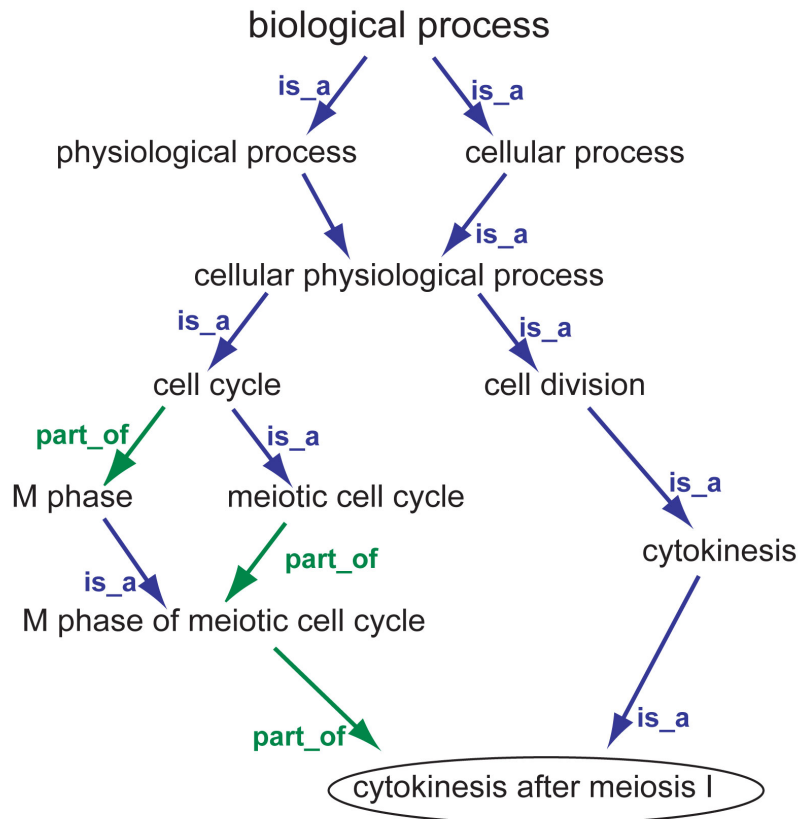
*D. melanogaster* gene CG3340 annotated as:  
"Kruppel" and *P. falciparum* gene  
PF3D7\_1209300 annotated a "putative  
KROX1"

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



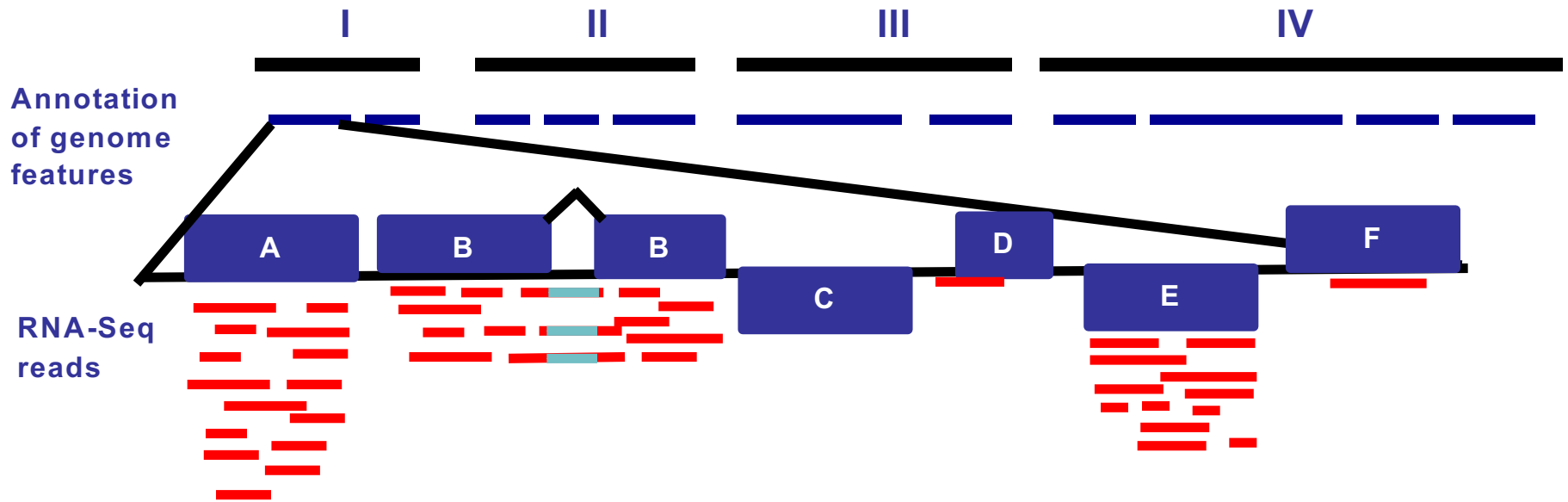
# Expression Profiles (RNA and Protein)

- 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

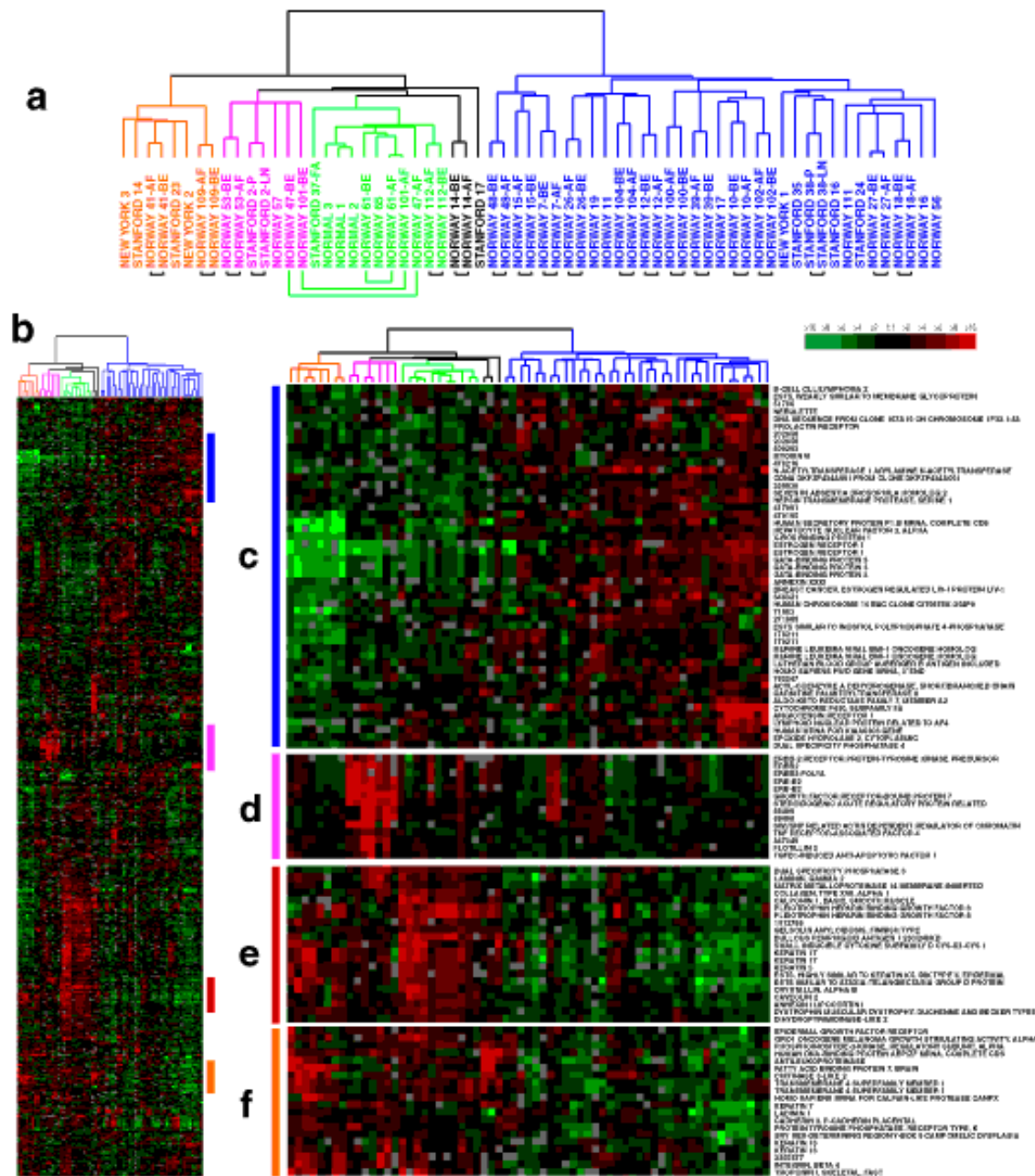
# 30,000 ft View - RNA-Seq



FPKM = Fragments per kilobase of exon per million fragments mapped

Figure 2

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

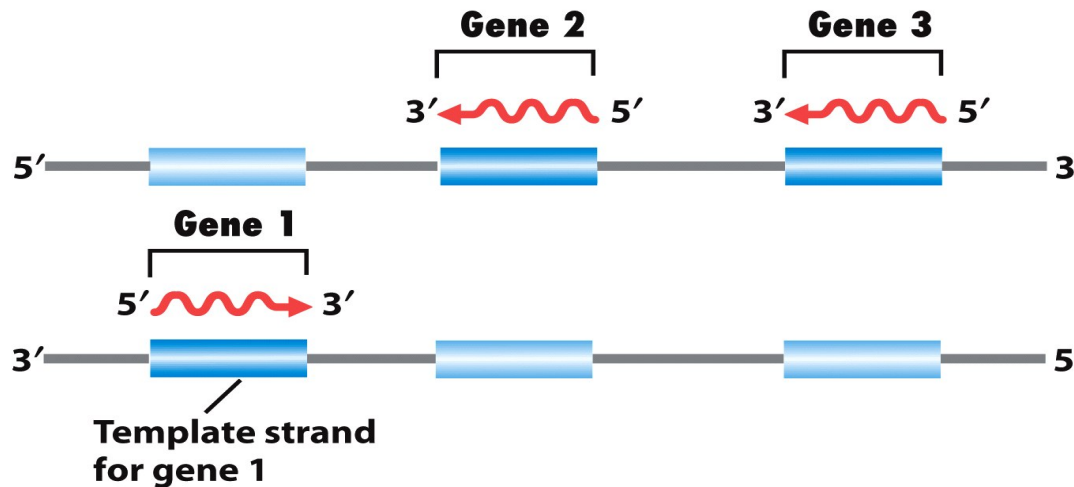


Figure 8-3  
 Introduction to Genetic Analysis, Ninth Edition  
 © 2008 W. H. Freeman and Company

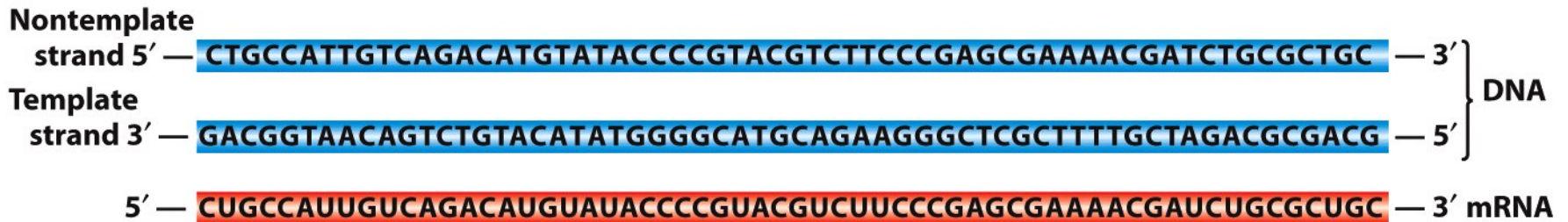


Figure 8-6  
 Introduction to Genetic Analysis, Ninth Edition  
 © 2008 W. H. Freeman and Company



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

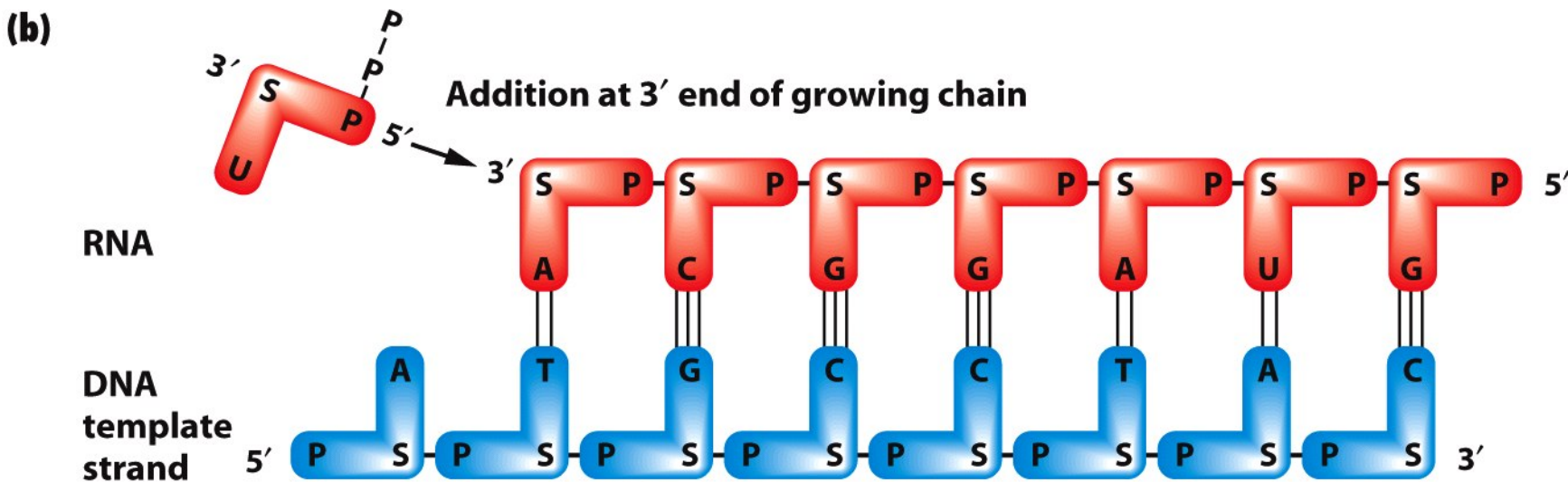
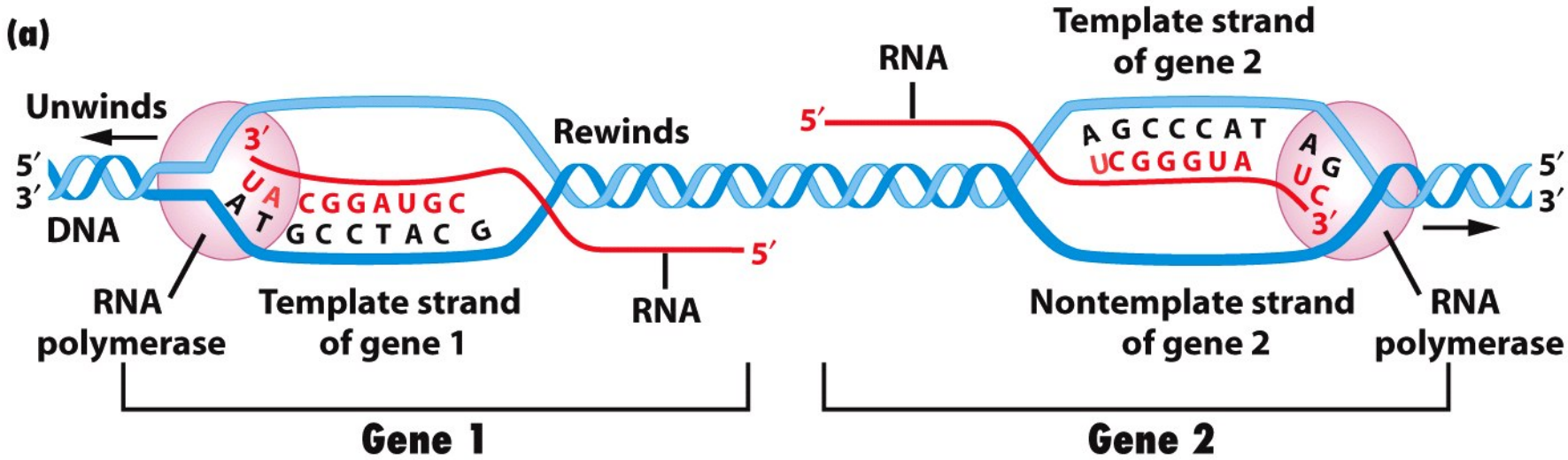


Figure 8-4  
*Introduction to Genetic Analysis, Ninth Edition*  
 © 2008 W. H. Freeman and Company

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

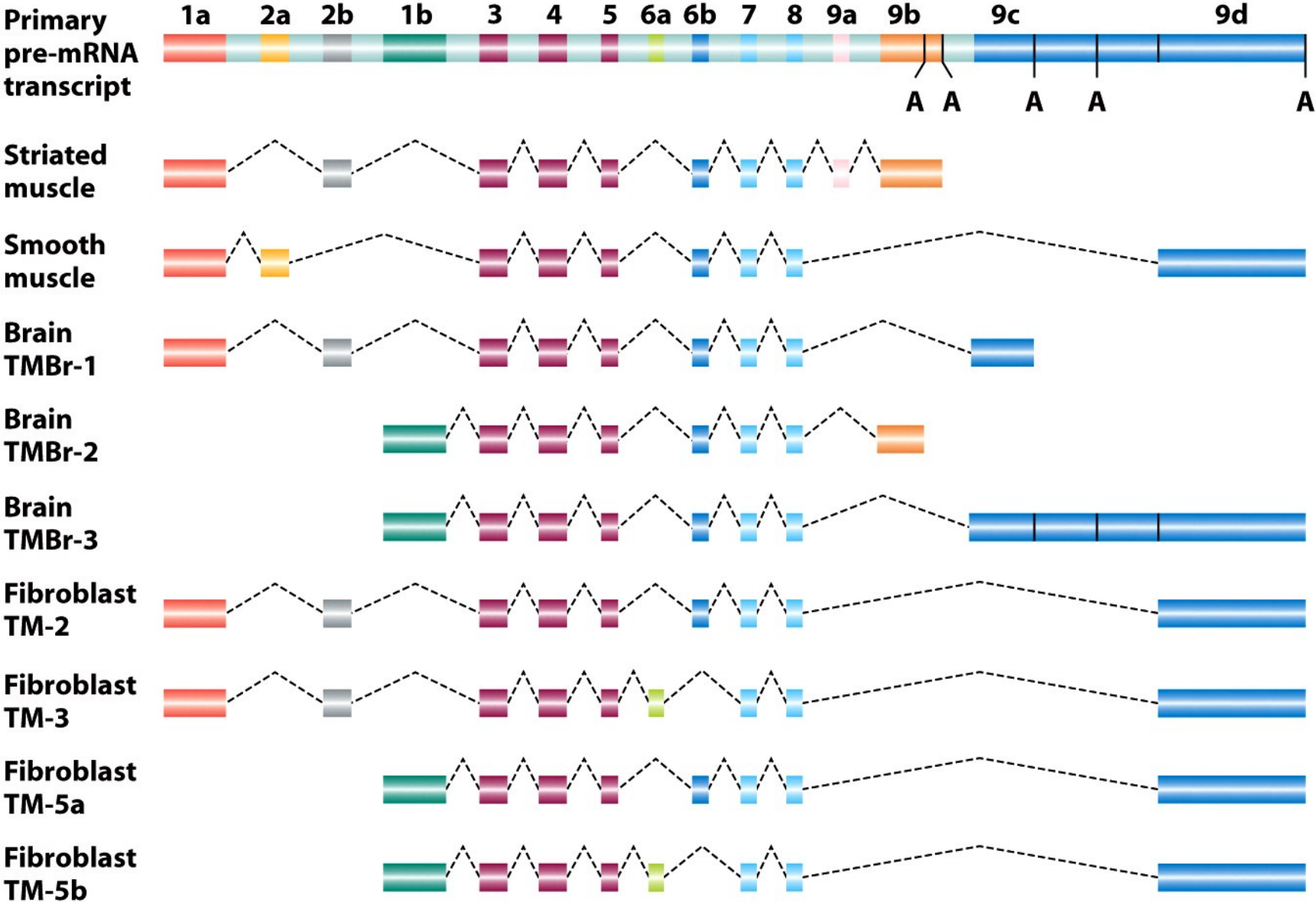
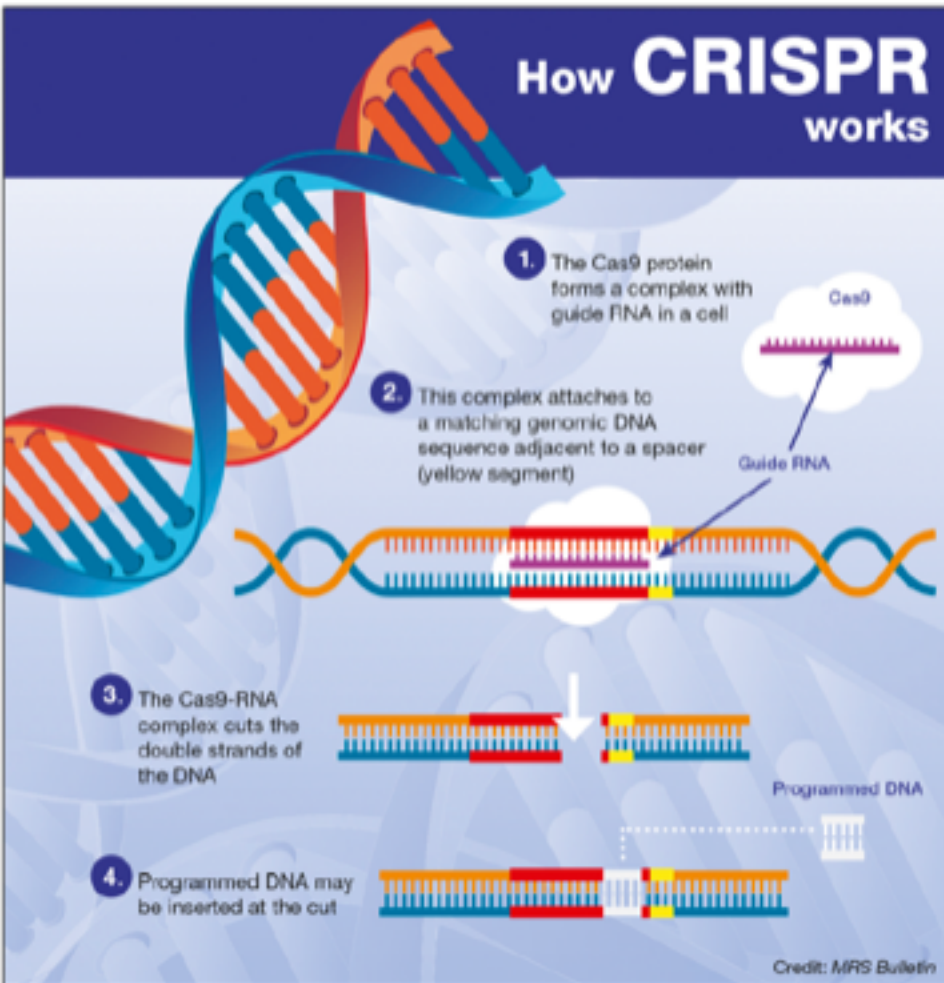


Figure 8-14  
*Introduction to Genetic Analysis, Ninth Edition*  
 © 2008 W. H. Freeman and Company

# CRISPR-CAS



- 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

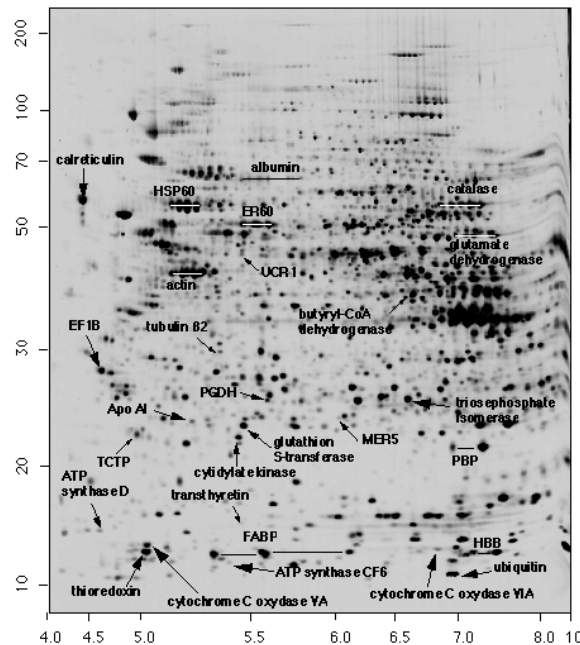
# Protein Expression/Sequence

## Data

- MW-Isoelectric point
- MW
- Sequence/spans

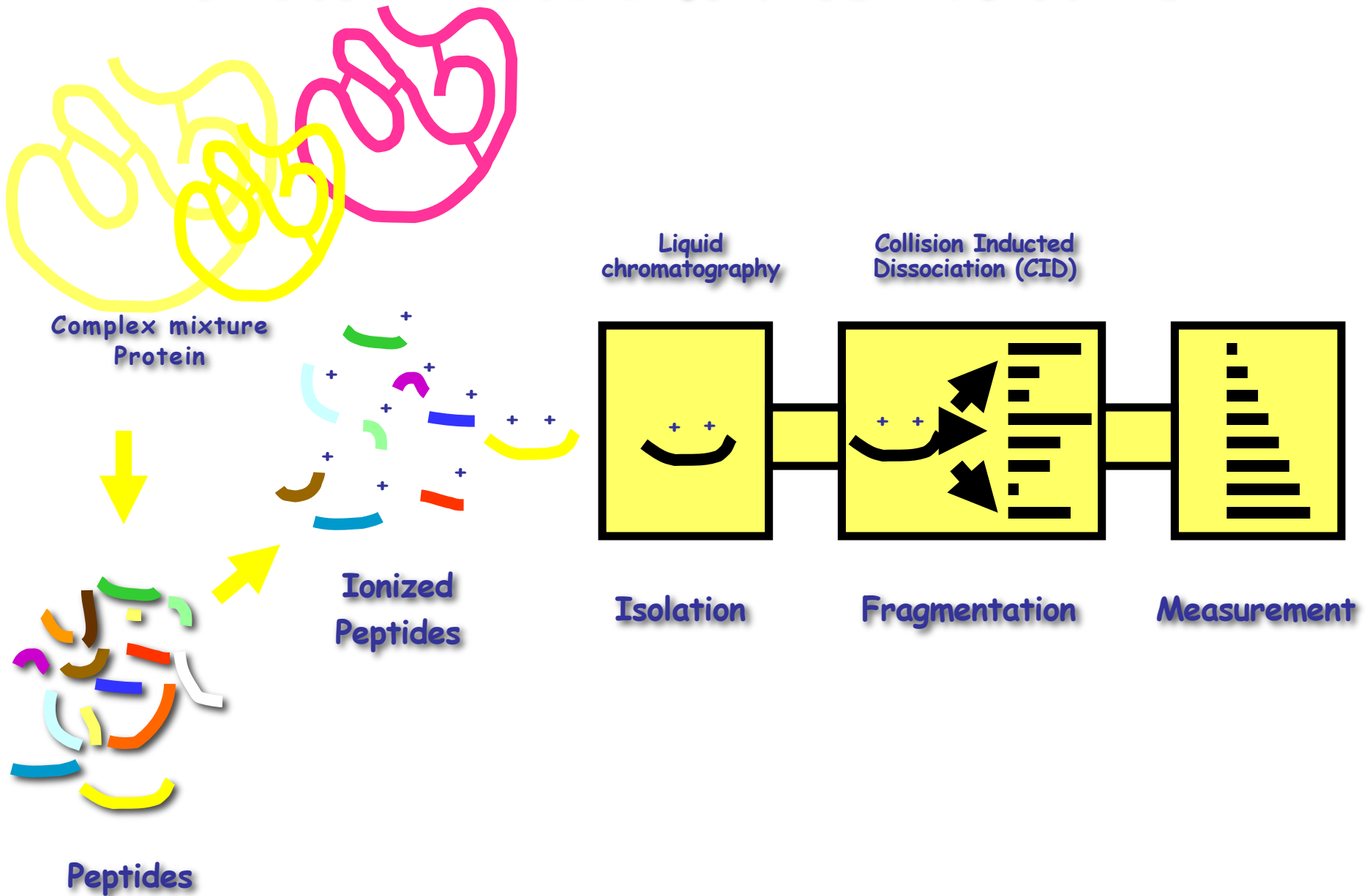
## Technology

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



Typical 2 D gel

# How Tandem MS Works



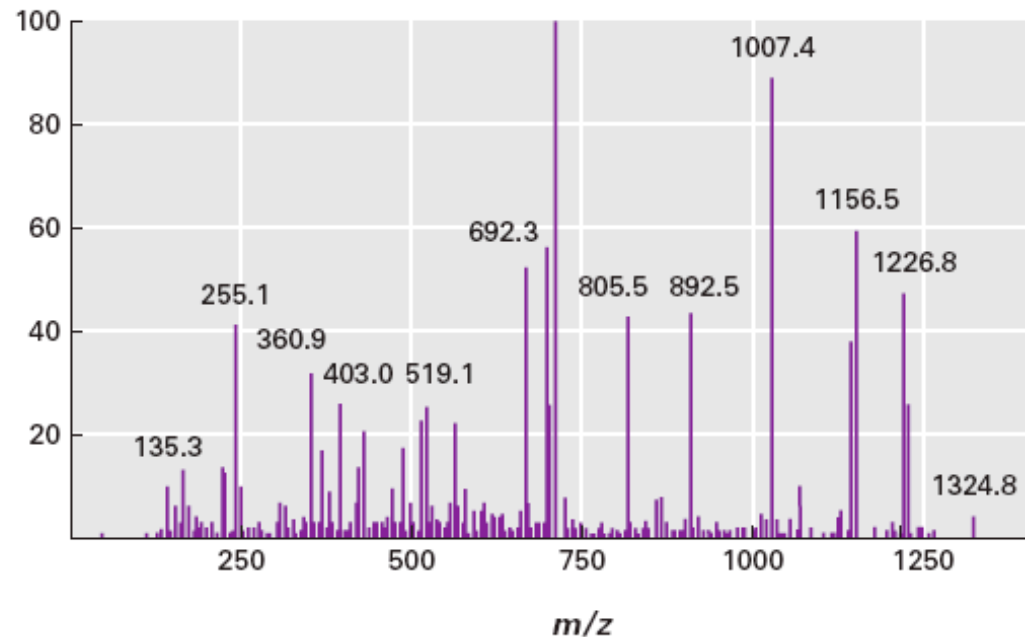
# Tandem MS protein data

a)

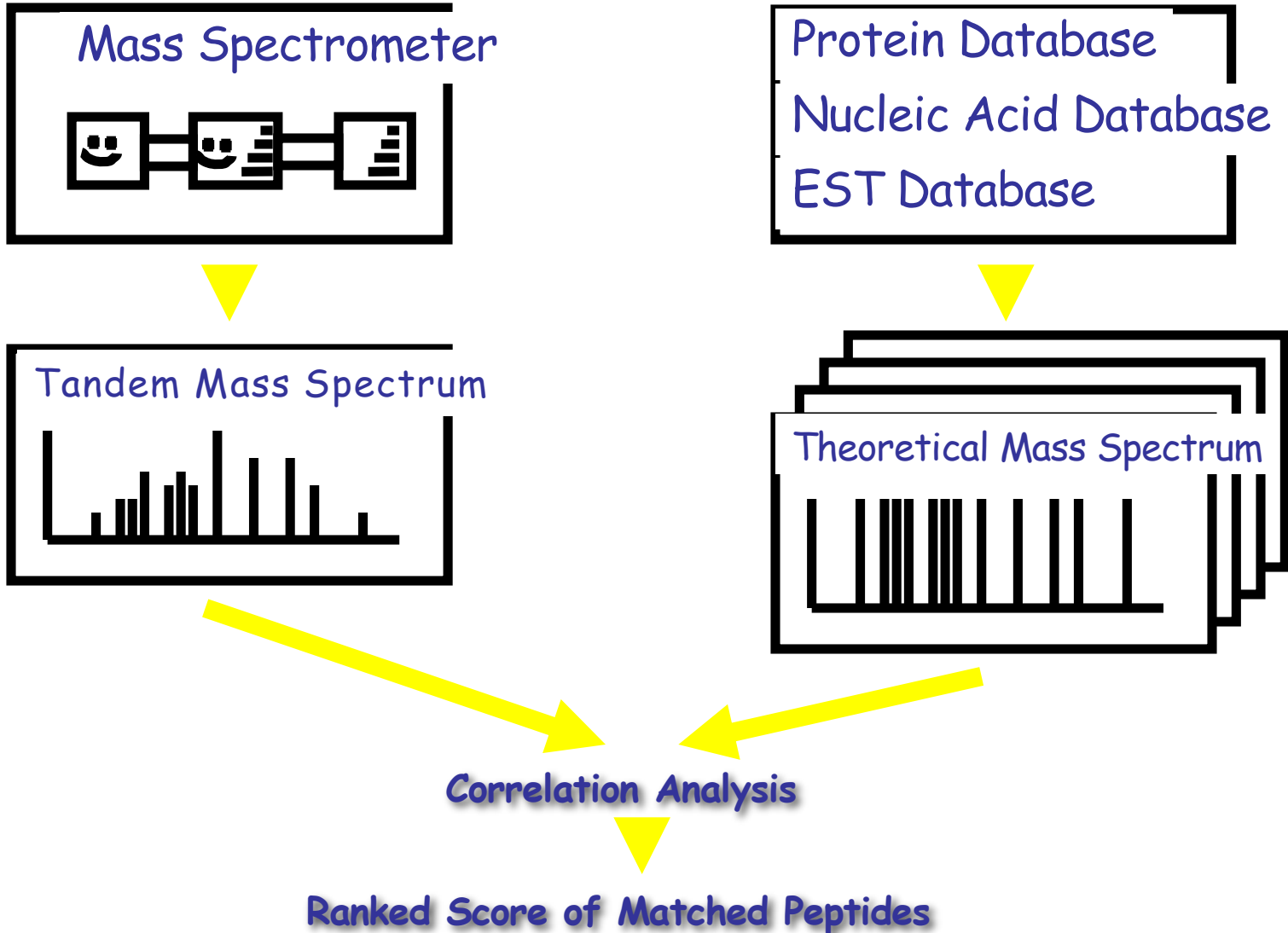
**S-P-A-F-D-S-I-M-A-E-T-L-K**  
(protonated mass 1410.6)

Mass <sup>+</sup>	b-ions	y-ions	Mass <sup>+</sup>
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)



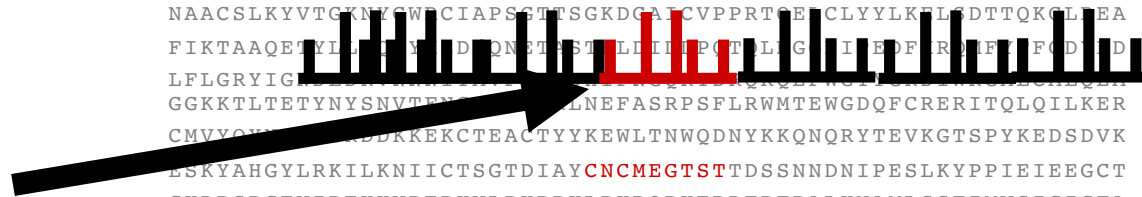
# Sequest Database Search



# Peptide database



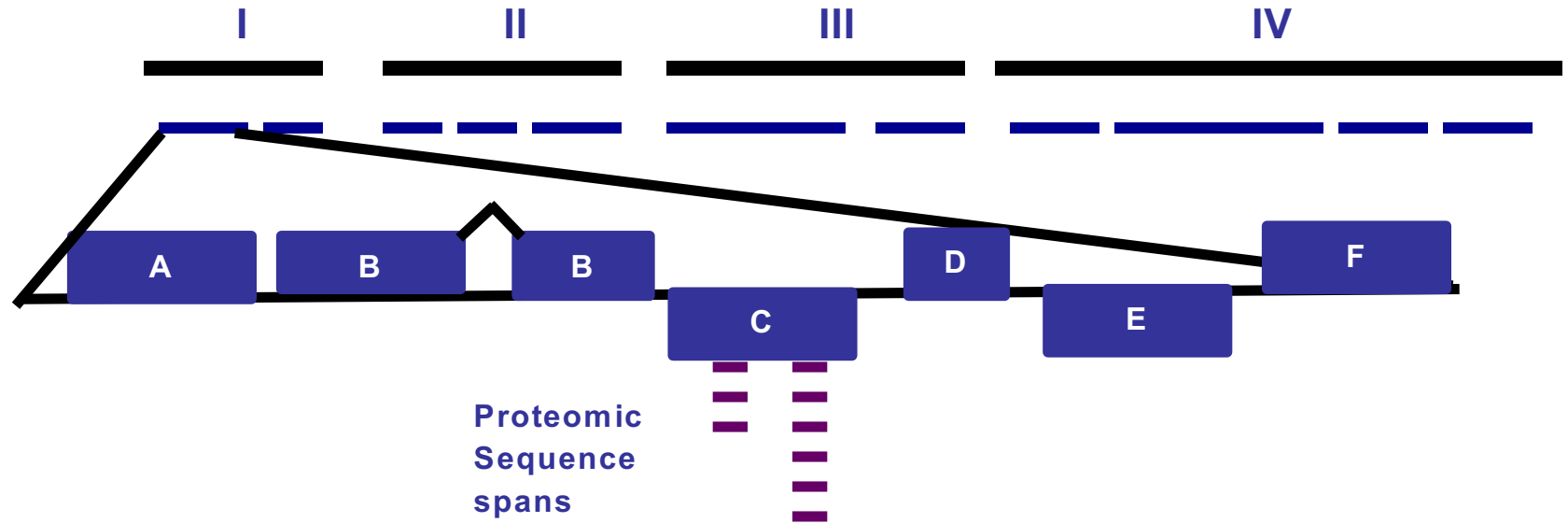
ENNPCKLQYDYNNTNVTHGFGQEYPCETDIVERFSDTEGAQCDDKKIKDENSEGACAPYRRL  
HVCVRNLENINDYSKINNKHNLLEVECLAAYEGESITGRYPQHQETNPDTKSQCLCTVLA  
RSFADIGDIIRGKDLYRGGNTKEKKRKKLEENLKTIFGHIYDELKNGKTNNGEELQKRY  
RGDKDNDFYQLREDWWDANRETVWKAITCNAQSYQYSQPTCGRGEIPYVTLKQOC IAGE  
VPTYFDYVPQYLRWFEEWAEDFCRKKKKKIPNVKTNCRQVQRGKEKYCDRDGYNCDGTIR  
KQYIYRLD TDCTKCSLACKTFAEWIDNQKEQFDKQKQKYQNEISGGGGRRQKRSTHSTKE  
YEGYEKHFNEELRNEGKDVRSFLQLLSKEKICKERIQVGEETANYGNFENESNTFSHTEY  
CDRCPLCGVDCSSDNCRKKPKDKSCDEQITDKEYPPENTTKIPKLTAEKRKTGILKKYEFK  
CKNSDGNNGGQIKKWECHYEKNDKDDGNGDINNCIQGDWKT SKNVYYPISYYSFFYGSII  
DMLNESIEWRERLKS CINDAKLGKCRKGCKNPCECYKRWVEKKKDEWDKIKEFFRKQKDL  
LKDIAGMDAGELLE FYLENIFLEDMKNANGDPKVIKFKELGKENEVQDPLKTKKTID  
DFLEKELNEAKNCVEKNPDNECPKQKAPGDGAAPSDPPREDITHHDGEHSSDEDEEEEE  
EEQQPPAEGTEQGEEKSESKEVVEQQETPQKDTEKTVPTTTPTVDVCDTVKLTALADTGS  
NAACSLKYVTCNYSWCIAPSGTSGKDAICVPPRTECLYLLKTLSDTTQKCLLEA  
FIKTAQEYLLMDDQNETSTLIIIPDTLGIIEDFRIFFDID  
LFLGRYIG  
GGKKTLETETYNYSNVTN  
LNEFASRPSFLRWMTEWGDQFCRERITQLQILKER  
CMVYQ  
DRREKCTEACTYYKEWLTNWQDNYKKQNRQRYTEVKGTSPYKEDSDVK  
ESKYAHGYLRKILKNIICTSGTDIAY **CNCMEGTST** TDSSNNDNIPESLKYPPIEIEEGCT  
CKDPSGPEVIPEKKVPEPKVLPKPPKLPKRQPKERDFPTPALKNAMLSSTIMWSIGIGFA  
TFTYFYLKKKTKSTIDLLRVINIPKSDYDIPTKLSPNRYIPYTSQKYRGKRYIYLEGDSG  
TDSGYTDHYS DITSSSESEYEELDINDIYAPRAPKYKTLIEVVLEPSGNNTTASGNNTPS  
DTQNDIQNDGIPSSKITDNEWNTLKDEFISQYLQSEQPNDVPNDYSSGDIPLNTQPNTLY  
FDNPDEKPFITSIHDRDLYSGEEYSYNVMVNTNNDIPISGKNGTYSgidLINDSLNSNN



Note: ORFs in addition to predicted Genes must be searched



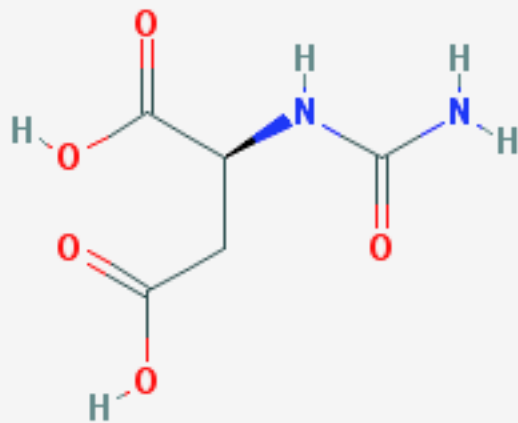
# 30,000 ft View - Proteomics



## Overview

<b>PubChem Compound ID:</b>	<b>CID:93072</b>
<b>PubChem Substance ID(s):</b>	<b>3727</b>
<b>Synonyms:</b>	<b>N-Carbamoyl-L-aspartate</b>
<b>Molecular Weight:</b>	<b>176.12742</b>
<b>Molecular Formula:</b>	<b>C<sub>5</sub>H<sub>8</sub>N<sub>2</sub>O<sub>5</sub></b>

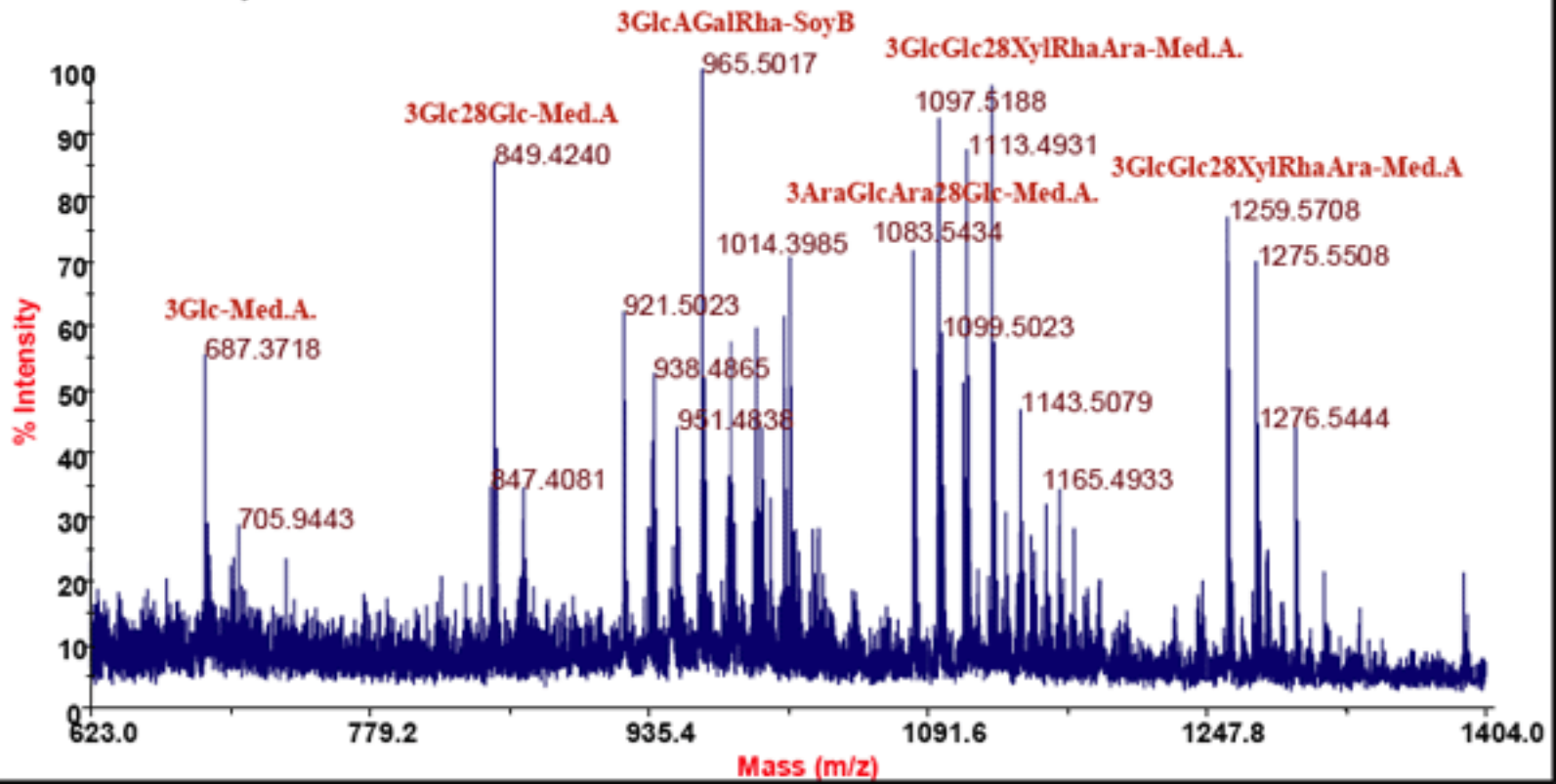
## 2D Structure



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

## ALANINE, ASPARTATE AND GLUTAMATE METABOLISM

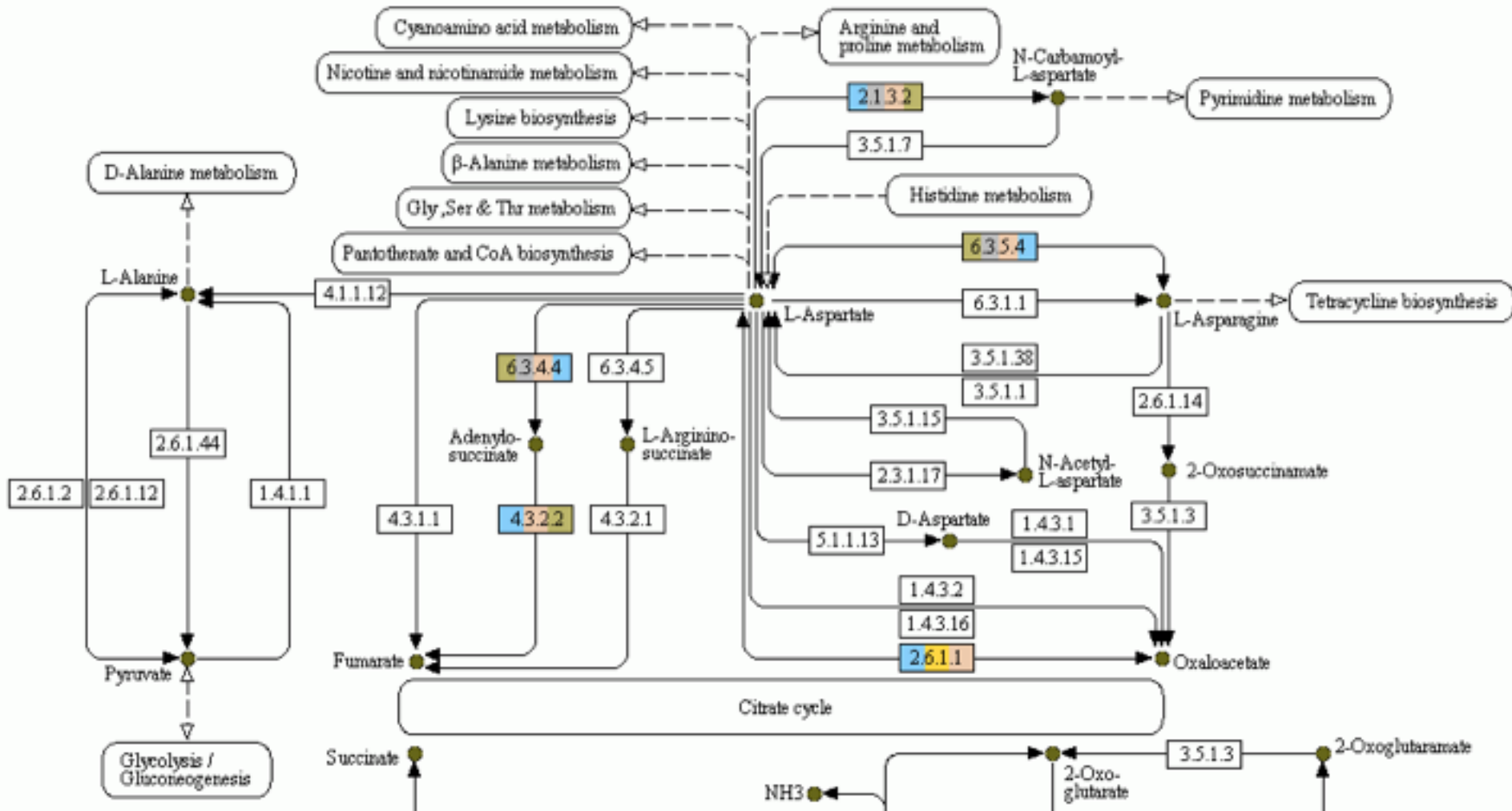
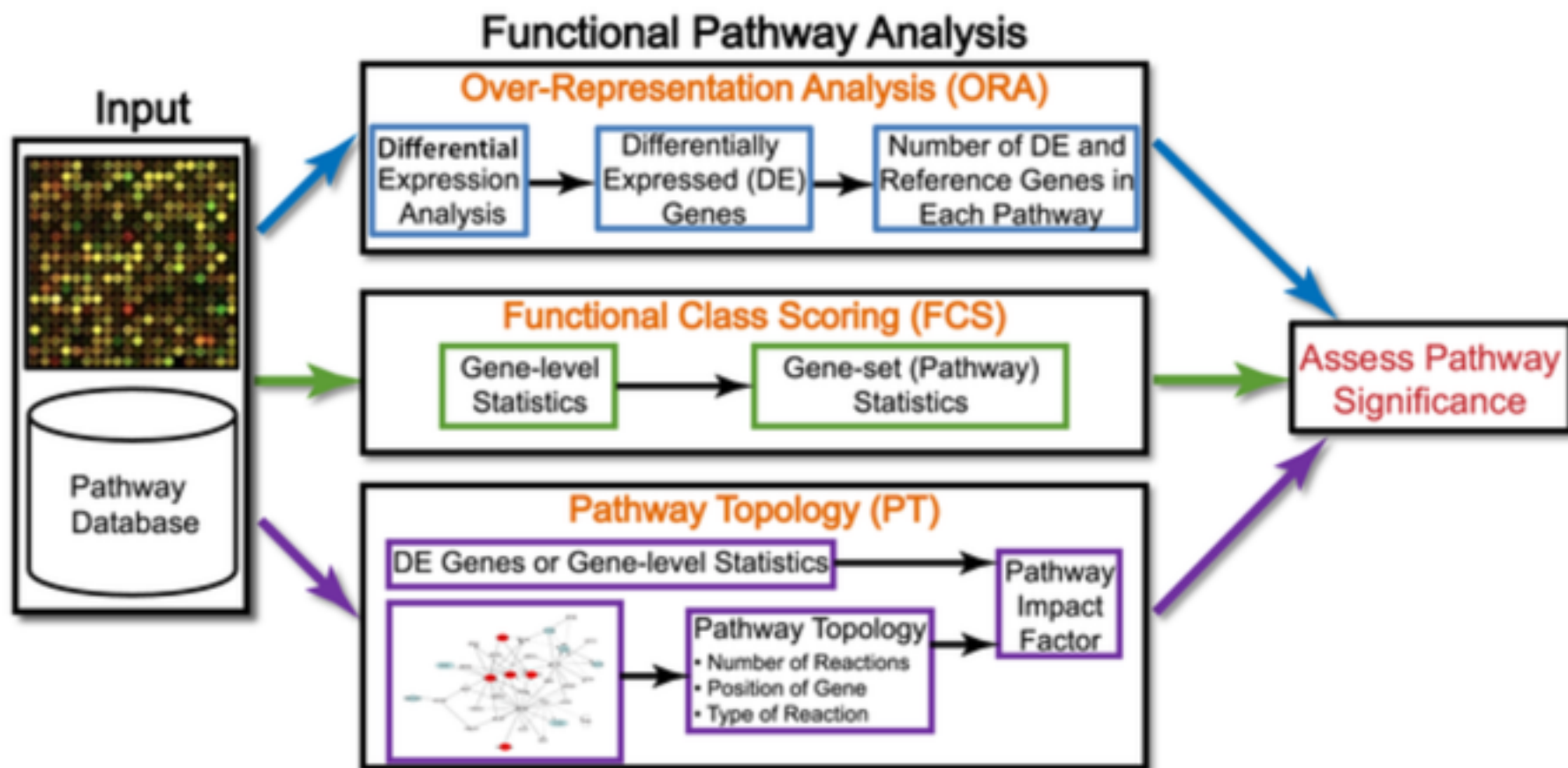


Figure 1. Overview of existing pathway analysis methods using gene expression data as an example.

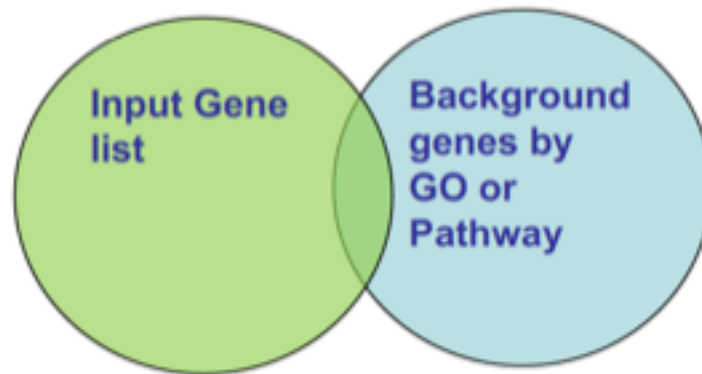


Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology 8(2): e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>  
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375>

# 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

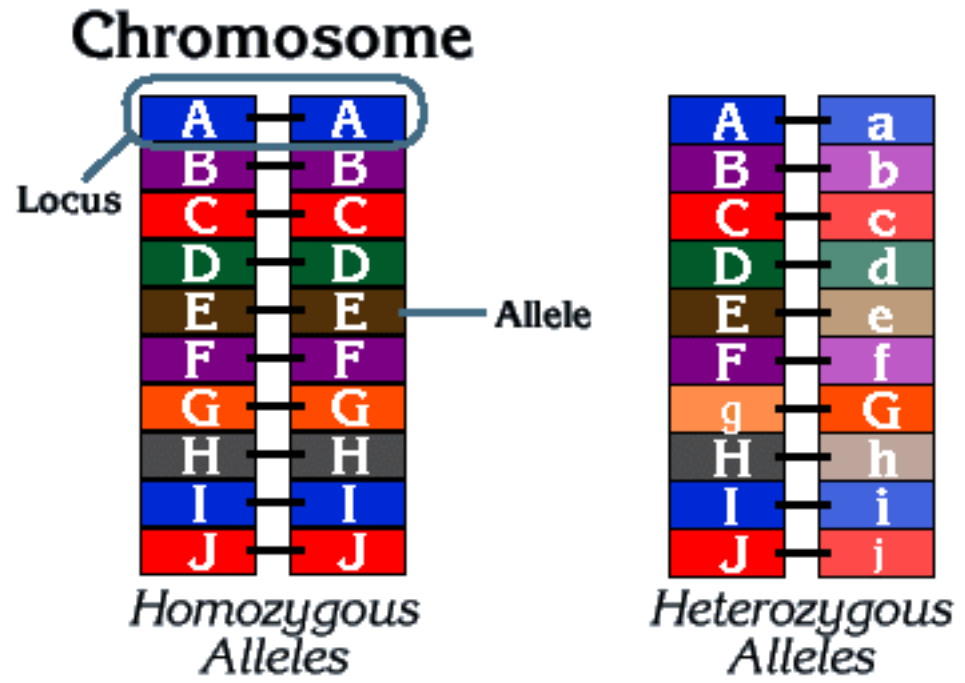
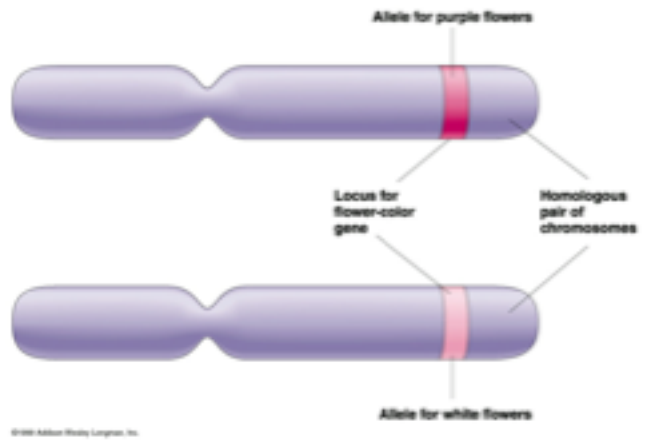
**Result: GO:ID or Pathway ID that is enriched**

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

# 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)



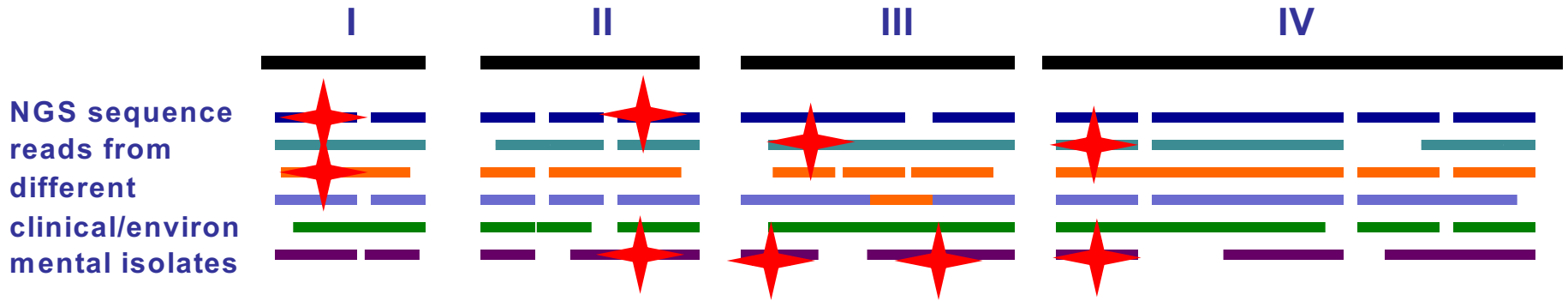
A    AAGCCTCATC

a    ACGCCTCATC

SNP = Single Nucleotide Polymorphism



# 30,000 ft View- NGS SNPs



 = SNP

reference	IGGTGATACT	AAGCTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
303.1	IGGTGATACT	AAGCTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
309.1	IGATAATNCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
RV_3600	IGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
RV_3606	IGATAATNCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
RV_3610	TGATGATTCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT119.09	IGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
SenT123.09	IGATRATTCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT140.08	IGGTGATACT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CGGTGCTTCA
SenT142.09	IGGTGATACT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT175.08	IGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA

Reference = A  
 6 isolate seq = A  
 2 isolate seq = T  
 2 isolate seq = N (no call)  
 % with base call = 80  
 Major allele = A  
 Major allele freq = 75% (6/8)

Reference = G  
 9 isolate seq = A  
 1 isolate seq = G  
 % with base call = 100  
 Major allele = A  
 Major allele freq = 90% (9/10)

Reference = C  
 8 isolate seq = C  
 2 isolate seq = T  
 % with base call = 100  
 Major allele = C  
 Major allele freq = 80% (8/10)

# Population data

## Data

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

## Technology

- Chip-Seq
- NGS



## Allele Frequencies for **Exon 3.48 bp VNTR**

Locus [D21S11](#)

[Information on this locus](#)

[Click on icon](#) for additional information.

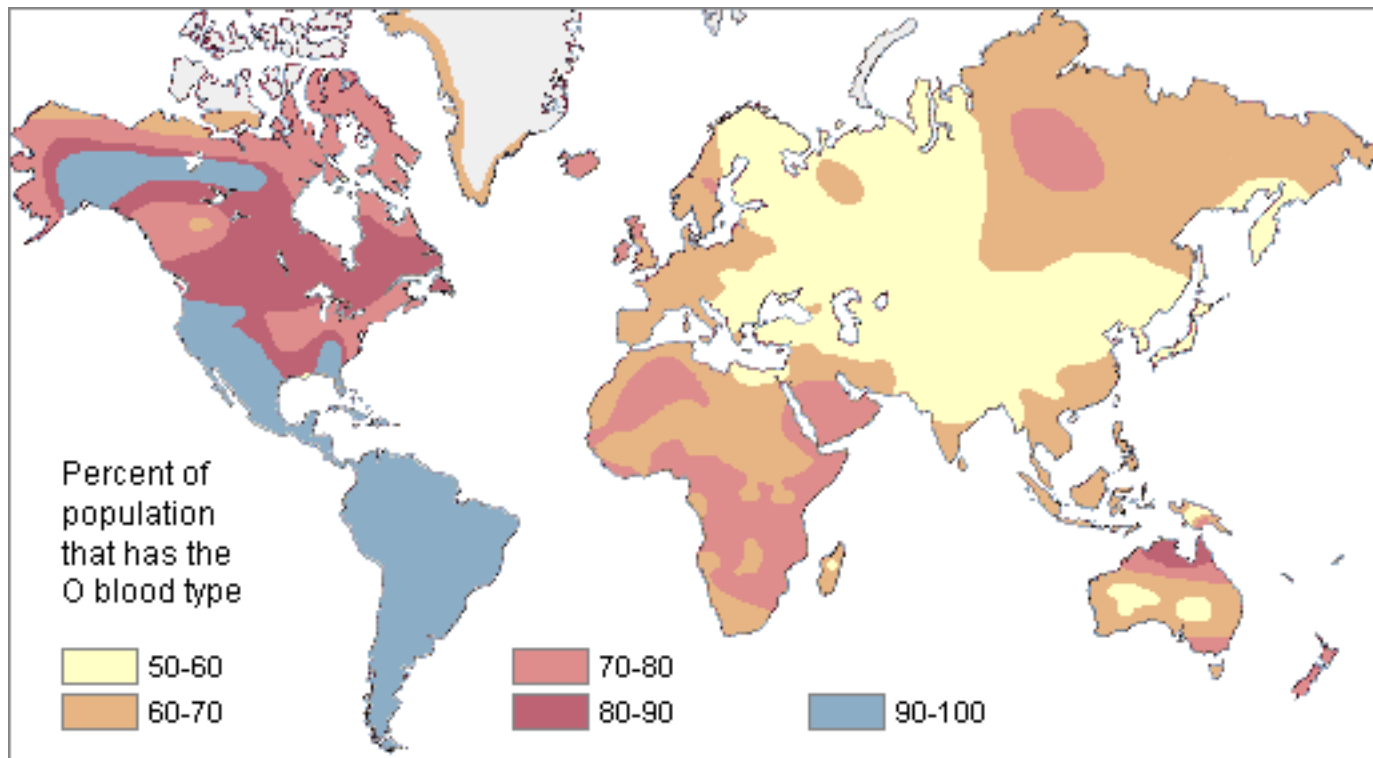
Geographic region	Population (Sample ID, Typed Sample Size (N), entry date) Add Info	
Africa	<a href="#">Bantu (LA00007F)</a> , 134, 9/19/2000	
Africa	<a href="#">Mbuti (LA00006C)</a> , 72, 9/19/2000	
Africa	<a href="#">Jewi, Ethiopean (LA00042F)</a> , 128, 9/19/2000	
Europe	<a href="#">Druze (LA00006T)</a> , 130, 9/19/2000	
Europe	<a href="#">Jewi, Ypresian (LA00014H)</a> , 80, 9/19/2000	
Europe	<a href="#">Samaritan (LA00009B)</a> , 78, 9/19/2000	
Europe	<a href="#">Ashazi (LA00017I)</a> , 108, 9/19/2000	
Europe	<a href="#">Dane (LA00007H)</a> , 108, 9/19/2000	
Europe	<a href="#">Europeans, Mixed (LA00002C)</a> , 174, 9/19/2000	
Europe	<a href="#">Europeans, Mixed (LA001775U)</a> , 146, 6/21/2004	
Europe	<a href="#">Finn (LA00011J)</a> , 46, 9/19/2000	
Asia	<a href="#">Indones (LA00131N)</a> , 154, 6/23/2004	
Asia	<a href="#">Keritao (LA00131P)</a> , 148, 6/23/2004	
Asia	<a href="#">Keritao (LA00131Q)</a> , 114, 6/23/2004	
Asia	<a href="#">Keritao (LA00131R)</a> , 138, 6/23/2004	
Asia	<a href="#">Marshay (LA00132Q)</a> , 114, 6/23/2004	
Asia	<a href="#">Kachao (LA00004E)</a> , 36, 9/19/2000	
East Asia	<a href="#">Ami (LA00006X)</a> , 80, 9/19/2000	
East Asia	<a href="#">Ainu (LA00021T)</a> , 84, 9/19/2000	
East Asia	<a href="#">Han (LA00000E)</a> , 94, 9/19/2000	
East Asia	<a href="#">Japanese (LA00001B)</a> , 180, 9/19/2000	
East Asia	<a href="#">Cambodian, Khmer (LA00012E)</a> , 58, 9/19/2000	
East Asia	<a href="#">Hakka (LA00001D)</a> , 32, 9/19/2000	
Oceania	<a href="#">Melanesian, New (LA00012D)</a> , 46, 9/19/2000	
Oceania	<a href="#">Melanesian (LA00001I)</a> , 58, 9/19/2000	
Siberia	<a href="#">Yakut (LA00011C)</a> , 92, 9/19/2000	
North America	<a href="#">Cherokee (LA00021P)</a> , 96, 9/19/2000	
North America	<a href="#">Pima, Arizona (LA00012H)</a> , 94, 9/19/2000	
North America	<a href="#">Pima, Mexico (LA00012G)</a> , 104, 9/19/2000	
North America	<a href="#">Iroquois of Androisians (LA00014G)</a> , 96, 9/19/2000	
North America	<a href="#">Mesa, Yucatan (LA00013E)</a> , 100, 9/19/2000	
South America	<a href="#">Garifuna (LA00148N)</a> , 58, 6/9/2005	
South America	<a href="#">Kari'na (LA00020K)</a> , 108, 9/19/2000	
South America	<a href="#">Tupi (LA00014E)</a> , 90, 9/19/2000	

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

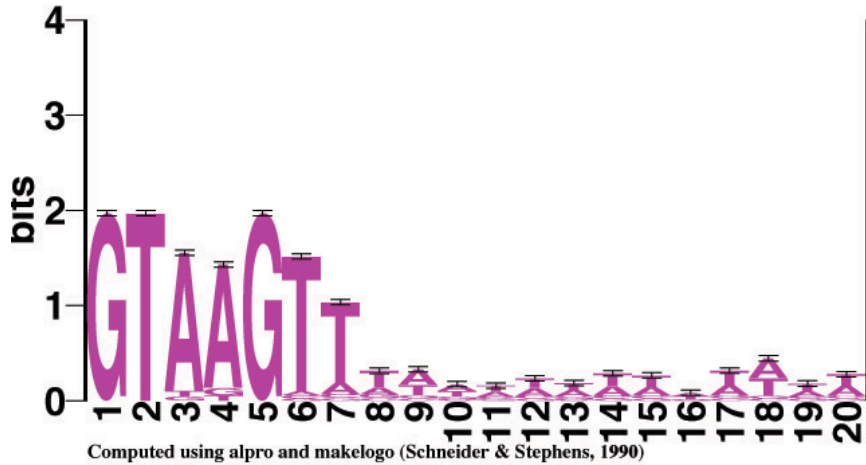
- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats (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 preceded by pyrimidine stretch
- Has a mean length of “X” as is observed in this species

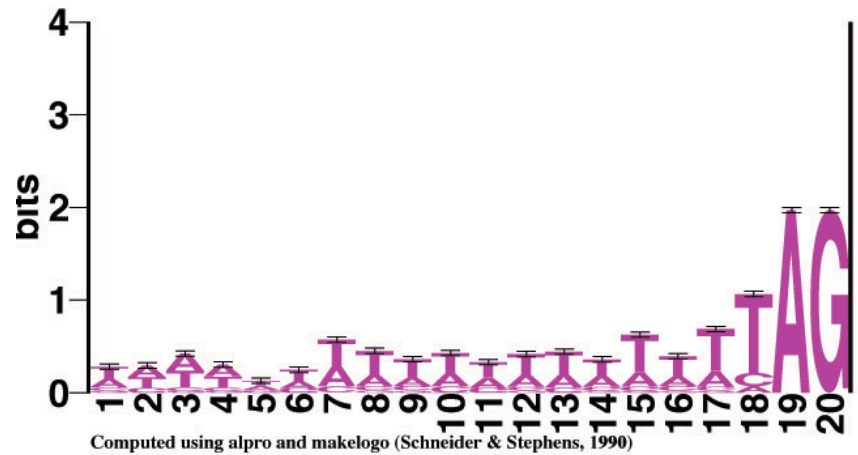
# Donor Site

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

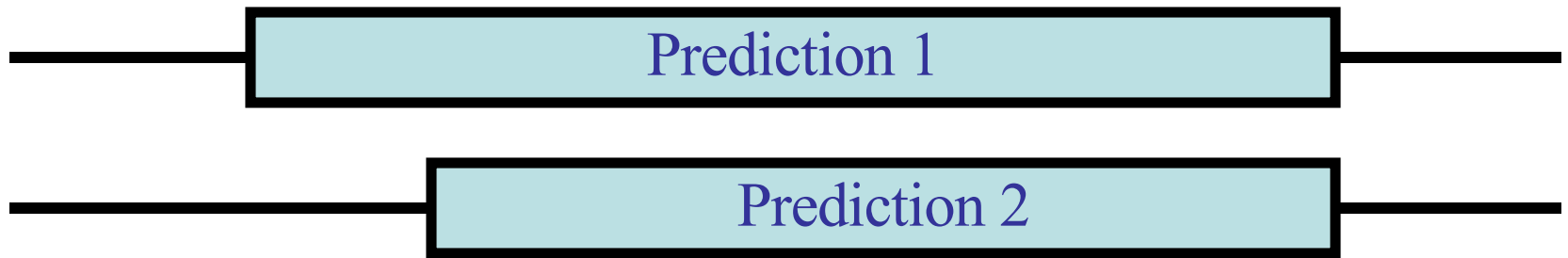


# Acceptor site

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!**



# 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

OPEN ACCESS Freely available online

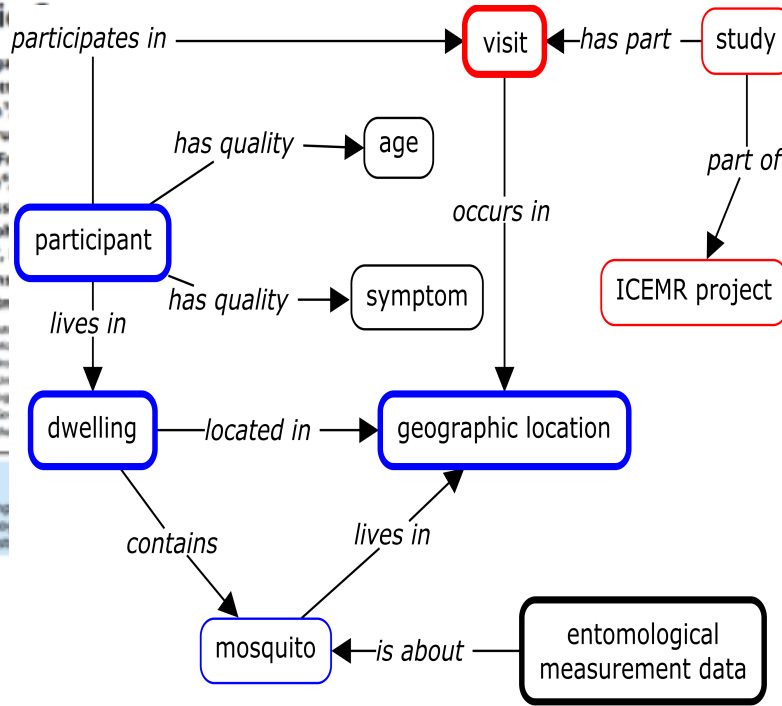
PLOS ONE

## Standardized Metadata for Human Pathogen/Vector Genomics

Vivien G. Dugan  
Brett E. Pickett  
Indresh Singh<sup>1</sup>  
Vincent M. Bruner  
Valentina Di Florio  
Florian Fricke<sup>2</sup>  
Mizrachi<sup>3</sup>, Jess  
Cheryl L. Murphy  
David Rasko<sup>4</sup>,  
Rick L. Stevens  
Jennifer Wortman

1.1. Craig Venter Institute  
United States of America, 9300  
Baltimore, Maryland, US  
Center for Biotechnology  
of America, 18 Holly Dr  
12 Department of Pathology

**Abstract**  
High throughput  
disease pathogen  
association in



	Core Project	Core Sample	Project Specific	Pathogen Specific	Sequencing Assay
Investigation	■				
Host Characterization		■	■		
Specimen Isolation		■	■		
Pathogen Characterization		■	■		
Specimen Processing		■	■		
Pathogen Detection		■			
Pathogen Isolation			■		
Sample Management			■		
Data Transformation			■		
Sample Shipment					
Sequencing Sample Preparation					■
Sequencing Assay					■

# 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.