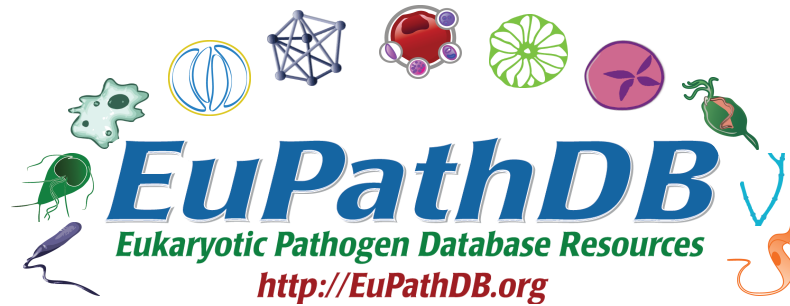


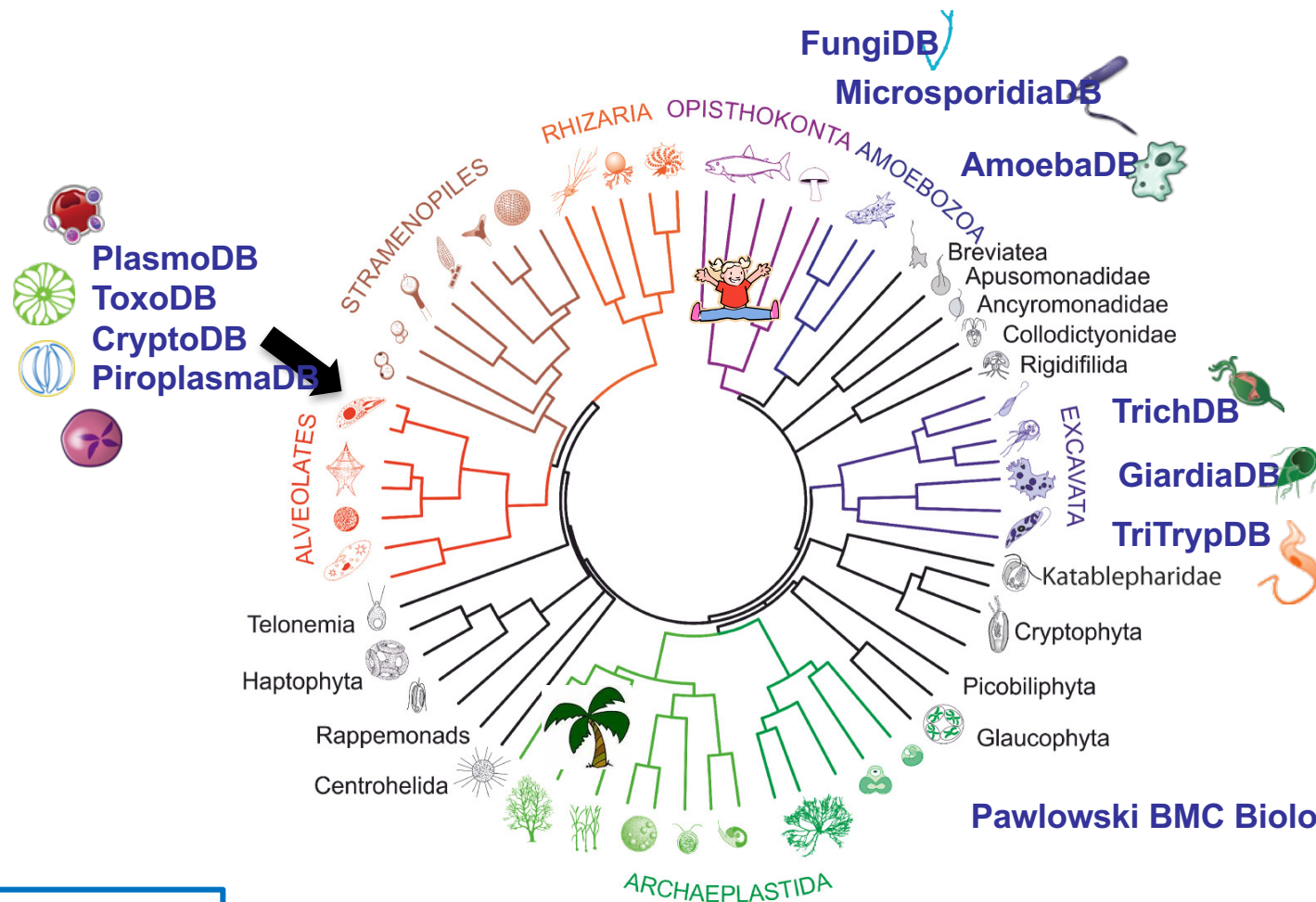


Crash Course in Omics Terminology, Concepts & Data Types

12th Annual EuPathDB Workshop: 2017

Jessie Kissinger

June 18, 2017

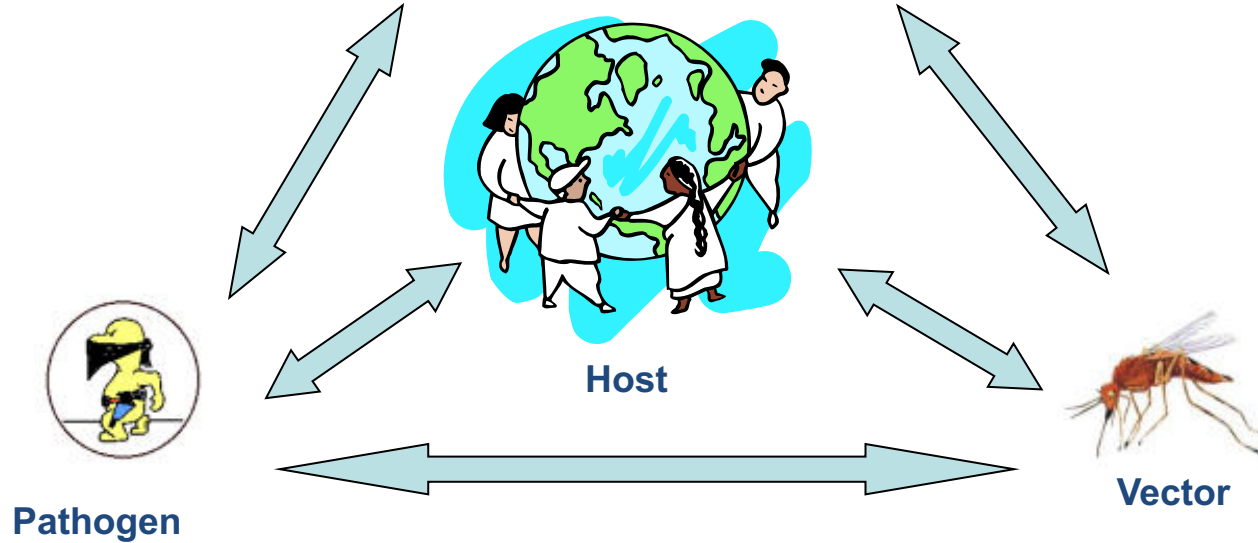


Pawlowski BMC Biology 2013

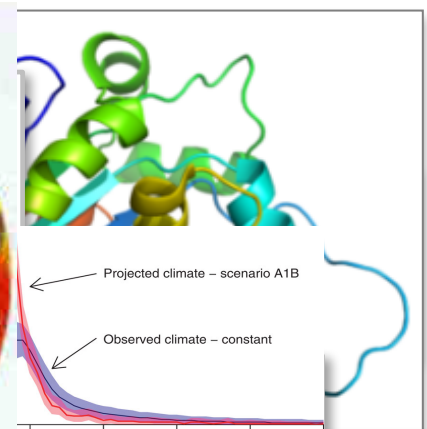
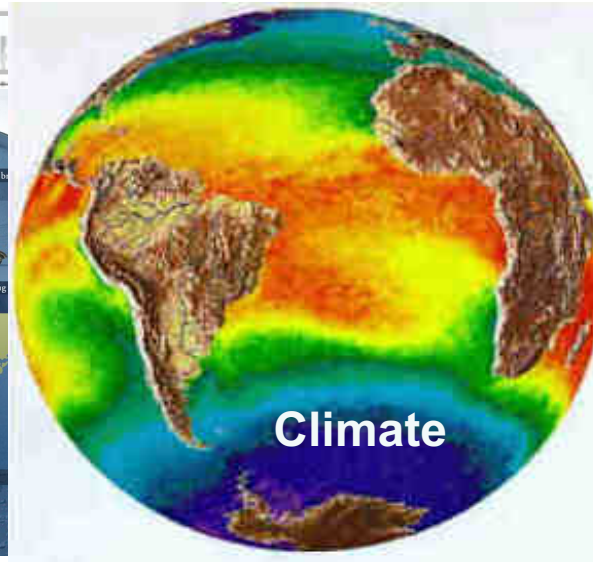
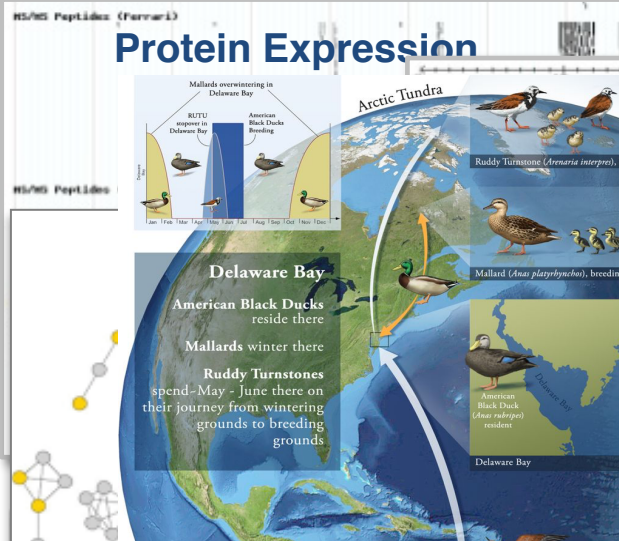
248 Organisms



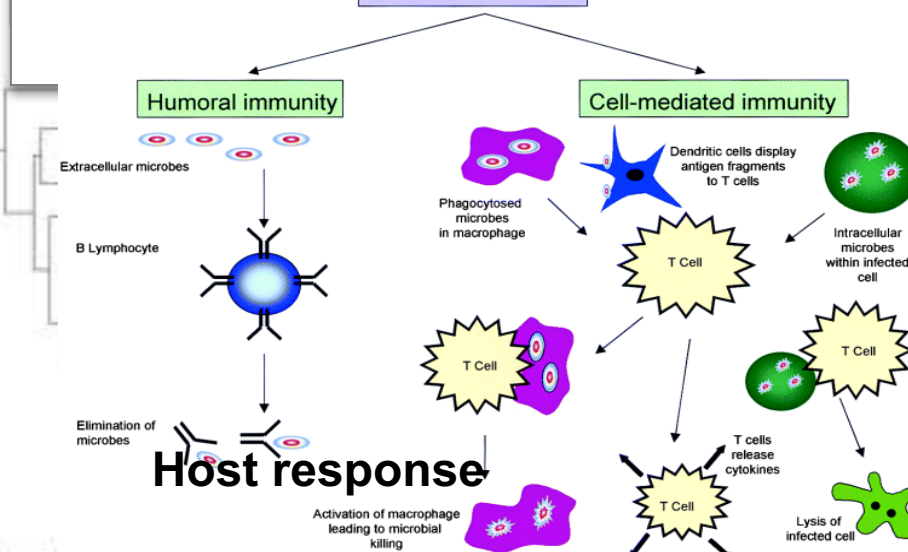
Infectious Disease Paradigm



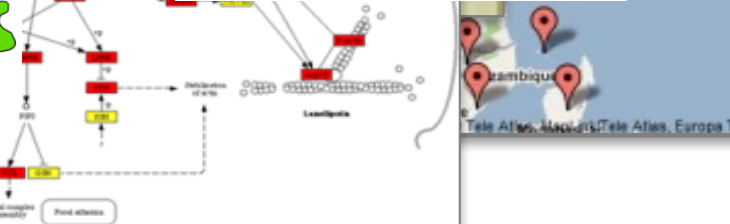
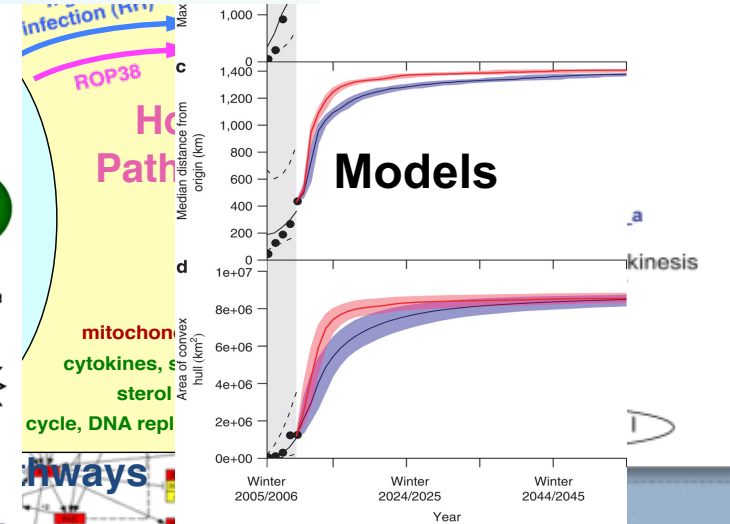
Protein Expression



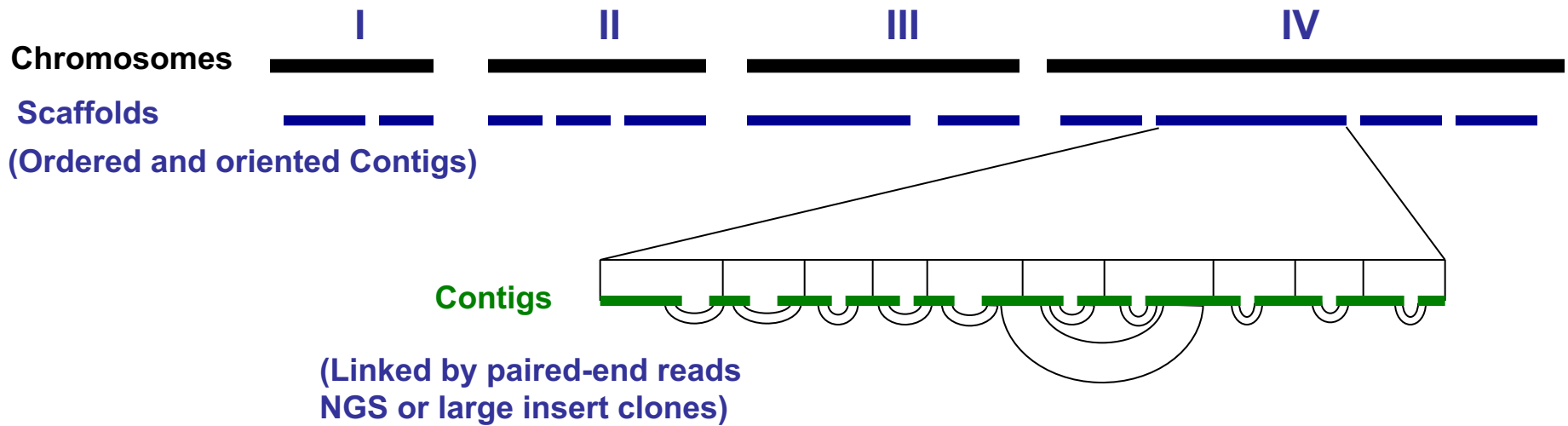
Adaptive Immunity



Host response

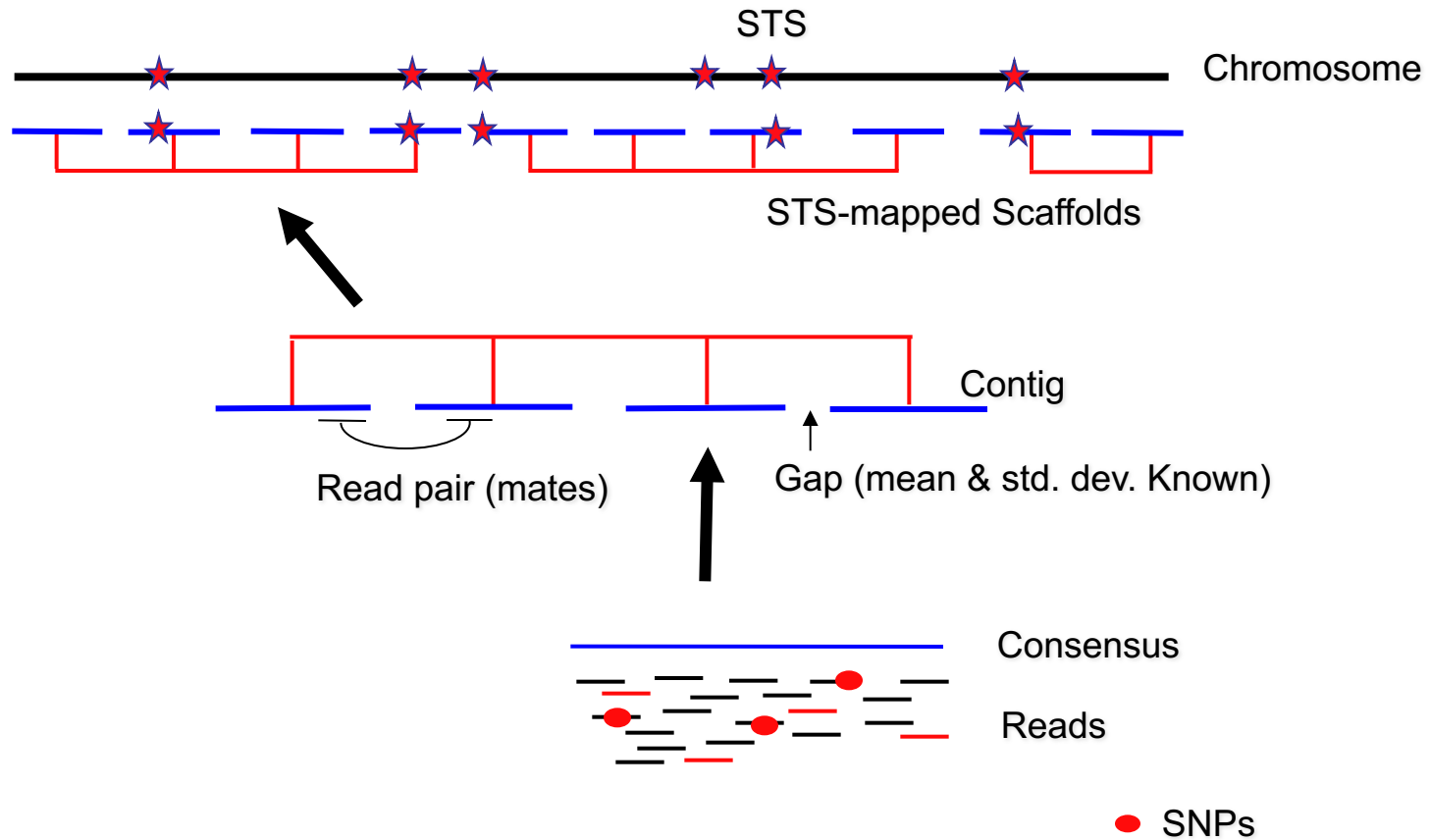


30,000 ft View - Genome Assembly

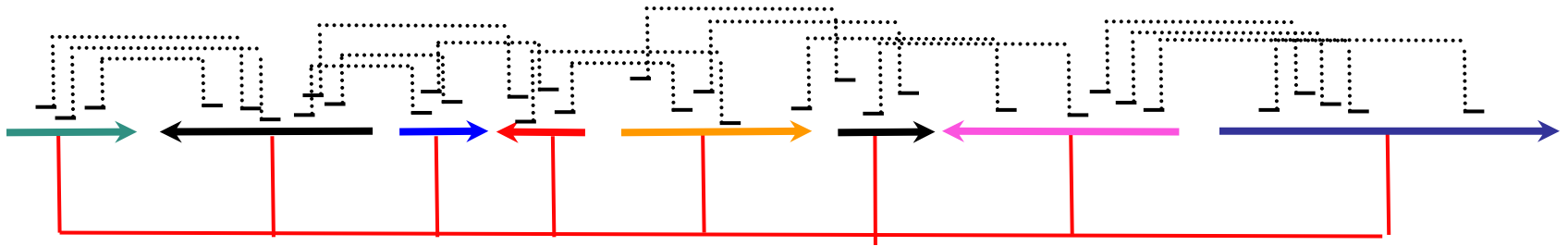


5X genome sequence means that sequences equivalent to 5X the genome size were generated e.g. Genome size = 10 Mbp, then 50Mbp of random sequences were generated

Anatomy of a WGS Assembly

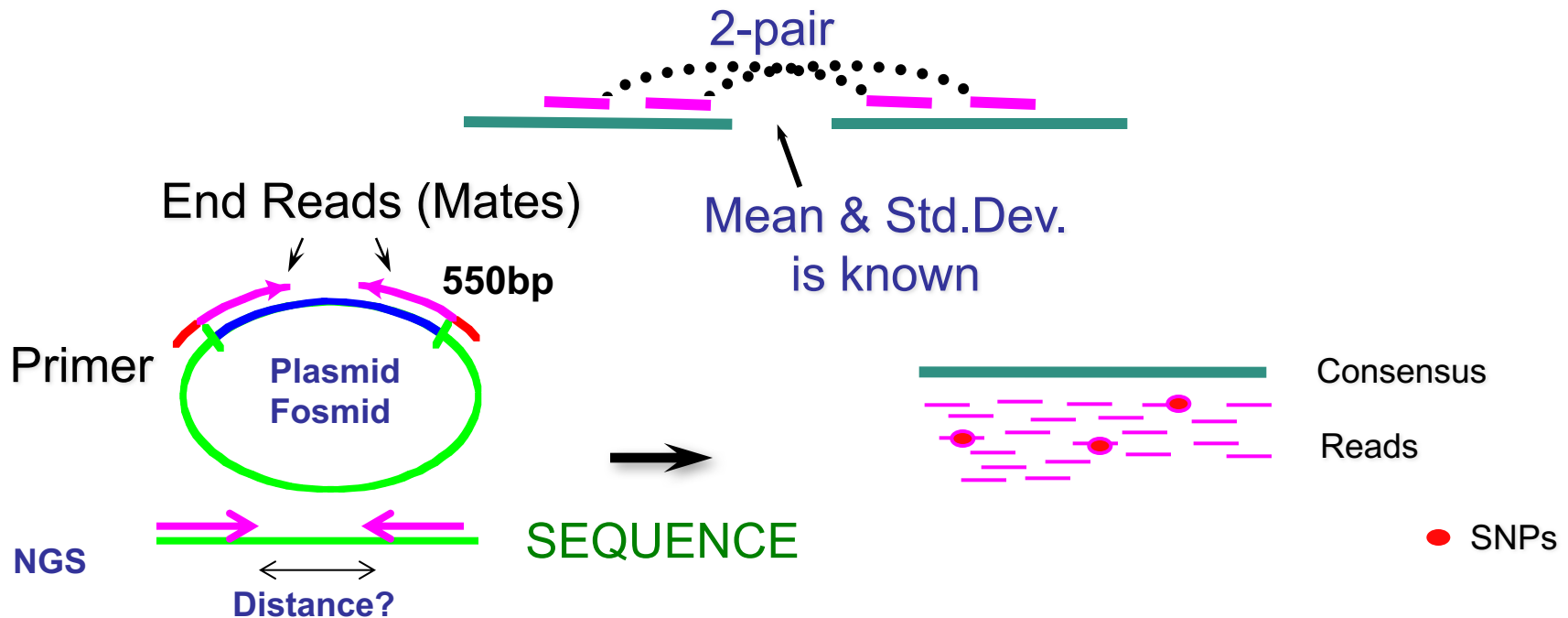


Pairs Give Order & Orientation

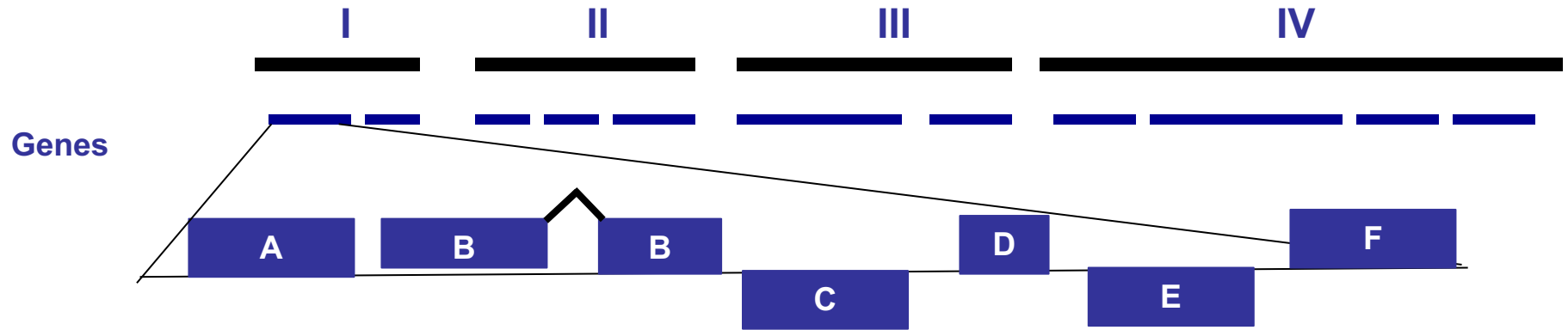


Scaffold

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



30,000 ft View - Annotation



[illegible]

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

```

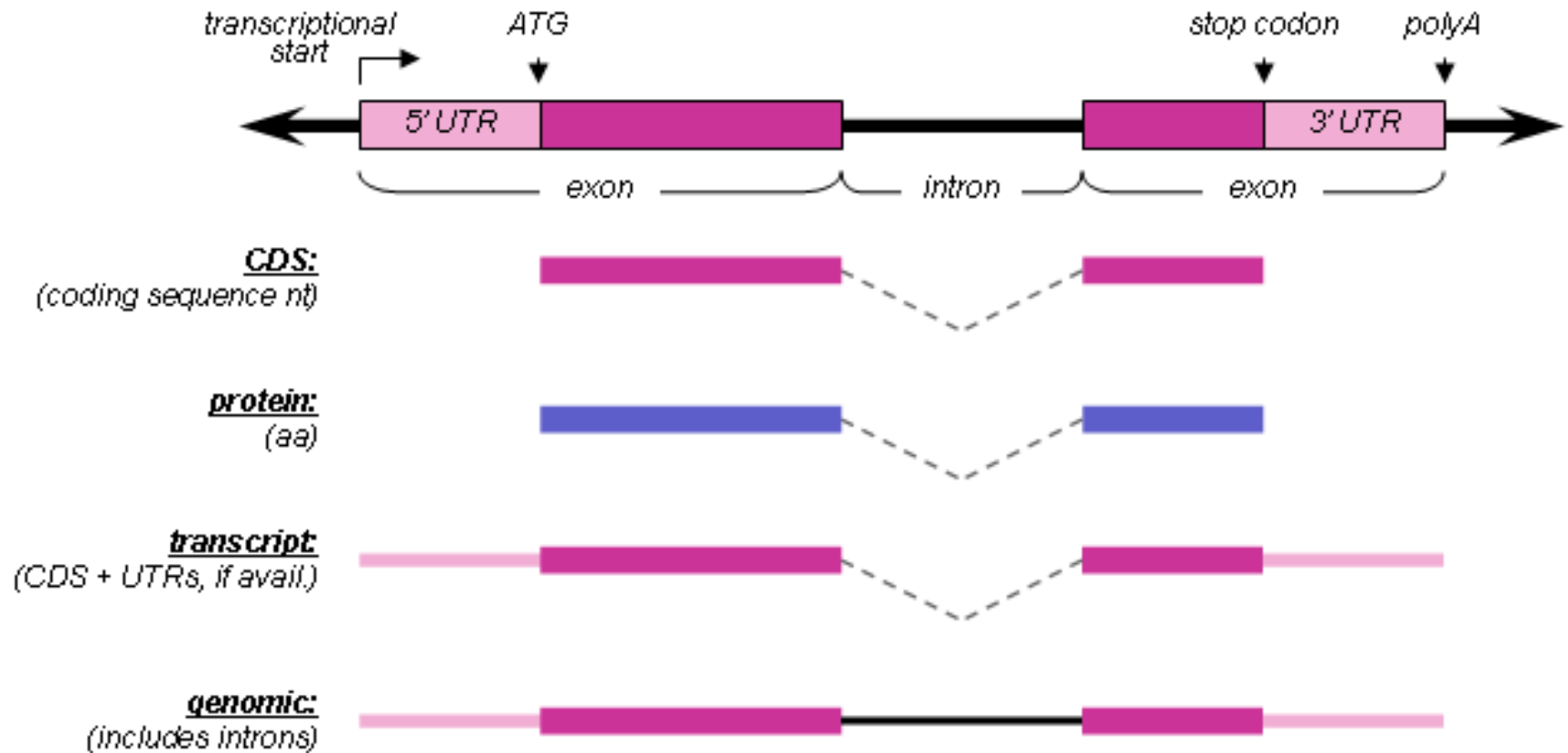
ORFs ≠ Genes

ATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC
ACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGTCGCCTGAACGGGTGGCT
TCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATG
GGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCTCTTCCC
TCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTGACACAGATTTTGTGTCGTTGGAGGAGCGGGACTGTACGAG
GCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGAGTTCCGTGCGACGTTTTCTTCC
CTGCGTTCCCCGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTTCGT
TCCCTTTTGTCCGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCA
GACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGAGCA
ACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACCTCCCGTGACGGGTGACAGGCTCCTTCTTCGGCCGACGCCAT
TGCCCCGGTGTGCGGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAAGTATTGCGGCCGTTCCGCAT
GTTCACTTTAGAGGCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATG
ACCGAACGG

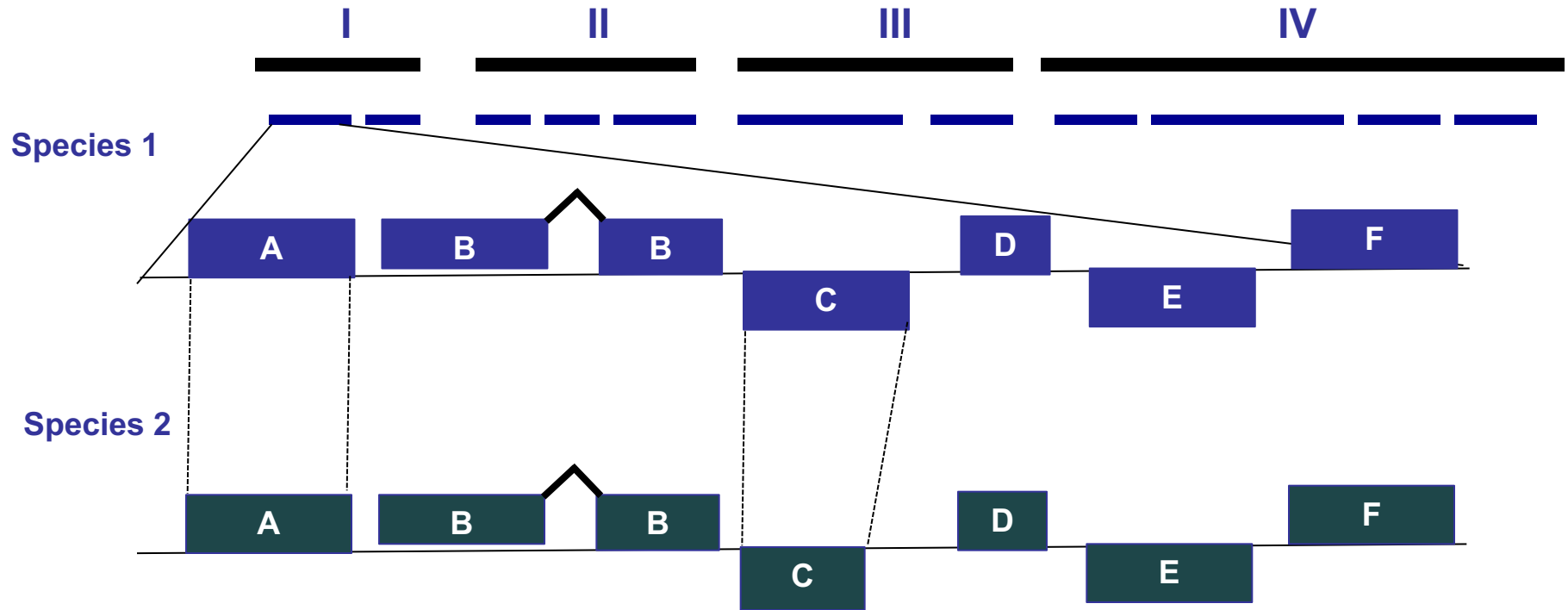
>Translation Frame 1

MQKPVCLVVAMTPKRGIGINNGLPWPHLTTFDKHFSRVTKTTPEEASRLN
GWLPRKFAKTGDSGLPSVSGKRFNAVVMGRKWTWESMPRKFRPLVDRLNI
VVSSSLKEEDIAAEKPQAEQQQVRVVCASLPAALSLLEEEYKDSVDQIFV
VGGAGLYEAALSLGVASHLYITRVAREFPDVFVPAFPGDDILSNKSTAA
QAAAPAESVFVPFCPELGREKDNEATYRPIFISKTFSDNGVPYDFVLEK
RRKTDDAATAEPSNAMSSLTSTRETTVPVHGLQAPSSAAAIAPVLAWMDEE
DRKKREQKELIRAVPHVHFRGHEEFQYLDLIADIINNGRTMDDRT

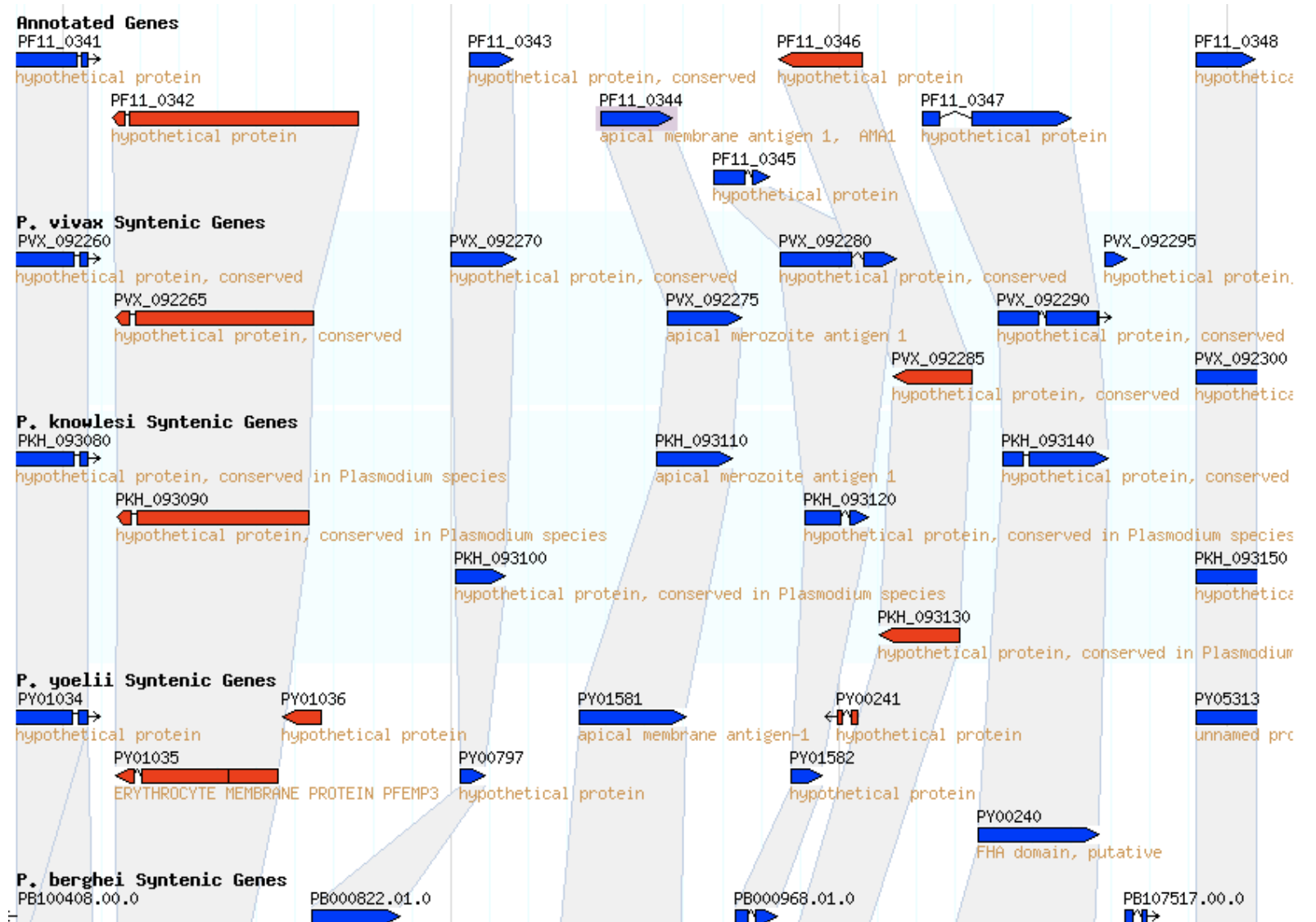
Terminology



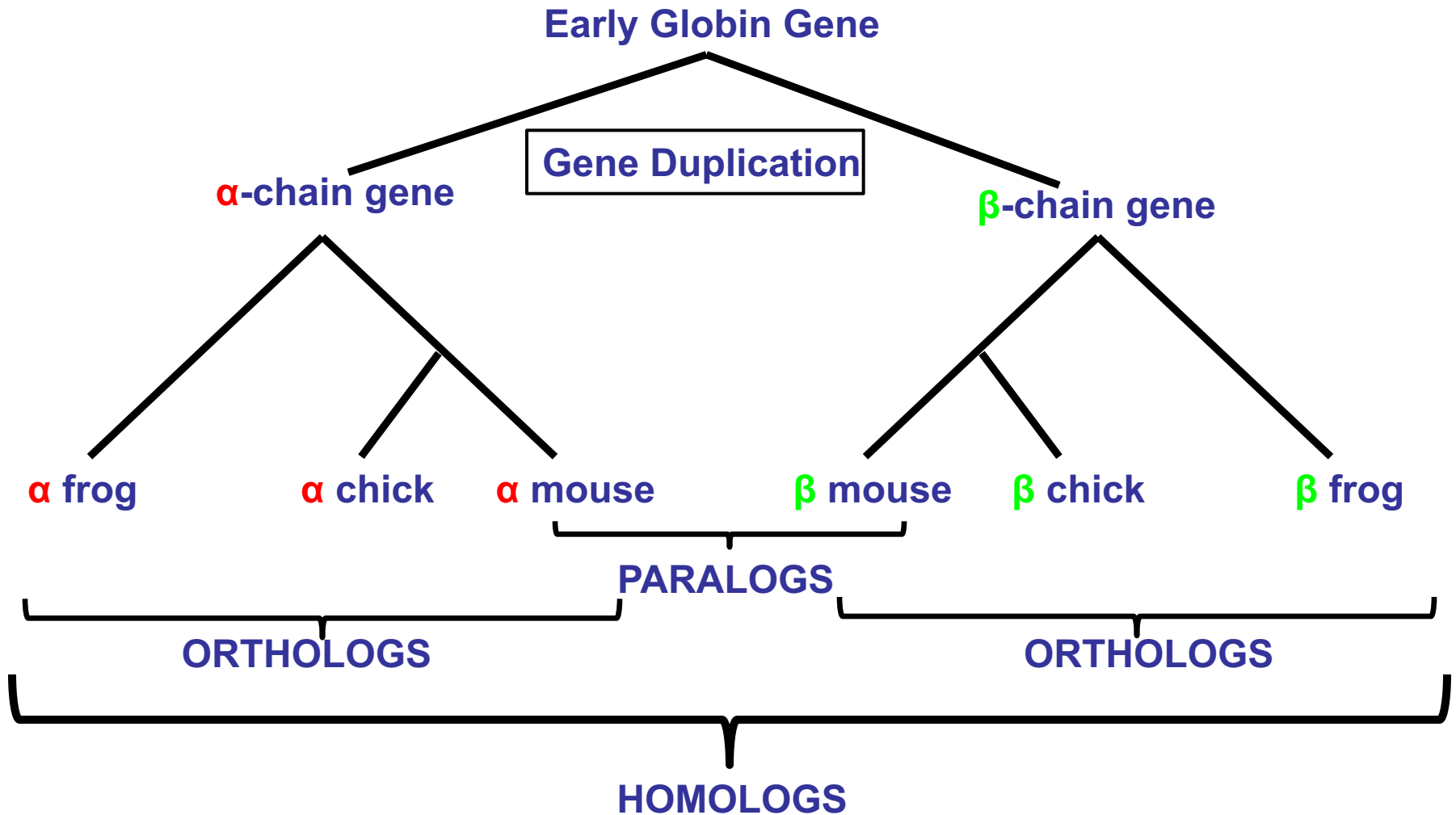
30,000 ft View - Synteny



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:

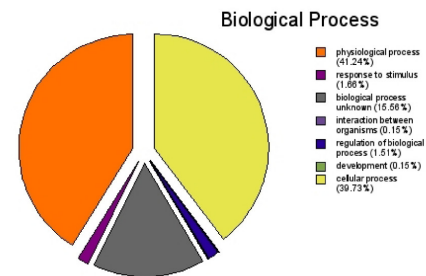
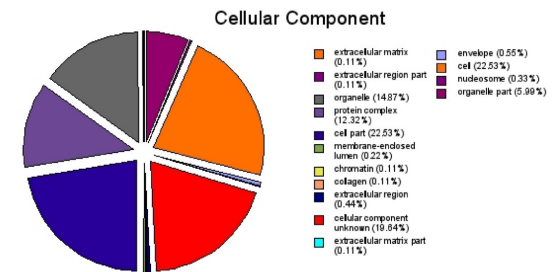
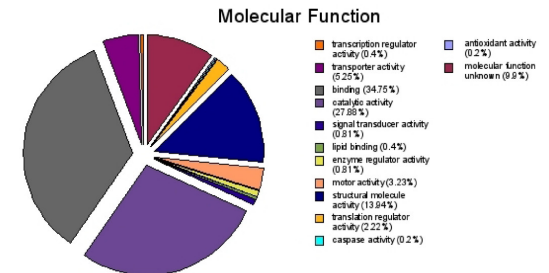
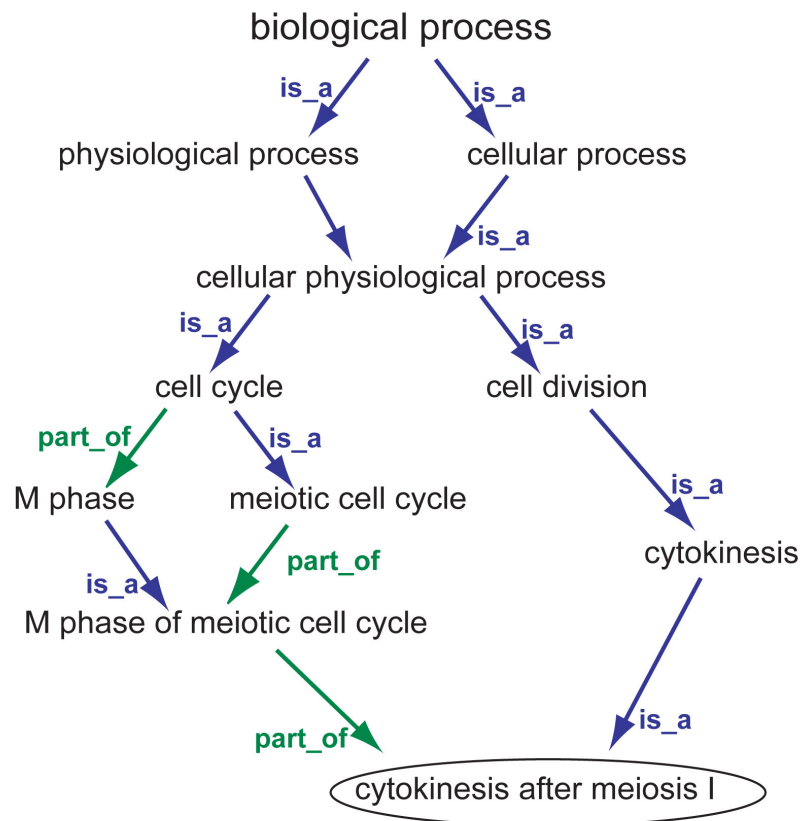
D. melanogaster gene CG3340 annotated as:
"Kruppel" and *P. falciparum* gene
PF3D7_1209300 annotated a "putative
KROX1"

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



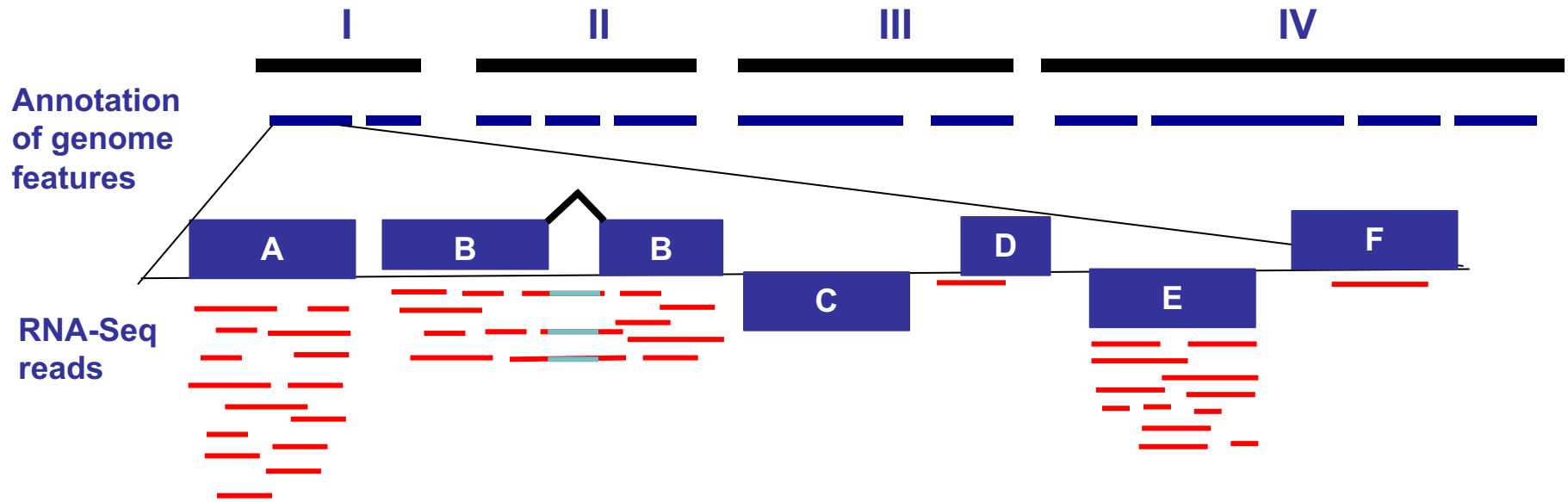
Expression Profiles (RNA and Protein)

- 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

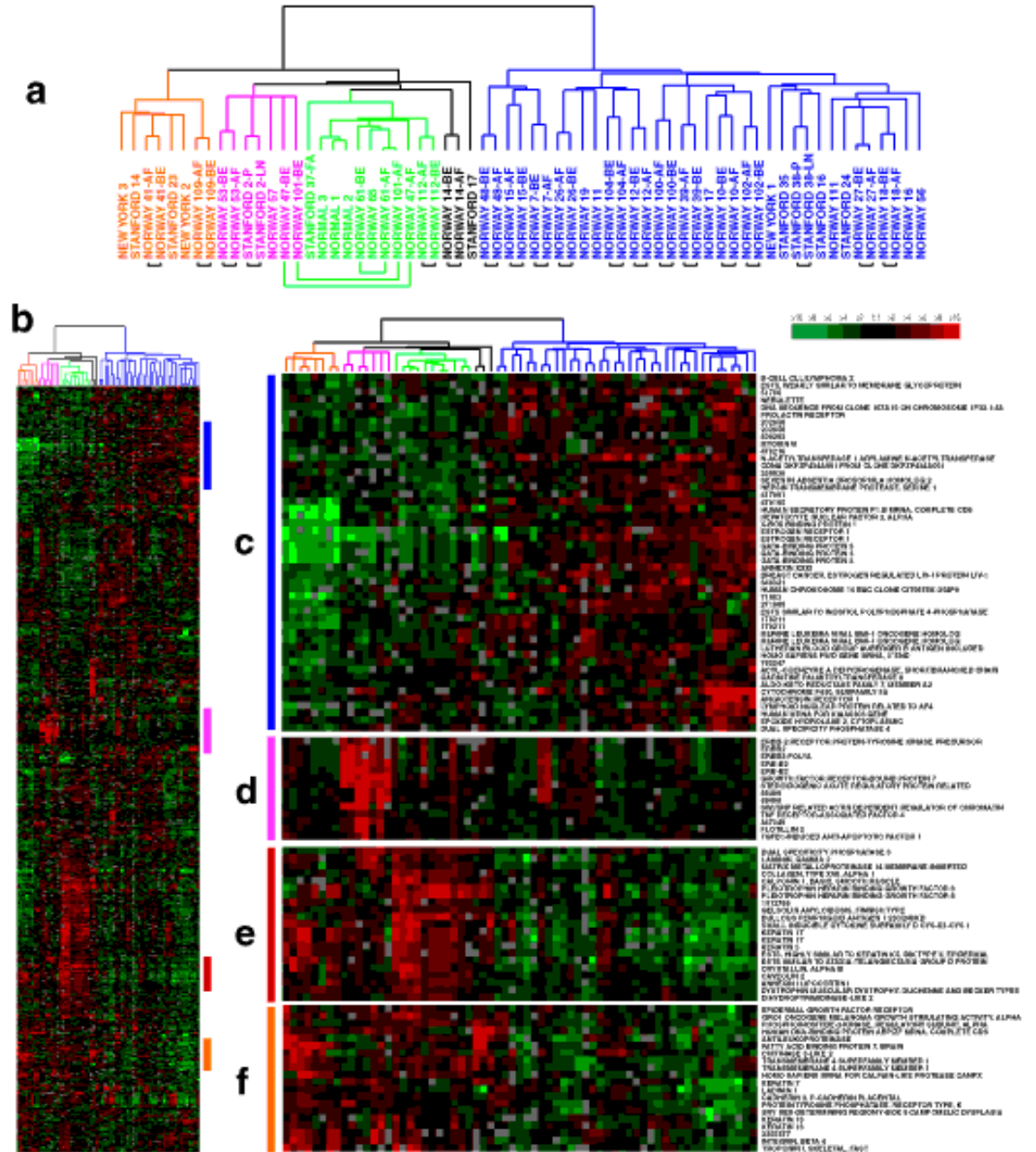


FPKM = Fragments per kilobase of exon per million fragments mapped

Clustered
Microarray
Data
Genes with
Similar
Expression
Profiles are
Grouped
together

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

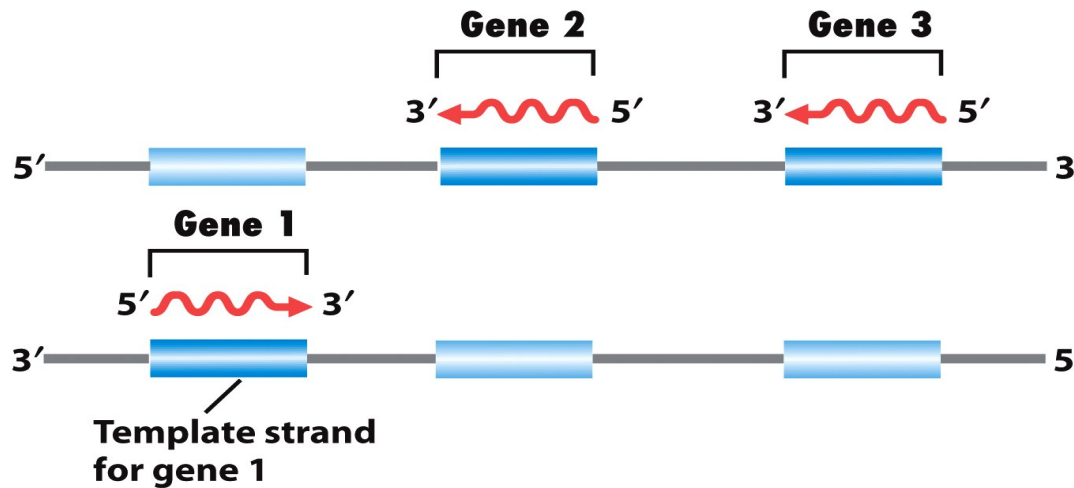


Figure 8-3
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

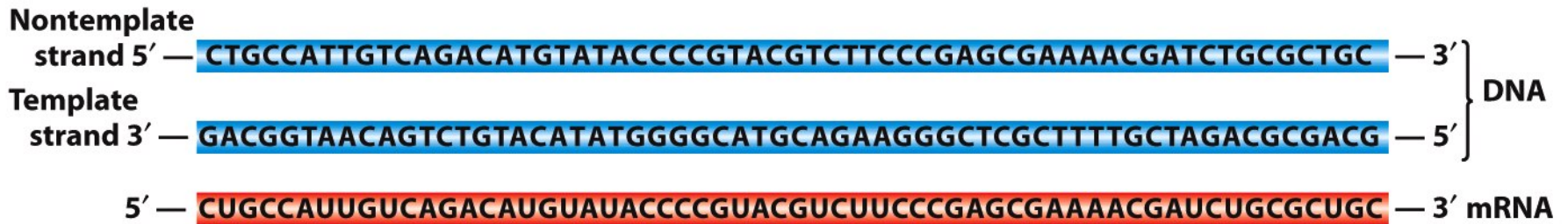


Figure 8-6
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

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

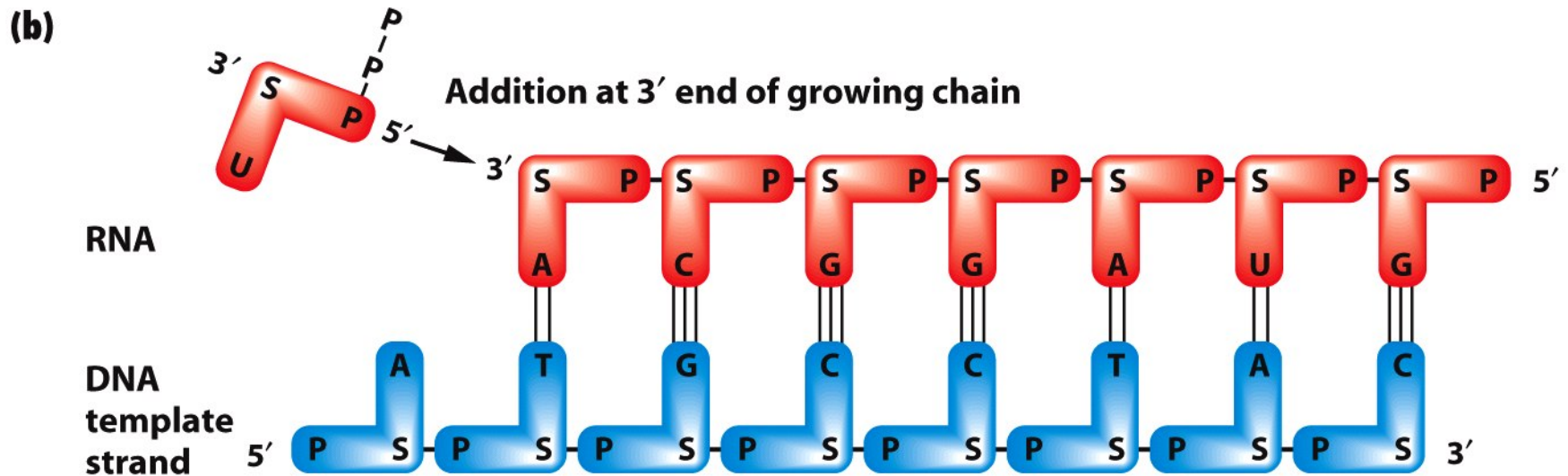
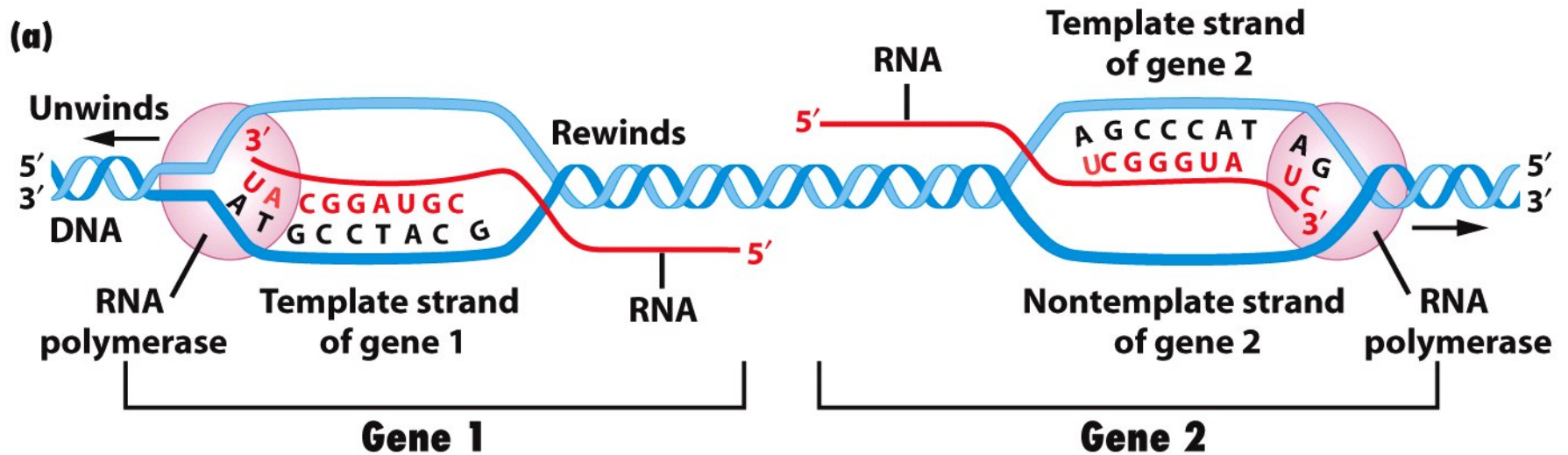


Figure 8-4
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

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

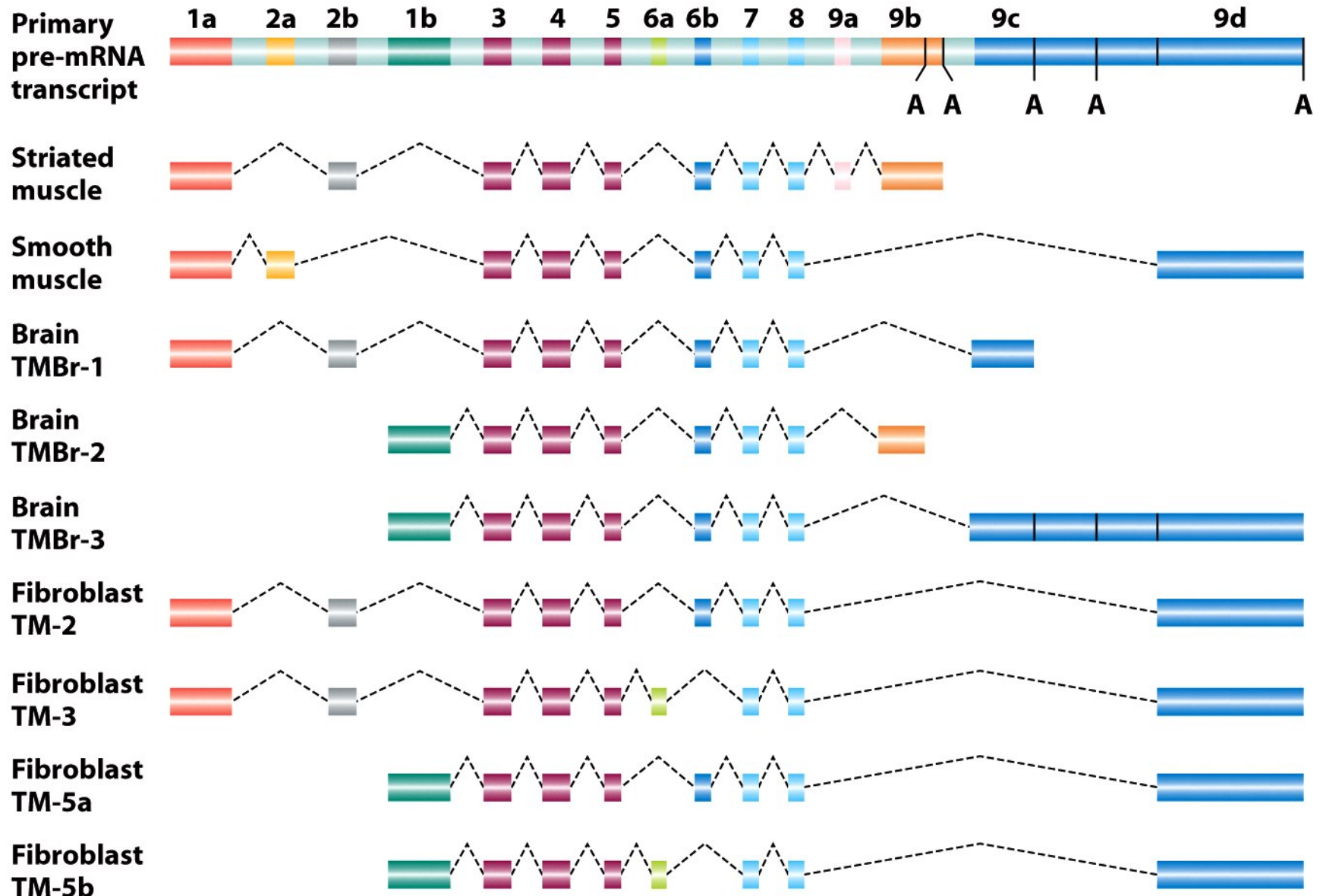
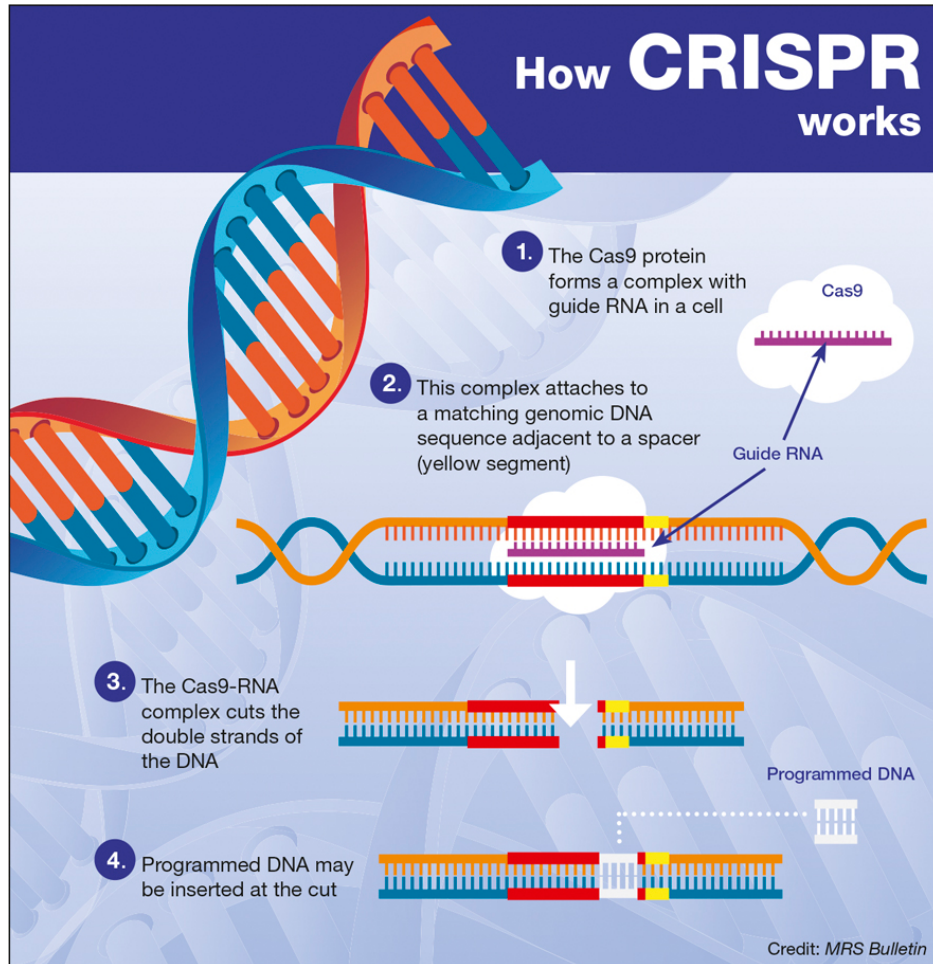


Figure 8-14

Introduction to Genetic Analysis, Ninth Edition

© 2008 W.H. Freeman and Company

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

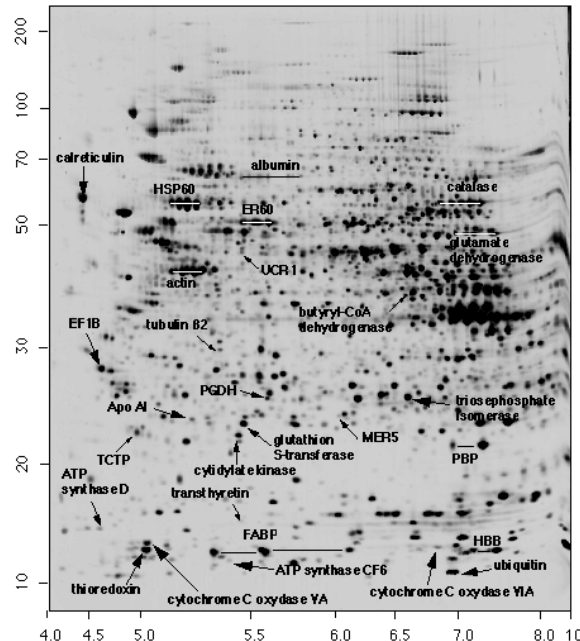
Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans

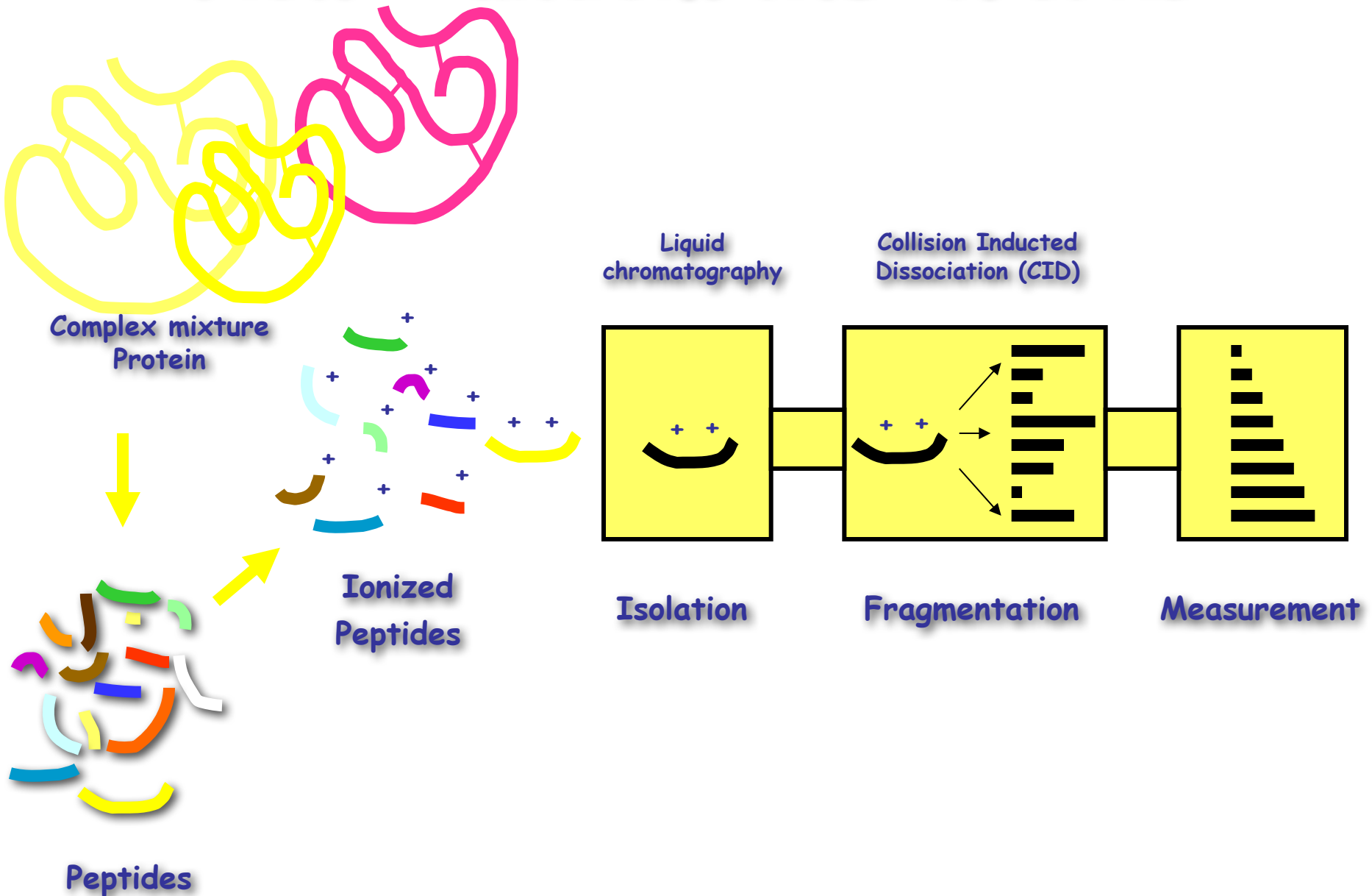
Technology

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



Typical 2 D gel

How Tandem MS Works



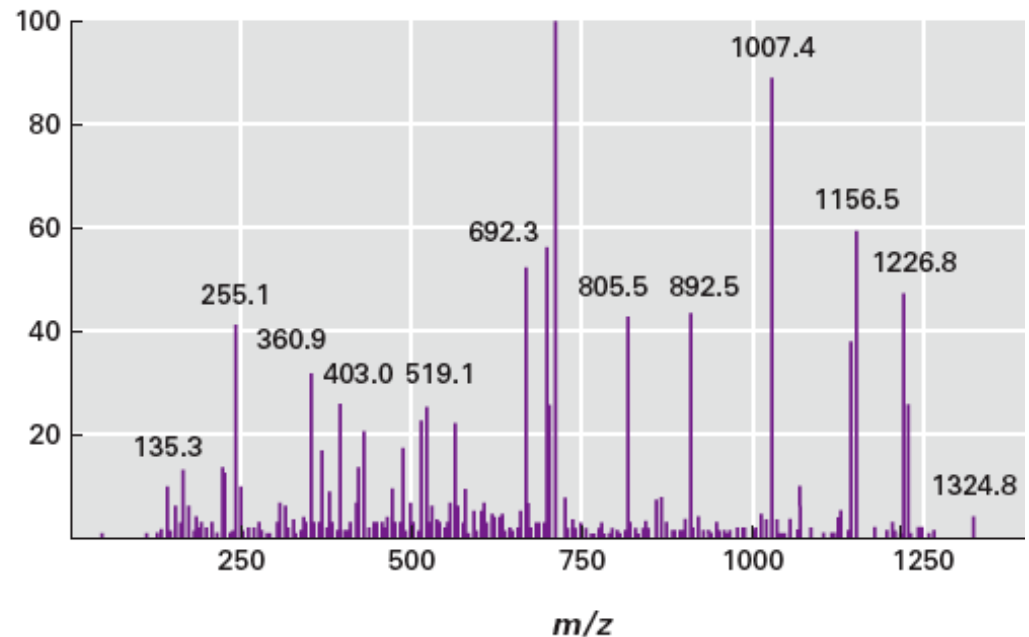
Tandem MS protein data

a)

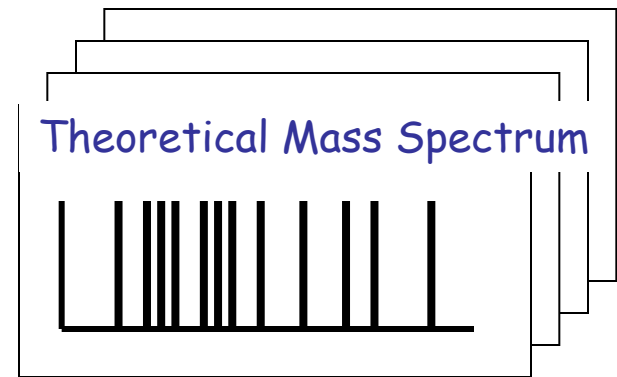
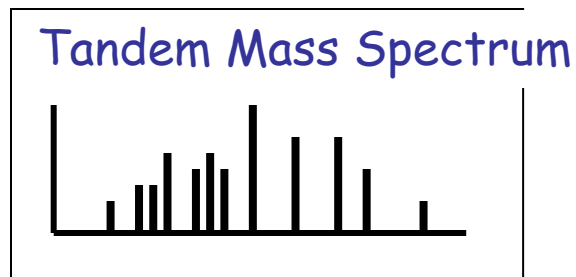
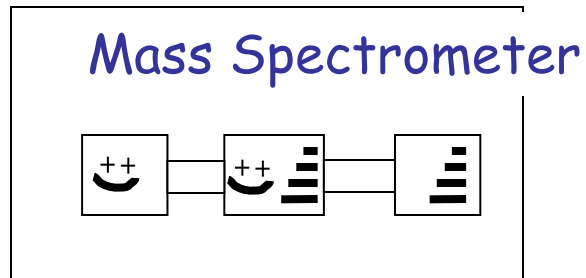
S-P-A-F-D-S-I-M-A-E-T-L-K
(protonated mass 1410.6)

Mass ⁺	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)



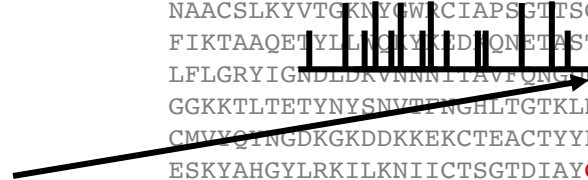
Sequest Database Search



Correlation Analysis

Ranked Score of Matched Peptides

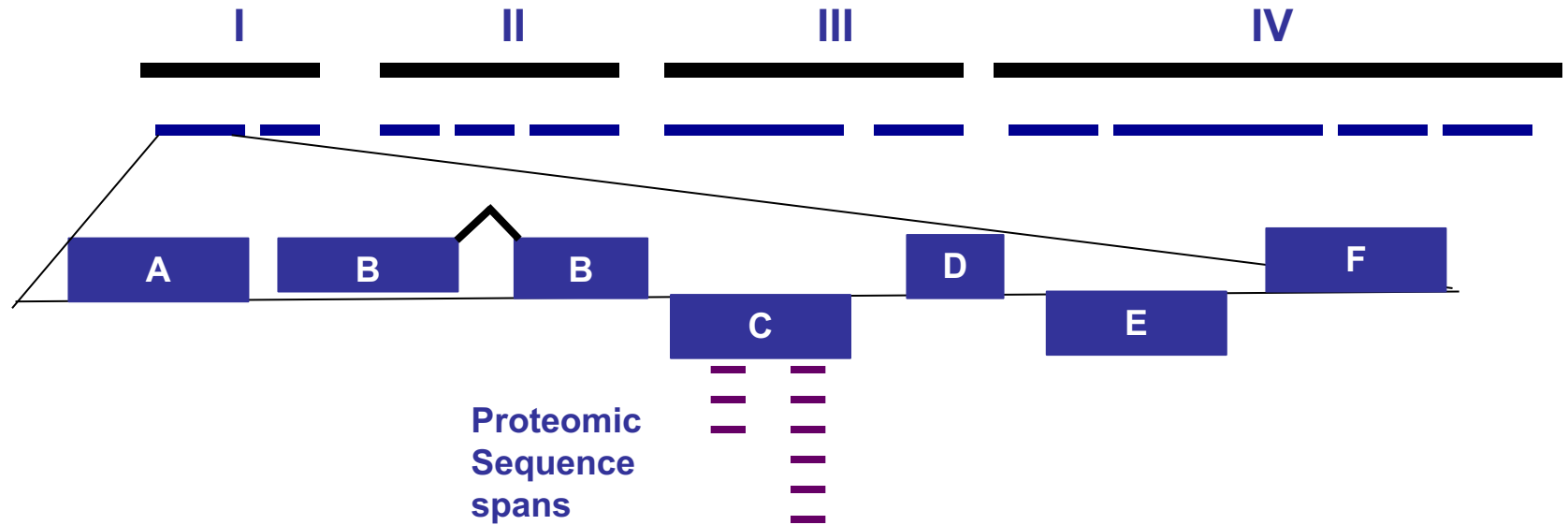
Peptide database



ENNPCKLQYDYNNTNVTHGFGQEYPCETDIVERFSDTEGAQCDDKKKIKDNSEGACAPYRRL
 HVCVRNLENINDYSKINNKHNLLEVCCLAAKYEGESITGRYPQHQETNPDTKSQCLTVLA
 RSFADIGDIIRGKDLRYGGNTKEKKRKKLEENLKTIFGHIYDELKNGKTNGEEELQKRY
 RGDKDNDFYQLREDWWDANRETVWKAITCNAQSYQYSQPTCGRGEIPYVTLSCQC IAGE
 VPTYFDYVPQYLRWFEEWAEDFCRKKKKKIPNVKTNCRQVQRGKEKYCDRDGYNC DGTIR
 KQYIYRLD'TDC'TKCSLACKTFAEWIDNQEQFDKQKQKYQNEISGGGGRRQKRSTHSTKE
 YEGYEKHFNEELRNEGKDVRSLQLLSKEKICKERIQQVEETANYGNFENESNTFSHTEY
 CDRCPLCGVDCSSDNCRKKPKDKSCDEQITDKEYPPENTTKIPKLTAEKRKTGILKKYEKF
 CKNSDGNNGGQIKKWECHYEKNDKDDGNGDINNCCIQGDWKTSKNVYYPISYYSFFYGSII
 DMLNESIEWRERLKSCINDAKLGKCRKGCKNPCECYKRWVEKKKDEWDKIKEFFRKQKDL
 LKDIAGMDAGELLE FYLENIFLED MNANGDPKVIKFKKEILGKENE EVDPLKTKKTID
 DFLEKELNEAKNCVEKNPDNECPKQKAPGDGAAPSDPPREDITHHDGEHSSDEDEDEDEEEE
 EEQQPPAEGTEQGEEKSESKEVVEQQETPQKDTEKTPVPTTTPTVDVCDTVKLTALADTGS
 NAACSLKYVTGKNYHWRCIAPSGTITSGKDGATCVPPRTQELCLYYLKELSDTTQKQLLEA
 FIKTAAQETIYLLNCKEDHQNETASTHLDIIPCTQLIGGFIPEDFKRMFYTFQDITD
 LFLGRYIGNDLDRVNNNTIAVTFONGTILPNGQKTRQRQEFWGTIGKDIWKGMLCALQEA
 GGKKTLTETYNYSNVTFNGHLTGTKLNEFASRPSFLRWMTEWGDQFCRERITQLQILKER
 CMVYQINGDKGKDDKKEKCTEACTYYKEWLTNWQDNYKKQNRQRYTEVKGTSPYKEDSDVK
 ESKYAHGYLRKILKNIICTSGTDIAY **CNCMEGTS** TDSSNNDNIPESLKYPPIEIEEGCT
 CKDPSPGEVIPEKKVPEPKVLPKPPKLPKRQPKERDFPTPALKNAMLSSTIMWSIGIGFA
 TFTYFYLKKKTSTIDLLRVINIPKSDYDIPTKLSPNRYIPYTSKGYRGKRYIYLEGDSG
 TDSGYTDHYS DITSSSESEYEELDINDIYAPRAPKYKTLIEVVLEPSGNNTTASGNNTPS
 DTQNDIQNDGIPSSKITDNEWNTLKDEFISQYLQSEQPNDVPNDYSSGDIPLNTQPNNTLY
 FDNPDEKPFITSIHDRDLYSGEYSYNVNMVN'TNNDIPISGKNGTYSIDLINDSLNSNN

Note: ORFs in addition to predicted Genes must be searched

30,000 ft View - Proteomics



Overview

PubChem Compound ID: [CID:93072](#)

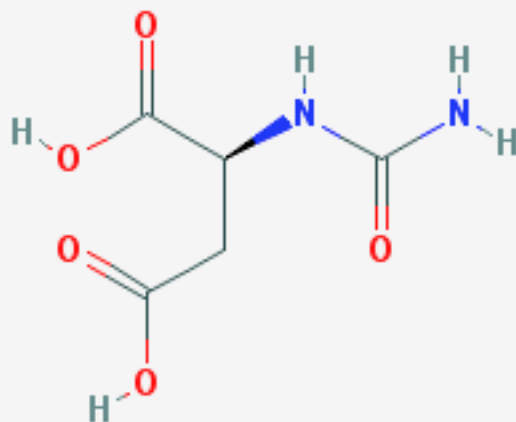
PubChem Substance ID(s): 3727

Synonyms: N-Carbamoyl-L-aspartate

Molecular Weight: 176.12742

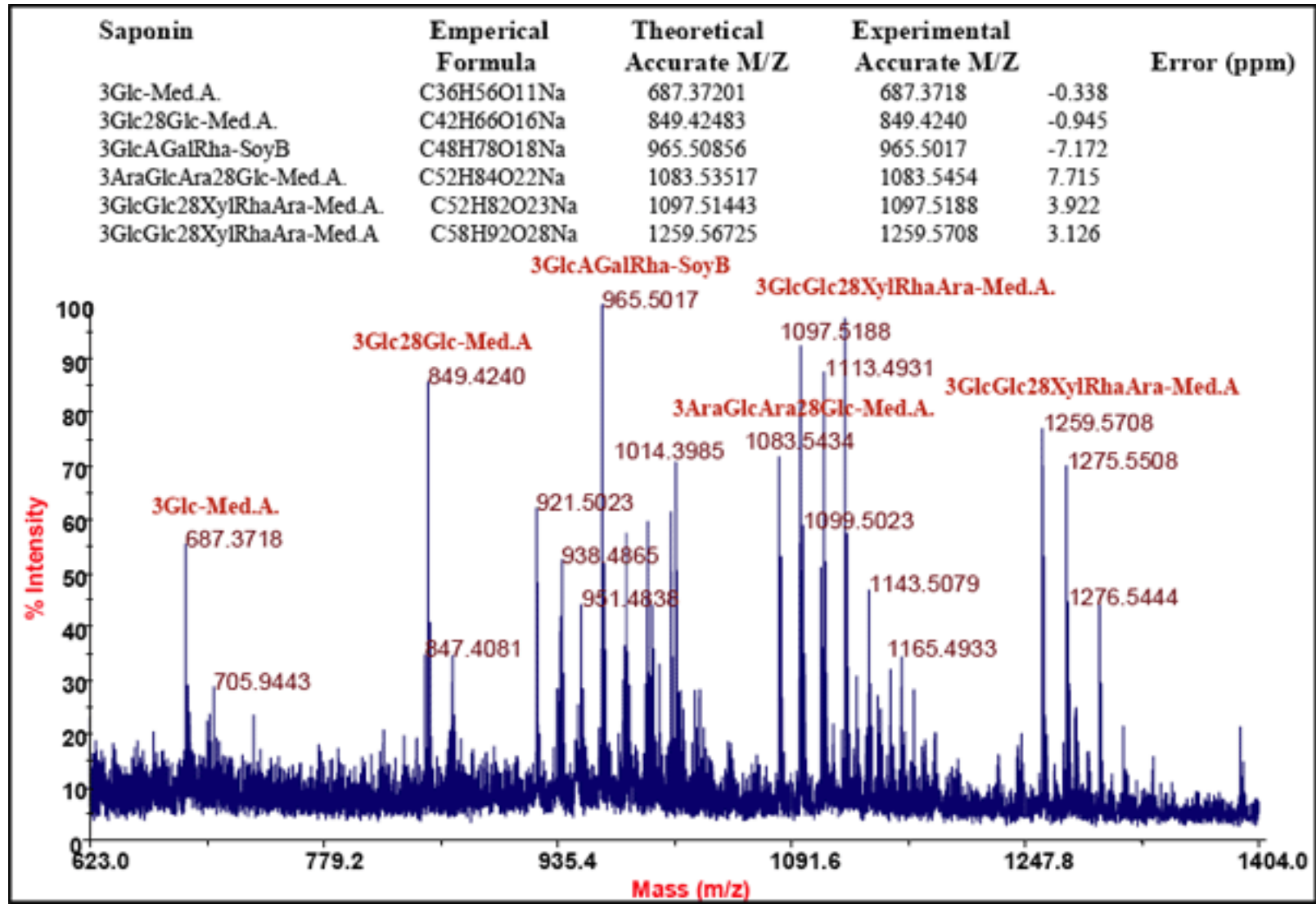
Molecular Formula: $C_5H_8N_2O_5$

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

ALANINE, ASPARTATE AND GLUTAMATE METABOLISM

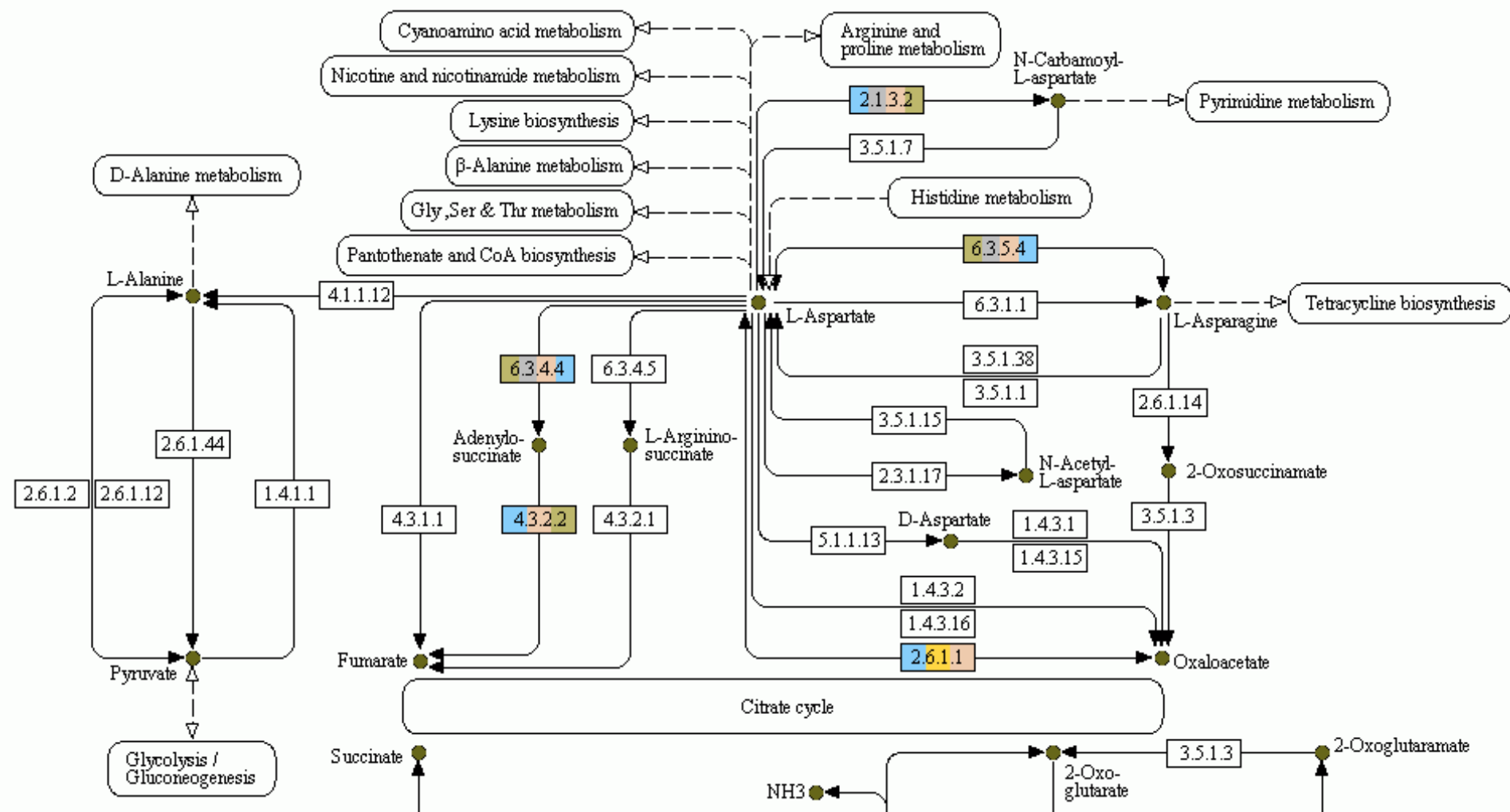
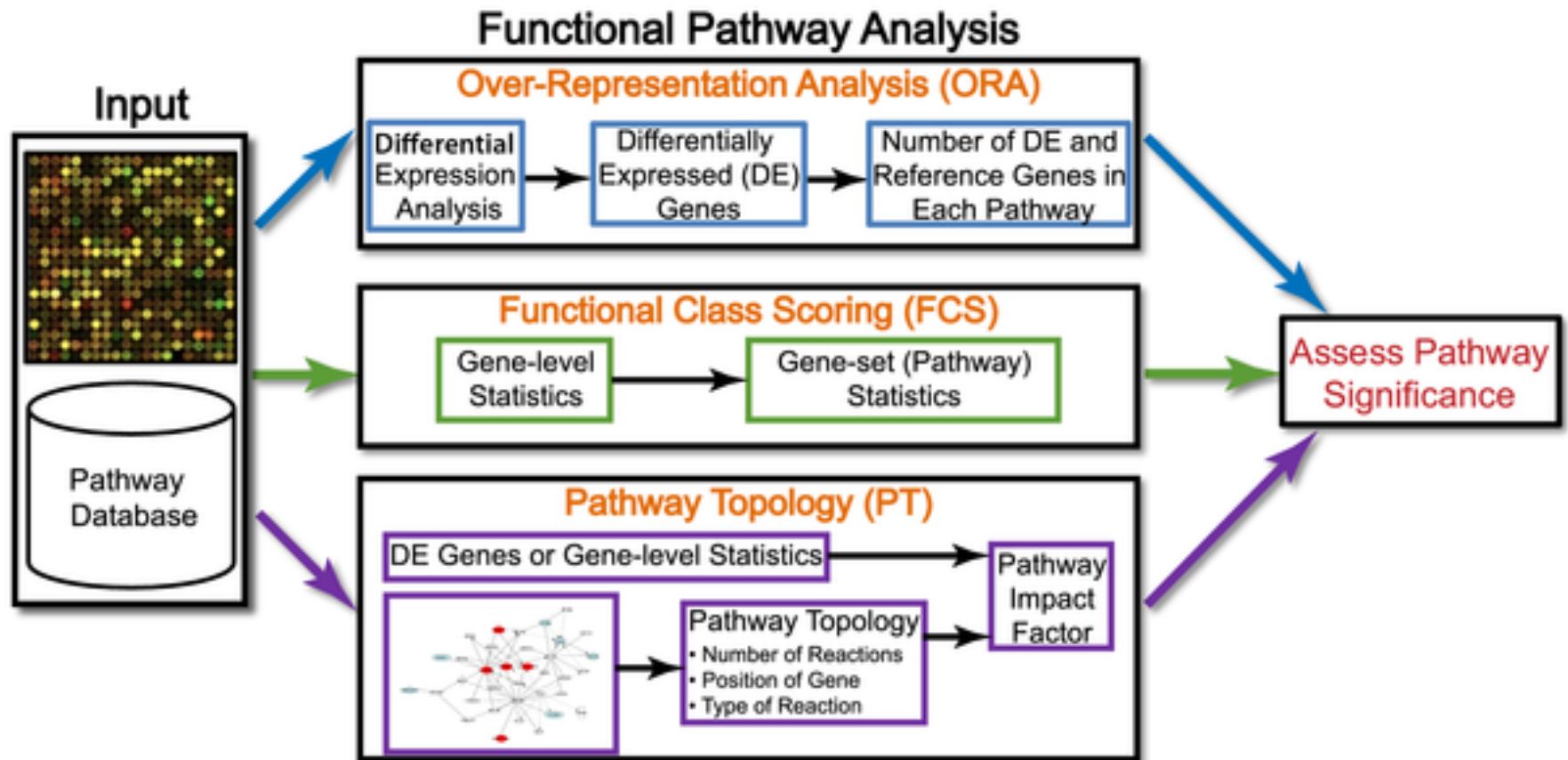


Figure 1. Overview of existing pathway analysis methods using gene expression data as an example.

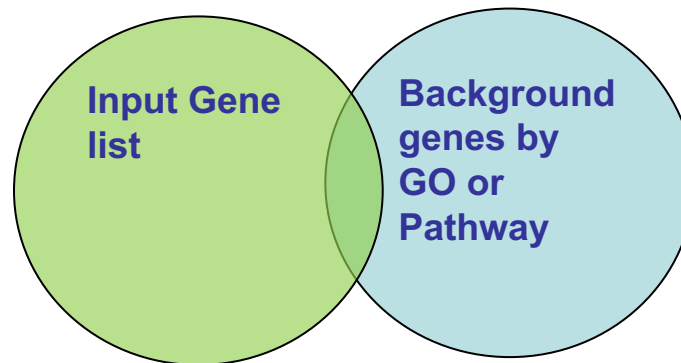


Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology 8(2): e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>
<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375>

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

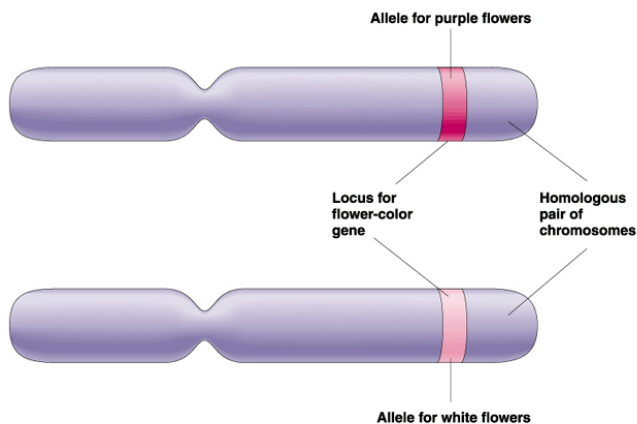
Result: GO:ID or Pathway ID that is enriched

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

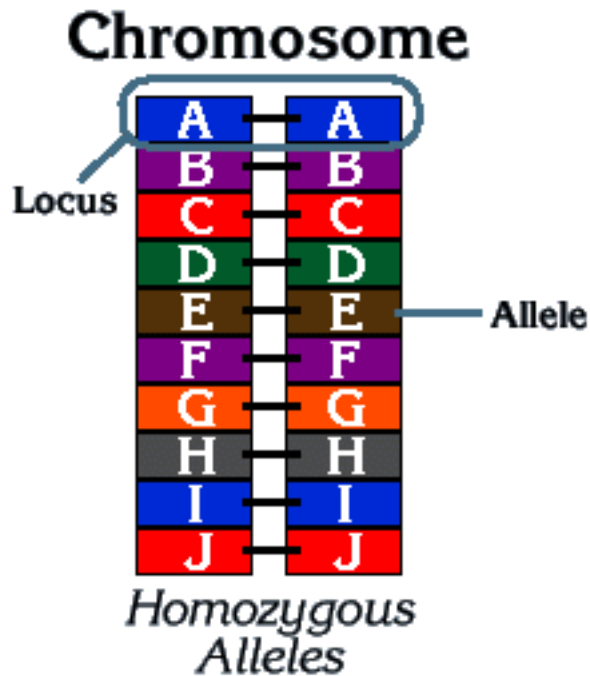
Alleles and Phenotype

- 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)



©1999 Addison Wesley Longman, Inc.



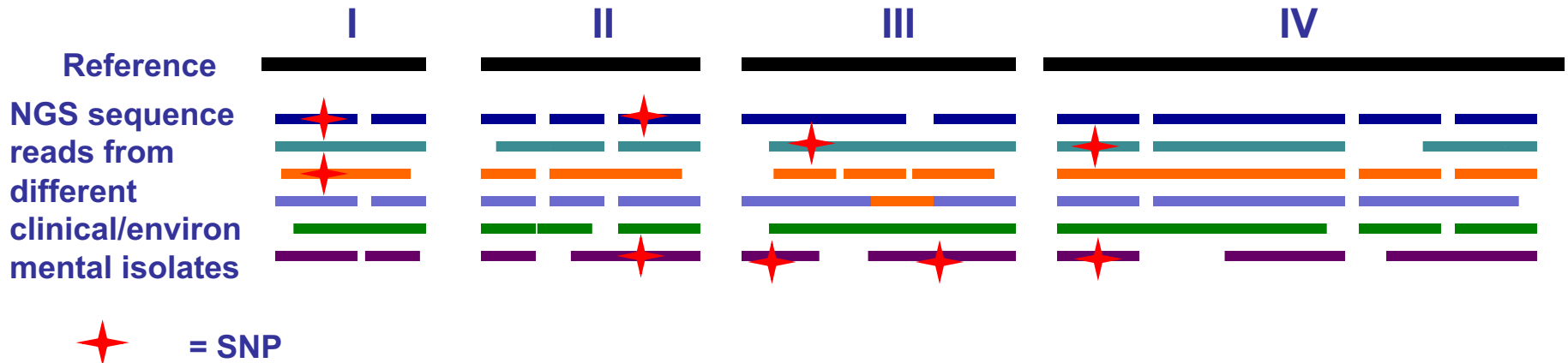
A AAGCCTCATC

a ACGCCTCATC

SNP = Single Nucleotide Polymorphism

30,000 ft View- NGS SNPs

Single Nucleotide Polymorphisms



reference	TGGTGATACT	AAGCTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
303.1	TGGTGATACT	AAGCTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTTA
309.1	TGATAATNCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
RV_3600	TGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
RV_3606	TGATAATNCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
RV_3610	TGATGATTCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT119.09	TGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTCA
SenT123.09	TGATRATTCT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT140.08	TGGTGATACT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CGGTGCTTCA
SenT142.09	TGGTGATACT	AAACTGGGAA	CTCCACTTCC	TTTTCTACTG	CAGTGCTTCA
SenT175.08	TGGTGATACT	AAACTGGGAA	CTCCACTTCT	TTTTCTACTG	CGGTGCTTTA

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

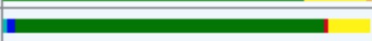
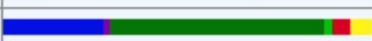
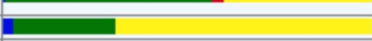
- Chip-Seq
- NGS

Allele Frequencies for Exon 3 48 bp VNTR

Locus [Dexamine receptor 1M](#)

[Information on Histograms](#)

Click on icon  for additional information.

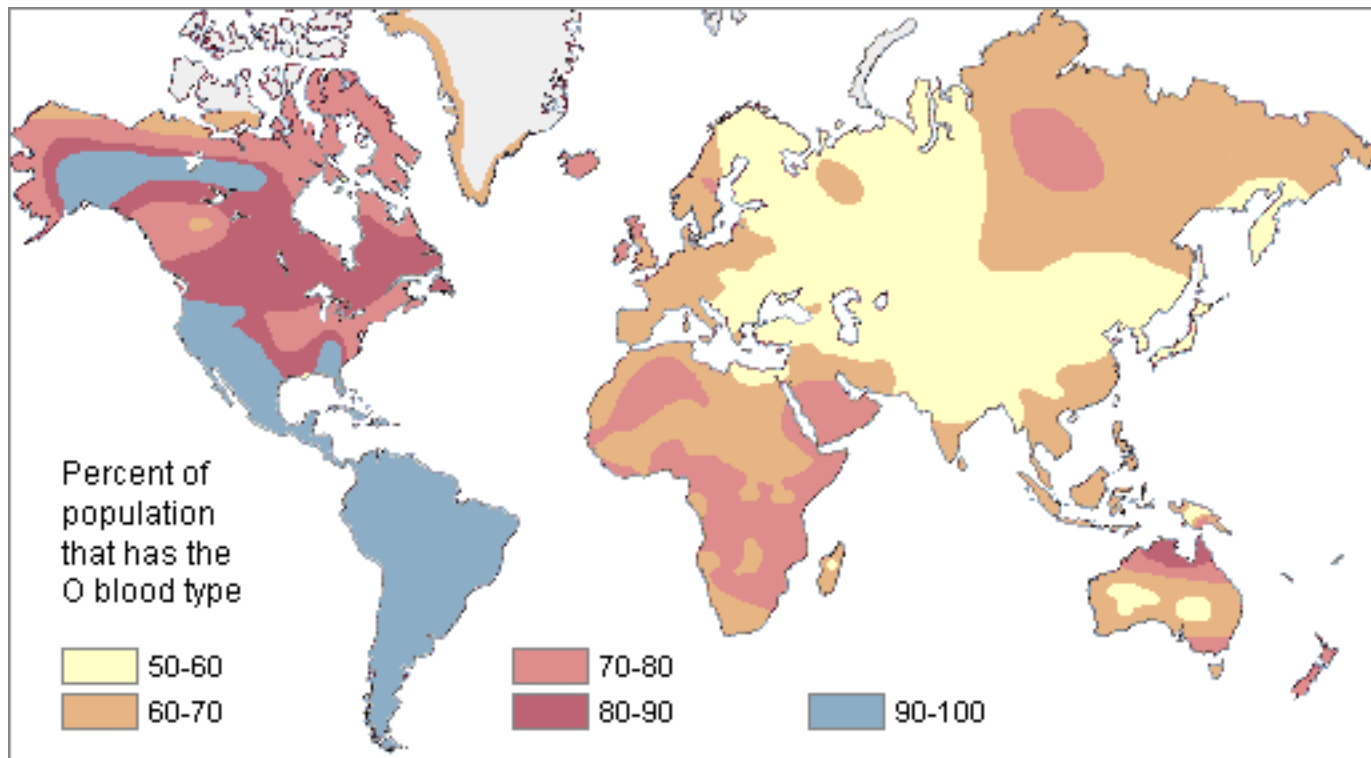
Geographic region	Population (Sample UID, Typed Sample Size (2N), entry date) Add Info	
Africa	Baka(SA000005F) , 134, 9/19/2000, 	
Africa	Mbuti(SA000006G) , 72, 9/19/2000, 	
Africa	Jews, Ethiopian(SA000042N) , 128, 9/19/2000, 	
Europe	Druze(SA000047L) , 130, 9/19/2000, 	
Europe	Jews, Yemenite(SA000016H) , 80, 9/19/2000, 	
Europe	Samaritans(SA000098R) , 78, 9/19/2000, 	
Europe	Adyghej(SA000017I) , 104, 9/19/2000, 	
Europe	Conej(SA000007H) , 100, 9/19/2000, 	
Europe	Europeans, Mixed(SA000020C) , 176, 9/19/2000, 	
Europe	Europeans, Mixed(SA001775U) , 146, 6/21/2006, 	
Europe	Finn(SA000018I) , 66, 9/19/2000, 	
Asia	Brahmin(SA001831N) , 156, 8/23/2006, 	
Asia	Keralite(SA001833P) , 148, 8/23/2006, 	
Asia	Keralite(SA001834Q) , 114, 8/23/2006, 	
Asia	Keralite(SA001835R) , 130, 8/23/2006, 	
Asia	Marathas(SA001832Q) , 116, 8/23/2006, 	
Asia	Kachan(SA000040E) , 36, 9/19/2000, 	
East Asia	Ami(SA000002C) , 80, 9/19/2000, 	
East Asia	Abnaji(SA000021D) , 84, 9/19/2000, 	
East Asia	Han(SA000000J) , 96, 9/19/2000, 	
East Asia	Japanese(SA000010B) , 100, 9/19/2000, 	
East Asia	Cambodians, Khmer(SA000023E) , 50, 9/19/2000, 	
East Asia	Hakka(SA000003D) , 32, 9/19/2000, 	
Oceania	Melanesian, Nasioi(SA000012D) , 46, 9/19/2000, 	
Oceania	Micronesians(SA000067I) , 58, 9/19/2000, 	
Siberia	Yakut(SA000011C) , 92, 9/19/2000, 	
North America	Cheyenne(SA0000023F) , 96, 9/19/2000, 	
North America	Pima, Arizona(SA000025H) , 94, 9/19/2000,	
North America	Pima, Mexico(SA000026I) , 104, 9/19/2000,	
North America	Southern Amerindians(SA000024G) , 86, 9/19/2000,	
North America	Maya, Yucatec(SA000013E) , 100, 9/19/2000,	
South America	Gaviao(SA000148N) , 58, 8/9/2005,	
South America	Karitiana(SA000028K) , 108, 9/19/2000,	
South America	Surui(SA000014F) , 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

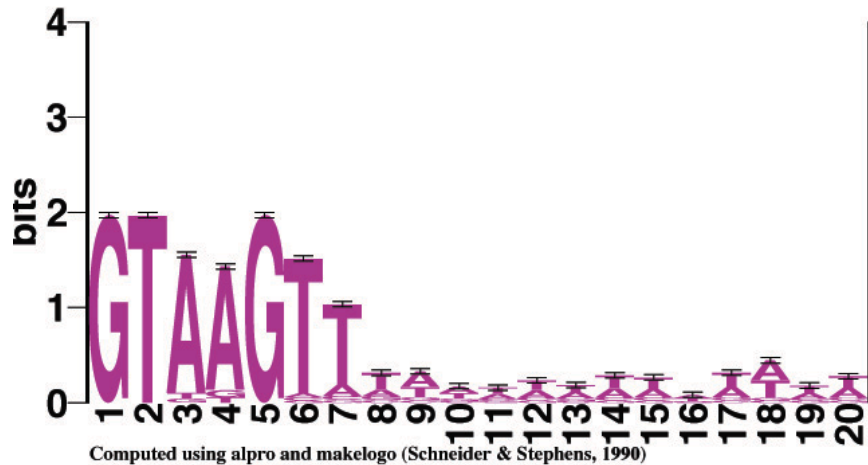
- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats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

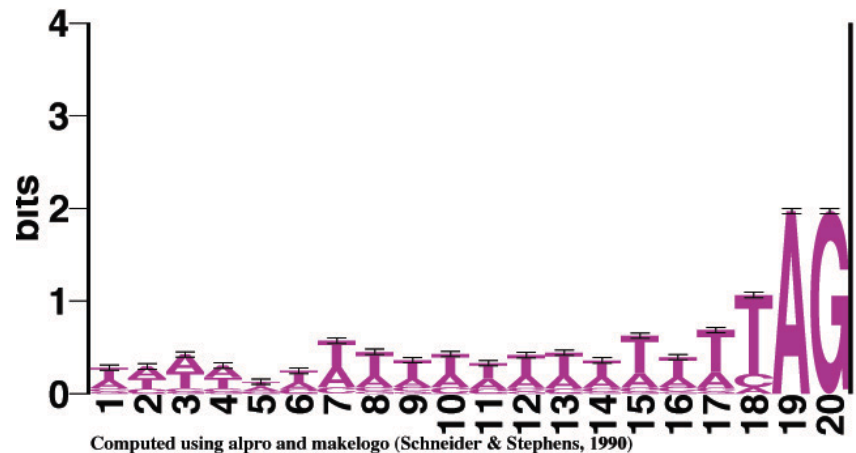
Donor Site

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

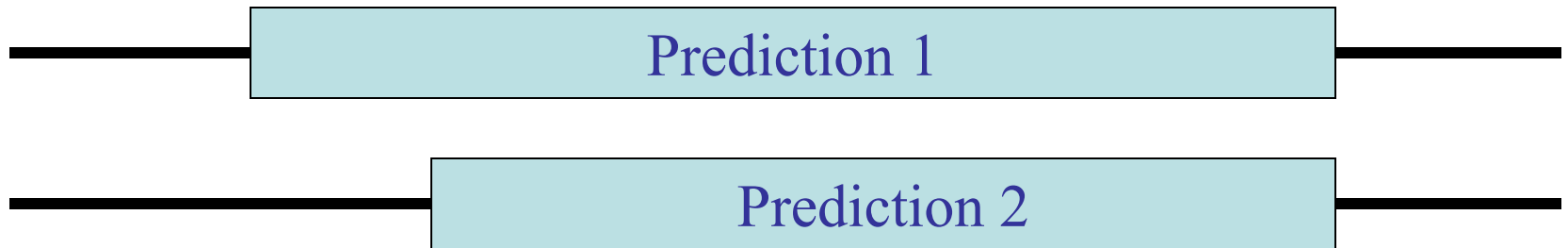


Acceptor site

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)
- 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 all week and we are happy to talk to you.
- These slides are posted on the course Schedule for your review