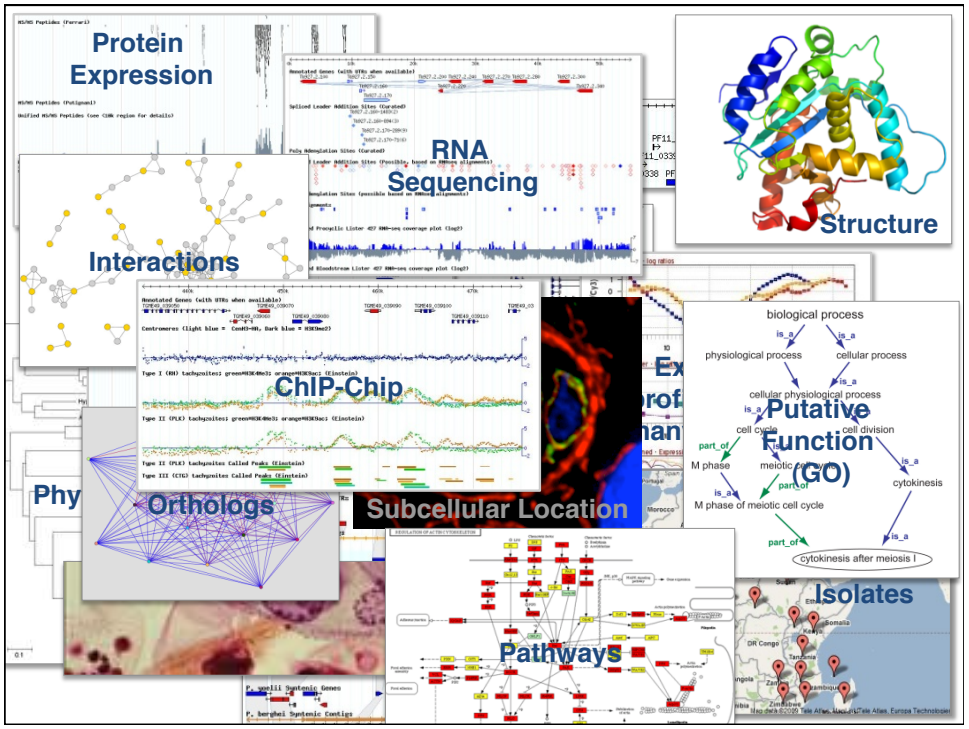
Welcome!

EuPathDB Workshop 2012

Crash Course in Omics Terminology, Concepts & Data Types

Jessie Kissinger

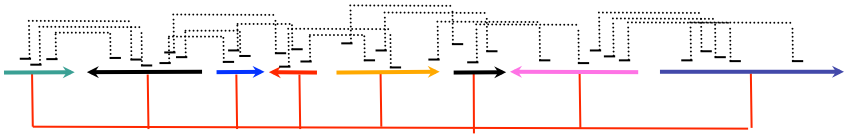
June 17, 2012



Genome assembly

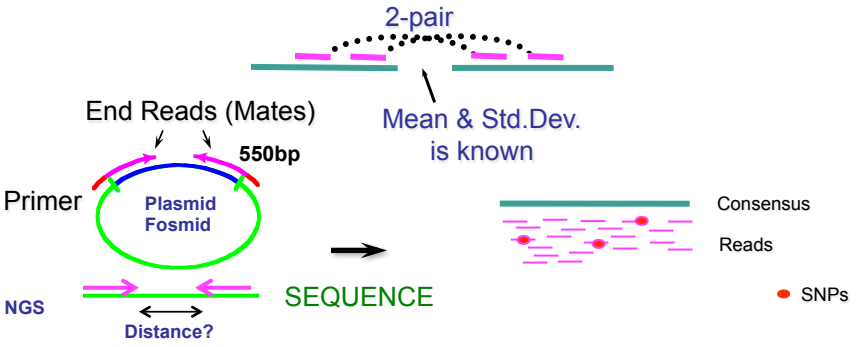
- 5X random genome shotgun
- Library insert size
- Paired-end? (Mated end pairs)
- Contigs
- Scaffolds

Pairs Give Order & Orientation

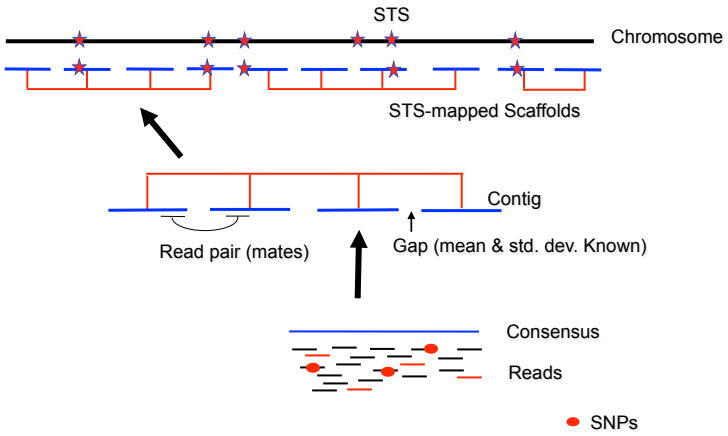


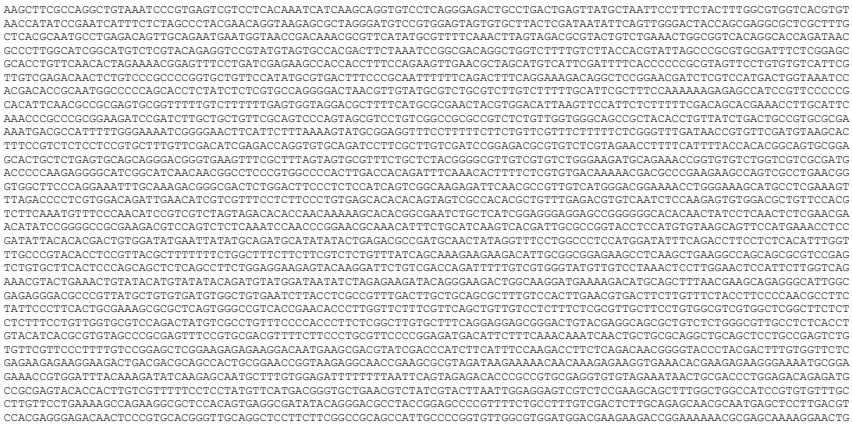
Scaffold

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

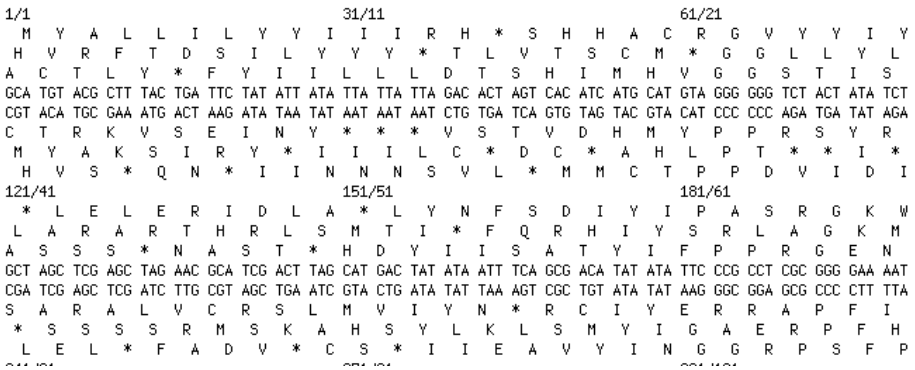


Anatomy of a WGS Assembly

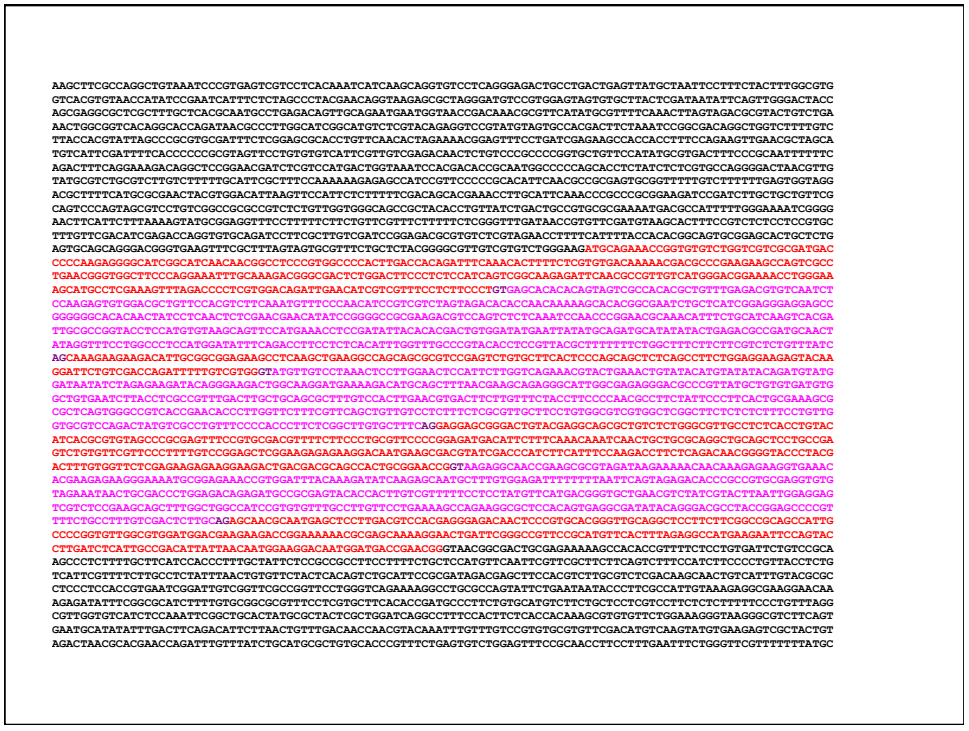


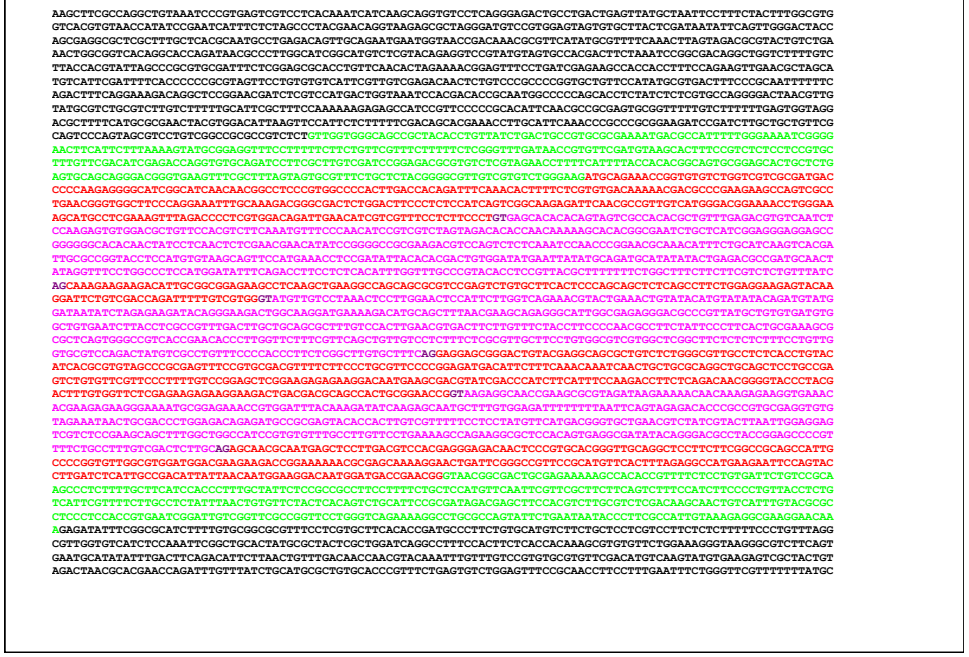


Six Frame Translation ORF-finding



ORFs ≠ Genes

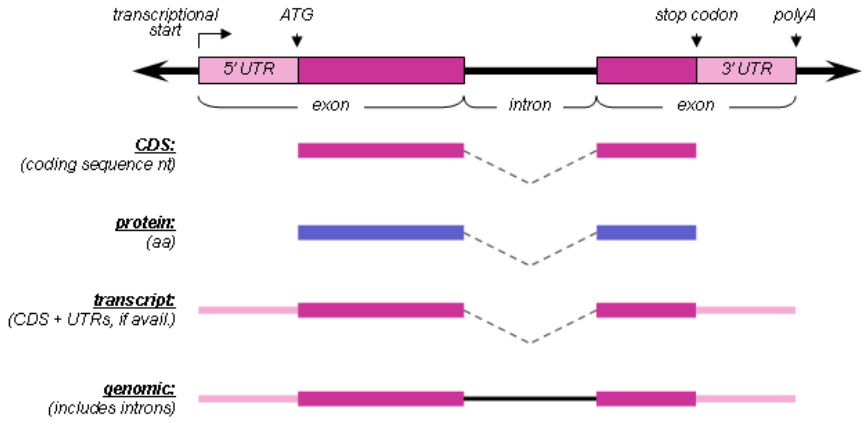




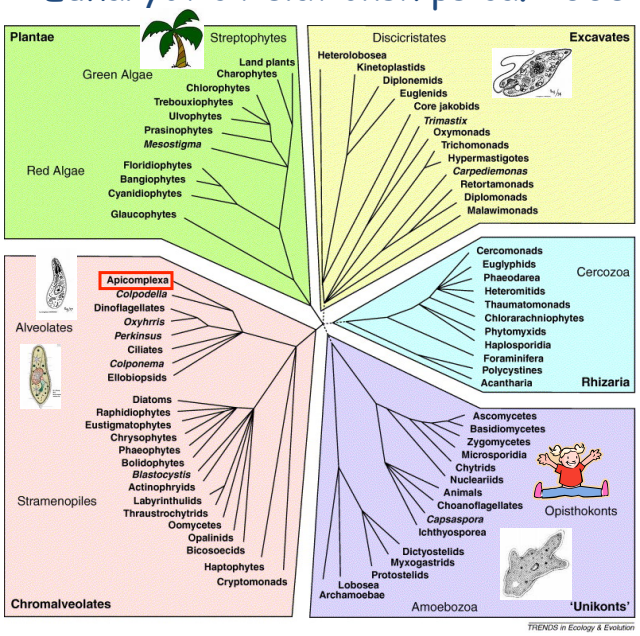
```
ATGCAGAAACCGGTGTGTCTGGTCGTCGCGATGACCCCCAAGAGGGGCATCGGCATCAACAACGGCCTCCCGTGGCCCC
ACTTGACCACAGATTTCAAACACTTTTCTCGTGTGACAAAAACGACGCCCGAAGAAGCCAGTCGCCTGAACGGGTGGCT
TCCCAGGAAATTTGCAAAGACGGGCGACTCTGGACTTCCCTCTCCATCAGTCGGCAAGAGATTCAACGCCGTTGTCATG
GGACGGAAAACCTGGGAAAGCATGCCTCGAAAGTTTAGACCCCTCGTGGACAGATTGAACATCGTCGTTTCCTCTTCCC
TCAAAGAAGAAGACATTGCGGCGGAGAAGCCTCAAGCTGAAGGCCAGCAGCGCGTCCGAGTCTGTGCTTCACTCCCAGC
AGCTCTCAGCCTTCTGGAGGAAGAGTACAAGGATTCTGTCGACCAGATTTTTGTCGTGGGAGGAGCGGGACTGTACGAG
GCAGCGCTGTCTCTGGGCGTTGCCTCTCACCTGTACATCACGCGTGTAGCCCGCGAGTTTCCGTGCGACGTTTTCTTCC
CTGCGTTCCCCGGAGATGACATTCTTTCAAACAAATCAACTGCTGCGCAGGCTGCAGCTCCTGCCGAGTCTGTGTTCGT
TCCCTTTTGTCCGGAGCTCGGAAGAGAGAAGGACAATGAAGCGACGTATCGACCCATCTTCATTTCCAAGACCTTCTCA
GACAACGGGGTACCCTACGACTTTGTGGTTCTCGAGAAGAGAAGGAAGACTGACGACGCAGCCACTGCGGAACCGAGCA
ACGCAATGAGCTCCTTGACGTCCACGAGGGAGACAACTCCCGTGCACGGGTTGCAGGCTCCTTCTTCGGCCGCAGCCAT
TGCCCCGGTGTTGGCGTGGATGGACGAAGAAGACCGGAAAAAACGCGAGCAAAAGGAACTGATTCGGGCCGTTCCGCAT
GTTCACTTTAGAGGCCATGAAGAATTCCAGTACCTTGATCTCATTGCCGACATTATTAACAATGGAAGGACAATGGATG
ACCGAACGG
```

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

Terminology



Eukaryotic Relationships ca. 2005



Keeling, 2005

Synteny = large regions of chromosomes containing the same genes

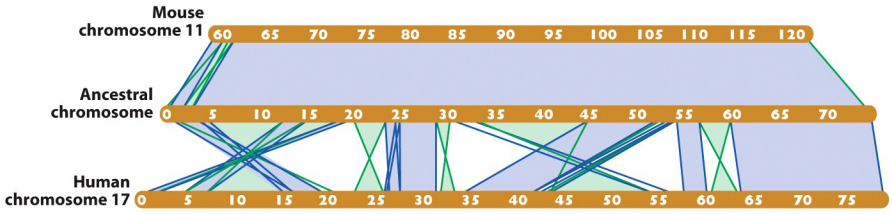
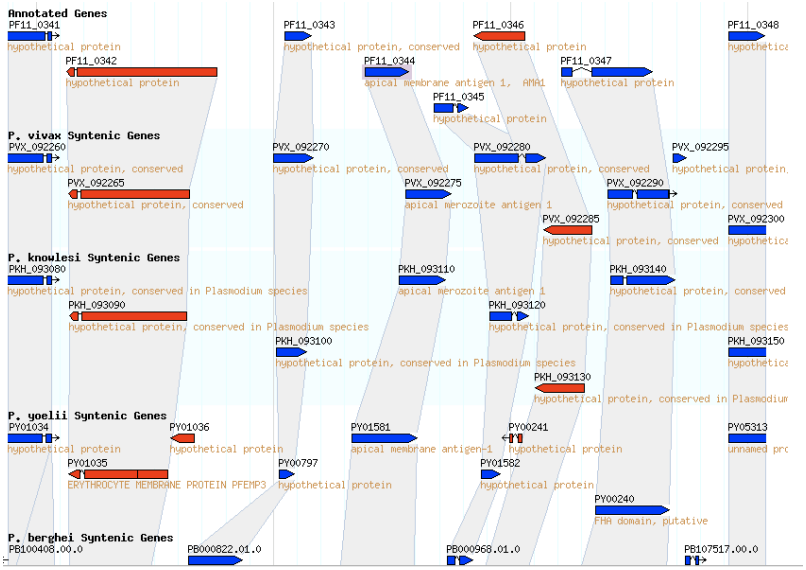
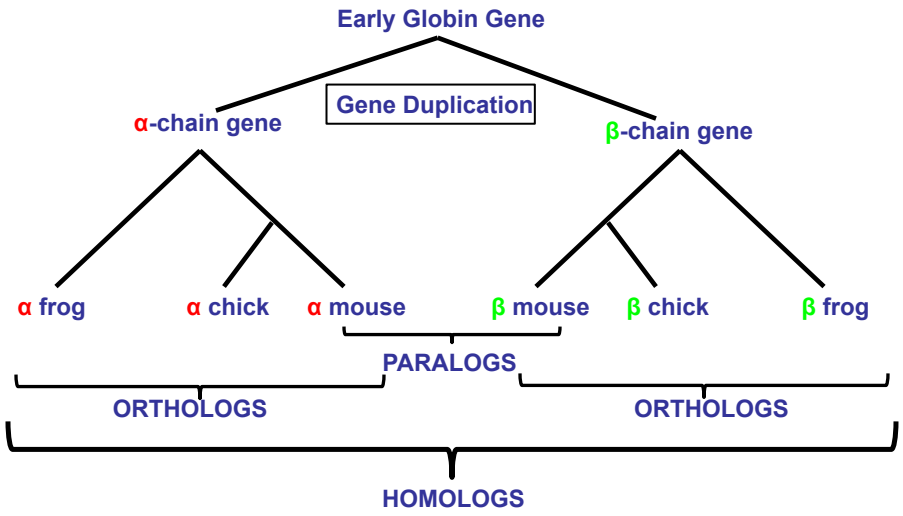


Figure 13-15
Introduction to Genetic Analysis, Ninth Edition

Synteny among Plasmodia



Homology



Evolutionary relationships

- Homology - related by evolutionary descent not equivalent to similarity
- Orthology - same gene in different organisms, e.g. alpha hemoglobin in humans and chimps
- Paralogy - genes within an organism related by gene duplication, e.g. alpha and beta hemoglobin in humans
- Xenology - genes related by gene transfer

RNA sequence/Expression Data

Data

- cDNA
- Expressed Sequence Tags (EST)
- RNA-Seq
- Microarray
- Ditag (SAGE-tags)
- Small RNA's (various types)

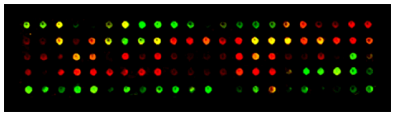
Technolgy

- Sanger
- Next-gen (454, Illumina, etc)
- Microarray-slides
- Microarray-chips
- SAGE

Expression Profiles

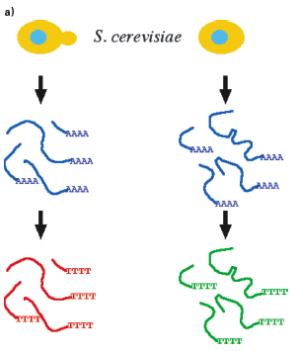
- The pattern of expression of one or more genes over time or a set of experimental conditions, e.g. during development or a drug treatment or in a genetic mutant such as a gene knock-out.
- Always… has a time and space component

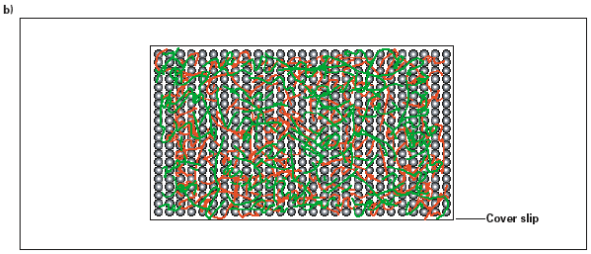
Microarrays



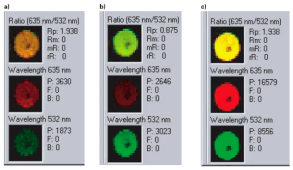
- cDNA microarrays
- "GeneChip" in situ synthesized oligonucleotide arrays
- Oligomer (~70mer) arrays

Experiments are almost always
Competitions between conditions or stages



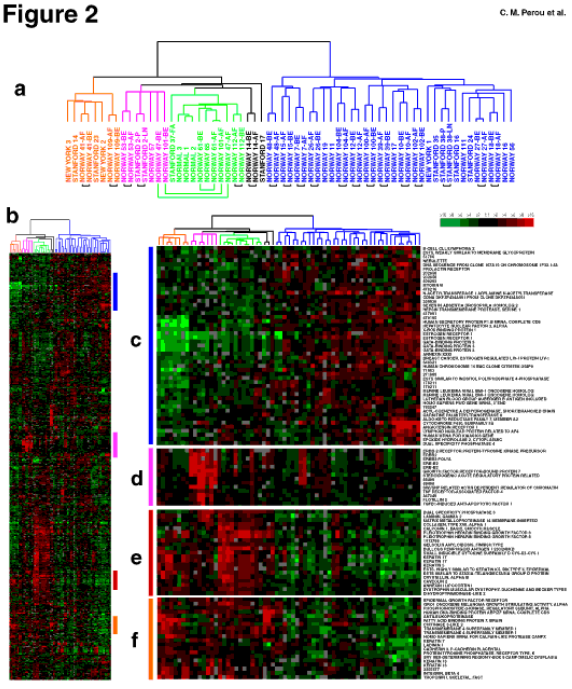


The RNA samples from the test and the control are labeled with different colors in a reverse-transcription reaction and then hybridized, together, competitively to a slide or chip containing gene sequences in multiple copies.



Ratios of experimental to control expression are often expressed as colors rather than numbers

Clustered Microarray Data Genes with Similar Expression Profiles are Grouped together



Other RNA expression

- 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)
  - Primary purpose is relative levels of gene expression
- RNA-Seq (NGS)
  - Little sequence bias
  - Quantitative
  - Can be strand-specific

Genes can be located on either DNA strand

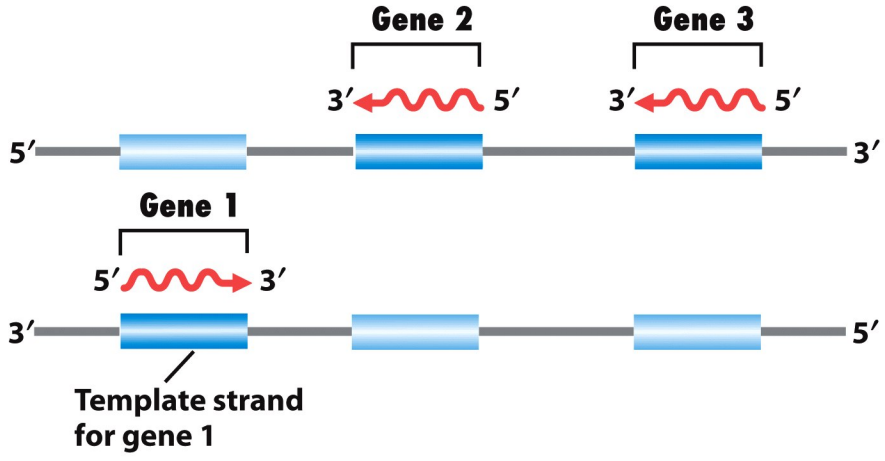


Figure 8-3
Introduction to Genetic Analysis, Ninth Edition

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

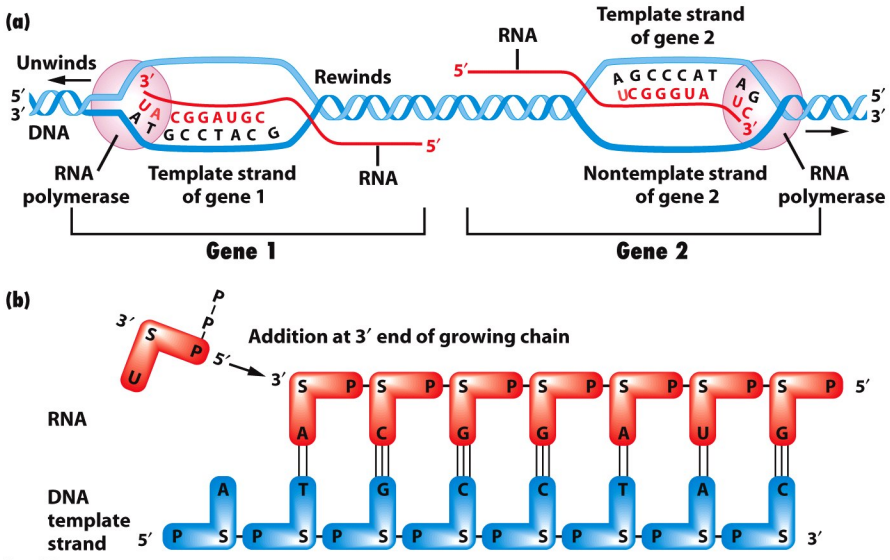


Figure 8-4
Introduction to Genetic Analysis, Ninth Edition

Sequences of DNA and transcribed RNA

Convention

Gene location = non-template strand,
i.e. same as the mRNA

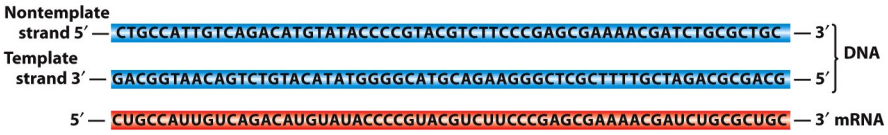


Figure 8-6
Introduction to Genetic Analysis, Ninth Edition

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

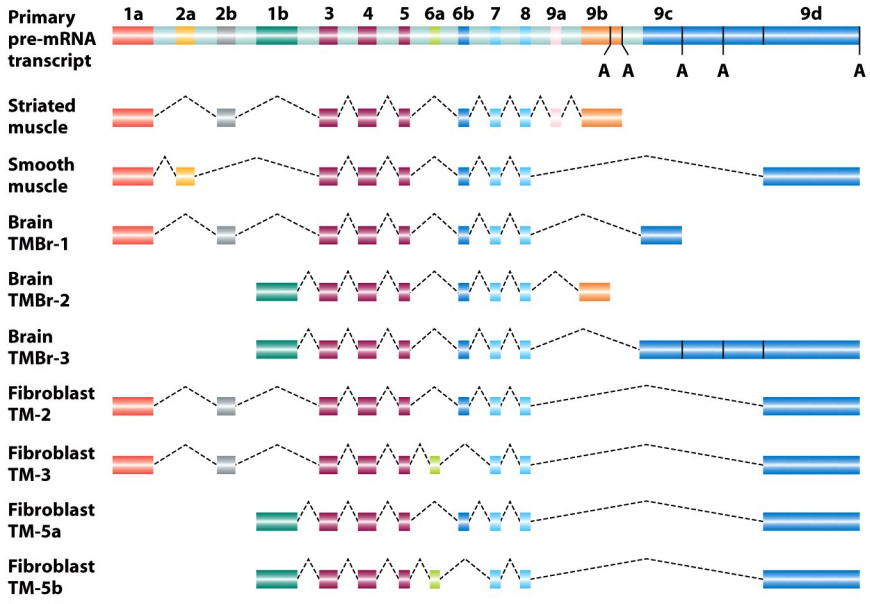


Figure 8-14
Introduction to Genetic Analysis, Ninth Edition

Bioinformatics uses algorithms

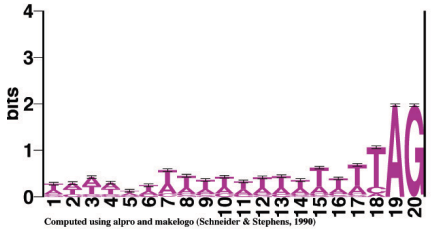
- Algorithms are sets of rules for solving problems or identifying patterns
- Algorithms can be general or case specific and often need to be trained
- Computational analysis, like wet-bench analyses are only as good as the tools, techniques and material allow, and all interpretations come with caveats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

How to find an intron

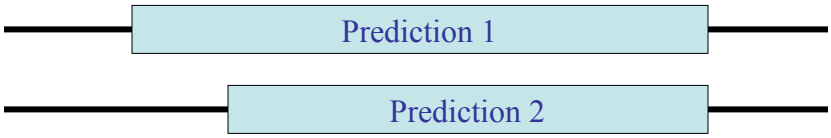
- Usually begins with GT and end with AG
- Must be longer than 19 nucleotides
- Must contain a branchpoint "A"
- Donor GT often followed by a sequence pattern. This pattern is species-specific
- Acceptor AG often proceeded by pyrimidine stretch
- Has a mean length of "X" as is observed in this species

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

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



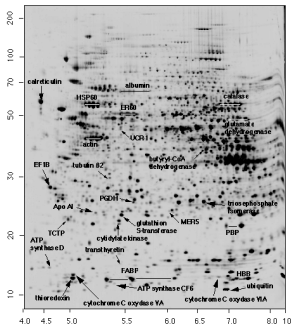
Different prediction methods often generate different results



Protein Expression/Sequence

Data

- MW-Isoelectric point
- MW
- Sequence/spans



Technology

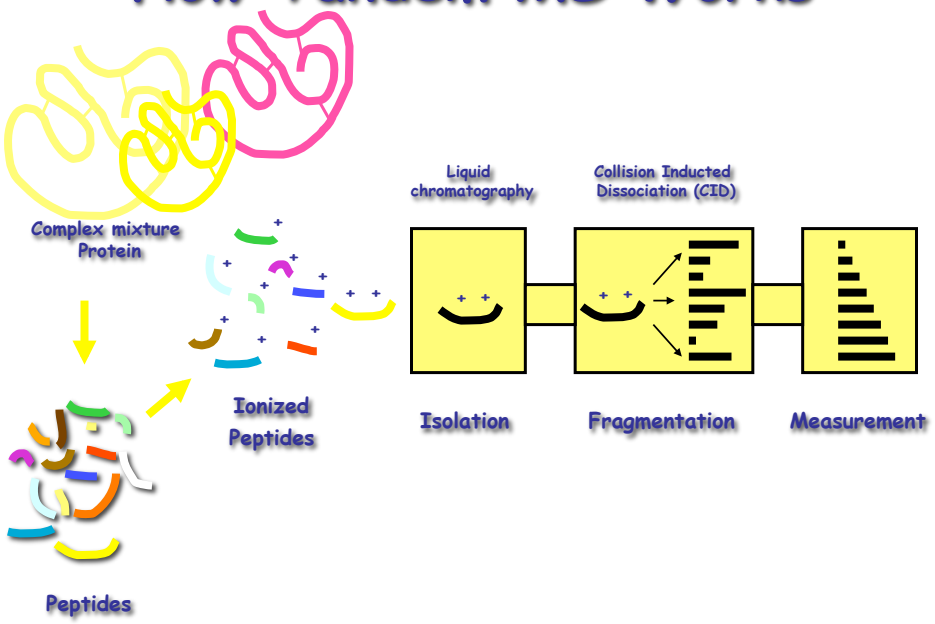
- 2D gel electrophoresis
- Mass spectrometry
- Tandem MS (MS-MS, LC MS-MS etc)

Typical 2 D gel

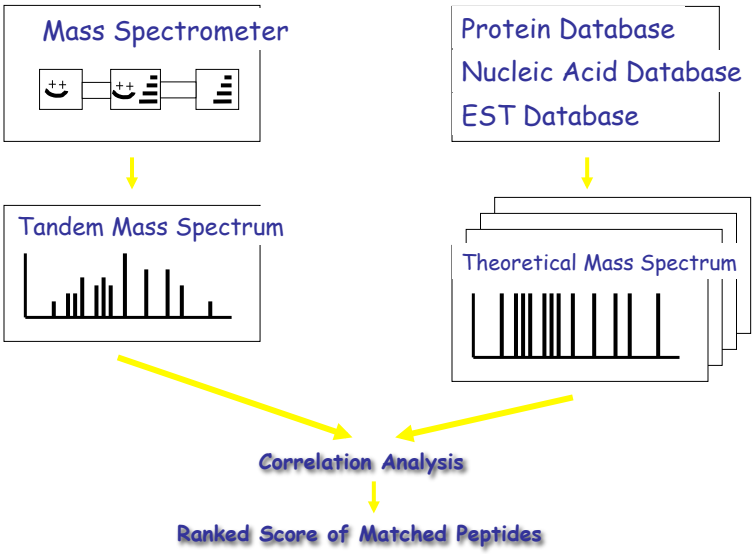
High throughput mass spectrometry

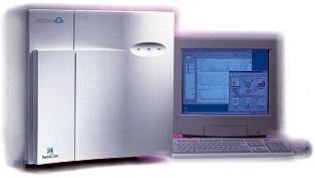
- Direct identification of proteins from biological sample.
- Capillary liquid chromatography apparatus (LC) coupled with…
- Electrospray tandem mass spectroscopy (MS/MS)
- "Sequest", Mascot, or other software links mass spectra with genomic sequence database.

How Tandem MS Works



Sequest Database Search

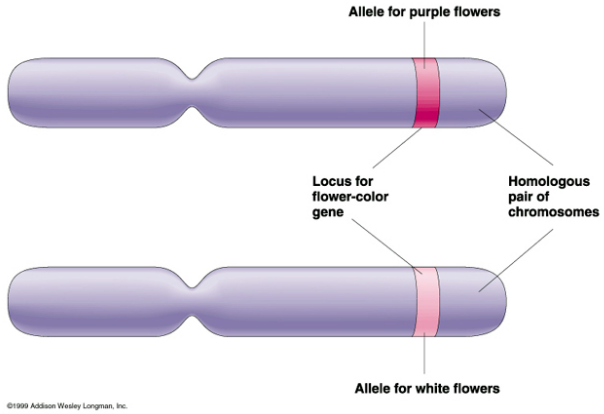




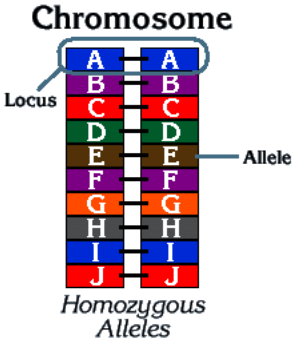
Peptide database

Note: ORFs in addition to predicted Genes must be searched

Homologous chromsomes (in a diploid)



Loci, alleles and SNPs in a population



A AAGCCTCATC

a ACGCCTCATC

SNP =Single Nucleotide Polymorphism

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)

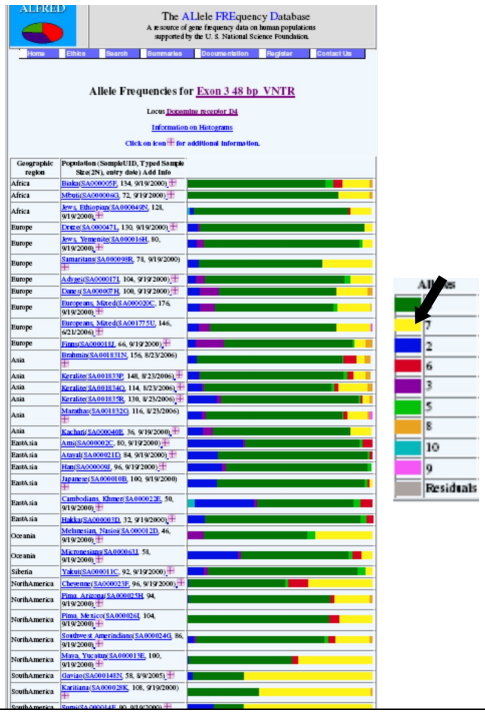
Population data

Data

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

Technology

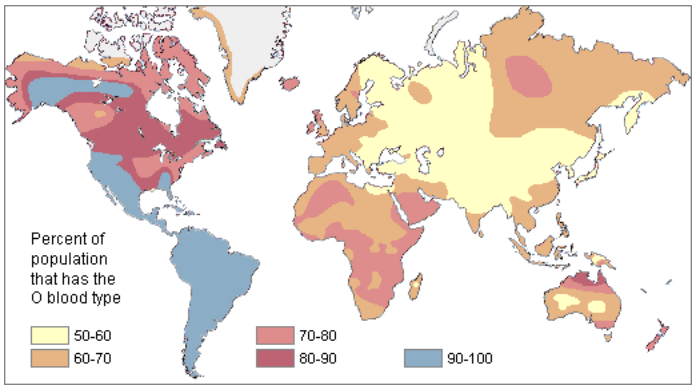
- Chip-Seq
- NGS



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

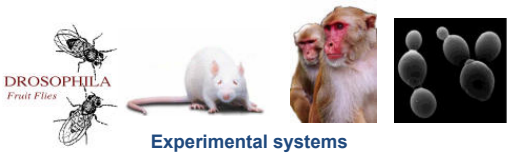
Data

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

Technology

- PCR-RFLP
- Sequencing
- SNP chip
- GPS

Infectious Disease Paradigm



Experimental systems