Genetics and Genomics

Study of Omics

Study of Omics

Study of Omics

- Genomics: all the genes
- Pharmacogenomics choice of personalized medicine
- Nutrigenomics choice of best diet
- Toxicogenomics prediction of toxicity

Study of Omics

Study of Omics

- Epigenomics: all epigenetic changes in genome
- Transcriptomics: all the mRNAs
- Proteomics : all the proteins

Study of Omics

Study of Omics

- Interactomics : all interactions between all proteins
- Metabolomics (or metabonomics): all metabolites

Genetics and Genomics

Genomics, Proteomics and Metabolomics

Genomics, Proteomics, Metabolomics

Genome and Genomics

- The complete set of DNA found in each cell is known as the genome and study is called as genomics.

Genomics, Proteomics, Metabolomics

Proteome and Proteomics

- The complete set of proteins found in each cell is known as the proteome.

Genomics, Proteomics, Metabolomics

Proteome and Proteomics

- Proteins concentration (and activity) may be different than gene expression due to post-translational modification

Genomics, Proteomics, Metabolomics

Metabolomics

- The complete set of metabolites found in each cell is known as the metabolome.
- Use of high-throughput mass spectrometry to analyze the metabolic components of cell.

Genomics, Proteomics, Metabolomics

Metabolomics

- Useful for determining the effects of the environment or gene transformation on the metabolism of the plants/animals.

Genomics, Proteomics, Metabolomics

Conclusion

- Genomics, proteomics and metabolomics will give an integrated, wholistic view of the cell.
- Can be used to monitor or modify organisms in a comprehensive way.

Genomics, Proteomics, Metabolomics

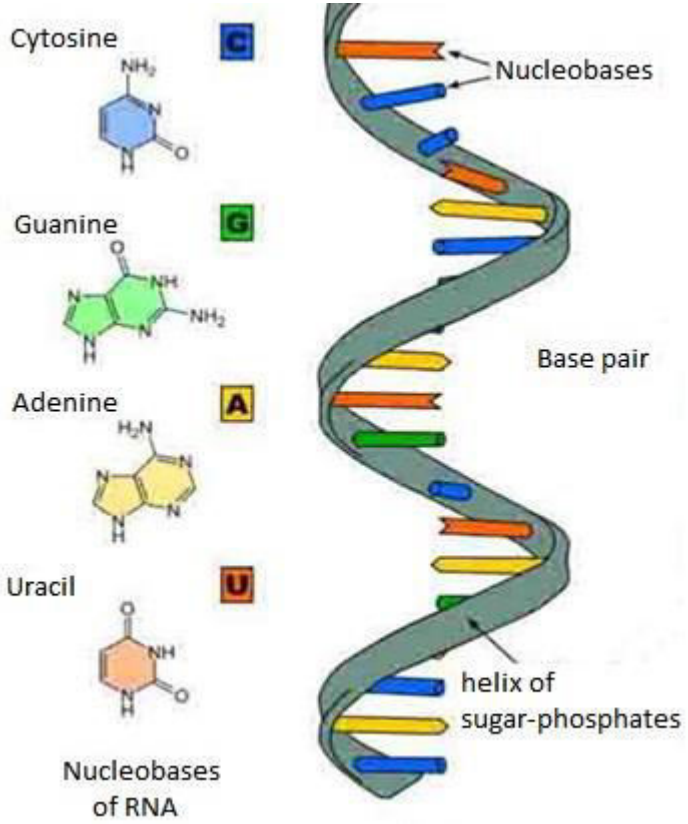
Conclusion

- Bioinformatics - the key to understand the plethora of information and modeling the cell.

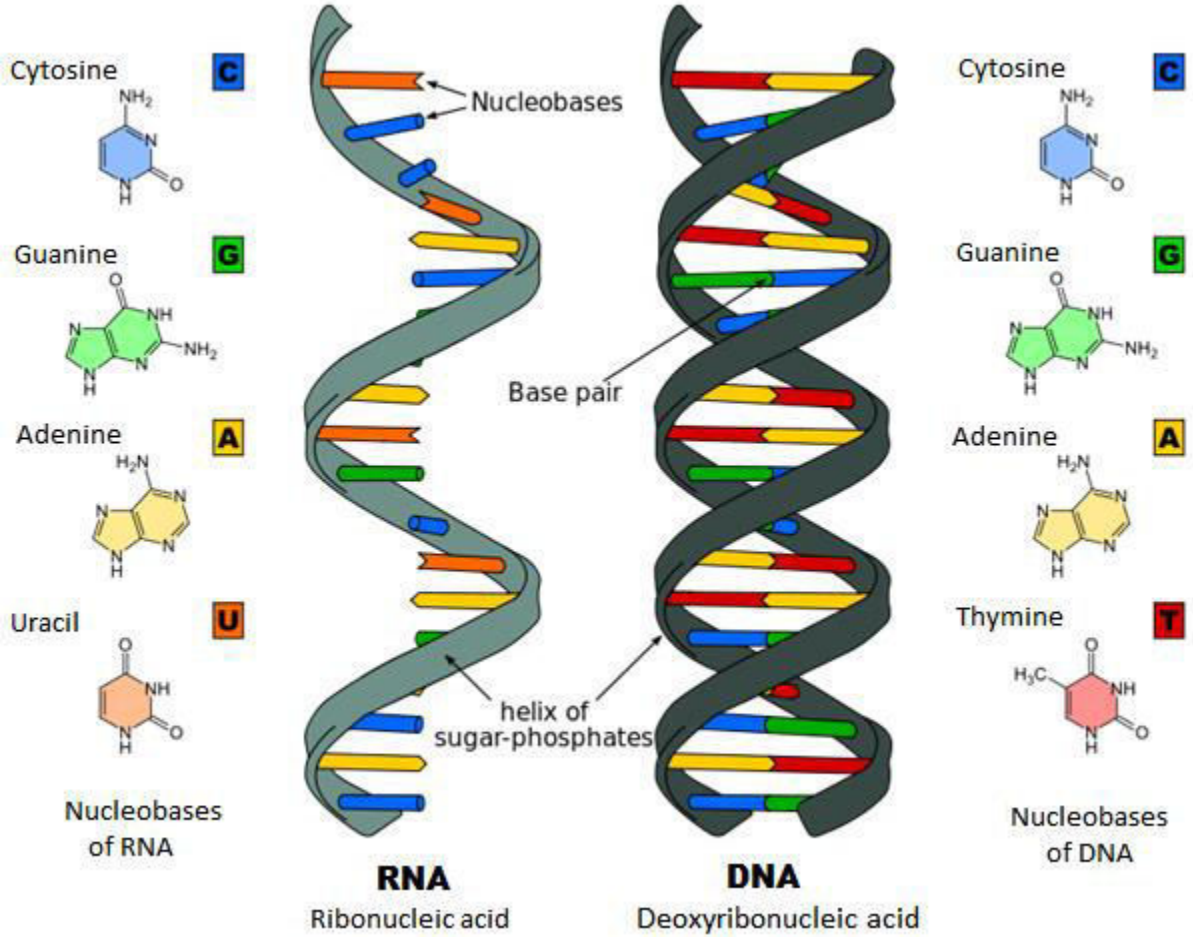
Advanced Computing Approaches

Structure of RNA

Structure of RNA



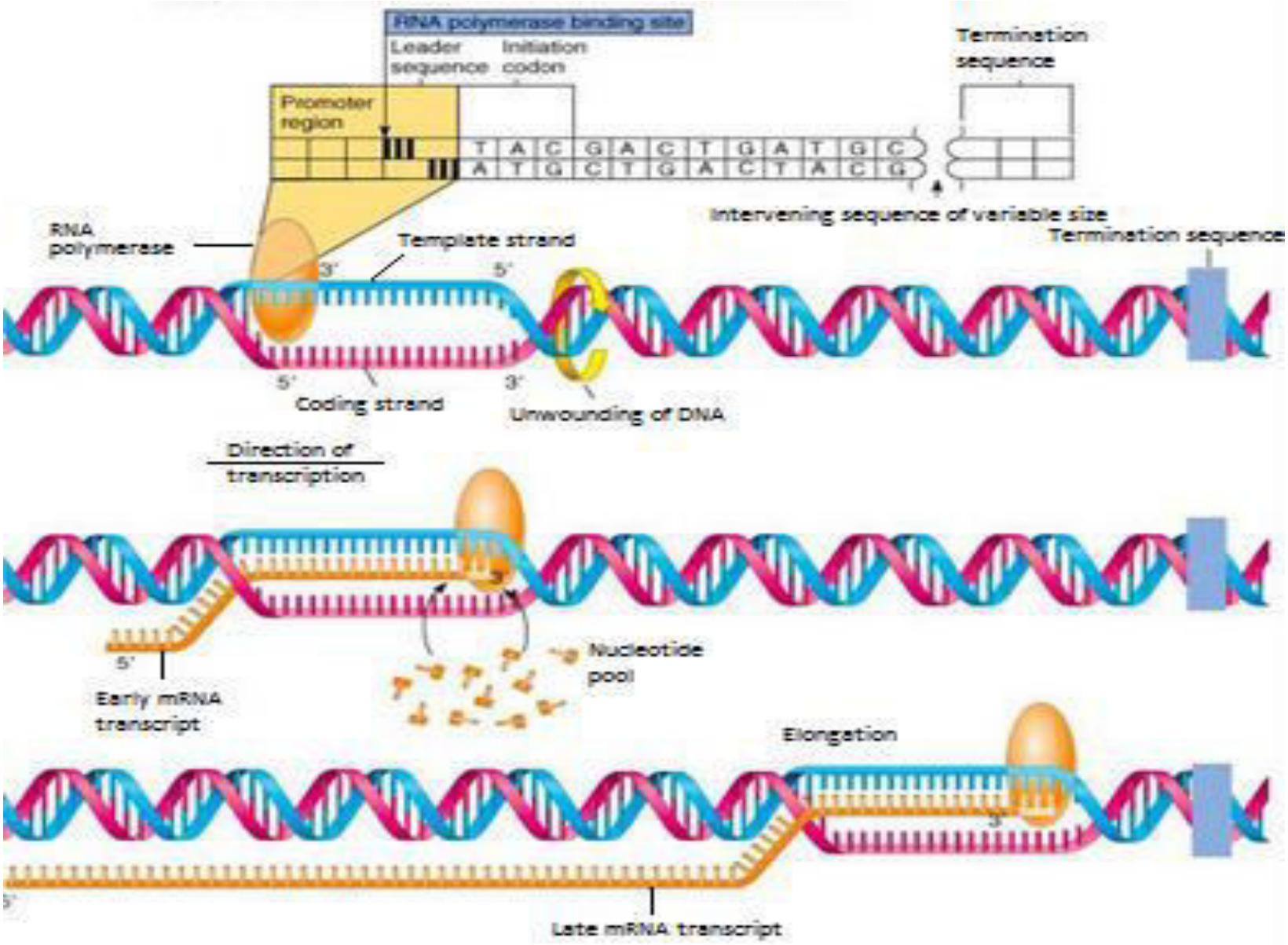
Structure of RNA



DNA Transcription

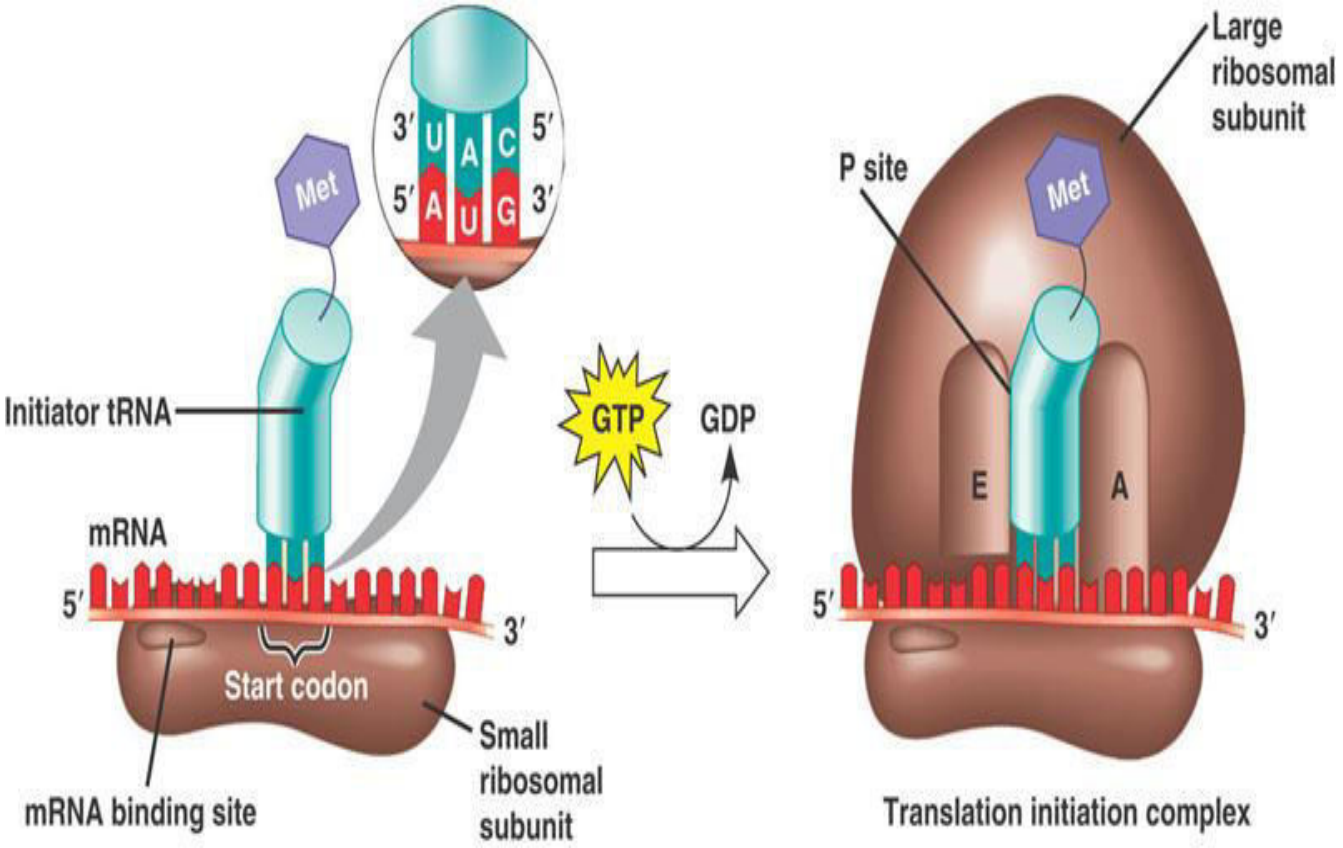
Transcription

DNA Transcription

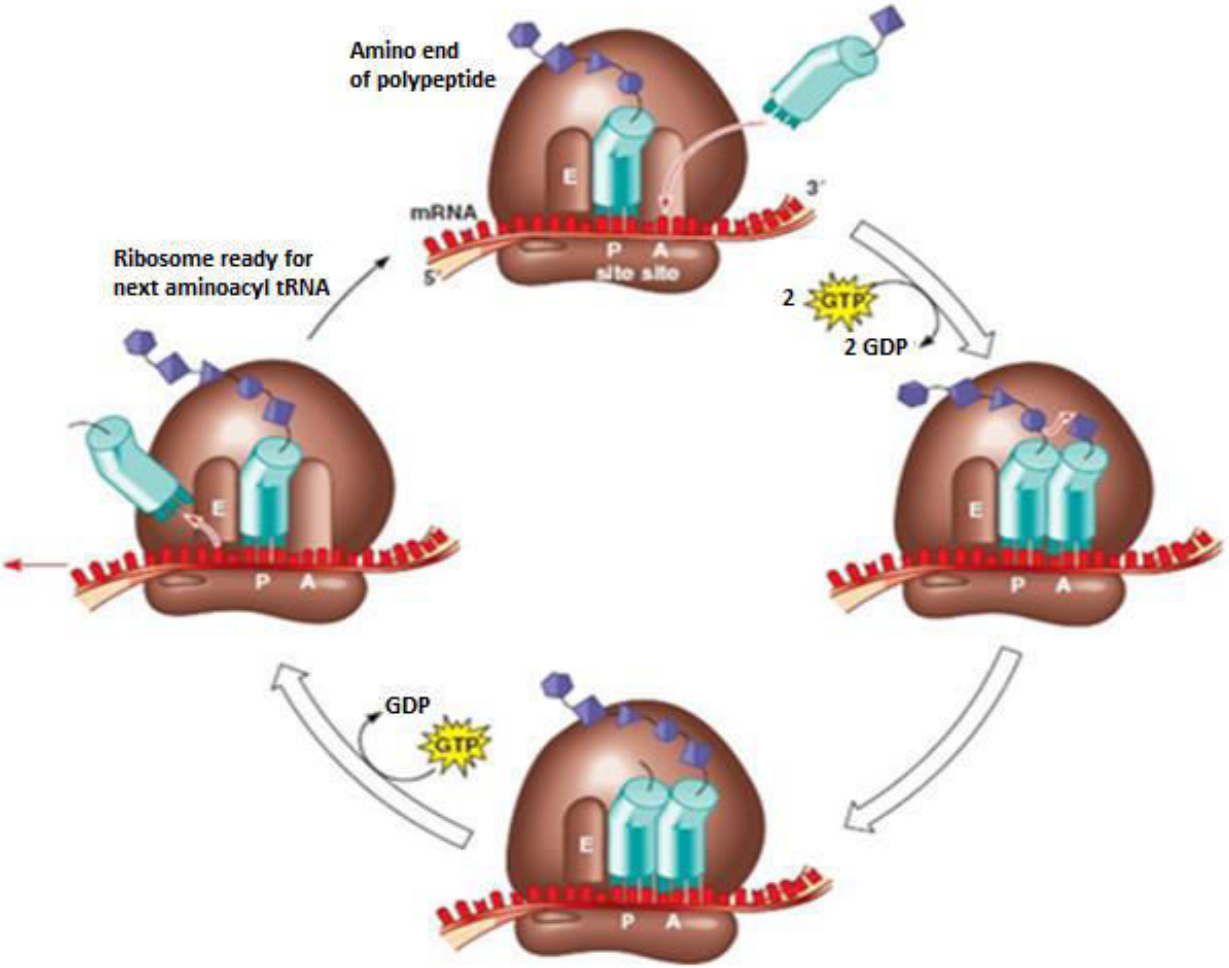


Protein Translation

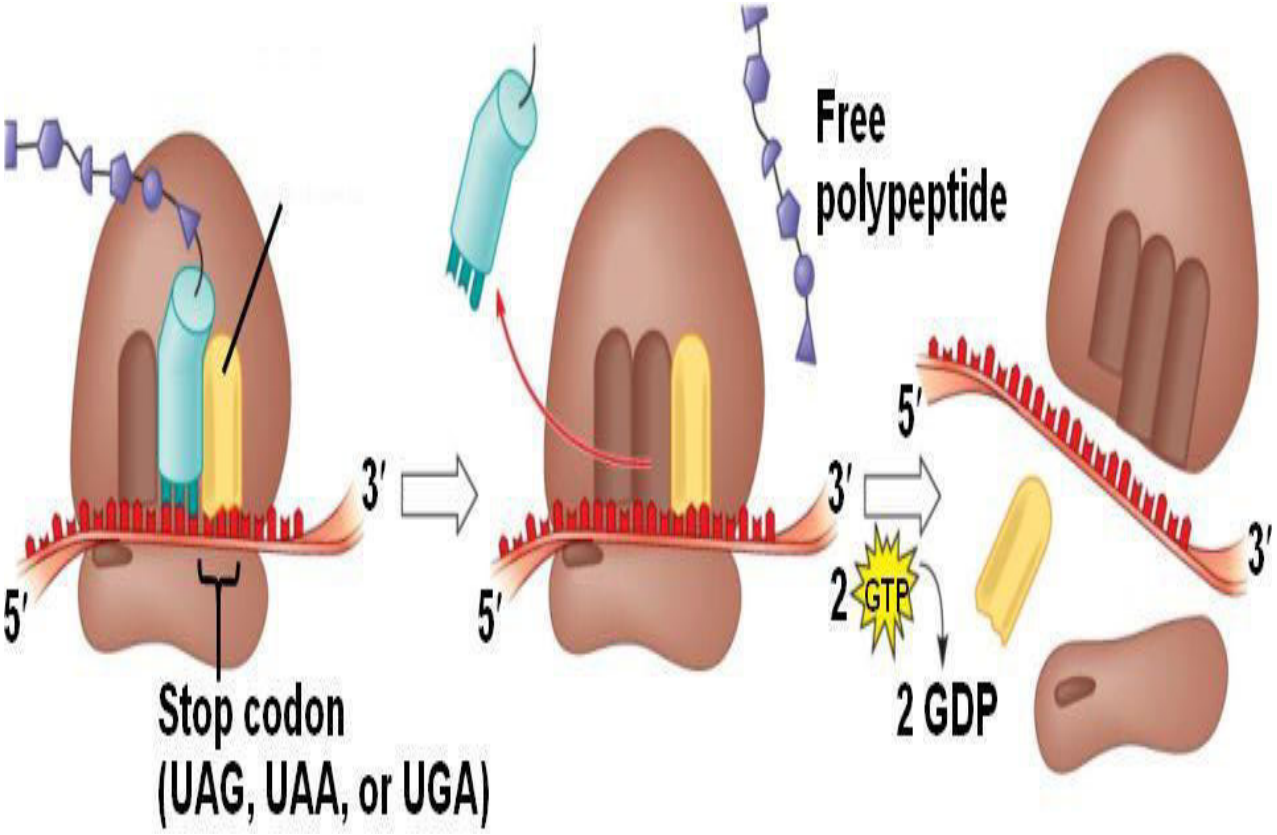
Initiation

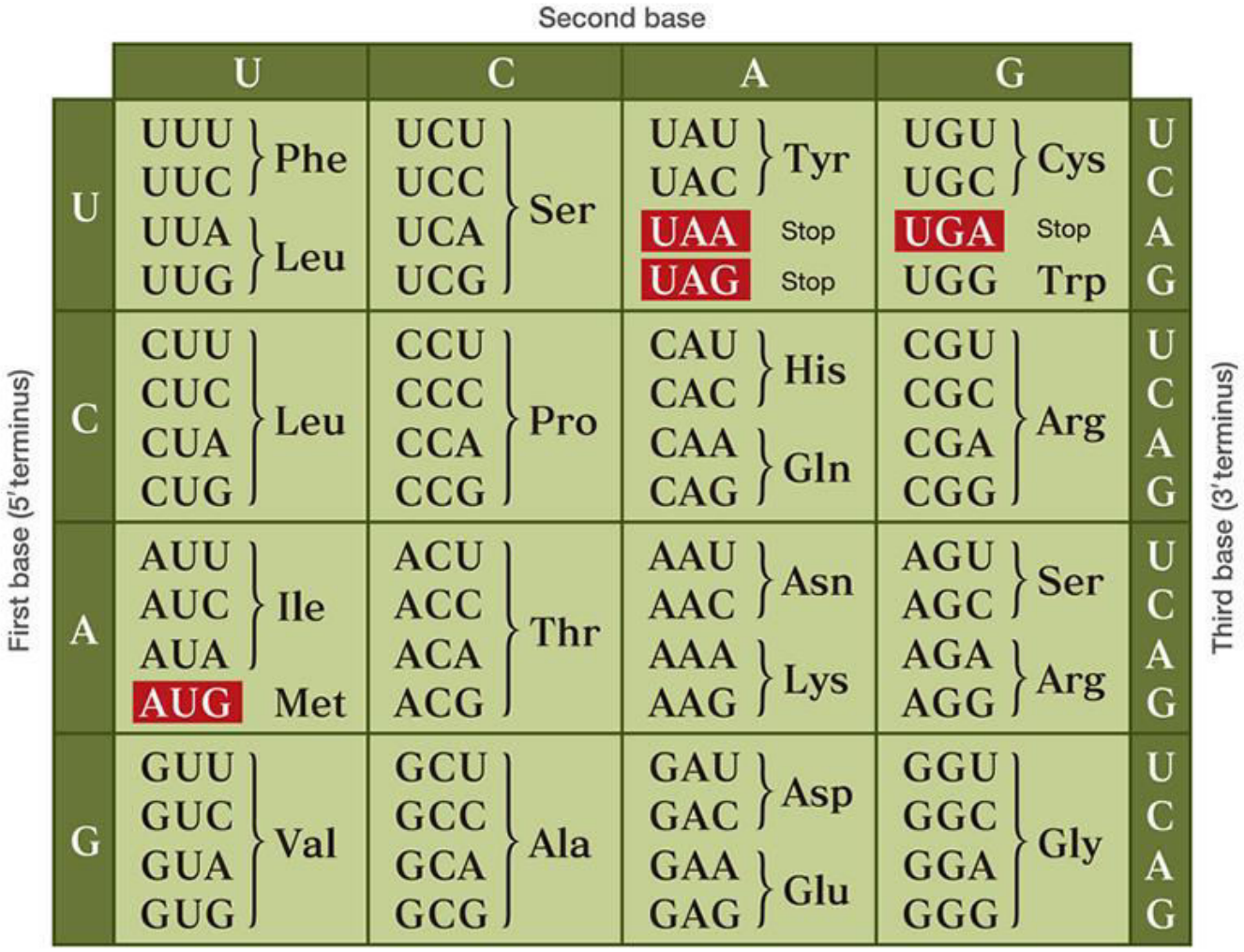

3′ U A C 5′
5′ A U G 3′
Met
Initiator tRNA
mRNA
5′
3′
Start codon
mRNA binding site
Small ribosomal subunit
GTP
GDP
Large ribosomal subunit
P site
Met
E
A
5′
3′
Translation initiation complex

Elongation


Amino end
of polypeptide
mRNA
5'
3'
E
P site
A site
Ribosome ready for
next aminoacyl tRNA
2 GTP
2 GDP
GDP
GTP
E
P
A

Termination





© CSLS/The University of Tokyo

Genetics and Genomics

What is Genomics

What is Genomics

Genome

- Total amount of DNA of a single cell of an organism (haploid cell in the case of a diploid).
- The whole hereditary information of an organism encoded by DNA.

What is Genomics

Genomics and Biology of Organisms

- Determination of entire genome sequence is a prerequisite to understand the complete biology of an organism.

What is Genomics

Genomics includes

- To generate physical, genetic, and sequence maps of different genomes.
- To sequence the genomes of model organisms.
- To develop new technologies for mapping/sequencing.

What is Genomics

Genomics helps in different ways

- Functions of genes
- Organizations of genomes.
- Structural make-up of genomes.
- Functions of coding and non-coding DNA

What is Genomics

Study of Genomics helps to understand

- The ethical, social, and legal issues and challenges posted by genomic information.

What is Genomics

Genomics Sub-disciplines

- Comparative genomics
- Structural genomics
- Functional genomics
- Population genomics

What is Genomics

Genomics Sub-disciplines

- Metagenomics
- Microbial genomics

What is Genomics

Genomics - Conclusion

- Total amount of DNA of a single cell of an organism – Genome and study is called as Genomics.

Genetics and Genomics

Anatomy and Organization of Genomes

Genomes Anatomy/Organization

Genome Anatomy

- Anatomy of different genomes differ from each other.
- Eukaryotes and prokaryotes genomes differ very significantly.

Genomes Anatomy/Organization

Genome Anatomy

- Size of genomes - 1000 fold difference between eukaryotes and prokaryotes.
- ~ 30 fold between genomes of different eukaryotes.

Genomes Anatomy/Organization

Genome Anatomy

- In humans ~ 23,000
- Bacterial genomes ~ 1,500 – 2,000 genes.

Genomes Anatomy/Organization

Genome Anatomy - Eukaryotes

- Eukaryotic genomes are full of simple repeats, numerous types of transposable elements and other sequences.

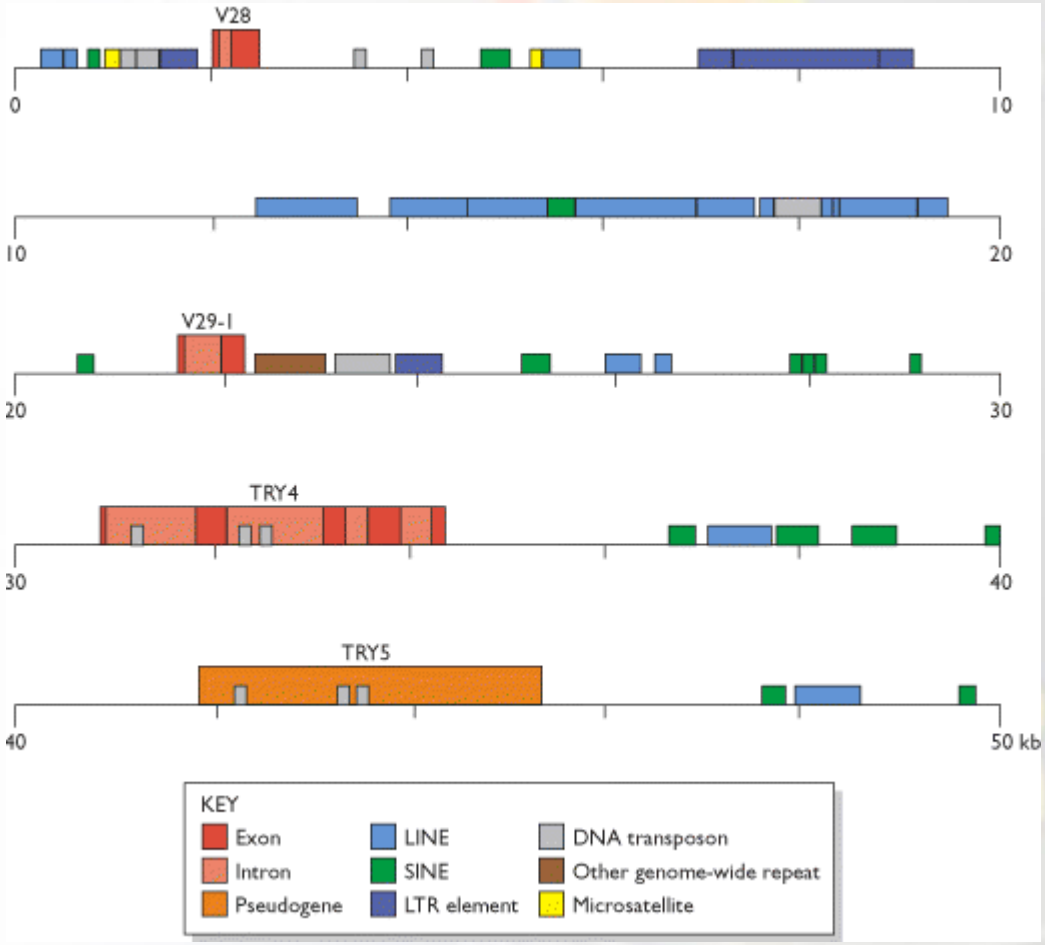
Genomes Anatomy/Organization

Genome Anatomy - Prokaryotes

- Prokaryotes have a few repeats and transposable elements and their genomes consist of mainly the genes.

Genomes Anatomy/Organization

Genome Anatomy: Segment of human chromosome 7



1 Gene: TRY4

2 Gene Segments: V28 & V29-1

1 Pseudogene: TRY5

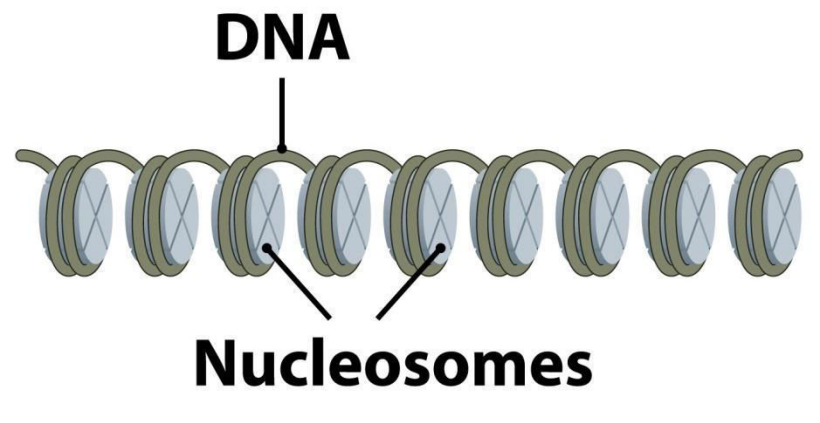
52 genome wide repeat sequences:

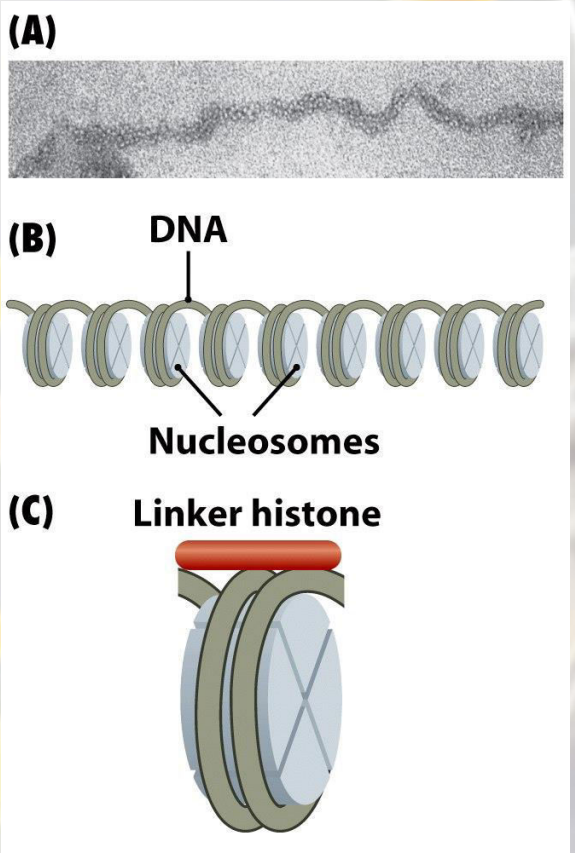
LINE, SINE, LTR, & DNA transposon.

Two Microsatellites

Genomes Anatomy/Organization

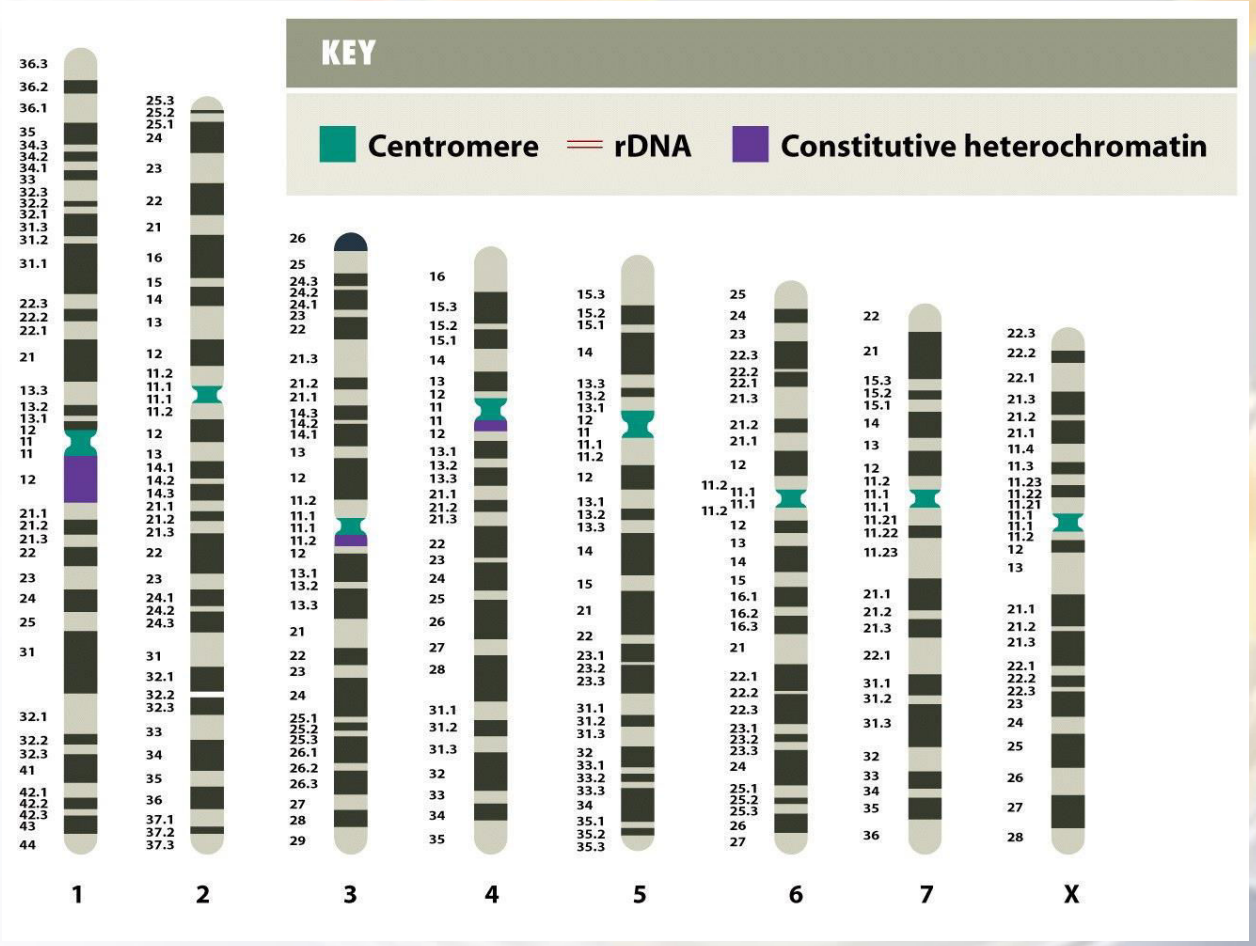
Genome Organization





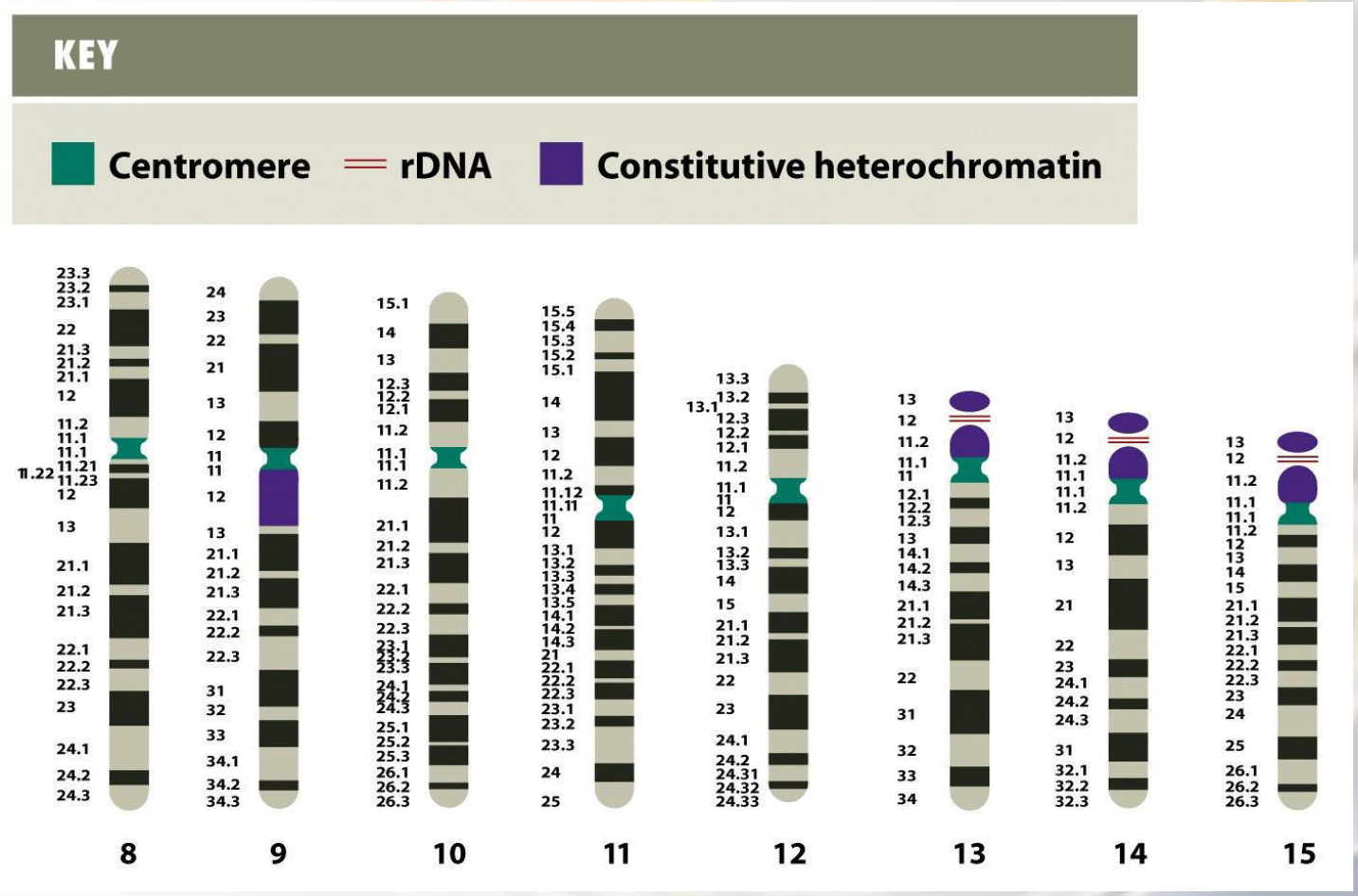
Genomes Anatomy/Organization

Genome Organization



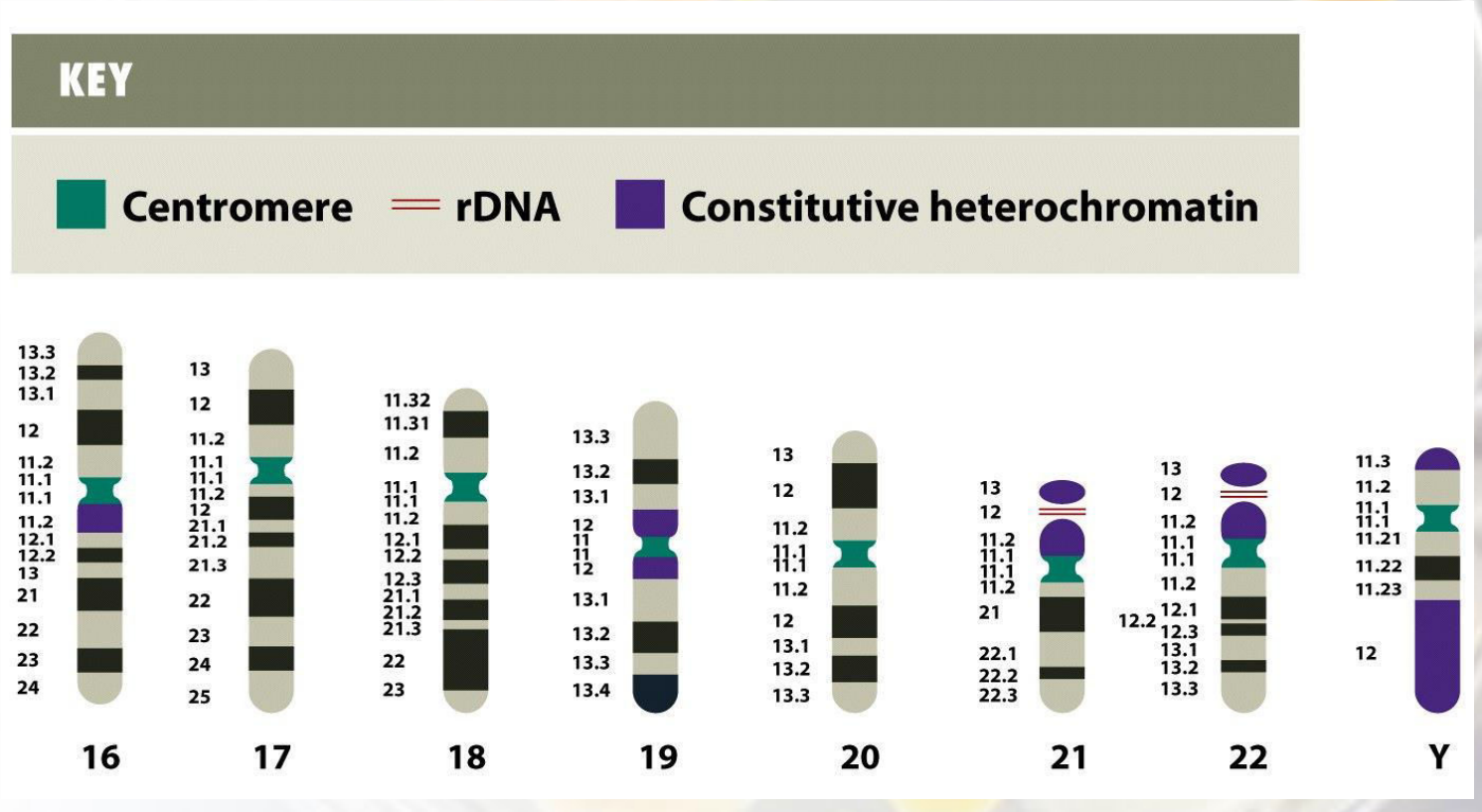
Genomes Anatomy/Organization

Genome Organization



Genomes Anatomy/Organization

Genome Organization



Genomes Anatomy/Organization

Genome Organization in Prokaryotes

Species	Genome organization		
	DNA molecules	Size (Mb)	Number of genes
Escherichia coli K12	One circular molecule	4.639	4405
Vibrio cholerae El Tor N16961	Two circular molecules		
	Main chromosome	2.961	2770
	Megaplasmid	1.073	1115
Deinococcus radiodurans R1	Four circular molecules		
	Chromosome 1	2.649	2633
	Chromosome 2	0.412	369
	Megaplasmid	0.177	145
	Plasmid	0.046	40

Genomes Anatomy/Organization

Genome Organization in Prokaryotes

Species	Size of genome (Mb)	Approximate number of genes
Bacteria		
Mycoplasma genitalium	0.58	500
Streptococcus pneumoniae	2.16	2300
Vibrio cholerae El Tor N16961	4.03	4000
Mycobacterium tuberculosis H37Rv	4.41	4000
Escherichia coli K12	4.64	4400
Yersinia pestis CO92	4.65	4100
Pseudomonas aeruginosa PA01	6.26	5700
Archaea		
Methanococcus jannaschii	1.66	1750
Archaeoglobus fulgidus	2.18	2500

Genomes Anatomy/Organization

Genome Organization: Comparisons

Species	Genome size (Mb)
Fungi	
Saccharomyces cerevisiae	12.1
Aspergillus nidulans	25.4
Protozoa	
Tetrahymena pyriformis	190
Invertebrates	
Caenorhabditis elegans	97
Drosophila melanogaster	180
Bombyx mori (silkworm)	490
Strongylocentrotus purpuratus (sea urchin)	845
Locusta migratoria (locust)	5000

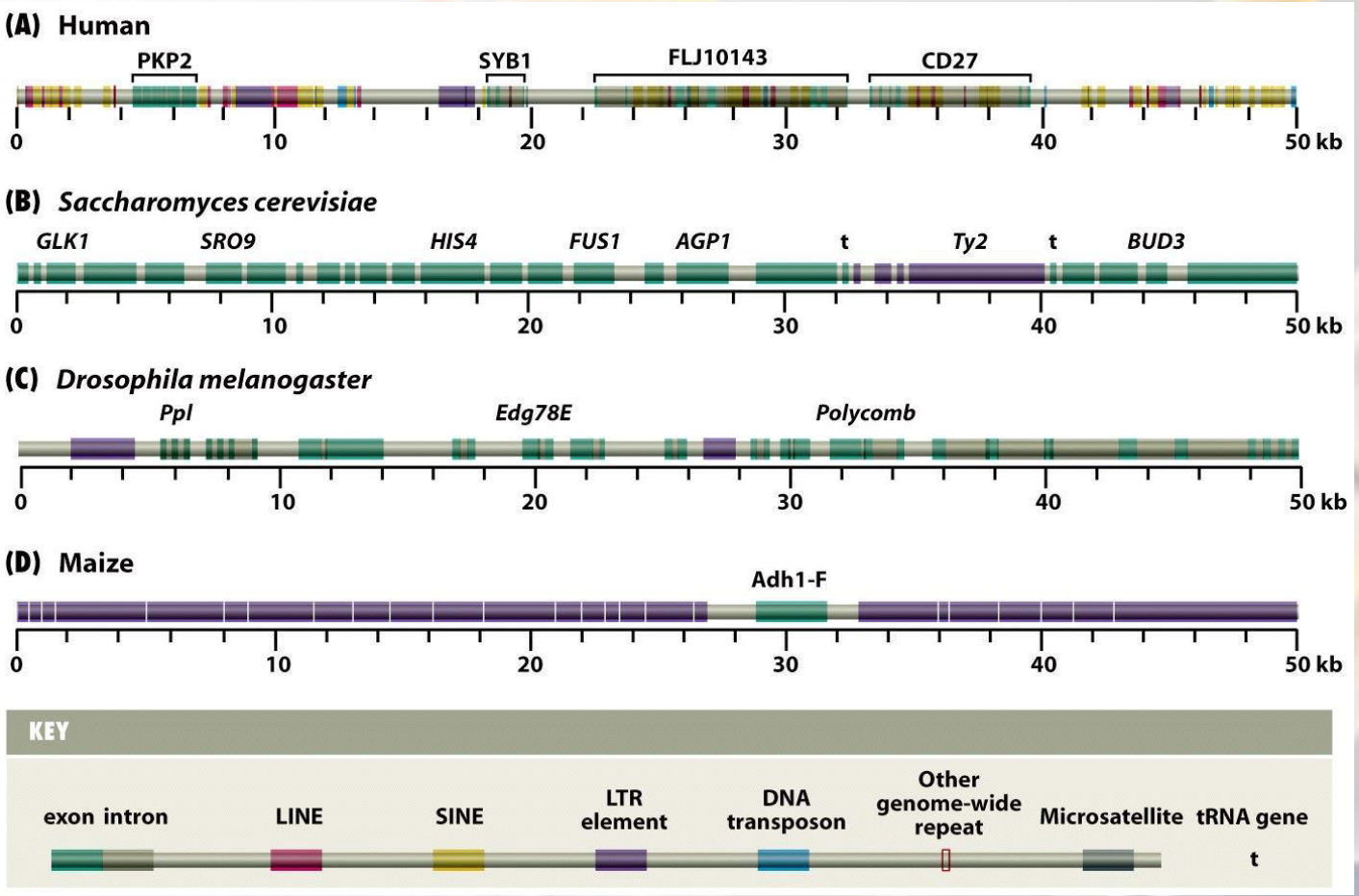
Genomes Anatomy/Organization

Genome Organization: Comparisons

Species	Genome size (Mb)
Vertebrates	
Takifugu rubripes (pufferfish)	400
Homo sapiens	3200
Mus musculus (mouse)	3300
Plants	
Arabidopsis thaliana (vetch)	125
Oryza sativa (rice)	466
Zea mays (maize)	2500
Pisum sativum (pea)	4800
Triticum aestivum (wheat)	16,000
Fritillaria assyriaca (fritillary)	120,000

Genomes Anatomy/Organization

Genome Organization: Human, Yeast, Fruit Fly , Maize



Genomes Anatomy/Organization

Compactness of Genomes

Feature	Yeast	Fruit fly	Human
Gene density (average number per Mb)	496	76	11
Introns per gene (average)	0.04	3	9
Amount of the genome that is taken up by genome-wide repeats	3.4%	12%	44%

Genomes Anatomy/Organization

Conclusion

- Anatomy of different genomes differ from each other.
- Eukaryotes and prokaryotes genomes differ very significantly.

Genetics and Genomics

Gene Anatomy

Gene Anatomy

What is Gene

- A piece of DNA (or RNA) that contains the primary sequence to produce a functional biological gene product (RNA or protein).

Gene Anatomy

What is Gene

- Entire nucleic acid sequence necessary for the synthesis of a functional polypeptide (protein chain) or functional RNA

Gene Anatomy

Genetic information is stored in DNA

- Segments of DNA that encode proteins or other functional products are called genes.
- Gene sequences are transcribed into messenger RNA (mRNA).

Gene Anatomy

mRNA is translated into Proteins

- mRNA intermediates are translated into proteins that perform most of the life functions.

Gene Anatomy

Gene Anatomy

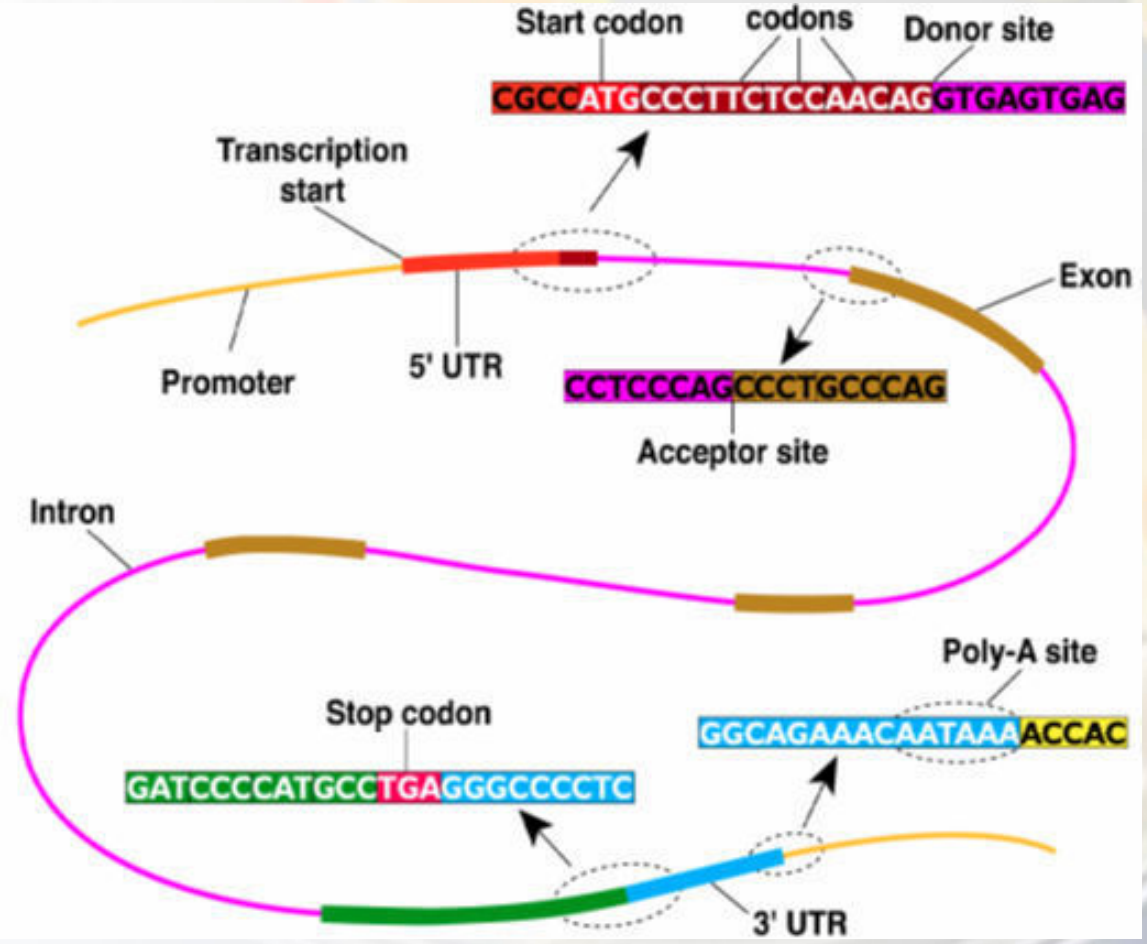
- Three components
- Open Reading Frame: From start codon (ATG) to stop (TGA, TAA, TAG)
- Upstream region with binding site. (e.g. TATA box, GC box, CAAT box).
- Poly-a tail

Gene Anatomy

Gene Anatomy – Typical Prokaryotic Gene

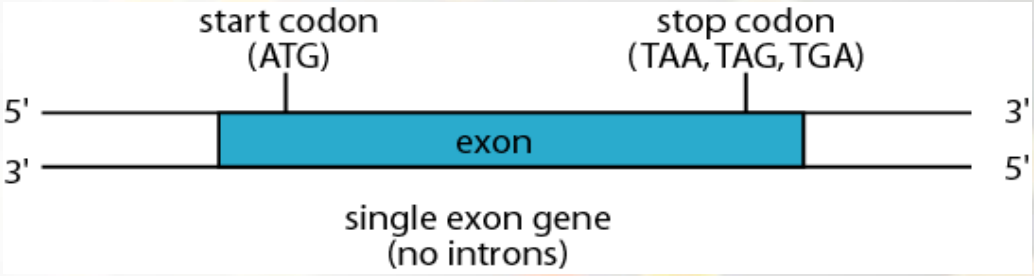
Gene Anatomy

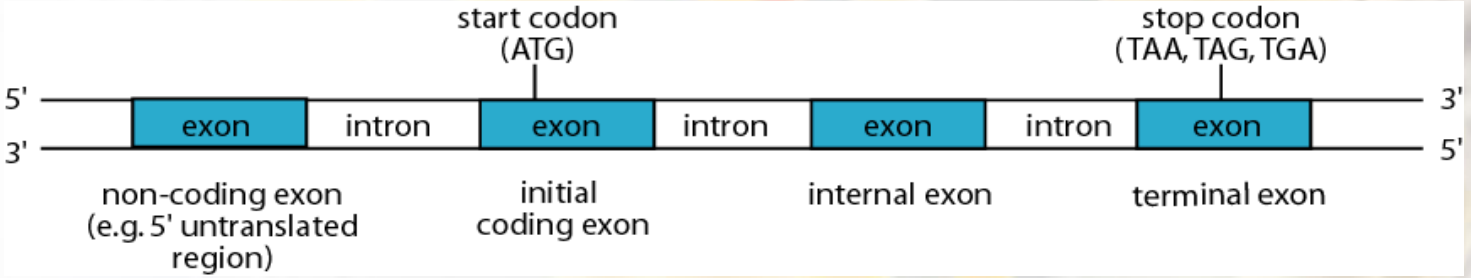
Gene Anatomy – Typical Eukaryotic Gene



Gene Anatomy

Single Exon Gene and Multiple Exons Gene





Gene Anatomy

What is Gene

- A piece of DNA (or RNA) that contains the primary sequence to produce a functional biological gene product (RNA or protein).

Genetics and Genomics

Prokaryotic Gene and Eukaryotic Gene

Prokaryotic Gene/Eukaryotic Gene

Bacterial Gene

- Most do not have introns
- Many are organized in operons: contiguous genes, transcribed as a single polycistronic mRNA, that encode proteins with related functions

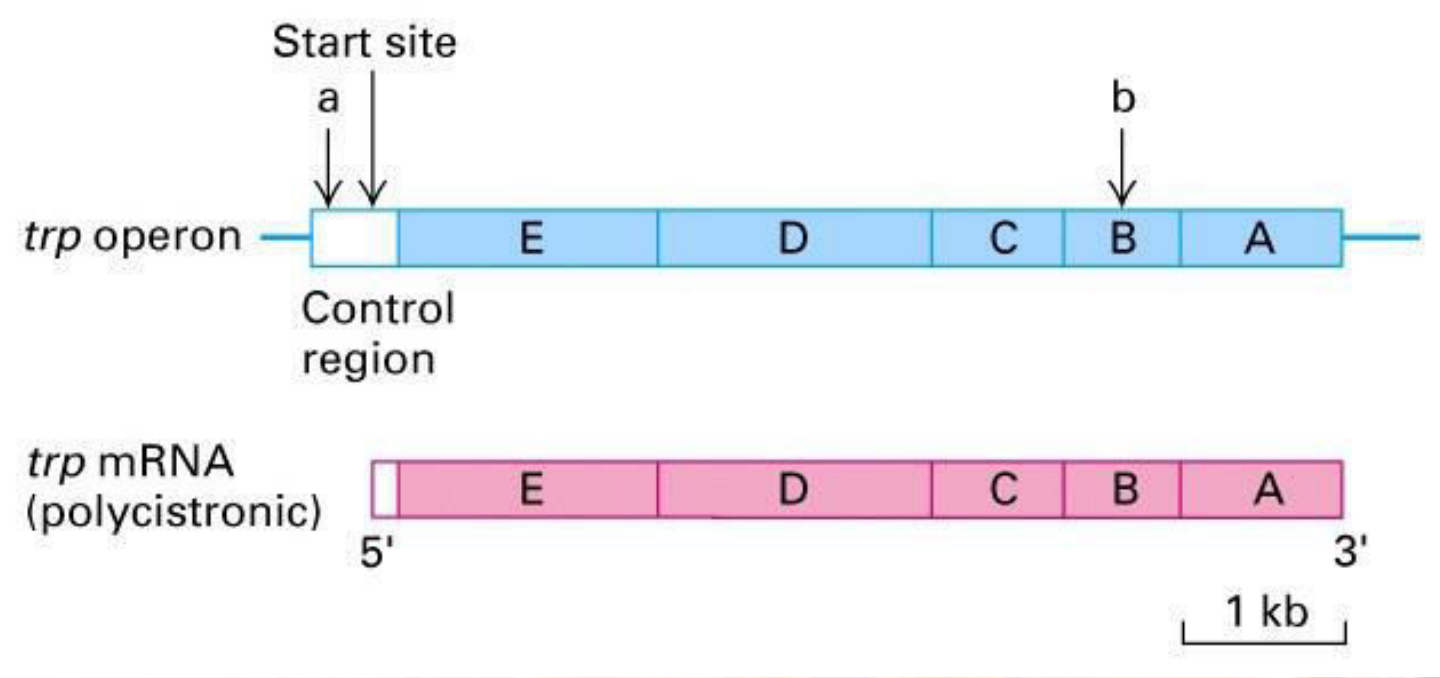
Prokaryotic Gene/Eukaryotic Gene

Bacterial Gene

- Polycistronic mRNA encodes several proteins

Prokaryotic Gene/Eukaryotic Gene

Bacterial Gene: Polycistronic mRNA encodes several proteins



Prokaryotic Gene/Eukaryotic Gene

Eukaryotic Gene: Exons and Introns

- Introns: intervening sequences within a gene that are not translated into a protein sequence.
- Exons: sequences within a gene that encode protein sequence.

Prokaryotic Gene/Eukaryotic Gene

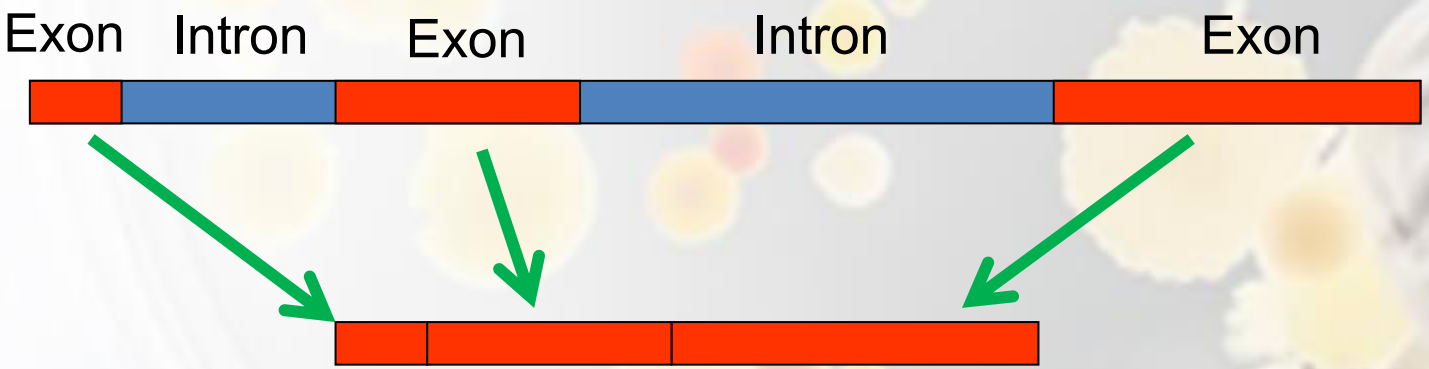
Eukaryotic Gene

- Splicing: Removal of introns from the mRNA molecule

Prokaryotic Gene/Eukaryotic Gene

Eukaryotic Gene

- Splicing: Removal of introns from the mRNA molecule



Prokaryotic Gene/Eukaryotic Gene

Eukaryotic Gene

- Organize expression of genes' (function calls)
- Promoter region (binding site), usually near coding region
- Binding can block (inhibit) expression

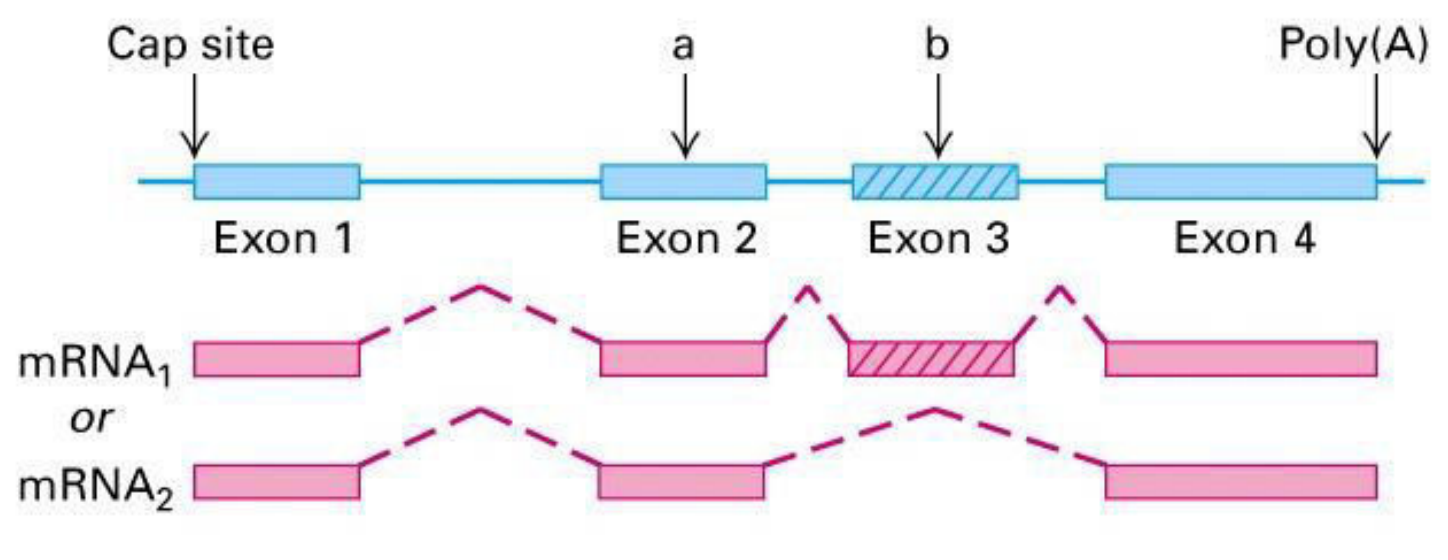
Prokaryotic Gene/Eukaryotic Gene

Alternative Splicing in Eukaryotic Genes

- Most have introns
- Produce monocistronic mRNA
- Large in size

Prokaryotic Gene/Eukaryotic Gene

Eukaryotic Gene: Alternative Splicing



Prokaryotic Gene/Eukaryotic Gene

Eukaryotic Gene

Computational challenges

Identify binding sites

Correlate sequence to expression

Prokaryotic Gene/Eukaryotic Gene

Conclusion

- Most of prokaryotic genes are without introns and in the are polycistronic.
- Eukaryotic genes have introns and alternative splicing.

Genetics and Genomics

Prokaryotic and Eukaryotic Gene Expression - Difference

Prokaryotic/Eukaryotic Gene Expression-Difference

Prokaryotic and Eukaryotic Gene Expression

- Bacterial genetics are different.
- Prokaryote genes are grouped in operons
- Prokaryotes have one type of RNA polymerase for all types of RNA

Prokaryotic/Eukaryotic Gene Expression-Difference

Difference between Prokaryotic and Eukaryotic gene Expression

- mRNA is not modified
- The existence of introns in prokaryotes is extremely rare

Prokaryotic/Eukaryotic Gene Expression-Difference

Difference

- Transcription in bacteria, sigma factors bind to RNA polymerases.
- RNA polymerases/ sigma factors complex bind to promoter about 40 bases prior to coding region of gene.

Prokaryotic/Eukaryotic Gene Expression-Difference

Difference between Prokaryotic and Eukaryotic gene Expression

- In prokaryotes, the newly synthesized mRNA is polycistronic (code for more than one polypeptide chain).

Prokaryotic/Eukaryotic Gene Expression-Difference

Difference

- In prokaryotes, transcription of a gene and translation of the resulting mRNA occur simultaneously.
- Many polysomes are found associated with an active gene.

Prokaryotic/Eukaryotic Gene Expression-Difference

Difference - Prokaryotic and Eukaryotic Gene Expression

Prokaryotes	Eukaryotes
• Operons • 27% of E. coli genes (Housekeeping genes not in operons) • Simultaneous transcription and translation	• No operons, but they still need to coordinate regulation • More kinds of control elements • RNA processing • Chromatin remodeling • Histones must be modified to loosen DNA

Prokaryotic/Eukaryotic Gene Expression-Difference

Prokaryotic/Eukaryotic Gene Expression

- Prokaryotes-operons simultaneously transcription and translation
- Eukaryotes – no operons, RNA processing, chromatin remodeling

Genetics and Genomics

Genes Families
and
Genes Clusters

Genes Families and Genes Clusters

Genes Clusters

- Many genes are arranged in groups along a chromosome - Genes Clusters
- Eukaryotic ribosomal RNA genes: tandem repeats

Genes Families and Genes Clusters

Genes Families

- Related genes may be organized as several clusters at different locations
- These are known as gene families e.g. globin genes family

Genes Families and Genes Clusters

Single Copy Genes and Genes Families

- Many eukaryotic genes are present in one copy per haploid set of chromosomes
- The rest of the genes occur in multigene families, collections of identical or very similar genes

Genes Families and Genes Clusters

Gene Family and Family members -

- Peroxiredoxin family
- *PRDX* is the root symbol
- Family members are *PRDX1*, *PRDX2*, *PRDX3*, *PRDX4*, *PRDX5* and *PRDX6*

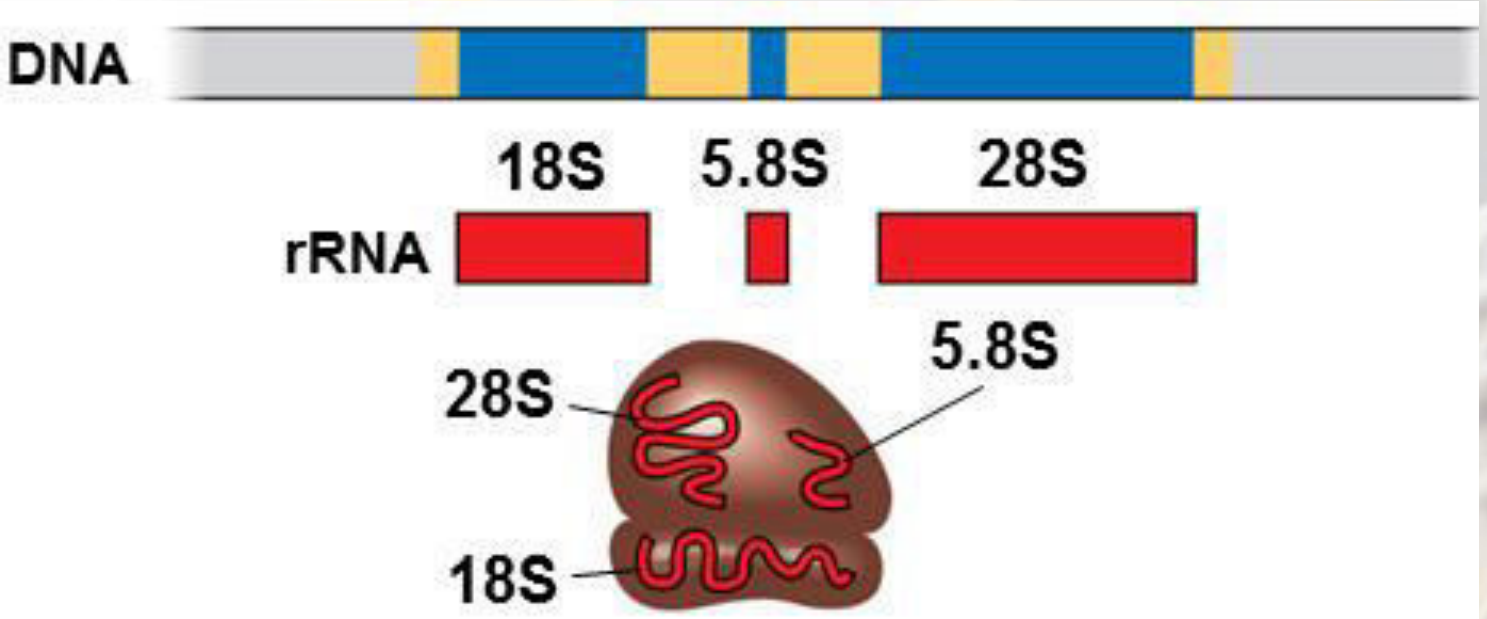
Genes Families and Genes Clusters

Genes Families

- Some multigene families consist of identical DNA sequences, usually clustered tandemly, such as those genes that code for rRNA products

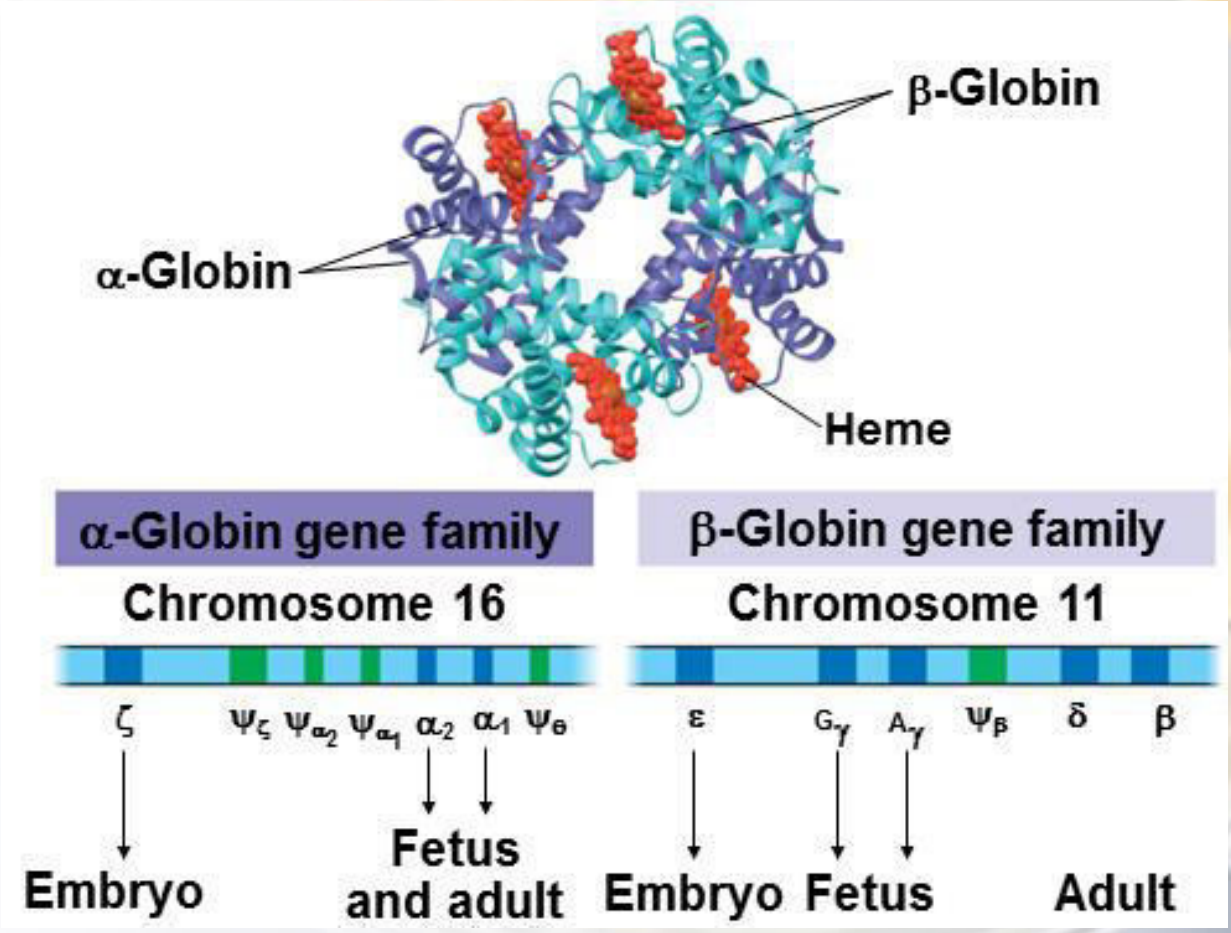
Genes Families and Genes Clusters

Ribosomal RNA Gene Family



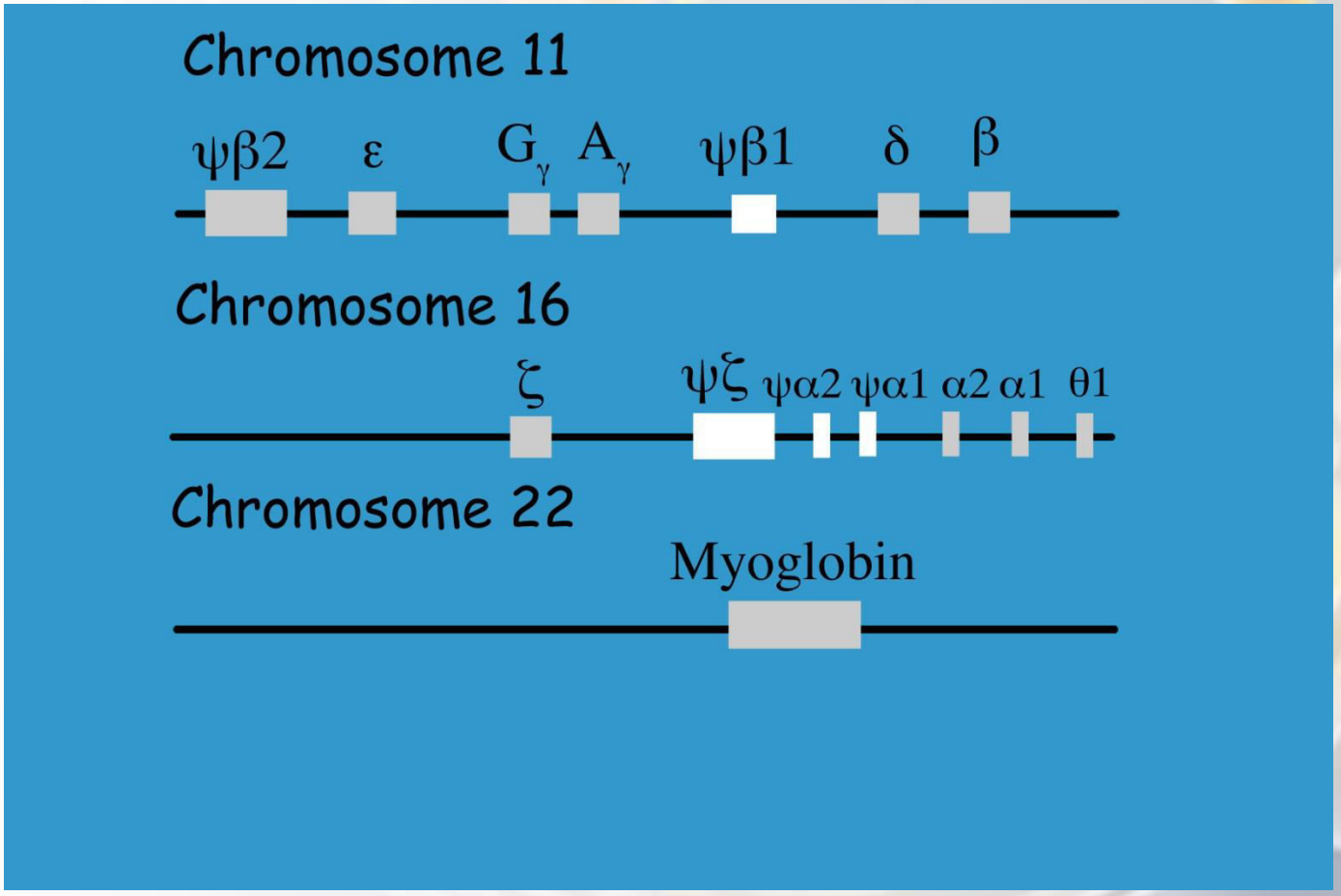
Genes Families and Genes Clusters

Human Alpha & Beta Globin Gene Families



Genes Families and Genes Clusters

Genes Families on Different Chromosomes



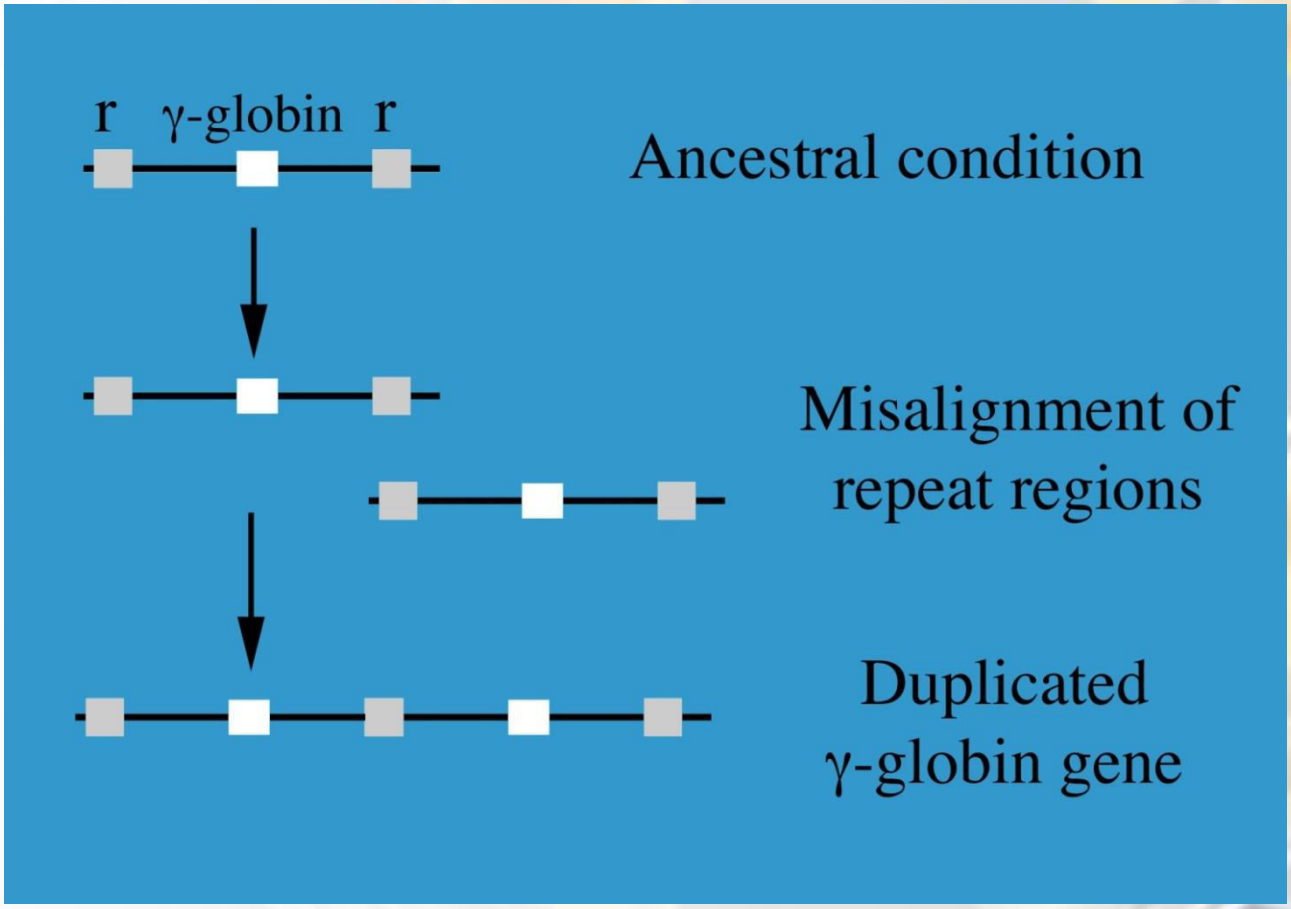
Genes Families and Genes Clusters

Genes Families

- The misalignment of genes during recombination is the most likely cause of gene duplication and clustering.
- Once repeats exist the probability of misalignment increases.

Genes Families and Genes Clusters

Genes Families: The misalignment - Globins



Genes Families and Genes Clusters

Super Families

- Super families are much larger than single multigene families.
- Super families contain hundreds of genes, including multiple multigene families and single individual genes

Genes Families and Genes Clusters

Super Families

- The large number of members allows super families to be widely dispersed with some genes clustered and some spread far apart.

Genes Families and Genes Clusters

Genes Families and Genes Clusters - Conclusion

- Single individuals Genes
- Gene families
- Gene families in form of clusters
- Gene families as Super families

Genetics and Genomics

Type of Proteins/ Families

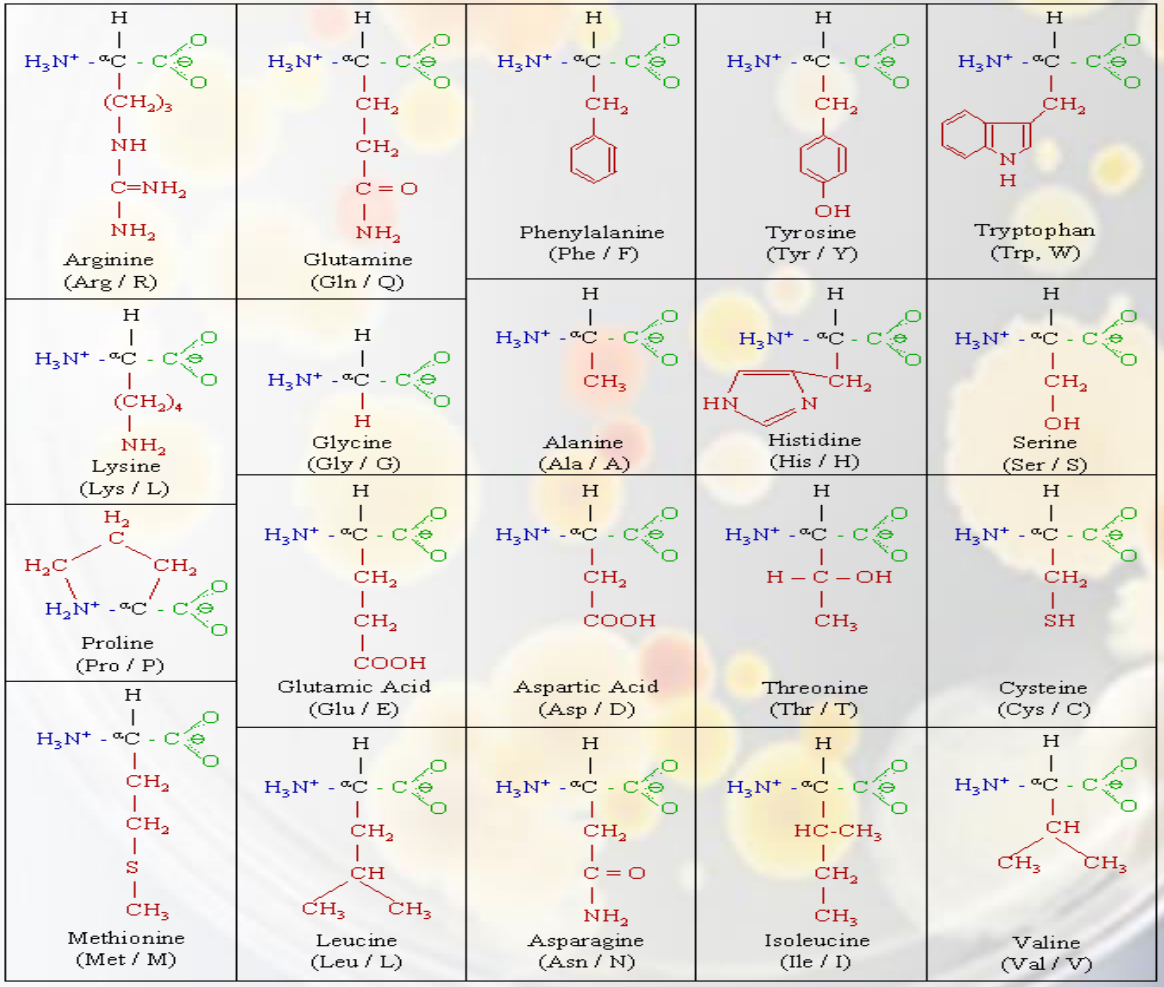
Type of Proteins/Families

Protein Structure

- Amino acids
  - 20 amino acids
  - Hydrophobic / hydrophylic
  - Charged / neutral
- Functions
  - Enzymes
  - Structure protein
  - Channel
  - Other functions

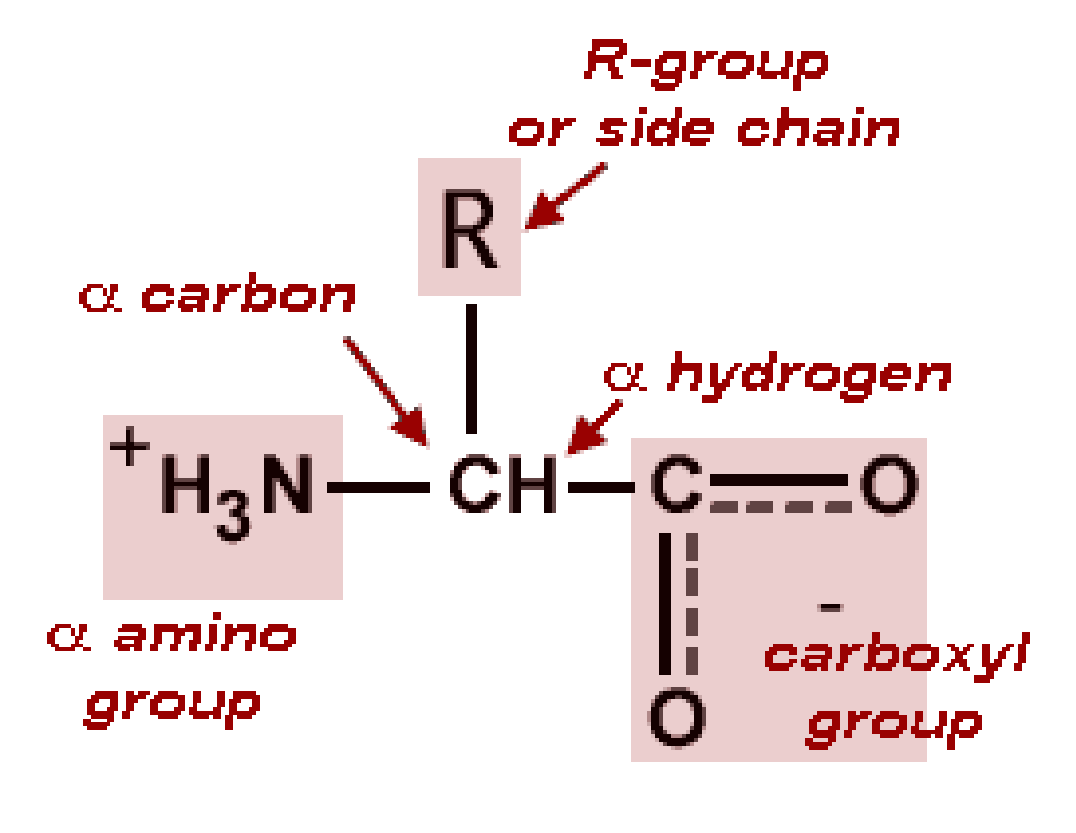
Type of Proteins/Families

Amino Acids: Building Blocks of Proteins



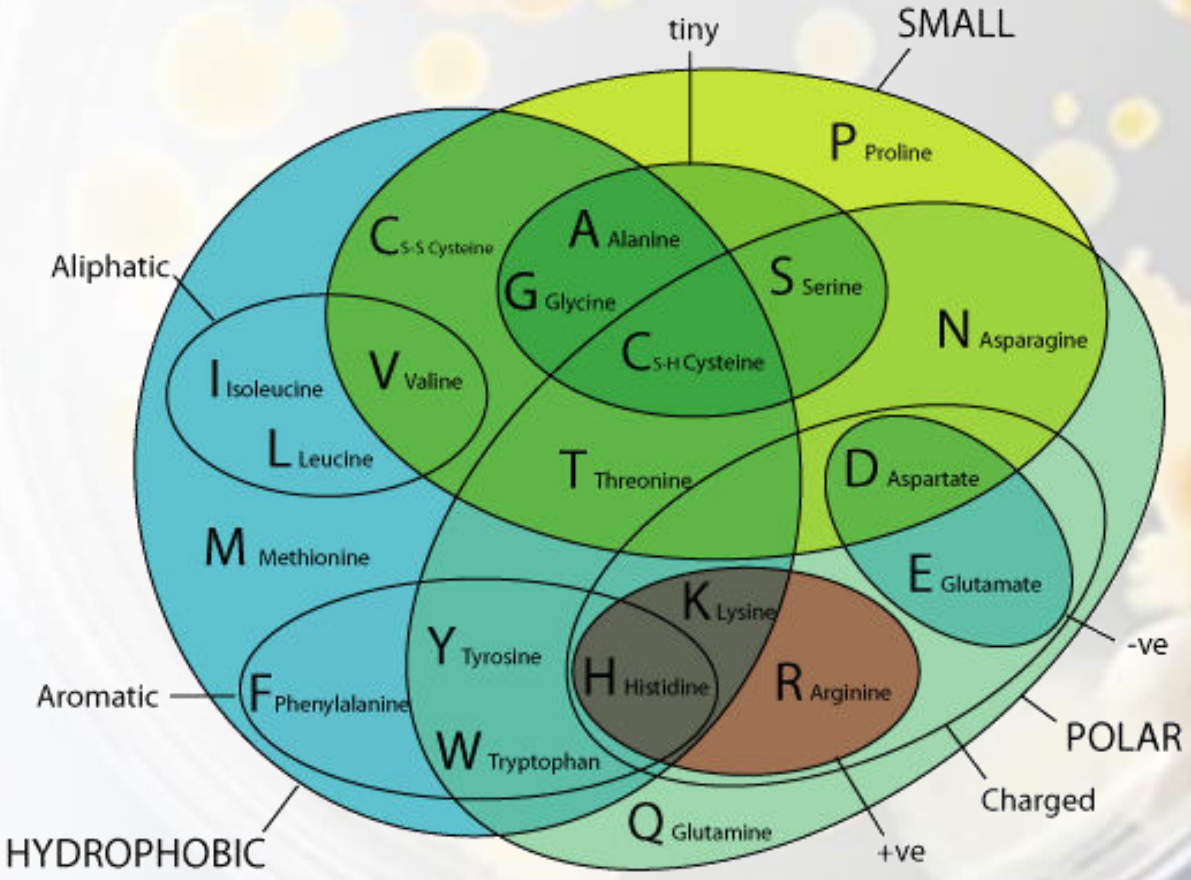
Type of Proteins/Families

Protein Structure: Amino Acid



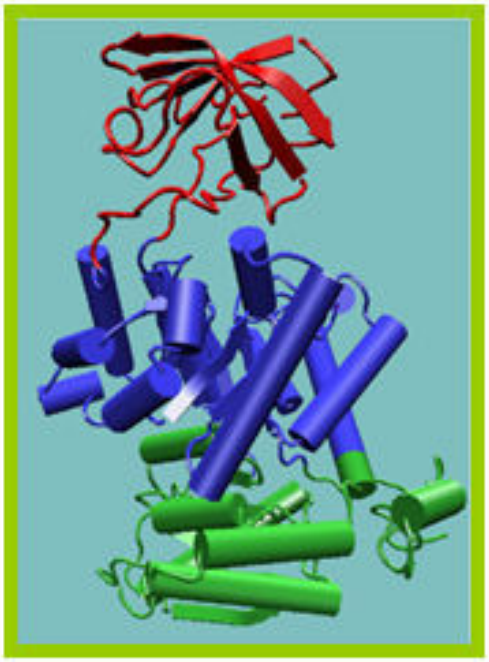
Type of Proteins/Families

Protein Structure: Types of Amino Acids



Type of Proteins/Families

Protein Domains: Domains are units of compact structure, function and evolution and folding



SKSHSEAGSAFIQTQQLHAAMADTFLEHMCRLDIDSAP
ITARNTGIICTIGPASRSVETLKEMIKSGMNVARMNFS
HGTHEYHAETIKNVRTATESFASDPILYRPVAVALDTK
GPEIRTGLIKGSGTAEVELKKGATLKITLDNAYMAACD
ENILWLDYKNICKVVEVGSKVYVDDGLISLQVKQKGPD
FLVTEVENGGFLGSKKGVNLPGAAVDLPAVSEKDIQDL
KFGVDEDVDMVFASFIRKAADVHEVRKILGEKGKNIKI
ISKIENHEGVRRFDEILEASDGIMVARGDLGIEIPAEK
VFLAQKMIIGRCNRAGKPVICATQMLESMIKKPRPTRA
EGSDVANAVLDGADCIMLSGETAKGDYPLEAVRMQHLI
AREAEAAMFHRKLFEELARSSSHSTDLMEAMAMGSVEA
SYKCLAAALIVLTESGRSAHQVARYRPRAPIIAVTRNH
QTARQAHLYRGIFPVVCKDPVQEAWAEDVDLRVNLAMN
VGKAAGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP

Type of Proteins/Families

Proteins

- Different method of categorizing proteins
- Three major categories
- Fibrous proteins
- Globular proteins
- Complexes with multiple components including proteins

Type of Proteins/Families

Fibrous Proteins – Cytoskeletal Proteins

- Actin
- Coronin
- Dystrophin
- Keratin
- Tubulin

Type of Proteins/Families

Fibrous Proteins – Extracellular Matrix Proteins

- Collagen
- Elastin
- Fibronectin

Type of Proteins/Families

Globular Proteins - Major Types

- Plasma proteins
- Hemoproteins
- Cell adhesion
- Transmembrane transport proteins
- Hormones and growth factors
- Receptors

Type of Proteins/Families

Globular Proteins - Major Types

- DNA-binding protein
- Immune system proteins
- Nutrient storage/transport
- Chaperone proteins
- Enzymes

Type of Proteins/Families

Complexes with multiple components including proteins

- Nucleosome
- Ribonucleoprotein (generic)
- Signal recognition particle
- Spliceosome

Type of Proteins/Families

Types of Proteins - Conclusion

- Three major categories
- Fibrous proteins
- Globular proteins
- Complexes with multiple components including proteins

Genome Informatics

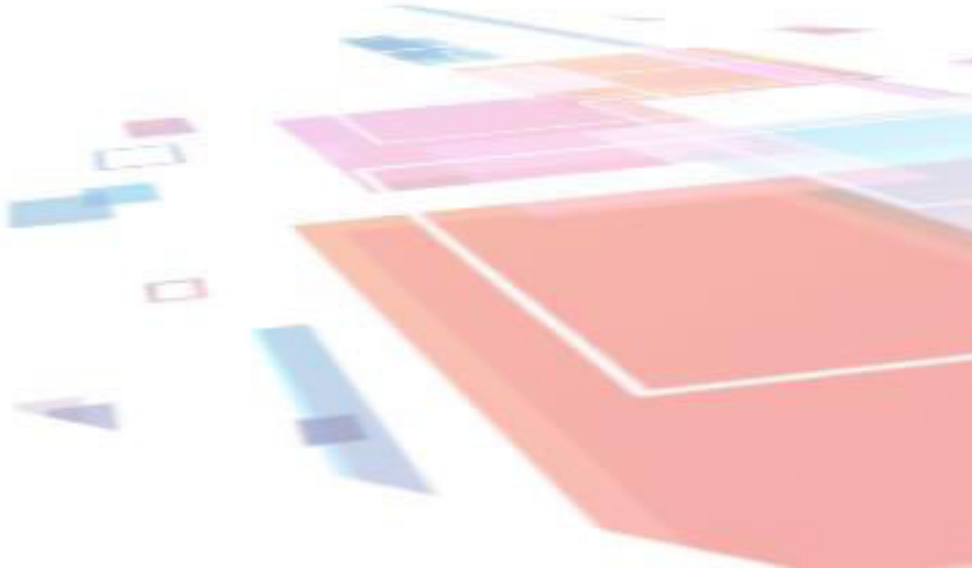
Introduction

- Genome sequencing provides the sequences of all the genes of an organism
- A major application of Bioinformatics is analysis of full genomes that have been sequenced

Genome Informatics

Introduction

- Challenge is to identify those genes that are predicted to have a particular biological function



Genome Informatics

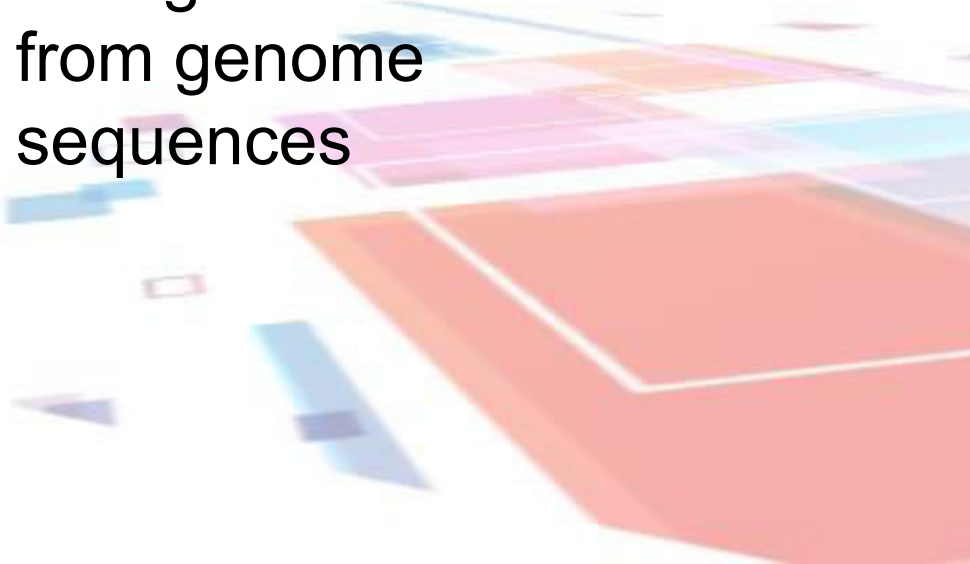
Genomics

Study of all of a person's genes (the **Genome**), including interactions of those genes with each other and with the person's environment **(NHGRI)**

Genome Informatics

Genome Informatics

Genome informatics is the field in which computational and statistical techniques are applied to derive biological information from genome sequences



Genome Informatics

Genome Informatics

Genome informatics includes methods to analyze DNA sequence information and to predict protein sequence and structure

(Iossifov, et al. 2014)

Genome Informatics

Genome Sequences

Availability of genome sequences facilitates;

- The discovery and utilization of sequence polymorphisms
- Opportunity to explore genetic variability both between and within the organisms

Genome Informatics

Genome Analysis

Following tasks;

- Sequencing
- Assembly
- Repeat identification and masking out
- Gene prediction
- Looking for EST and cDNA sequences
- Genome annotation

Genome Informatics

Genome Analysis

- Expression analysis
- Metabolic pathways and regulation studies
- Functional genomics
- Gene location/gene map identification

Genome Informatics

Genome Analysis

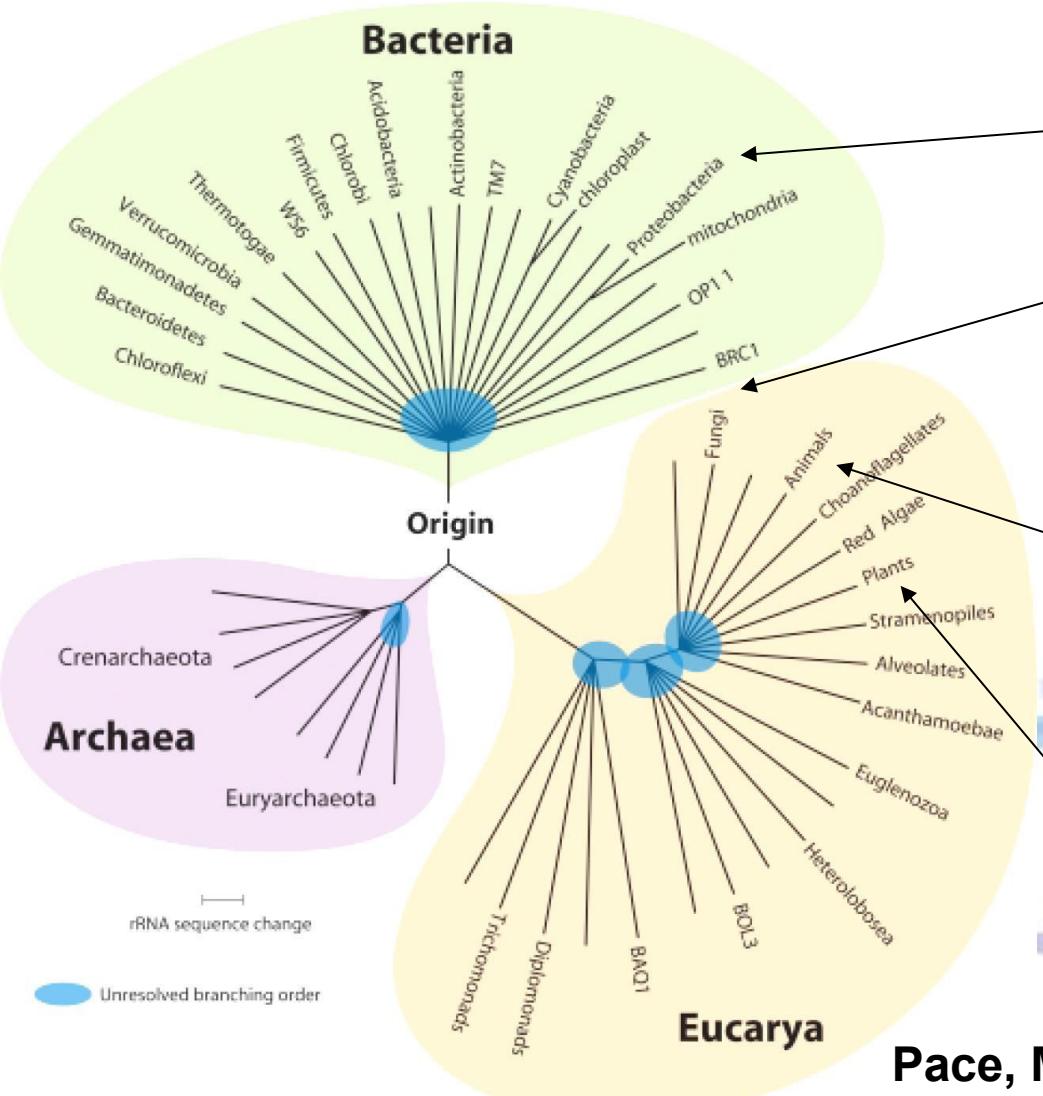
- Comparative genomics
- Identify clusters of functionally related genes
- Evolutionary modeling
- Self-comparison of proteome

Genome Informatics

Model organisms

- E. coli – bacteria
- S. cerevisiae – yeast
- C. elegans – worm
- D.melanogaster – fly
- Danio rerio – zebrafish
- Mus musculus - mouse
- Homo sapiens – you and me
- Arabidopsis - plant

Genome Informatics



E. coli is here

Yeast (*Saccharomyces*)

Worms, flies, fish, mice

Arabidosis, rice, soybean

Pace, Microbiol. Mol. Biol. Rev. 73:565 (2009)

Genome Informatics

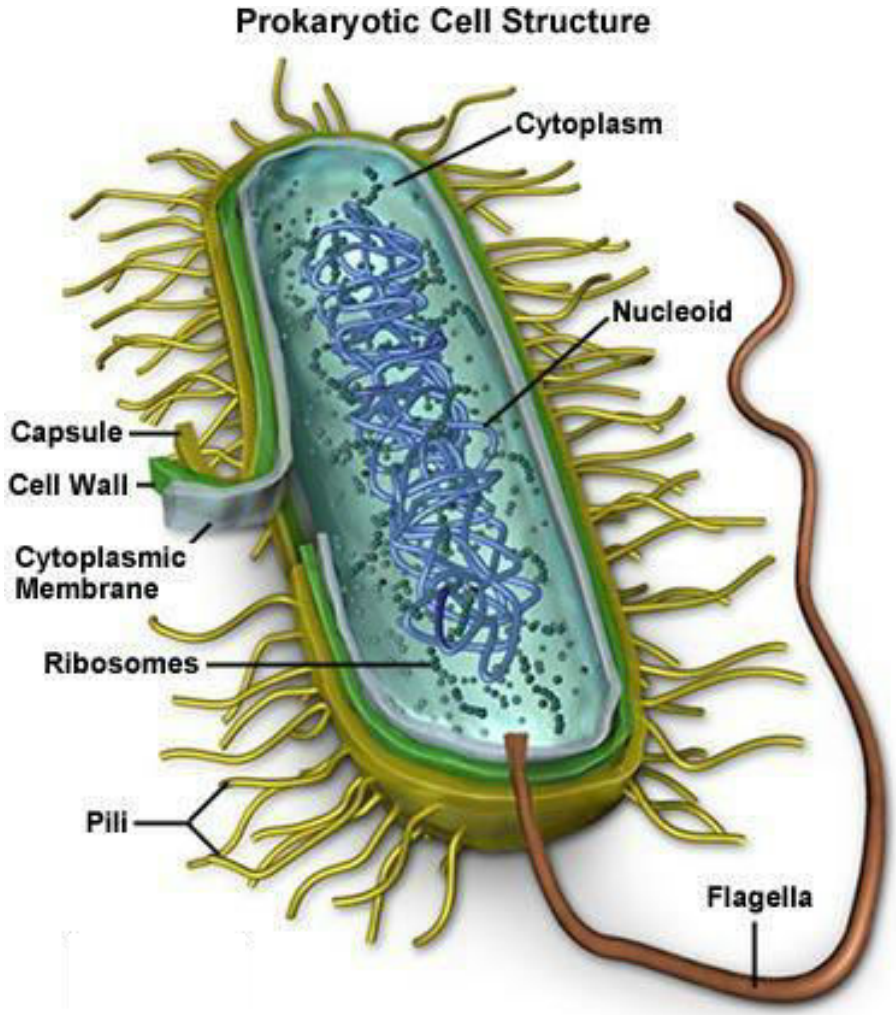
Conclusions

- Sequencing and analysis of full genomes paves the way for future discoveries
- Different model organisms can help explore our Genome and what maters most for us

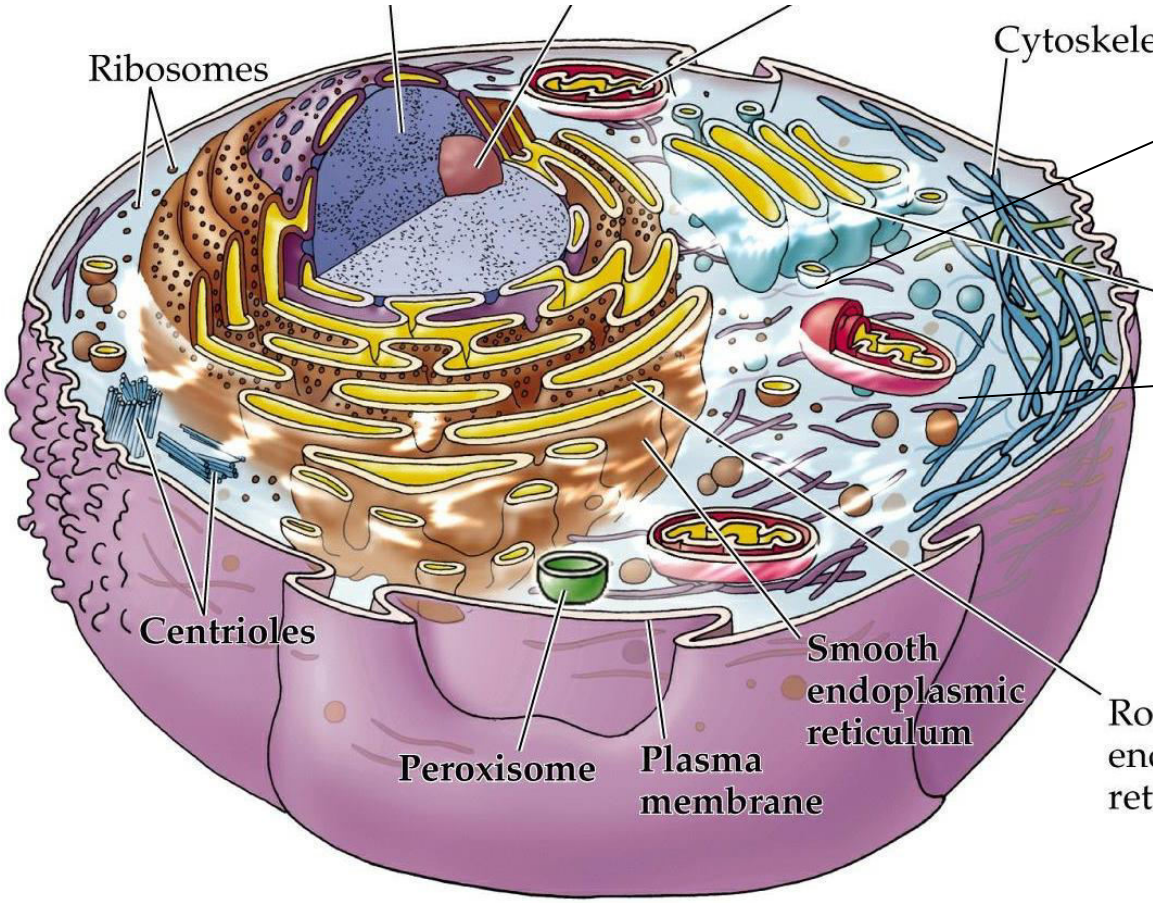
Prokaryotic Genome

- Prokaryotes are the organisms whose Genetic material (DNA) is not enclosed in a nuclear membrane
- No membrane bound organelles

Prokaryotic Genome



Prokaryotic Genome



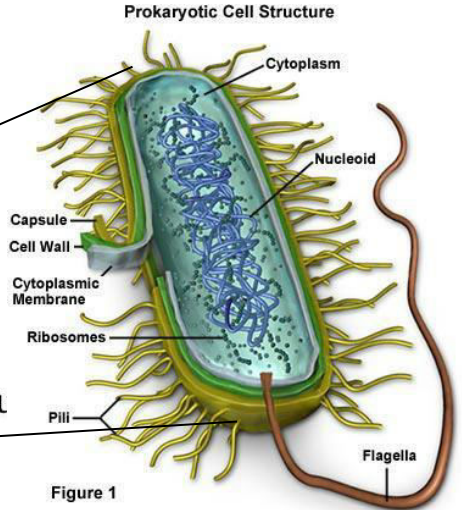
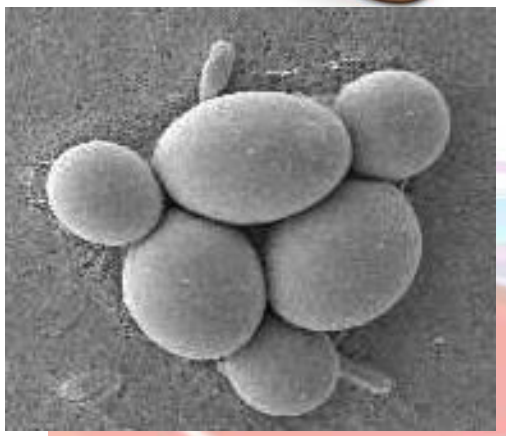


Figure 1



10 um

© 2001 Sinauer Associates, Inc.

Mitochondria evolved from a bacterial endosymbiont

Prokaryotic Genome

- First prokaryotic Genome sequenced was that of *Hemophilus influenzae*
- Paved the way for sequencing of many other organisms

Prokaryotic Genome

Selection criteria

Following are the criteria for organisms selected for sequencing;

- They had been subjected to a detailed biological analysis and thus were model organisms

Prokaryotic Genome

Selection criteria

- important human pathogens
- They were of phylogenetic interest
- Sequences were annotated as they were sequenced

Prokaryotic Genome

Table 10.2. *Features of representative prokaryotic genomes*

Organism (reference)	Phylogenetic group	Genome size (Mbp) (no. protein-encoding genes)	Novel functions
Escherichia coli (Blattner et al. 1997)	Bacteria	4.6 (4288)	model organism
Methanococcus jannaschii (Bult et al. 1996)	Archaea	1.66 (1682)[a]	grows at high temperature and pressure and produces methane
Hemophilus influenzae (Fleischmann et al. 1995)	Bacteria	1.83 (1743)	human pathogen
Mycoplasma pneumoniae (Himmelreich et al. 1996)	Bacteria	0.82 (676)	human pathogen that grows inside cells; metabolically weak
Bacillus subtilis (Kunst et al. 1997)	Bacteria	4.2 (4098)	model organism
Aquifex aeolicus (Deckert et al. 1998)	Bacteria	1.55 (1512)[b]	ancient species, grows at high temperature and can grow in a hydrogen, oxygen, carbon dioxide atmosphere in the presence of only mineral salts
Synechocystis sp. (Kaneko et al. 1996a,b)	Bacteria	3.57 (3168)	ancient organism that produces oxygen by light-harvesting; may have oxygenated atmosphere

Prokaryotic Genome

Conclusions

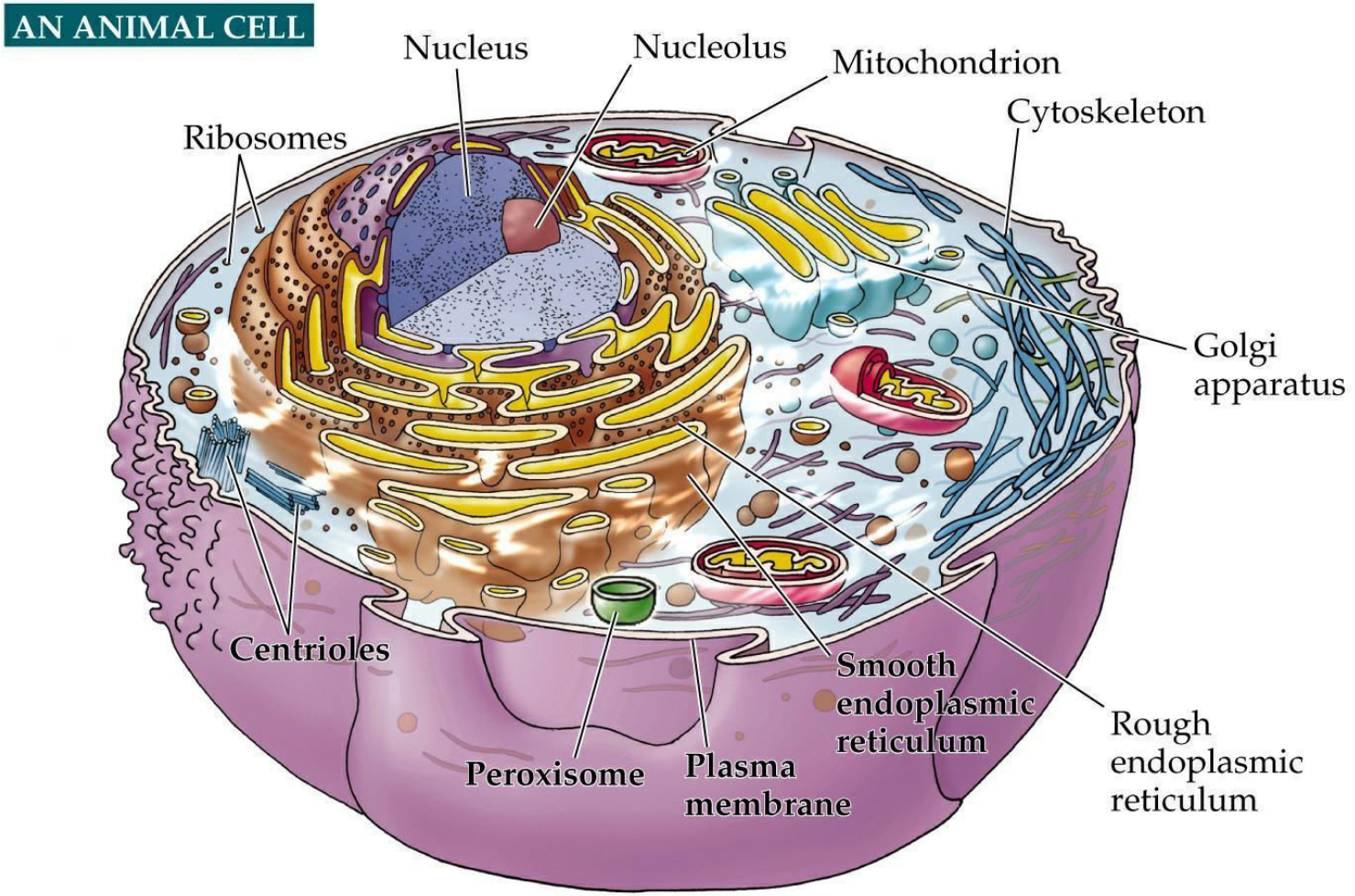
- Prokaryotes are simple Genomes
- Easy models to study Biochemistry and Molecular biology of life processes
- Sequencing is done on economically important organisms

Eukaryotic Genome

Introduction

- Eukaryotes have larger genomes
- Have tandem repeats
- Have introns in their protein-coding genes
- Heterochromatin and euchromatin region

Eukaryotic Genome



Eukaryotic Genome

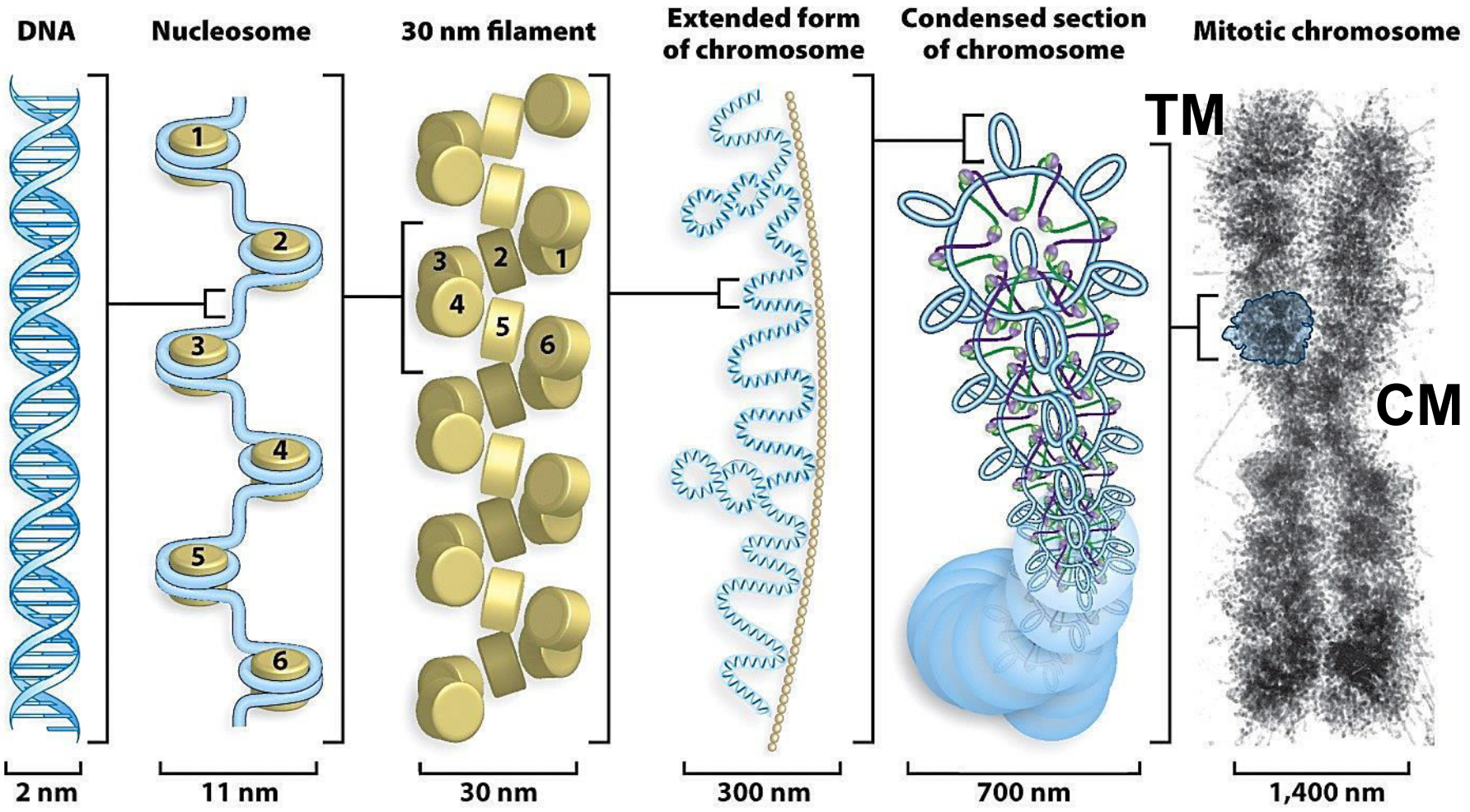


Figure 10-16
Molecular Biology: Principles and Practice

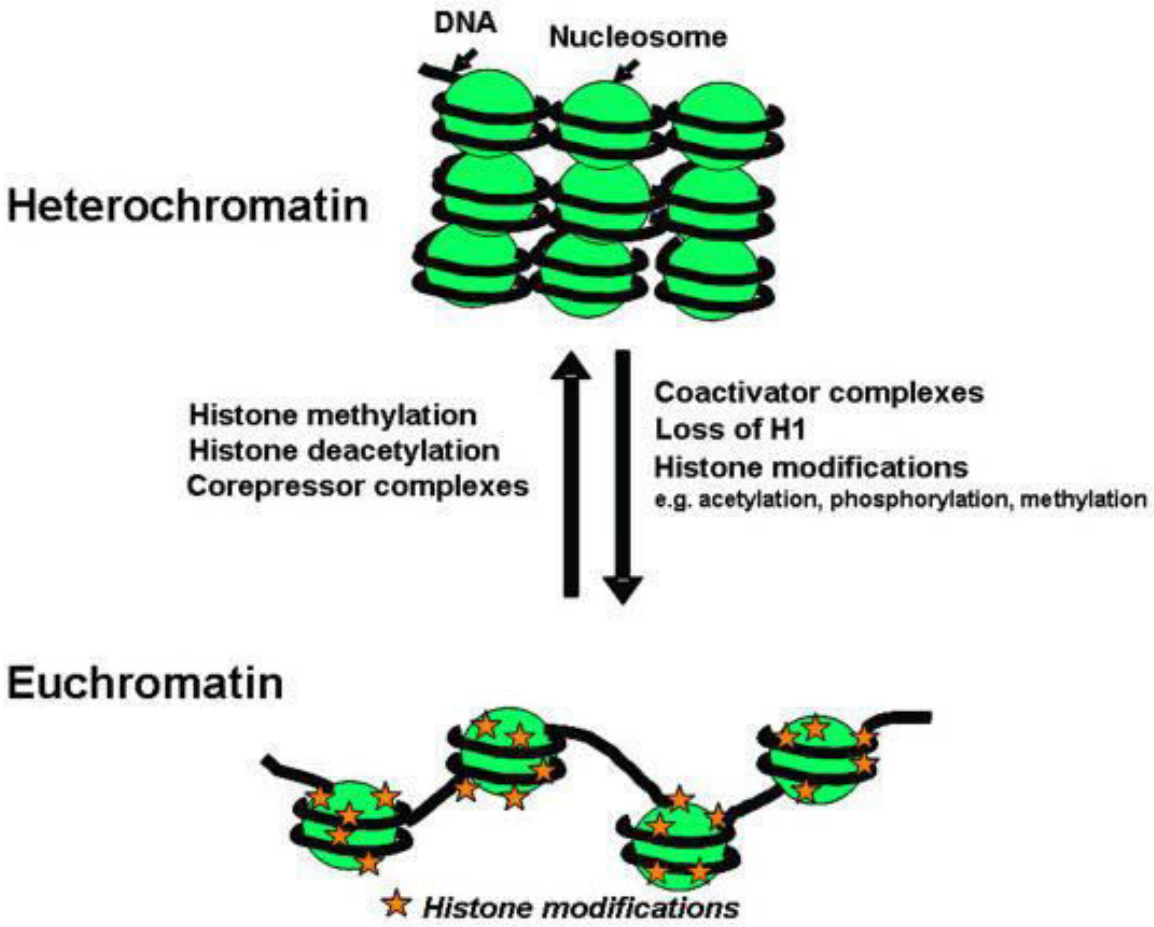
CM: Centromere TM: Telomere

Eukaryotic Genome

Staining with dyes

- Dense heterochromatin
- Light Euchromatin
- Heterochromatic regions are poorly transcribed (expressed)
- Euchromatic regions are highly expressed

Eukaryotic Genome



http://pixshark.com/euchromatin-and-heterochromatin.htm

Eukaryotic Genome

Conclusions

- Eukaryotes are distinguished by the presence of prominent nuclei
- have larger genomes, tandem repeats and introns in their protein-coding genes

Epichromosomal Elements (EEs)

Introduction

- Genome is the total collection of genetic material and is made up of
  - Chromosomes
  - Epichromosomal Elements

Epichromosomal Elements (EEs)

Prokaryotic EEs

- Plasmids
  - Self replicating
  - Additional rings
- Bacteriophages
  - Host colonization
- Transposons
  - Parasitic DNA elements

Epichromosomal Elements (EEs)

Eukaryotic EEs

- Organellar DNA
  - Mitochondria
  - Chloroplasts
- Plasmids
  - yeast
- Transposons
- Viral genomes
  - retroviruses

Epichromosomal Elements (EEs)

Eukaryotic EEs

- membrane-bound organelles
- hundreds to thousands of them
- Site of respiration (mitochondria) or photosynthesis (chloroplast)
- mtDNA
- cpDNA

Epichromosomal Elements (EEs)

Endosymbiont hypothesis

- originally proposed in 1883 by Andreas Schimper, but extended by Lynn Margulis in the 1980s
- Mitochondria and Chloroplast are derived from endosymbiotic bacteria

Epichromosomal Elements (EEs)

Organelle Genome

- Circular
- Double stranded
- Supercoiled
- No histones
- Multiple copies

of mtDNA/cell or cpDNA/cell

Epichromosomal Elements (EEs)

Genome	Size & Organization
Plant plastid	150 kb circle
Plant mitochondria	150 – 2000 kb multipartite
Human mitochondria	17 kb circle
saccharomyces mitochondria	75 kb circle

Epichromosomal Elements (EEs)

Organelle Genome

- Encode only a subset of genes required to elaborate a functional organelle
  - rRNAs, tRNAs, ribosomal proteins, membrane-associated respiratory or photosynthetic components

Epichromosomal Elements (EEs)

Organelle Genome

- Other components encoded by nuclear genome, translated in the cytosol and imported into the organelle
- 10% of nuclear genes devoted to mitochondrial function; 15% to plastid function

Epichromosomal Elements (EEs)

Conclusions

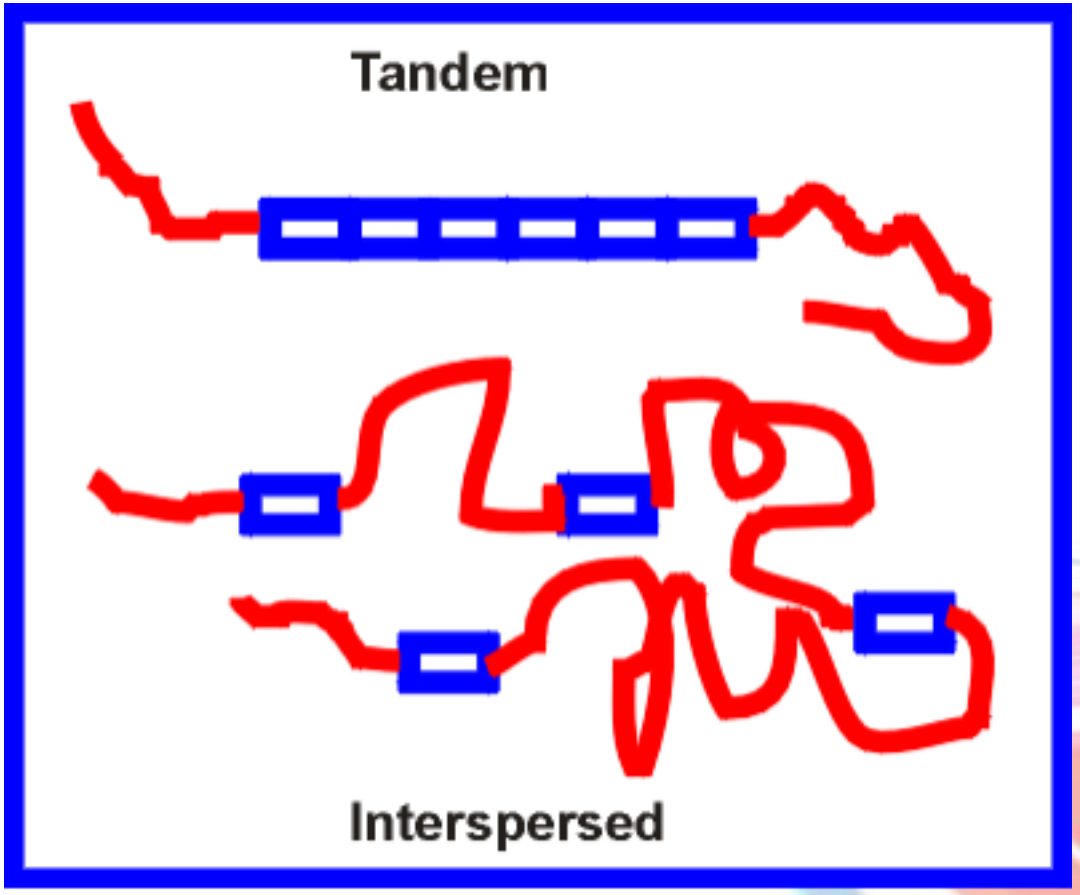
- Organelle genome is similar to prokaryotes
- High copy number
- mtDNA and cpDNA depends on Nuclear DNA

Genome Repeats

Sequence Repeats

- Repeats skew the base composition
- Contribute to differences in buoyant densities
- Repeat containing DNA can be separated as **satellite DNA** on the bases of these densities

Genome Repeats



http://mcb1.ims.abdn.ac.uk/djs/web/lectures/repeats1.html#anchor10305

Genome Repeats

Satellite DNA

- One to several thousand bp long
- Tandem array of 100 million bases long
- Near centromere and telomere
- Minisatellite and Nicrosatellite

Genome Repeats

Minisatellite

- 15 bases long in array of several hundred to thousands kb
- Typically in euchromatin
- VNTR e.g, used to identify human individuals in forensics

Genome Repeats

Microsatellite

- 2-6 bases long in arrays of 10-100 bases
- Inherited to offsprings
- Useful markers for genetic analysis and evolutionary studies
- Found in telomeres TTAGGG
- SSRs and STRs

Genome Repeats

Transposable Elements (TEs)

- Large portion of eukaryotic genome
- Thought to play important role in evolution of these genomes
- Move (Jump) from one location to another faster than chromosome replicate

Genome Repeats

Transposable Elements (TEs)

- Have a potential to increase in number
- Make up a large proportion of eukaryotic genome
- Detectable until blend into genome due to mutations

Genome Repeats

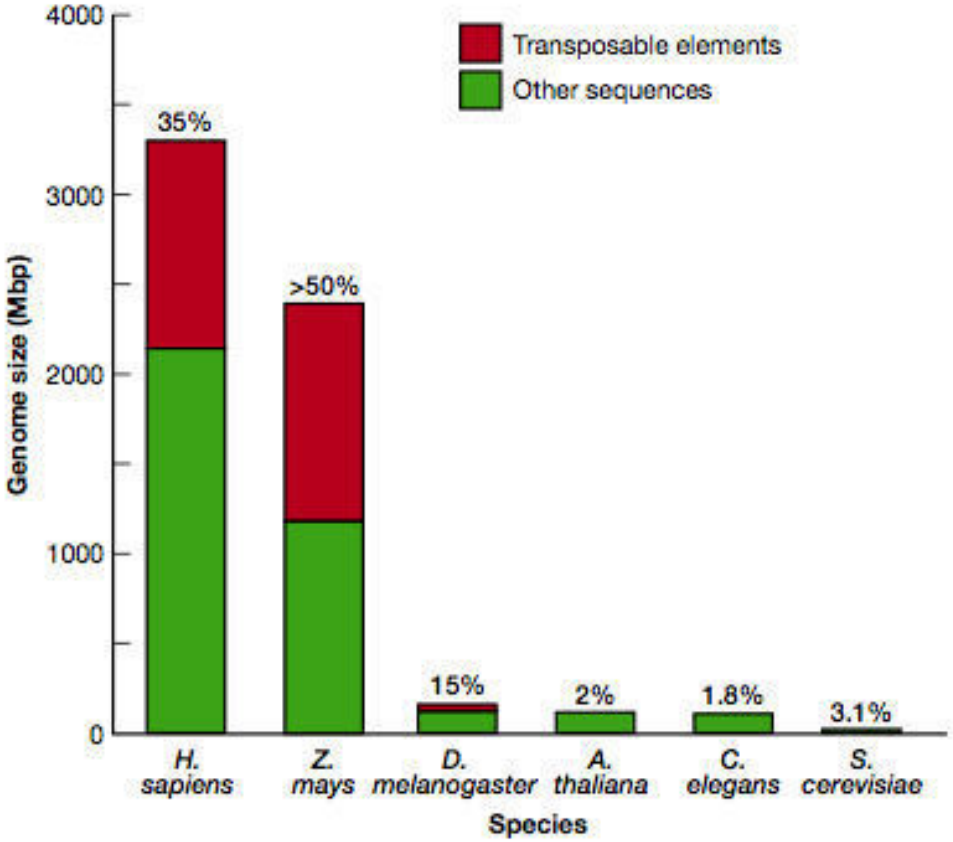


Figure 10.1. Percentages of representative genomes that are made up of transposable elements. The genomes include those of humans, maize, the fruit fly *Drosophila*, the model plant *Arabidopsis*, the nematode *C. elegans*, and budding yeast *S. cerevisiae*, respectively. (Adapted, with permission, from Kidwell and Lisch 2000 [copyright Elsevier Science].)

Genome Repeats

Conclusions

- Large proportion of eukaryotic genome is composed of repeats
- Different repeats act as markers to detect genetic variation and are also used to study evolution

Transposable Elements (TEs)

Introduction

Any segment of DNA, able to transpose

- **Insertion Sequences (IS elements)**
  - The simplest transposable elements
  - Code only for the ability to transpose
  - prokaryotes

Transposable Elements (TEs)

IS elements

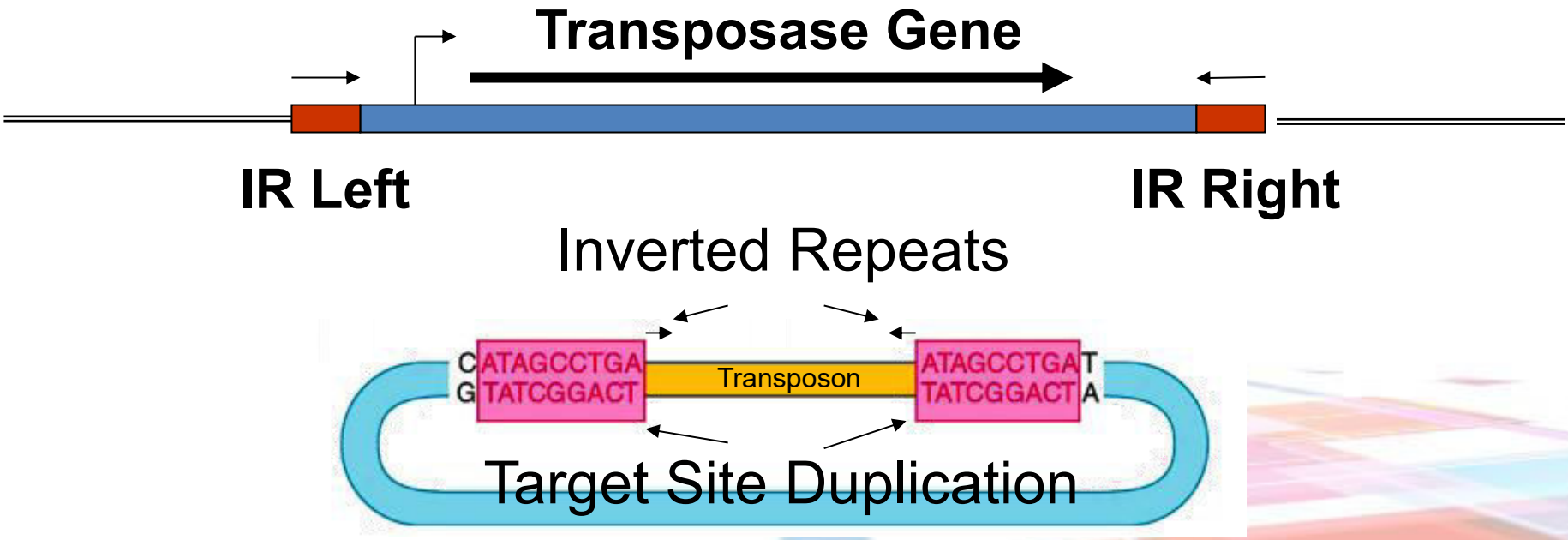
- usually very small (< 1 Kb to 2 Kb)
- Flanked by inverted repeat sequences (IRs)
- Encode at least one gene that provides their own transposition functions

Transposable Elements (TEs)

IS elements

- Do not code for noticeable (phenotypic) traits
- Can cause mutations by transposition into genes

Transposable Elements (TEs)



Transposable Elements (TEs)

- **Transposons**
  - More complex transposable elements
  - Code for the ability to transpose
  - Code for other detectable traits as well as ability to transpose

Transposable Elements (TEs)

Eukaryotic TEs

- 2 Classes; Class I and Class II
- Class I
  - Encode reverse transcriptase
  - Use RNA-mediated mechanisms of transcription

Transposable Elements (TEs)

Eukaryotic TEs

- **Class I**

Three classes

- LTR (Long Terminal Repeats) retrotransposons
- Retrotransposons
- Retrovirus like Elements

Transposable Elements (TEs)

Eukaryotic TEs

- **SINES** (Short Interspersed Nuclear Elements
  - 80-300 bp
  - ***Alu* repeats**
- **LINES** (Long Interspersed Nuclear Elements)
  - 6-8 KB long

Transposable Elements (TEs)

Eukaryotic TEs

- **Class II TEs**
  - Ac-Ds in maize
  - P elements in Drosophila
- **MITES** (Miniature Inverted repeat TEs)
  - Features of both Class I and II
  - 400 bp

Transposable Elements (TEs)

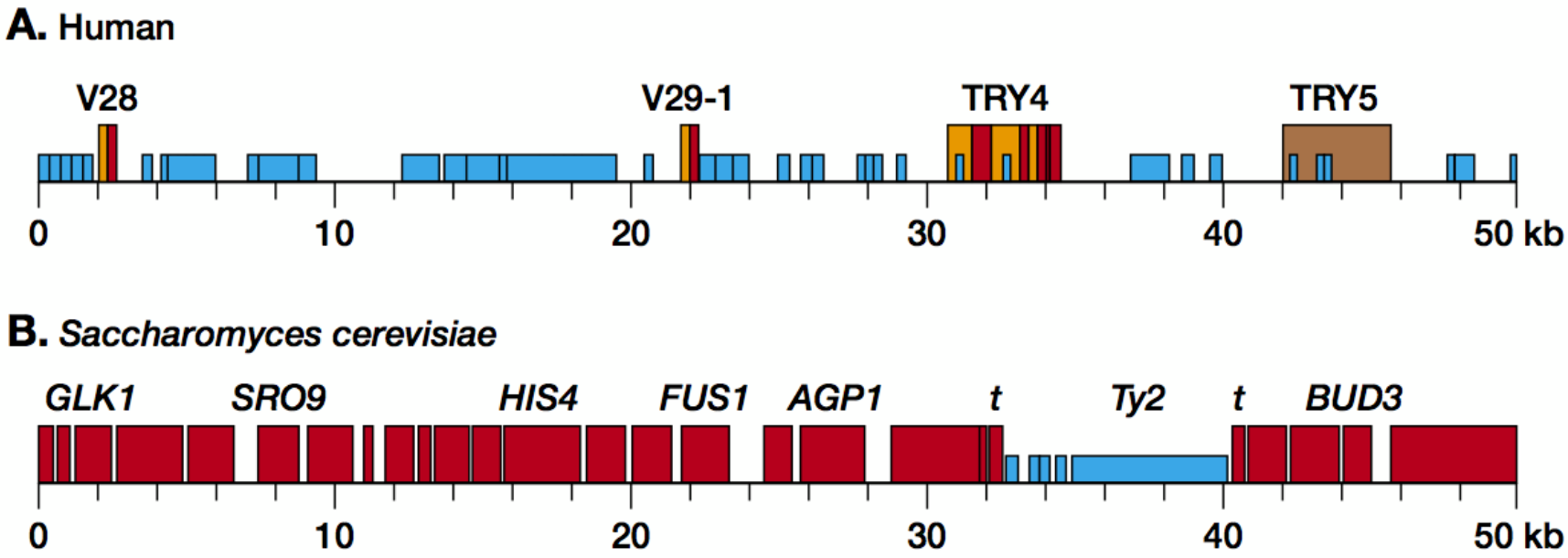


Figure 10.3. Comparison of genome composition in four genomes. (*A*) Human β T-cell receptor locus on chromosome 7. V28 and V29.1 encode parts of the β T-cell receptor proteins that are joined during development of the immune system (Rowen et al. 1996). TRY4, the gene for trypsinogen, and TRY5, a pseudogene related to the trypsinogen family, are not related to the receptor sequence. Why they are located here is not known. (*B*) Segment of yeast chromosome III (Oliver et al. 1992). (*C*, *D*) 50-kb fragments of the maize and *E. coli* chromosomes, respectively (SanMiguel et al. 1996; Blattner et al. 1997). The maize repeats are LTR retrovirus-like elements (Fig. 10.2) that have inserted within the last 3 million years (SanMiguel et al. 1998). (Redrawn, with permission, from Brown 1999 [BIOS Scientific].)

Transposable Elements (TEs)

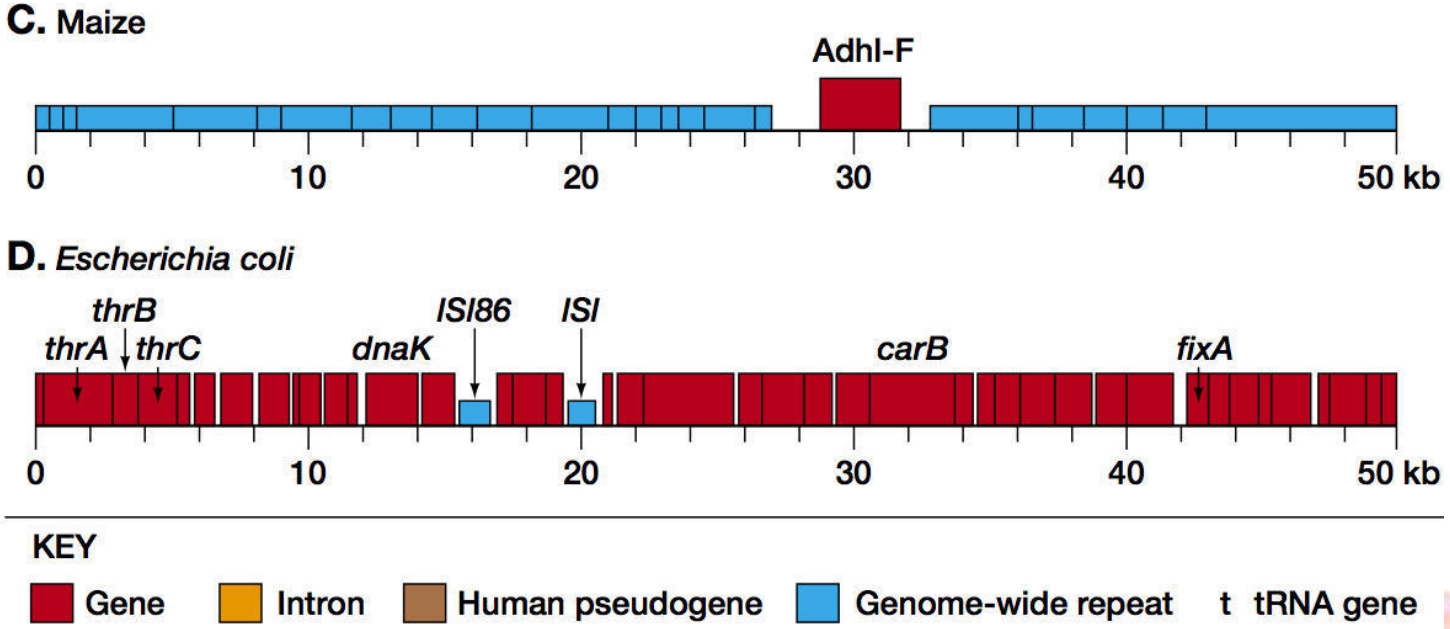


Figure 10.3. Comparison of genome composition in four genomes. (*A*) Human β T-cell receptor locus on chromosome 7. V28 and V29.1 encode parts of the β T-cell receptor proteins that are joined during development of the immune system (Rowen et al. 1996). TRY4, the gene for trypsinogen, and TRY5, a pseudogene related to the trypsinogen family, are not related to the receptor sequence. Why they are located here is not known. (*B*) Segment of yeast chromosome III (Oliver et al. 1992). (*C*, *D*) 50-kb fragments of the maize and *E. coli* chromosomes, respectively (SanMiguel et al. 1996; Blattner et al. 1997). The maize repeats are LTR retrovirus-like elements (Fig. 10.2) that have inserted within the last 3 million years (SanMiguel et al. 1998). (Redrawn, with permission, from Brown 1999 [BIOS Scientific].)

Transposable Elements (TEs)

Conclusions

- Transposable elements make up a significant part of eukaryotic genome
- Move within and across genomes
- Cause genome expansion

Eukaryotic Gene Structure

Eukaryotic genes

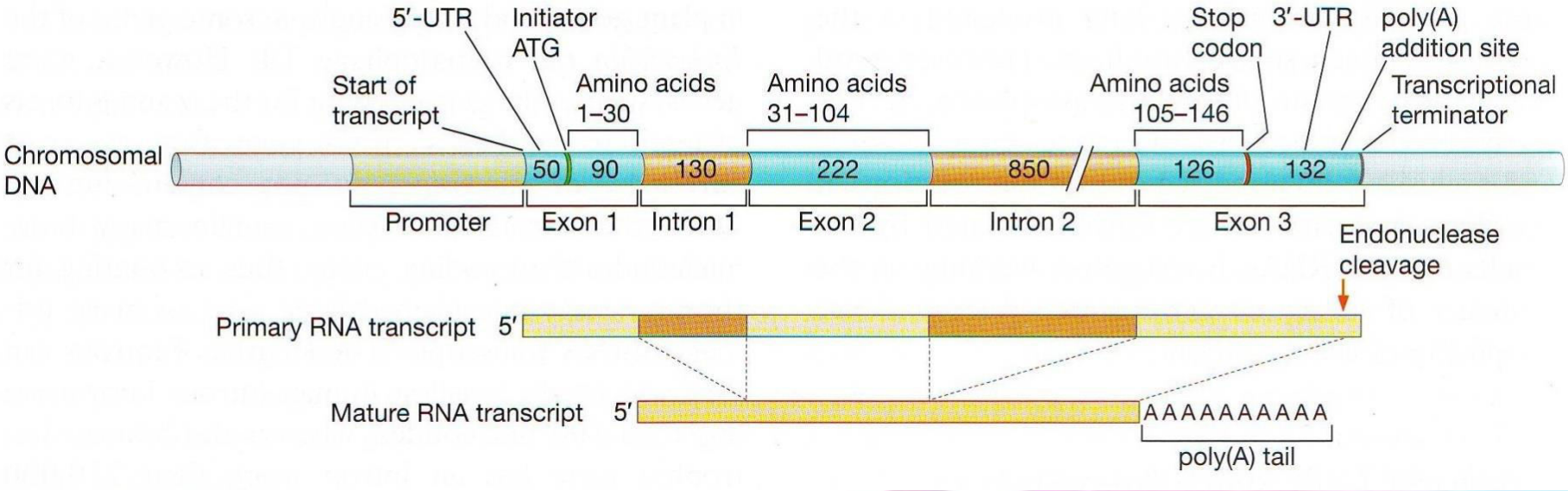
- Eukaryotic genes are complicated
- They possess exons which are protein coding regions and are interrupted with introns

Eukaryotic Gene Structure

Eukaryotic genes

- Whole gene is transcribed as a large transcript
- Introns are spliced leaving an ORF that is translated

Eukaryotic Gene Structure



Human β-globin gene

Eukaryotic Gene Structure

Eukaryotic genes

- Introns start on the 5’ ends as GT and end as AG towards 3’ end **GT----------AG**
- This trend is highly conserved

Eukaryotic Gene Structure

Introns proportion

- A small proportion in yeast, only 239 introns in genome
- Human genes have hundreds, often 95% of genes
- Might have embedded genes

Eukaryotic Gene Structure

Intron origin

- **Intron-early**
  - used to assemble first genes from ancient conserved introns
- **Intron-late**
  - Broke up previously continuous genes by inserting into them

Eukaryotic Gene Structure

Number of Genes

- Degree of compactness contributes to variation in gene density
- Compact genomes have higher genome density

Eukaryotic Gene Structure

Organism	Genome size (haploid MB)	Predicted genes
A. thaliana (plant)	130	~25,000
C. elagans (worm)	100	18,424
Drosphila melangaster	180	13,601
Escherichia coli	4.7	4,288
Homo sapiens (human)	3000	45,000 – 120,000
S. cerevisiae (yeast)	13.5	6,241

Mount: Table 10.3, 11.3 2nd edn

Eukaryotic Gene Structure

Pseudogenes

- Non functional genes
- Derived from functional genes through mutations following genome duplication
- Processed pseudogenes lack introns and promoters

Eukaryotic Gene Structure

Gene families

- Set of genes having similar sequences and functions
- Gene families arise from gene duplication and subsequent divergence

Eukaryotic Gene Structure

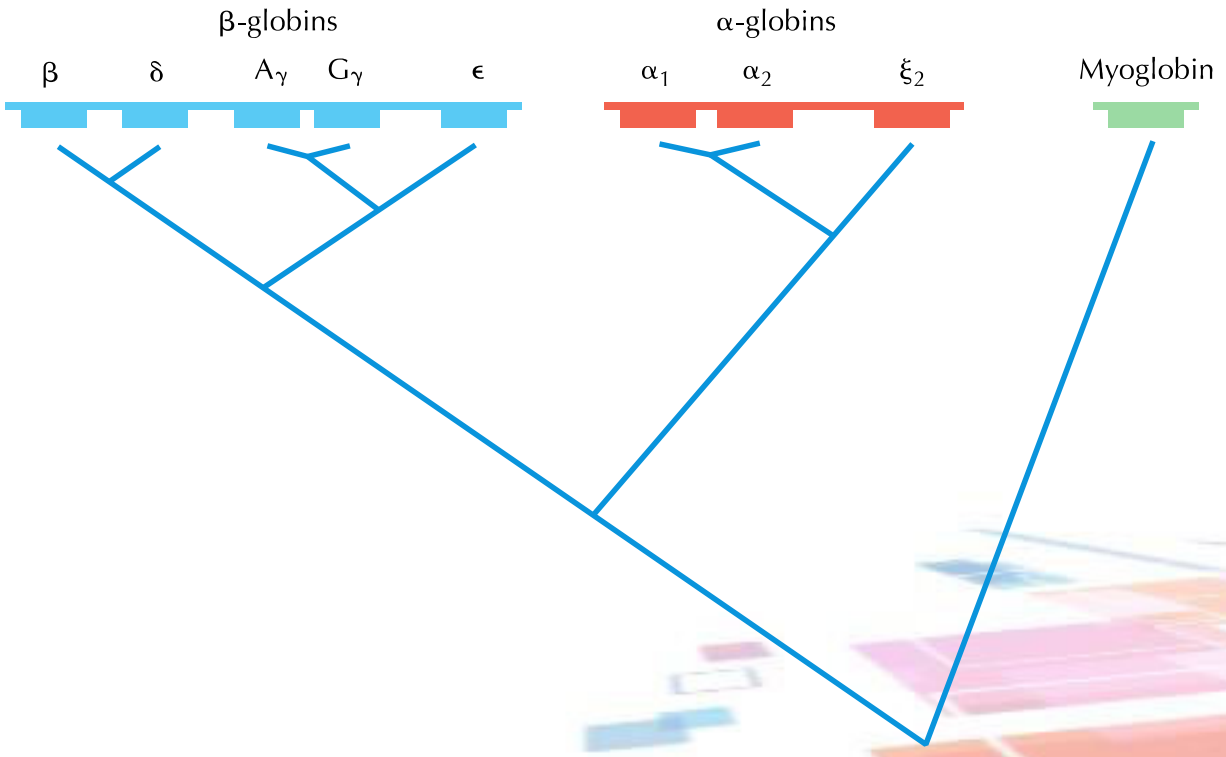


FIGURE 27.29. Gene duplications during the evolution of the human globin gene families. The initial split gave rise to two lineages, one leading to the modern gene for myoglobin and the other to the globin genes. Subsequently, the proto-α-globin and proto-β-globin lineages split following a duplication. Other duplications took place within the α and β lineages. (Modified from Strachan T. and Read A.P. *Human Molecular Genetics 2*, Fig. 14.16, © 1999 Garland Science.)

Evolution © 2008 Cold Spring Harbor Laboratory Press

Eukaryotic Gene Structure

Conclusions

- Eukaryotic genes have exons and introns
- Introns make up a significant portion of Human genome
- Pseudogenes are non functional genes
- Similar genes make gene families

Comparative Genomics

Introduction

- Comparison of gene number, gene content and gene location in both prokaryotic and eukaryotic groups of organisms

Comparative Genomics

Introduction

- Availability of genome makes possible a comparison of all the proteins (proteome) encoded by one organism with those of another

Comparative Genomics

Orthologs

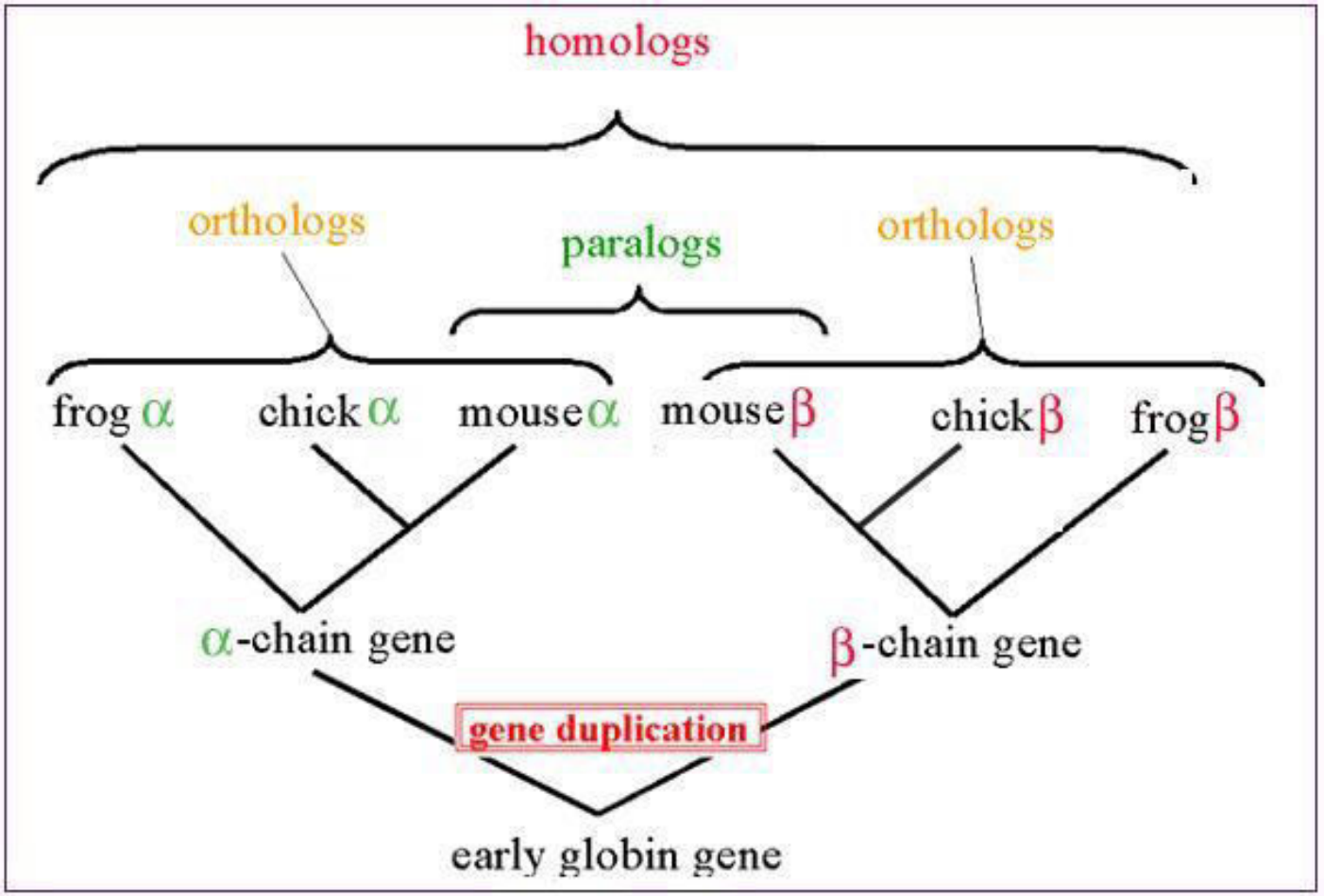
- Genes in two organisms that are so similar that they must have the same function and evolutionary history are called orthologs (Fitch, 1970)

Comparative Genomics

Paralogs

- Gene families originating from rare gene duplication events over the evolutionary time are called paralogs
- Unlike pseudogenes, 2nd copy of gene remains functional

Comparative Genomics



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Comparative Genomics

Drosophila and yeast

- *Drosophila* has core proteome only twice the size of that of yeast
- *Drosophila* proteome is more similar to mammalian proteomes than worm or yeast

Comparative Genomics

Drosophila and *C. elegans*

- Despite the large differences between fly and worm, they use a core proteome of similar size
- Nearly 30% of fly genes have putative orthologs in the worm

Comparative Genomics

Drosophila and Human

- Some human disease genes absent in *Drosophila*
- A number of previously unknown counterparts to human cancer and neurological disorders genes

Comparative Genomics

Conclusions

- Comparative genomics reveals the relationship among different organisms
- Fruit fly has more similarities with mammals

Genetics and Genomics

Functional Genomics

Functional Genomics

Functional Genomics

- Functions of genes, their regulation and end products.
- Functional genomics analyzes all genes in genomes to determine their functions and their gene control and expression.

Functional Genomics

Functional Genomics

- Classically, genetics analysis begins with a phenotype and moves for identification of genes.
- New approaches are needed to work in the opposite direction, from genes to phenotype.

Functional Genomics

Functional Genomics relies on Molecular Biology, Biochemistry, Genetics and Bioinformatics tools

- Functional genomics relies on molecular biology lab research and sophisticated computer analysis by bioinformatics tools.

Functional Genomics

Functional Genomics

- Fusion of biology with maths and computer science is used for many things. Examples:
- Finding genes within a genomic sequence.
- Aligning DNA/proteins sequences.

Functional Genomics

Functional Genomics – what is included

- Subtracted cDNA libraries
- Differential display
- Representational difference analysis
- Suppression subtractive hybridization

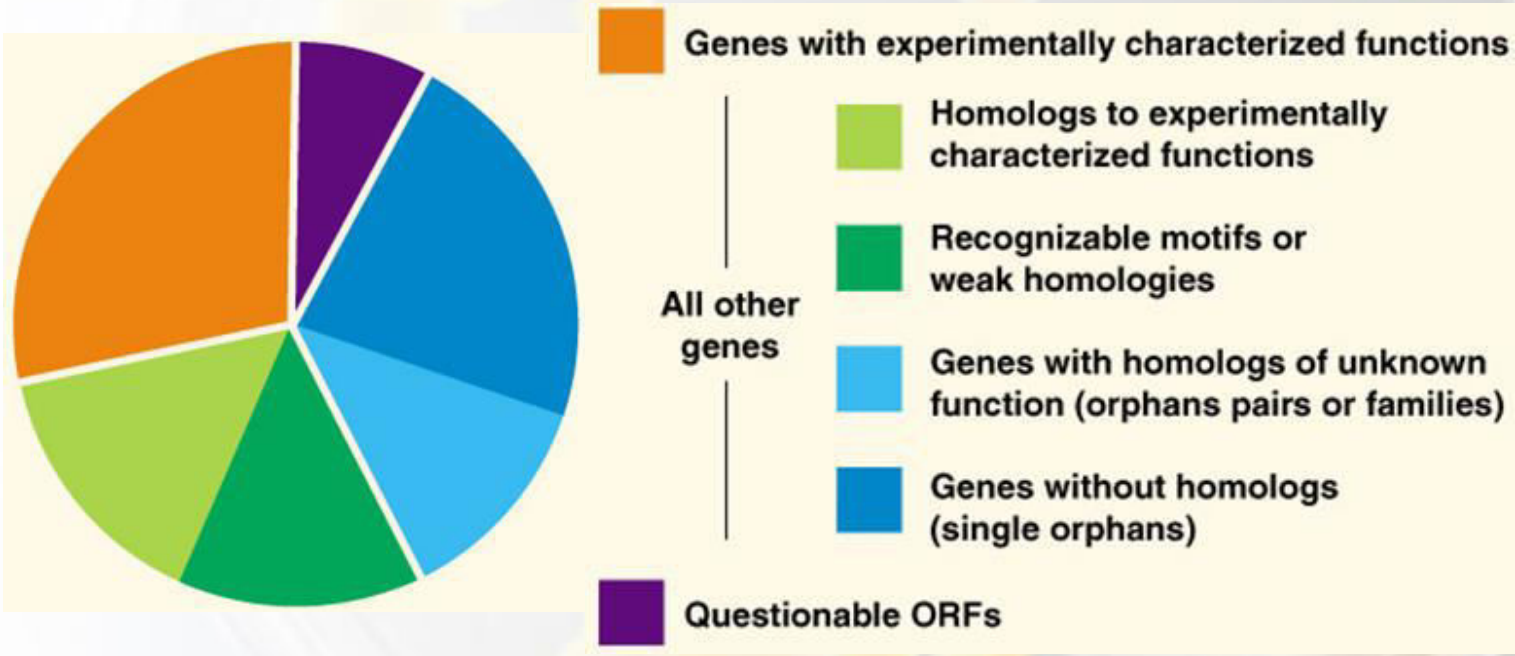
Functional Genomics

Functional Genomics – what is included

- cDNA Microarrays
- Serial analysis of gene expression
- 2-D Gel electrophoresis

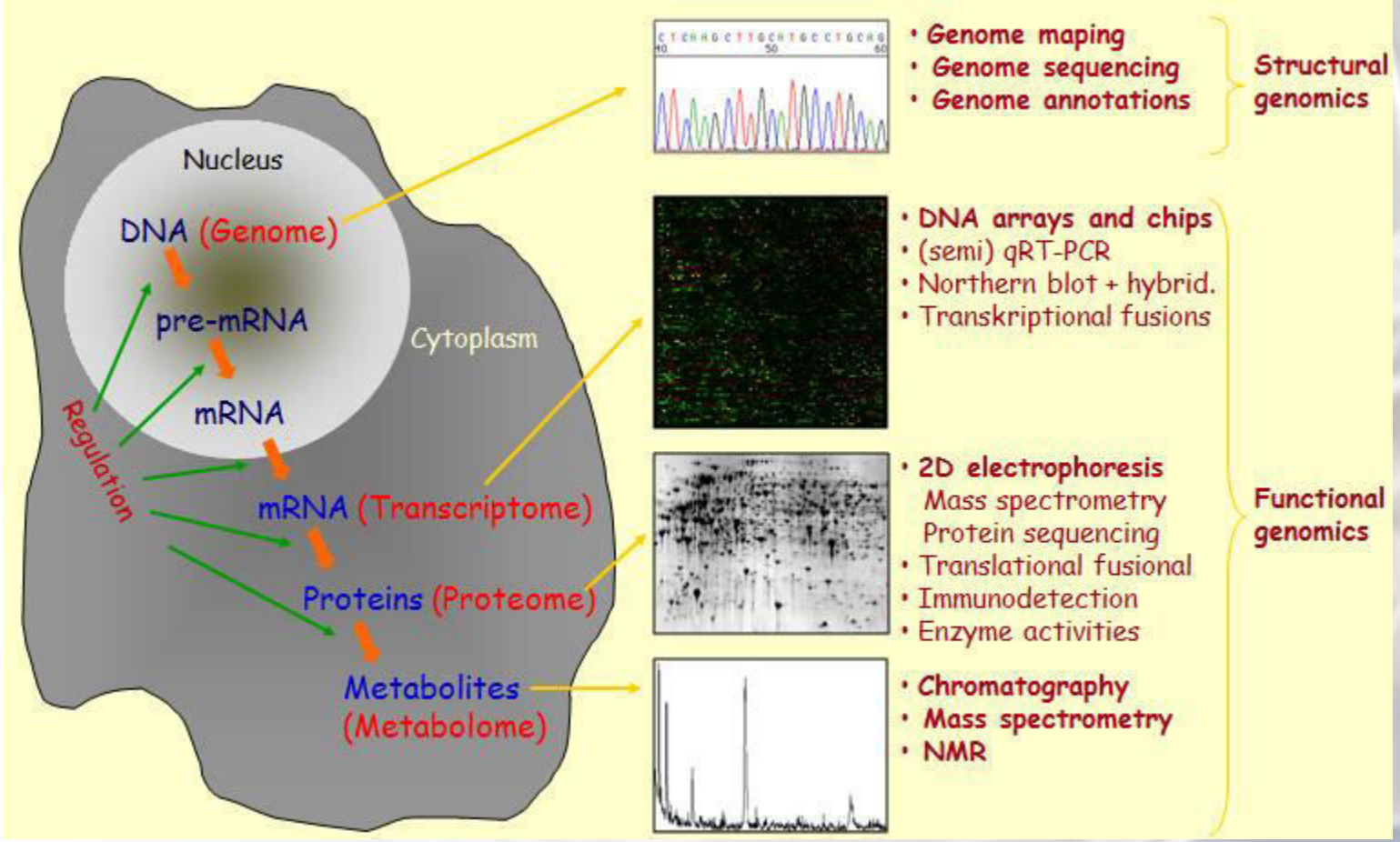
Functional Genomics

Yeast Genome with Experimentally Characterized Functions



Functional Genomics

Functional Genomics



Functional Genomics

Functional Genomics - Conclusion

- Function of gene products.
- Describing interactions between genes and gene products in the cell, between cells and between organisms.

Functional Genomics

Functional Genomics - Conclusion

- Considering phylogenetic relationships.

Genetics and Genomics

Structural Genomics

Structural Genomics

Structural Genomics

- The ultimate goal of genomic studies is to determine the nucleotides sequences of entire genomes of organisms.
- The genetic and physical mapping and sequencing of chromosomes.

Structural Genomics

Genetic Mapping

- Genetic mapping - approximate locations of genes, relative to the locations of other genes, based on the rates of recombinations.

Structural Genomics

Physical Mapping

- Physical mapping is based on the direct analysis of DNA.
- Physical mapping places genes on the genomes in relation to distances measured in bp, kbp, and mbp.

Structural Genomics

Structural Genomics incudes

- Distinct components of genomes
- Abundance and complexity of mRNA
- Genome sequences
- Gene numbers
- Coding and non-coding DNA

Structural Genomics

Structural Genomics – Complex Genomes

- Complex genomes have roughly 10x to 30x more DNA than is required to encode all the RNAs or proteins in the organism

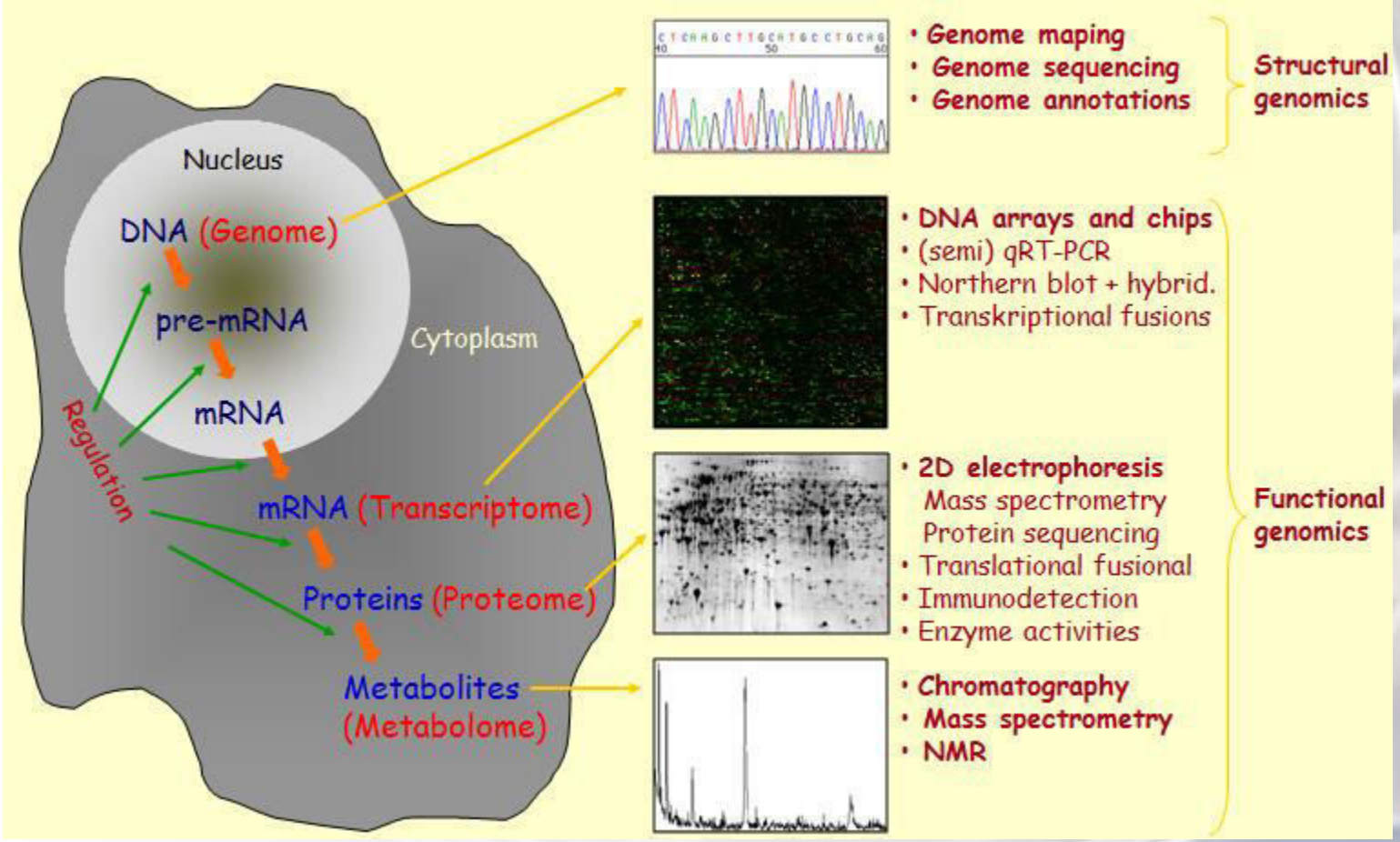
Structural Genomics

Structure of Complex Genomes

- Introns in genes
- Regulatory elements of genes
- Multiple copies of genes, including pseudogenes
- Intergenic sequences
- Interspersed repeats

Structural Genomics

Structural Genomics



Structural Genomics

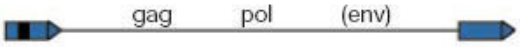
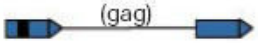
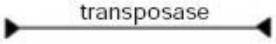
Structural Genomics – Transposable Elements

Vast majority of TEs can be classified into four families:

- LINEs (Long Interspersed Nuclear Elements, autonomous)
- SINEs (Short Interspersed Nuclear Elements, use LINE proteins for life cycle)
- LTR elements (Long Terminal Repeats; derived from retroviruses)
- DNA transposons (replicate without RNA intermediary)

Structural Genomics

Structural Genomics – Repetitive DNA in Humans

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
SINEs	Non-autonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Structural Genomics

Structural Genomics - Conclusion

- To determine the nucleotides sequences of genomes.
- To study genetic maps
- To study physical maps.

Genetics and Genomics

Comparative Genomics

Comparative Genomics

Comparative Genomics

- Comparison of gene numbers, gene locations and biological functions of genes, in the genomes of different organisms.

Comparative Genomics

Comparative Genomics

- To identify and compare groups of genes that play a unique biological role in different organisms.

Comparative Genomics

Homology

- Homology is the relationship of any two characters (genes or proteins) that have descended, usually through divergence, from a common ancestor/ ancestral character.

Comparative Genomics

Homologues

- Homologues are thus components or characters (such as genes/proteins with similar sequences) that can be attributed to a common ancestor of the two organisms during evolution.

Comparative Genomics

Homologoues can be

- Orthologues
- Paralogues
- Xenologues
- Analogues

Comparative Genomics

Orthologues

- Orthologues are homologues that have evolved from a common ancestral gene by speciation.
- They usually have similar functions.

Comparative Genomics

Paralogues

- Paralogues are homologues that are related or produced by duplication within a genome followed by subsequent divergence.
- They often have different functions.

Comparative Genomics

Xenologues

- Xenologues are homologous that are related by an interspecies (horizontal transfer) of the genetic material for one of the homologues.
- The functions of the xenologues are quite often similar.

Comparative Genomics

Analogues

- Analogues are non-homologues genes/proteins that have descended convergently from an unrelated ancestor.
- Similar function, different sequence or structure.

Comparative Genomics

Comparative Genomics – Evolutionary Trends

- Comparison of genomes belonging to relatively similar group, like mammals, reveals some evolutionary trends.

Comparative Genomics

Comparative Genomics – Functional elements

- Provides a powerful and general approach to identify functional elements without previous knowledge of functions.

Comparative Genomics

Comparative Genomics – Ancestral Genome

- Reconstruction of an ancestral genome for a group of organism.

Comparative Genomics

Conclusion

- Comparative genomics is the comparison of gene numbers, gene locations and biological functions of genes, in the genomes of different organisms.

Genetics and Genomics

Population Genomics

Population Genomics

Population Genomics

- Study of genomes of a specific population, strains, varieties or organisms.
- Study about the genetic diversity.
- Understanding new insights into disease and drug response.

Population Genomics

Population Genomics

- Deeper insights about genomic size.
- Further complexities in the genomes of the organisms.
- Variations between individuals or strains.

Population Genomics

Population Genomics

- Human genome is 200 times larger than yeast but 200 times smaller than Amoeba.
- Less than 2% of human genome is coding sequence

Population Genomics

1000 Genomes - Population Genomics

- International research consortium sequenced the genomes of at least 1000 people from around the world.
- Detailed and medically useful pictures of human genome variations.

Population Genomics

1000 Genomes - Population Genomics

- Any two humans are more than 99% identical at genetic level.
- Genetic variations may explain individual differences in susceptibility to diseases, responses to drugs.

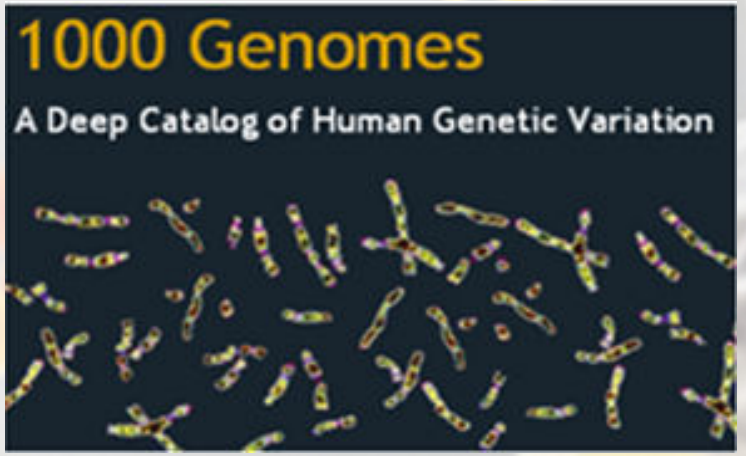
Population Genomics

HapMap Project - Population Genomics

- The HapMap project has already discovered many regions of the genome containing genetic variations associated with common human diseases.

Population Genomics

Population Genomics - HapMap Project and 1000 Genomes Project



Population Genomics

Goals of Population Genomics

- Produce a catalog of variants present at 1% or greater frequency in the human population.
- Down to 0.5 percent or lower within genes.
- Increase sensitivity of disease discovery.

Population Genomics

Goals of Population Genomics

- Provide better understanding of very rare genetic diseases.
- Understand contribution of common variants to common diseases like diabetes and heart diseases.

Population Genomics

Goals of Population Genomics

- Identify SNP but also large differences like rearrangements, deletions or duplications

Genetics and Genomics

Metagenomics

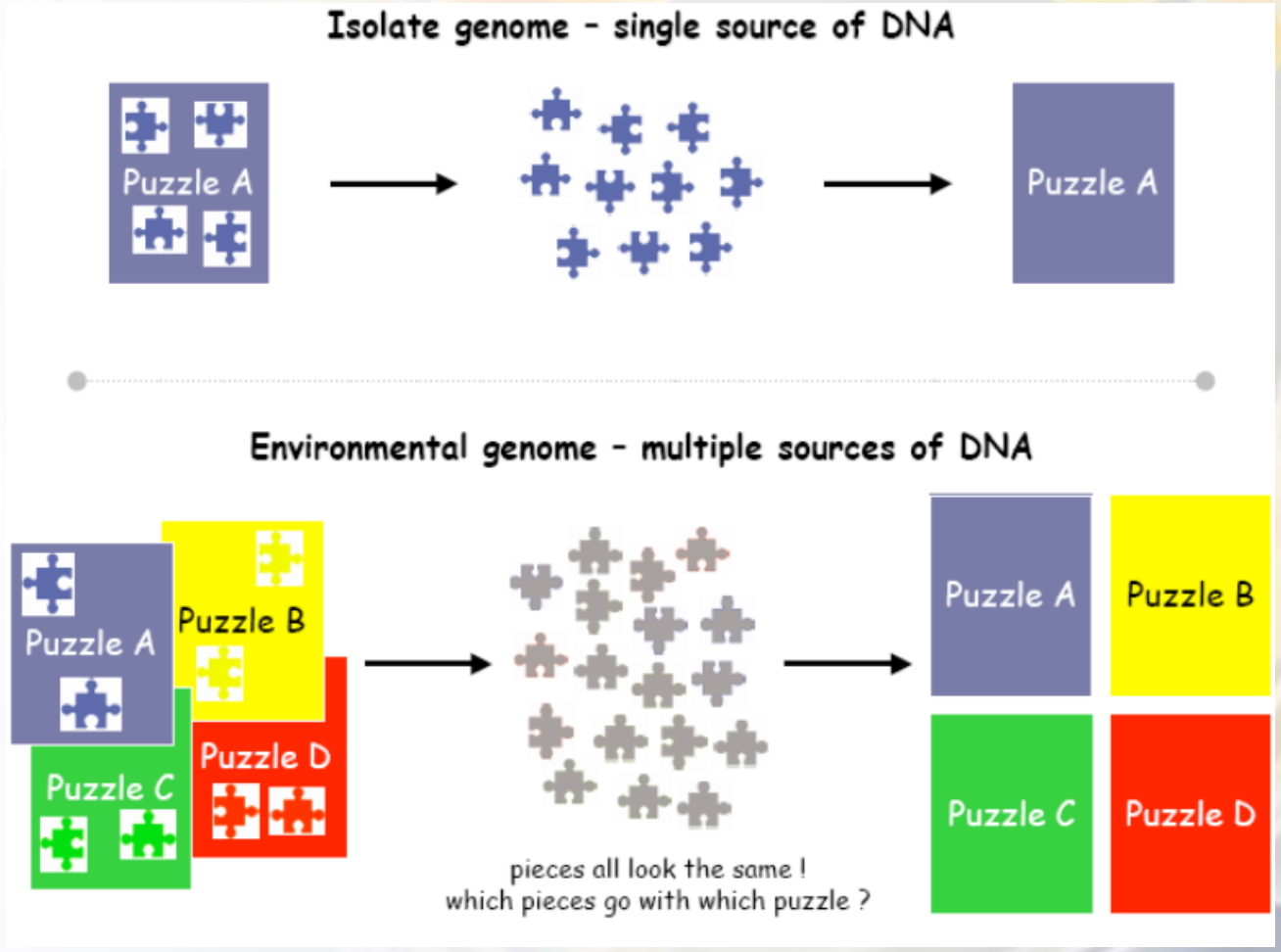
Metagenomics

Metagenomics

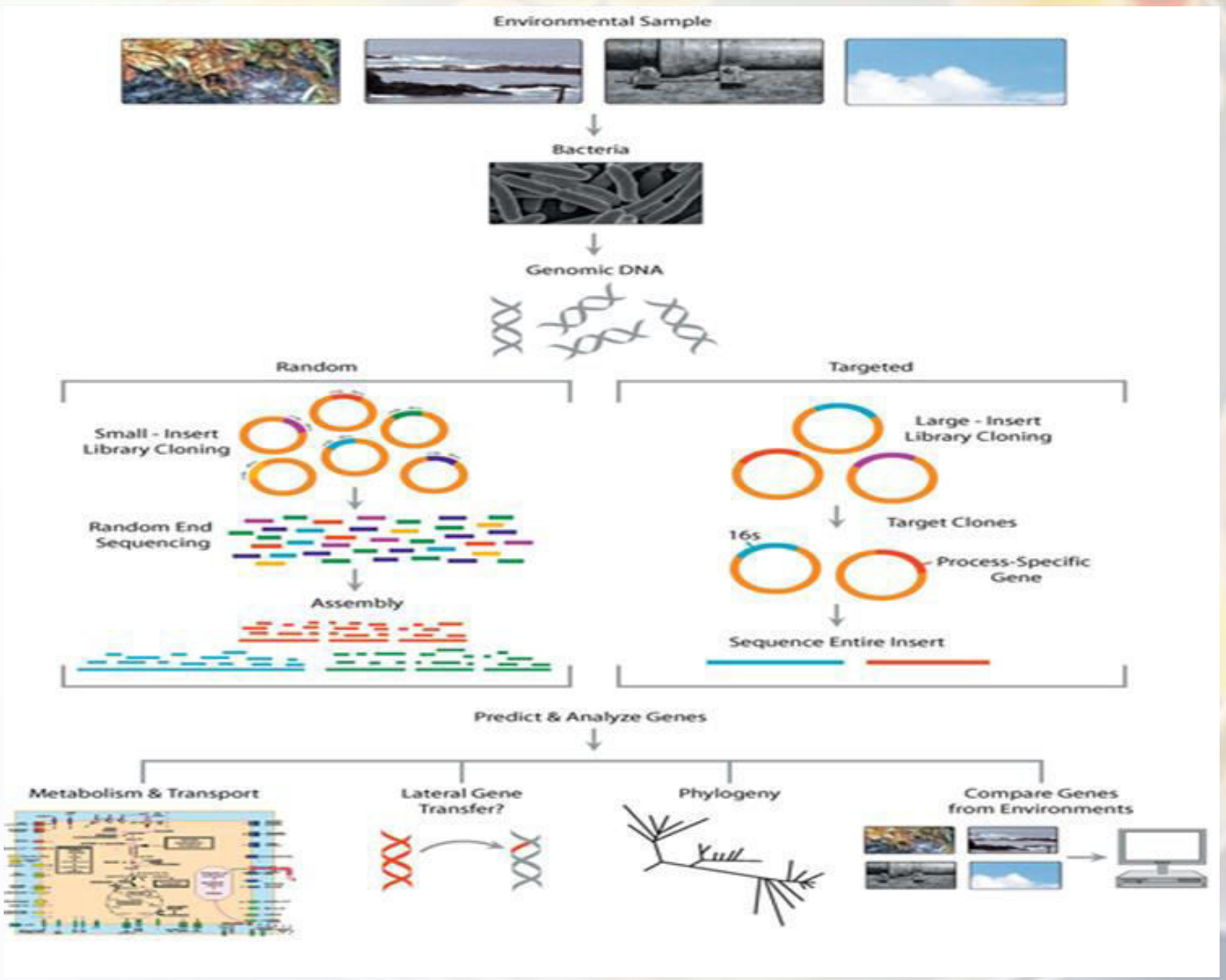
- Metagenome- environmental genome.
- Collection of genes sequenced from the environment could be analyzed in a way analogous to the study of a single genome.

Metagenomics

Metagenomics

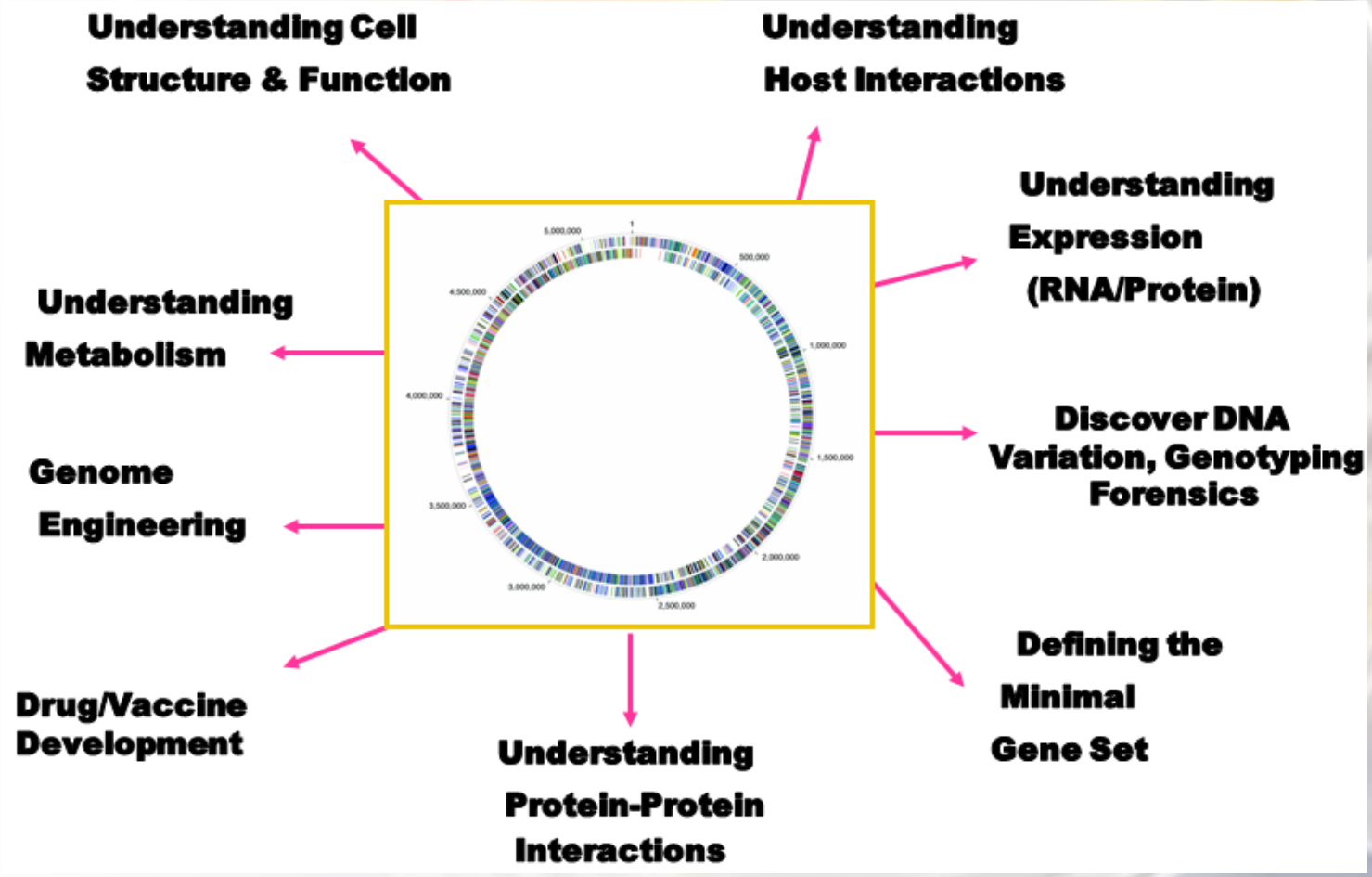


Metagenomics



Metagenomics

Why do Metagenomics



Metagenomics

Objectives of Metagenomics

- Examining phylogenetic diversity using 16s rRNA
- Diversity patterns of microorganisms for monitoring and predicting environmental conditions/change.

Metagenomics

Objectives of Metagenomics

- Examining genes/operons for desirable enzymes (cellulases, lipases, antibiotics, other natural products).
- Exploited for industrial or medical applications.

Metagenomics

Objectives of Metagenomics

- Examining secretory, regulatory, and signal transduction mechanisms associated with samples or genes of interest.

Metagenomics

Objectives of Metagenomics

- Examining bacteriophage or plasmid sequences. These potentially influence diversity and structure of microbial communities.

Metagenomics

Objectives of Metagenomics

- Examining potential lateral gene transfer events. Knowledge of genome plasticity may give us an idea of selective pressures for gene capture and evolution within a habitat.

Metagenomics

Objectives of Metagenomics

- Examining metabolic pathways.
- Directed approach towards designing culture media.
- Examining genes that predominate in a given environment compared to others.

Metagenomics

Metagenomics Data

- Metagenomic data can be used towards designing low and high throughput experiments focused on defining the roles of genes and microorganisms in the establishment of dynamic microbial community.

Metagenomics

Conclusion

- Metagenome- environmental genome

Genetics and Genomics

Why Sequence Genomes

Why Sequence Genomes

Why Sequence Genomes

- To identify gene numbers, their locations on genomes, and to study their functions.
- Genes regulation

Why Sequence Genomes

Why Sequence Genomes

- DNA sequence
- Genome organization
- Chromosomal structure and organization

Why Sequence Genomes

Why Sequence Genomes

- Noncoding DNA types, amount, distribution and functions.
- Coordination of gene expression, protein synthesis, and post-translational events.

Why Sequence Genomes

Why Sequence Genomes

- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation

Why Sequence Genomes

Why Sequence Genomes

- Proteins structure and function.
- Proteomes (total protein content and function) in organisms.
- Correlation of SNPs with health and disease

Why Sequence Genomes

Why Sequence Genomes

- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases

Why Sequence Genomes

Novel Diagnostics

- Complex systems biology, developmental genetics.
- To provide platform for microchips and DNA microarrays.
- Gene expression - RNA

Why Sequence Genomes

Novel Diagnostics

- Complex systems biology, developmental genetics and genomics

Why Sequence Genomes

Novel Therapeutics

- Drug target discovery
- Rational drug design
- Molecular docking
- Gene therapy
- Stem cell therapy

Why Sequence Genomes

Understanding Metabolism

- To understand the metabolism of cells and tissues within different organisms.

Why Sequence Genomes

Understanding mechanism of diseases

- Inherited diseases
- Infectious diseases
- Pathogenic bacteria
- Viruses

Why Sequence Genomes

Conclusions

- Better understanding of the genomes would be possible by sequencing of the genomes.

Genetics and Genomics

Major Techniques used for Genomes Characterization

Genomes Characterization -Techniques

Major Techniques used for Genomes Characterization

- Cloning
- Hybridization
- PCR amplification
- Sequencing
- Computational tool

Genomes Characterization - Techniques

Genomes Characterization

Techniques - Cloning

- Genomes digested with restriction enzymes and inserted in vectors to produce genomic libraries.
- BACs
- YACs

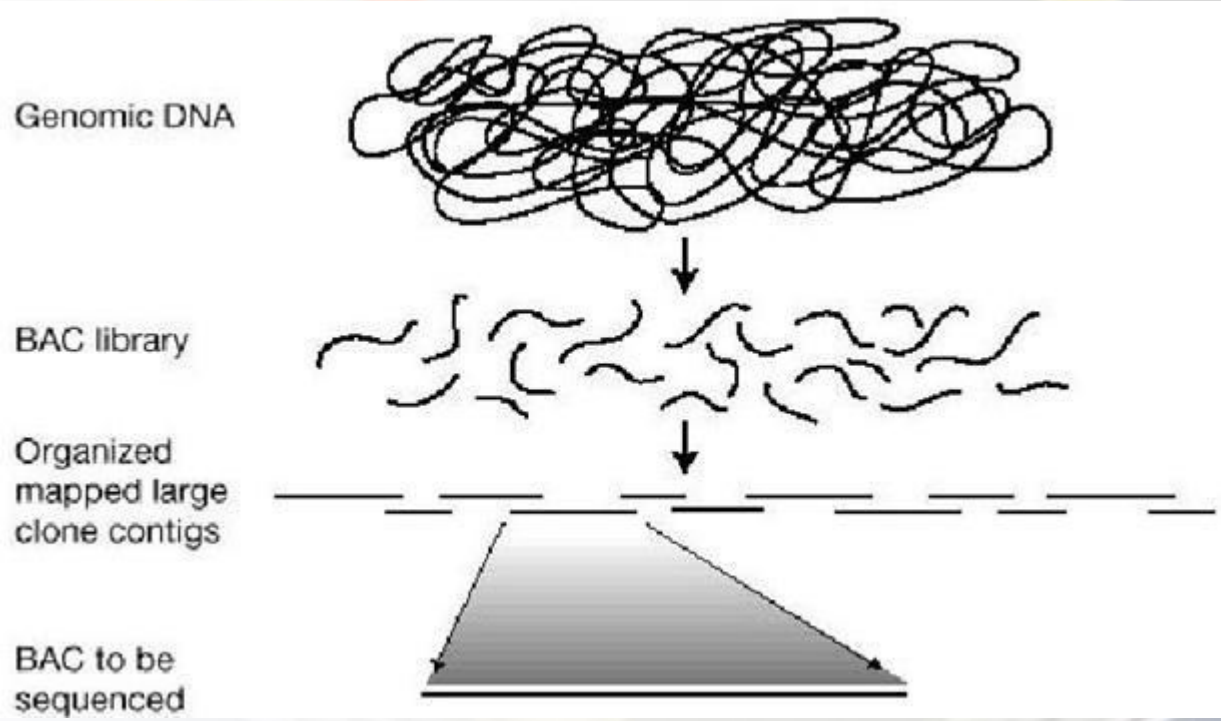
Genomes Characterization - Techniques

Genomes Characterization Techniques - Hybridization

- To arrange large contigs of genomes to produce genetic maps and physical maps of genomes.

Genomes Characterization - Techniques

Genomes Characterization Techniques – Hybridization: To arrange large contigs of genomes to produce genetic and physical maps



Genomes Characterization - Techniques

Genomes Characterization Techniques – PCR

- Technique to amplify the DNA. Different variants of the technique used

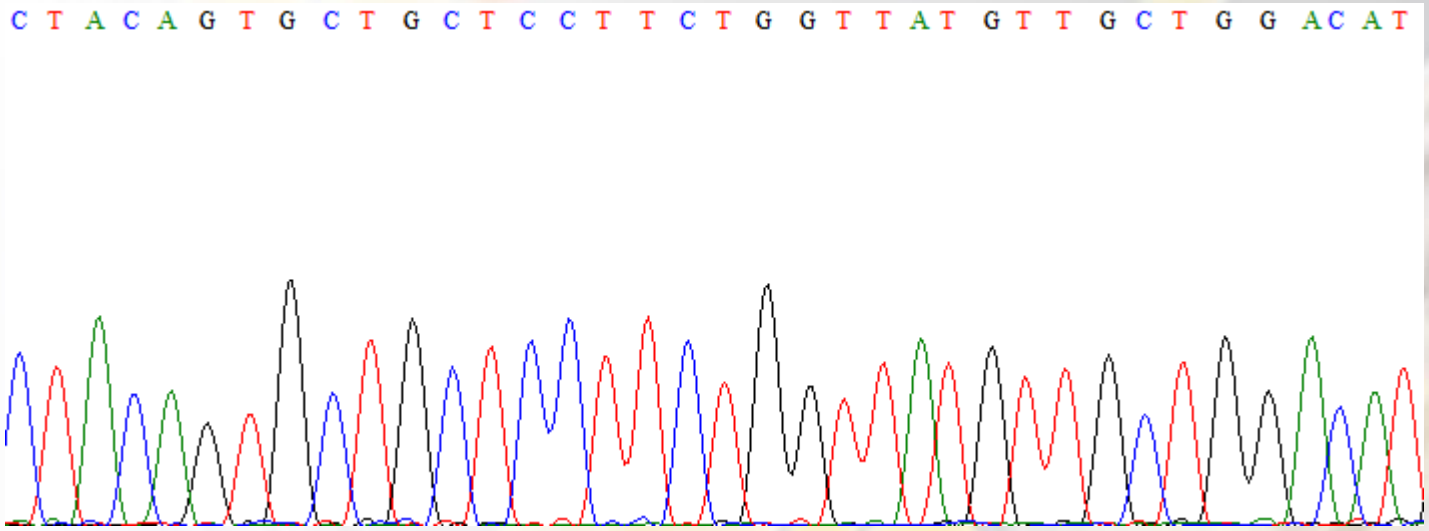
Genomes Characterization - Techniques

Genomes Characterization Techniques – DNA Sequencing

- One of the important technique used to characterize the genomes
- To study structure and function of genomes.

Genomes Characterization - Techniques

Genomes Characterization Techniques – DNA Sequencing: DNAs are amplified and sequenced



Genomes Characterization - Techniques

Genomes Characterization Techniques – Computational Tools

- Used to align the sequenced DNA to produce physical maps of the genomes.

Genomes Characterization - Techniques

Genomes Characterization Techniques – Conclusion

- Different techniques used for genomes characterizations.

Genetics and Genomics

Steps of Genomes Analysis

Genomes Analysis - Steps

Steps of Genomes Analysis

- Genome sequence assembled
- Identify repetitive sequences – mask out
- Gene prediction – train a model for each genome

Genomes Analysis - Steps

Steps of Genomes Analysis

- Look for EST and cDNA sequences
- Genome annotation
- Microarray analysis
- Metabolic pathways and regulation
- Protein 2D gel electrophoresis

Genomes Analysis - Steps

Steps of Genomes Analysis

- Functional genomics
- Gene location/gene map
- Self-comparison of proteome
- Comparative genomics

Genomes Analysis - Steps

Steps of Genomes Analysis

- Identify clusters of functionally related genes
- Evolutionary modeling

Genetics and Genomics

Steps of Genomes Analysis

Genomes Analysis - Steps

Steps of Genomes Analysis

- Genome sequence assembled
- Identify repetitive sequences – mask out
- Gene prediction – train a model for each genome

Genomes Analysis - Steps

Steps of Genomes Analysis

- Look for EST and cDNA sequences
- Genome annotation
- Microarray analysis
- Metabolic pathways and regulation
- Protein 2D gel electrophoresis

Genomes Analysis - Steps

Steps of Genomes Analysis

- Functional genomics
- Gene location/gene map
- Self-comparison of proteome
- Comparative genomics

Genomes Analysis - Steps

Steps of Genomes Analysis

- Identify clusters of functionally related genes
- Evolutionary modeling

Genetics and Genomics

Genes and Size of Genomes

Genes and Size of Genomes

Size of Genomes

- Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb).
- Genomes of eukaryotes are usually larger

Genes and Size of Genomes

Size of Genomes

- Most plants and animals have genomes greater than 100 Mb.
- Humans have genome size of 3,000 Mb

Genes and Size of Genomes

Size of Genomes

- Within each domain there is no systematic relationship between genome size and phenotype

Genes and Size of Genomes

Size of Genomes

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
Bacteria			
Haemophilus influenzae	1.8	1,700	940
Escherichia coli	4.6	4,400	950
Archaea			
Archaeoglobus fulgidus	2.2	2,500	1,130
Methanosarcina barkeri	4.8	3,600	750

Genes and Size of Genomes

Size of Genomes

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
Eukaryotes			
Saccharomyces cerevisiae (yeast, a fungus)	12	6,300	525
Caenorhabditis elegans (nematode)	100	20,100	200
Arabidopsis thaliana (mustard family plant)	120	27,000	225
Drosophila melanogaster (fruit fly)	165	13,700	83
Oryza sativa (rice)	430	42,000	98
Zea mays (corn)	2,300	32,000	14
Mus musculus (house mouse)	2,600	22,000	11
Ailuropoda melanoleuca (giant panda)	2,400	21,000	9
Homo sapiens (human)	3,000	<21,000	7

Genes and Size of Genomes

Conclusion

- Although most eukaryotes have large size of genomes.
- Within each domain there is no systematic relationship between genome size and phenotype

Genetics and Genomics

Viral Genomes

Viral Genomes

Genomes of Viruses

- Viral genomes can be
- ssRNA
- dsRNA
- ssDNA
- dsDNA
- Linear
- Ciruclar

Viral Genomes

Viruses Genomes

- A viral genome is the genetic material of the virus.
- Also termed the viral chromosome.
- Viral genomes vary in size -few thousand to more than a hundred thousand nucleotides.

Viral Genomes

Viruses with RNA Genomes

- Almost all plants viruses and some bacterial and animal viruses
- Genomes are rather small (a few thousands nucleotides)

Viral Genomes

Viruses with DNA Genomes

- Often a circular genome
- lambda = 48,502 bp

Viral Genomes

Replicative form of Viral Genomes

- All ssRNA viruses produce dsRNA molecules
- Many linear DNA molecules become circular

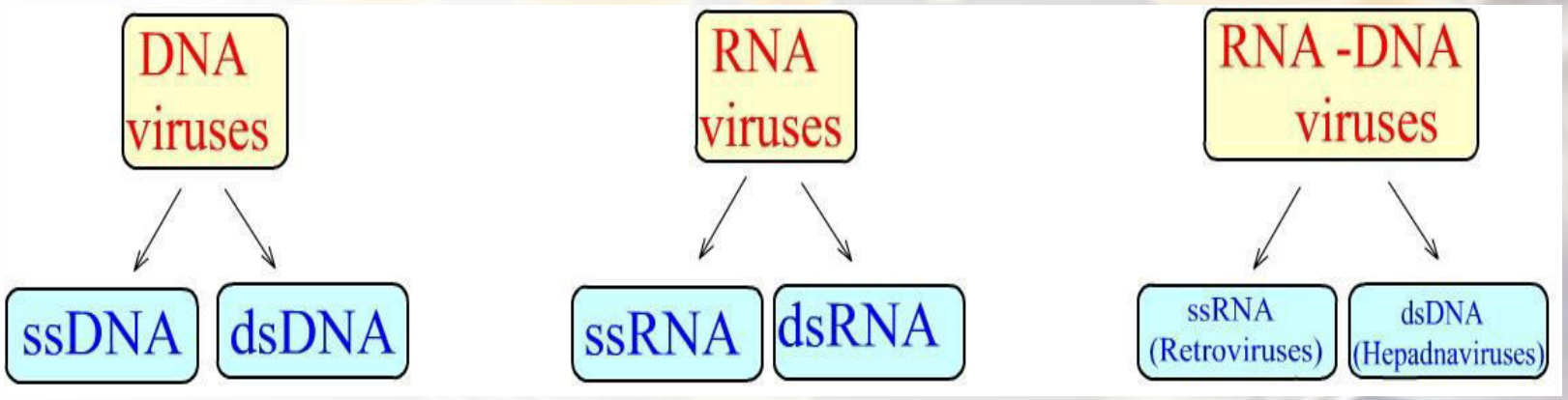
Viral Genomes

Viruses and Kingdoms

- Many plants viruses contain ssRNA genomes.
- Many fungal viruses contain dsRNA genomes.
- Many bacterial viruses contain dsDNA genomes.

Viral Genomes

Genomes in Virions: The genomes of viruses can be composed of either DNA or RNA, and some use both as their genomic material at different stages in their life cycle. However, only one type of nucleic acid is found in the virion of any particular type of virus.



Viral Genomes

Viruses and Number of Genes

Virus	Host	Type of Nucleic Acid	Number of Genes
Parvovirus	Mammals	ssDNA	5
Phage fd	*E. coli*	ssDNA	10
Lambda	*E. coli*	dsDNA	36
T4	*E. coli*	dsDNA	>190
Qβ	*E. coli*	ssRNA	4
TMV	Many plants	ssRNA	6
Influenza virus	Mammals	ssRNA	12

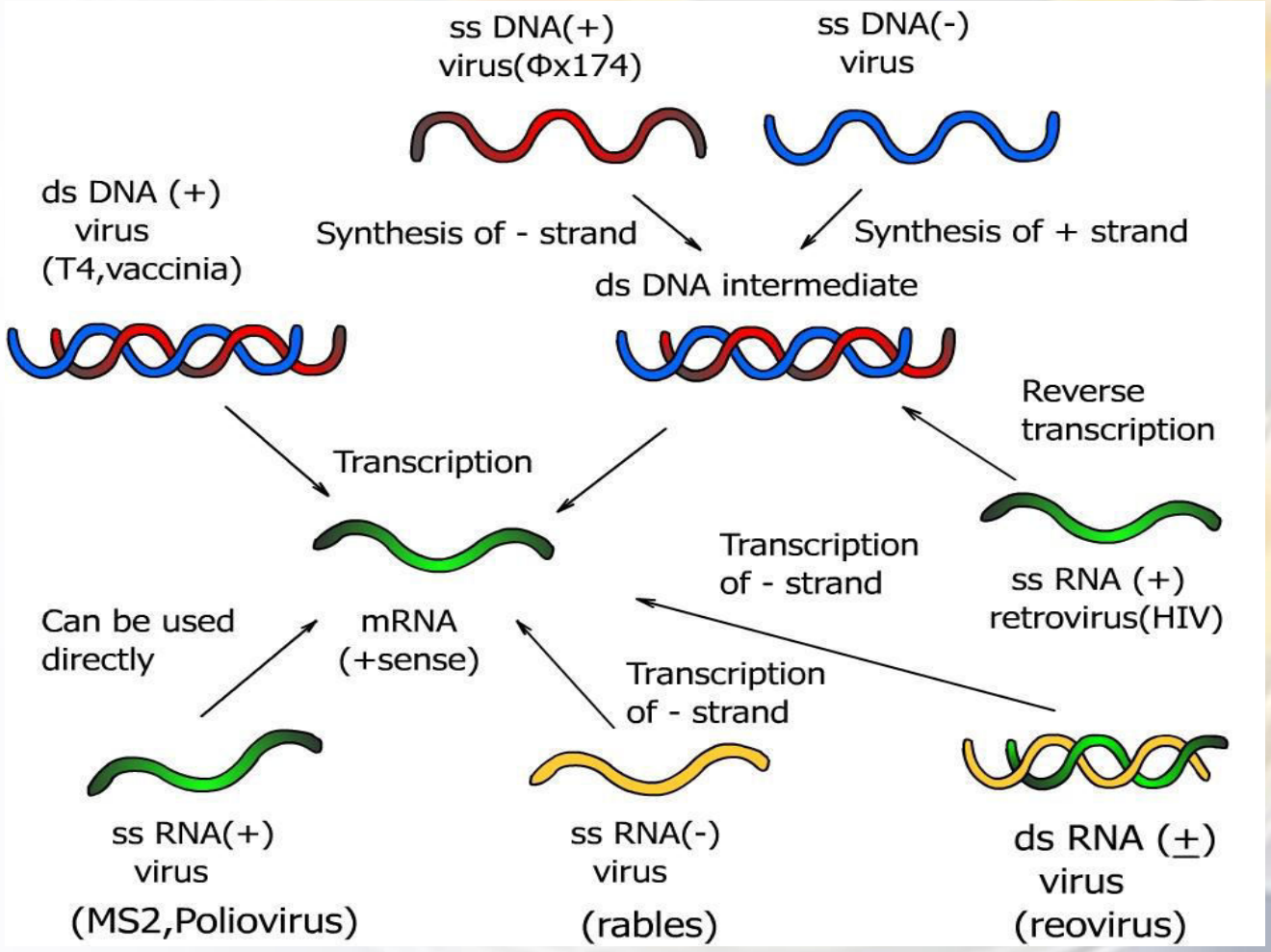
Viral Genomes

Viruses and Number of Genes

Virus	Genome structure	Genome size (kb)	Number of genes
Adenovirus	Double-stranded linear DNA	36.0	30
Hepatitis B	Partly double-stranded circular DNA	3.2	4
Influenza virus	Single-stranded segmented linear RNA	22.0	12
Parvovirus	Single-stranded linear DNA	1.6	5
Poliovirus	Single-stranded linear RNA	7.6	8
Reovirus	Double-stranded segmented linear RNA	22.5	22
Retroviruses	Single-stranded linear RNA	6.0–9.0	3
SV40	Double-stranded circular DNA	5.0	5
Tobacco mosaic virus	Single-stranded linear RNA	6.4	6
Vaccinia virus	Double-stranded circular DNA	240	240

Viral Genomes

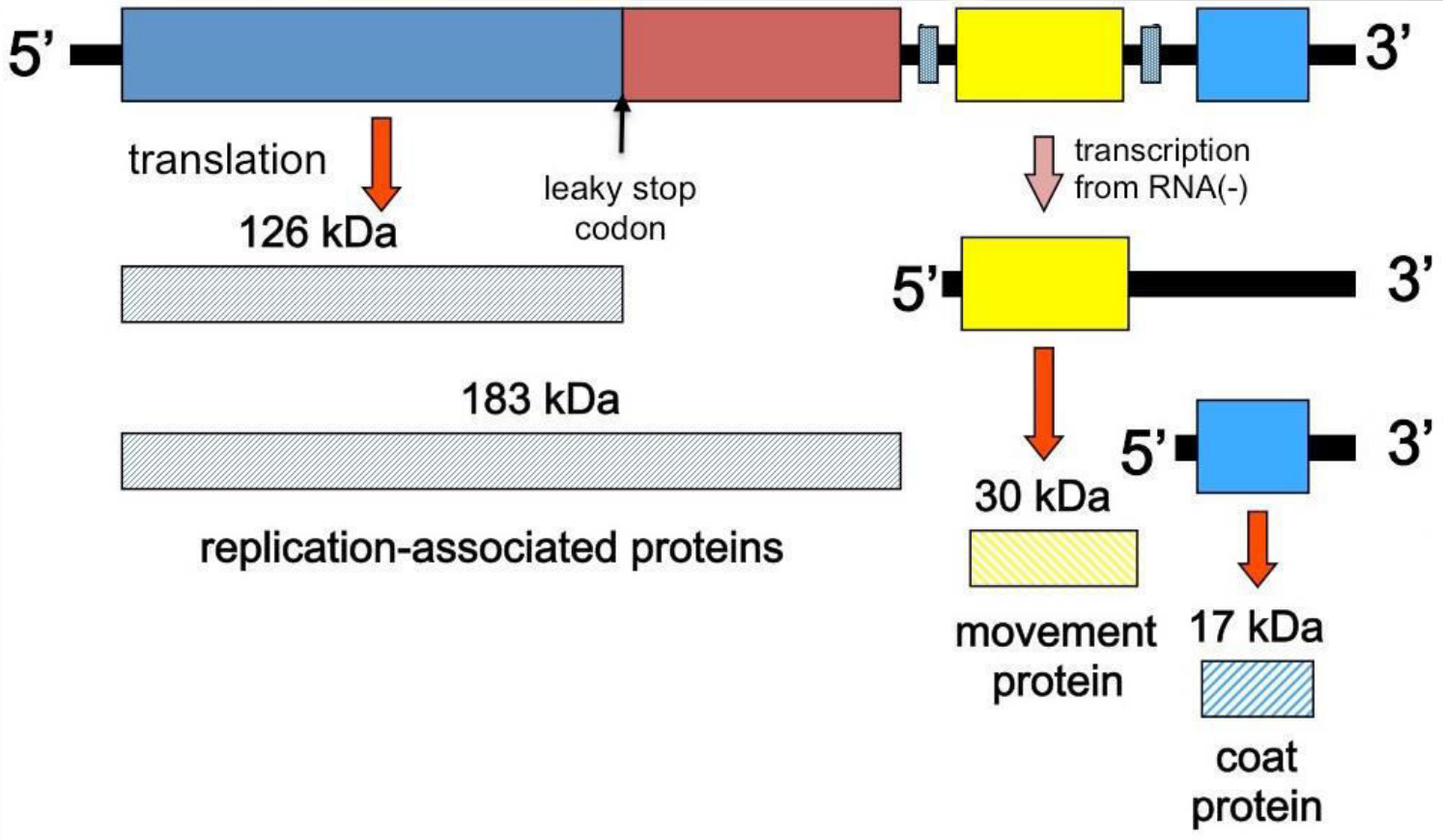
Genomes in Virions



Viral Genomes

Genome of Tobacco Mosaic Virus

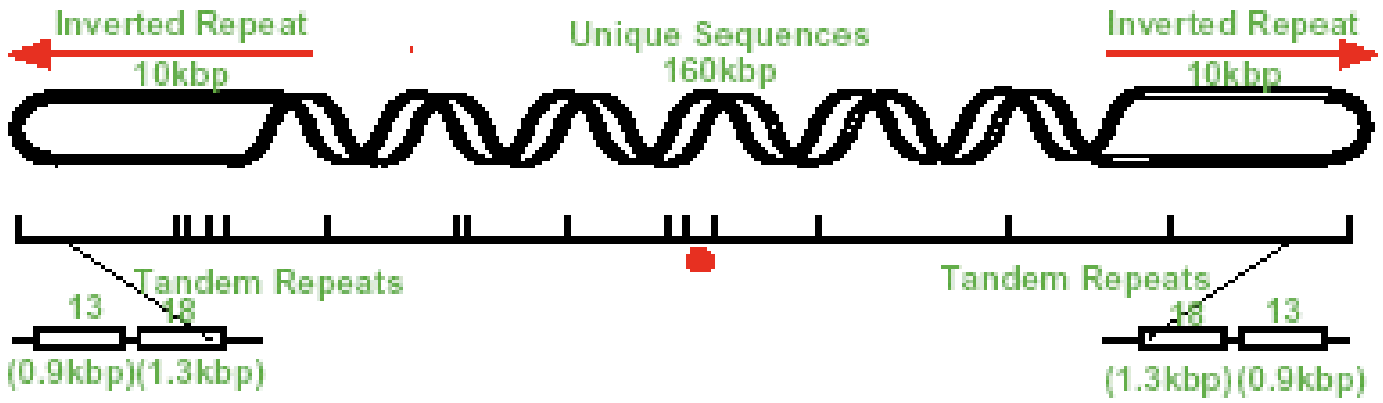
- Single, 6400 nucleotides RNA, 3 Essential Genes



Viral Genomes

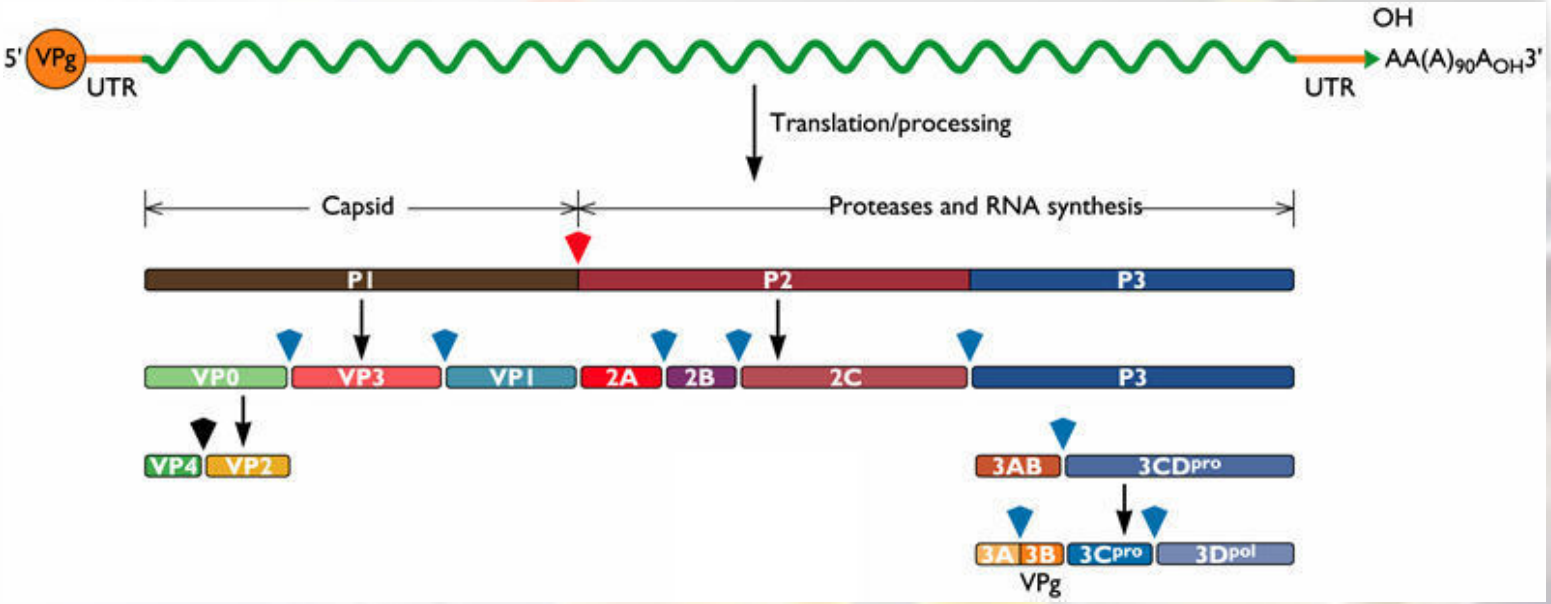
Genome of Poxvirus –A typical large dsDNA Virus

- 180 kb DNA, >100 Essential Genes



Viral Genomes

Genome of Polio Virus: Single-stranded positive-sense RNA genome that is about 7500 nucleotides long



Viral Genomes

Genome of Pox Virus

- Linear dsDNA 130-375 kbp; covalently closed termini.
- Large hairpin structure at each terminus - up to 10 kb total at each end is repeat sequence.
- Encode 150-300 proteins.
- Coding regions are closely spaced, no introns.
- Coding regions are on both strands of genome, and are not tightly clustered with respect to time of expression or function.

Viral Genomes

Viral Genomes

- Viral genomes can be
- ssRNA
- dsRNA
- ssDNA
- dsDNA
- Linear
- Ciruclar

Genetics and Genomics

Bacterial Genomes

Bacterial Genomes

Genomes of Bacteria

- Small organisms carry high coding density (85-90%)
- 1 gene per 1000 bases in prokaryotes
- Large variation in genome size between bacteria

Bacterial Genomes

Genomes of Bacteria – Large Variation

- *Tremblaya princeps* 140kb, 121 coding sequences
- *Sorangium cellulosum*
- 14000kb
- 11599 coding sequences

Bacterial Genomes

Genomes of Bacteria

Comparison of regulatory genes in bacterial genomes

Microorganism	# Genes in the Genome	# Regulatory Proteins	% of Total
Pseudomonas aeruginosa	5570	468	8.4
Escherichia coli	4289	250	5.8
Bacillus subtilis	4100	217	5.3
Mycobacterium tuberculosis	3918	117	3.0
Helicobacter pylori	1566	18	1.1

Bacterial Genomes

Distribution of genes among selected bacterial genomes and their sizes

Organism	Genome Size (Mbp)	No. of ORFs (% coding)		Unknown Function		Unique ORFs	
Aeropyrum pernix K1	1.67	1,885	(89%)				
A. aeolicus VF5	1.50	1,749	(93%)	663	(44%)	407	(27%)
A. fulgidus	2.18	2,437	(92%)	1,315	(54%)	641	(26%)
B. subtilis	4.20	4,779	(87%)	1,722	(42%)	1,053	(26%)
B. burgdorferi	1.44	1,738	(88%)	1,132	(65%)	682	(39%)
Chlamydia pneumoniae AR39	1.23	1,134	(90%)	543	(48%)	262	(23%)
Chlamydia trachomatis MoP_n	1.07	936	(91%)	353	(38%)	77	(8%)
C. trachomatis serovar D	1.04	928	(92%)	290	(32%)	255	(29%)
Deinococcus radiodurans	3.28	3,187	(91%)	1,715	(54%)	1,001	(31%)
E. coli K-12-MG1655	4.60	5,295	(88%)	1,632	(38%)	1,114	(26%)
H. influenzae	1.83	1,738	(88%)	595	(35%)	237	(14%)
H. pylori 26695	1.66	1,589	(91%)	744	(45%)	539	(33%)
Methanobacterium thermotautotrophicum	1.75	2,008	(90%)	1,010	(54%)	496	(27%)

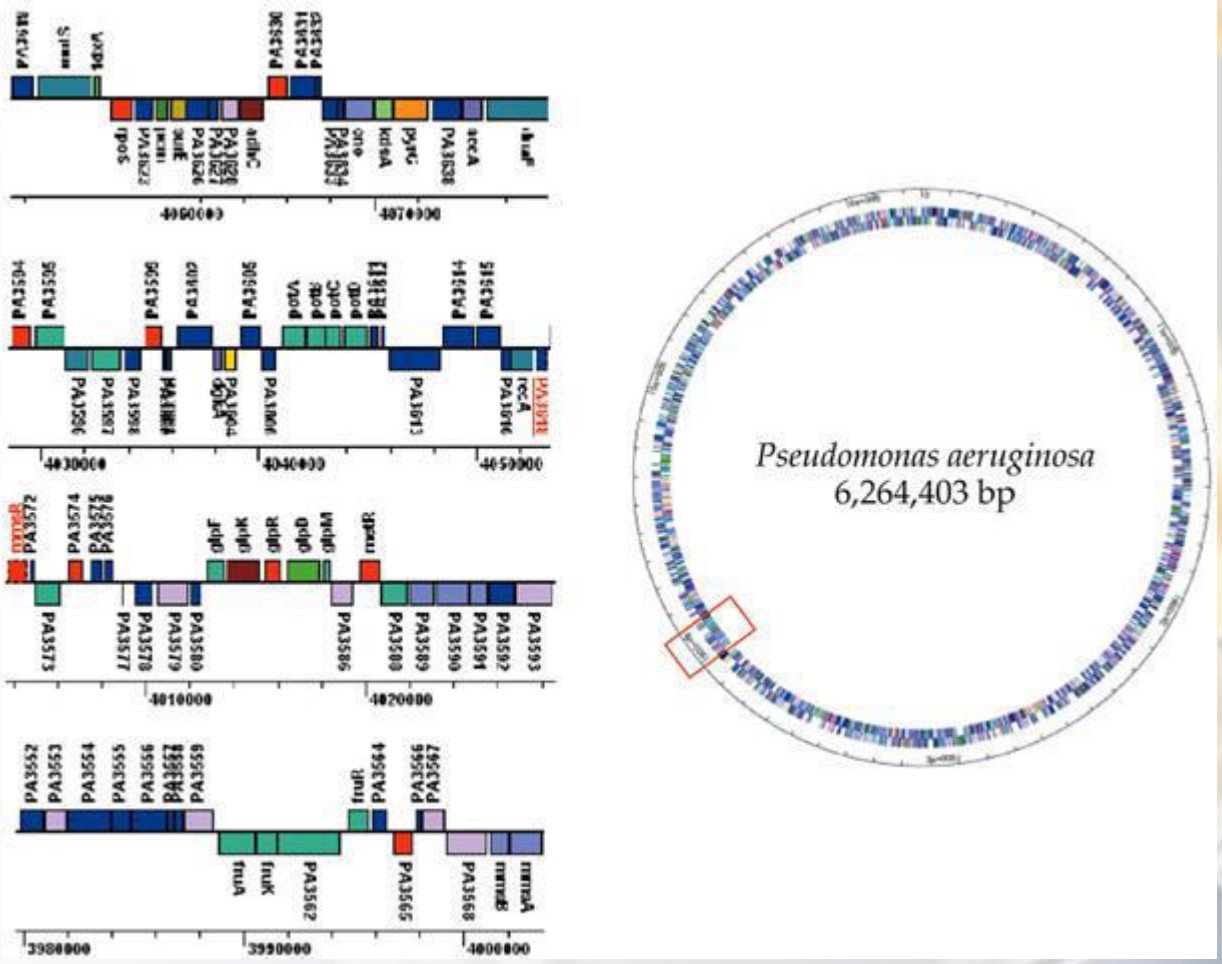
Bacterial Genomes

Distribution of genes among selected bacterial genomes and their sizes

Organism	Genome Size (Mbp)	No. of ORFs (% coding)		Unknown Function		Unique ORFs	
Methanococcus jannaschii	1.66	1,783	(87%)	1,076	(62%)	525	(30%)
M. tuberculosis CSU#93	4.41	4,275	(92%)	1,521	(39%)	606	(15%)
M. genitalium	0.58	483	(91%)	173	(37%)	7	(2%)
M. pneumoniae	0.81	680	(89%)	248	(37%)	67	(10%)
N. meningitidis MC58	2.24	2,155	(83%)	856	(40%)	517	(24%)
Pyrococcus horikoshii OT3	1.74	1,994	(91%)	589	(42%)	453	(22%)
Rickettsia prowazekii Madrid E	1.11	878	(75%)	311	(37%)	209	(25%)
Synechocystis sp.	3.57	4,003	(87%)	2,384	(75%)	1,426	(45%)
T. maritma MSB8	1.86	1,879	(95%)	863	(46%)	373	(26%)
T. pallidum	1.14	1,039	(93%)	461	(44%)	280	(27%)
Vibrio cholerae El Tor N1696	4.03	3,890	(88%)	1,806	(46%)	934	(24%)
	50.60	52,462	(89%)	22.358	(43%)	12,161	(23%)

Bacterial Genomes

Genes in a portion of a bacterial genome



Bacterial Genomes

Bacterial Genomes - Conclusion

- Small organisms carry high coding density.
- Large variation in genome size between bacteria.

Genetics and Genomics

Yeast Genome

Yeast Genome

Yeast Genome

- The nuclear genome consists of 16 chromosomes.
- In addition, there is a mitochondrial genome and a plasmid, 2 micron circle.

Yeast Genome

Yeast Genome

- The haploid yeast genome consists of ~ 12.1 Mb
- Yeast genome was completely sequenced by 1996

Yeast Genome

Yeast Genome - Characteristics

- Small and compact
- Small intergenic sequences
- Few transposable elements
- Few introns
- Limited RNA interference

Yeast Genome

Yeast Genome

- The yeast genome is predicted to contain about 6,200 genes
- 274 tRNA
- 287 introns
- Small percentage of yeast genes have introns
- The intergenic space between genes is only between 200bp - 1,000bp

Yeast Genome

Yeast Genome: Genome of Yeast Cell

Characteristic	Chromosomes	Plasmid	Mitochondria
Relative amount (%)	85	5	10
Number of copies	2 x 16	60-100	~50 (8-130)
Size (kbp)	~ 12,100	6.318	70-76

Yeast Genome

Yeast Genome

- The largest known regulatory sequences are spread over about 2,800bp
- MUC1/FLO11

Yeast Genome

Yeast Genome

- Yeast genes have names consisting of three letters and up to three numbers
- GPD1, HSP12, PDC6
- Usually they are meaningful

Yeast Genome

Yeast Genome – Genes Nomenclature

- Wild type genes are written with capital letters in italics: *TPS1*, *RHO1*, *CDC28*
- Recessive mutant genes are written with small letters in italics: *tps1, rho1, cdc28*

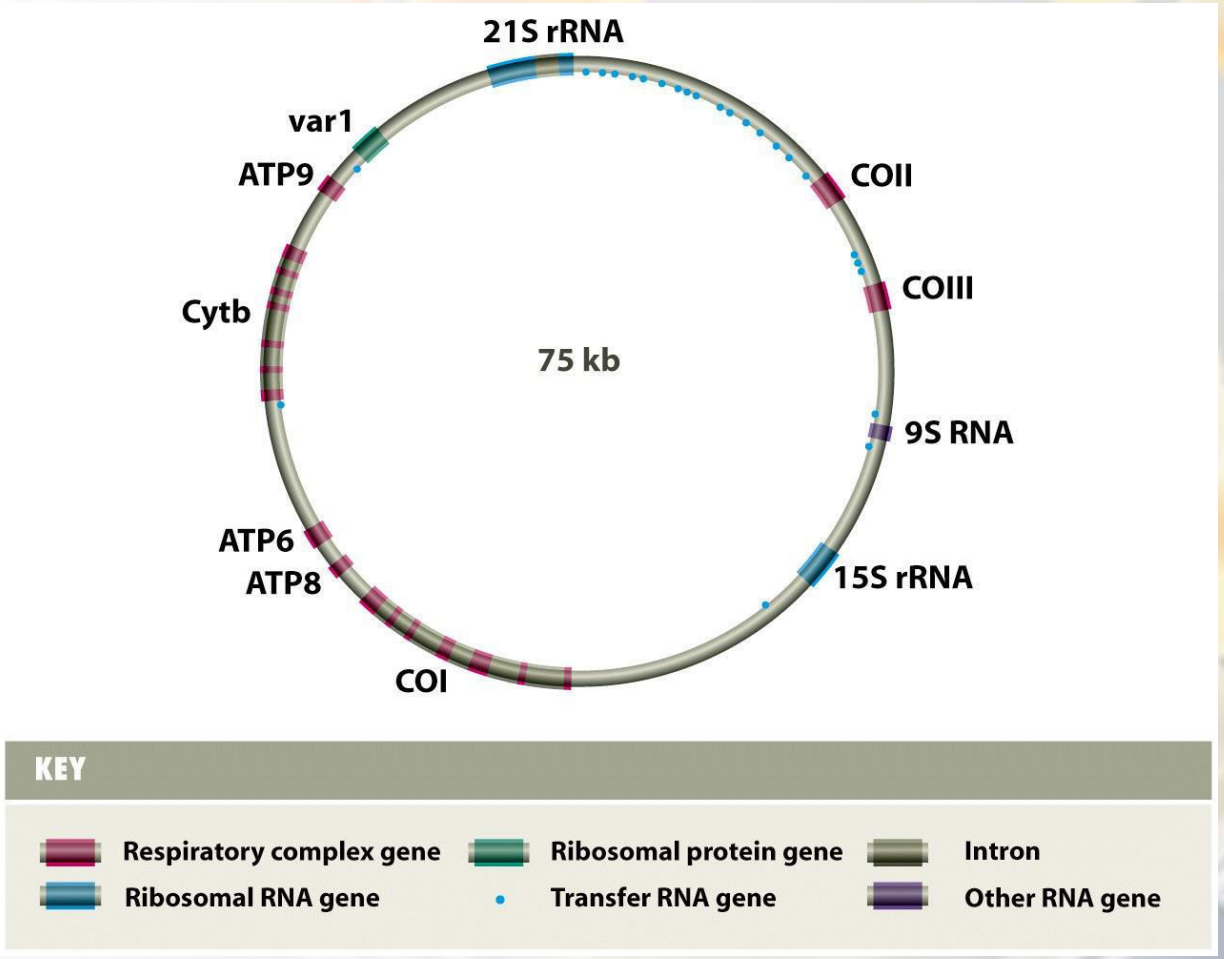
Yeast Genome

Yeast Genome – Genes Nomenclature

- Three letters provides information about a function, mutant phenotype, or process related to that gene.
- CDC - Cell Division Cycle ; ADE-ADEnine biosynthesis

Yeast Genome

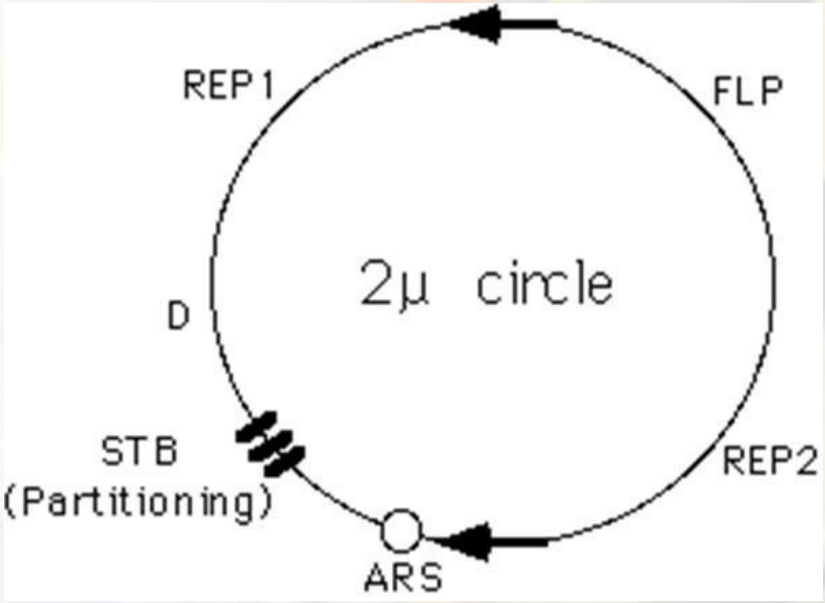
Yeast – Mitochondrial DNA



Yeast Genome

Yeast – Plasmid DNA

- The 2u circle is a 6.3 kb
- 50 to 100 copies per haploid genome of the yeast cells
- ARS, the FLP gene, the three genes which encode proteins required for regulation of FLP expression (REP2, REP1, and D)
- Set of small direct repeats (called "STB") required for partitioning into daughter cells during mitosis and meiosis.



Yeast Genome

Yeast Genome

- Yeast nuclear genome has 16 chromosomes.
- A mitochondrial genome.
- A plasmid.

Mitochondrial Genome

Mitochondrial Genome

Mitochondrial Genome

- Multiple identical circular chromosomes
- ~15-16 Kb in animals
- ~ 200 kb to 2,500 kb in plants

Mitochondrial Genome

Mitochondrial Genome

- Over 95% of mitochondrial proteins are encoded in the nuclear genome.
- Often A+T rich genomes.

Mitochondrial Genome

Human Mitochondrial Genome

- Circular, double stranded, 16.6 kb
- The two strands are notably different in base composition, leading to one strand being heavy (H strand) and the other light (L strand).

Mitochondrial Genome

Human Mitochondrial Genome

- Both strands encode genes, although more are on the H strand.
- A short region (1121 bp), the D loop is a DNA triple helix: two overlapping copies of the H strand.

Mitochondrial Genome

Human Mitochondrial Genome

- The D loop is also the site where most of replication and transcription is controlled.
- Genes are tightly packed, with almost no non-coding DNA outside of D loop.

Mitochondrial Genome

Human Mitochondrial Genome

- Human mitochondrial genes contain no introns, although introns are found in the mitochondria of other groups (plants)

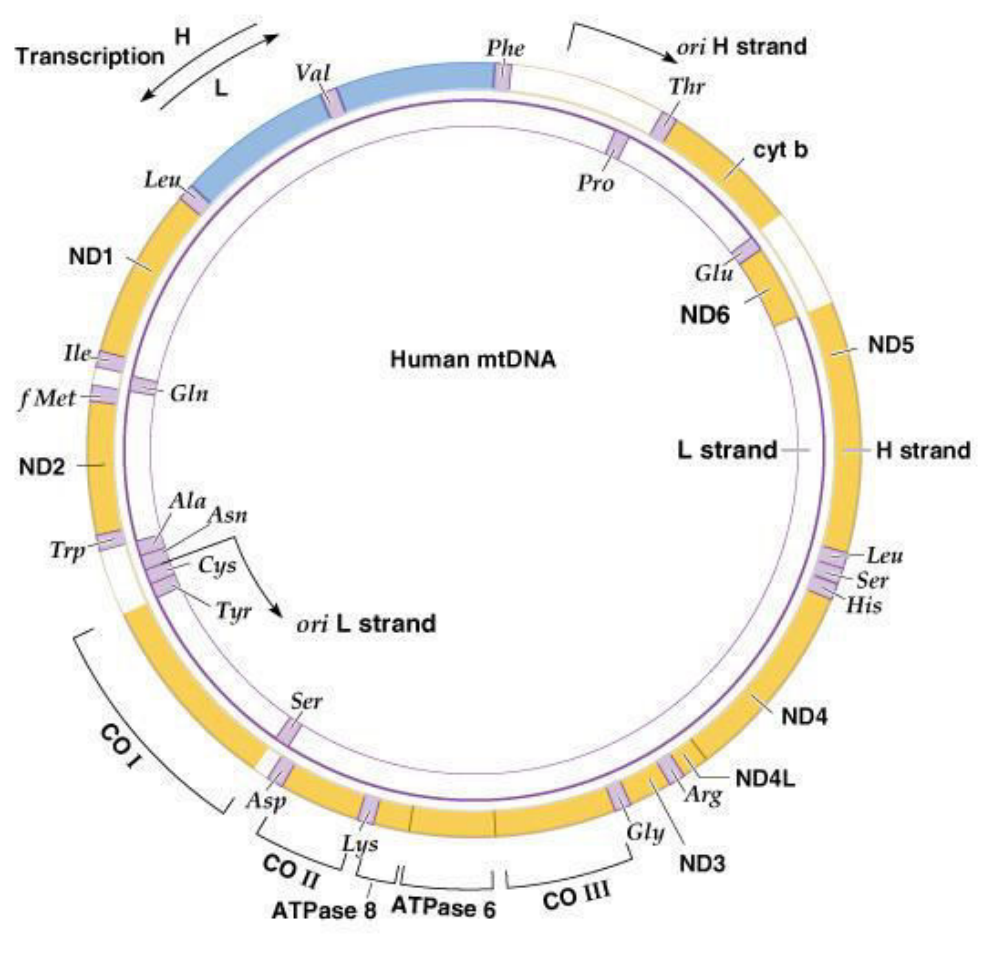
Mitochondrial Genome

Human Mitochondrial Genome

- 37 Genes
- 22 tRNAs
- 2 rRNAs
- 13 polypeptides
- tRNA: only 60 of the 64 codons code for amino acids.

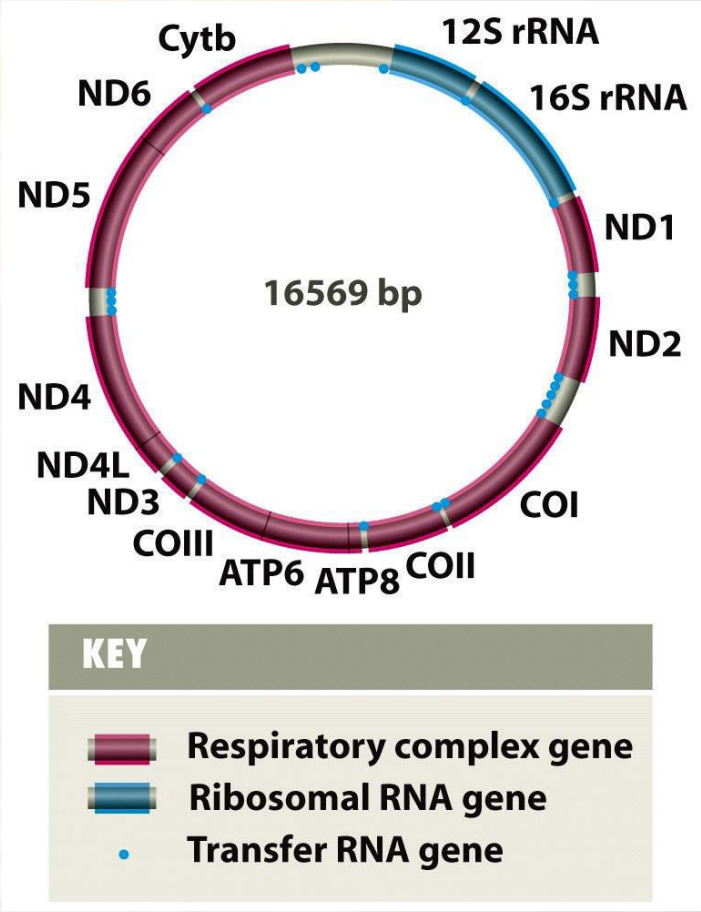
Mitochondrial Genome

Mitochondrial Genome - Humans



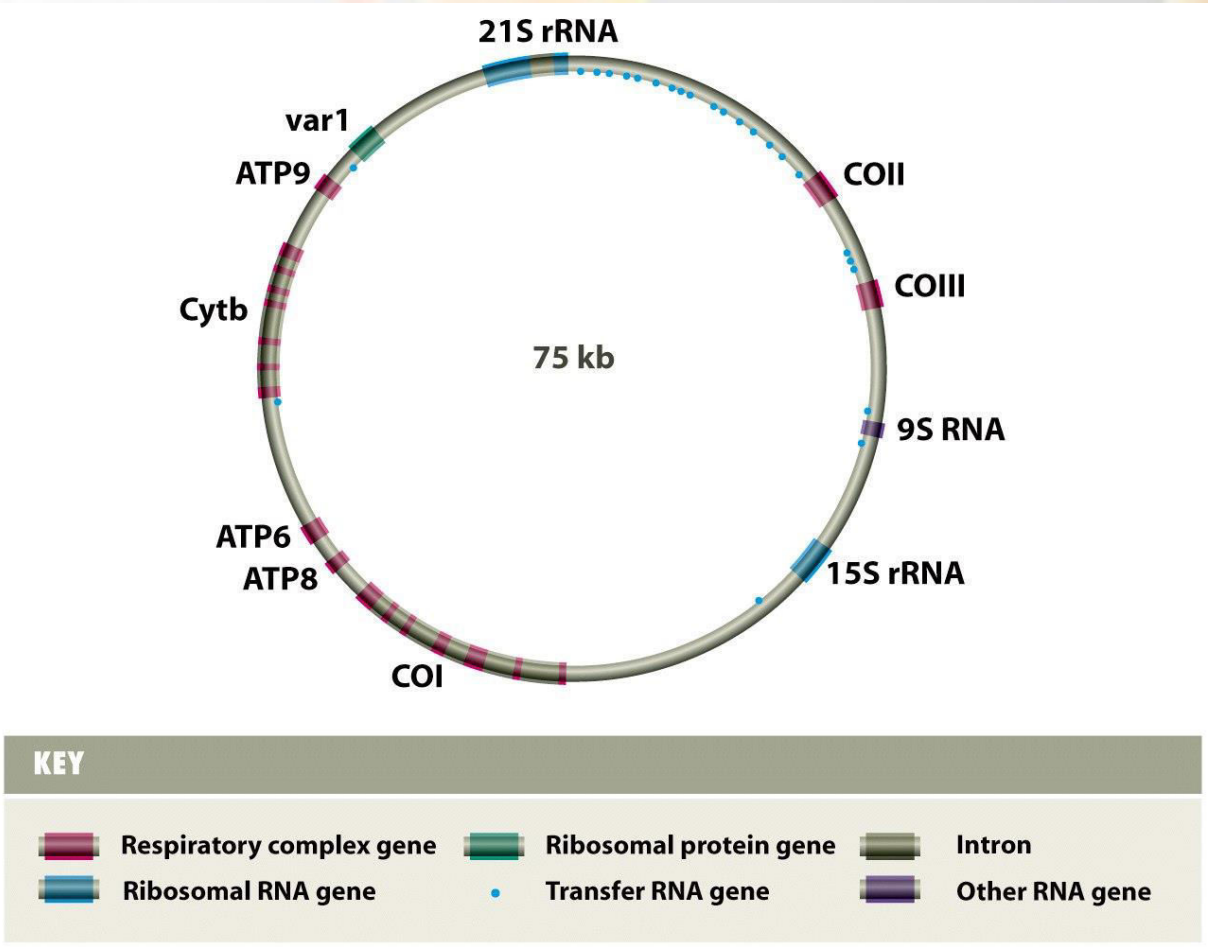
Mitochondrial Genome

Mitochondrial Genome - Humans



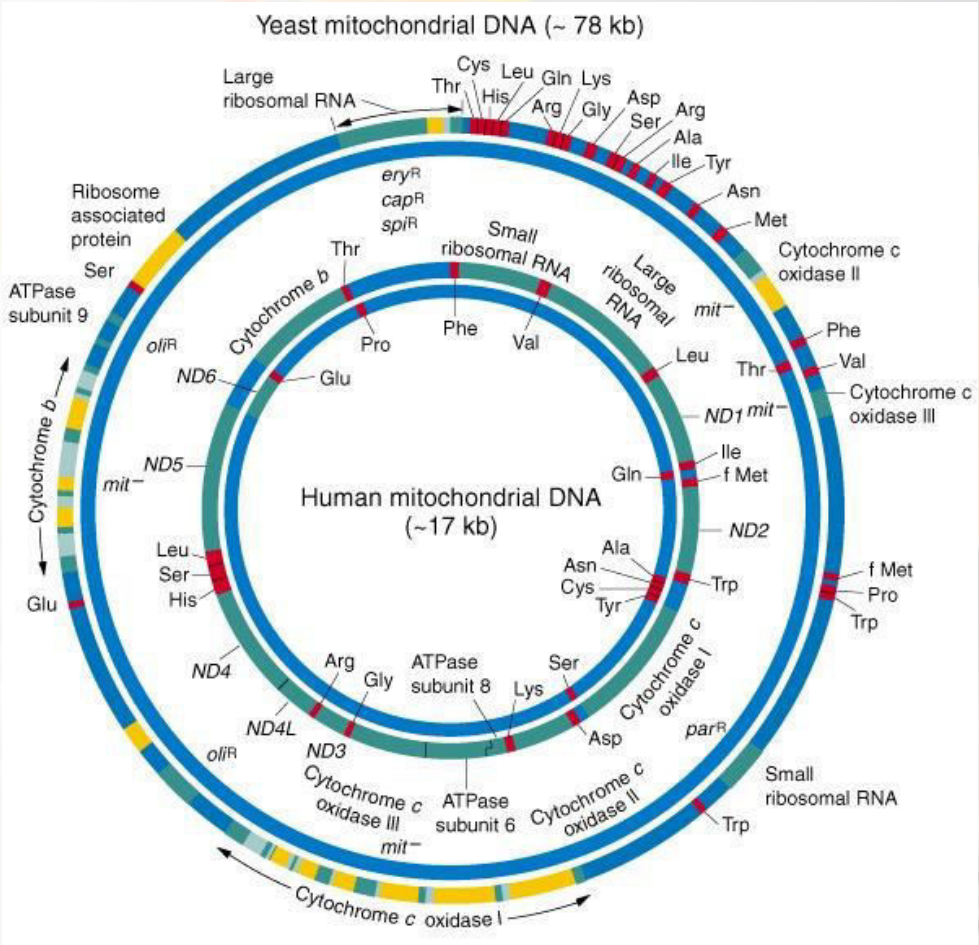
Mitochondrial Genome

Mitochondrial Genome – Yeast



Mitochondrial Genome

Mitochondrial Genome – Human and Yeast



Mitochondrial Genome

Mitochondrial Genomes

Species	Type of organism	Genome size (kb)
Mitochondrial genomes		
Plasmodium falciparum	Protozoan (malaria parasite)	6
Chlamydomonas reinhardtii	Green alga	16
Mus musculus	Vertebrate (mouse)	16
Homo sapiens	Vertebrate (human)	17
Metridium senile	Invertebrate (sea anemone)	17
Drosophila melanogaster	Invertebrate (fruit fly)	19
Chondrus crispus	Red alga	26
Aspergillus nidulans	Ascomycete fungus	33
Reclinomonas americana	Protozoa	69
Saccharomyces cerevisiae	Yeast	75
Suillus grisellus	Basidiomycete fungus	121
Brassica oleracea	Flowering plant (cabbage)	160
Arabidopsis thaliana	Flowering plant (vetch)	367
Zea mays	Flowering plant (maize)	570
Cucumis melo	Flowering plant (melon)	2500

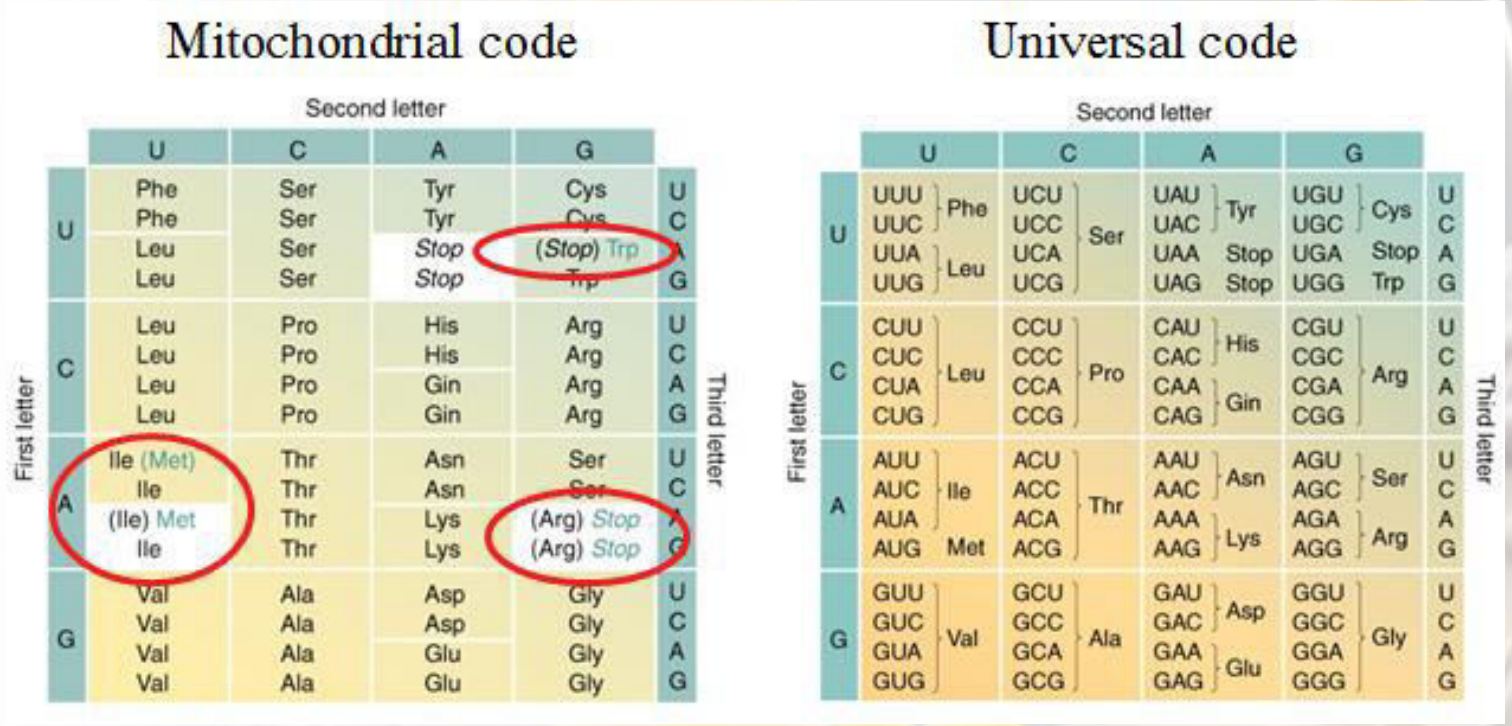
Mitochondrial Genome

Mitochondrial Genomes

Feature	*Plasmodium falciparum*	*Chlamydomonas reinhardtii*	*Homo sapiens*	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Reclinomonas americana*
Total number of genes	5	12	37	35	52	92
Types of genes						
Protein-coding genes	3	7	13	8	27	62
Respiratory complex	3	7	13	7	17	24
Ribosomal proteins	0	0	0	1	7	27
Transport proteins	0	0	0	0	3	6
RNA polymerase	0	0	0	0	0	4
Translation factor	0	0	0	0	0	1
Functional RNA genes	2	5	24	27	25	30
Ribosomal RNA genes	2	2	2	2	3	3
Transfer RNA genes	0	3	22	24	22	26
Other RNA genes	0	0	0	1	0	1
Number of introns	0	1	0	8	23	1
Genome size (kb)	6	16	17	75	367	69

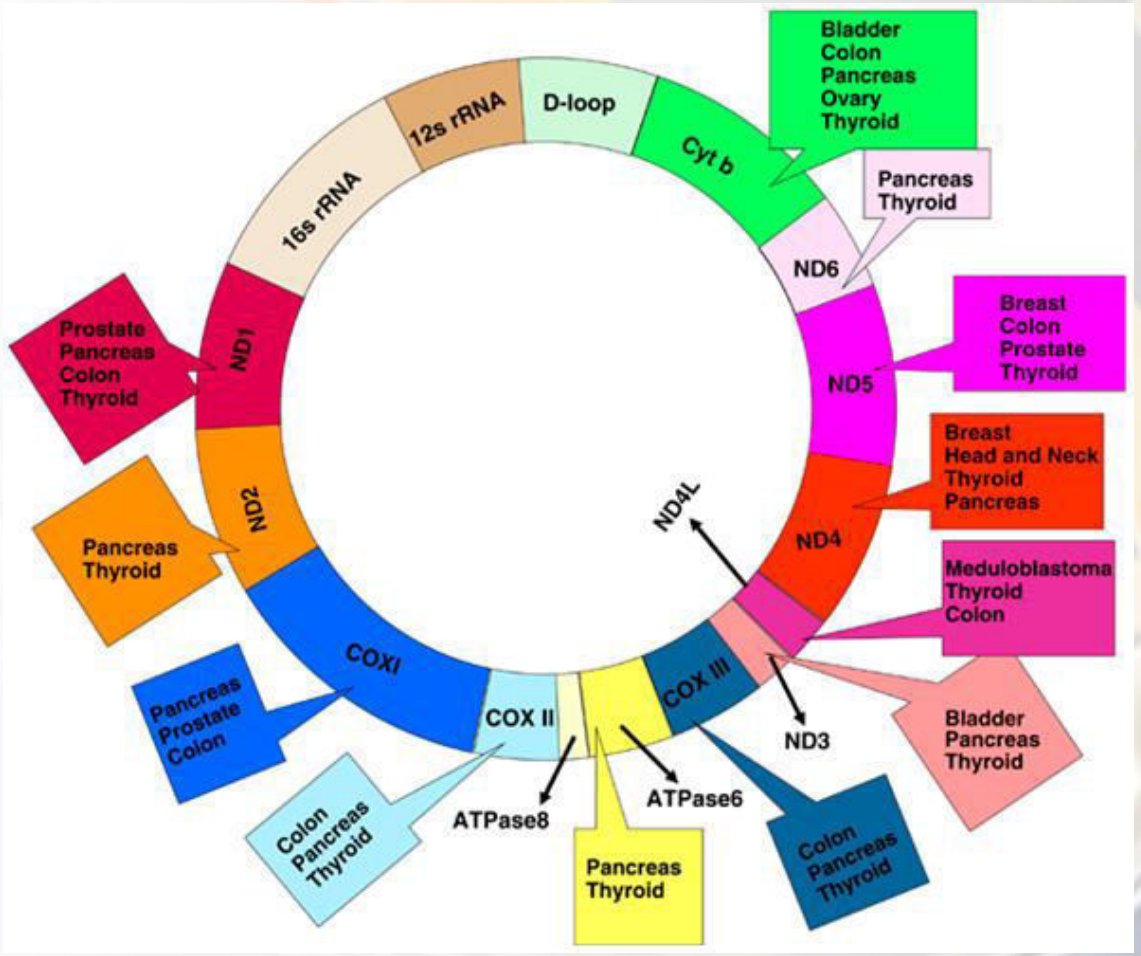
Mitochondrial Genome

Universal Code and Mitochondrial Code



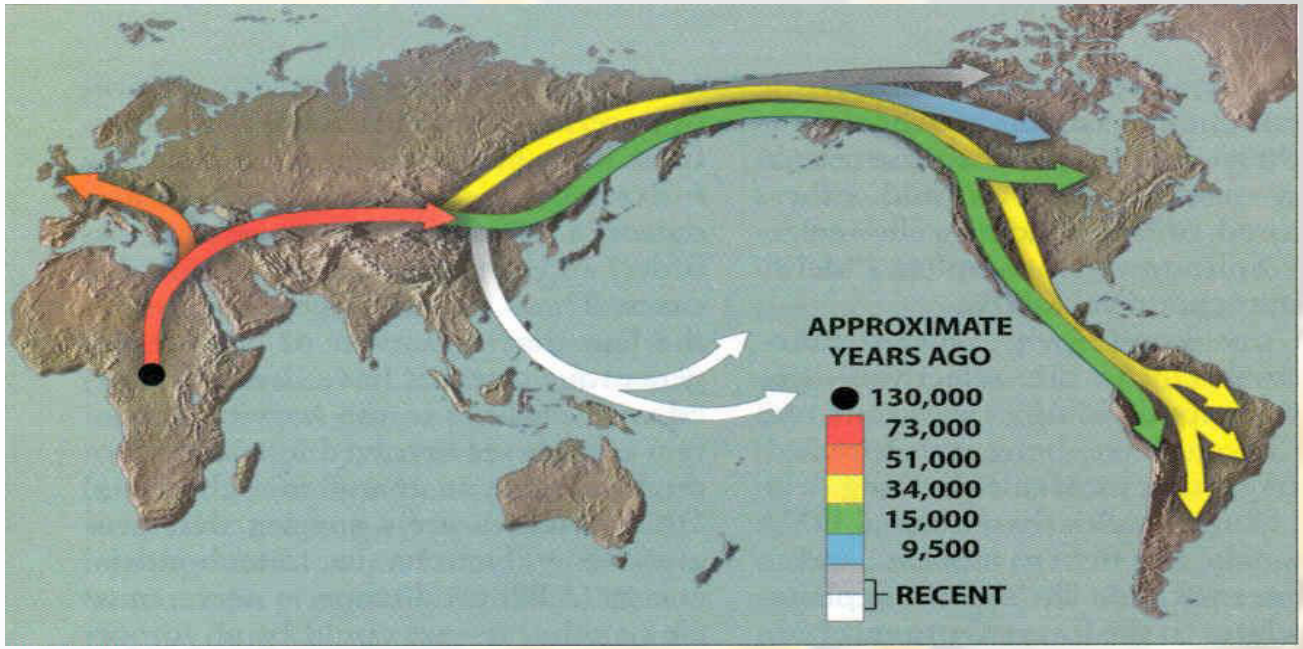
Mitochondrial Genome

Mitochondrial Genome – Tumors due to mutations



Mitochondrial Genome

Mitochondrial DNA polymorphisms track human migrations



All humans descend from a small group of Africans
This group originated in central Africa ~200,000 years ago
The founding group was small (10^2-10^4 people)

Mitochondrial Genome

Conclusion

- Multiple identical circular chromosomes
- 15-16 Kb in animals
- 200 kb to 2,500 kb in plants

Genetics and Genomics

Genomes Comparisons

Genomes Comparisons

Genomes vary in size

- Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb)
- Most plants and animals have genomes greater than 100 Mb; humans have 3,000 Mb

Genomes Comparisons

Genomes vary in genes numbers

- Free-living bacteria and archaea have 1,500 to 7,500 genes
- Fungi have about 5,000 genes and multicellular eukaryotes upto 40,000 genes

Genomes Comparisons

Genomes vary in genes numbers

- Number of genes is not correlated to genome size
- Nematode *C. elegans* has 100 Mb and 20,000 genes, while *Drosophila* has 165 Mb and 13,700 genes

Genomes Comparisons

Genomes vary in genes numbers

- Vertebrate genomes can produce more than one polypeptide per gene because of alternative splicing of RNA transcripts

Genomes Comparisons

Humans and Mammals have low gene density

- Humans and other mammals have the lowest gene density, or number of genes, in a given length of DNA. Multicellular eukaryotes have many introns within genes and noncoding DNA between genes

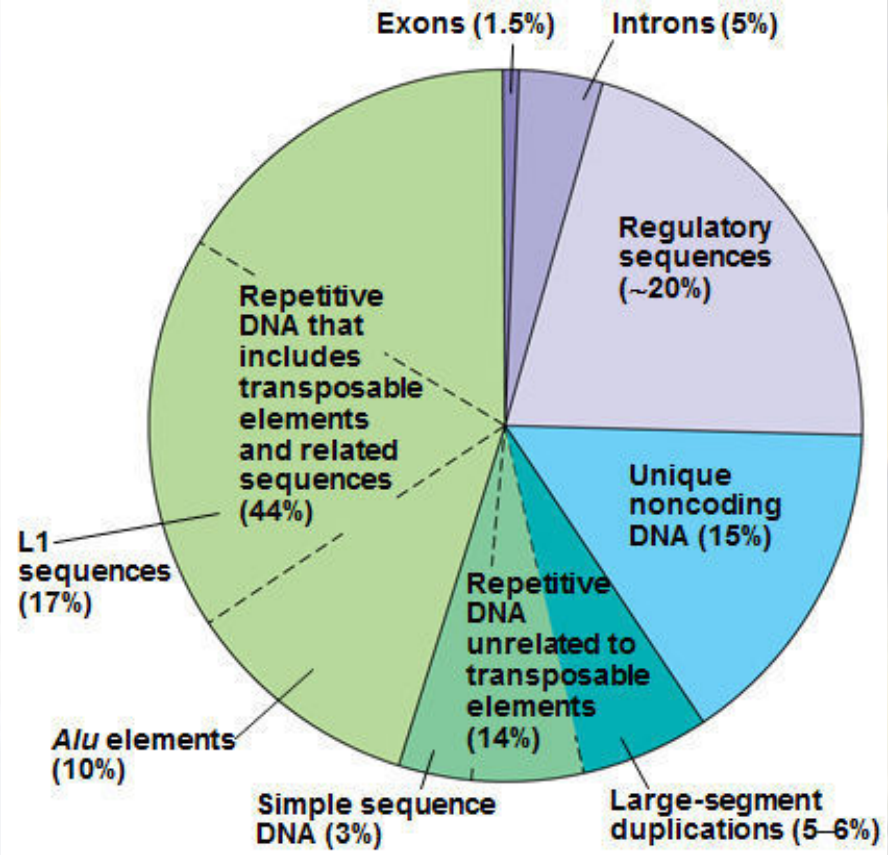
Genomes Comparisons

Multicellular eukaryotes have much noncoding DNA and multigene families

- Most of eukaryotic genomes neither encodes proteins nor functional RNAs
- Evidence indicates that noncoding DNA plays important roles in the cell

Genomes Comparisons

Human Genome: Distribution of coding and non-coding DNA



Genomes Comparisons

Comparing Genomes

- Significant similarity between genomes of ”distant” species (Man – Yeast 23%)
- Similarity increases for taxonomically close species.

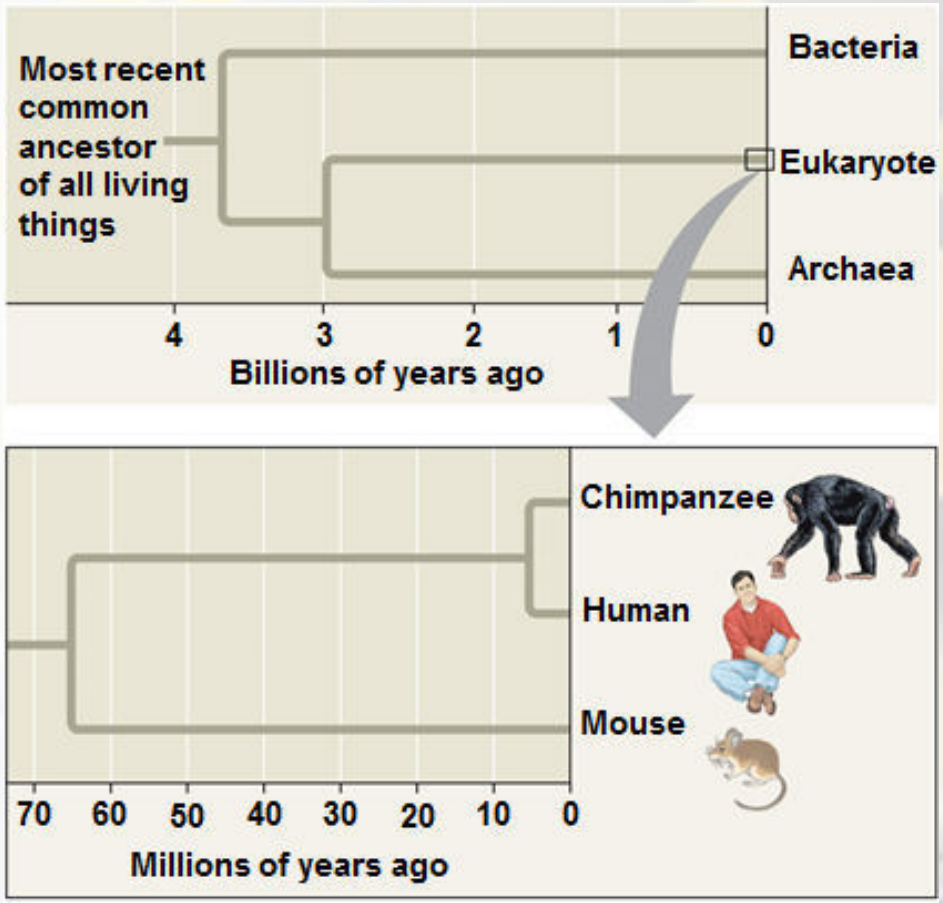
Genomes Comparisons

Comparing Genomes

- Closely related species help us understand recent evolutionary events
- Distantly related species help us understand ancient evolutionary events

Genomes Comparisons

Comparing Genomes: **Bacteria, archaea, and eukaryotes diverged from each other between 2 and 4 billion years ago**



Genomes Comparisons

Comparing Genomes

	Bacteria	Archaea	Eukaryotes
Genome size	Most are 1–6 Mb		Most are 10–4,000 Mb, but a few are much larger
Number of genes	1,500–7,500		5,000–40,000
Gene density	Higher than in eukaryotes		Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)
Introns	None in protein-coding genes	Present in some genes	Unicellular eukaryotes: present, but prevalent only in some species Multicellular eukaryotes: present in most genes
Other noncoding DNA	Very little		Can be large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes

Genomes Comparisons

Comparing Genomes

- Human and chimpanzee genomes differ by 1.2%, at single base-pairs, and by 2.7% because of insertions and deletions

Genetics and Genomics

Comparing distantly/closely related species

Distantly/Closely Related Species

Comparing distantly related species

- Highly conserved genes have changed very little over time
- These help to clarify relationships among species that diverged from each other long ago

Distantly/Closely Related Species

Comparing distantly related species

- Bacteria, archaea, and eukaryotes diverged from each other 2 and 4 billion years ago
- Highly conserved genes can be studied in one model organism.

Distantly/Closely Related Species

Comparing closely related species

- Genetic differences between closely related species can be correlated with phenotypic differences

Distantly/Closely Related Species

Comparing closely related species

- Genetic comparison of several mammals with non-mammals helps to identify what make mammals

Distantly/Closely Related Species

Comparing closely related species

- Human and chimpanzee genomes differ by 1.2%, at single base-pairs, and by 2.7% because of insertions and deletions

Distantly/Closely Related Species

Comparing closely related species

- Several genes are evolving faster in humans than chimpanzees

Distantly/Closely Related Species

Comparing closely related species

- Genes involved in defense against malaria and tuberculosis and in regulation of brain size, genes code for transcription factors

Distantly/Closely Related Species

Comparing closely related species

- Humans and chimpanzees differ in the expression of the *FOXP2* gene, whose product turns on genes involved in vocalization

Distantly/Closely Related Species

Comparing closely related species

- Differences in the *FOXP2* gene may explain why humans but not chimpanzees communicate by speech

Distantly/Closely Related Species

Conclusion

- Highly conserved genes have changed very little over the time
- These help to clarify relationships among species that diverged from each other long ago

Genetics and Genomics

Genome Mapping

Genome Mapping

Genome mapping

- Different types
- Genetic mapping
- Physical mapping

Genome Mapping

Genetic Mapping of Genomes

- Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome.

Genome Mapping

Genetic Mapping

- Genetic techniques include cross-breeding experiments
- Case of humans, the examination of family histories (pedigrees).

Genome Mapping

Physical Mapping of Genomes

- Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes.

Genome Mapping

Mapping of Genomes

Mapping Strategy	Requires	Resolution	How to increase resolution
Genetic	Polymorphic Markers, and Pedigrees	Medium to High	Increase number of markers or people
Restriction	Restriction Enzymes	High	Increase number of enzymes
Somatic Cell Hybrid	Somatic Hybrid Panel, and STSs	Low to Medium	Increase number of deletion hybrids
Radiation Hybrid	Radiation Hybrid Panel, and STSs	High	Use additional hybrids, or make a new panel

Genome Mapping

Genome Mapping

- Different types
- Genetic mapping
- Physical mapping

Genetic Mapping of Genomes

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- Linkage analysis is the basis of genetic mapping
- Recombination fraction is a measure of the distance between two loci.

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- Two loci that show 1% recombination are defined as being 1 centimorgan (cM) apart on a genetic map.
- 1 map unit = 1 cM centimorgan

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- Genetic mapping involves determining, the location of genes on a chromosome relative to other genes, using genetic crosses and pedigree analysis.

Genetic Mapping of Genomes

Genetic Mapping

- Genetic map of the human genome has 24 different maps; one for each autosome, X and Y.
- Marker alleles in genetic crosses help to determine crossover rate between linked genes

Genetic Mapping of Genomes

Genetic Mapping of Genomes - Procedure

- Individuals with different alleles at two or more loci are crossed, and their offspring examined.
- Most of the offspring will have phenotypes corresponding to the linked alleles. A few progeny will be recombinant.
- The frequency of the recombinant phenotype is calculated as a percentage of the total offspring, giving the recombination frequency or genetic distance.

Genetic Mapping of Genomes

Genetic Mapping in Humans

- Experimental crosses are not done in humans and so genetic mapping relies on pedigree analysis, and is limited by rarity of large, multigenerational pedigrees showing segregation of defined linked traits.
- Usually, the lod (logarithm of odds) score method is used for statistical analysis of pedigree data.

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- A lod score compares the expected distributions of traits if they are linked or not linked.
- The lod score is the log10 of the ratio of the two probabilities. The higher the lod score, the closer the two genes.
- The map distance for linked markers is computed from the recombination frequency

Genetic Mapping of Genomes

Genetic Mapping

- Mapped features that are not genes are called as DNA markers.
- A DNA marker must have at least two alleles to be useful.
- Mostly, three types RFLP, SSLP, SNP

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- To test linkage between the genes for two traits, certain types of matings are examined whether or not the pattern of the combinations of traits exhibited by the offspring follows the law of independent assortment.
- If not, the gene pairs for those traits must be linked, that is they must be on the same chromosome pair.

Genetic Mapping of Genomes

Recombinations Frequencies (RF)

- Two genes that undergo independent assortment have RF of 50 % and are located on non-homologous chromosomes
- They are unlinked

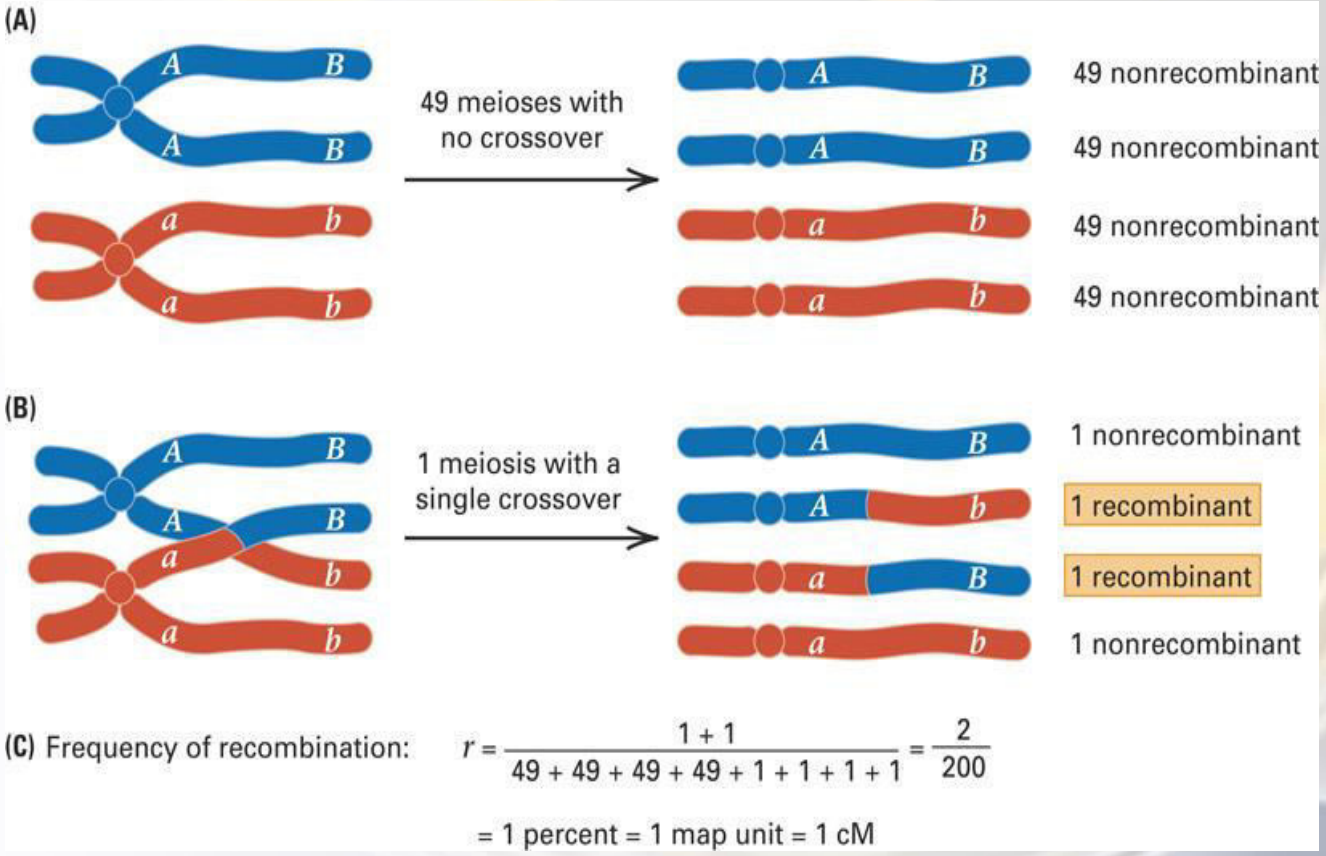
Genetic Mapping of Genomes

Recombinations Frequencies

- Genes with recombination frequencies less than 50 percent are on the same chromosome
- They are linked

Genetic Mapping of Genomes

Genetic Mapping of Genomes – Recombinations Frequency Calculated as:



Genetic Mapping of Genomes

Genetic mapping of Genomes – Recombinations Frequency

$$\text{recombinant frequency} = \frac{\text{number of recombinant progeny}}{\text{total number of progeny}} \times 100\%$$

Genetic Mapping of Genomes

Genetic Mapping of Genomes – Recombinations frequency which is calculated as follows:

- The LOD score is calculated as follows:
- LOD = Z = $\log_{10} \frac{\text{probability of birth sequence with a given linkage}}{\text{probability of birth sequence with no linkage}}$

- A LOD score greater than 3.0 is considered evidence for linkage.
- A LOD score less than -2.0 is considered evidence to exclude linkage.

Genetic Mapping of Genomes

Genetic Mapping of Genomes

- Linkage analysis is the basis of genetic mapping
- Recombination fraction is a measure of the distance between two loci.

Physical Mapping of Genomes

Physical Mapping of Genomes

Physical Map of Genomes

- A physical map of a chromosome or genome that shows the physical locations of genes and other DNA sequences

Physical Mapping of Genomes

Genetic Mapping provides in-sufficient information about exact locations of genes

- Genetic Information by genetic map rarely sufficient for directing the sequencing phase of a genome project.

Physical Mapping of Genomes

Genetic Mapping – limited accuracy

- Two reasons
- The resolution of a genetic map depends on the number of crossovers that have been scored.
- Genetic maps have limited accuracy

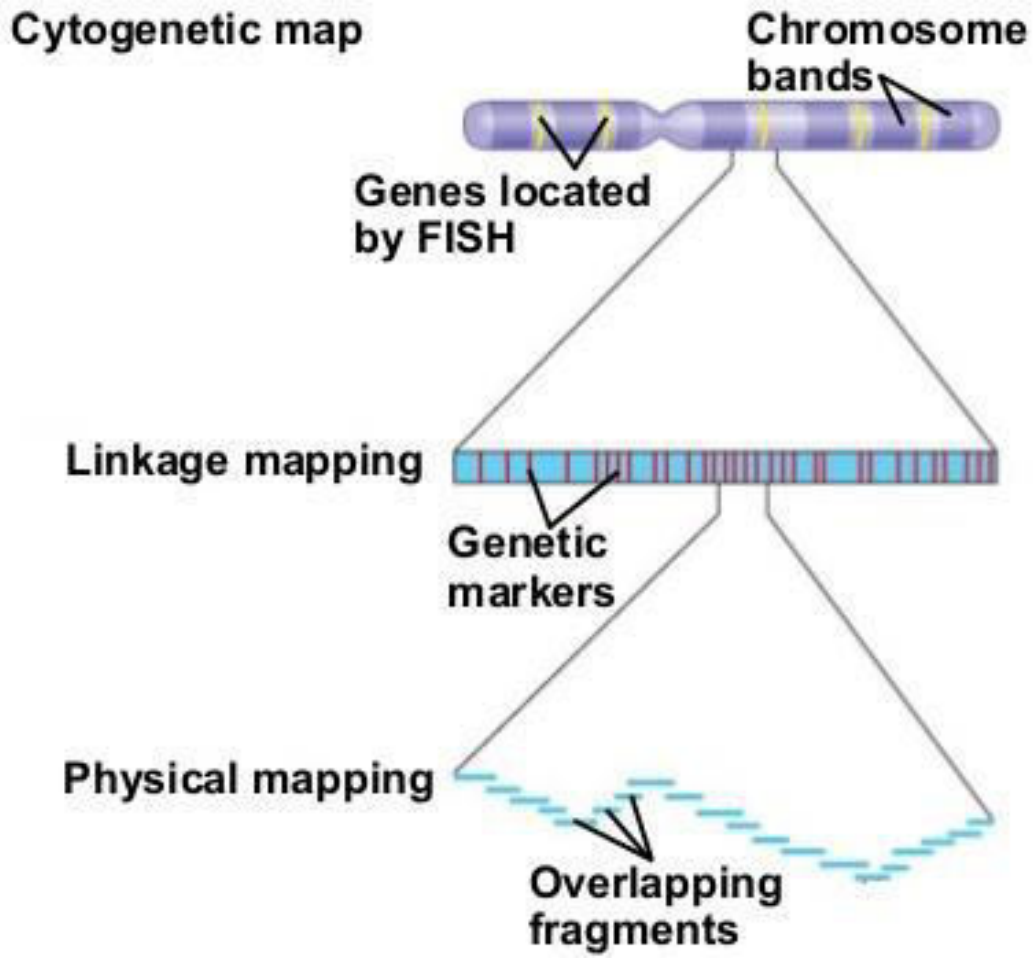
Physical Mapping of Genomes

Physical Mapping

- In humans, even the detailed genetic map lacks the required resolution.
- Therefore, a physical map derived directly from genomic DNA rather than analysis of recombinants has been generated.

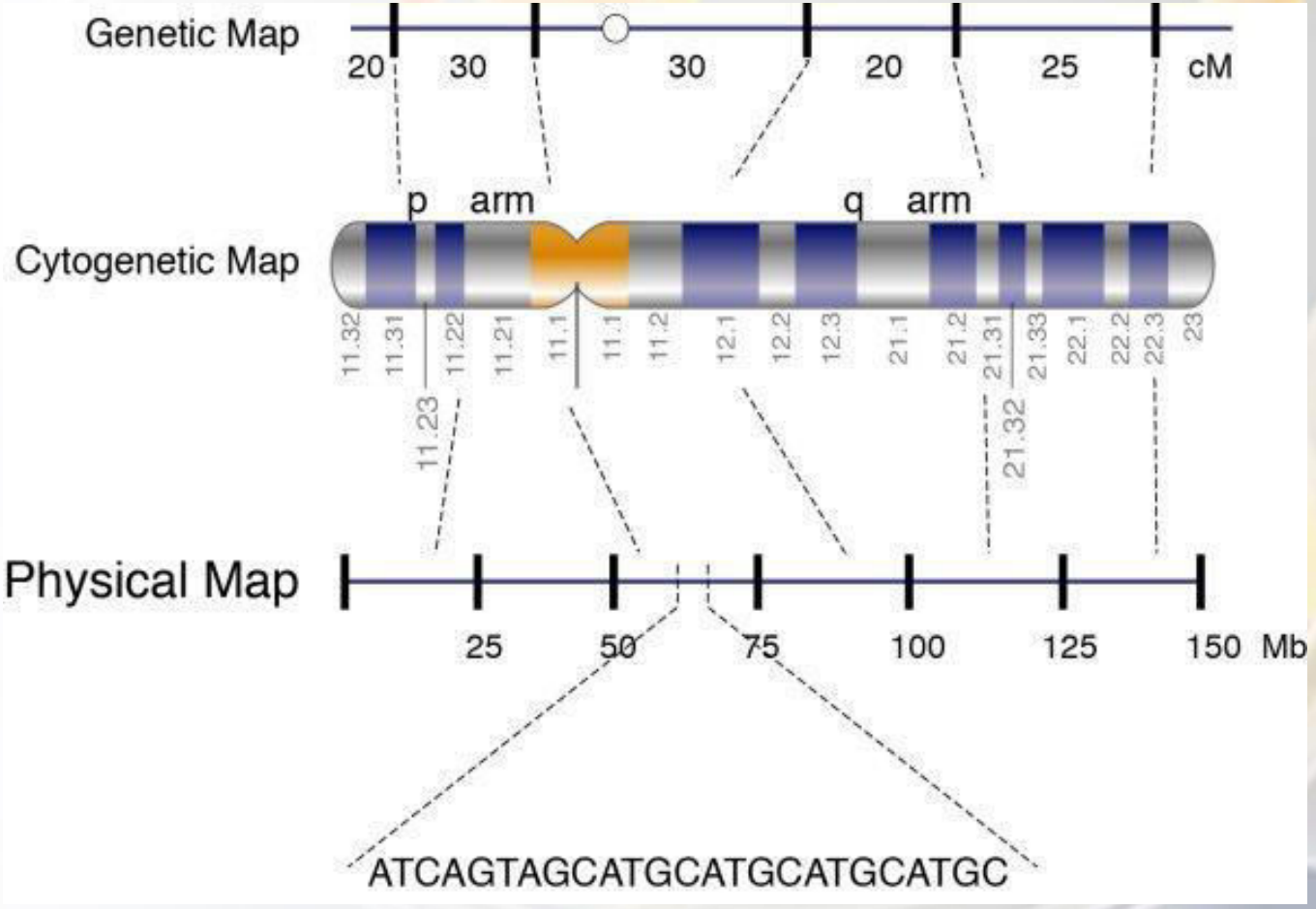
Physical Mapping of Genomes

Cytogenetic Map, Genetic Map, Physical Map



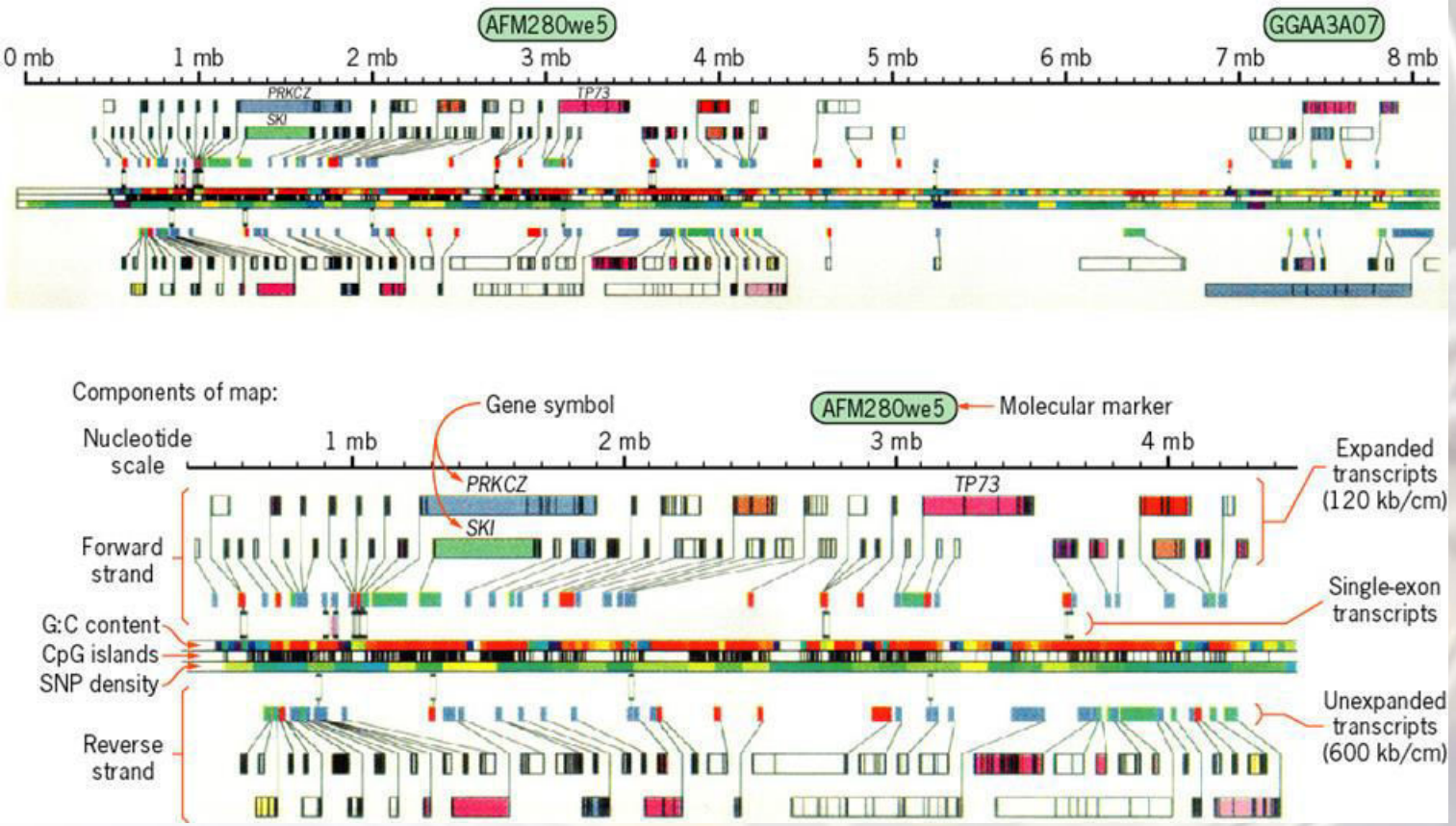
Physical Mapping of Genomes

Genetic Map and Physical Map



Physical Mapping of Genomes

Physical Map



Physical Mapping of Genomes

Physical Mapping of Genomes

- As in human, there are 24 physical maps, 22 autosomes plus X and Y.
- Types of physical maps are presented in order of increasing resolution:

Physical Mapping of Genomes

Physical Mapping Techniques

- Cyotogenetic map
- A restriction map
- Fluorescent in situ hybridization
- Sequence tagged site (STS) map
- Nucleotide sequence map

Physical Mapping of Genomes

Physical Mapping Techniques

- Plasmid
- Phagemid
- Cosmid
- YACs
- BACs

Physical Mapping of Genomes

Physical Map of Genomes

- A physical map of a chromosome or genome that shows the physical location of genes/DNA
- Low resolution - banding patterns; highest-resolution nucleotide sequence

Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes

Cytogenetic Mapping

- Microscopic examination of stained chromosome reveals a banding patterns
- Regions - designated based on their chromosomal position relative to the centromere

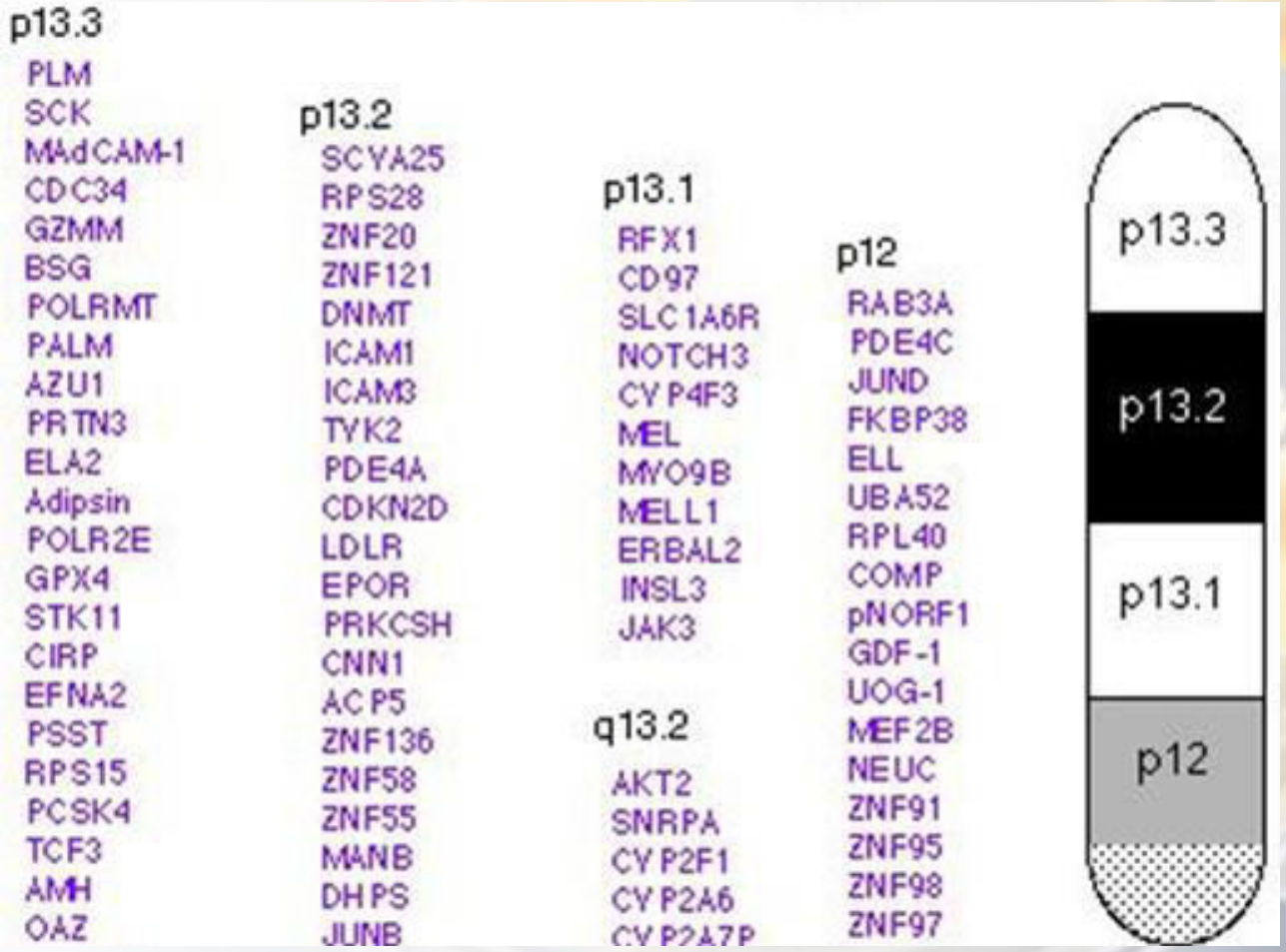
Cytogenetic Mapping of Genomes

Cytogenetic Mapping

- Chromosomal banding pattern.
- These methods allow a rough determination of locations, but not to yield a direct measure of distance.

Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes



Cytogenetic Mapping of Genomes

Cytogenetic Mapping

- Regions designated “q” are on the chromosome’s long arm. Regions designated “p” are on the short arm
- Regions are numbered from the centromere outward, with q1 and p1

Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes

- Cytogenetic
  - Visual study of chromosomes at microscopic level
- Karyotype
  - Chromosome complement
  - also applied to picture of chromosomes
- Idiogram
  - Stylised form of karyotype

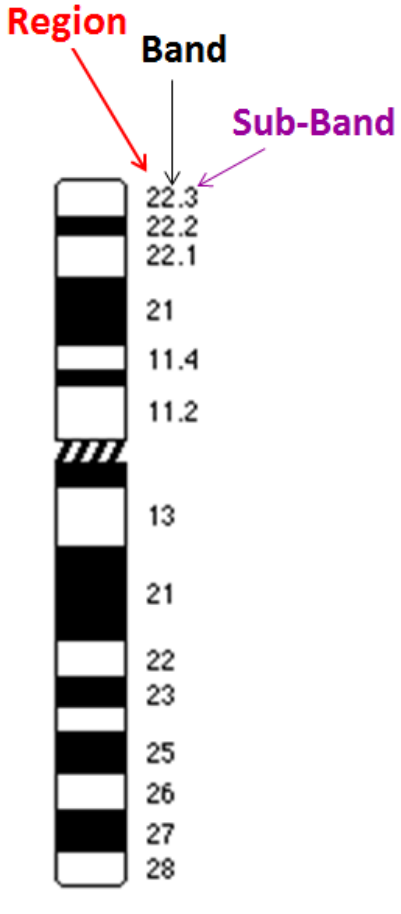
Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes

- Classified according to position of centromere
- Central centromere - metacentric
- Sub-terminal centromere - acrocentric
  - have satellites which contain multiple copies of genes for ribosomal RNA on short arm
- Intermediate centromere – submetacentric
- Terminal centromere - Telocentric

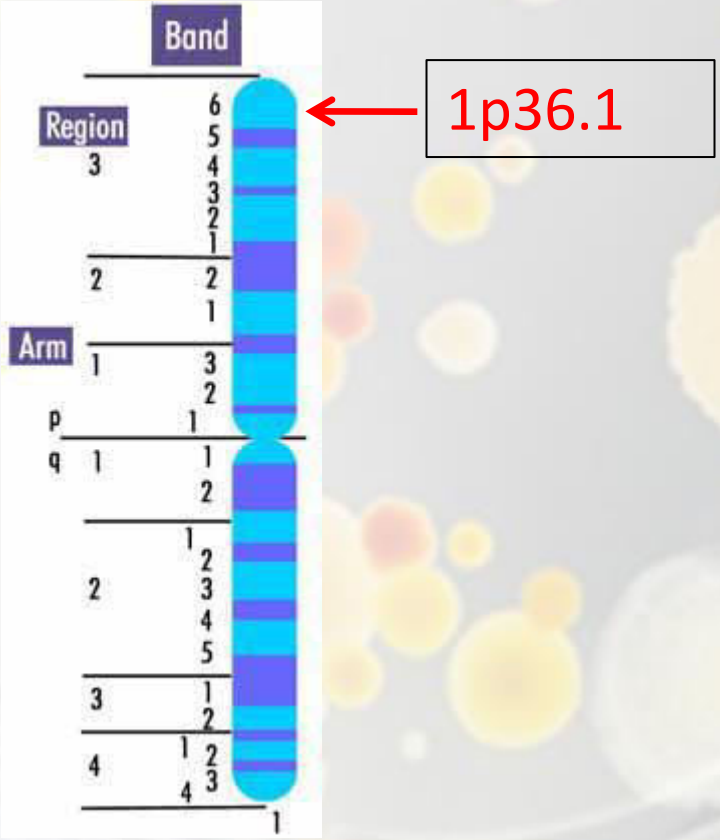
Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes



Cytogenetic Mapping of Genomes

Cytogenetic Mapping of Genomes: Gene is located at Chromosome 1p36.1



Cytogenetic Mapping of Genomes

Cytogenetic Mapping

- Microscopic examination of stained chromosome reveals a pattern of bands.
- Regions are designated based on their chromosomal position relative to the centromere

FISH Mapping of Genomes

FISH Mapping of Genomes

FISH Mapping

- Fluorescence in situ hybridization (FISH) is a type of mapping that uses fluorescent probes that bind to only those parts of the chromosome with a high degree of sequence complementarity.

FISH Mapping of Genomes

FISH Mapping

- It is used to detect and localize the presence or absence of specific DNA sequences on chromosomes.

FISH Mapping of Genomes

FISH Mapping

- Fluorescence microscopy used to find out where fluorescent probe is bound to the chromosomes.
- Used in genetic counseling, medicine and species identification.

FISH Mapping of Genomes

FISH Mapping

- Individual metaphase chromosomes are probed in situ with specific fluorescently labeled DNA sequences, identifying homologous sequences in the chromosome

FISH Mapping of Genomes

FISH Mapping

- Different probes labeled with different fluorescent dyes may be used in same experiment.
- Fluorescence microscopy provides data for computer imaging analysis to determine binding site for each probe.

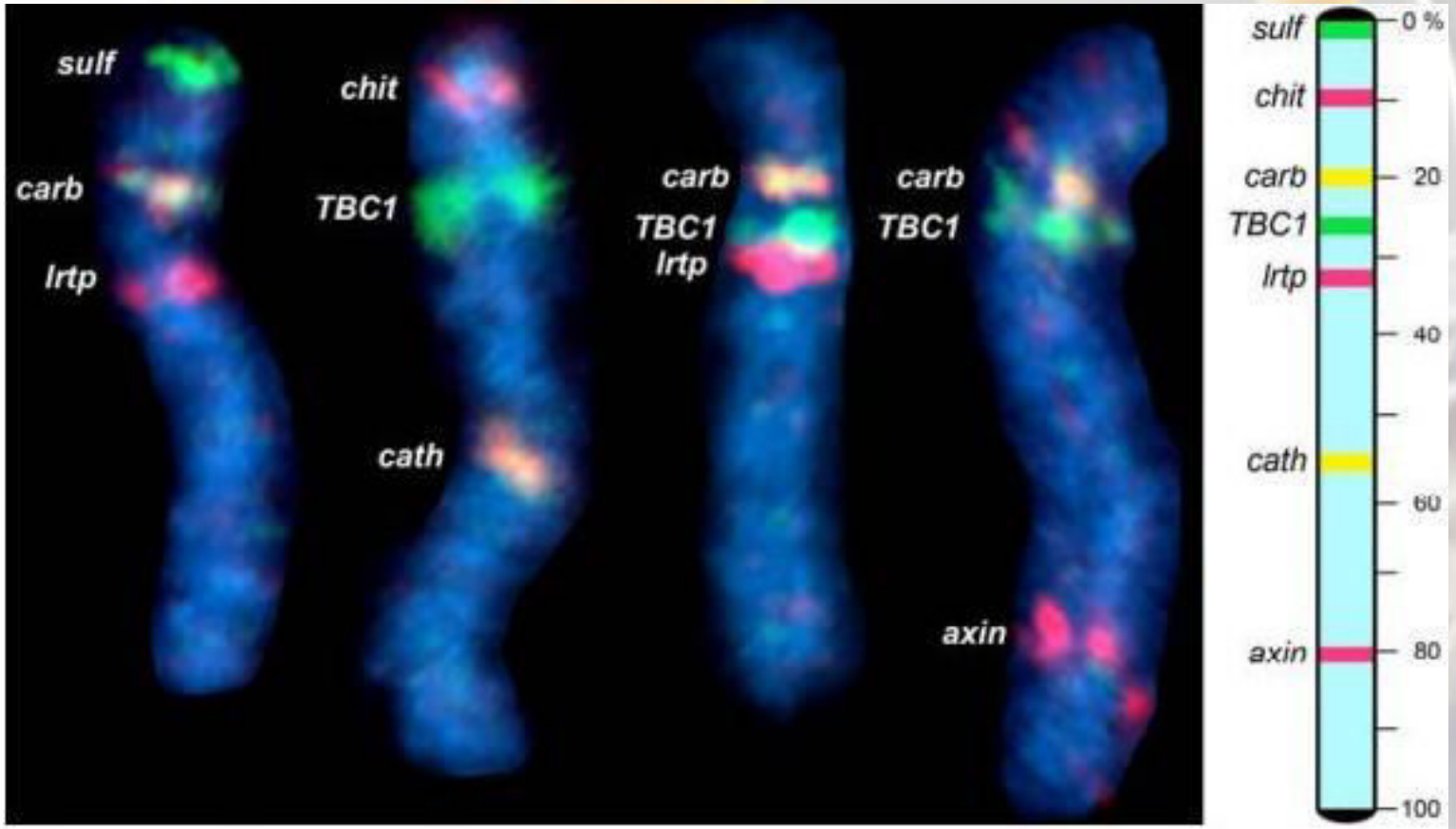
FISH Mapping of Genomes

FISH Mapping

- With a resolution of 2-5 Mb in metaphase chromosomes, FISH can localize markers to sub-regions of chromosomal bands.
- Less condensed chromosomes may be resolved in the 5-700 kb range.

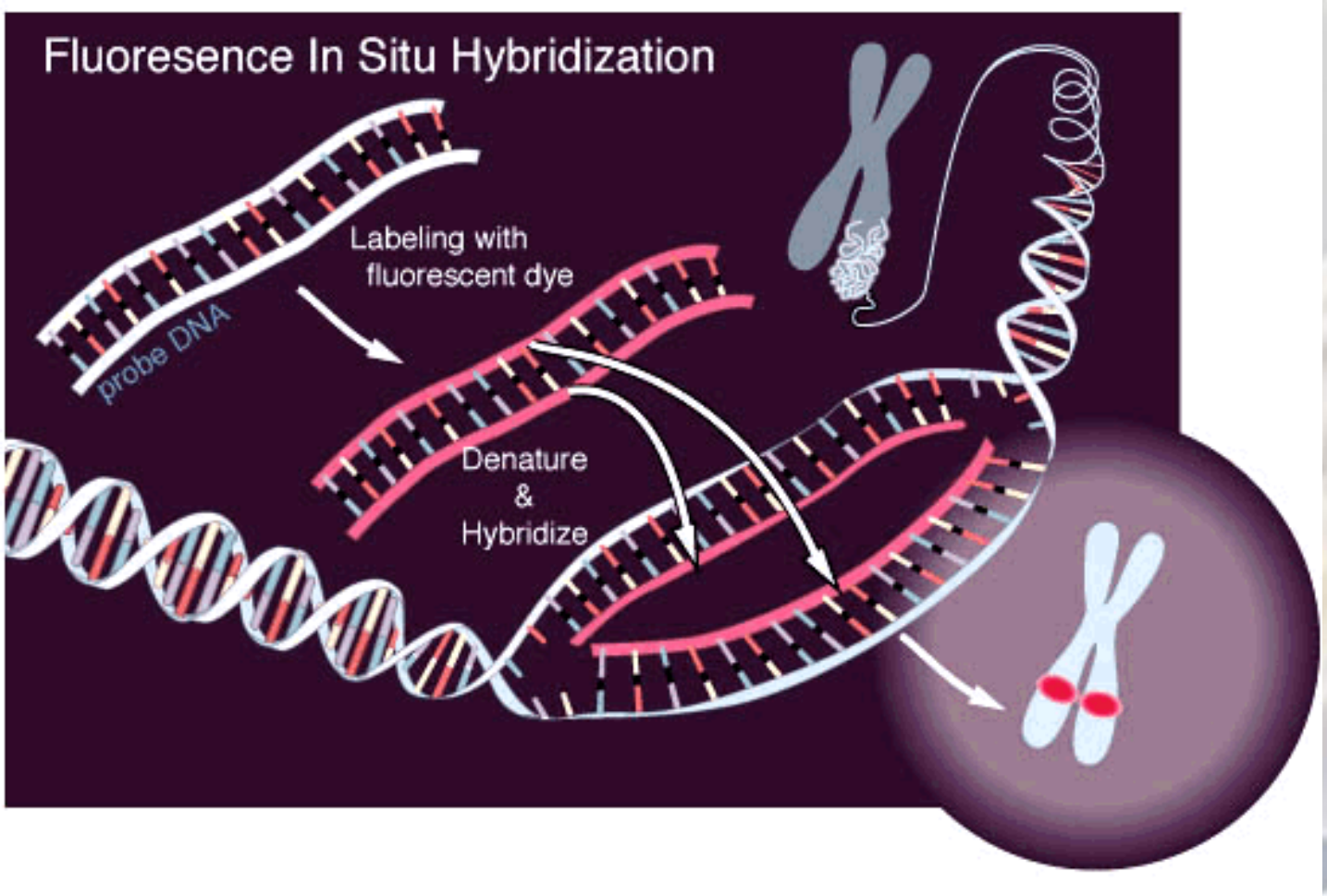
FISH Mapping of Genomes

FISH Mapping of Genomes



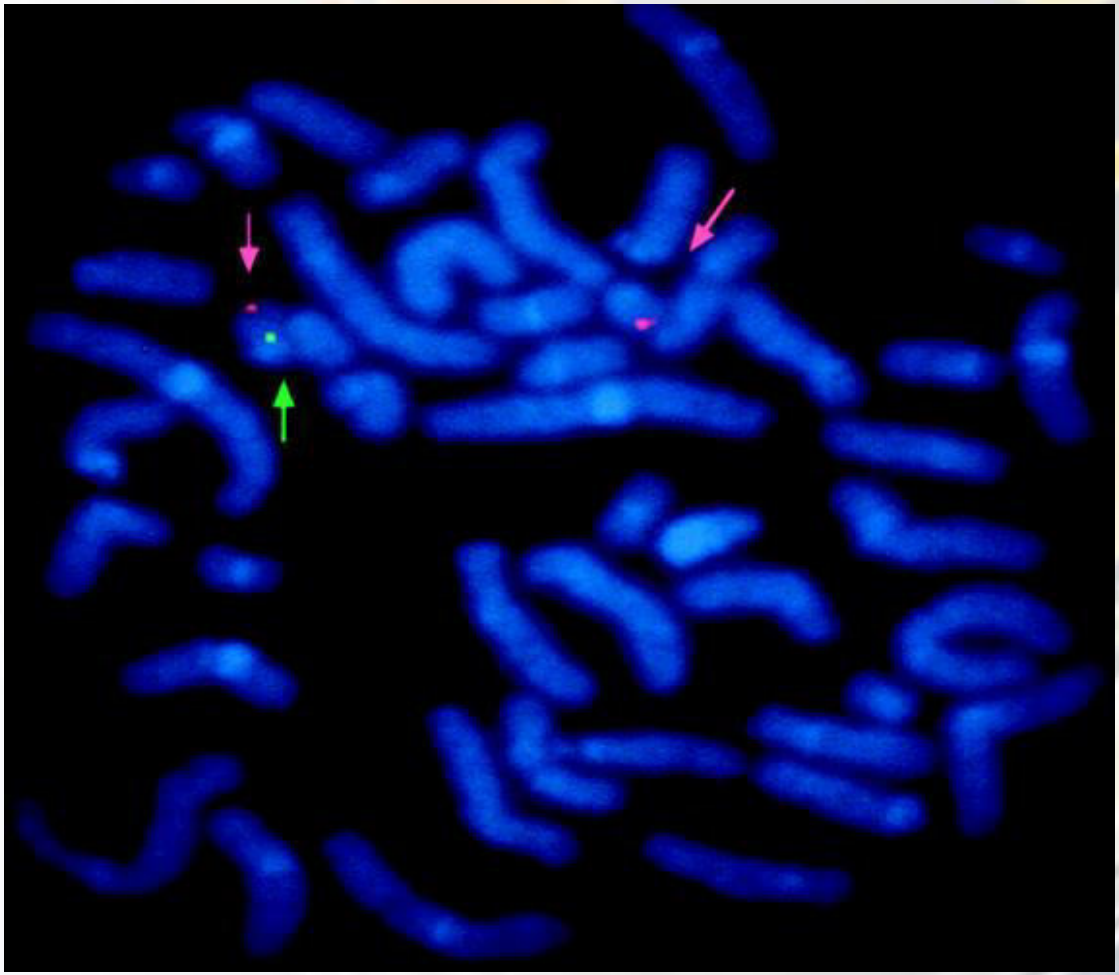
FISH Mapping of Genomes

FISH Mapping of Genomes



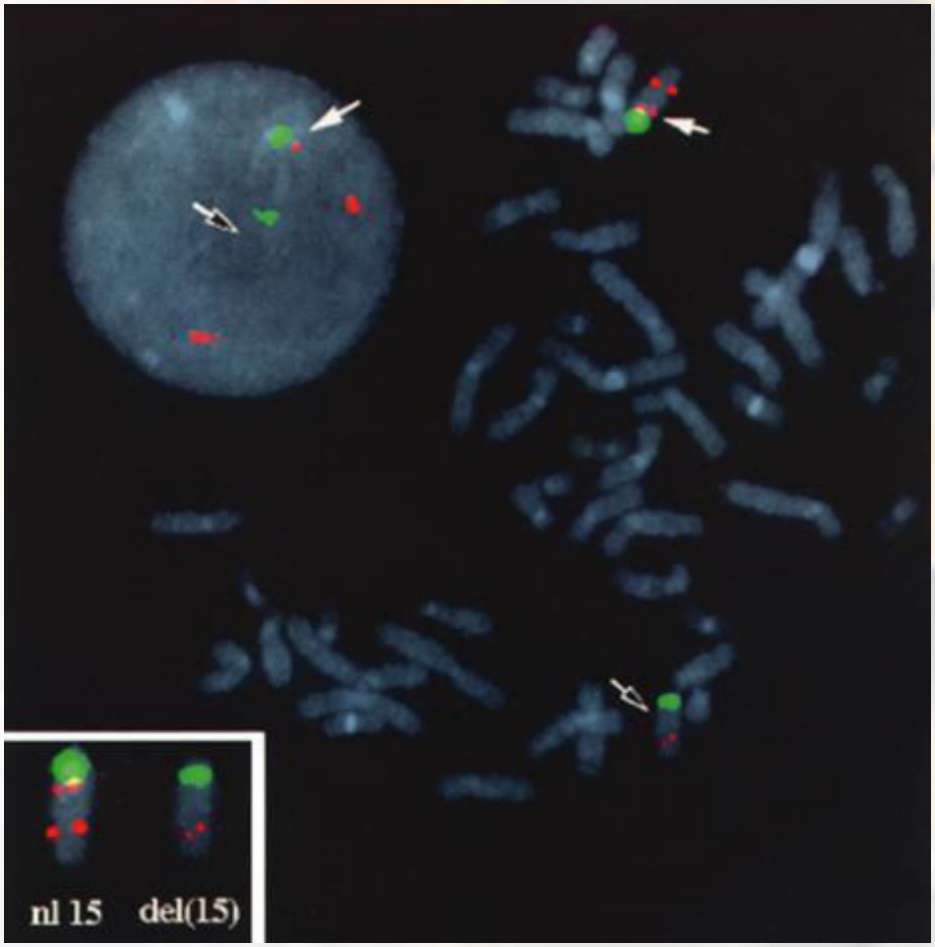
FISH Mapping of Genomes

FISH Mapping of Genomes - Microdeletions



FISH Mapping of Genomes

FISH Mapping of Genomes - Microdeletions



FISH Mapping of Genomes

FISH Mapping of Genomes

- Advantages

  Highly specific, Microdeletions /Microduplications

- Disadvantages

  500-600 probes needed to match the power of karyotyping

Genetics and Genomics

Restriction Mapping of Genomes

Restriction Mapping of Genomes

Restriction Mapping

- A restriction map is a map of known restriction sites within a sequence of DNA.
- Restriction mapping requires the use of restriction enzymes.

Restriction Mapping of Genomes

Restriction Mapping

- Restriction enzymes are used that cut DNA due to the recognition sequence in the DNA/genome under study

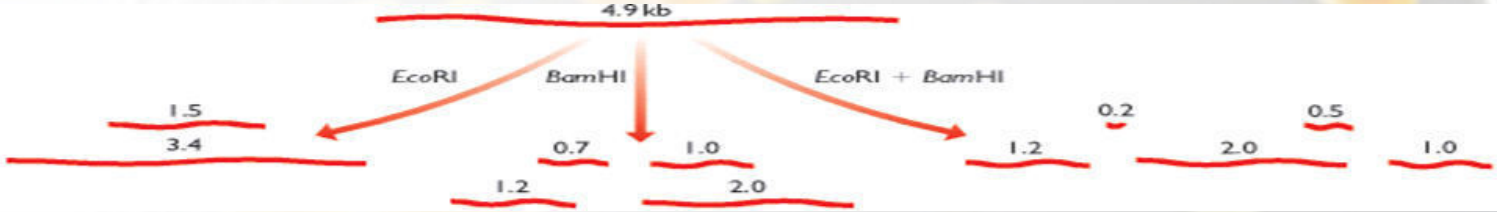
Restriction Mapping of Genomes

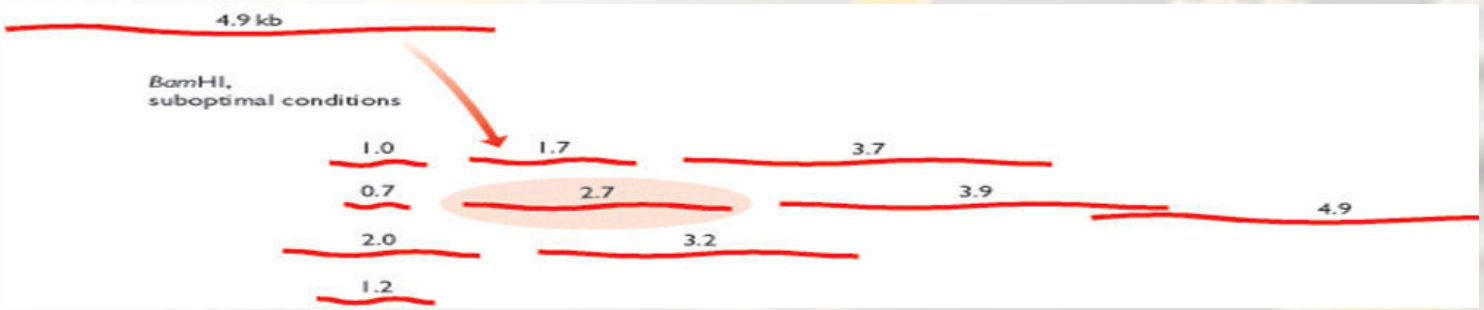
Restriction Mapping

- To construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with different restriction enzymes that recognize different target sequences.

Restriction Mapping of Genomes

Restriction Mapping of Genomes

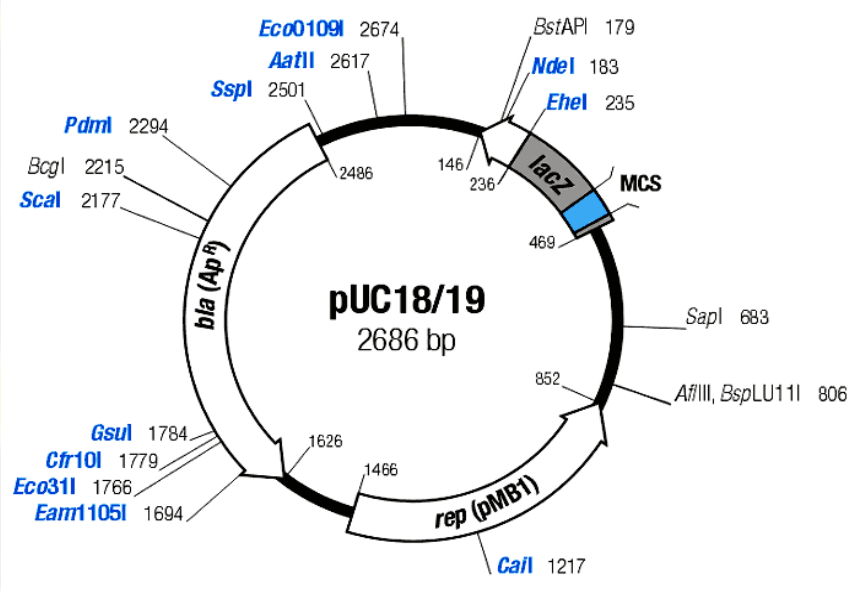




Restriction Mapping of Genomes

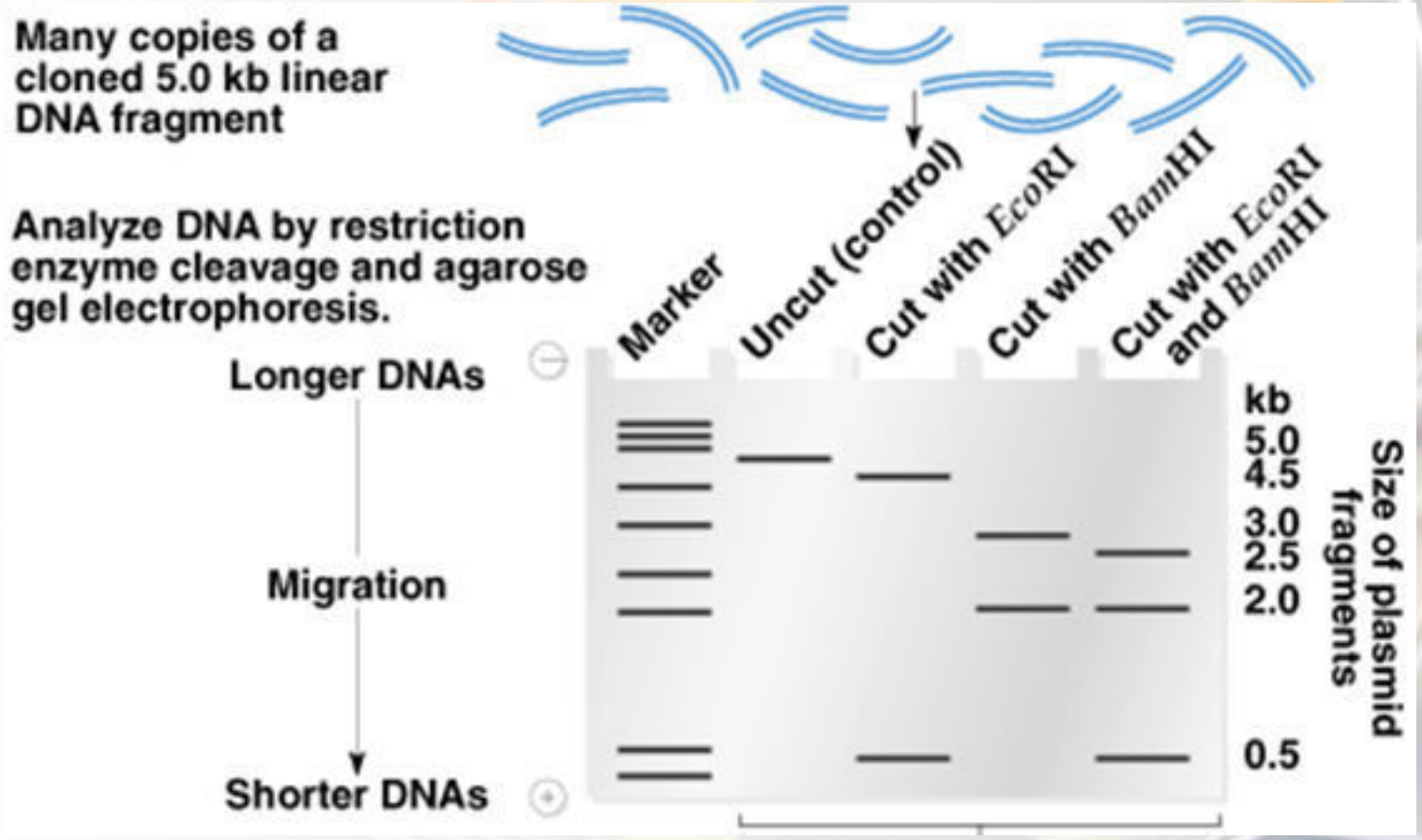
Restriction Mapping of Genomes

- A map showing positions of restriction sites in a DNA sequence
- If DNA sequence is known then construction of restriction map is a trivial exercise
- In early days of molecular biology DNA sequences were often unknown



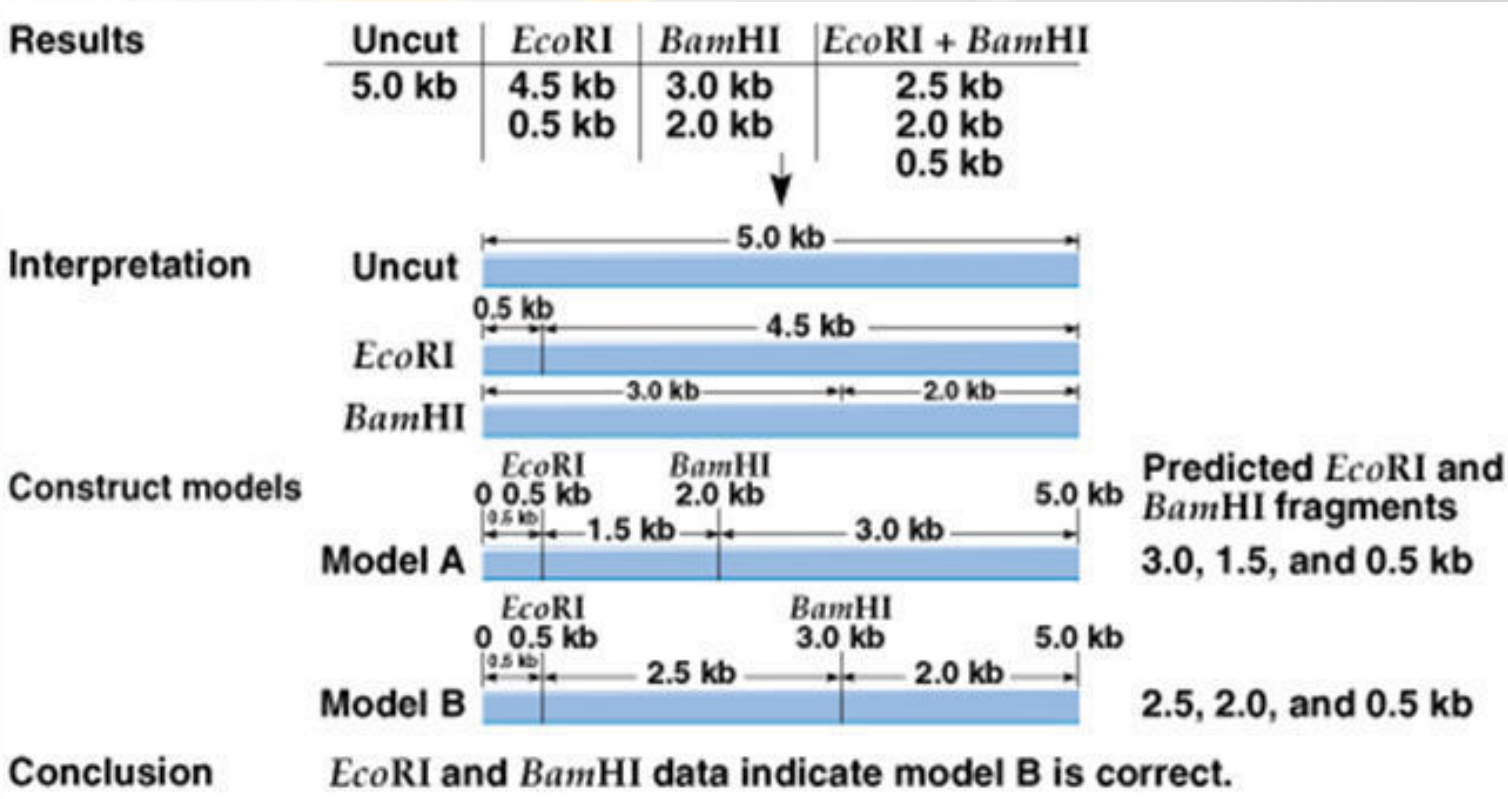
Restriction Mapping of Genomes

Restriction Mapping of Genomes: Example



Restriction Mapping of Genomes

Restriction Mapping of Genomes: How to Construct Restriction Enzyme Map



Restriction Mapping of Genomes

Limitations of Restriction Mapping

- Restriction mapping is more applicable to small rather than large molecules.

Restriction Mapping of Genomes

Limitations of Restriction Mapping

- In practice, if a DNA molecule is less than 50 kb in length it is usually possible to construct an restriction map.
- To construct map for large genomic size is difficult

Restriction Mapping of Genomes

Limitations of Restriction Mapping

- The limitations of restriction mapping can be eased slightly by choosing enzymes expected to have infrequent cut sites(rare cutter) in the target DNA molecule.

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping of Genomes

- A radiation hybrid is a rodent cell line carrying a small genomic DNA molecule from another organism e.g. a human.

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping of Genomes

- Exposure to X rays breaks the DNA in human cells.
- The fragments become smaller with more X ray exposure
- Fragments length determines the map resolution

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping of Genomes

- Irradiation kills the human cells, which are then fused with rodent cells, rescuing chromosomal fragments that are typically a few Mb in length.

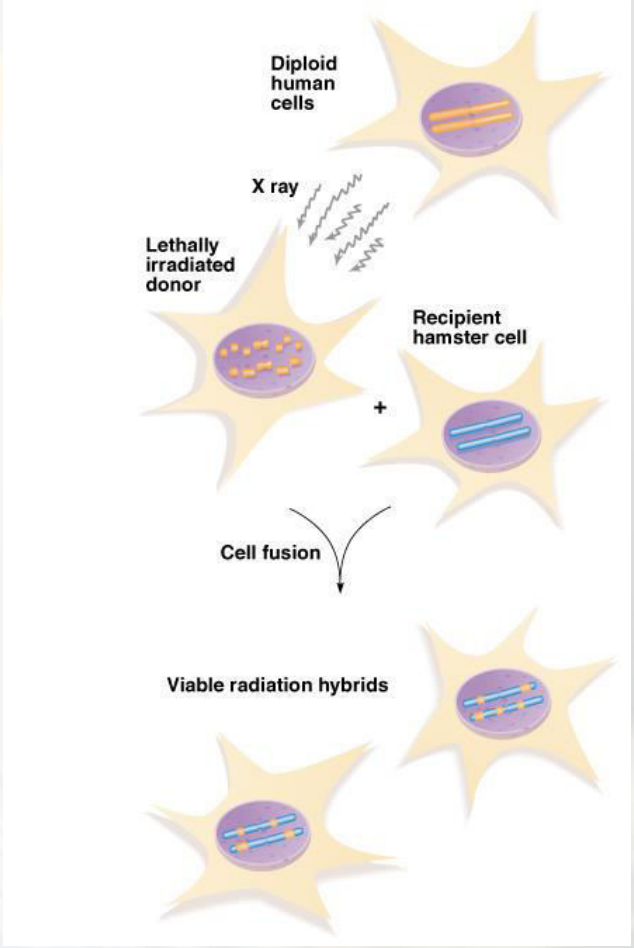
Radiation Hybrid Mapping of Genomes

Radiation Hybrid

- Human DNA in the RH is analyzed for gene and/or DNA markers.
- Closer the two markers to each other – chromosome; the more likely they are to be found together in an RH.

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping of Genomes



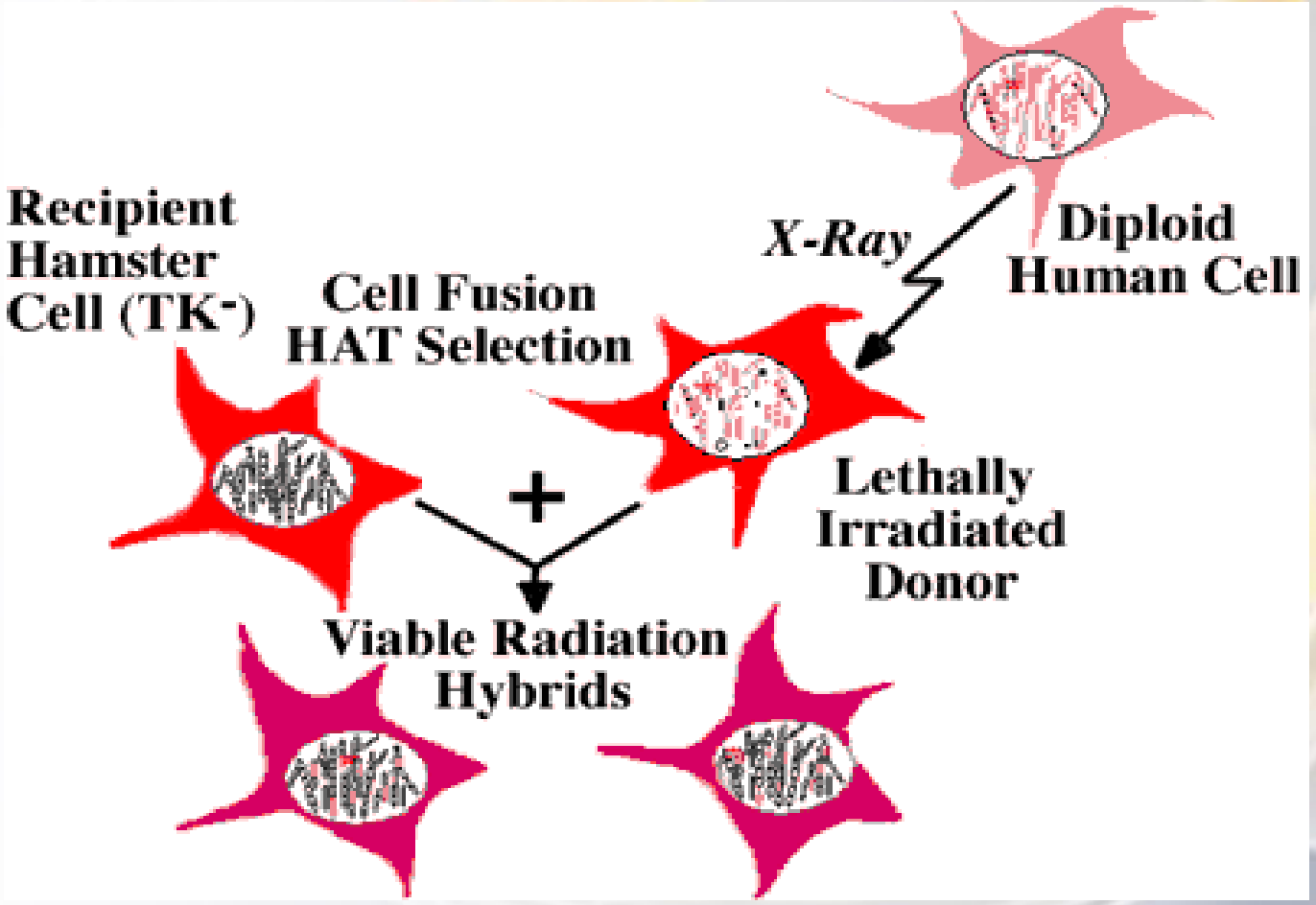
Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping: Methodology

- Standard somatic cell fusions contain entire human chromosomes. To locate a gene more closely, you need to use chromosome fragments.
- Start by irradiating human cells with a controlled dose of X-rays: chromosomes break up. Then, fuse the cells to rodent cells.
- The human chromosome fragments get integrated into mouse/rodent chromosomes
- Hybrid cell lines contain random human chromosome fragments.

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping



Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping

- Radiation hybrid mapping is a method for high-resolution mapping.
- Exploits the ability of rodent cells (hamster cells) to stably incorporate genetic material from fused cells.

Radiation Hybrid Mapping of Genomes

Radiation Hybrid Mapping

- Resolution is tunable
- Relatively cheap
- Difficult to compare results from different groups

Clone Contigs Mapping of Genomes

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- A partial restriction digest produces a set of large, overlapping DNAs, which are cloned into YAC/BAC vector

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- Shearing may also be used to make high-molecular-weight DNA that is blunt-end cloned into a YAC/BAC.

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- An entire genome or single chromosome may be represented in a YAC clone library
- YAC clones are then assembled into a map.

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- Maps can be generated either by matching with a FISH-generated chromosome map or by DNA fingerprinting
- Assembly of clone contigs is based on clone overlaps.

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- Non-polymorphic short tandem sequence are especially useful for YAC contig mapping

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- A complete library should yield a complete contig map that indicates the order in which the cloned fragments occur in the chromosome

Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- Problems arise when some of the YAC inserts contain DNA from more than one chromosomal location.
- Complicated the efforts at generating a YAC contig map of human chromosome

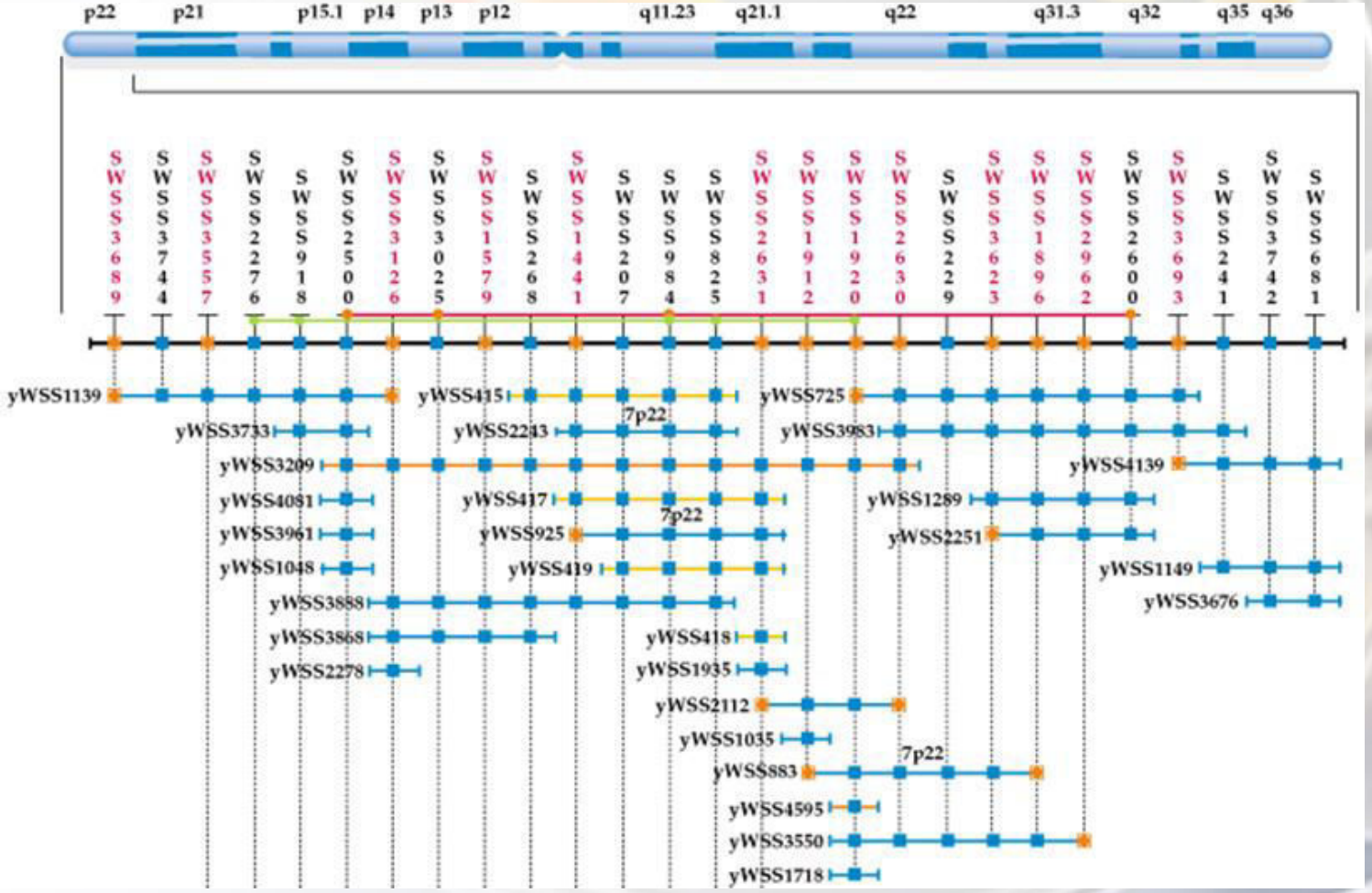
Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- Many labs have switched to BAC vectors with a capacity of 300 kb and the ability to replicate in *E. coli* as a resource for their sequencing projects.

Clone Contigs Mapping of Genomes

YAC Contigs Map of Chromosome 7



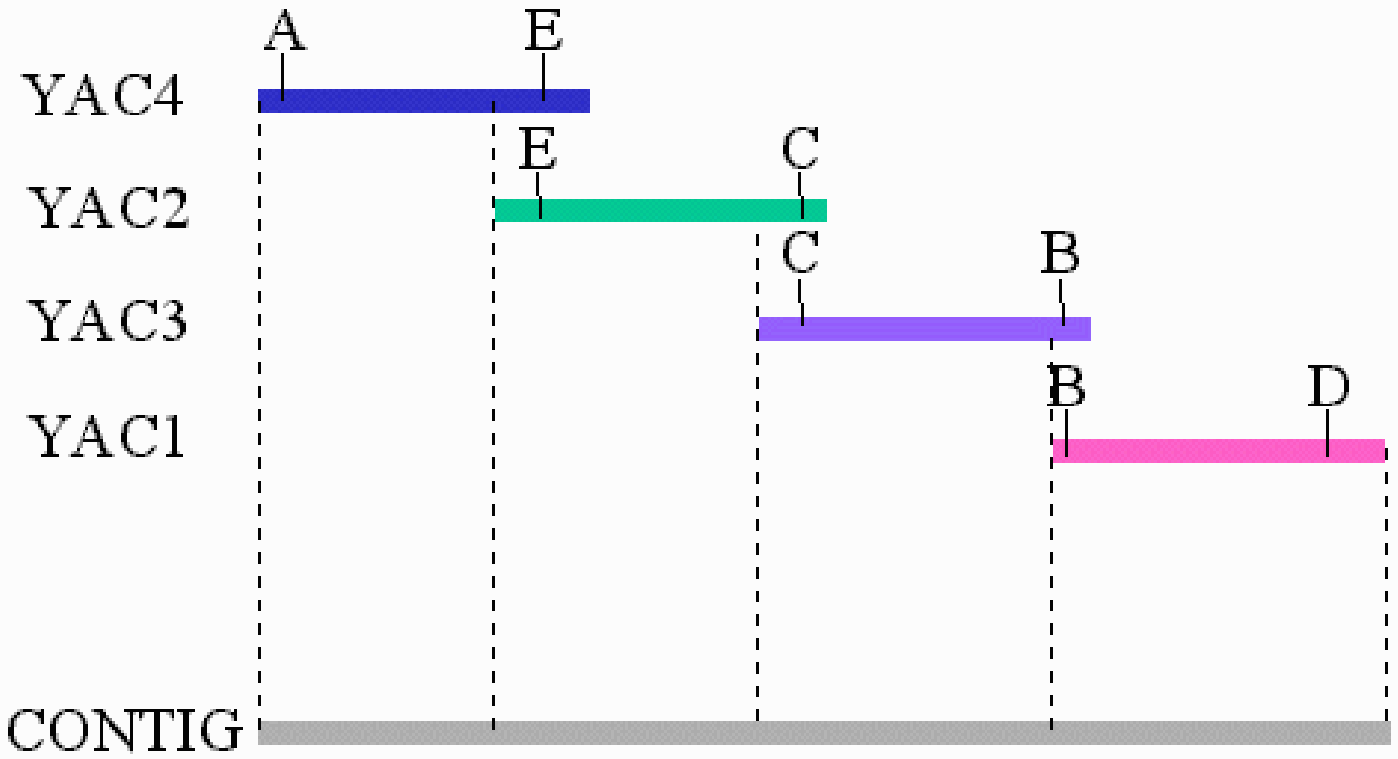
Clone Contigs Mapping of Genomes

Clone Contigs Mapping of Genomes

- A contig is a set of partially overlapping clones, a contiguous set of clones. No gaps between them.
- Contigs allow you to build up the sequence of the chromosome over much larger regions than any single clone.
- The first reasonably complete physical map of the human genome involved contigs generated by YACs.

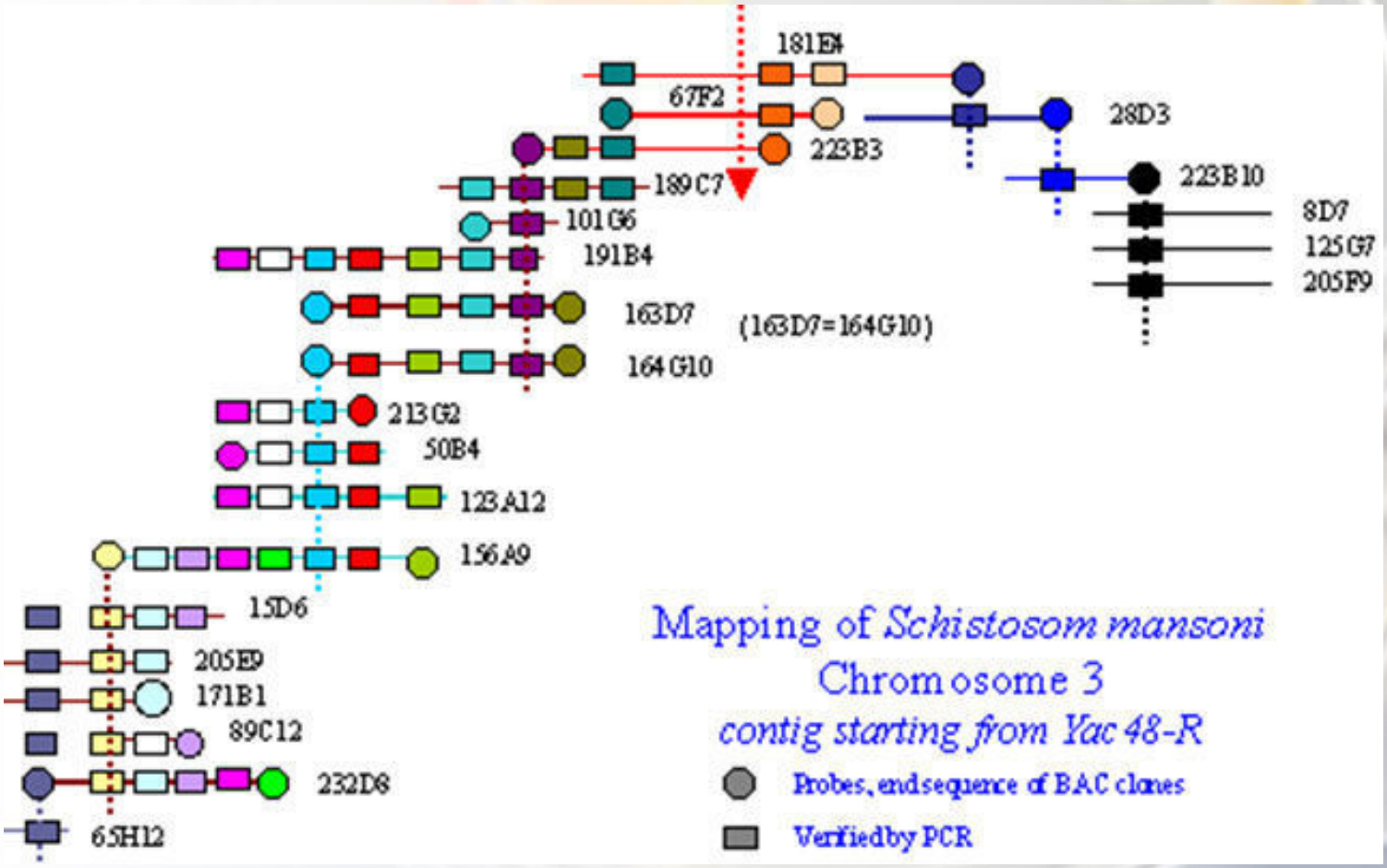
Clone Contigs Mapping of Genomes

Clone Contigs Mapping of Genomes



Clone Contigs Mapping of Genomes

Clone Contigs Mapping of Genomes: Yac48-R



Clone Contigs Mapping of Genomes

Clone Contigs Mapping

- An entire genome or single chromosome may be represented in a YAC/BAC clone library

Genetics and Genomics

Generating Sequence of Genomes

Generating Sequence of Genomes

Generating Sequence of Genomes

- Dideoxy sequencing used to sequence the genomes

Generating Sequence of Genomes

Generating Sequence of Genomes

- One sequencing reaction is limited to on average 500 nucleotides, and for accurate sequences both strands were sequenced.

Generating Sequence of Genomes

Generating Sequence of Genomes

- Two competing sequencing strategies used
- Clone-by-clone
- Whole-genome shotgun

Generating Sequence of Genomes

Generating Sequence of Genomes: Clone-by-Clone Shotgun Sequencing

- Map construction
- Clone selection
- Sub-clone library construction
- Random shotgun phase
- Directed finishing phase and sequence authentication

Generating Sequence of Genomes

Generating Sequence

- Human genome sequencing by the mapping approach used BACs
- BAC insert too large to sequence in one reaction.
- BAC inserts were sequenced using a shotgun approach

Generating Sequence of Genomes

Generating Sequence of Genomes

- Each insert is cut from the vector, sheared into fragments that will be partially overlapping and cloned into a plasmid vector.

Generating Sequence of Genomes

Generating Sequence of Genomes

- Each subclone is sequenced, and overlaps are used by a computer to assemble the data into one contiguous sequence representing the BAC insert.

Generating Sequence of Genomes

Generating Sequence of Genomes

- Using the chromosomal map for BAC clones, the BAC insert sequences are put in order to yield the complete chromosome sequence

Genetics and Genomics

Human Genome Project

Human Genome Project

HGP

- 1990 – Human Genome Project was started (NHGRI)
- Later many institutes of UK, France, Japan, Germany, China involved.
- In 1998, Celera announced a 3-year plan to sequence human genome

Human Genome Project

Challenges to Sequence Human Genome

- Size
- Polymorphism
- Repeats - smaller repeats are technically difficult to sequence
- Some DNA sequence are repeated all over the genome

Human Genome Project

Challenges to Sequence Human Genome

- Relies on cloning (Some regions are difficult to clone - Heterochromatin).
- Some sequences rearrange or are deleted when cloned.

Human Genome Project

Human Genome Project - Objectives

- Create a genetic and physical map of the 24 human chromosomes (22 autosomes, X & Y)
- Identify the entire set of genes & map them all to their chromosomes

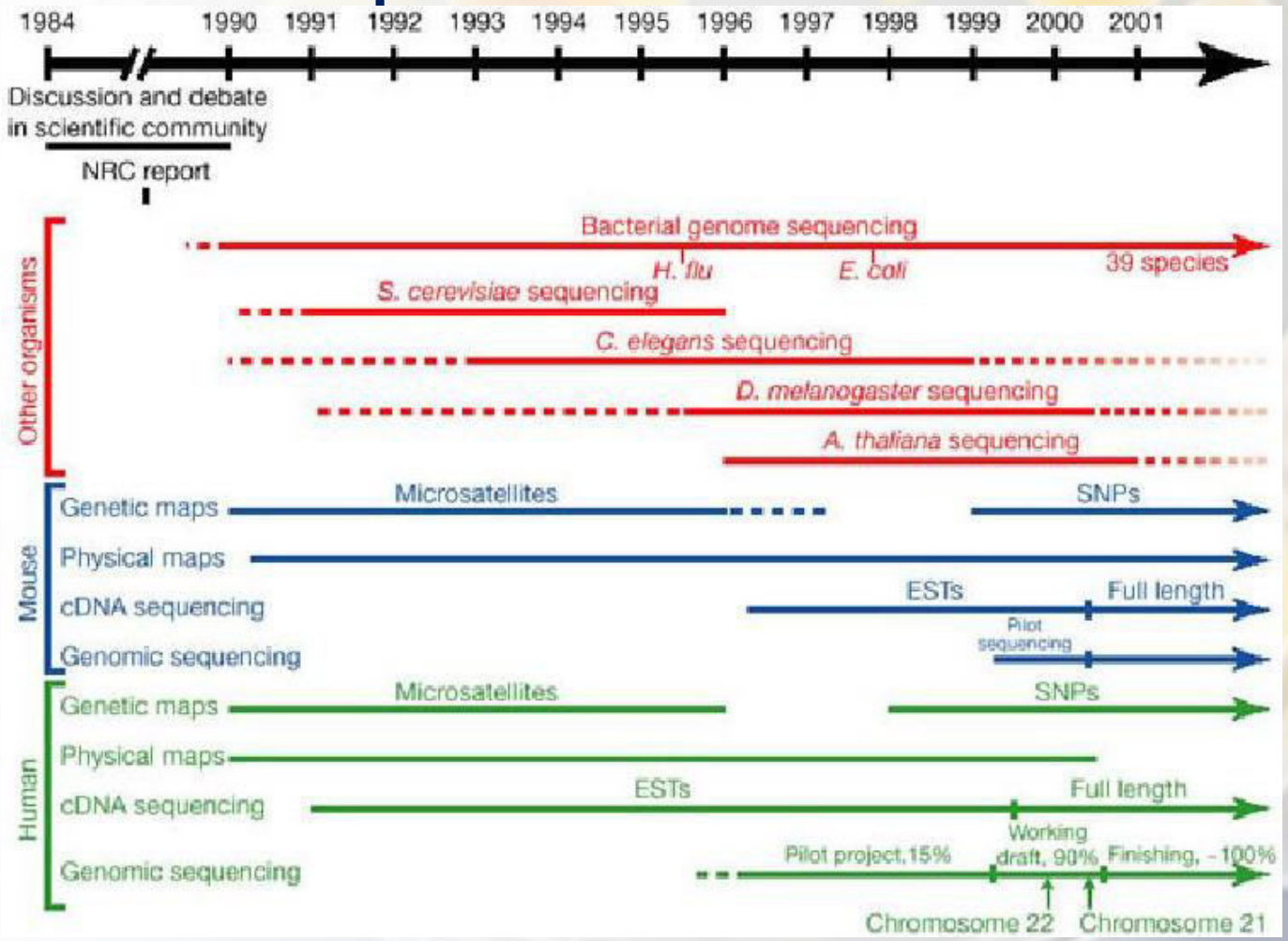
Human Genome Project

Human Genome Project - Objectives

- Determine the nucleotide sequence of the estimated 3 billion base pairs
- Analyze genetic variations in humans
- Map and sequence the genomes of model organisms

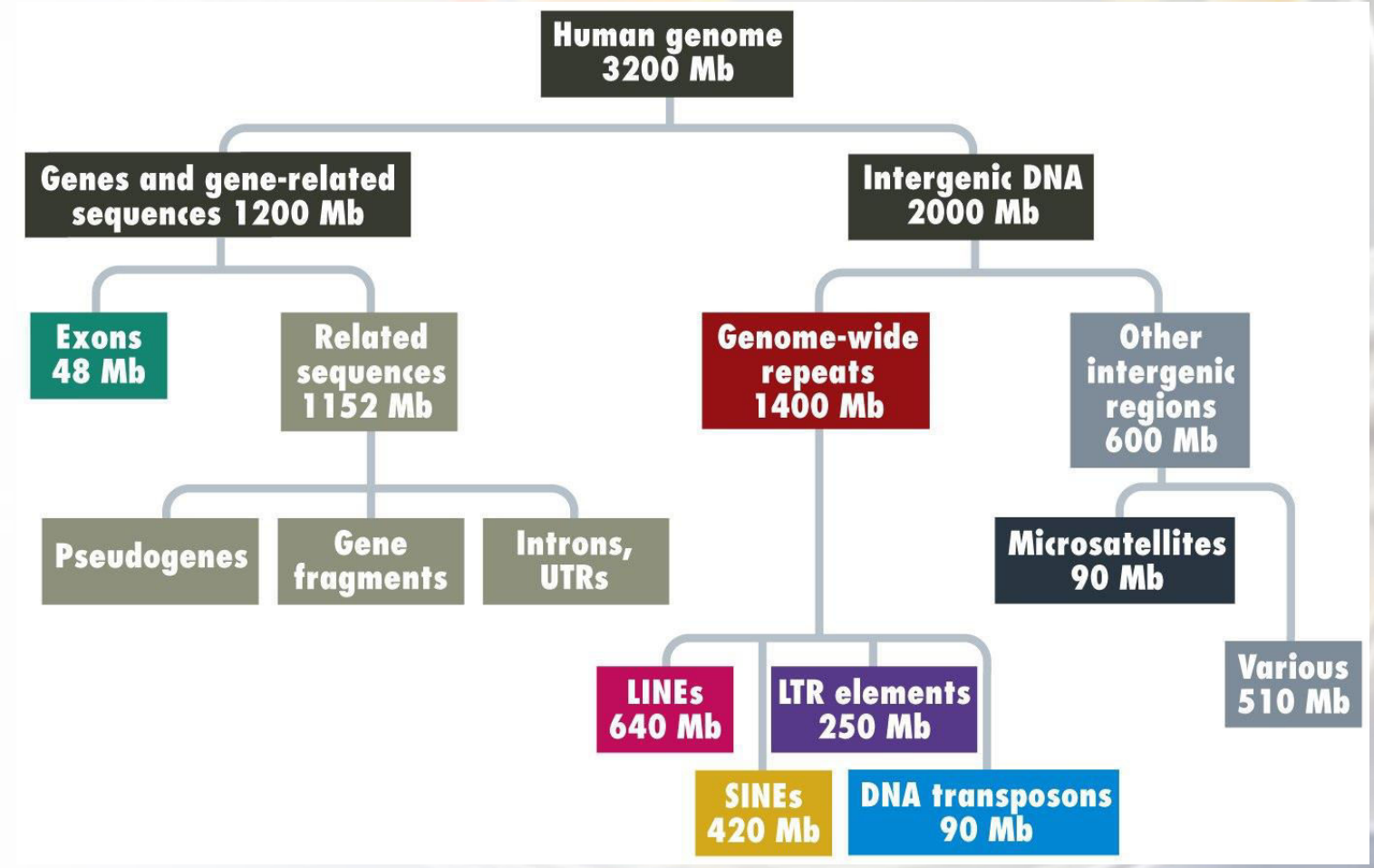
Human Genome Project

Time line to sequence Human Genome



Human Genome Project

Human Genome



Human Genome Project

Human Genome Project

- Two strategies used to sequence human genome
- Hierarchical Shotgun Sequencing
- Whole Genome Shotgun Sequencing

Human Genome Project

Human Genome Project - conclusion

- Determine the nucleotide sequence of the estimated 3 billion base pairs

Genetics and Genomics

HGP – Hierarchical Shotgun Sequencing

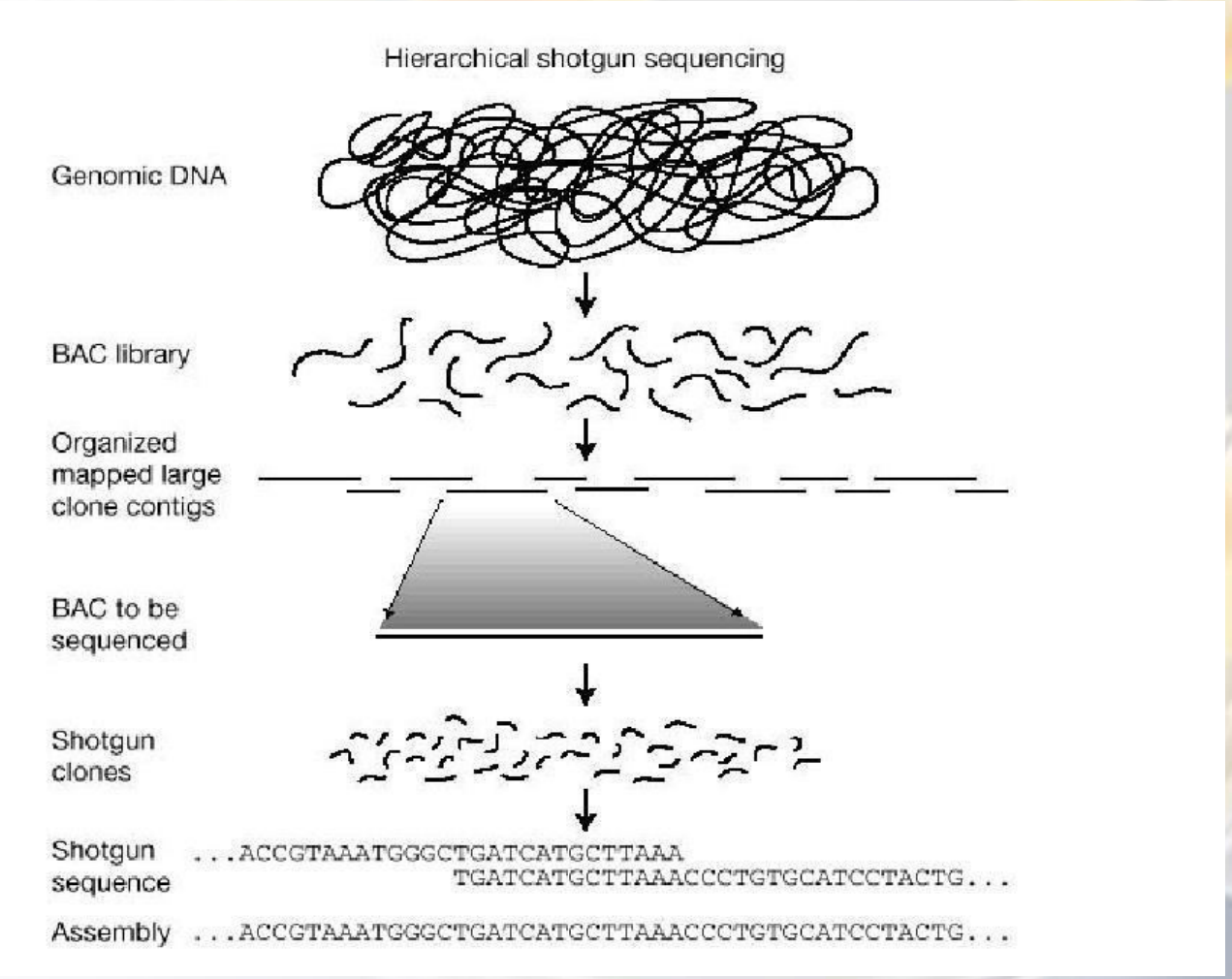
HGP–Hierarchical Shotgun Sequencing

HGP – Hierarchical Shotgun Sequencing

- Map Construction
- Clone selection
- Sub-clone library construction
- Random shotgun phase
- Directed finishing phase and sequence authentication

HGP–Hierarchical Shotgun Sequencing

HGP – Hierarchical Shotgun Sequencing



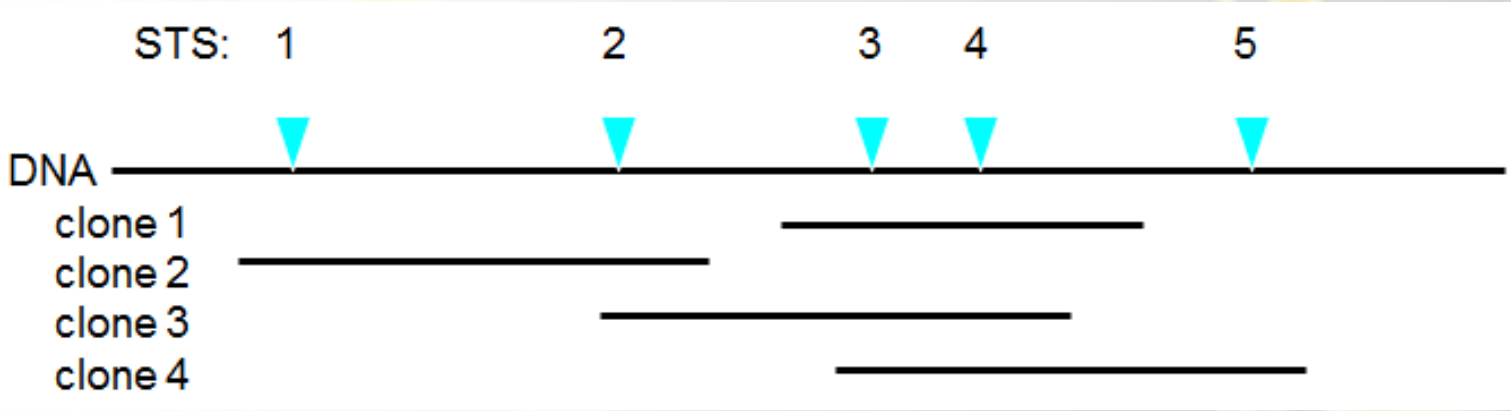
HGP–Hierarchical Shotgun Sequencing

Hierarchical Shotgun Sequencing: Map Construction

- Clone genomic DNA in YACs (~1MB) or BACs (~200-300Kb)
- Map the relative location of clones
  - Sequenced-tagged sites (STS, e.g. EST) mapping
  - PCR or probe hybridization to screen STS
  - Restriction site fingerprint
- Most time consuming
  - 1990-98 to generate physical maps for human

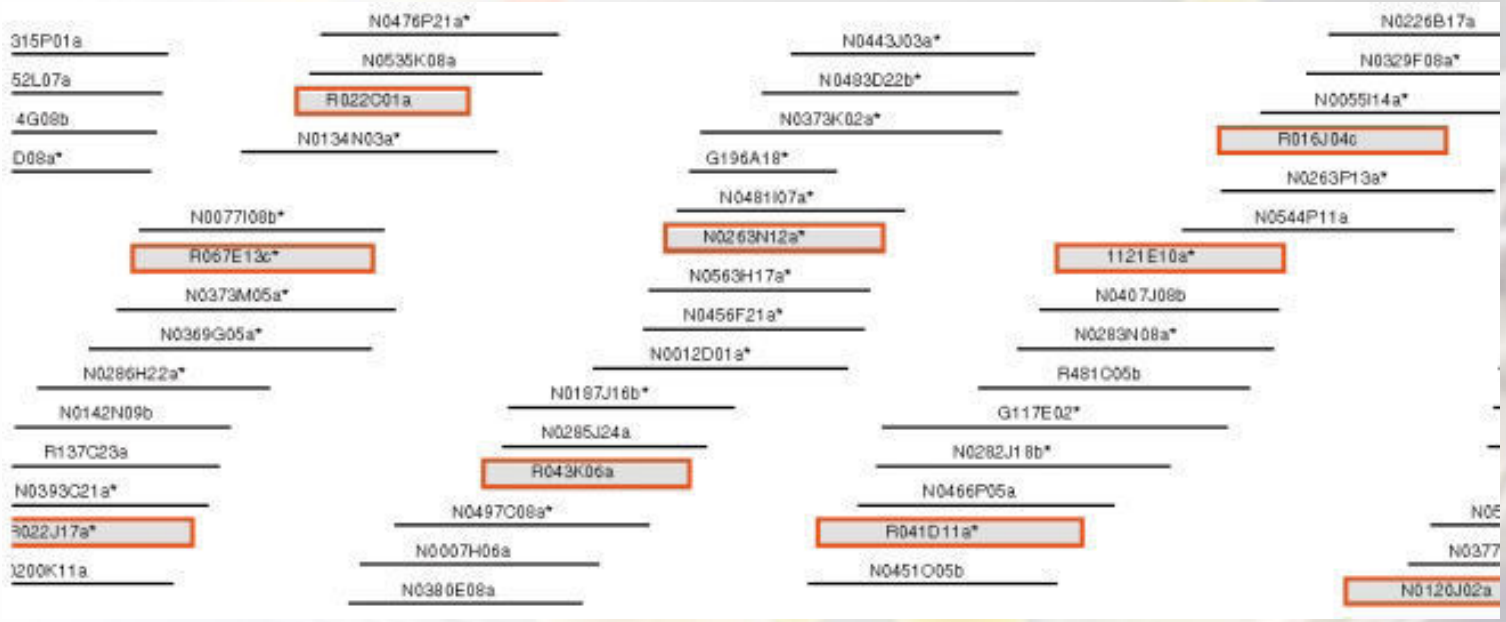
HGP–Hierarchical Shotgun Sequencing

Hierarchical Shotgun Sequencing: Resolve Clone Relative Location



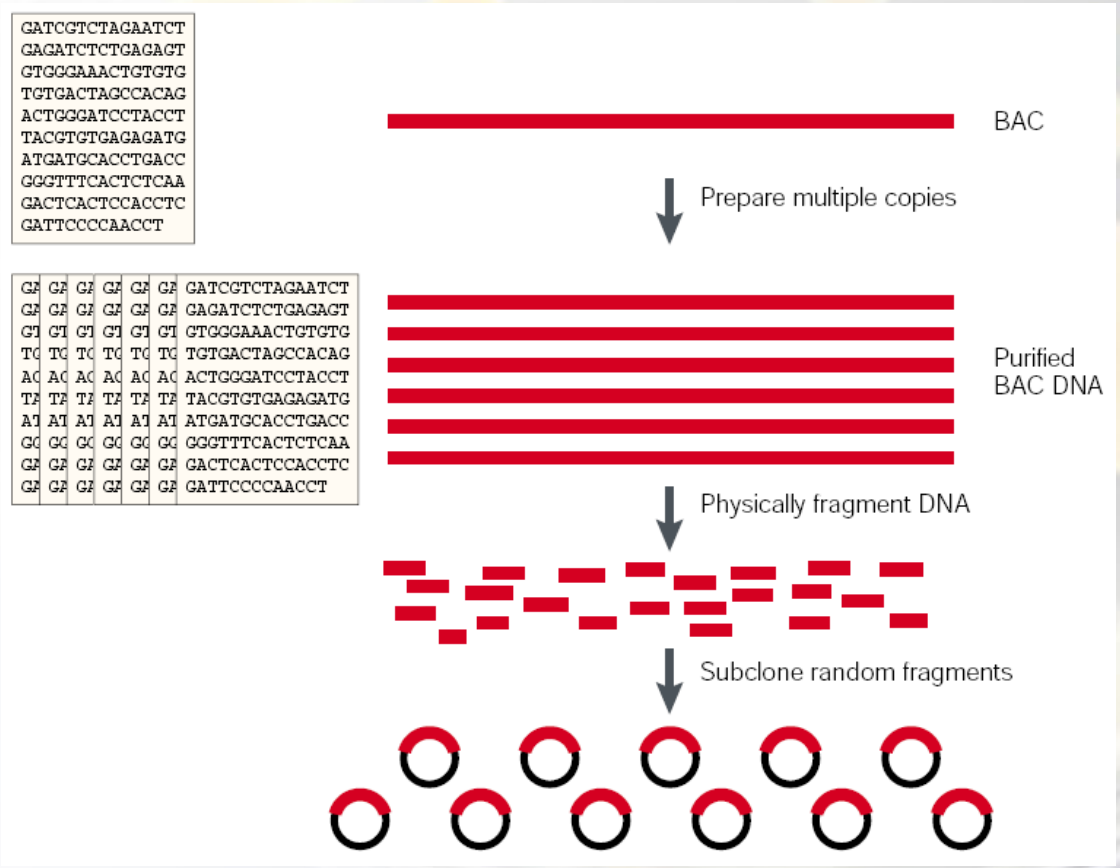
HGP–Hierarchical Shotgun Sequencing

Hierarchical Shotgun Sequencing: Clone Selection: Based on clone map, select authentic clones to generate a minimum tiling path



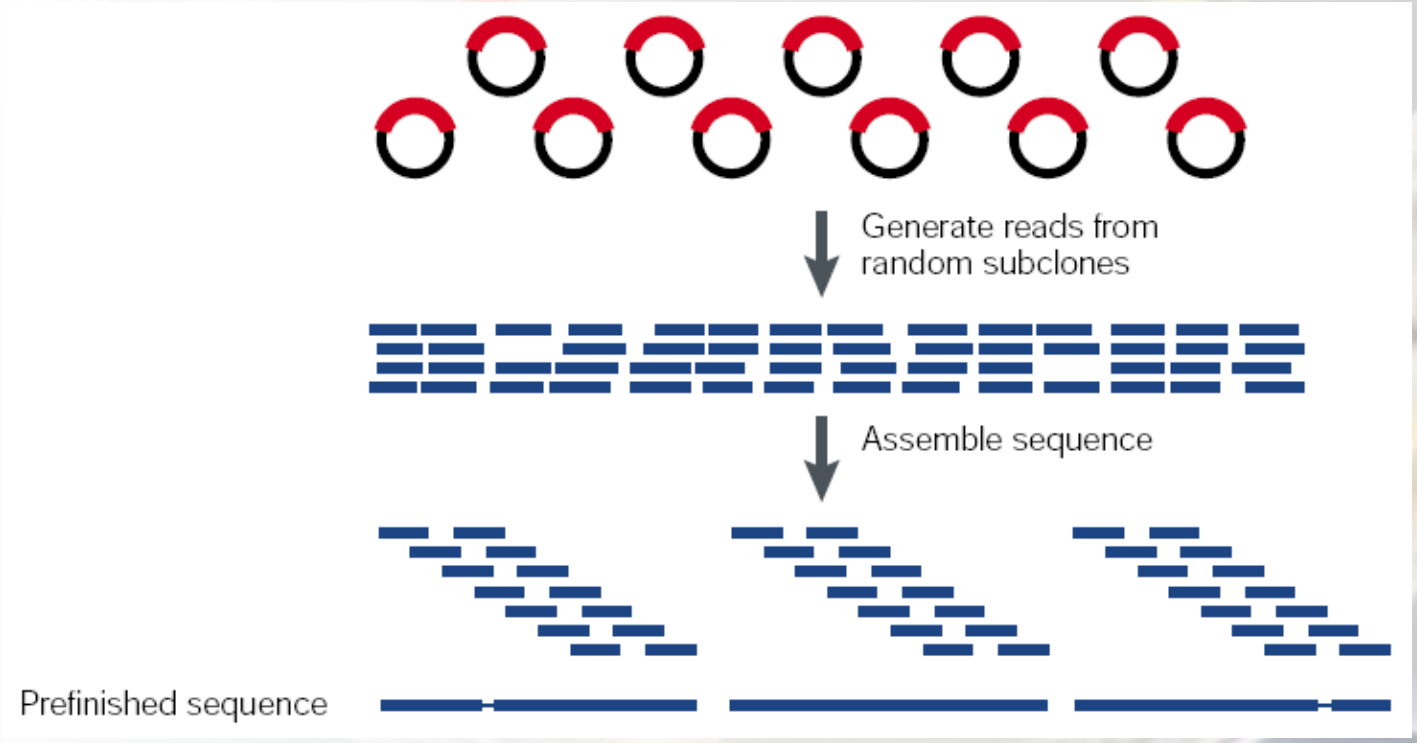
HGP–Hierarchical Shotgun Sequencing

Subclone Library Construction: DNA fragmented by sonication or RE cut; Fragment size ~ 2-5 kb



HGP–Hierarchical Shotgun Sequencing

Random Shotgun Phase: Dideoxy termination reaction; Informatics programs; Coverage and contigs



HGP–Hierarchical Shotgun Sequencing

HGP – Hierarchical Shotgun Sequencing

- Map Construction
- Clone selection
- Subclone library construction
- Random shotgun phase
- Directed finishing phase and sequence authentication

Genetics and Genomics

HGP – Whole Genome Shotgun Sequencing

HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- Whole genome randomly digested three times
- Plasmid library constructed with ~ 2kb inserts and ~10 kb inserts
- BAC library with ~ 200 kb inserts

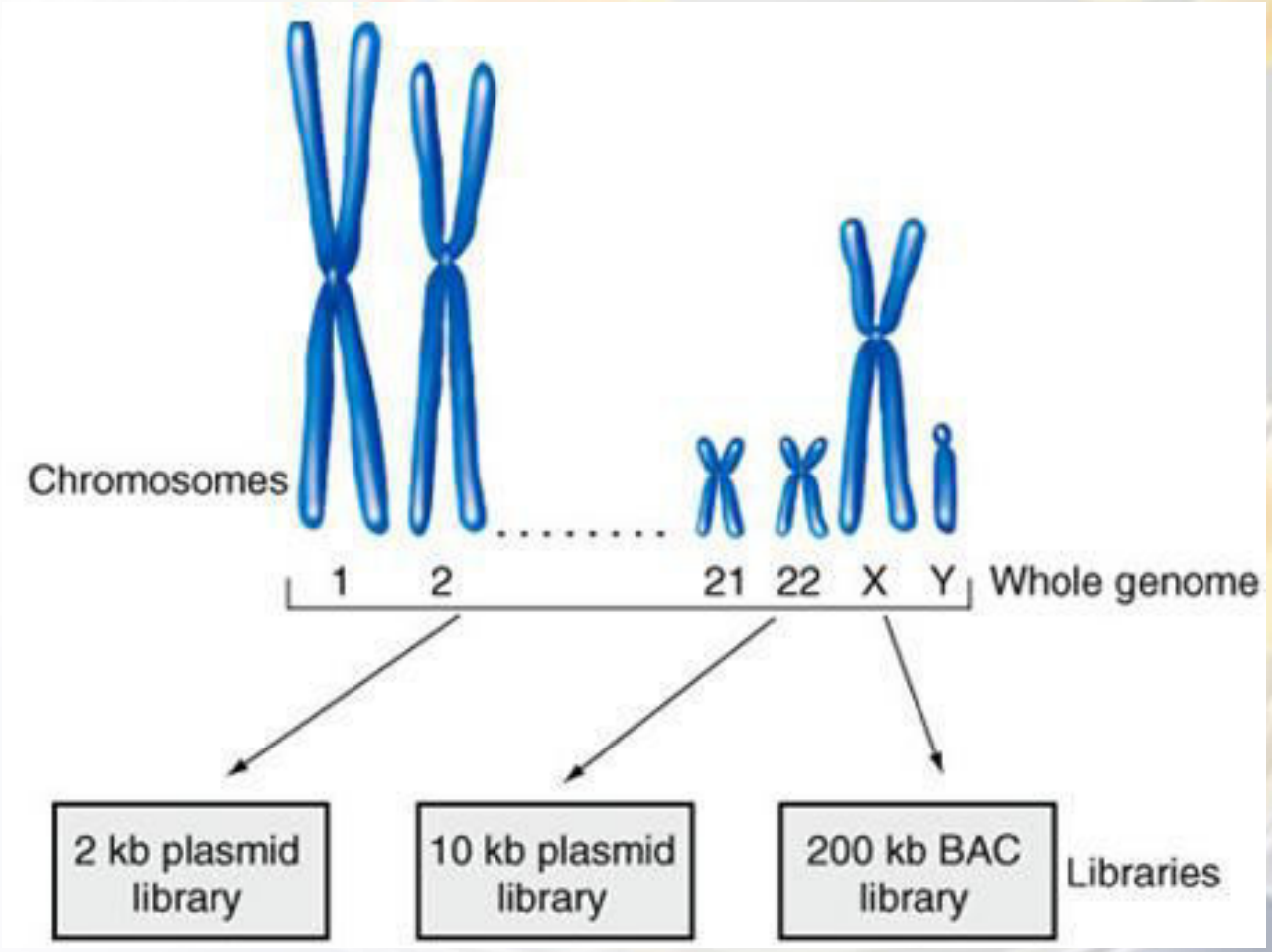
HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- Computer program assembles sequences into chromosomes
- No physical map construction
- Reduces problems of repeat sequences

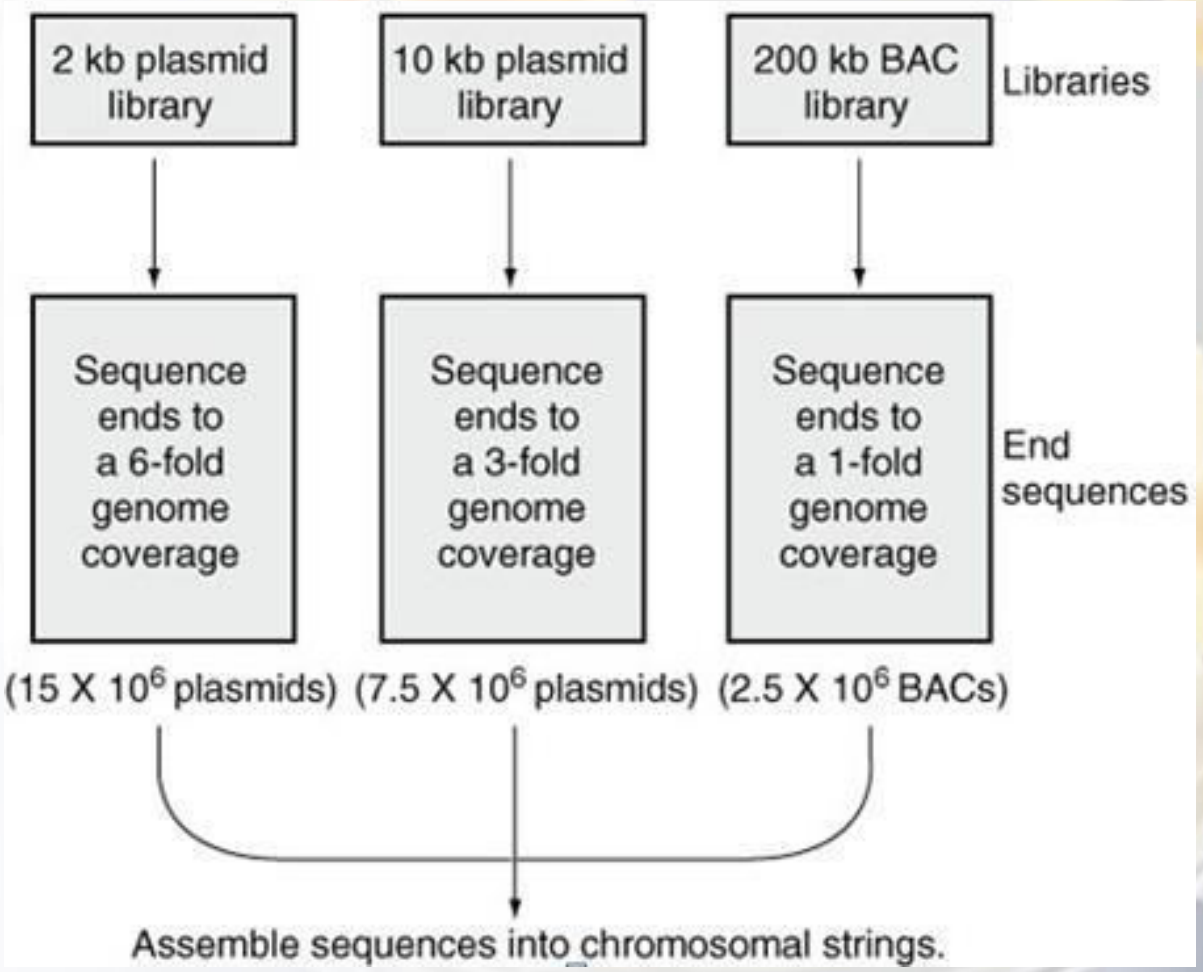
HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing



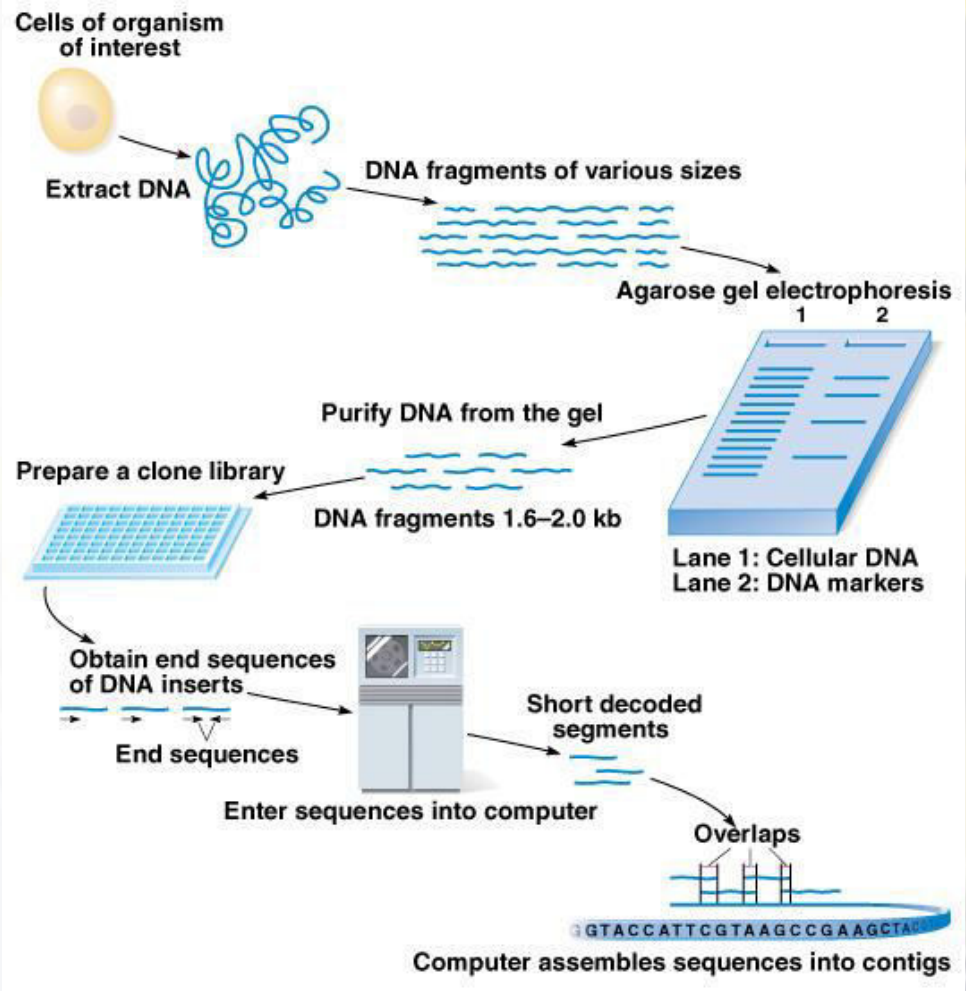
HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing



HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing



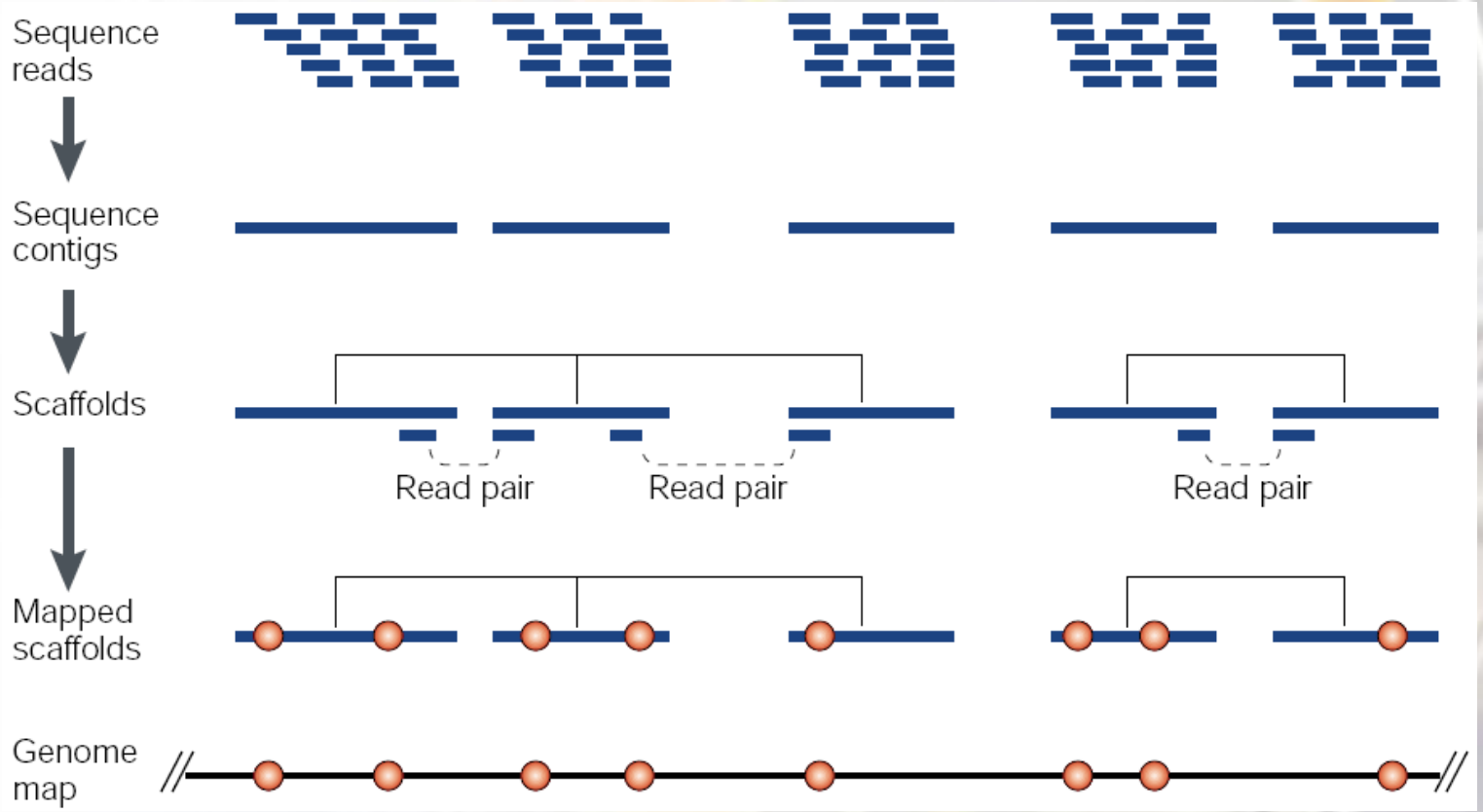
HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- After the sequence is shotgunned, 10 million fragments of the genomic sequences need to be recompiled into the readable base pairs in the proper order.

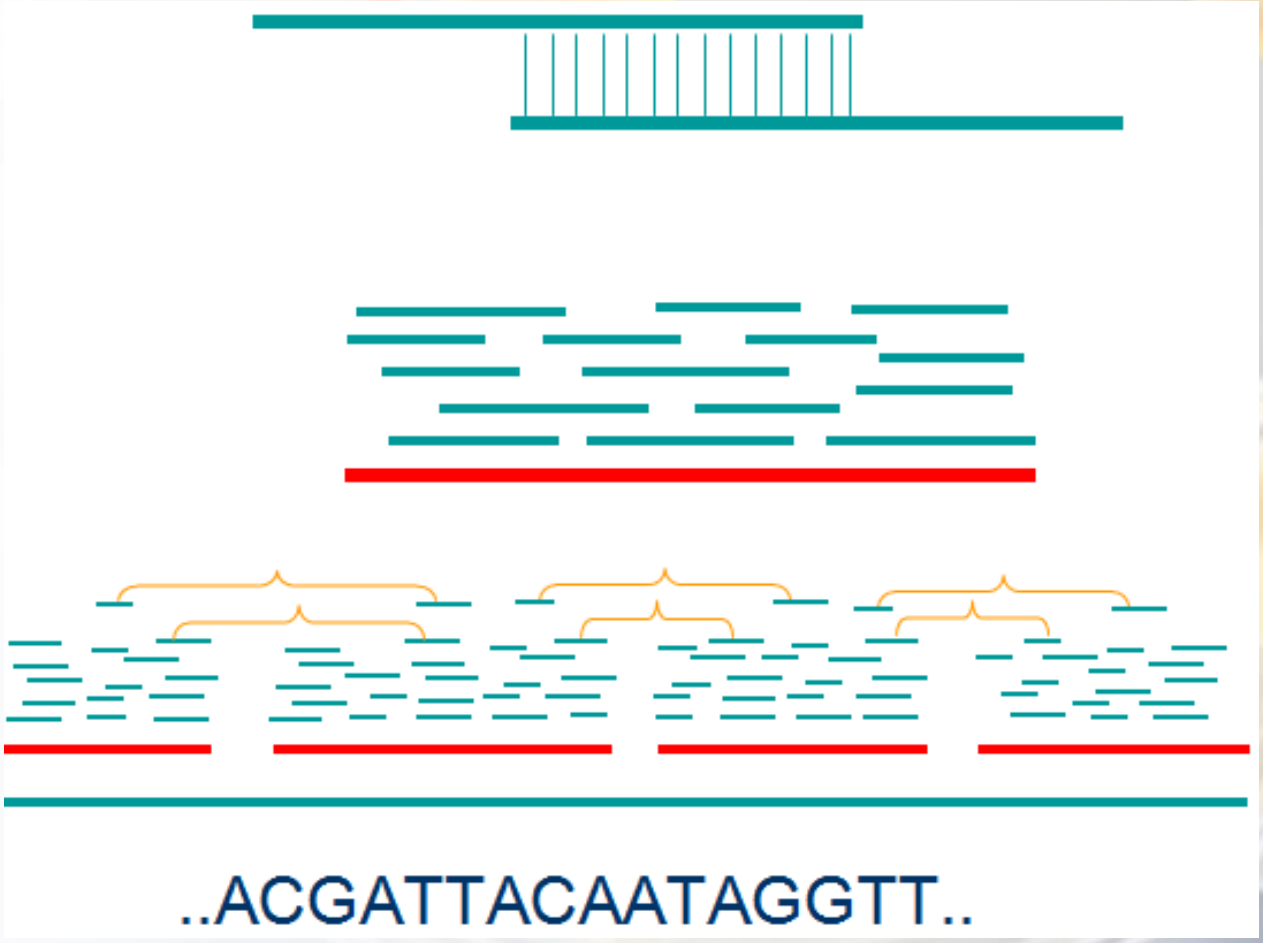
HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing



HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing



HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- The Celera Assembler was one of the core competencies and makes the task possible.

HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- Shotgunned fragments are compared against each other and equivalent sequences greater than 40 base pairs long are identified.
- These 40 base pairs matches are statistically impossible to occur by chance. These matches are then determined to be true or repeat induced.
- True matches are overlapping sections and are the desired fragments

HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- The assembler then searches for overlapping fragments that have a common sequence and are not contested elsewhere in the dataset.
- The uncontested data is assembled into unitigs containing approximately 30 fragments
- These assembled unitigs are 99 % accurate
- Unitigs passing this filter are ready for ordering.

HGP–Whole Genome Shotgun Sequencing

HGP – Whole Genome Shotgun Sequencing

- By looking at these contigs and orientation the scaffold become complete
- Mapped scaffolds were arranged to prepare chromosome maps

Genetics and Genomics

Human Genome - Characteristics (A)

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Human Genome contain ~3 billion bp
- The average gene consists of 3000 bases, but sizes vary
- Largest known human gene dystrophin at 2.4 million bases

Human Genome-Characteristics (A)

Human Genome - Characteristics

- The total number of genes is estimated at around 23,000-much lower than previous estimates of 30,000.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Human genome's gene-dense areas are predominantly composed of the DNA building blocks G and C.

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Gene poor areas are rich in the DNA building blocks A and T.
- GC- and AT-rich regions can be seen through microscope as light and dark bands-chromosomes

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Genes appear to be concentrated in random areas along the genome
- Large areas of noncoding DNA between the genes

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Stretches of up to 30,000 C-G bases repeating over and over often occur adjacent to gene-rich areas.
- CpG islands are believed to regulate gene activity

Human Genome-Characteristics (A)

Human Genome - Characteristics

- Chromosome 1 has the most genes ~ (2100), and the Y chromosome has the fewest ~ (458).

Human Genome-Characteristics (A)

Human Genome - Characteristics

- 3,000 Mb
- ~ 23,000 genes
- Exons ~ 1.5%
- Introns ~ 3.5%
- Repeats ~ 45%

Genetics and Genomics

Human Genome - Characteristics (B)

Human Genome-Characteristics (B)

Human Genome - Characteristics

- ~ 1.5% of the genome codes for proteins.
- Repeated sequences make up at least 45 % of the human genome

Human Genome-Characteristics (B)

Human Genome - Characteristics

- Average size of an exon is 120 -145bp.
- Average number of exons is 7-9.
- Average coding sequence encodes proteins of 370 - 440 amino acids

Human Genome-Characteristics (B)

Human Genome - Characteristics

Chromosome	Length (mm)	Base pairs	Variations	Confirmed proteins	Putative proteins	Pseudogenes	miRNA	rRNA	snRNA	snoRNA	Misc ncRNA
1	85	249,250,621	4,401,091	2,012	31	1,130	134	66	221	145	106
2	83	243,199,373	4,607,702	1,203	50	948	115	40	161	117	93
3	67	198,022,430	3,894,345	1,040	25	719	99	29	138	87	77
4	65	191,154,276	3,673,892	718	39	698	92	24	120	56	71
5	62	180,915,260	3,436,667	849	24	676	83	25	106	61	68
6	58	171,115,067	3,360,890	1,002	39	731	81	26	111	73	67
7	54	159,138,663	3,045,992	866	34	803	90	24	90	76	70
8	50	146,364,022	2,890,692	659	39	568	80	28	86	52	42
9	48	141,213,431	2,581,827	785	15	714	69	19	66	51	55
10	46	135,534,747	2,609,802	745	18	500	64	32	87	56	56
11	46	135,006,516	2,607,254	1,258	48	775	63	24	74	76	53
12	45	133,851,895	2,482,194	1,003	47	582	72	27	106	62	69
13	39	115,169,878	1,814,242	318	8	323	42	16	45	34	36
14	36	107,349,540	1,712,799	601	50	472	92	10	65	97	46
15	35	102,531,392	1,577,346	562	43	473	78	13	63	136	39

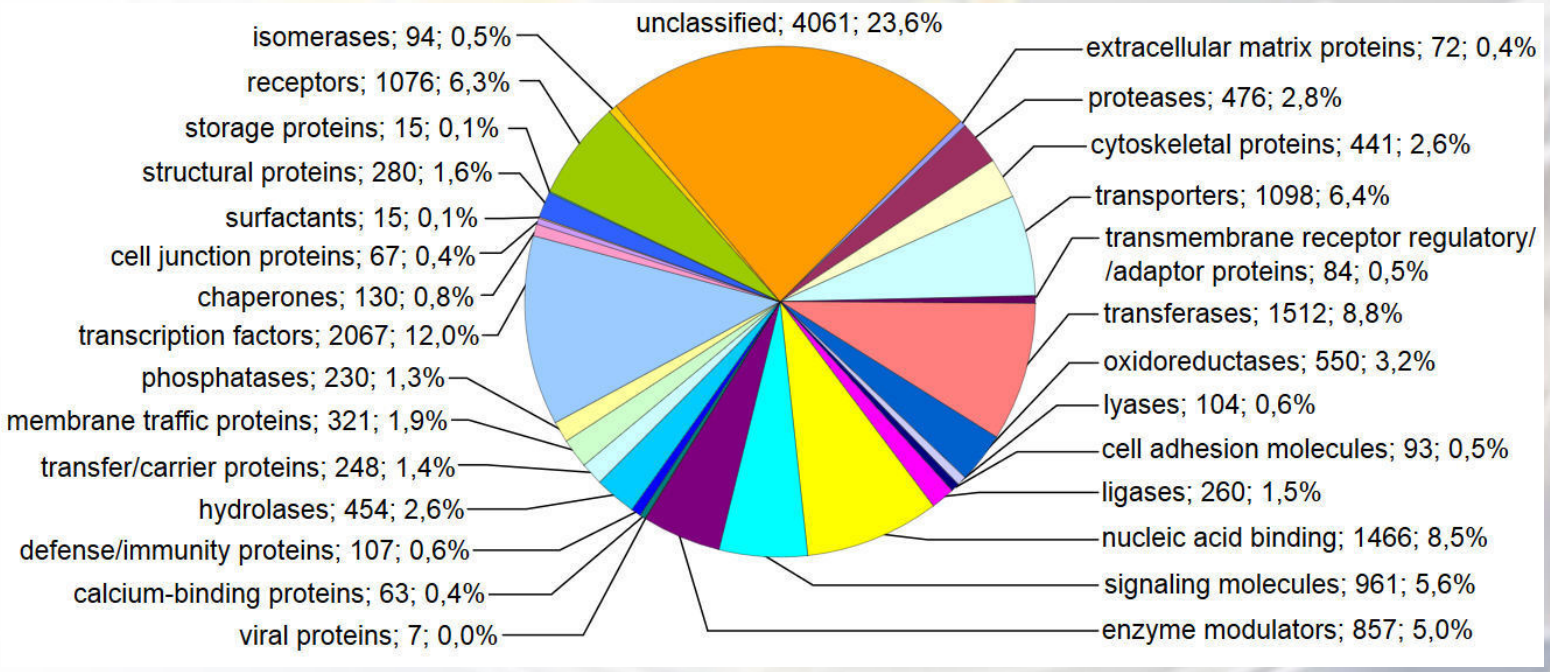
Human Genome-Characteristics (B)

Human Genome - Characteristics

Chromosome	Length (mm)	Base pairs	Variations	Confirmed proteins	Putative proteins	Pseudogenes	miRNA	rRNA	snRNA	snoRNA	Misc ncRNA
16	31	90,354,753	1,747,136	805	65	429	52	32	53	58	34
17	28	81,195,210	1,491,841	1,158	44	300	61	15	80	71	46
18	27	78,077,248	1,448,602	268	20	59	32	13	51	36	25
19	20	59,128,983	1,171,356	1,399	26	181	110	13	29	31	15
20	21	63,025,520	1,206,753	533	13	213	57	15	46	37	34
21	16	48,129,895	787,784	225	8	150	16	5	21	19	8
22	17	51,304,566	745,778	431	21	308	31	5	23	23	23
X	53	155,270,560	2,174,952	815	23	780	128	22	85	64	52
Y	20	59,373,566	286,812	45	8	327	15	7	17	3	2

Human Genome-Characteristics (B)

Human Genome - Human genes categorized by function of the transcribed proteins, given both as number of encoding genes and percentage of all identified genes



Human Genome-Characteristics (B)

Human Genome - Characteristics

- The human genome has many different regulatory sequences which are crucial to controlling gene expression.
- Conservative estimates indicate that these sequences make up 8% of the genome.

Human Genome-Characteristics (B)

Human Genome - Characteristics

- About 8% of the human genome consists of tandem DNA arrays or tandem repeats

Human Genome-Characteristics (B)

Human Genome - Characteristics

- Mobile elements:
- LTR retrotransposons (8.3% of total genome)
- SINEs (13.1% of total genome) including Alu elements
- LINEs (20.4% of total genome)

Human Genome-Characteristics (B)

Human Genome - Characteristics

- Pericentromeric and sub-telomeric regions of chromosomes filled with large transposable elements

Human Genome-Characteristics (B)

Human Genome - Characteristics

- Chimpanzee genome differs from that of the human genome by 1.23% in direct sequence comparisons

Human Genome-Characteristics (B)

Human Genome - Characteristics

- Male mutation rate about twice female
- Most mutations occurs in males
- Recombination rates much higher in distal regions of chromosomes

Genetics and Genomics

Genome Browser - UCSC

Genome Browser - UCSC

Genome Browsers

- The genomes are so large that useful information is hard to find.
- Researchers at UCSC decided to make a computational microscope to help scientists search the genomes.

Genome Browser - UCSC

Genome Browsers

- Researchers can use the UCSC Genome Browser to find information in the human genome and other genomes that have been sequenced

Genome Browser - UCSC

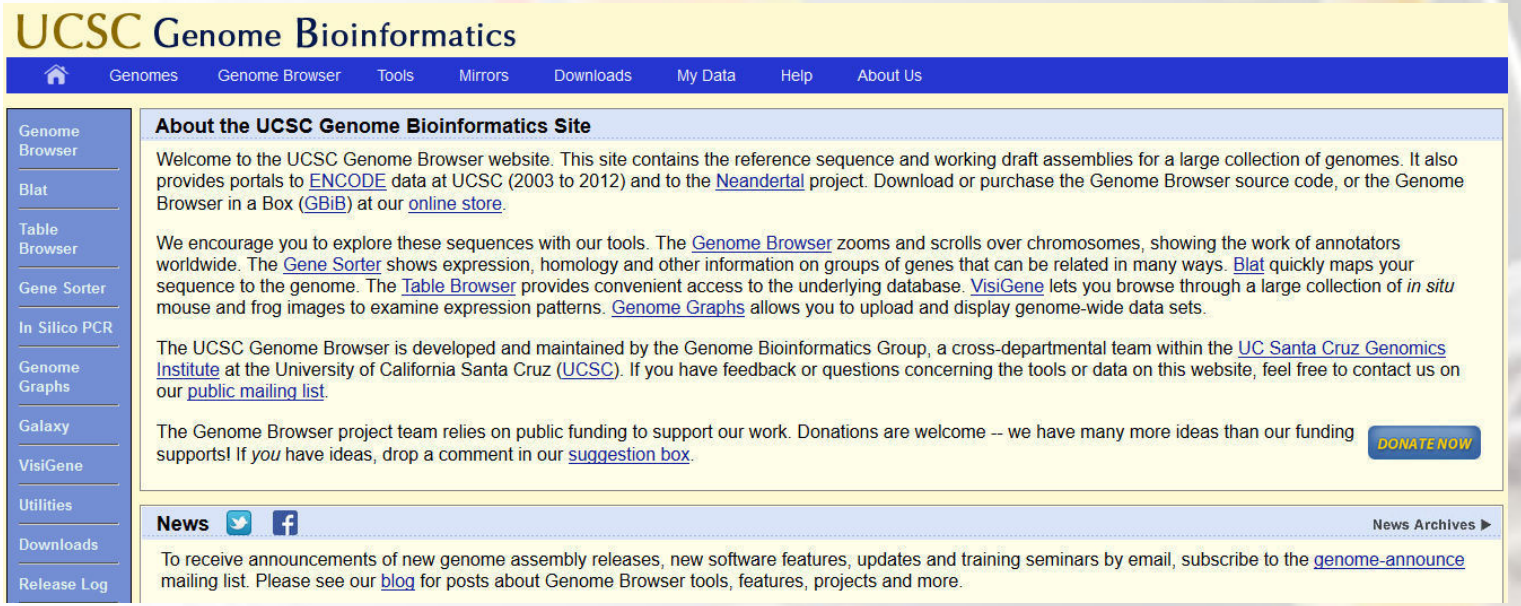
Genome Browsers

- UCSC Genome Browser

  http://genome.ucsc.edu

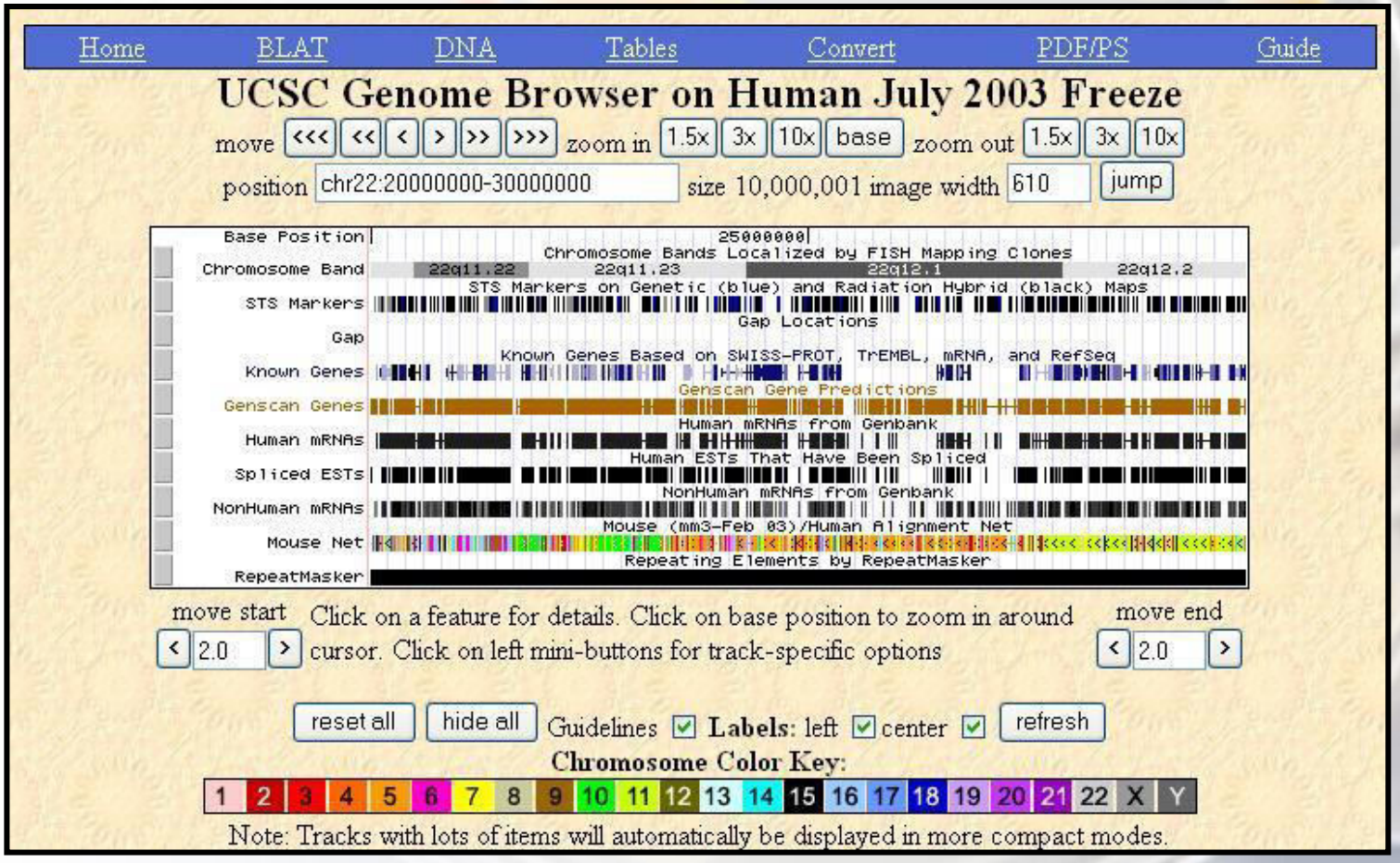
Genome Browser - UCSC

The UCSC Genome Browser - Homepage



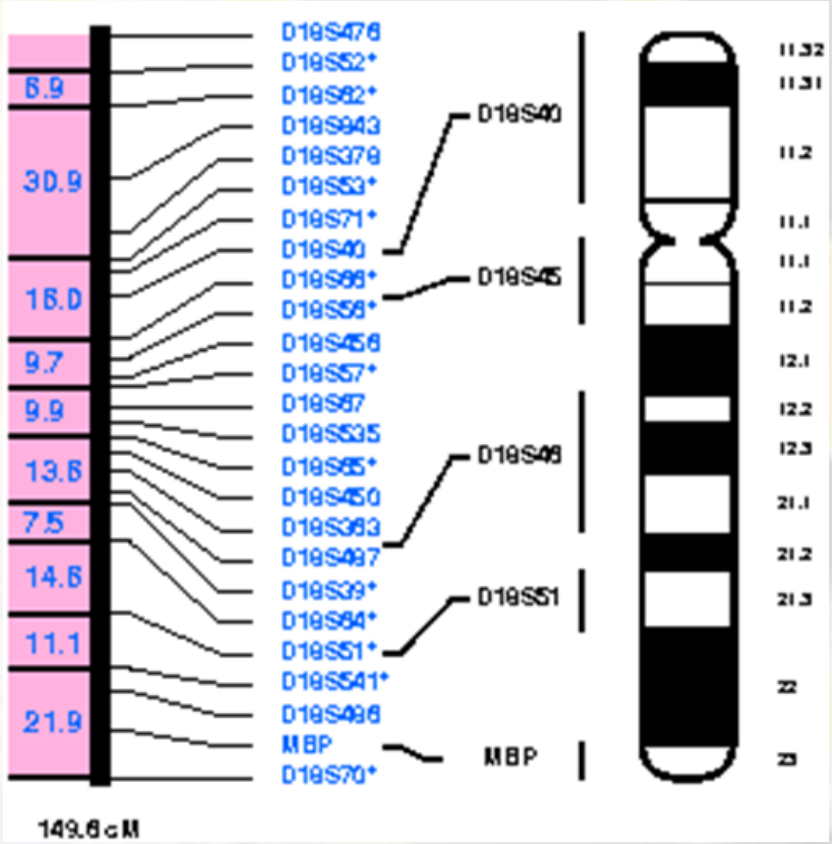
Genome Browser - UCSC

The UCSC Genome Browser



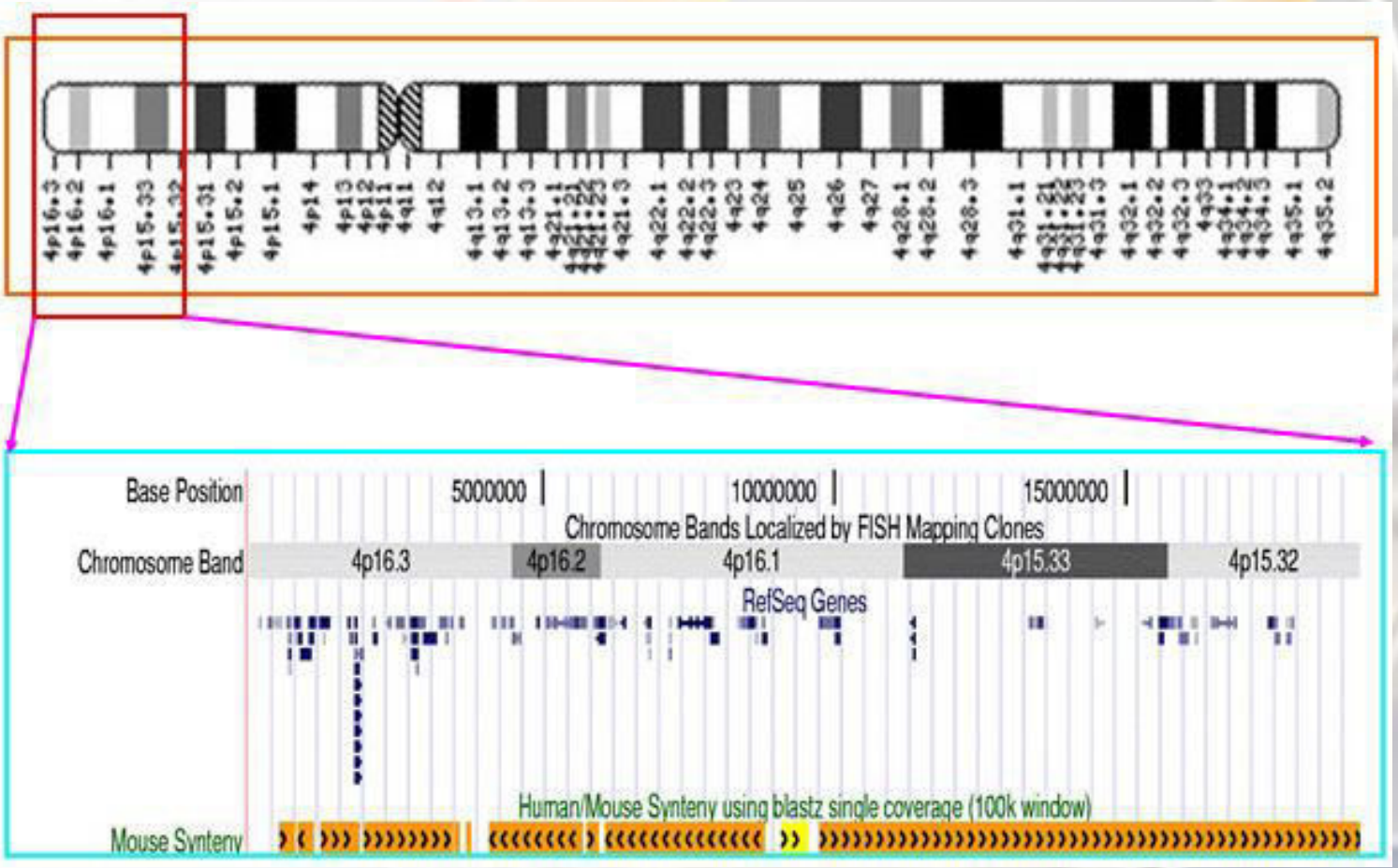
Genome Browser - UCSC

The browser takes you from early maps of the genome . . .



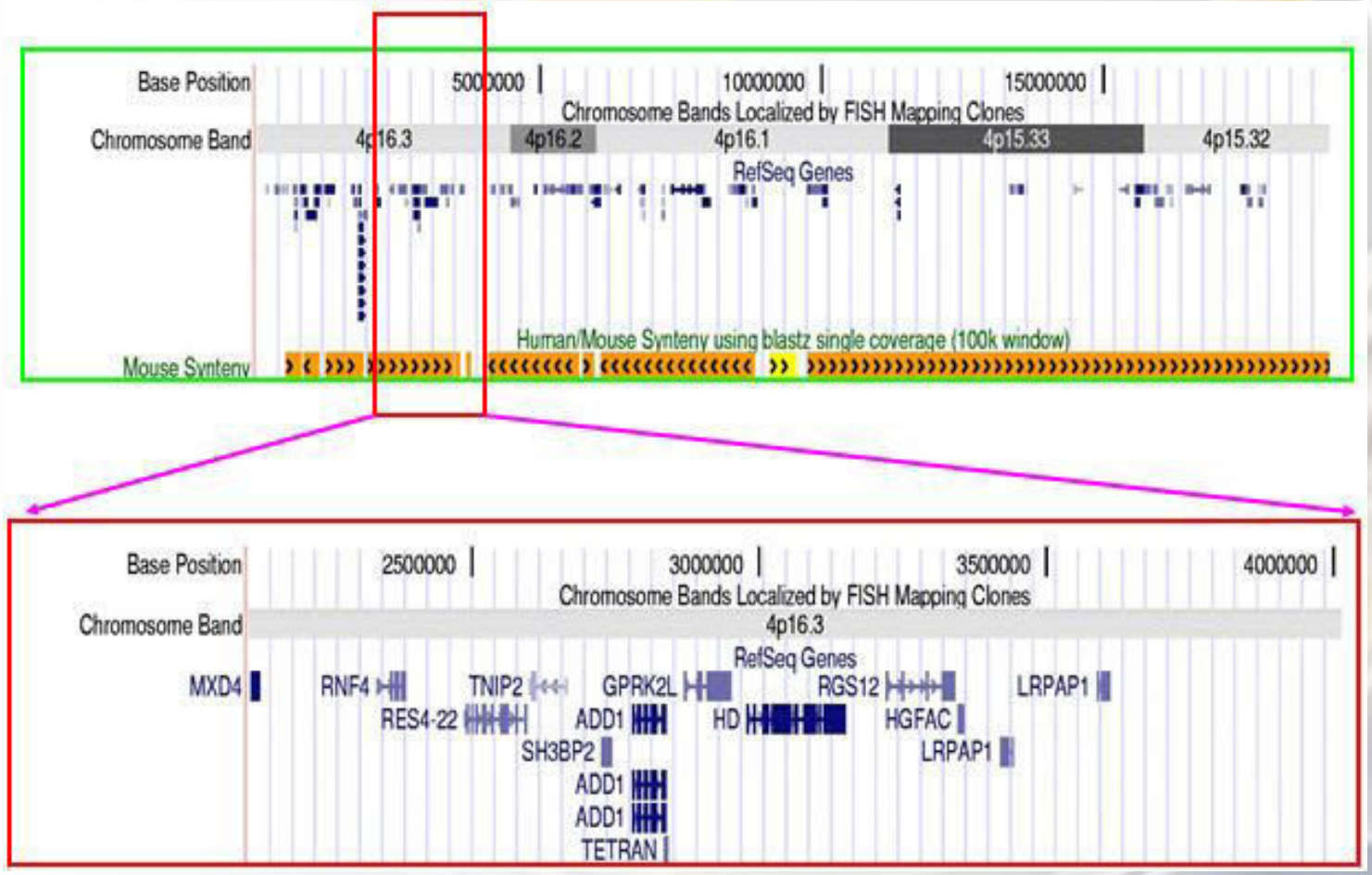
Genome Browser - UCSC

Multi Resolution View



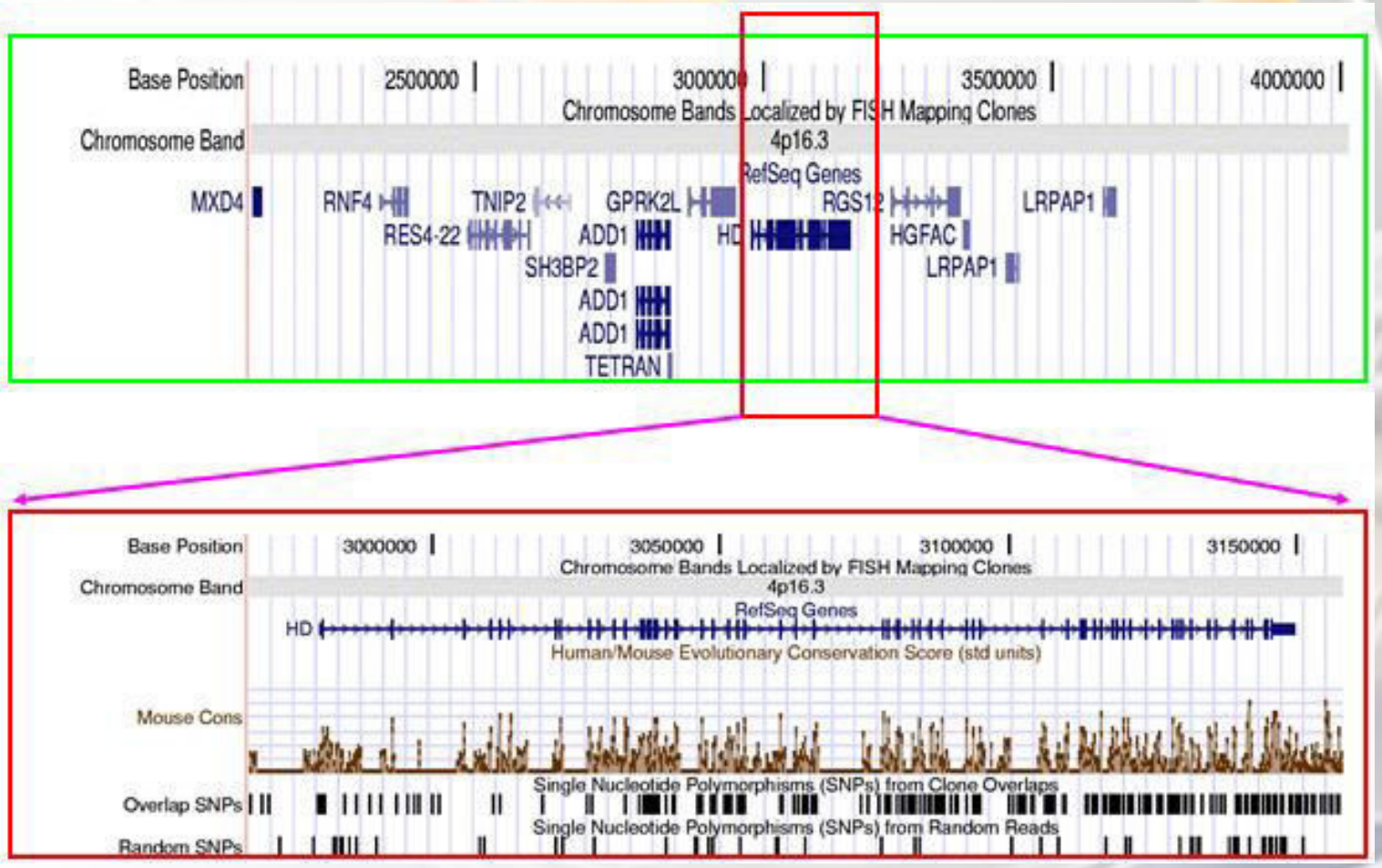
Genome Browser - UCSC

Gene Cluster View



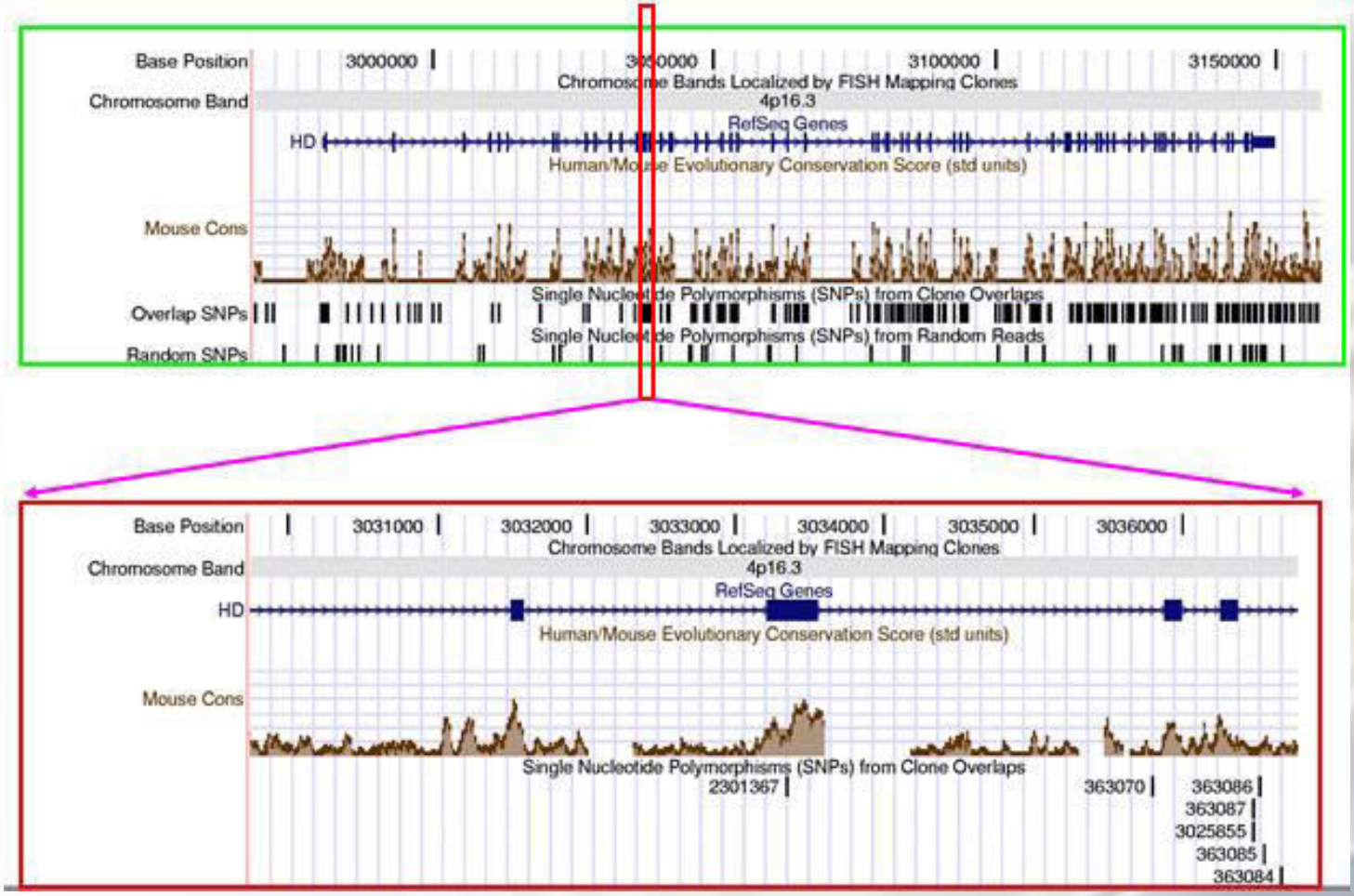
Genome Browser - UCSC

Single Gene View



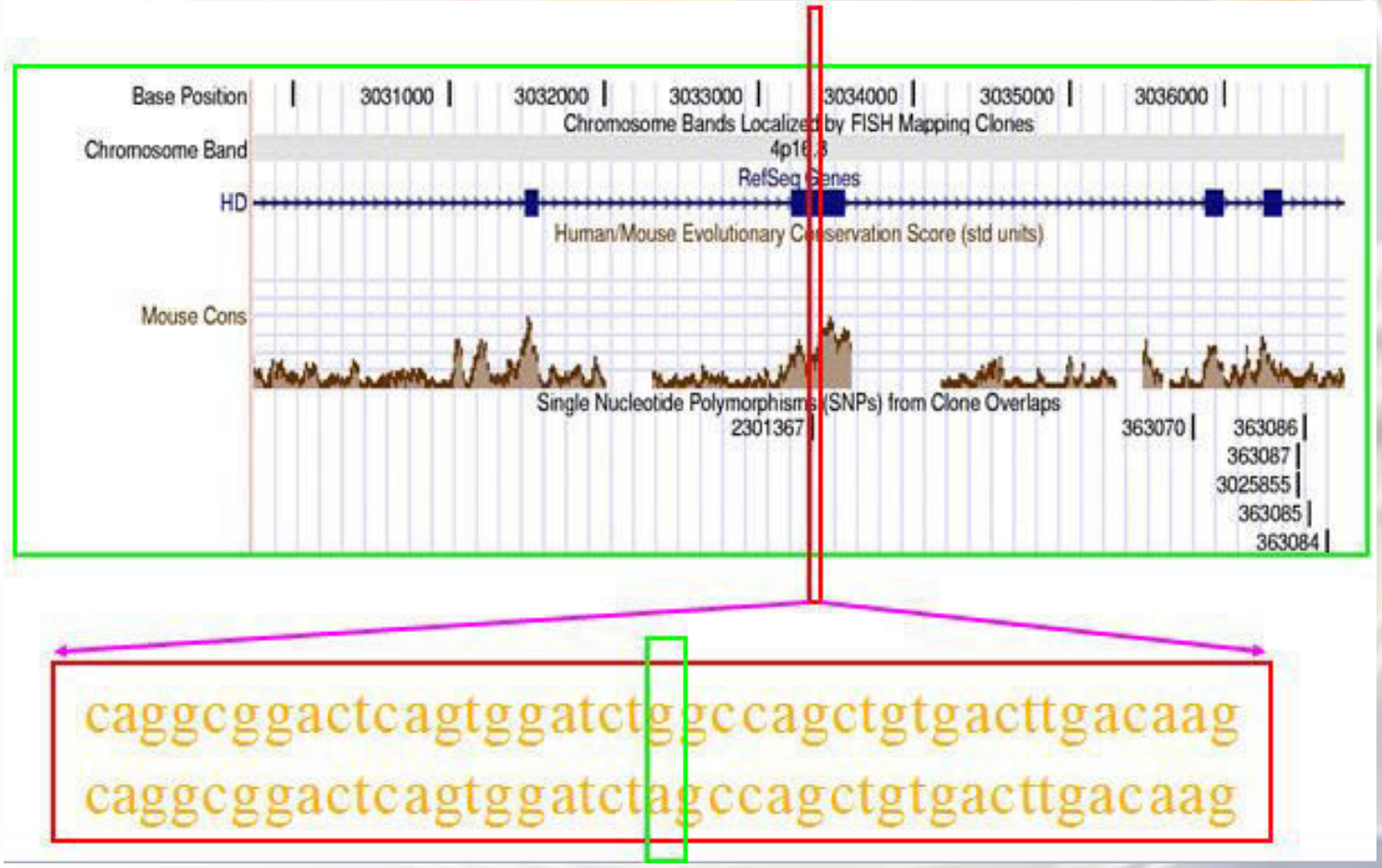
Genome Browser - UCSC

Single Exon View



Genome Browser - UCSC

Single Base/Nucleotide View



Genome Browser - UCSC

Genome Browser - UCSC

- Researchers can use the UCSC genome browser to find information in the human genome and other genomes that have been sequenced

Genetics and Genomics

Genome Browser - Ensembl

Genome Browser - Ensembl

Ensembl Genome Browser

- Ensembl is a joint project between 3 organizations
  - EMBL
  - EBI
  - WTSI

Genome Browser - Ensembl

Ensembl Genome Browser

- Ensembl does not gather any genome project directly
- Works in relation with the sequencing centers that generate the genome assembly

Genome Browser - Ensembl

Ensembl Genome Browser

- 4 Main Databases
  - Ensembl Core Database
  - Ensembl EST Database
  - Ensembl Compara Database
  - Ensembl Variation Database

Genome Browser - Ensembl

Ensembl

- Ensembl core databases store genome sequence and annotation information
- Gene, transcript, and protein models
- Databases stores information about cDNA and protein alignments

Genome Browser - Ensembl

Ensembl Genome Browser

- Within one genome: regulatory elements, gene order etc
- Comparative studies: Evolution, conserved regions rearrangements
- Gene quality and prediction

Genome Browser - Ensembl

Ensembl Genome Browser

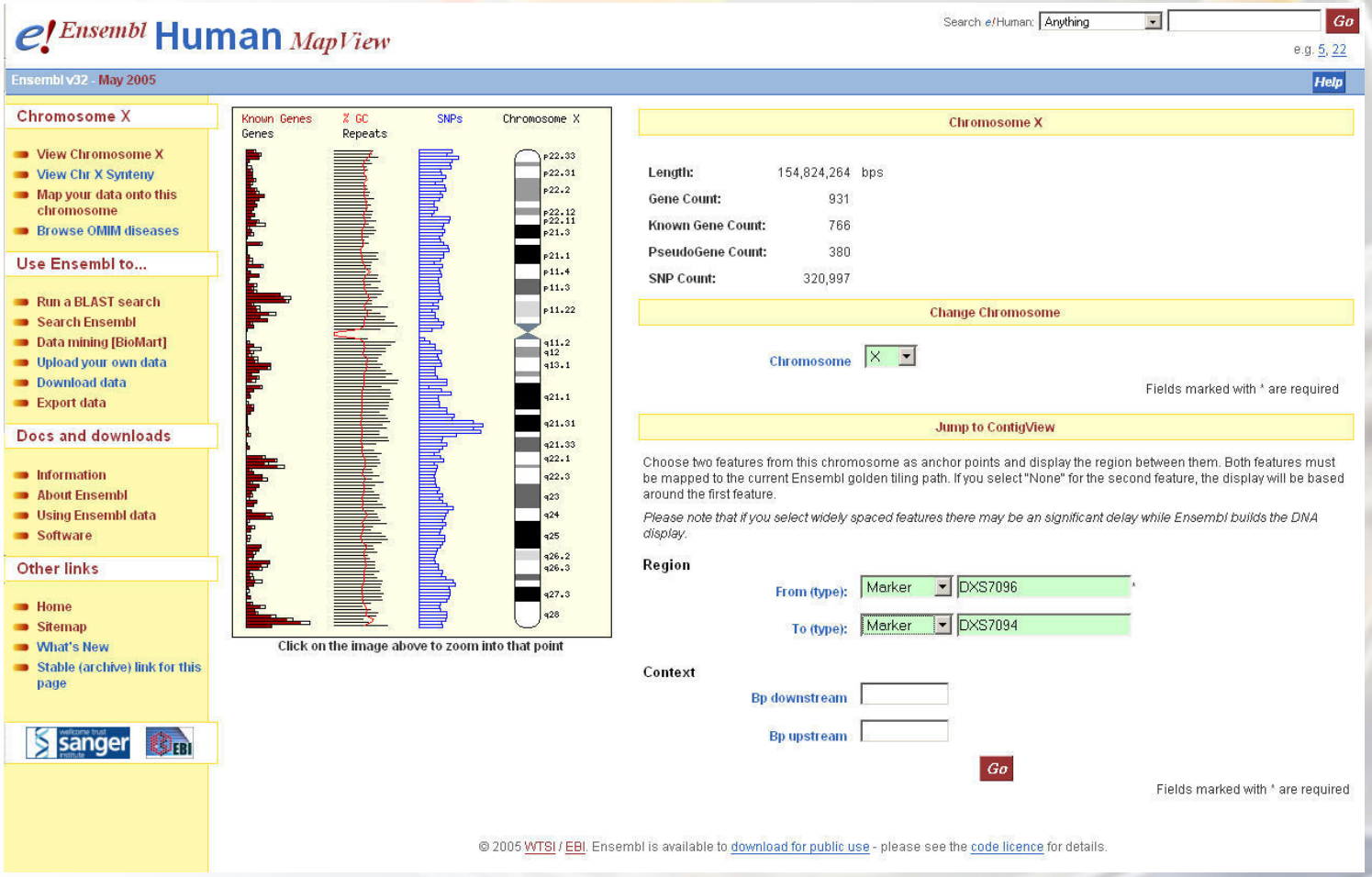
- Ensembl Genome browser
  http://www.ensembl.org

Genome Browser - Ensembl

Ensembl Genome Browser – Home page

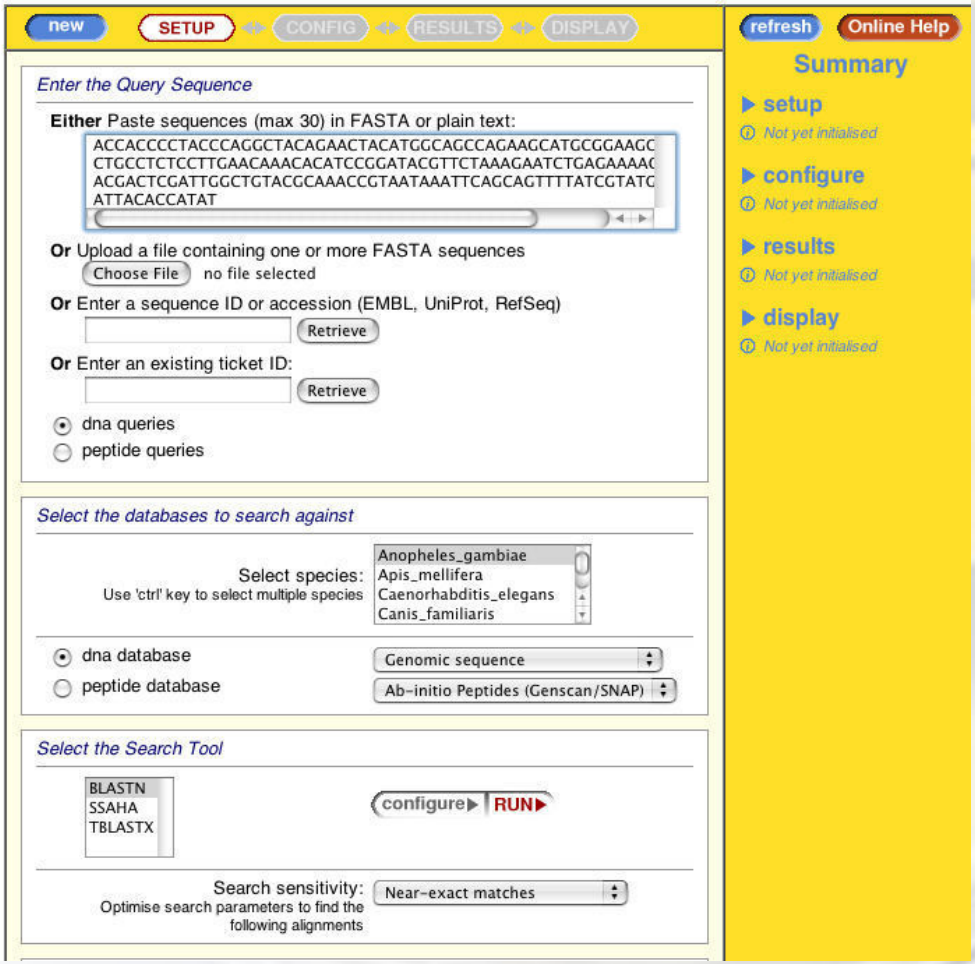
Genome Browser - Ensembl

Ensembl Genome Browser – Map View



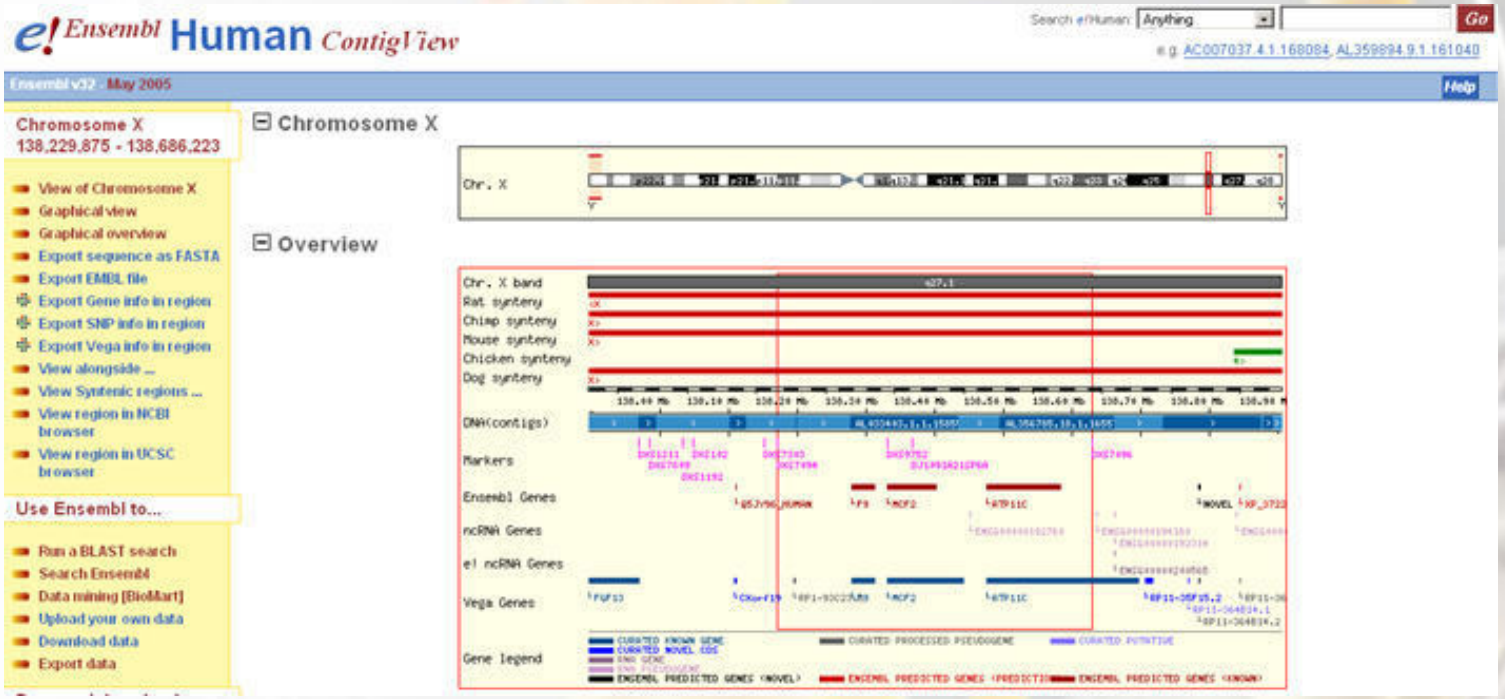
Genome Browser - Ensembl

Ensembl Genome Browser – Blast



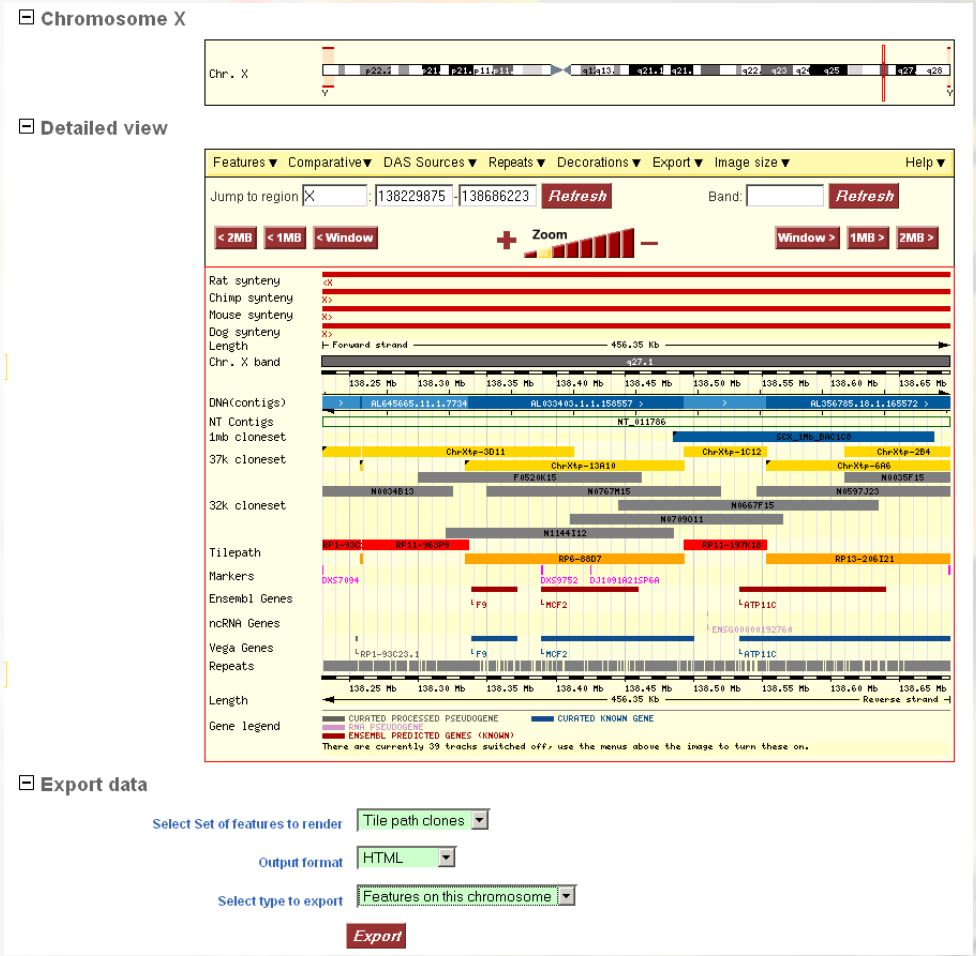
Genome Browser - Ensembl

Ensembl Genome Browser – Contig View



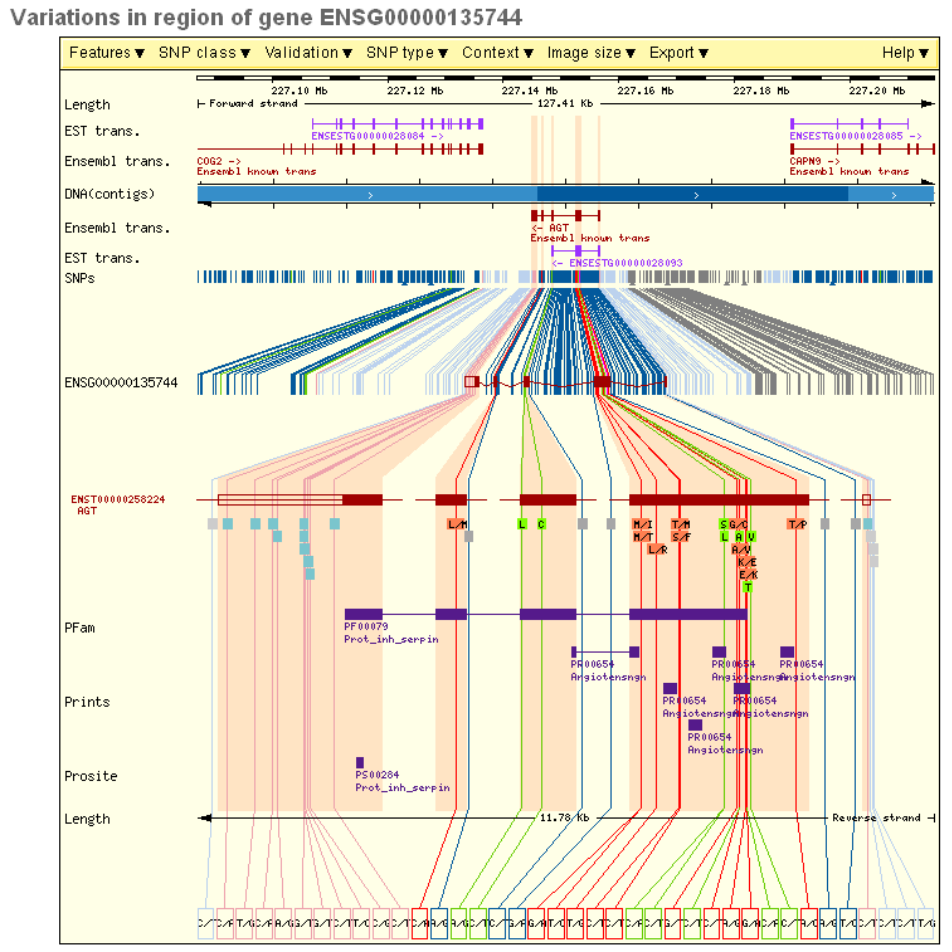
Genome Browser - Ensembl

Ensembl Genome Browser – Cyto View



Genome Browser - Ensembl

Ensembl Genome Browser – SNP View



Genome Browser - Ensembl

Ensembl Genome Browser – Marker View

Chromosome Map Marker DXS9752

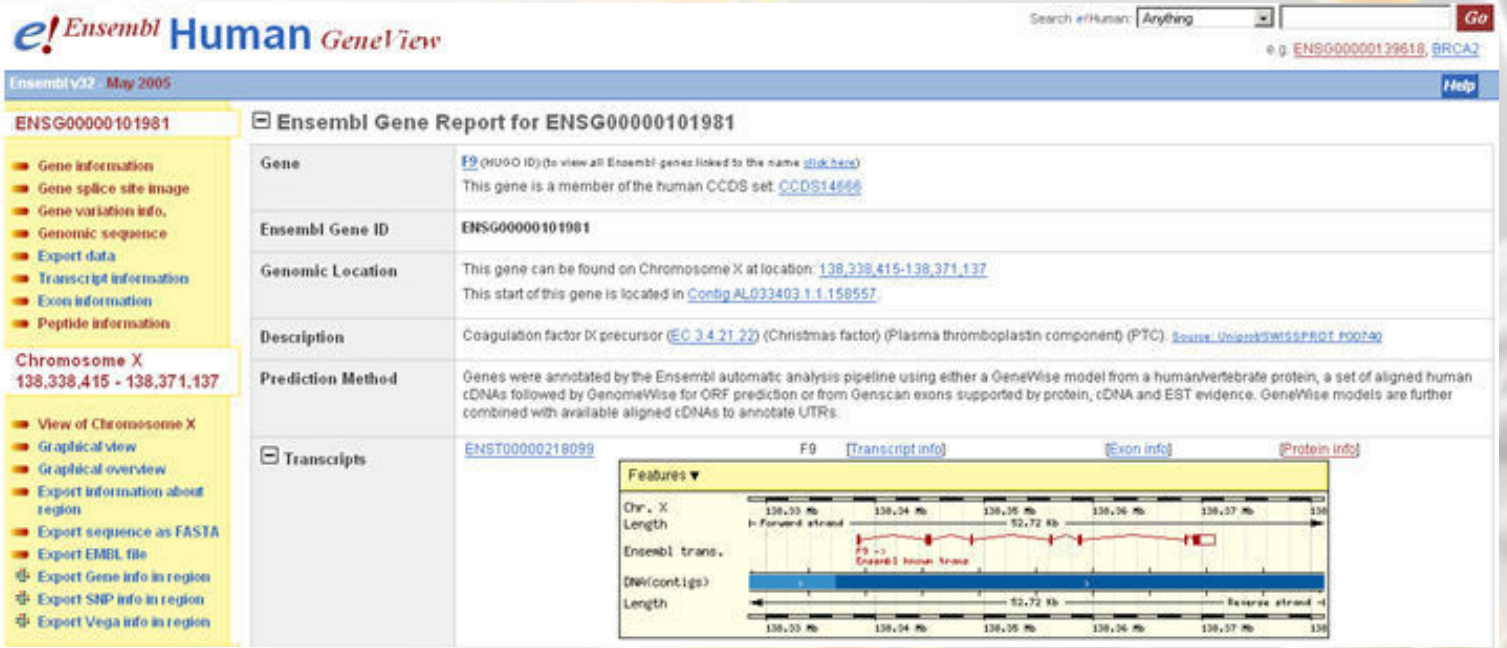
Marker Source	82913 (database: unists)
Marker Location	Basepairs 138389786 - 138390044 on chromosome X [Export data]
Marker Synonyms	**Gdb :** GDB:737728 GDB:738733 **Genbank :** G13636 **Other :** SHGC-11927 DXS9752 RH8108
Marker Primers	**Expected Product Size** 259 **Left Primer** TTTTCAGGTTAATGGACACGC **Right Primer** CCATTTTGCAGCCGTAATTT

Marker DXS9752 map locations

Map Name	Synonym	Chromosome	Position	LOD Score
gm99g3	RH8108	X	4259	3.5

Genome Browser - Ensembl

Ensembl Genome Browsers – Gene View



Genome Browser - Ensembl

Ensembl Genome Browsers – Exons View

Ensembl Exon Report

Transcript	**F9** (HUGO ID) (to view all Ensembl genes linked to the name click here) This transcript is a member of the human CCDS set: CCDS14666
Ensembl Transcript ID	**ENST00000218099**
Transcript information	**Exons:** 8 **Transcript length:** 2,802 bps **Translation length:** 461 residues This transcript is a product of gene: ENSG00000101981
Genomic Location	This transcript can be found on Chromosome X at location: 138,338,415-138,371,137 This start of this transcript is located in Contig AL033403.1.1.158557.
Description	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (Plasma thromboplastin component) (PTC). Source: UniprotSWISSPROT P00740

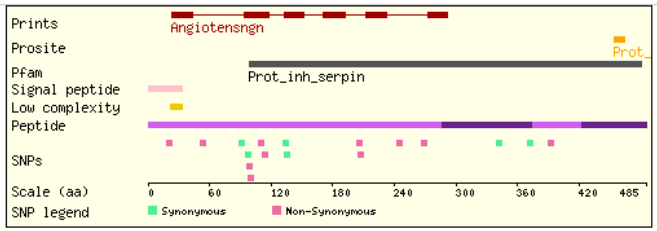
Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Length	Sequence
	5' upstream sequence						..........ccagaagtaaatacagctcagcttgtactttggtacaactaatcgacctt
1	ENSE00000677288	X	1	138,338,415	138,338,531	117 bp	ACCACTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAAT CACCAGGCCTCATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAG
	Intron 1-2	X	1	138,338,532	138,344,688	6,157 bp	gtttgtttccttttttaaaatacat.....aaagaattattcttttacatttcag
2	ENSE00000677287	X	1	138,344,689	138,344,852	164 bp	TTTTTCTTGATCATGAAAACGCCAACAAAATTCTGAATCGGCCAAAGAGGTATAATTCAG GTAAATTGGAAGAGTTTGTTCAAGGGAACCTTGAGAGAGAATGTATGGAAGAAAAGTGTA GTTTTGAAGAAGCACGAGAAGTTTTTGAAAACACTGAAAGAACA
	Intron 2-3	X	1	138,344,853	138,345,040	188 bp	gtgagtatttccacataataccctt.....atttgtcttctttttattctttatag
3	ENSE00001173335	X	1	138,345,041	138,345,065	25 bp	ACTGAATTTTGGAAGCAGTATGTTG
	Intron 3-4	X	1	138,345,066	138,348,754	3,689 bp	gtaagcaattcattttatcctctag.....ttcaatttcttaacctatctcaaag
4	ENSE00001173327	X	1	138,348,755	138,348,868	114 bp	ATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGTTGCAAGGATGACATTA ATTCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGAACTGTGAATTAG
	Intron 4-5	X	1	138,348,869	138,356,041	7,173 bp	gtaagtaactatttttgaatactc.....ttactgtctattttgcttctttttag
5	ENSE00001173320	X	1	138,356,042	138,356,170	129 bp	ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATA ACAAGGTGGTTTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAAGTCCTGTG AACCAGCAG
	Intron 5-6	X	1	138,356,171	138,358,740	2,570 bp	gtcataatctgaataagatttttta.....actaatttttcttctatttttctag
6	ENSE00001173315	X	1	138,358,741	138,358,943	203 bp	TGCCATTTCCATGTGGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGA CTGTTTTTCCTGATGTGGACTATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACA

Genome Browser - Ensembl

Ensembl Genome Browser – Proteins View

Ensembl Protein Report

Peptide	**AGT** (HUGO ID) (to view all Ensembl genes linked to the name click here) This peptide is a member of the human CCDS set: CCDS1585
Ensembl Peptide ID	**ENSP00000258224**
Translation information	This peptide is a translation of transcript: ENST00000258224, which is a product of gene: ENSG00000135744 .
Genomic Location	This peptide can be found on Chromosome 1 at location: 227,145,622-227,153,331 This start of this peptide is located in Contig AL158214.33.1.181175.
Description	Angiotensinogen precursor [Contains: Angiotensin I (Ang I); Angiotensin II (Ang II); Angiotensin III (Ang III) (Des-Asp[1]- angiotensin II)]. Source: Uniprot/SWISSPROT P01019
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomeWise for ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
InterPro	IPR000227 Angiotensinogen - [View other genes with this domain] IPR000215 Proteinase inhibitor I4, serpin - [View other genes with this domain]
Protein Family	ENSF00000005553 : ANGIOTENSINOGEN PRECURSOR [CONTAINS: ANGIOTENSIN I ANG I ; ANGIOTENSIN II ANG II ; ANGIOTENSIN III ANG III DES ASP[1] ANGIOTENSIN II] This cluster contains 1 Ensembl gene member(s)
Protein Features	
Peptide Sequence	MRKRAPQSEMAPAGVSLRATILCLLAWAGLAAGDRVYIHPFHLVIHNESTCEQLAKANAG KPKDPTFIPAPIQAKTSPVDEKALQDQLVLVAAKLDTEDKLRAAMVGMLANFLGFRIYGM HSELWGVVHGATVLSPTAVFGTLASLYLGALDHTADRLQAILGVPWKDKNCTSRLDAHKV LSALQAVQGLLVAQGRADSQAQLLLSTVVGVFTAPGLHLKQPFVQGLALYTPVVLPRSLD FTELDVAAEKIDRFMQAVTGWKTGCSLMGASVDSTLAFNTYVHFQGKMKGFSLLAEPQEF WVDNSTSVSVPMLSGMGTFQHWSDIQDNFSVTQVPFTESACLLLIQPHYASDLDKVEGLT FQQNSLNWMKKLSPRTIHLTMPQLVLQGSYDLQDLLAQAELPAILHTELNLQKLSNDRIR VGEVLNSIFFELEADEREPTESTQQLNKPEVLEVTLNRPFLFAVYDQSATALHFLGRVAN PLSTA

Genome Browser - Ensembl

Ensembl Genome Browser

- Homo Sapiens - Human
- Pan troglodytes- Chimpanzee
- Macaca mulatta - Rhesus monkey
- Mus musculus - Mouse
- Rattas norvegicus - Rat
- Canis familiaris - Dog

Genome Browser - Ensembl

Ensembl

- All data generated is free for download
- Includes genes sequence, transcript, protein predictions
- Ensembl provides a dedicated Export View page, can be exported into HTML, text or zipped format

Genetics and Genomics

HGP – Ethical, Legal and Social Issues

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- The ability to identify human genes raises complex ethical issues involving
- The right to information about one's own genome

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Right to genetic information by employers, insurance companies and government agencies, and concerns about the ability to diagnose but not treat genetic disorders

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Privacy and confidentiality of genetic information
- Fairness in the use of genetic information by insurers, employers, courts, schools, adoption agencies

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Psychological impact and discrimination due to an individual's genetic differences.

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Reproductive issues including adequate and informed consent and use of genetic information in reproductive decision making.

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Clinical issues including the education of doctors and other health-service providers

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- People identified with genetic conditions, and the general public about capabilities, limitations, and social risks.

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Uncertainties associated with gene tests for susceptibilities and complex conditions (e.g., heart disease, diabetes, and alzheimer's disease).

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Fairness in access to advanced genomic technologies.

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Health and environmental issues concerning genetically modified (GM) foods and microbes.

HGP – Ethical, Legal, Social Issues

HGP- Ethical, Legal and Social Issues

- Commercialization of products including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

Genetics and Genomics

Genome Annotations – Proteins Coding Genes

Genome Annotations-Proteins Coding Genes

Genome Annotations

- To make use of the genome sequences, all components of genomes need to understand.
- Assigning identities and functions to sequences within the genome is called genome annotation.

Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Genes transcribed, which means that we can identify them
- Traditionally cDNA /EST sequencing
- More recently by microarray

Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Structural Annotation – Finding genes and other biologically relevant sites

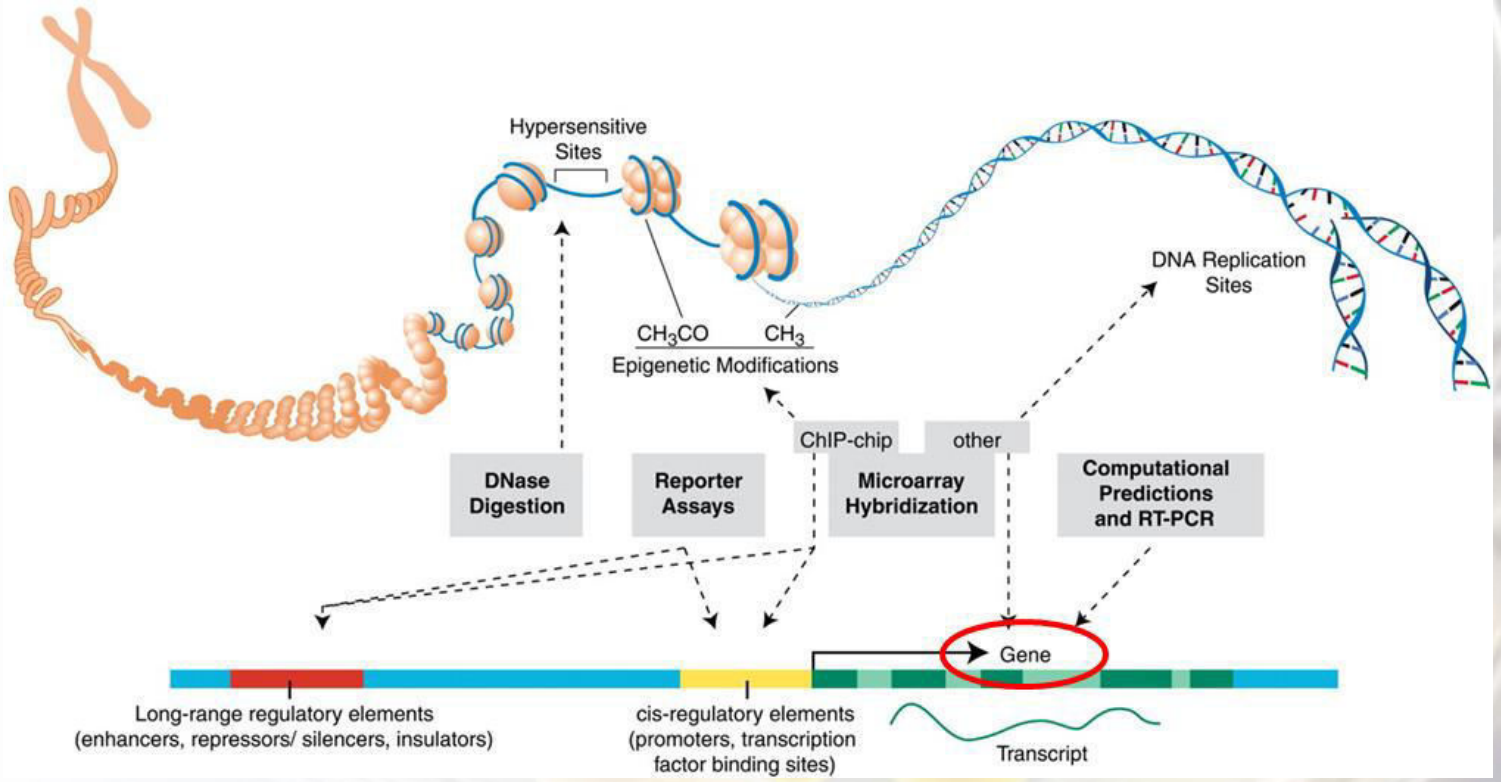
Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Annotation/function can be mapped to different levels:
- Organism, cellular, molecular level

Genome Annotations-Proteins Coding Genes

Genome Annotations



Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Genes can be identified *in vitro* using computational methods
- Protein-coding genes have recognizable features

Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Open reading frames
- Codon bias
- Known transcription and translational start and stop motifs
- Splice consensus sequences at intron-exon boundaries

Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Protein-coding genes recognizable feature
- Software to scan the genome and identify these features
- Some of these programs work quite well, in bacteria and simple eukaryotes

Genome Annotations-Proteins Coding Genes

Genome Annotations-Protein Coding Genes

- Harder for the higher eukaryotes where there are a lot of long introns, genes can be found within introns of other genes

Genome Annotations-Proteins Coding Genes

Validating Predictions & Refining Gene Models

- Validation of predictions include;
- Match to previously annotated cDNA
- Match to EST from same organism
- Similarity of nucleotide or translated protein sequence to GenBank
- Protein structure prediction match - PFAM domain
- Associated with recognized promoter sequences, i.e TATA box, CpG island

Genome Annotations-Proteins Coding Genes

Finding Non-Protein-Coding Genes

- Harder to find
- No poly-A tailed—
- No ORF
- Rely on sequence divergence at nucleotide not protein level, so homology is harder to detect

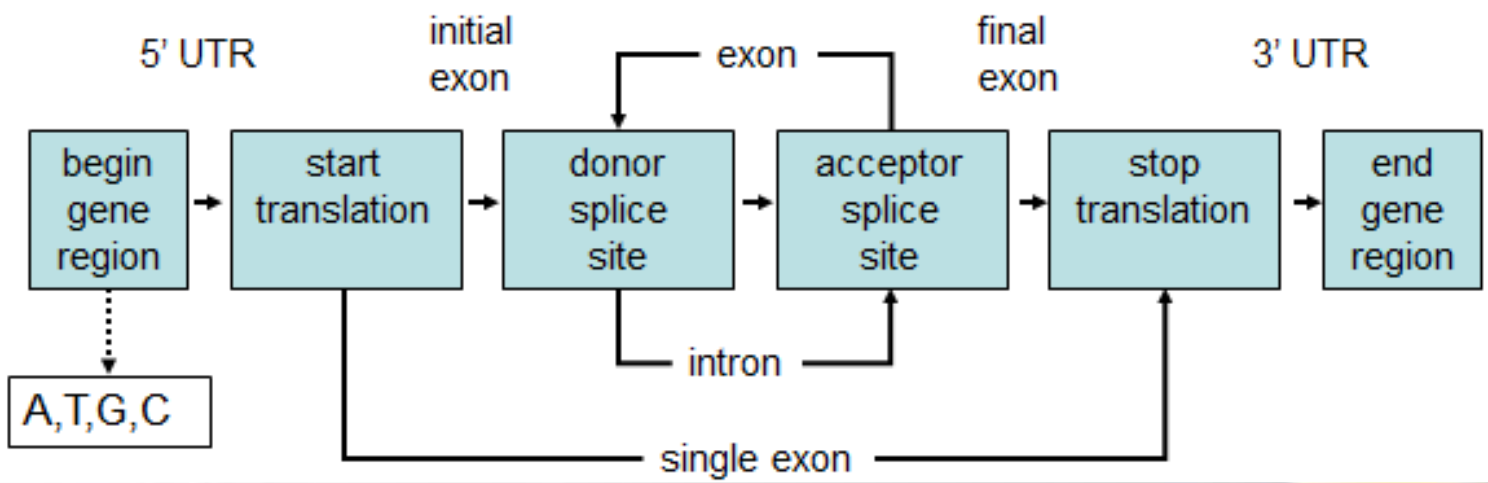
Genome Annotations-Proteins Coding Genes

Gene Discovery Programs

- Most gene-discovery programs makes use of machine learning algorithm.
- Two approaches;
- Artificial Neural Networks and Hidden Markov Models.

Genome Annotations-Proteins Coding Genes

Gene Discovery Programs- Diagram for an HMM for Gene Discovery



Examples of known gene models; provide the probability that a stretch of sequence is a gene.

Genome Annotations-Proteins Coding Genes

Genome Annotations

- Assigning identities and functions to sequences within the genome is called genome annotation.

Genetics and Genomics

Genome Annotations – Transcription Factors

Genome Annotations-Transcription Factors

Genome Annotations

- Other than protein coding genes
- Structural sequences
- Regulatory sequences
- Non-functional junk
- Annotate regulatory sequences such as transcription factor binding sites (Cis-)

Genome Annotations-Transcription Factors

Genome Annotations

- Transcription factors (TFs) are proteins that bind to the DNA and help to control gene expression.
- Sequences to which TF bind called transcription factor binding sites (Cis-acting elements)

Genome Annotations-Transcription Factors

Genome Annotations

- Transcription factors bind to specific DNA sequences
- Most transcription factors can bind to a range of similar sequences.

Genome Annotations-Transcription Factors

Genome Annotations

- Once we know the binding site, we can search the genome to find all of the (predicted) binding sites.

Genome Annotations-Transcription Factors

Genome Annotations

7 characterized binding sites for certain transcription factor

Consensus Sequence

Frequency matrix and its graphical depiction, a sequence logo

TCCGGAAGC
TCCGGATGC
TCCGGATCT
CATGGATGC
CCAGGAAGT
GGTGGATGC
ACCGGATGC

$^{T}_{C}C^{C}_{T}$GGAAGC

A	1	1	1	0	0	7	2	0
T	3	0	2	0	0	0	5	0
G	1	1	0	7	7	0	0	6
C	2	5	4	0	0	0	0	1

Genome Annotations-Transcription Factors

Genome Annotations - Conclusion

- Genome annotations can also be performed by searching transcription factor binding sites

Genetics and Genomics

Arabidopsis Thaliana Genome – A Model Organism in Plant Biology and Genetics

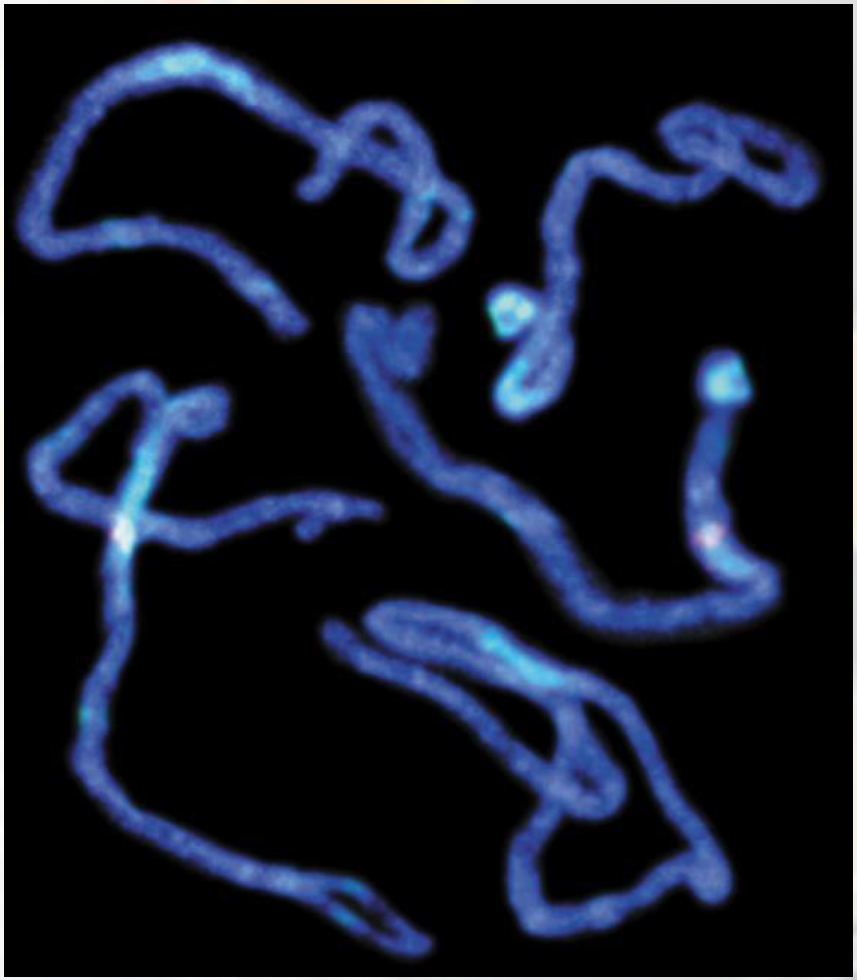
Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- Arabidopsis Thaliana – A model Plant
- Arabidopsis has 5 chromosomes (2n=10)

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Chromosomes



Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- Contains about 135 megabases of sequence
- Encodes approximately 27,000 genes and 35,000 proteins.

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- Has 35% unique genes
- Has 37% genes that exist as members of large gene families (families of 5 or more members)

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- An estimated 58-60% of the Arabidopsis genome exists as large segmental duplications

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- Arabidopsis genome contains genes encoding RNA polymerase subunits not seen in other eukaryotic organisms

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

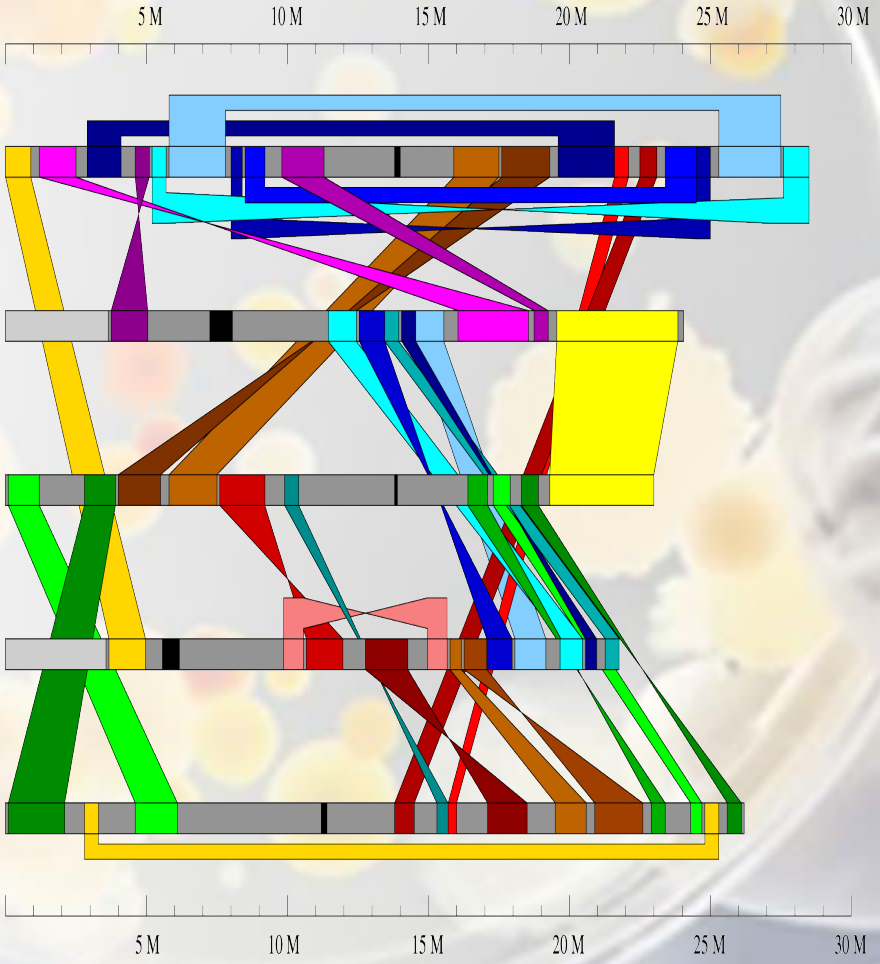
- Arabidopsis has genes unique to plants – approximately 150 unique protein families were found, including 16 unique families of transcription factors

Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome:

Segmental duplications in *Arabidopsis* thaliana.

Colored bands serve as representations of duplicated segments.



Arabidopsis Thaliana Genome

Arabidopsis Thaliana Genome

- Arabidopsis has 5 chromosomes (2n=10)
- Contains about 135 Mb of sequence
- Encodes ~ 27,000 genes and 35,000 proteins.

Genetics and Genomics

Mouse Genome – A Model Organism in Animal Biology and Genetics

Mouse Genome

Mouse Genome

- Number of chromosome 2n=40
- Human and mouse genomes have conserved blocks of genetic material
- Humans and mice suffer from similar diseases

Mouse Genome

Mouse genome

- Estimated genome Size 2.6 billion bp
- Number of estimated coding genes 22,500
- At the nucleotide level, approximately 40% of the human genome aligned to the mouse genome.

Mouse Genome

Mouse Genome: Chromosomes and their Sizes

Chromosome	cM	Mbp
1	98.5	195
2	103.9	182
3	82.7	160
4	88.6	157
5	90.2	152
6	79.0	150
7	89.1	145
8	76.2	129
9	75.1	125
10	77.9	131

Chromosome	cM	Mbp
11	88.0	122
12	63.9	120
13	67.3	120
14	66.4	125
15	59.0	104
16	57.8	91
17	61.3	95
18	59.4	91
19	56.9	61

Mouse Genome

Mouse Genome: Chromosomes and their Sizes

Type	Name	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	1	195.47	41.3	4,613	-	37	2,010	2,709	557
Chr	2	182.11	42.2	6,093	-	8	2,712	3,529	596
Chr	3	160.04	40.7	3,476	-	40	1,643	2,268	469
Chr	4	156.51	42.5	4,637	-	8	1,925	2,621	484
Chr	5	151.84	42.7	4,509	-	10	1,884	2,543	391
Chr	6	149.74	41.6	3,835	1	53	1,578	2,616	537
Chr	7	145.44	43.2	6,198	-	11	2,011	3,769	906
Chr	8	129.4	42.6	3,609	32	7	1,682	2,222	366
Chr	9	124.6	42.9	4,239	-	7	1,598	2,323	368
Chr	10	130.7	41.6	3,455	-	11	1,609	2,119	381
Chr	11	122.08	44.0	5,511	-	47	1,987	2,851	366
Chr	12	120.13	42.0	2,509	-	3	1,542	1,997	501
Chr	13	120.42	41.9	2,494	-	106	1,553	2,161	470
Chr	14	124.9	41.4	2,840	-	12	1,492	2,145	459
Chr	15	104.04	42.2	2,802	-	4	1,356	1,644	272
Chr	16	98.21	41.2	2,385	-	2	1,035	1,380	255
Chr	17	94.99	42.9	3,581	-	12	1,406	2,058	422
Chr	18	90.7	41.7	1,803	-	-	923	1,244	260
Chr	19	61.43	43.1	2,309	-	10	841	1,312	210
Chr	X	171.03	39.7	3,005	-	17	942	2,282	873
Chr	Y	91.74	39.3	238	-	-	84	372	140

Mouse Genome

Mouse Genome

- Mouse genome contains fewer CpG islands (15,500)
- The (G+C) content for each of the mouse chromosome is relatively similar

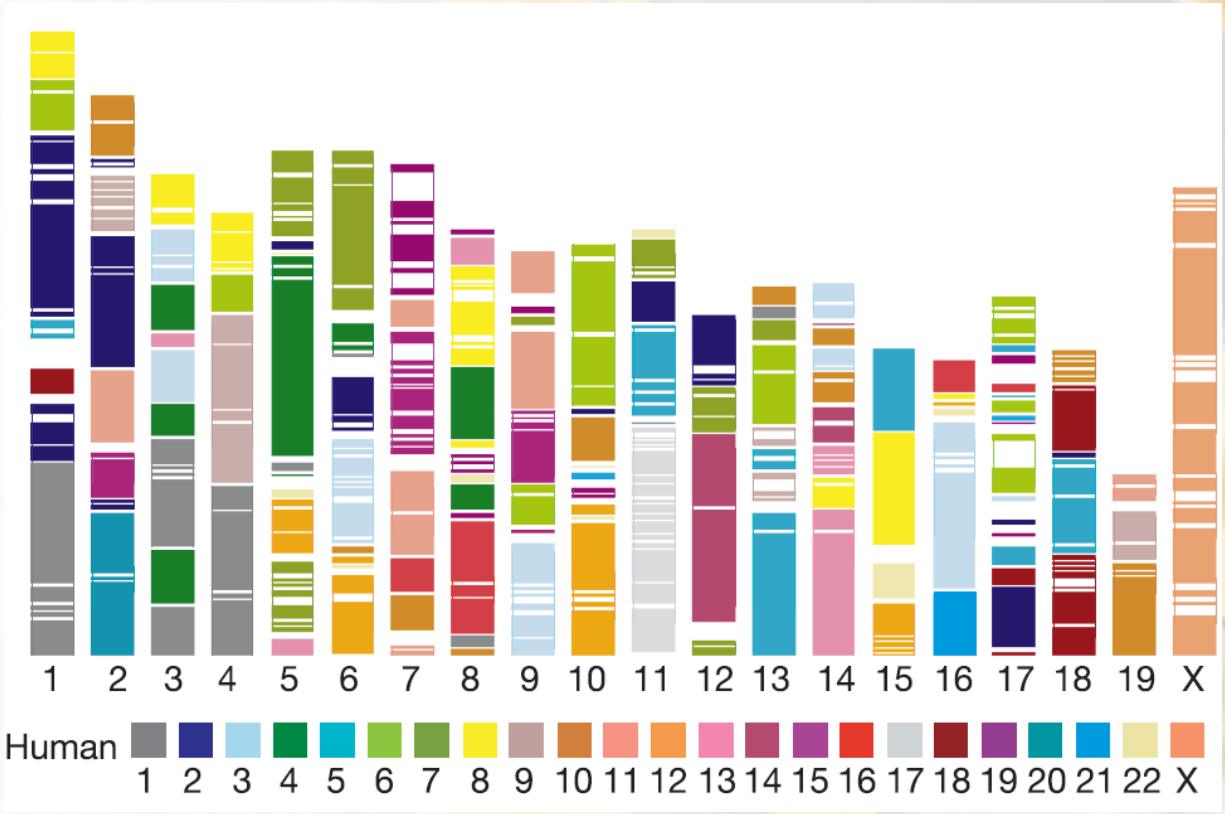
Mouse Genome

Mouse Genome

- Approximately 99% of mouse genes have a homologue in the human genome

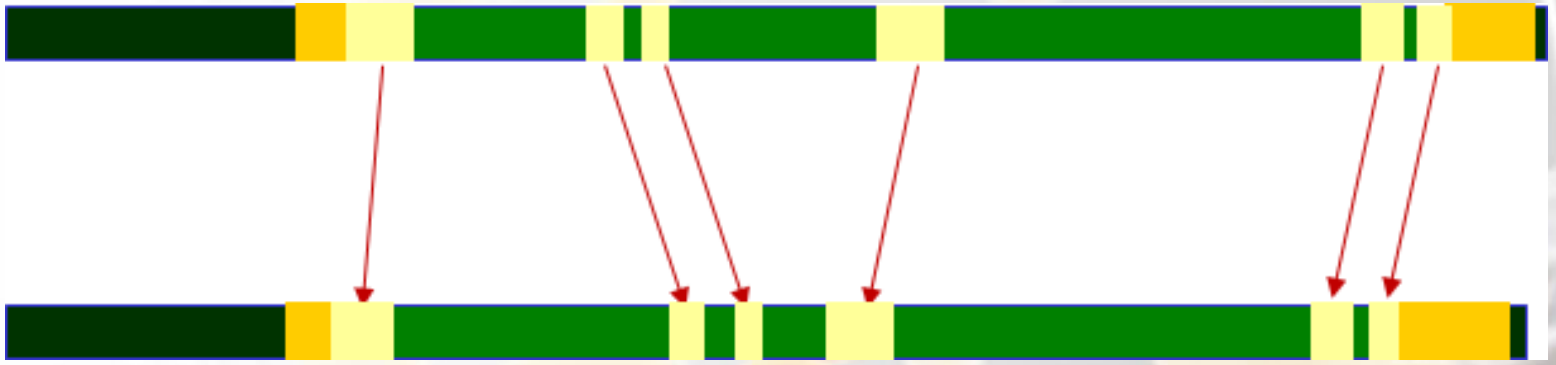
Mouse Genome

Human and Mouse Chromosomes



Mouse Genome

Human and Mouse: Chromosomes Orthologs



- Most human genes have mouse orthologs
- Coding exons usually correspond 1-1
- Coding sequence similarity ~ 85%

Mouse Genome

Mouse Genome Contigs

- Most mouse BAC contigs contained multiple mouse markers.
- Coverage of the mouse genome (2.6 billion bp) in mapped BACs is virtually complete: 296 contigs of average size 9.3 Mb.

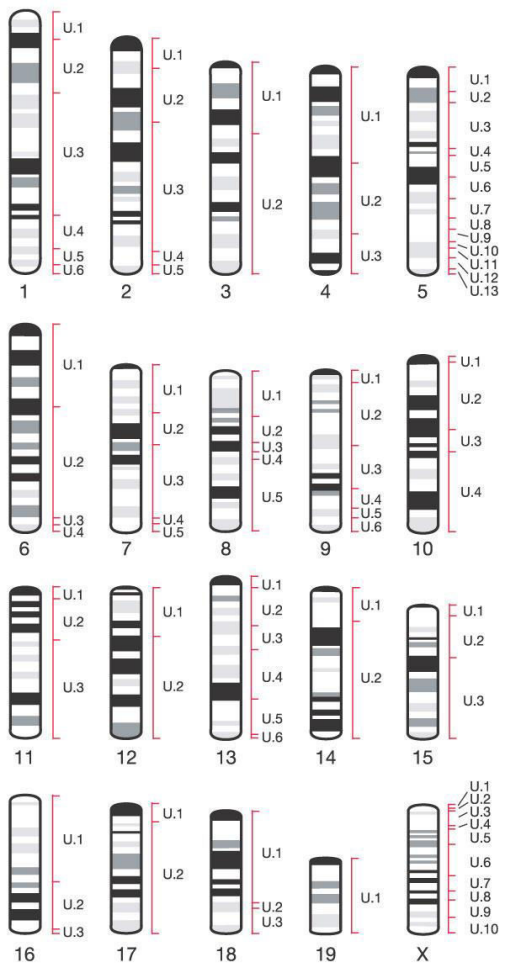
Mouse Genome

Mouse Genome

- All mouse chromosomes are acrocentric, with the centromeric end at the top of each chromosome.

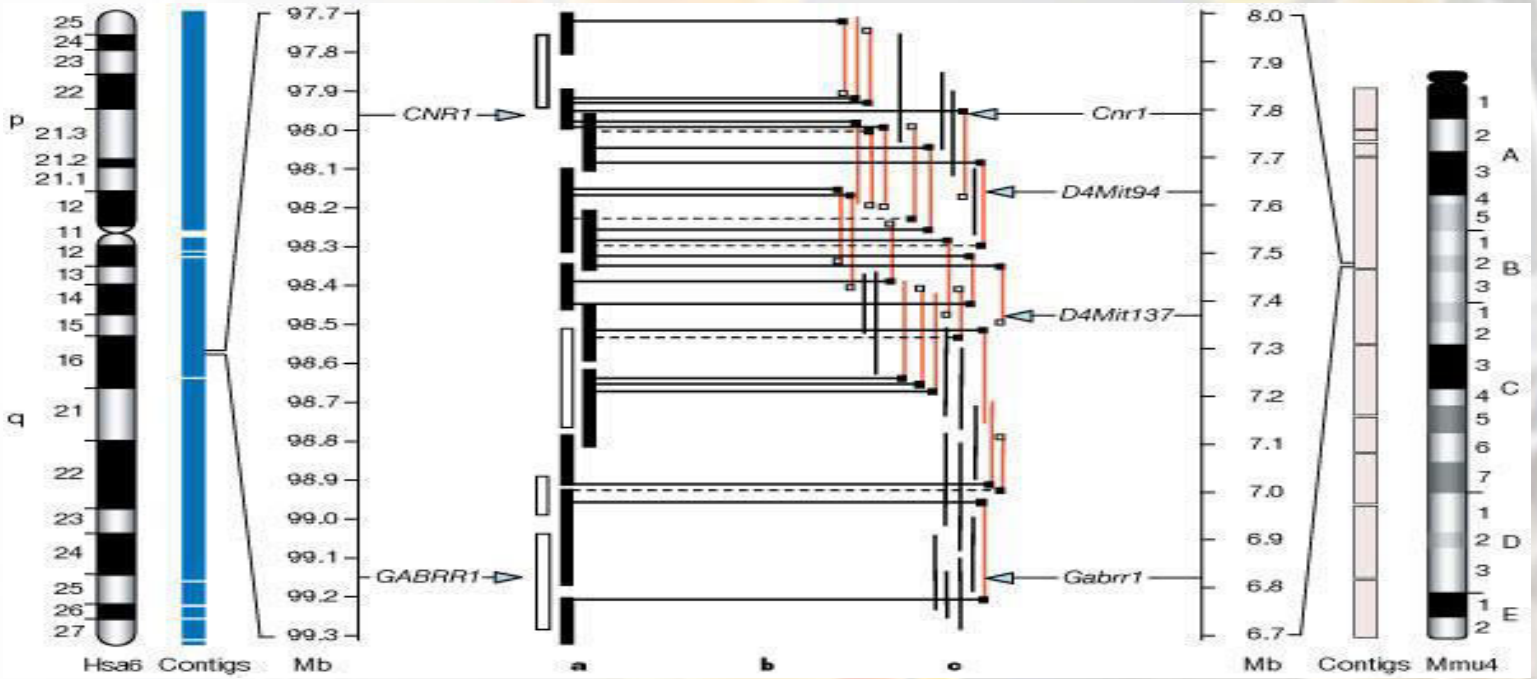
Mouse Genome

Cytogenetic View of Chromosomes



Mouse Genome

Human–Mouse Homology Clone Map



Alignment between part of human chromosome 6 and mouse chromosome 4 . 1.6-Mb interval is enlarged, showing part of 16.1 aligned to a 1.3-Mb mouse BAC contig. 11 of the 15 segments of human sequence match to 29 of segments of mouse BAC contig

Mouse Genome

Mouse Gene and Human Gene

Human Disease and Mouse Model Detail

Human Disease	Term: **Von Hippel-Lindau Syndrome; VHL** OMIM ID: 193300
Associated Genes	Orthologous mouse and human markers where mutations in one or both species have been associated with phenotypes characteristic of this disease.

Mouse Gene	Human Gene	Characteristics of this human disease are associated with mutations in...
Vhl	VHL	...both mouse and human orthologous genes.
Hif1a	HIF1A	...the mouse gene. *OMIM data currently do not associate this disease with the orthologous human gene.*
Ccnd1	CCND1	...the human gene. *MGI data currently do not associate this disease with a mouse model.*

Mouse Models

Genotype			Ref(s)
Allelic Composition	Note	Genetic Background	
Models with phenotypic similarity to human disease where etiologies involve orthologs.[1]			
Hif1a^{tm3Rsio}/Hif1a^{tm3Rsio} Hprt1$^{tm1(Pck1-cre)Vhh}$/Y Vhl^{tm1Jae}/Vhl^{tm1Jae}	2	involves: 129 * BALB/c * C57BL/6	J:97652, J:106705
Hif1a^{tm3Rsio}/Hif1a^{tm3Rsio} Vhl^{tm1Jae}/Vhl^{tm1Jae} Tg(Alb-cre)21Mgn/0	2	involves: 129 * BALB/c * C57BL/6 * DBA	J:97652

Mouse Genome

Mouse Genome

- Number of chromosome 2n=40
- Humans and mice suffer from similar diseases
- Estimated genome Size 2.6 billion bp
- Number of estimated coding genes 22,500

Genetics and Genomics

Genome Comparison - Human and Mouse

Genome Comparison-Human and Mouse

Comparison

- The mouse genome is about 14% smaller than the human genome
- Over 90% of mouse and human genomes can be partitioned into corresponding regions of conserved synteny

Genome Comparison-Human and Mouse

Comparison

- At the nucleotide level, 40% of the human genome can be aligned to the mouse genome

Genome Comparison-Human and Mouse

Comparison

- Mouse and human genomes each seem to contain ~ 23,000 protein-coding genes.

Genome Comparison-Human and Mouse

Comparison

- The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) is < 1%.

Genome Comparison-Human and Mouse

Comparison

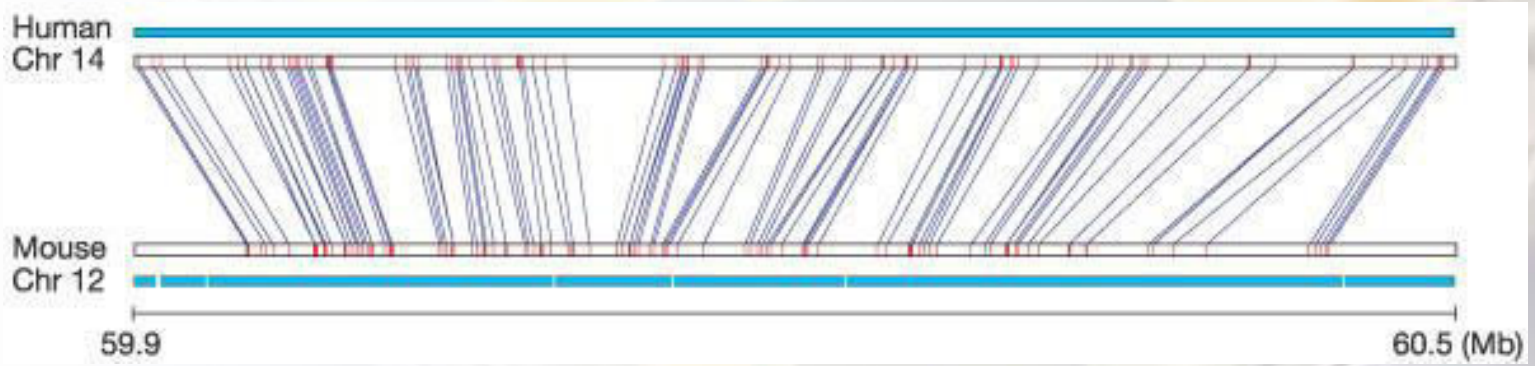
- Despite marked differences in activity of transposable elements, similar types of repeat sequences have accumulated in the corresponding genomic regions

Genome Comparison-Human and Mouse

Comparison - Human and Mouse Genome

A typical 510-kb segment of mouse chromosome 12 that shares common ancestry with a 600-kb section of human chromosome 14.

Blue lines connect the reciprocal unique matches in the two genomes.



Genome Comparison-Human and Mouse

Comparison – Human and Mouse Genome

- Mouse genome contains fewer CpG islands than human.
- (G+C) content and density of CpG islands shows more variability in human than mouse

Genome Comparison-Human and Mouse

Comparison

- Human chromosomes show more variation (G+C) on chromosomes 16, 17, 19 and 22 have higher (G+C) content, and chromosome 13 lower (G+C) content.

Genome Comparison-Human and Mouse

Comparison

- The density of CpG islands is relatively homogenous for all mouse chromosomes and more variable in human.

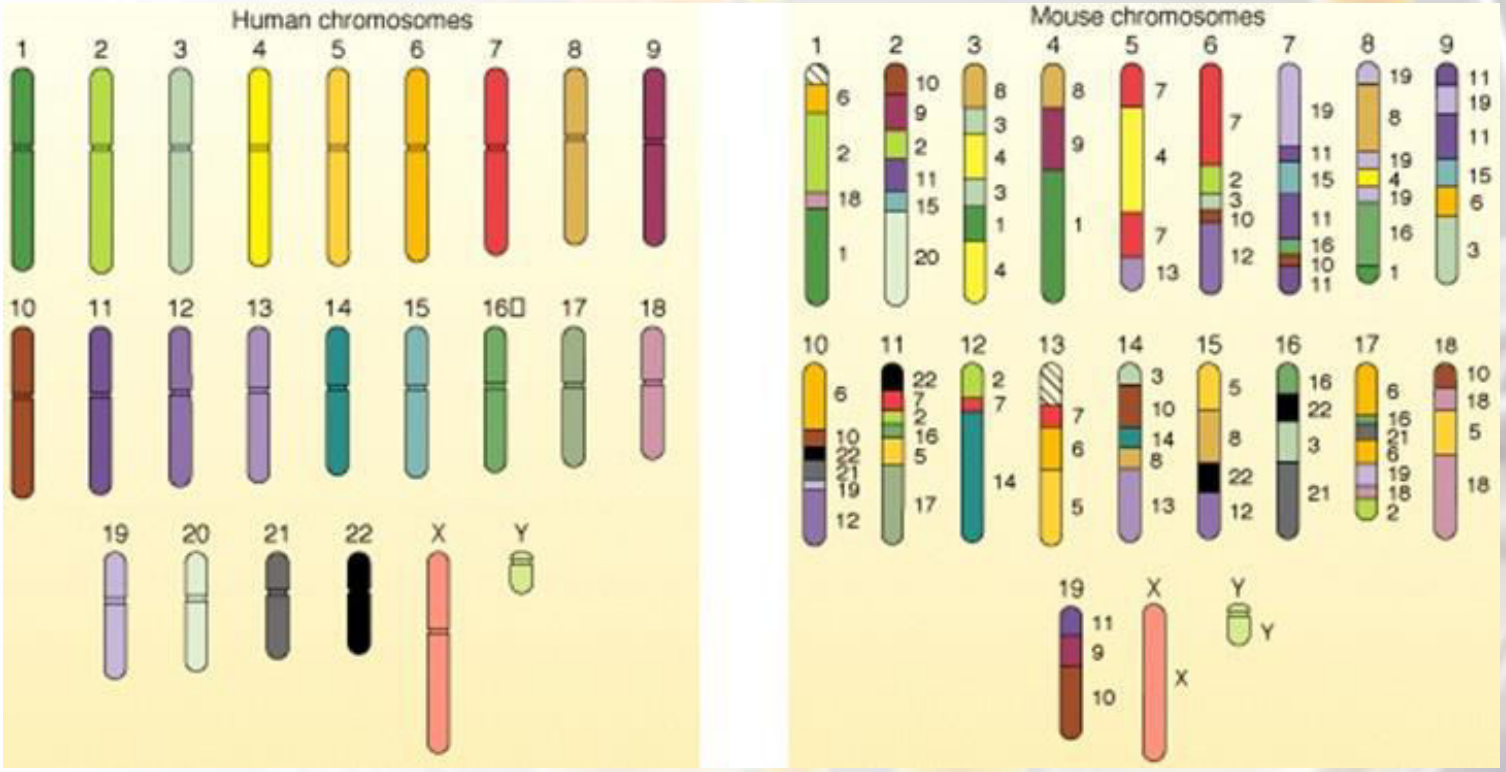
Genome Comparison-Human and Mouse

Comparison

- Approximately 99% of mouse genes have a homologue in the human genome

Genome Comparison-Human and Mouse

Comparison – Human and Mouse Genome



Genome Comparison-Human and Mouse

Comparison - Conclusion

- The mouse genome is about 14% smaller than the human genome
- Mouse and human genomes each seem to contain ~ 23,000 protein-coding genes.

Genetics and Genomics

Mapping of Disease Genes

Mapping of Disease Genes

Mapping of Disease Genes

- Disease gene mapping is one of the main objectives of genotyping
- Different approaches for identification of disease genes

Mapping of Disease Genes

Position Independent Method

- No information about the location of the gene.
- Starting at the phenotype, determining which protein was involved and getting to gene through the protein.

Mapping of Disease Genes

Position Dependent Method

- Starting from the approximate location of gene, to finding the gene itself, then translating it to learn about the protein and its function.
- Positional cloning or reverse genetics

Mapping of Disease Genes

Position Independent Method

- Mostly identifying and isolating the protein product of the gene.
- Such genes usually produce large amounts of well-known and well studied proteins.

Mapping of Disease Genes

Position Independent Method

- Two common methods used
- Gene-specific oligonucleotides
- Antibody Methods

Mapping of Disease Genes

Position Dependent Method

- Recombination mapping and/or somatic cell mapping to defined the region of interest as tightly as possible.

Mapping of Disease Genes

Position Dependent Method

- Chromosomes are inherited as intact units, so it is reasoned that the alleles of some pairs of genes will inherit together because they are on the same chromosome.

Mapping of Disease Genes

Mapping of Disease Genes

- Different approaches are used for identification of disease genes

Genetics and Genomics

DNA Microarray

DNA Microarrays

DNA Microarrays

- Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known DNA sequence.

DNA Microarrays

DNA Microarrays

- There are several synonyms of DNA microarrays such as
- DNA chips
- Gene chips
- DNA arrays
- Gene arrays
- Biochips

DNA Microarrays

Microarray - Principle

- The principle of DNA microarrays lies on the hybridization between the nucleotides.

DNA Microarrays

Microarray

- Presence of one genomic or cDNA sequence in 100,000 or more sequences can be screened in a single hybridization.

DNA Microarrays

DNA Microarray

- DNA microarray allows us to analyze thousands of genes in one experiment
- DNA microarrays are solid supports; glass or silicon, upon which DNA is attached

DNA Microarrays

Simultaneous Studies of Many Genes

- Patterns or clusters of genes are more informative regarding total cellular function than looking at one or two genes – can figure out new pathways

DNA Microarrays

Simultaneous Studies of Many Genes

- Just sequencing genomes is not sufficient
- Thousands of genes remain without an assigned function
- Each spot of DNA, (probe), represents a single gene

DNA Microarrays

Microarray Use

- Determine which genes are active in a cell and at what levels
- Compare the gene expression profiles of a control vs treated
- Assessing gene expression levels

DNA Microarrays

Microarray Use

- Genome-wide studies and genotyping
- Evaluating microRNA levels
- Determine which genes have biological significance in a system

DNA Microarrays

Microarray Use

- Discovery of new genes, pathways, and cellular trafficking

Types of Microarrays

Types of Microarrays

Types of Microarrays

- DNA microarrays
- Exon arrays
- Comparative Genomic Hybridization array

Types of Microarrays

Types of Microarrays

- DNA microarrays- such as cDNA microarrays and oligonucleotide microarrays, SNPs
- Exon arrays- Alternative splice variant detection

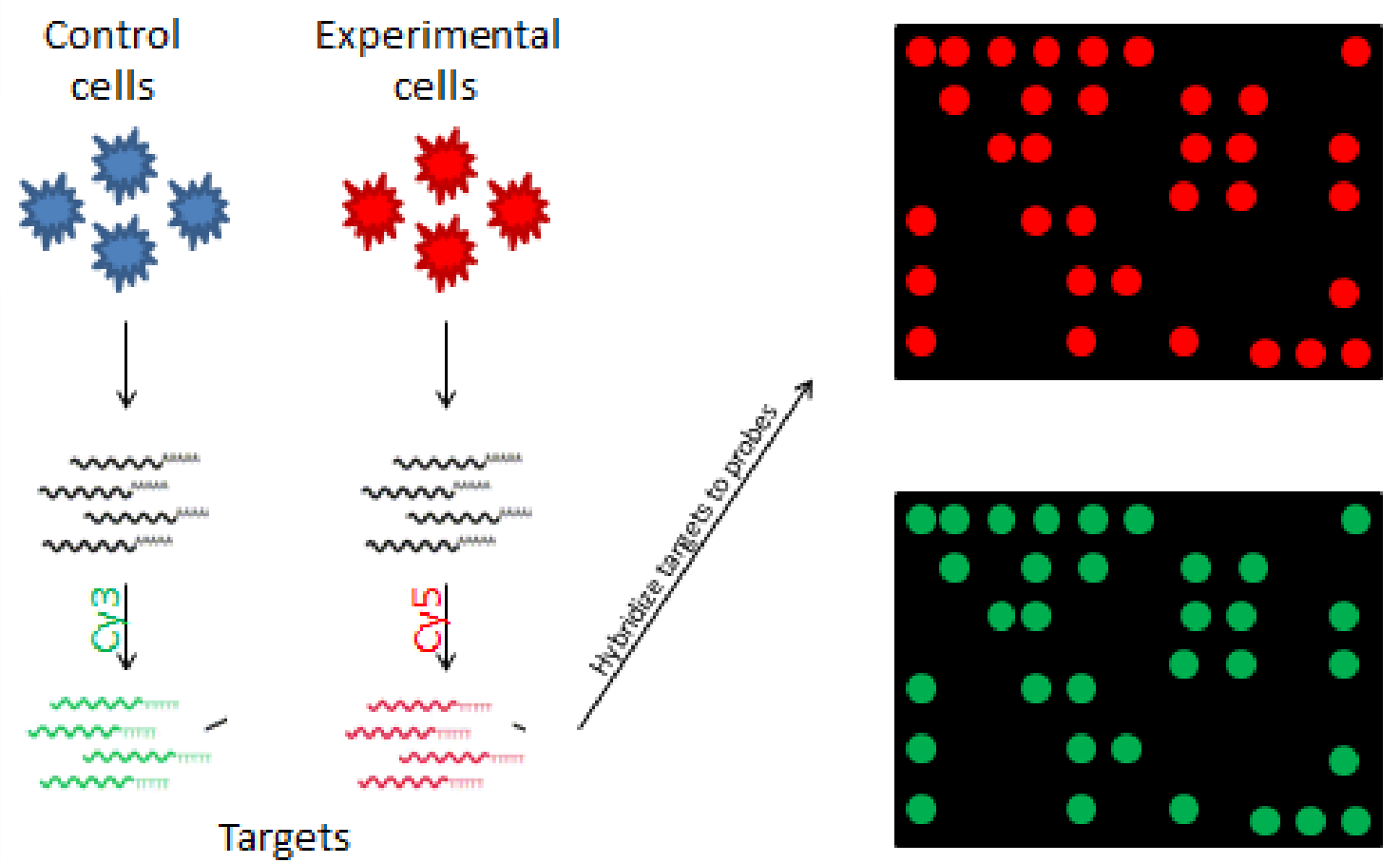
Types of Microarrays

Types of Microarrays

- CGH arrays- Comparative genomic hybridization

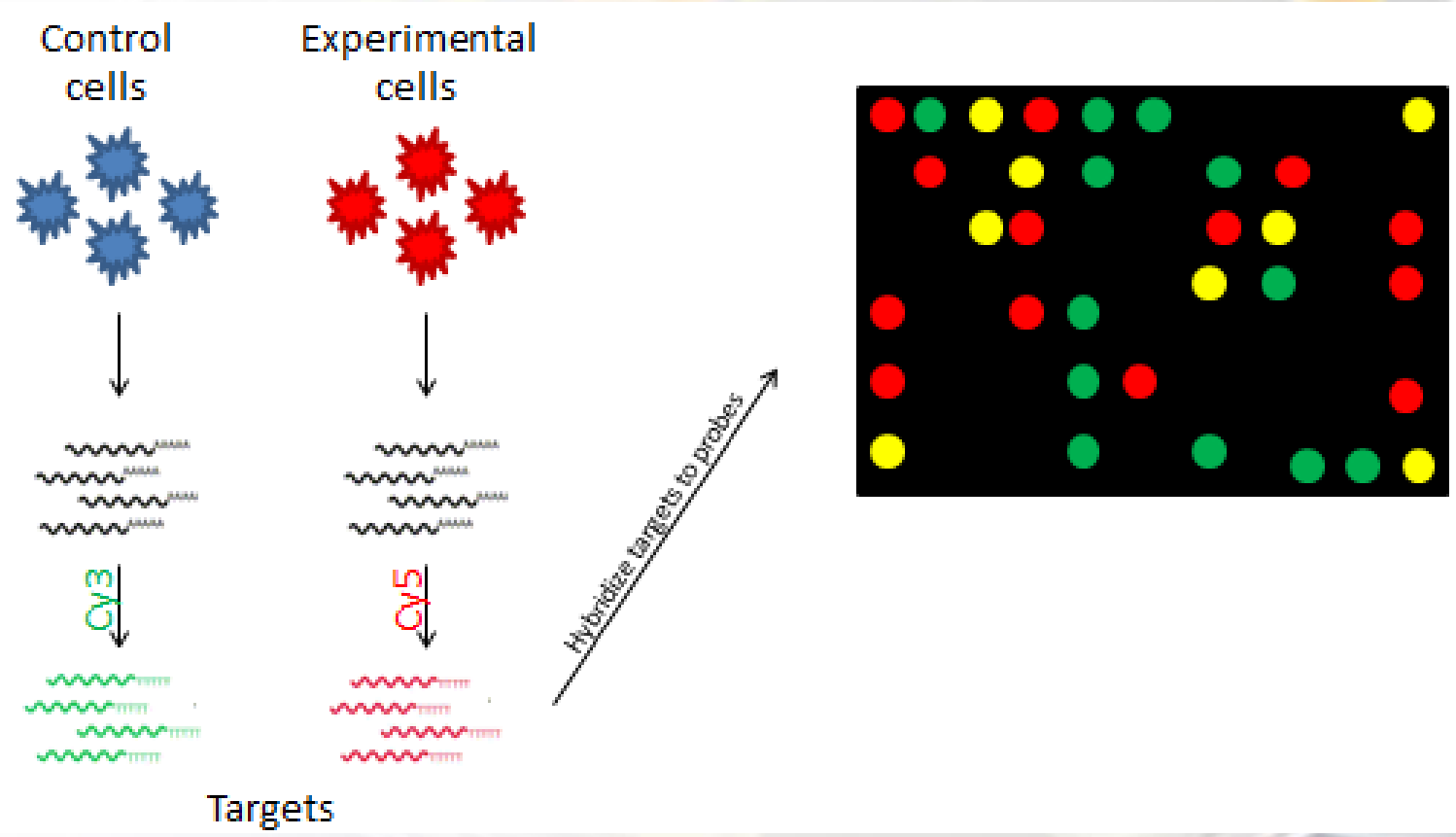
Types of Microarrays

Oligonucleotide Arrays



Types of Microarrays

cDNA Arrays



Types of Microarrays

Types of Microarray

- Protein microarrays (protein-protein interactions)
- Tissue microarrays
- Cellular microarrays (also called transfection microarrays
- Expression Arrays

Types of Microarrays

Types of Microarray

- Chemical compound microarrays
- Antibody microarrays (proteomics)
- Carbohydrate arrays (glycoarrays)
- Re-sequencing arrays

Types of Microarrays

Microarray Require

- Microarrays combine genomics, silicon chip manufacturing, DNA and Protein chemistry, signal and image processing, statistics, software skills and traditional molecular biology experiments.

Types of Microarrays

Microarray and Genes Expression Profiling

- Gene expression profiling can be monitored for thousands of genes simultaneously
- CGH arrays assesses genome content in different cells or related organisms

Types of Microarrays

Microarray and SNPs

- SNP array: Identifying single nucleotide polymorphism among alleles within or between populations

Types of Microarrays

Methylation Microarrays

- Methylation arrays determine methylated DNA
- Determining which regions of DNA are methylated ultimately to study epigenetics

Microarray Formats

Microarrays Formats

Main Formats

- Cartridge-based
  - Spotted
  - Electronic
- Spotted Glass Slide
- Tissue Section Slide

Microarrays Formats

Cartridge-Based Chips

- High density arrays of DNA oligos within a plastic housing
- One sample = One chip (Affymetrix, Agilent, Applied Biosystems etc)

Microarrays Formats

Cartridge-Based Chips

- Generally used with expression and DNA arrays

Microarrays Formats

Cartridge-Based Expression Microarrays

- Involves Fluorescently tagged biotinylated cRNA
- One chip per sample
- Uses single fluorescent dye

Microarrays Formats

Spotted Glass Arrays

- Uses cDNA, Oligonucleotide, protein, antibody
- Robotically spotted cDNAs or Oligonucleotides
- Printed on Nylon, Plastic, or Glass microscope slide

Microarrays Formats

Spotted cDNA and Oligo Glass Arrays

- Involves two dyes on the same slide
  - Red dye-Cy5
  - Green dye-Cy3
- Control and experimental cDNA

Microarrays Formats

Main Formats

- Cartridge-based
  - Spotted
  - Electronic
- Spotted Glass Slide
- Tissue Section Slide

Genetics and Genomics

Microarray Procedure

Microarray - Procedure

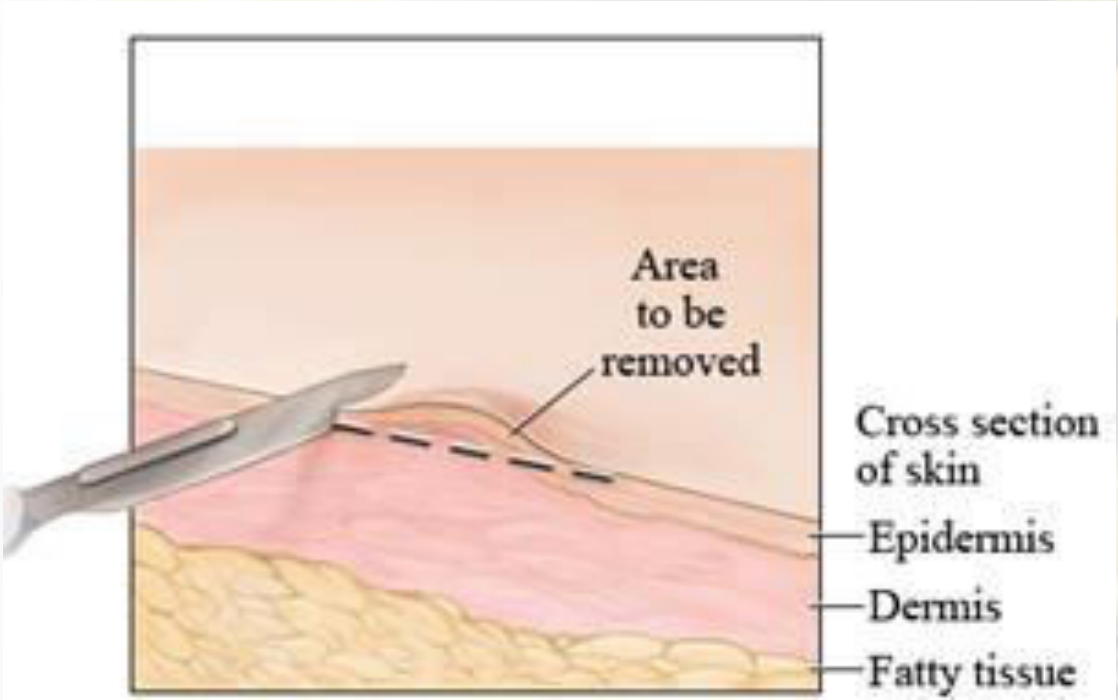
Microarray Procedure

- Collect Samples
- Isolate mRNA
- Create Labelled DNA
- Hybridization
- Microarray Scanner
- Analyze Data

Microarray - Procedure

Microarray Procedure

- Samples: This can be from a variety of organisms. Two samples – cancerous human skin tissue & healthy human skin tissue



Microarray - Procedure

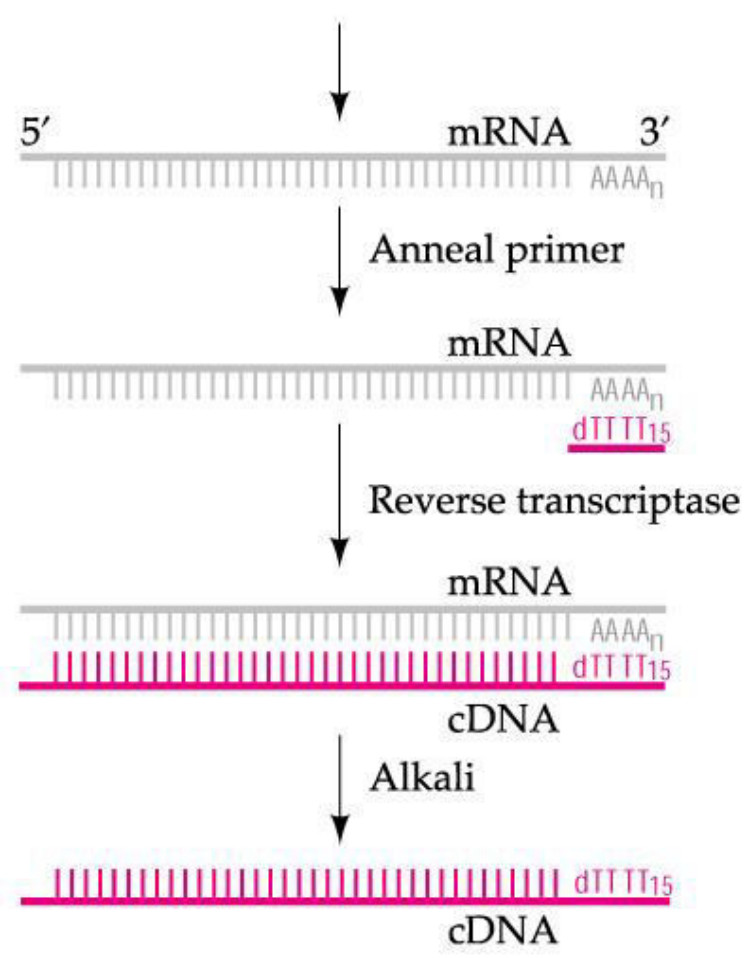
Microarray Procedure

- Extract the RNA from the samples.
- After isolating the RNA, isolate the mRNA from the rRNA and tRNA.
- mRNA has a poly-A tail, use a column containing beads with poly-T tails to bind the mRNA
- Rinse with buffer to release the mRNA from the beads. The buffer disrupts the pH, disrupting the hybrid bonds.

Microarray - Procedure

Microarray Procedure - Create Labeled cDNA

Add a labelling mix to the RNA. The labelling mix contains poly-T (oligo dT) primers, reverse transcriptase (to make cDNA), and fluorescently dyed nucleotides.
Add cyanine 3 (fluoresces green) to the healthy cells and cyanine 5 (fluoresces red) to the cancerous cells.



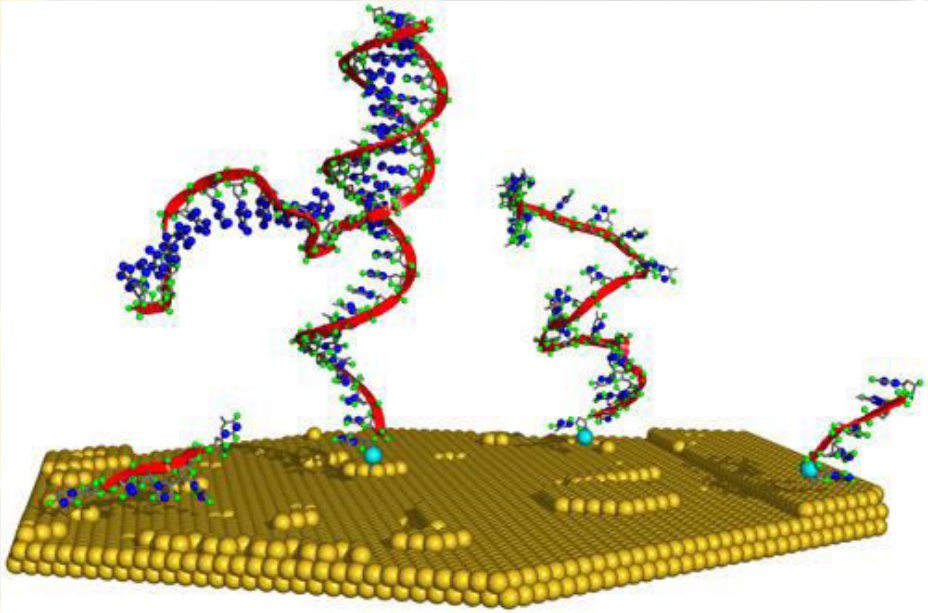
Microarray - Procedure

Microarray Procedure - Hybridization

Apply the cDNA to a microarray plate.

When comparing two samples, apply both samples to the same plate.

The ssDNA will bind to the cDNA already present on the plate.



Microarray - Procedure

Microarray Procedure – Lasers

- Laser scans array and produces images
- One laser for each color, e.g. one for green, one for red

Microarray - Procedure

Microarray Procedure – Microarray Scanner

The scanner has a laser, a computer, and a camera.

The laser causes the hybrid bonds to fluoresce.

The camera records the images produced when the laser scans the plate.

The computer allows to immediately view the results and it also stores the data.

Microarray - Procedure

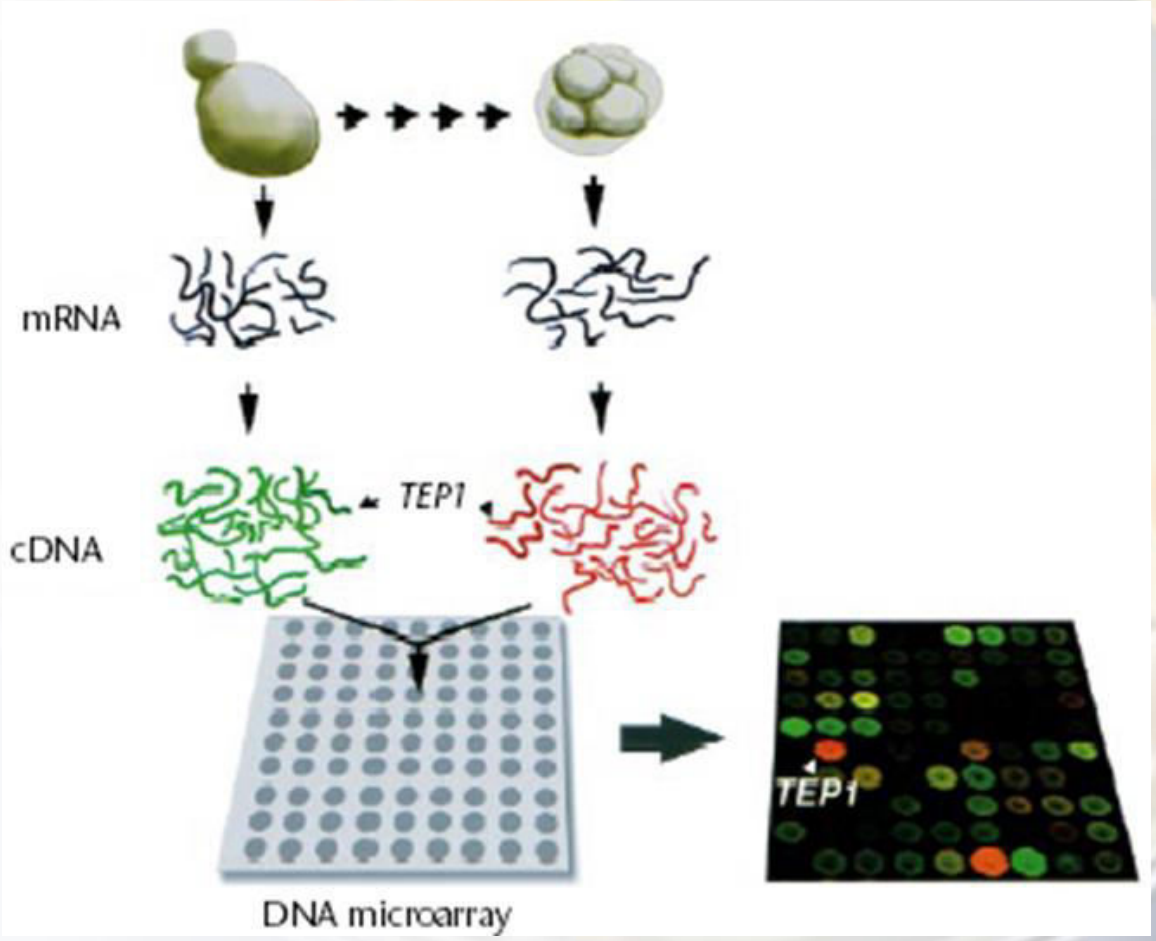
Microarray Procedure – Microarray Scanner

GREEN – the healthy sample hybridized more than the diseased sample.

RED – the diseased/cancerous sample hybridized more than the nondiseased sample.

Microarray - Procedure

Microarray Procedure – Microarray Scanner

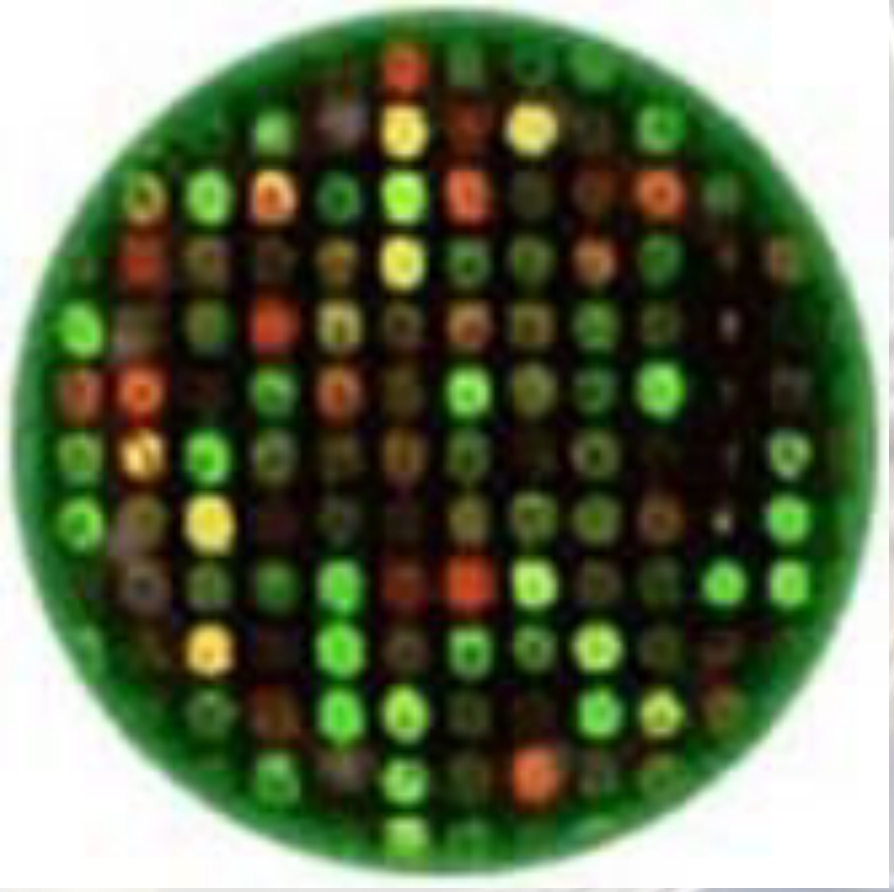


Microarray - Procedure

Microarray Procedure – Microarray Scanner

YELLOW - both samples hybridized equally to the target DNA.

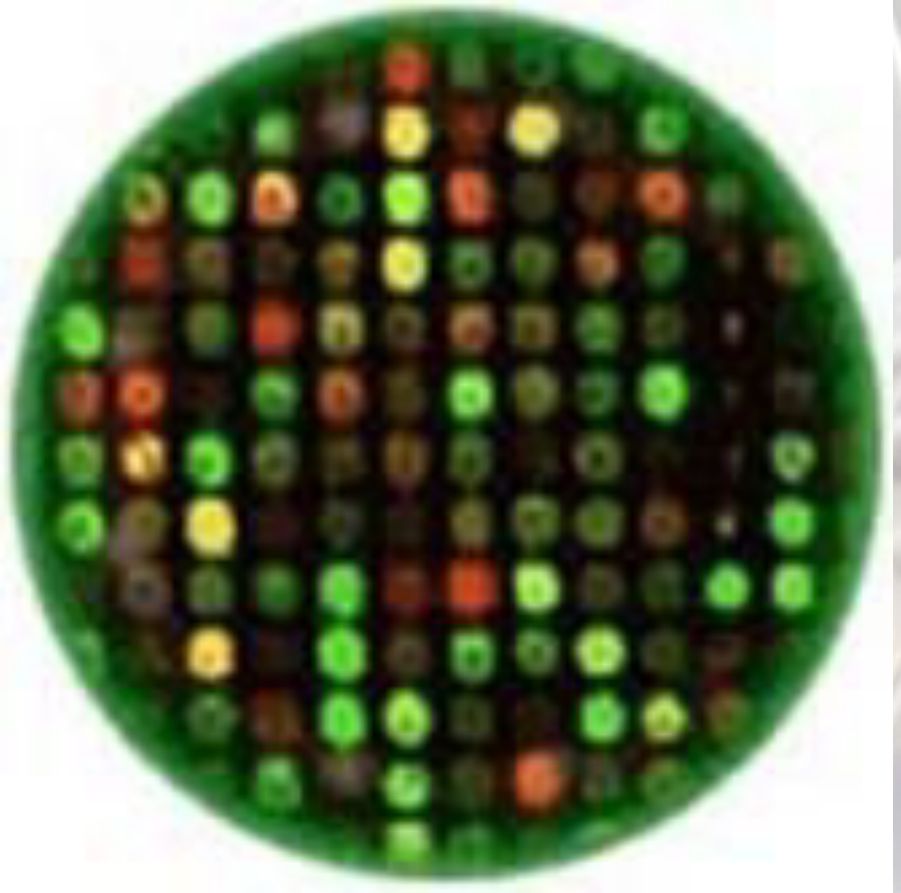
BLACK - areas where neither sample hybridized to the target DNA.



Microarray - Procedure

Microarray Procedure – Microarray Scanner

By comparing the differences in gene expression between the two samples, we can understand more about the genomics of a disease.



Microarray - Procedure

Microarray Procedure

- Collect Samples
- Isolate mRNA
- Create Labelled DNA
- Hybridization
- Microarray Scanner
- Analyze Data

Genetics and Genomics

Microarray Chips

Microarray Chips

Microarray Chips

- There are two types of DNA Chips/Microarrays:
- cDNA based microarray
- Oligonucleotide based microarray

Microarray Chips

cDNA Based Microarray Chips

- Chips are prepared by using cDNA- cDNA chips or cDNA microarray.
- The cDNAs are amplified. Then these immobilized on a solid support made up of nylon filter of glass slide

Microarray Chips

cDNA Based Microarray Chips

- The probe DNA are loaded by capillary action.
- Small volume of this DNA is spotted on solid surface
- DNA is delivered mechanically or in a robotic manner.

Microarray Chips

cDNA Based Microarray Chips

- When one DNA spotting is done, the pin is washed and loaded with fresh DNA to start the second cycle

Microarray Chips

DNA Based Microarray Chips

- Almost similar type procedure
- Gene expression profiling
- Discovery of drugs
- Diagnostics and genetic engineering
- Alternative splicing detection

Microarray Chips

DNA Based Microarray Chips

- Functional genomics
- DNA sequencing
- Toxicological research (Toxicogenomics)

Microarray Chips

Microarray Chips

- There are 2 types of DNA Chips/Microarrays:
- cDNA based microarray
- Oligonucleotide based microarray

Genetics and Genomics

Microarray Applications

Microarray Applications

Microarray - Advantages

- Provides data for thousands of genes
- One experiment instead of many
- Fast and easy to obtain results
- Closer to discovering cures for diseases and cancer

Microarray Applications

Microarray Advantages

- Different parts of DNA can be used to study gene expression

Microarray Applications

Disadvantages

- The biggest disadvantage of DNA chips is that they are expensive to create.
- The production of too many results at a time requires long time for analysis, which is quite complex in nature.

Microarray Applications

Disadvantages

- The DNA chips do not have very long shelf life, which proves to be another major disadvantage of the technology.

Microarray Applications

Microarray - Limitations

- Cross-hybridization of sequences with high identity
- Chip to chip variation
- The real limitation is Bioinformatics
- Expensive – repeat experiments

Microarray Applications

Gene Discovery and Disease Diagnostics

- Gene discovery and disease diagnosis
- Classify the types of cancer on the basis of the patterns of gene activity in the tumor cells

Microarray Applications

Pharmacogenomics

- Study of correlations between therapeutic responses to drugs and the genetic profiles of the patients

Microarray Applications

Toxicogenomics

- Microarray technology allows us to research the impact of toxins on cells.
- Some toxins can change the genetic profiles of cells.

Microarray Applications

Biological Applications

- Many biological discovery
- New and better molecular diagnostics
- New molecular targets for therapy
- Finding and refining biological pathways

Genetics and Genomics

Types of RNA and RNomics

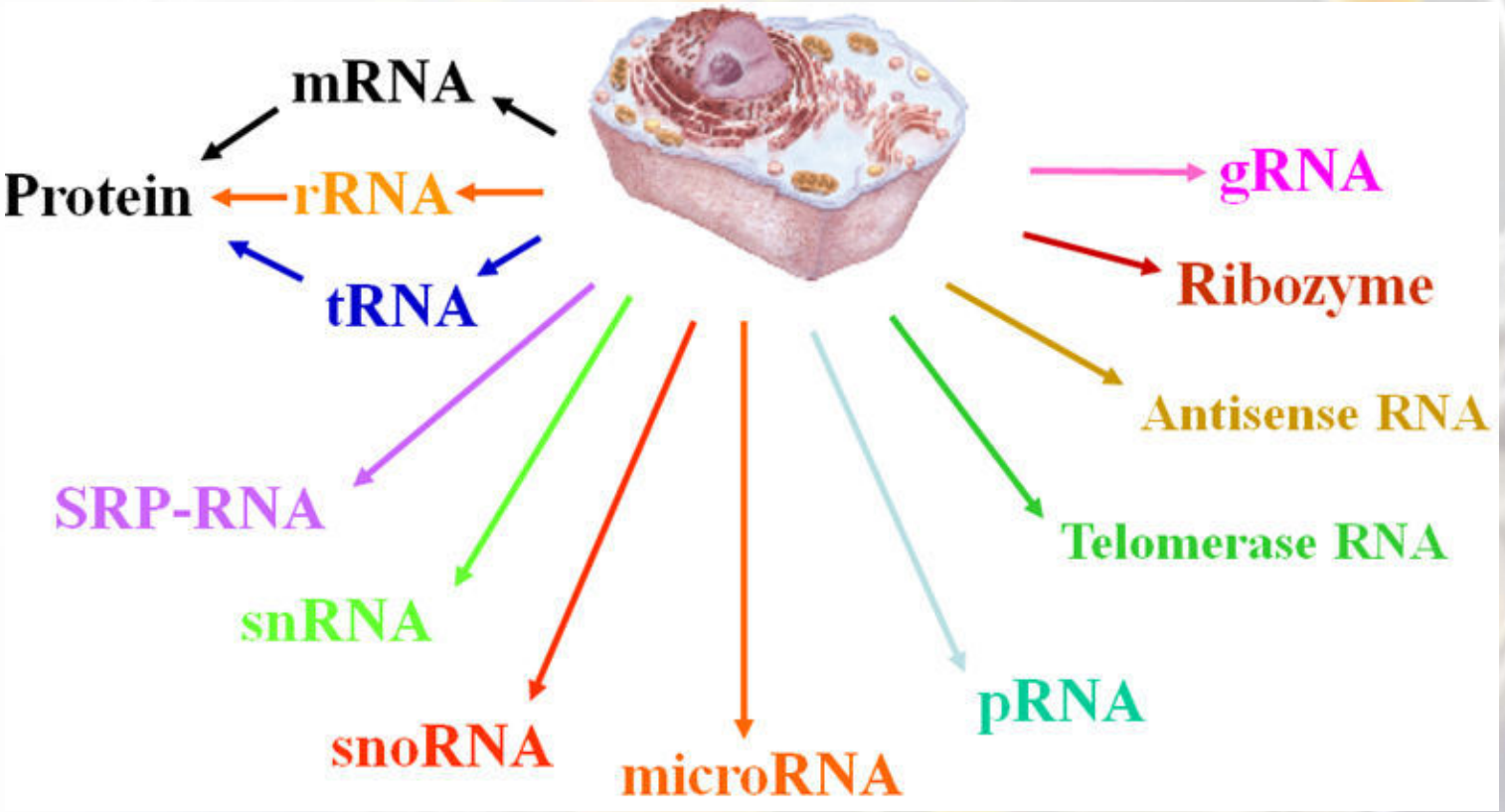
Types of RNA and RNomics

Types of RNA

- Wide range of RNAs
- Some of them are coding while others are noncoding

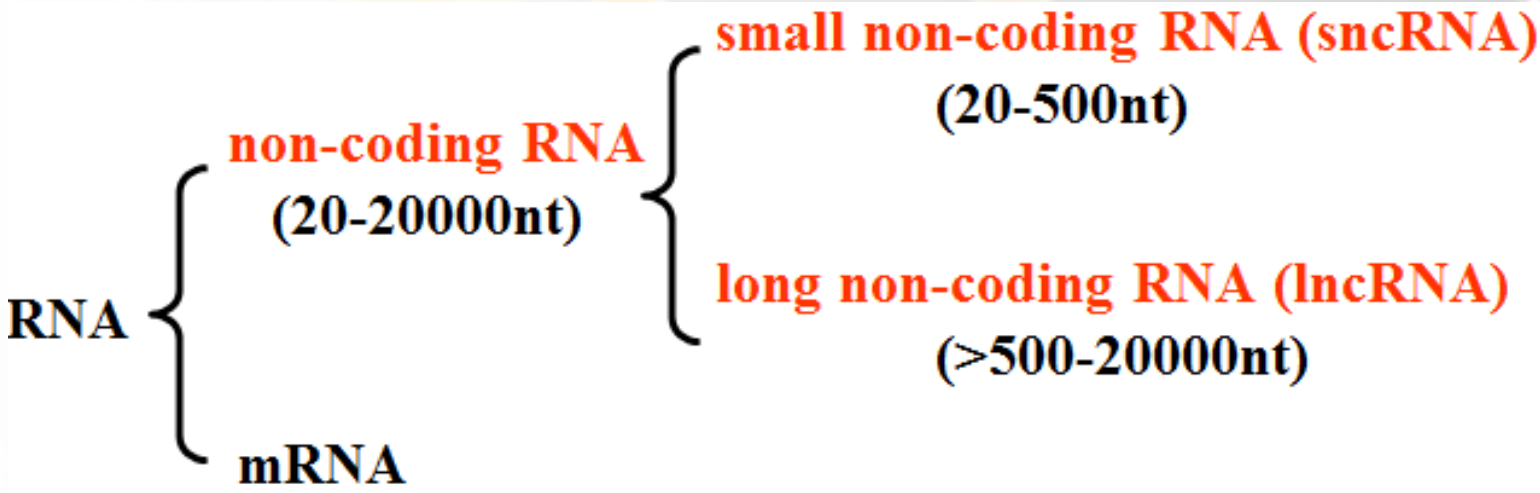
Types of RNA and RNomics

Types of RNA



Types of RNA and RNomics

Types of RNA - Classification



Types of RNA and RNomics

Types of RNA – Housekeeping RNAs

Category	Name
Housekeeping RNAs	Transfer RNAs
	Ribosomal RNAs
	Small nucleolar RNAs
	Small nuclear RNAs

Types of RNA and RNomics

Types of RNA – Small Non-coding RNAs

Category	Name
Small ncRNAs (200 bp or less in size)	MicroRNAs
	Tiny transcription initiation RNAs
	Repeat associated small interfering RNAs
	Promoter-associated short RNAs
	Termini-associated short RNAs
	Antisense termini associated short RNAs
	Transcription start site antisense RNAs
	Retrotransposon-derived RNAs
	3'UTR-derived RNAs
	Splice-site RNAs

Types of RNA and RNomics

Types of RNA – Long Non-coding RNAs

Category	Name
Long ncRNA (over 200 bp in size)	Long or large intergenic ncRNAs
	Transcribed ultraconserved regions
	Pseudogenes
	Enhancer RNAs
	Repeat-associated ncRNAs
	Long intronic ncRNAs
	Antisense RNAs
	Promoter-associated long RNAs
	Long stress-induced non-coding transcripts

Types of RNA and RNomics

RNomics is the study

- Identification
- Expression
- Biogenesis
- Structure
- Regulation of expression
- Targets
- Biological functions of RNAs

Types of RNA and RNomics

Computational RNomics

- Searching conserved intronic sequences by comparative analysis of introns
- Searching conserved intergenic sequences

Types of RNA and RNomics

Experimental RNomics

- PAGE separation of non-coding RNAs and sequencing
- Non-coding RNA enriched cDNA libraries and sequencing

Types of RNA and RNomics

Types of RNA

- Wide range of RNAs
- Some of them are coding RNA while others are noncoding.

Genetics and Genomics

Non-coding DNA

Non-coding DNA

Types of Non-coding DNA

- Non- coding functional RNA
- Cis- and Trans-regulatory elements
- Introns
- Pseudogenes

Non-coding DNA

Types of Non-coding DNA

- Repeat sequences, transposons and viral elements
- Telomeres

Non-coding DNA

Non-coding Functional RNA

- The RNA molecules which are not translated into proteins.
- Examples: Ribosomal RNA, transfer RNA & micro RNA

Non-coding DNA

Cis- and Trans-Regulatory Elements

- Those sequences that control the transcription of a nearby gene.
- Located within 5’ or 3’ untranslated regions or within introns.

Non-coding DNA

Introns

- They are non-coding sections of a gene.
- Transcribed in the precursor mRNA sequence but is ultimately removed by RNA splicing.
- Most of the introns appear to be mobile genetic elements.

Non-coding DNA

Pseudogenes

- They are related to known genes, that have lost their protein-coding ability or are otherwise no longer expressed in the cell.
- Arise from retrotransposition or genomic duplication of functional genes.

Non-coding DNA

Repeat Sequences, Transposons & Viral Elements

- Transposons and Retrotransposons are mobile genetic element
- Retrotransposons: LINEs, SINEs

Non-coding DNA

Telomeres

- Telomeres are regions of repetitive DNA
- Located at the end of a chromosome
- They provide protection from chromosomal deterioration during DNA replication

Non-coding DNA

Functions

- Essential for chromosome structure
- Genome Protection
- Enhancers, silencers, promotors, insulator
- Genetic switches
- Regulation of gene expression

Genetics and Genomics

Non-coding RNA

Non-coding RNA

Non-coding RNA

- Non-coding RNA is a RNA molecule that functions without being translated into a protein
- Large number of genes of noncoding RNA
- Noncoding RNA - different functions

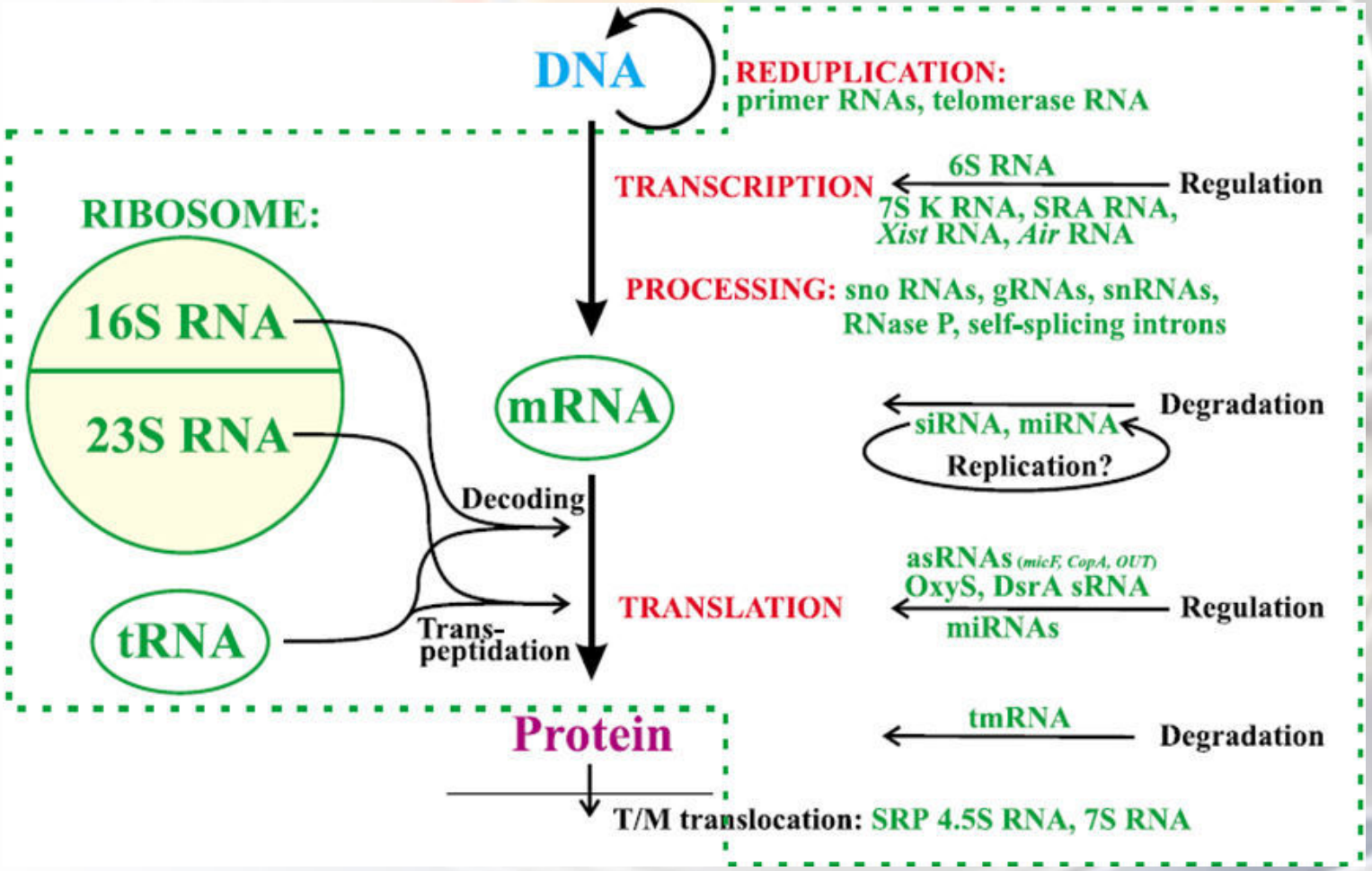
Non-coding RNA

Types of Non-coding RNA

- Ribosomal RNA
- Transfer RNA
- Small nuclear RNA
- Small nucleolar RNA
- Short interfering
- Micro RNA
- Long non-coding RNA

Non-coding RNA

Functional Diversity of Non-coding RNA



Non-coding RNA

Non-coding RNAs in Human Genome

tRNA	600
18S rRNA	200
5.8S rRNA	200
28S rRNA	200
5S rRNA	200
snoRNA	300
miRNA	250
U1	40
U2	30
U4	30
U5	30
U6	20
U4atac	5
U6atac	5
U11	5
U12	5

SRP RNA	1
RNase P RNA	1
Telomerase RNA	1
RNase MRP	1
Y RNA	5
Vault	4
7SK RNA	1
Xist	1
H19	1
BIC	1
Antisense RNAs	1000s?
Cis reg regions	100s?
Others	?

Non-coding RNA

Non-coding RNA

- Non-coding RNA (ncRNA) is a RNA molecule that functions without being translated into a protein

MicroRNA (miRNA)

MicroRNA (miRNA)

MicroRNA (miRNA)

- Small non-coding double stranded RNAs
- Approximately 19-22 nucleotides long
- Repress activity of complementary mRNAs

MicroRNA (miRNA)

MicroRNA (miRNA)

- Regulate 30% of mammalian genes
- Described in invertebrates and vertebrates: worms, fungi, plants, and mammals
- Many are conserved between vertebrates and invertebrates

MicroRNA (miRNA)

MicroRNA (miRNA)

- miRNA originates from ssRNA that forms a hairpin secondary structure
- miRNA regulates post-transcriptional gene expression and is often not 100% complementary to the target

MicroRNA (miRNA)

MicroRNA (miRNA)

- Originate from capped & polyadenylated full length precursors (pri-miRNA)
- Hairpin precursor ~70 nt (pre-miRNA)
- Mature miRNA ~22 nt (miRNA)

MicroRNA (miRNA)

MicroRNA (miRNA)

- Drosha and Pasha are part of the "Microprocessor" protein complex (~600 - 650kDa)
- Drosha and Dicer are RNase III enzymes

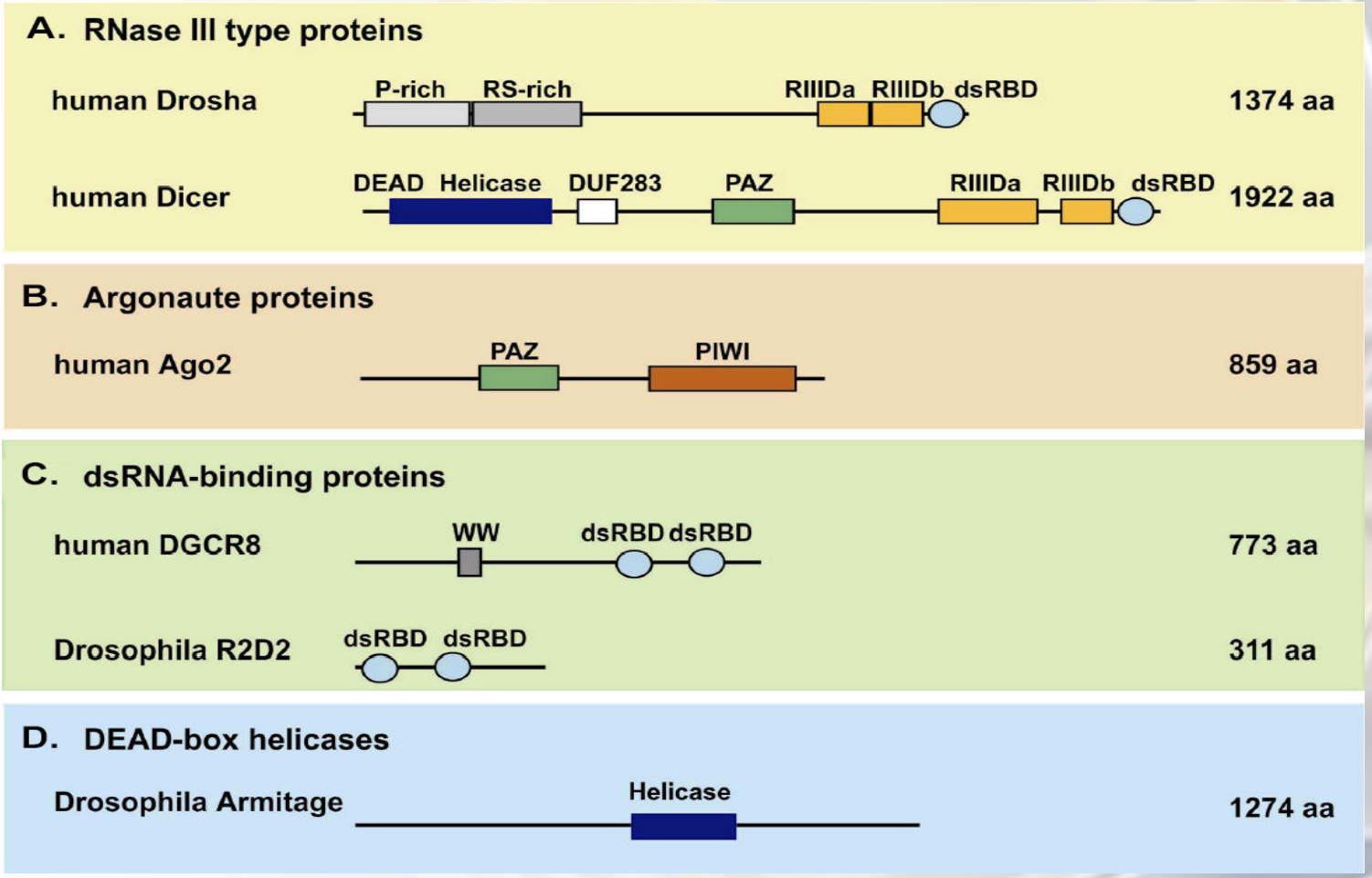
MicroRNA (miRNA)

MicroRNA (miRNA)

- Pasha is a dsRNA binding protein
- Exportin 5 is a member of the karyopherin nucleocytoplasmic transport factors
- Argonautes are RNase H enzymes

MicroRNA (miRNA)

MicroRNA (miRNA)



MicroRNA (miRNA)

MicroRNA (miRNA)

- ~1900 discovered in Homo sapiens
- Mouse and human have highly conserved miRNA genes

MicroRNA (miRNA)

MicroRNA (miRNA) Conclusion

- Small non-coding double stranded RNAs
- Approximately 19-22 nucleotides long
- 30% of mammalian mRNAs are regulated by miRNAs

Biogenesis of MicroRNA (miRNA)

Biogenesis of MicroRNA (miRNA)

MicroRNA (miRNA)

- Small non-coding double stranded RNAs
- Approximately 19-22 nucleotides long
- miRNA originates with ssRNA that forms a hairpin secondary structure

Biogenesis of MicroRNA (miRNA)

MicroRNA (miRNA)

- Originate from capped & polyadenylated full length precursors (pri-miRNA)

Biogenesis of MicroRNA (miRNA)

Biogenesis of miRNA

- Primary-miRNA is transcribed in the nucleus, and is usually several kilobases long; having 5’ cap and a poly-A tail.
- Cleaved in the nucleus by Drocha enzyme to 70nt hairpin transcript (pre-miRNA).
- Transported to the cytoplasm by Exportin 5 through nuclear pores.
- Cleaved by Dicer enzyme (RNase III enzyme) into 19-22nt ds-transcripts.

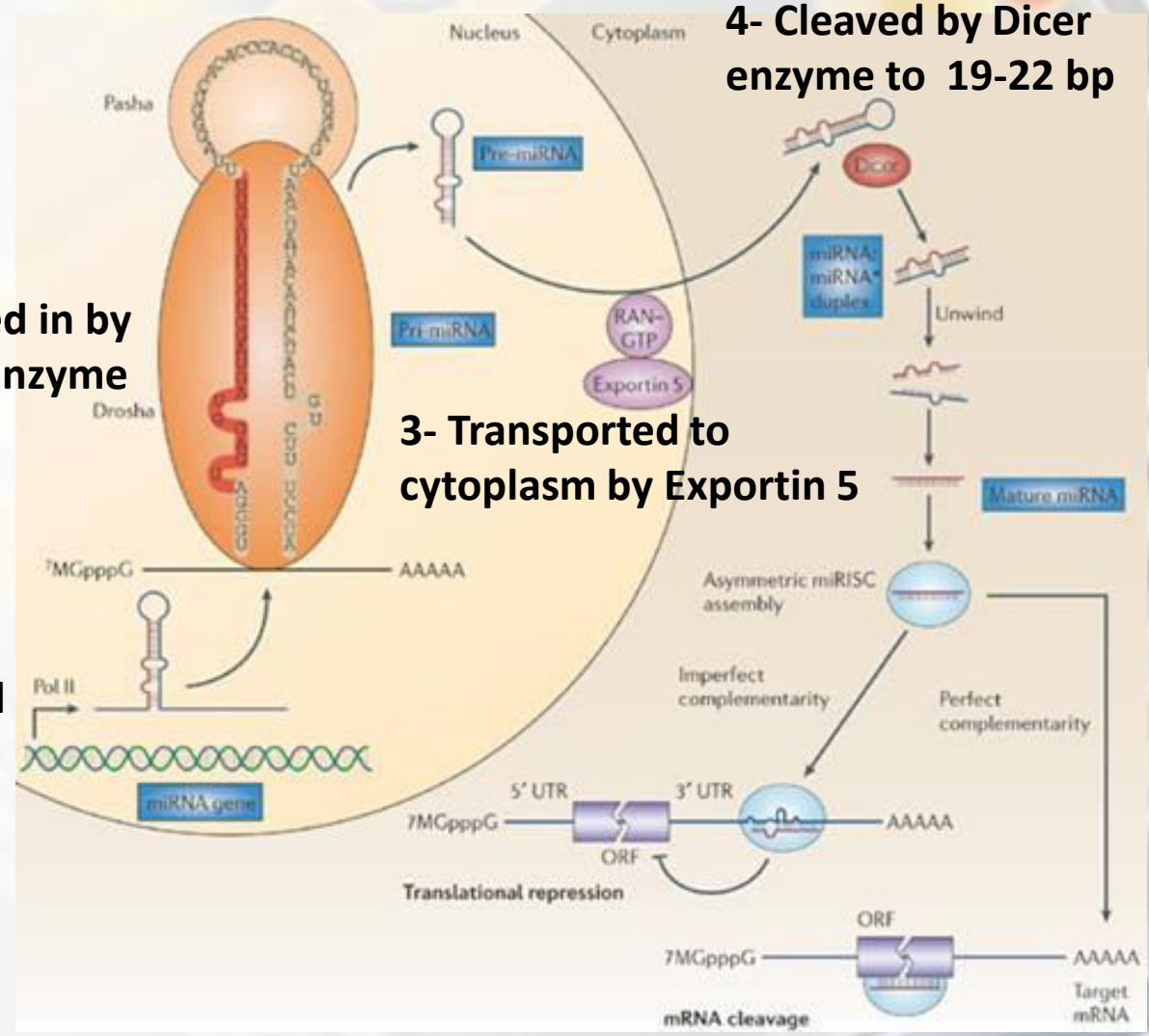
Biogenesis of MicroRNA (miRNA)

1- Transcribed in nucleus

2- Cleaved in by Drocha enzyme

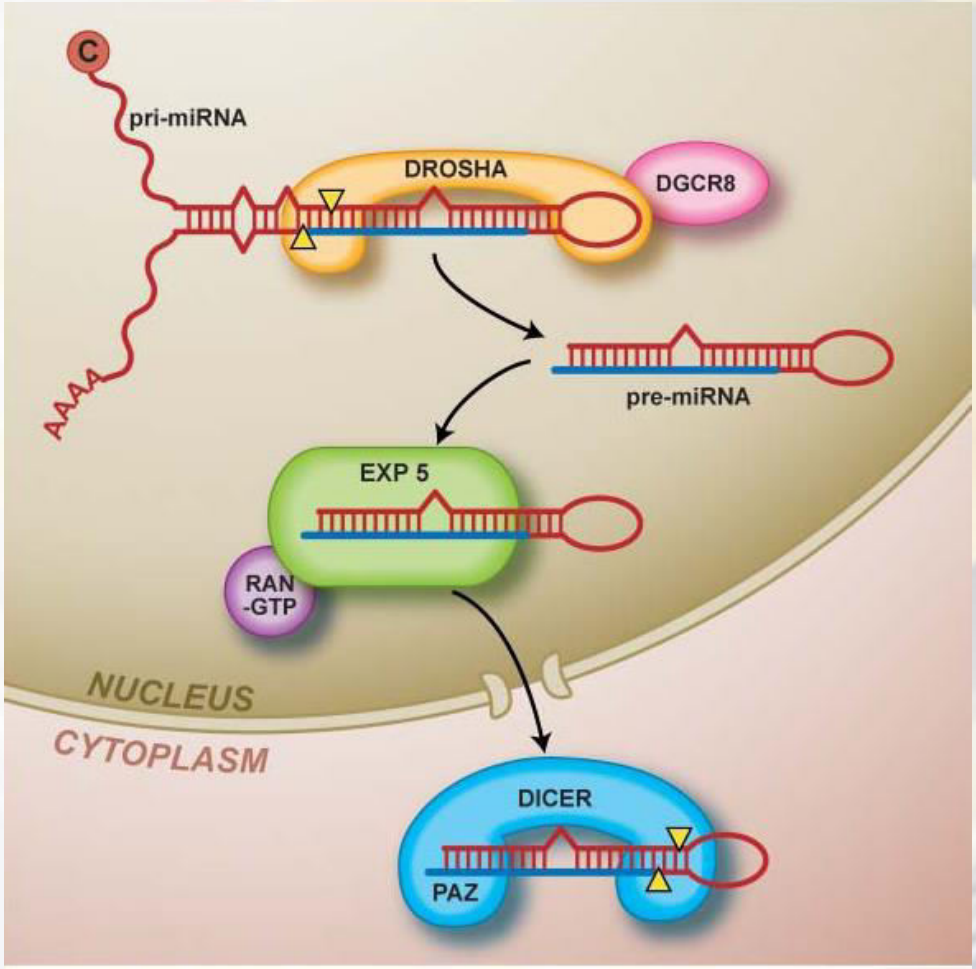
3- Transported to cytoplasm by Exportin 5

4- Cleaved by Dicer enzyme to 19-22 bp



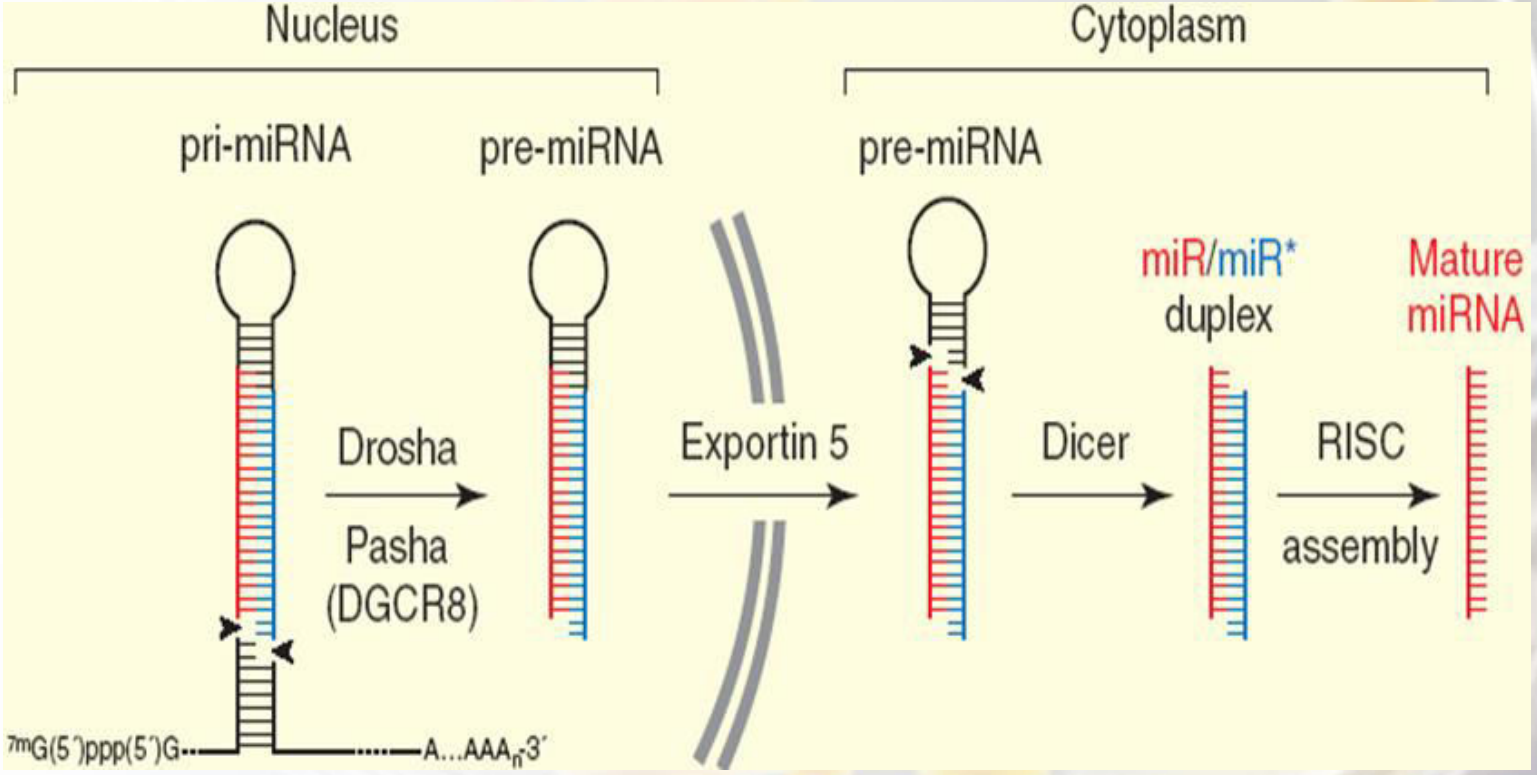
Biogenesis of MicroRNA (miRNA)

Illustration of miRNA Processing



Biogenesis of MicroRNA (miRNA)

Illustration of miRNA Processing - Another View



Biogenesis of MicroRNA (miRNA)

MicroRNA (miRNA)

- Small non-coding double stranded RNAs
- Approximately 19-22 nucleotides long

Functions of miRNA

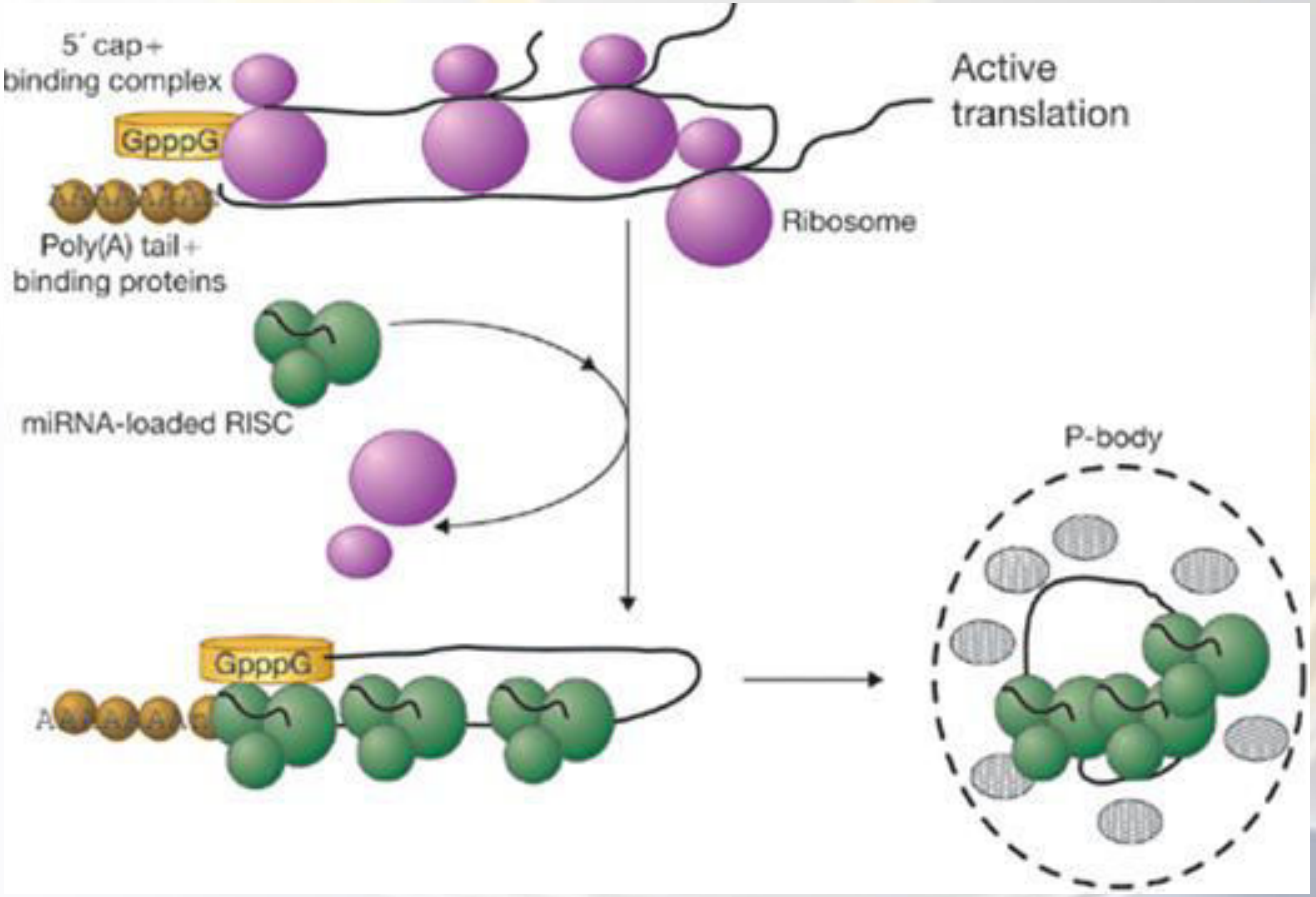
Functions of miRNA

Functions of miRNA

- Involved in the post-transcriptional regulation of gene expression
- Important in development
- Metabolic regulation (miR-375 & insulin secretion)

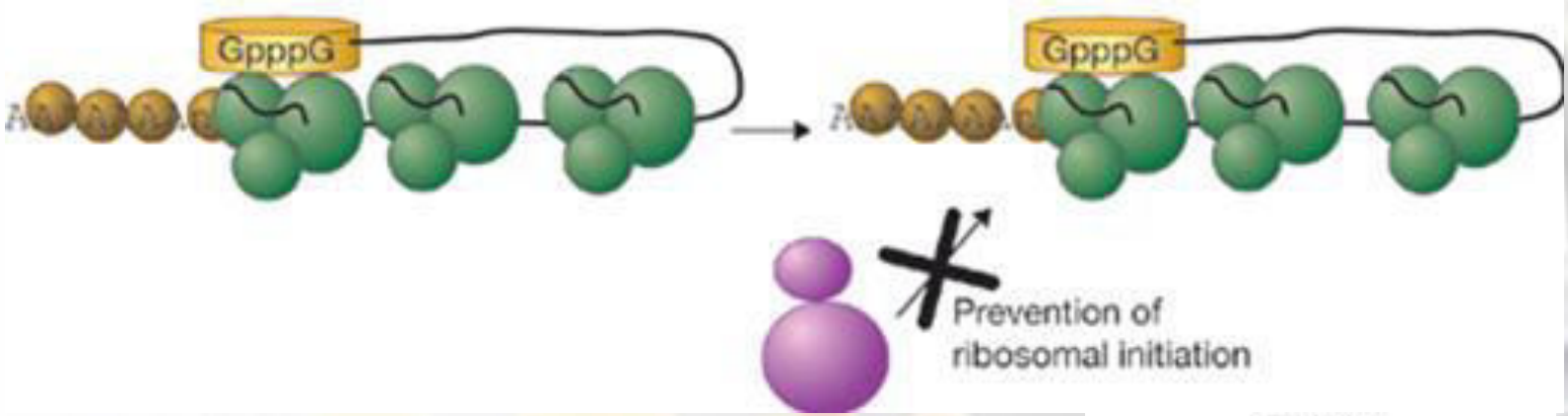
Functions of miRNA

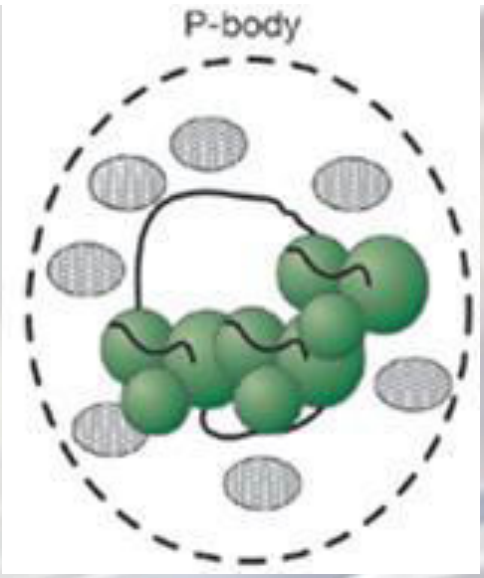
Functions of miRNA: Processing bodies are sites of storage and/or degradation of mRNA



Functions of miRNA

Functions of miRNA





Functions of miRNA

Functions of miRNA

miRNA	Target	Function
miR-15/miR-16	Bcl2	Apoptosis
miR-1	GJA1/KCNJ2	Cardiac Arrhythmia
miR-146	IRAK1/TRAF6	Bacterial Infectious Response; TLR-NFκB
miR-520h	ABCG2	Stem Cell Differentiation
miR-106a	Rb1	Cancer Pathogenesis
miR-let7	Multiple	Cell Cycle Regulation
miR-155	-	Adaptive Immunity
miR-223	-	Granulocyte Regulation
miR-208	-	Stress Response (Heart)

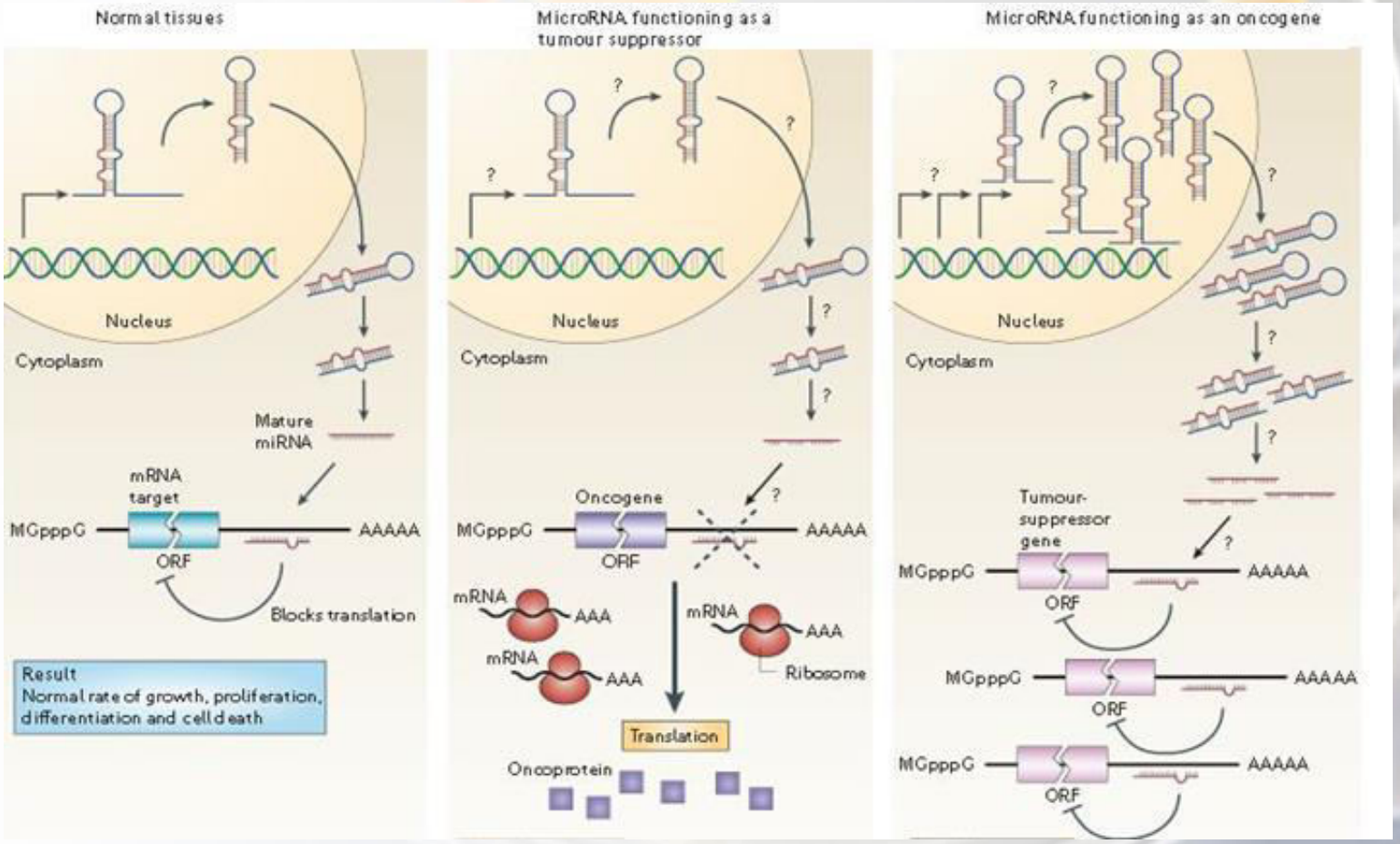
Functions of miRNA

Tumor Suppressors and Oncogenes

- miRNAs can function as tumour suppressors and oncogenes.
- Gene therapies that use miRNAs might be an effective approach to blocking tumour progression.
- miRNAs such as let-7, which has been shown to negatively regulate the Ras oncogenes, and miR-15 and miR-16, which negatively regulate BCL2, are promising candidates for cancer treatment.

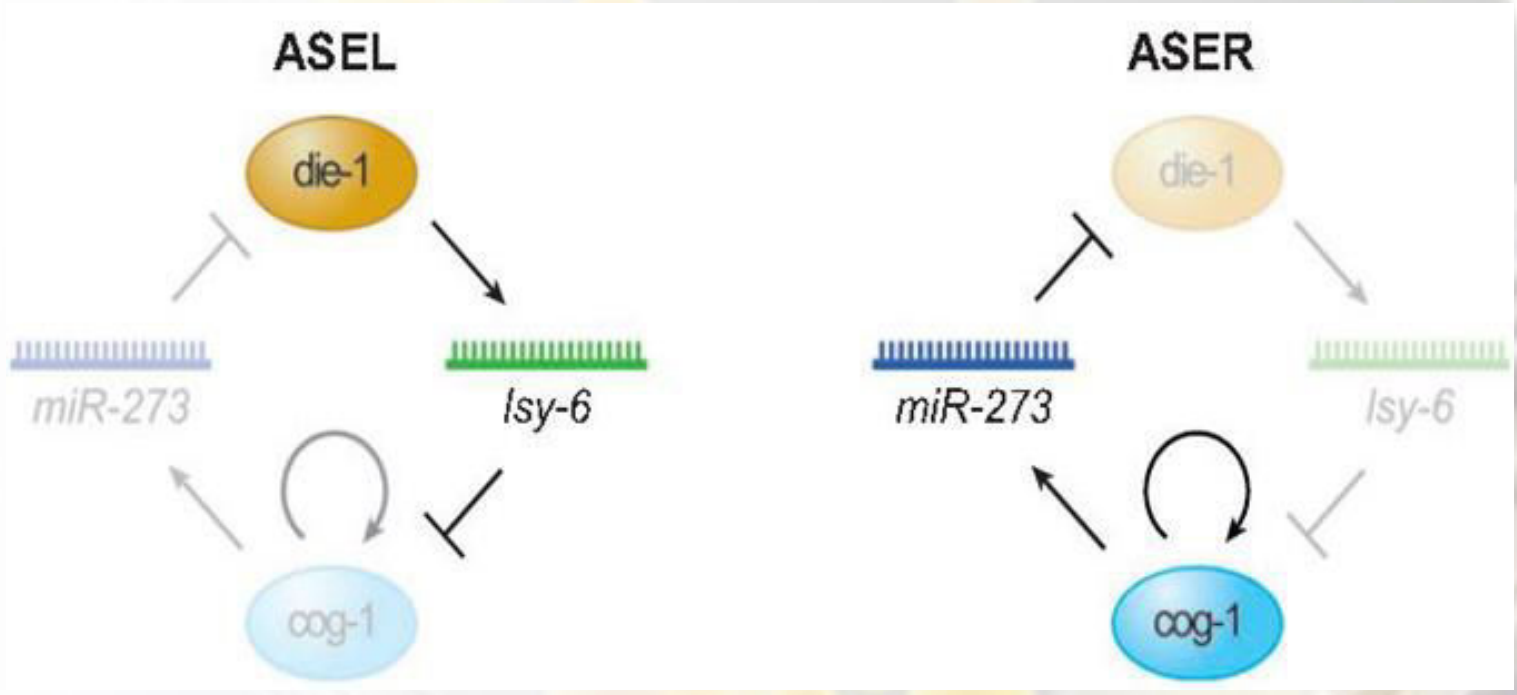
Functions of miRNA

Tumor Suppressors and Oncogenes



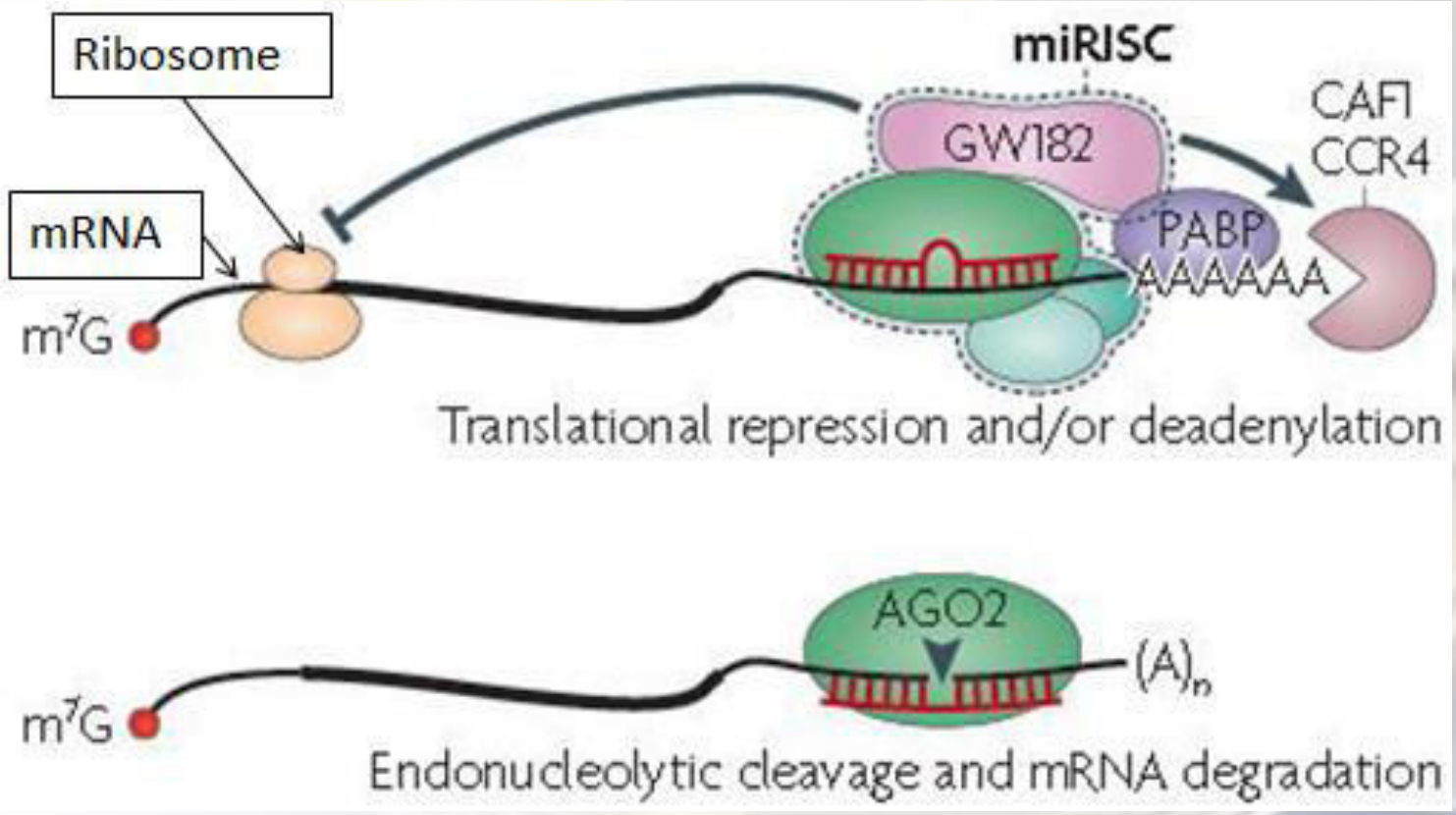
Functions of miRNA

Function as Developmental Switches



Functions of miRNA

Functions of miRNA: Gene Silencing in Plants and Animals



Functions of miRNA

Functions of miRNA

- Involved in the post-transcriptional/translational regulation of gene expression
- Metabolic regulation

Genetics and Genomics

miRNA
Mode of Action

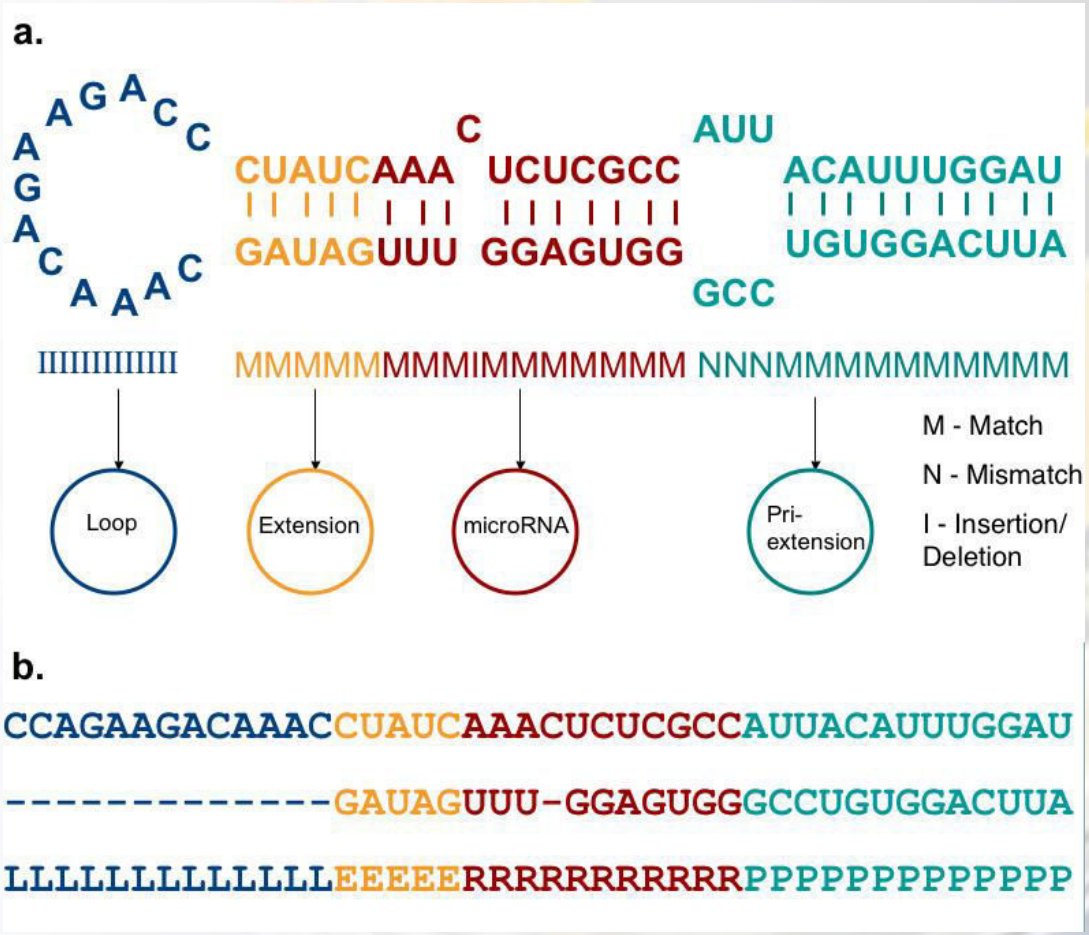
miRNA - Mode of Action

Mode of Action of miRNA

- Three different mode of action of miRNA
- mRNA degradation
- Transcriptional regulation
- Translational regulation

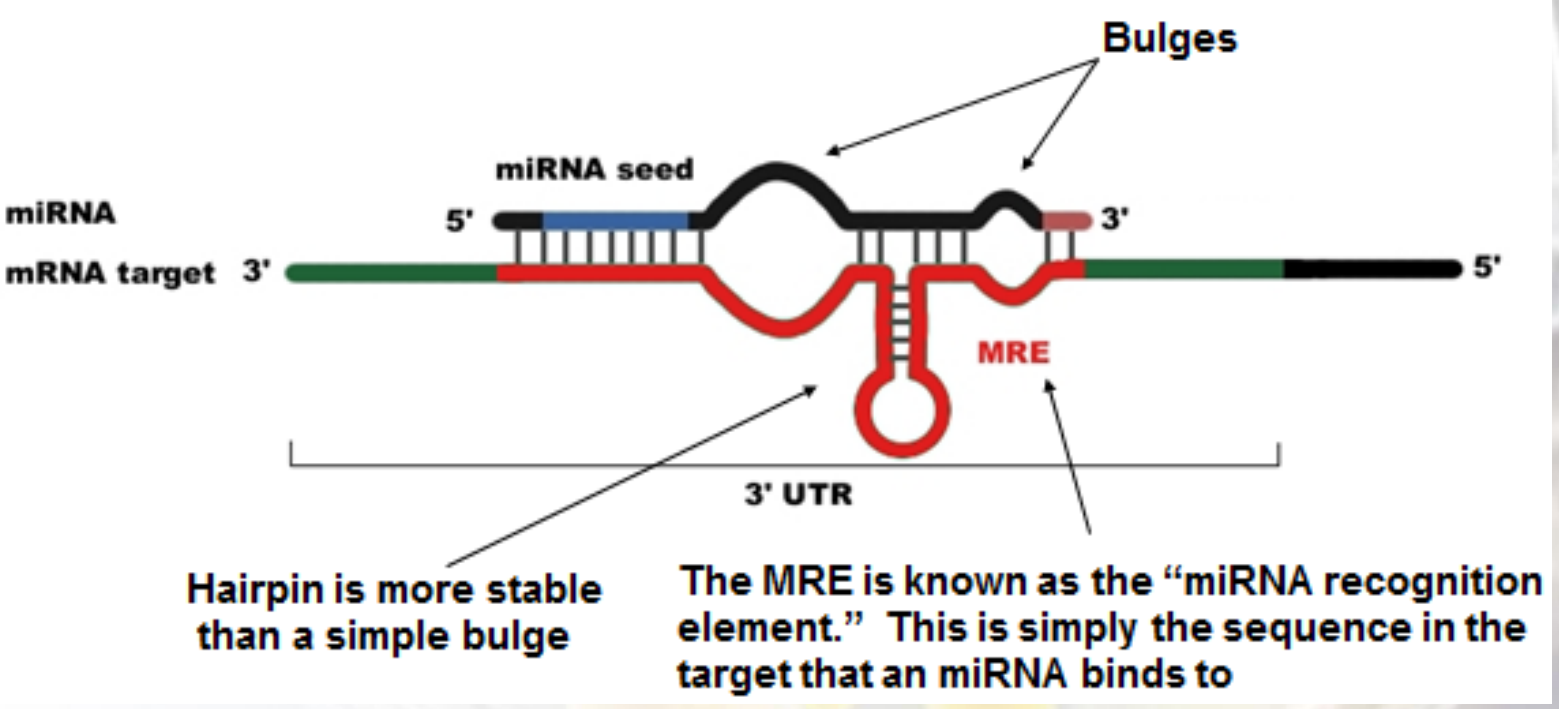
miRNA - Mode of Action

Structure of miRNA at Nucleotide level



miRNA - Mode of Action

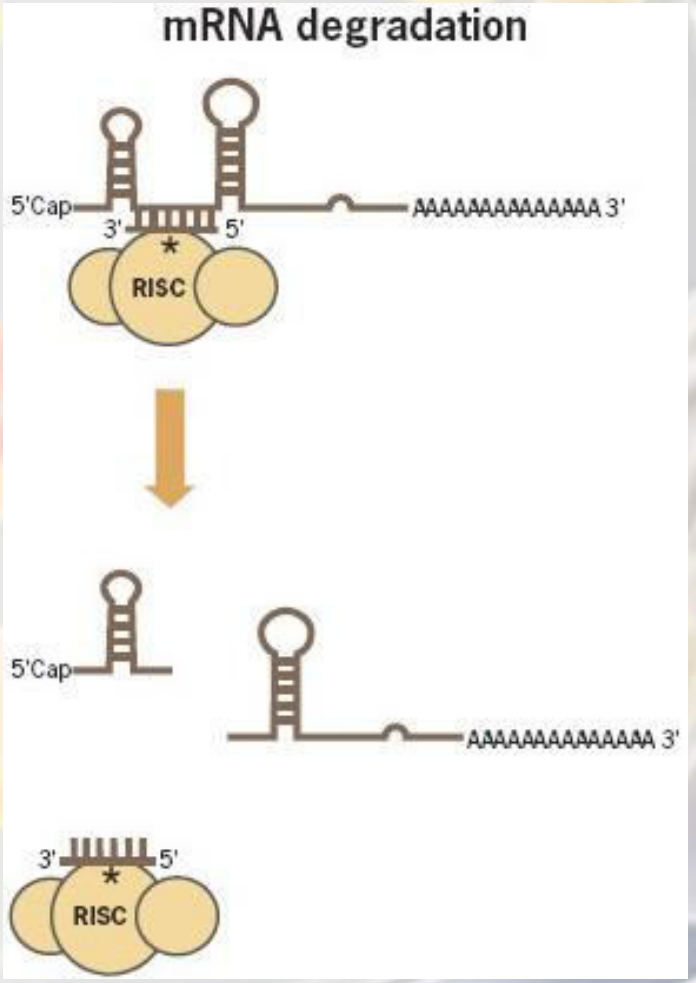
miRNA Binding to Target (mRNA)



miRNA - Mode of Action

Mode of Action of miRNA - mRNA Degradation

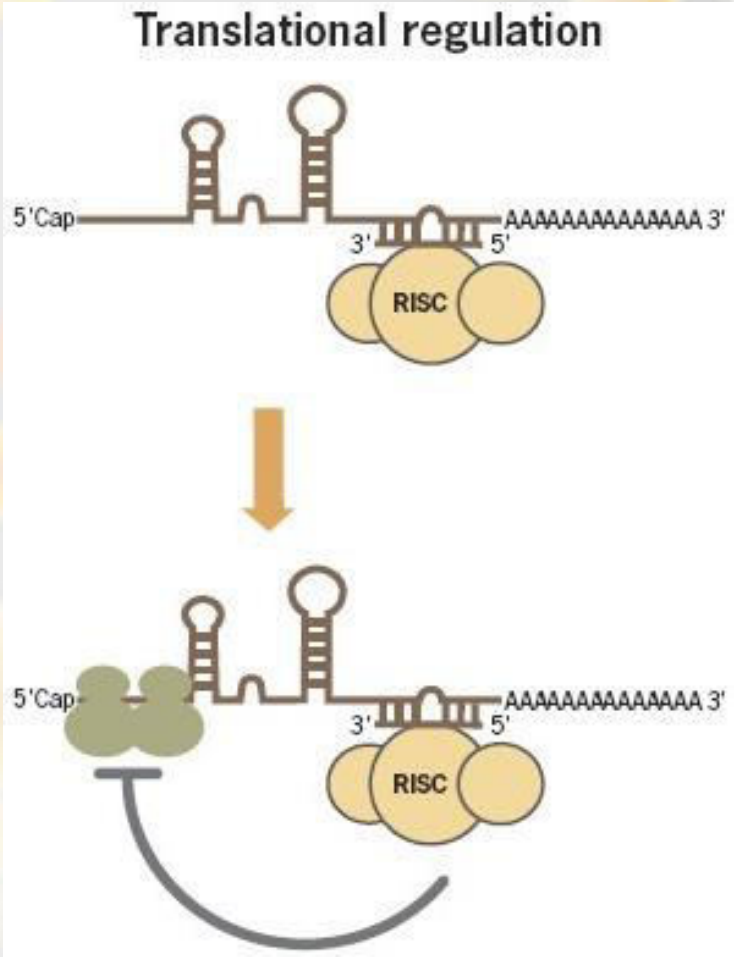
- Common in Plants



miRNA - Mode of Action

Mode of Action - Translational Regulation

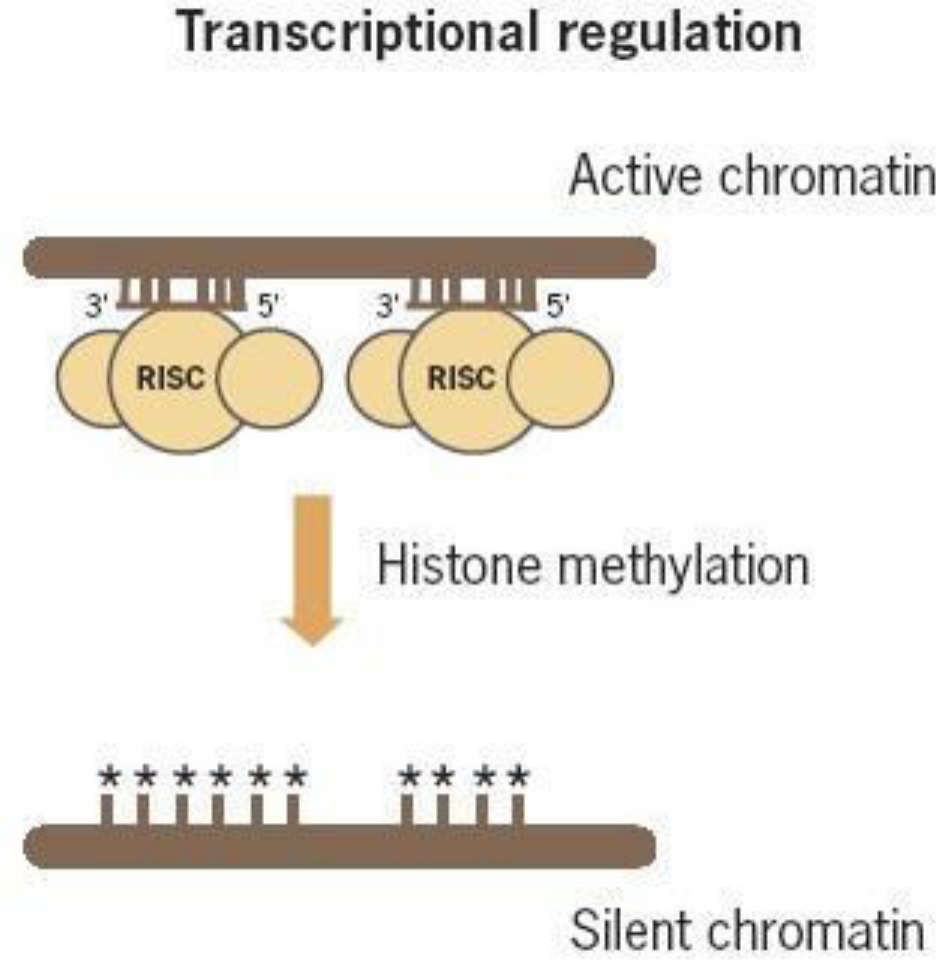
- **Common in Animals**



miRNA - Mode of Action

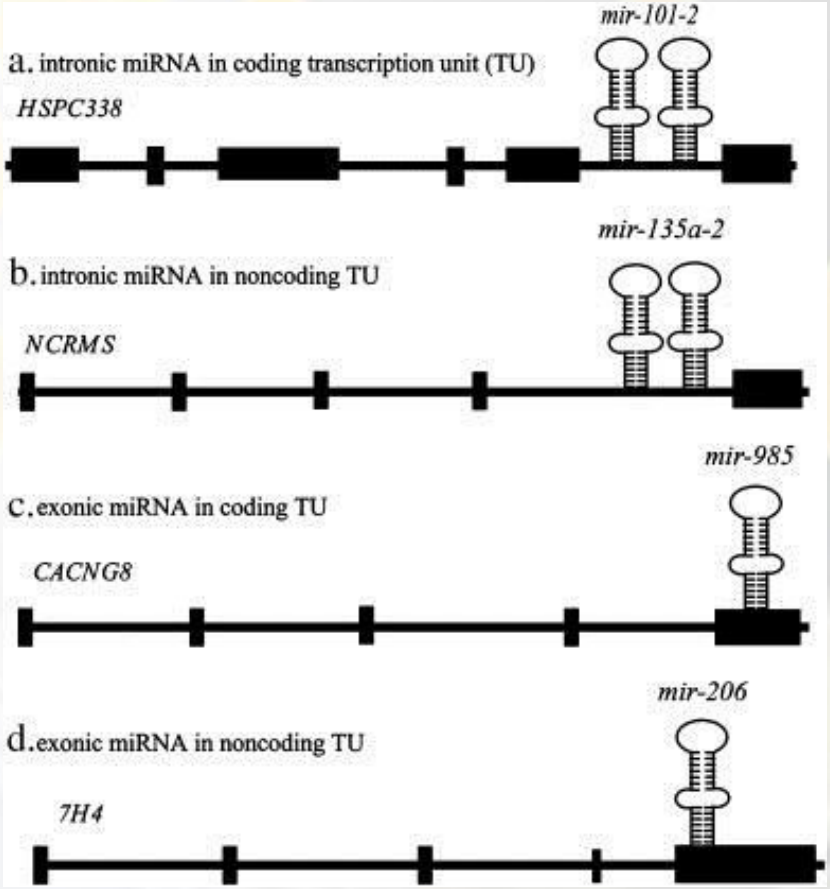
Mode of Action – Transcriptional Regulation

- Common in Yeast and Plants
- Sometimes in Animals



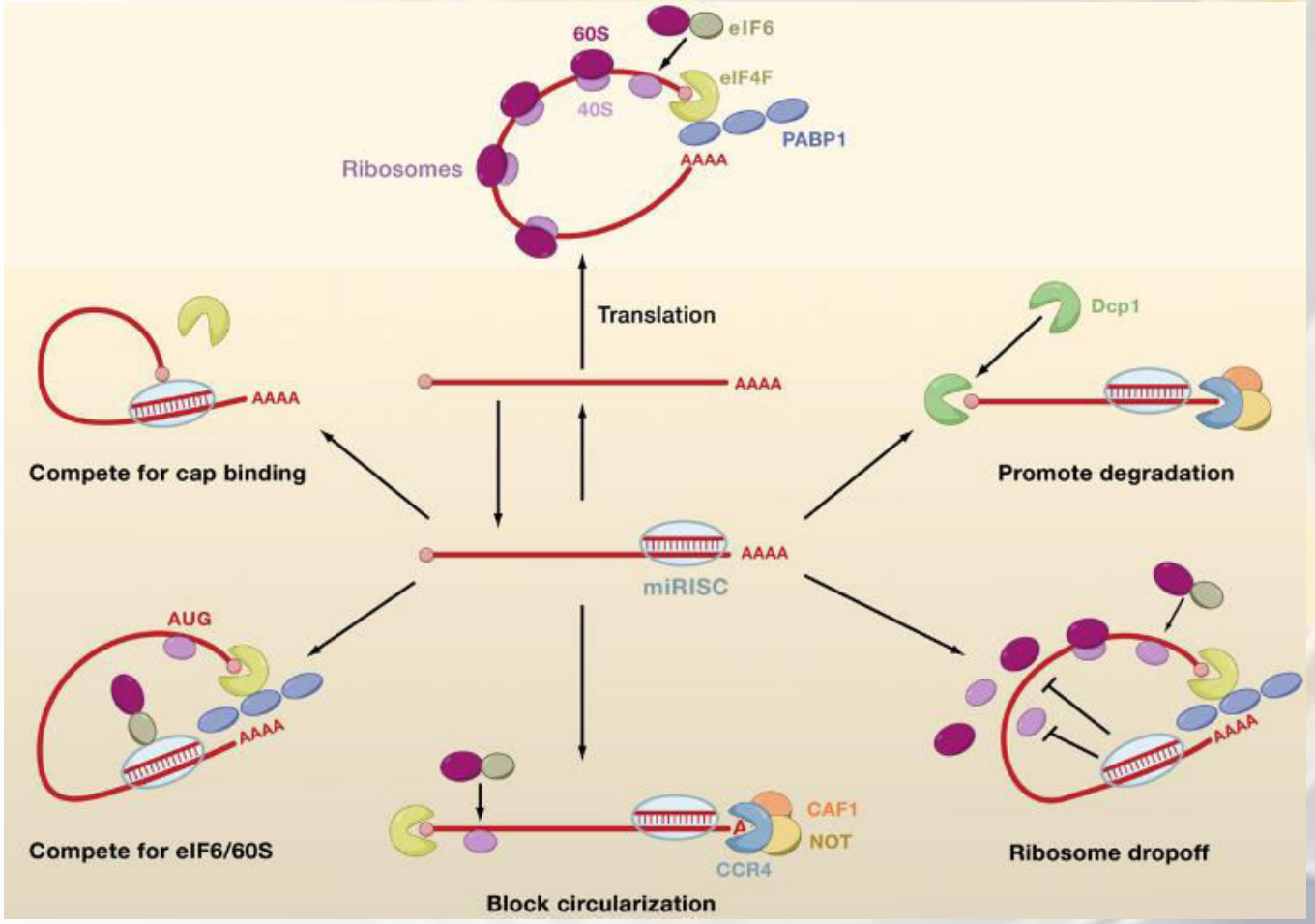
miRNA - Mode of Action

Schematic Illustration of the Genomic Organization and Structure of miRNA Genes



miRNA - Mode of Action

Mechanisms of miRISC-Mediated Repression



miRNA - Mode of Action

Mode of action of miRNA

- Three different mode of action of miRNA
- mRNA degradation
- Transcriptional regulation
- Translational regulation

Non-coding RNA and Silencing of X Chromosome

Non-coding RNA- X Chromosome Silencing

Silencing of one X chromosome

- X chromosome silencing is mediated by Xist – a 16,000 nt long ncRNA
- Xist ncRNA recruited complex has one entry site in X chromosome, corresponding to Xist gene itself

Non-coding RNA- X Chromosome Silencing

Silencing of one X chromosome

- Xist appears to recruit a specific histone isoform which maintains the chromosome in inactive state

Non-coding RNA- X Chromosome Silencing

Silencing of one X chromosome

- Additionally, Xist containing complexes recruit histone deacetylases and methylases

Non-coding RNA- X Chromosome Silencing

Silencing of one X chromosome

- Xist activity is regulated by another 40,000 nt long ncRNA – Tsix, which contains anti-sense sequence of Xist and therefore is able to regulate Xist activity by base-pairing to it

Non-coding RNA- X Chromosome Silencing

shRNA and Genetic Imprinting

- Activity of small heterochromatic RNAs (shRNAs) appear to be essential for establishing and maintaining the imprinted status of genes

RNA Induced Silencing Complex

RNA Induced Silencing Complex

RNA Induced Silencing Complex

- RISC is a large (~500-kDa) RNA multi-protein complex, which triggers mRNA degradation in response to siRNA

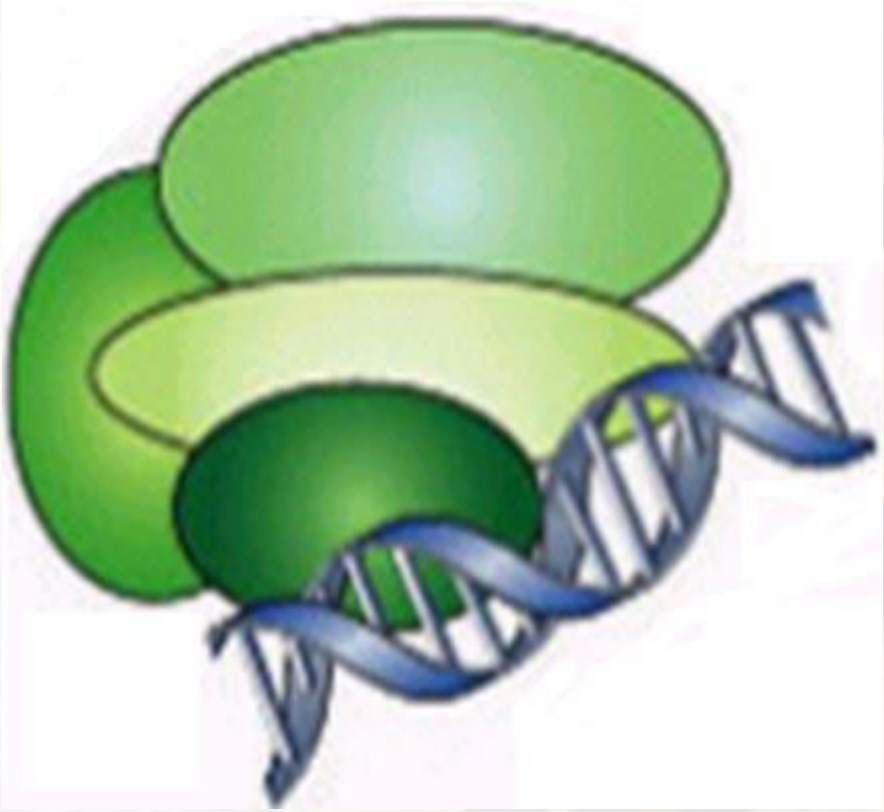
RNA Induced Silencing Complex

RISC

- Unwinding of double-stranded siRNA by ATP independent helicase.
- Active components of RISC are endonucleases called argonaute proteins which cleave the target mRNA strand.

RNA Induced Silencing Complex

RNA Induced Silencing Complex - Structure



RNA Induced Silencing Complex

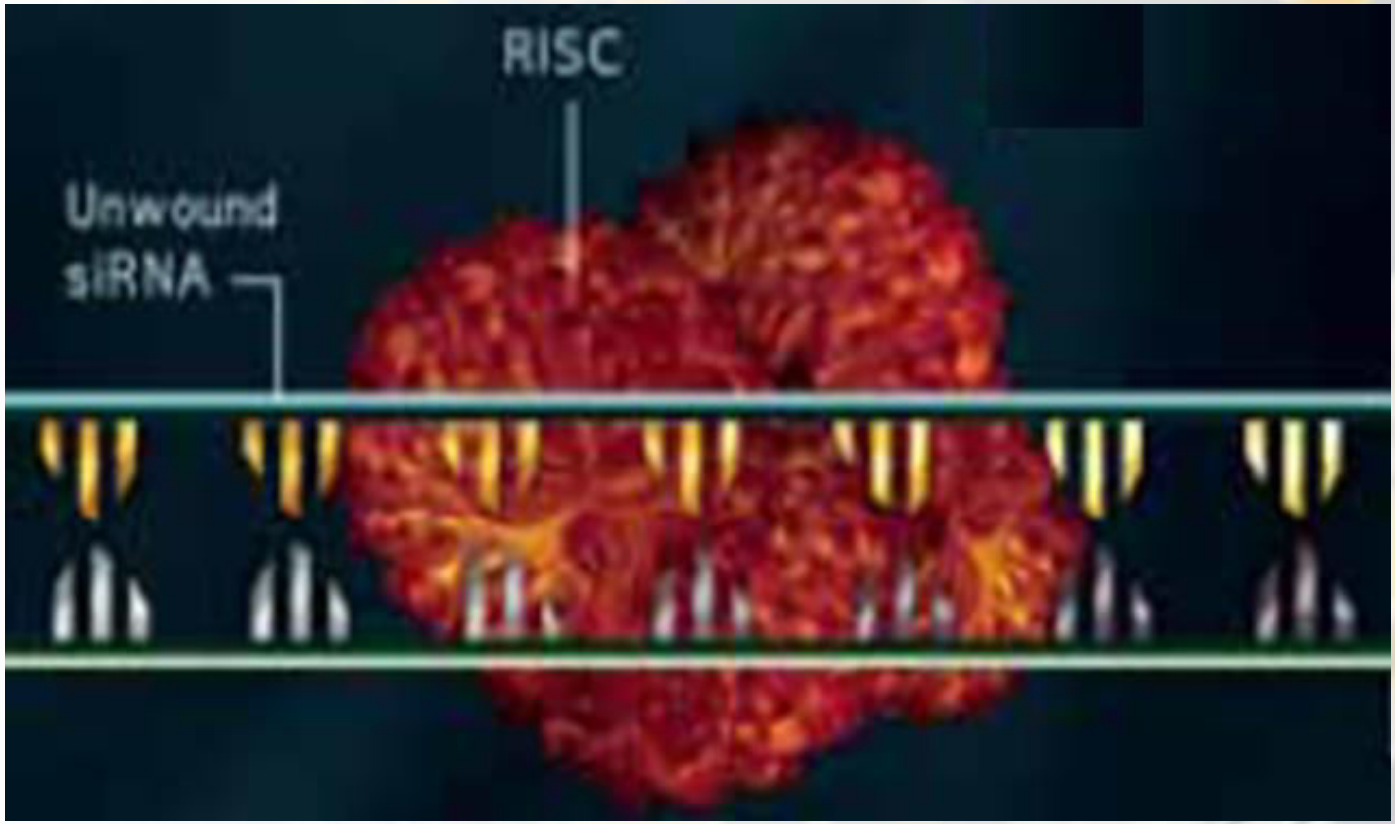
RNA Induced Silencing Complex

RISC is a large (~500-kDa) RNA-multiprotein complex, which triggers mRNA degradation

Some components have been defined by genetics, but function is unknown, e.g. - unwinding of double-stranded siRNA (Helicase) ribonuclease component cleaves mRNA (Nuclease).

RNA Induced Silencing Complex

RNA Induced Silencing Complex



RNA Induced Silencing Complex

RNA Induced Silencing Complex

RNA Induced Silencing Complex

RNA Induced Silencing Complex

- RISC bound to partially complementary mRNA induces translational repression: results in destabilization of target mRNA
- The more RISC bound; greater the inhibitory effect
- RISC bound to perfectly complementary mRNA (to miRNA) are cleaved and degraded by RISC in a process similar to RNAi

RNA Induced Silencing Complex

RNA Induced Silencing Complex

- RISC is a large (~500-kDa) RNA-multiprotein complex, which triggers mRNA degradation in response to siRNA

Dicer and Drosha

Dicer and Drosha

Dicer

- Dicer or helicase with RNase motif, is an enzyme of RNase III family
- In humans, it is encoded by the DICER1 gene.

Dicer and Drosha

Dicer

- Dicer cleaves double-stranded RNA (dsRNA) and pre-microRNA into short double-stranded RNA fragments called small interfering RNA and microRNA

Dicer and Drosha

Dicer

- It is able to digest dsRNA into uniformly sized small RNAs (siRNA)
- Dicer family proteins are ATP-dependent nucleases.
- RNase III enzyme acts as a dimer

Dicer and Drosha

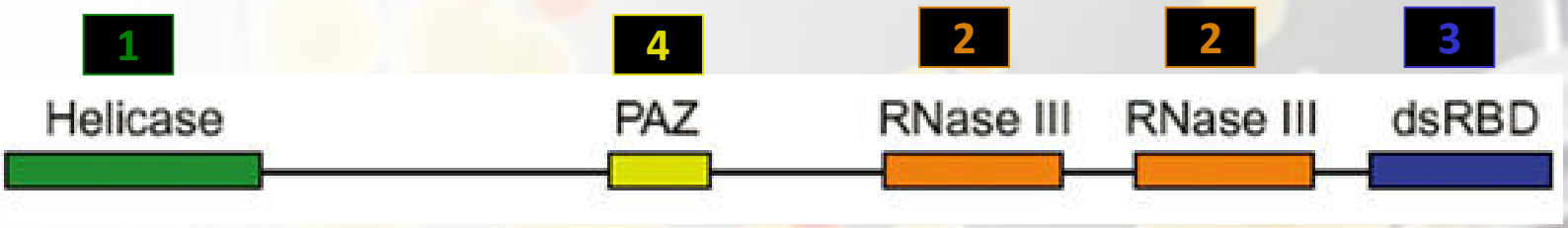
Dicer

- Dicer homologs exist in many organisms including C. elegans, Drosphila, yeast and humans

Dicer and Drosha

Dicer:

Dicer is a Ribonuclease (RNase III family) with 4 Distinct Domains



Amino-terminal helicase domain
Dual RNase III motifs in the carboxy terminal segment
dsRNA binding domain
PAZ domain - 110-130 amino acids

Dicer and Drosha

Drosha

- Drosha is a class 2 ribonuclease III enzyme that in humans is encoded by the gene DROSHA (formerly RNASEN) gene

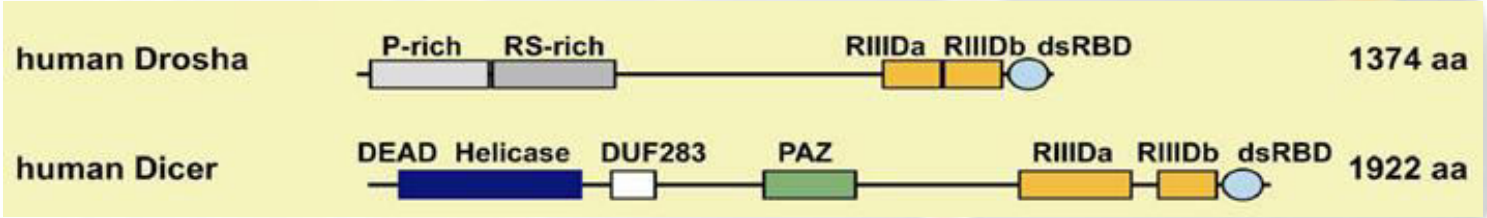
Dicer and Drosha

Drosha

- The RNase III drosha is the core nuclease that executes the initiation step of microRNA (miRNA) processing in the nucleus

Dicer and Drosha

Dicer and Drosha - Humans



Dicer and Drosha

Dicer and Drosha

- Dicer is an enzyme that is part of the RNase III family
- Drosha belongs to class 2 ribonuclease III enzyme family

Small Interfering RNA (siRNA)

Small Interfering RNA (siRNA)

siRNA

- Small interfering RNA (siRNA), sometimes known as short interfering RNA, are a class 20-25 nucleotide-long RNA molecules that interfere with the expression of genes.

Small Interfering RNA (siRNA)

siRNA

- They are naturally produced as part of the RNA interference (RNAi) pathway by the enzyme Dicer.
- Exogenously introduced by investigators to bring about knockdown of a particular gene.

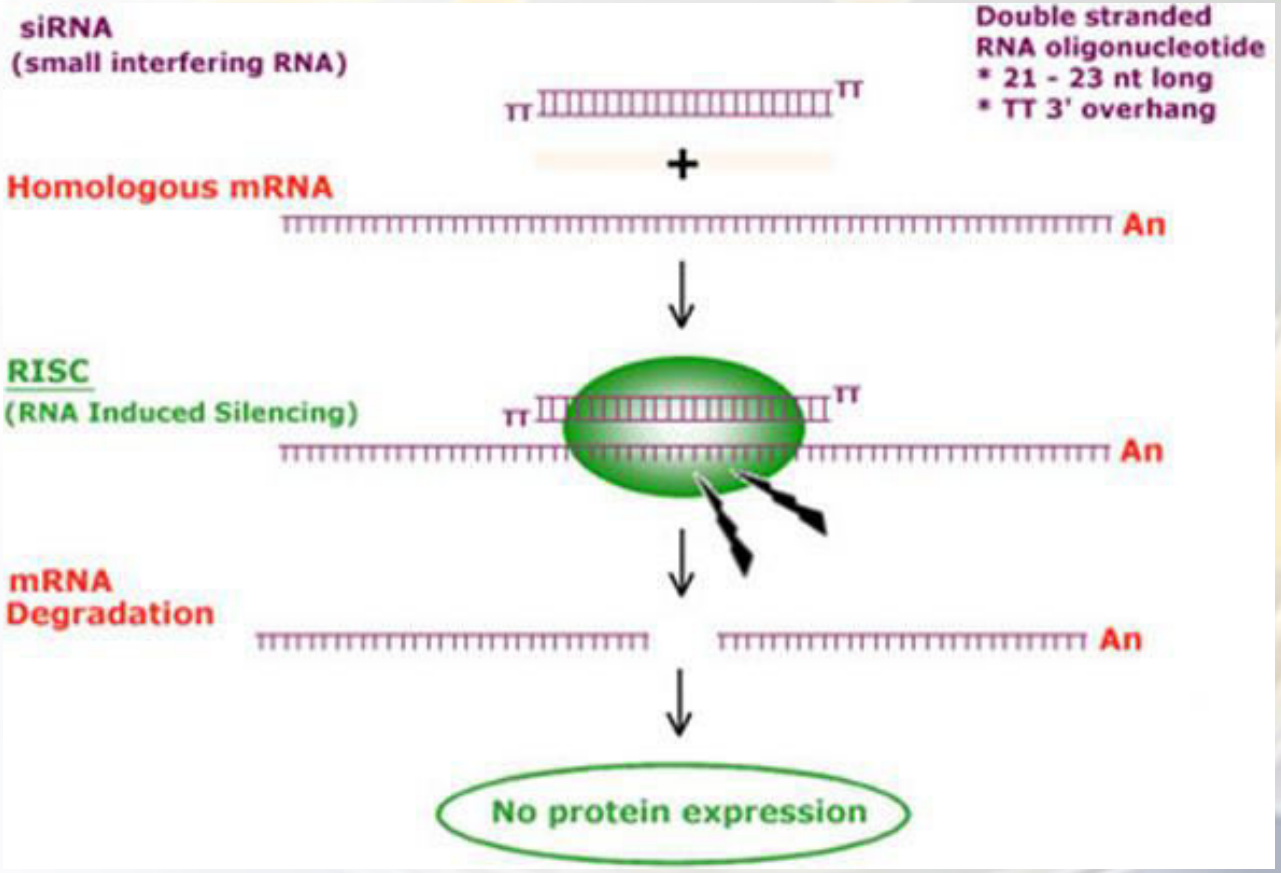
Small Interfering RNA (siRNA)

siRNA

- siRNA's have a well defined structure.
- A short (usually 21-nt) double-strand of RNA (dsRNA) with 2-nt overhangs on either end, including a 5' phosphate group and a 3' hydroxy (-OH) group.

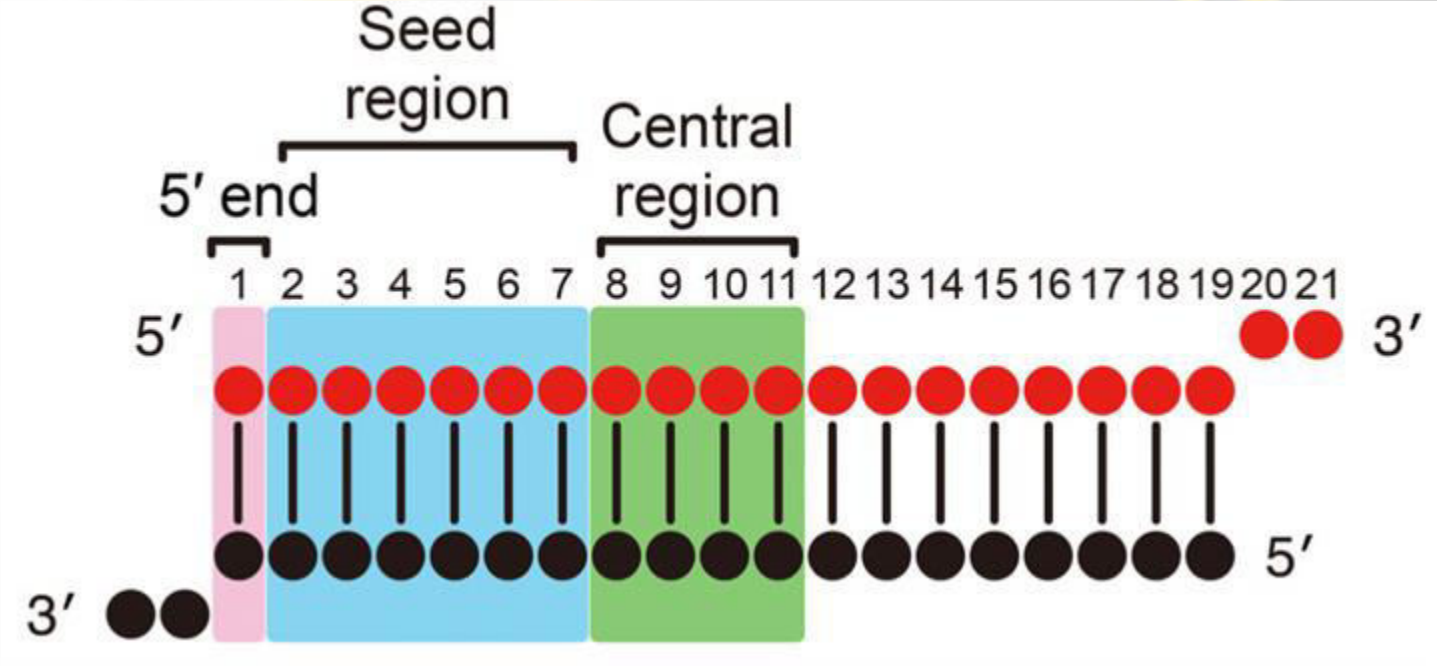
Small Interfering RNA (siRNA)

siRNA: short (usually 21-nt) double-strand of RNA (dsRNA) with 2-nt overhangs on either end.



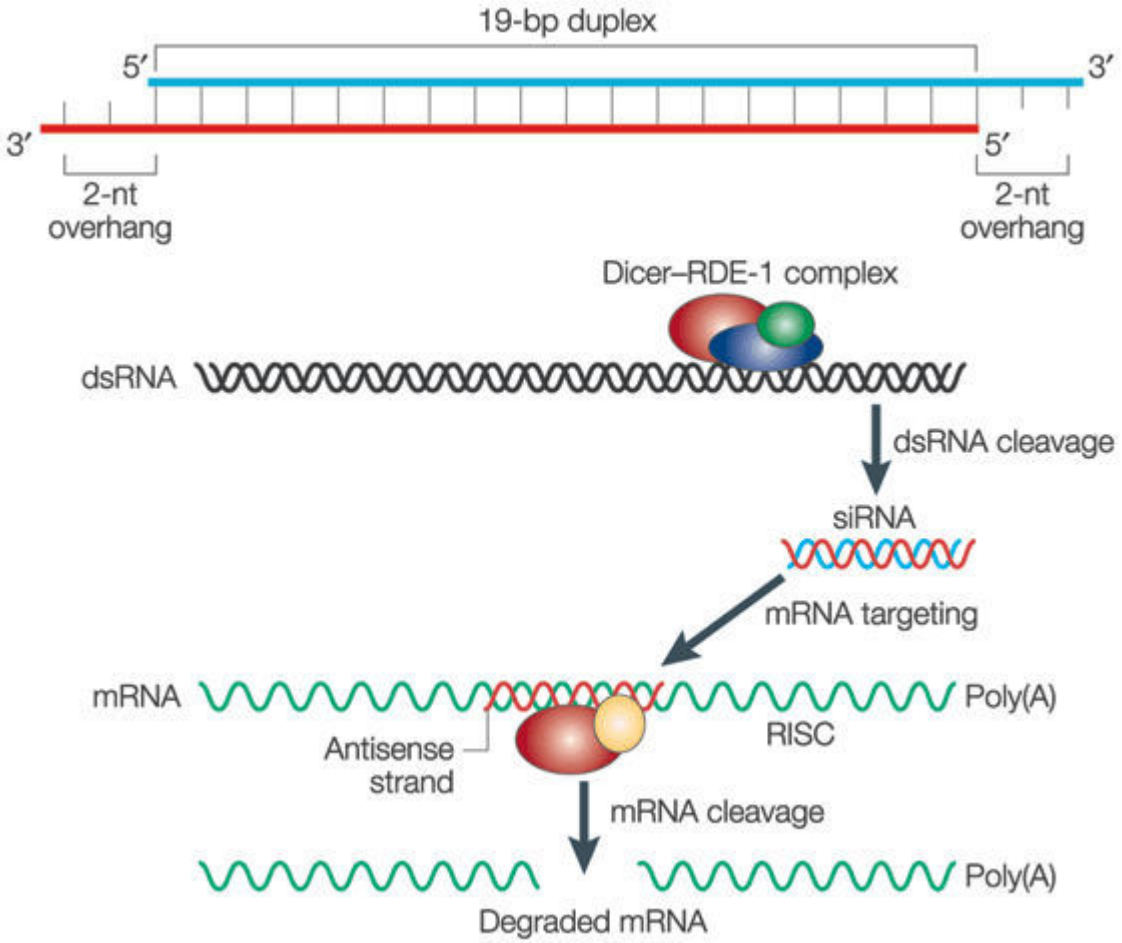
Small Interfering RNA (siRNA)

siRNA: Short double-strand of RNA (dsRNA) with 2-nt overhangs on either end.



Small Interfering RNA (siRNA)

siRNA:



Small Interfering RNA (siRNA)

siRNA

- Small interfering RNAs that have an integral role in the phenomenon of RNA interference(RNAi),a form of post-transcriptional gene silencing

Small Interfering RNA (siRNA)

siRNA

- A single base pair difference between the siRNA template and the target mRNA is enough to block the process.

Small Interfering RNA (siRNA)

Small interfering RNA

- Several different methods of expression
- Several different methods of delivery

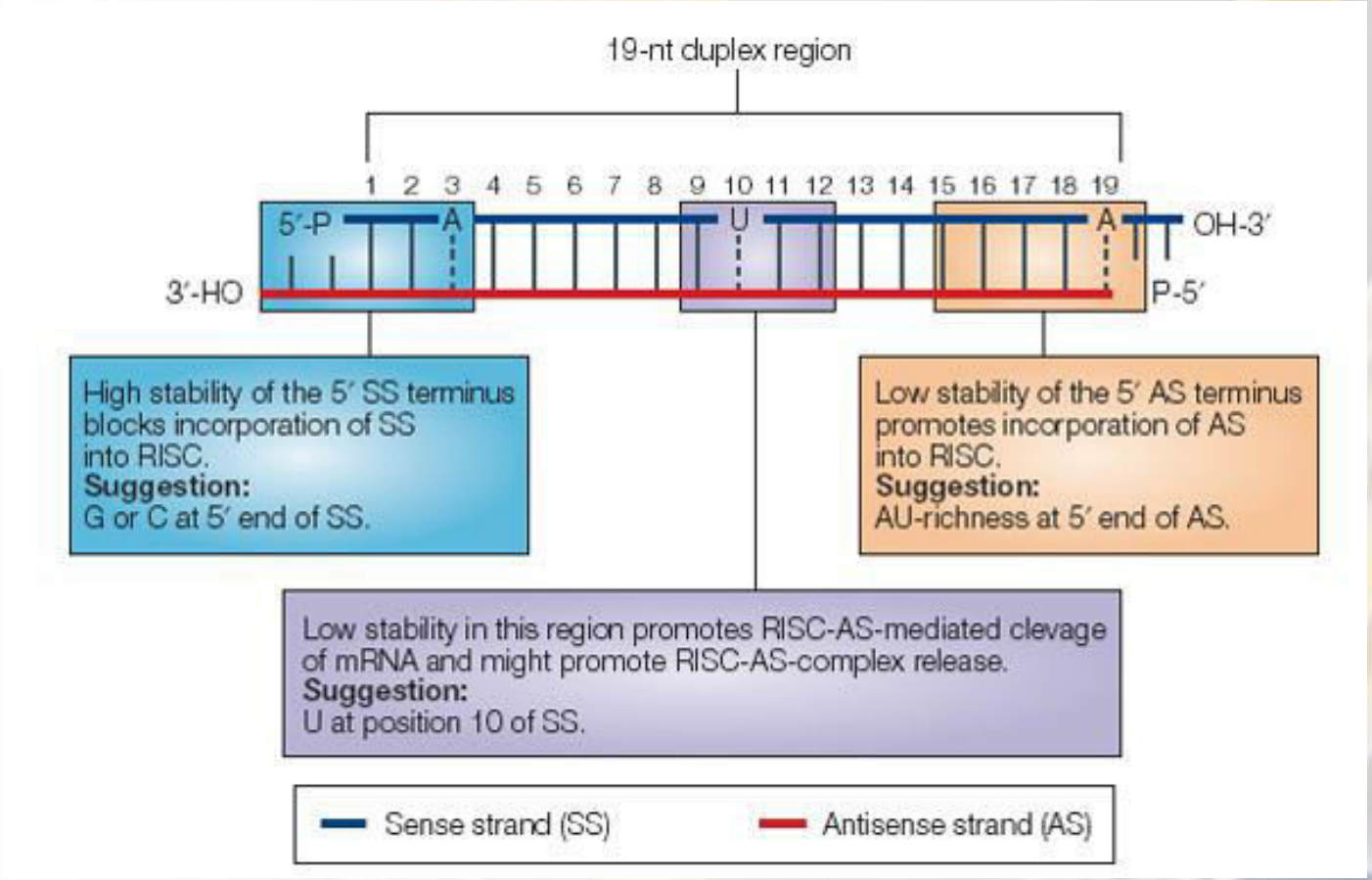
Small Interfering RNA (siRNA)

Small interfering RNA

- Dicer (type III RNAse III) cleaves long dsRNA into siRNA 21-25nt dsRNA from exogenous sources

Small Interfering RNA (siRNA)

Small interfering RNA: siRNA design



Small Interfering RNA (siRNA)

siRNA

- Small interfering RNA (siRNA), sometimes known as short interfering RNA, are a class 20-25 nucleotide-long RNA molecules that interfere with the expression of genes.

Design of siRNA

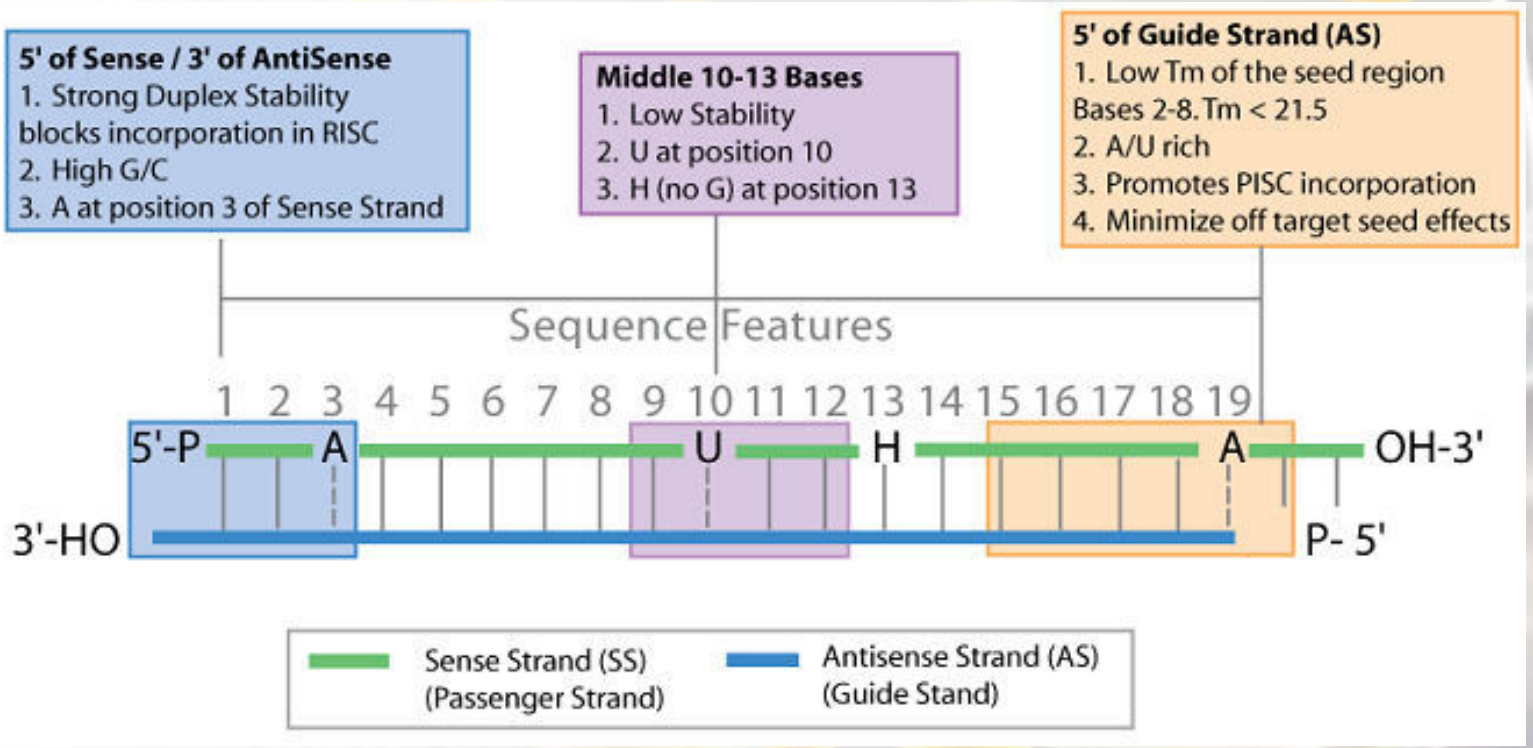
Design of siRNA

Design of siRNA

- 21-23 nt dsRNA,
- GC% slightly < 50%
- Perfect complimentary to target mRNA
- Targeting 3’ UTR works better than 5’ UTR

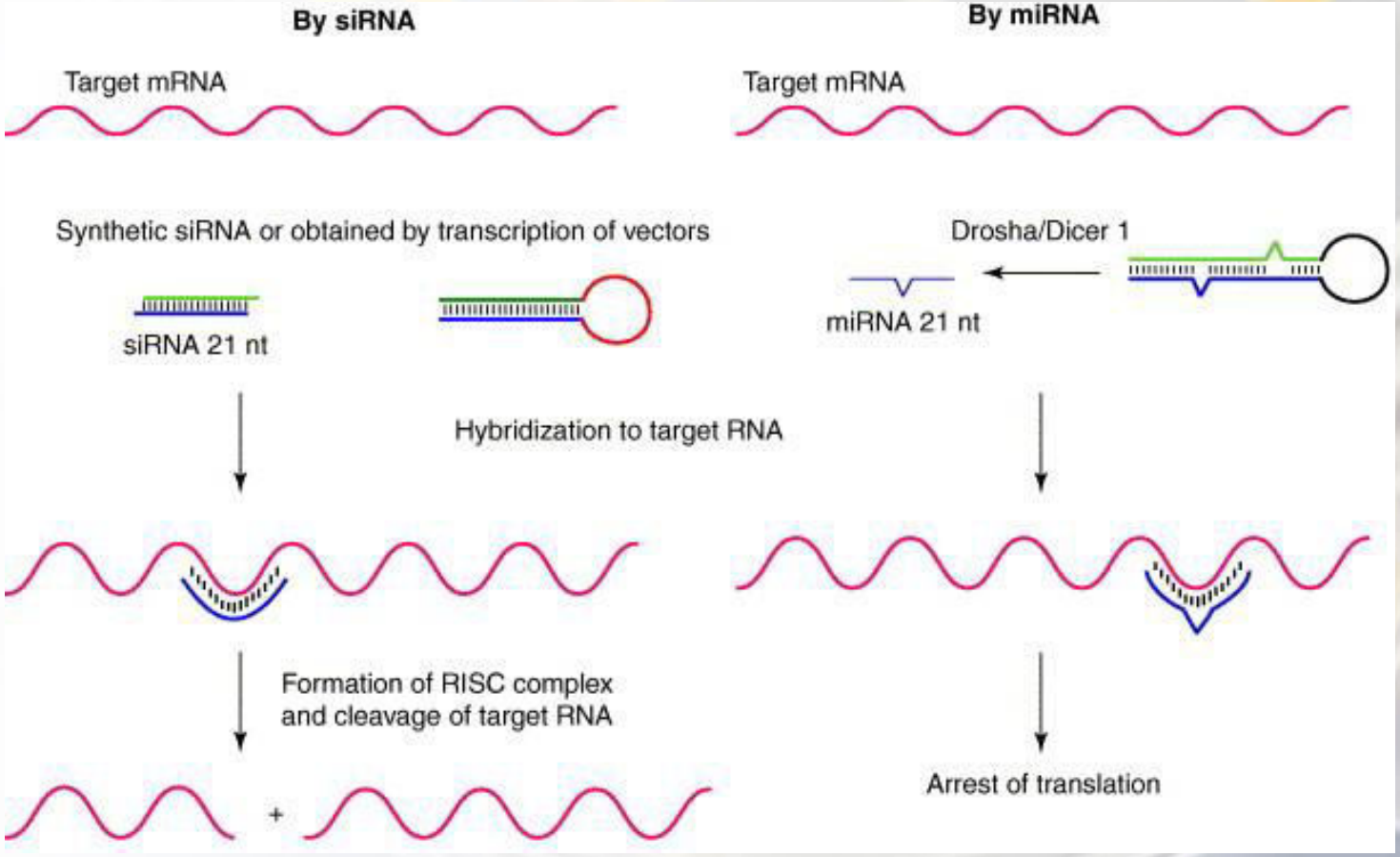
Design of siRNA

Design of siRNA



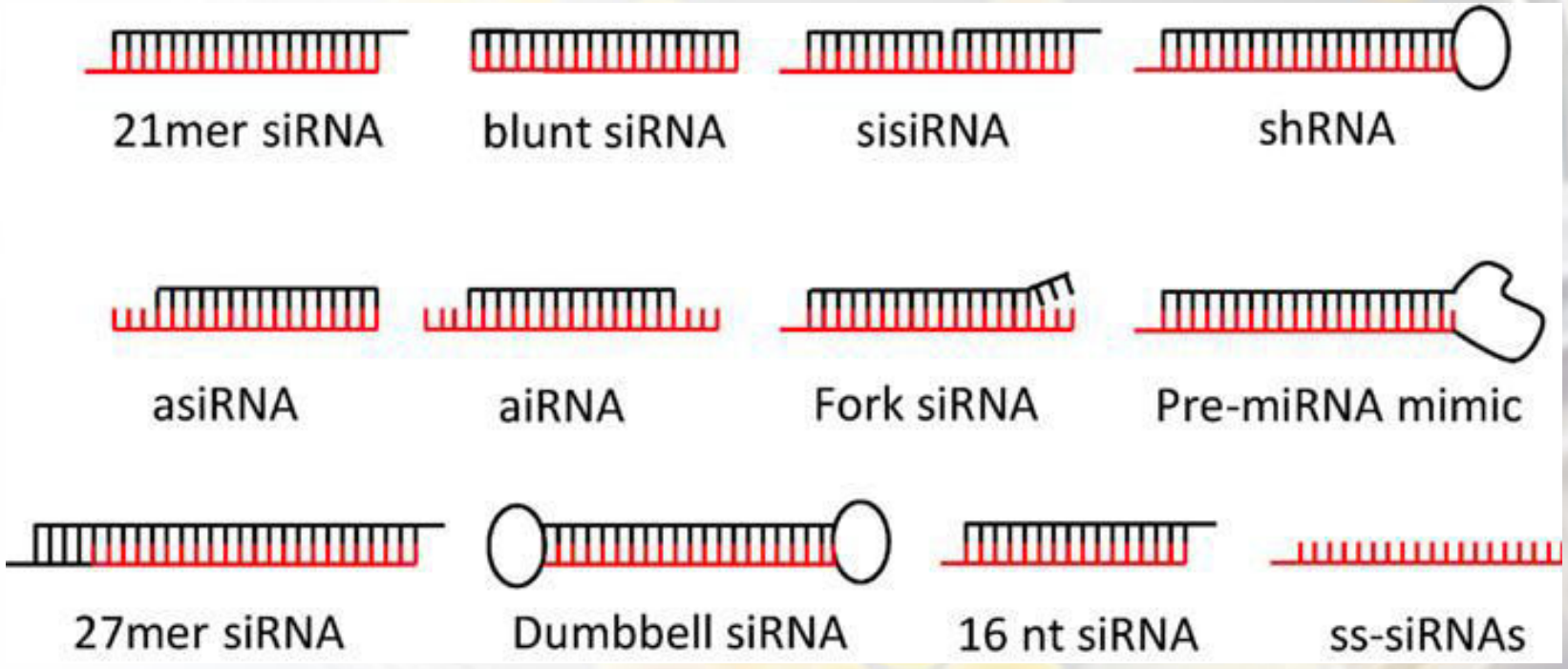
Design of siRNA

siRNA/miRNA: Difference in mode of action



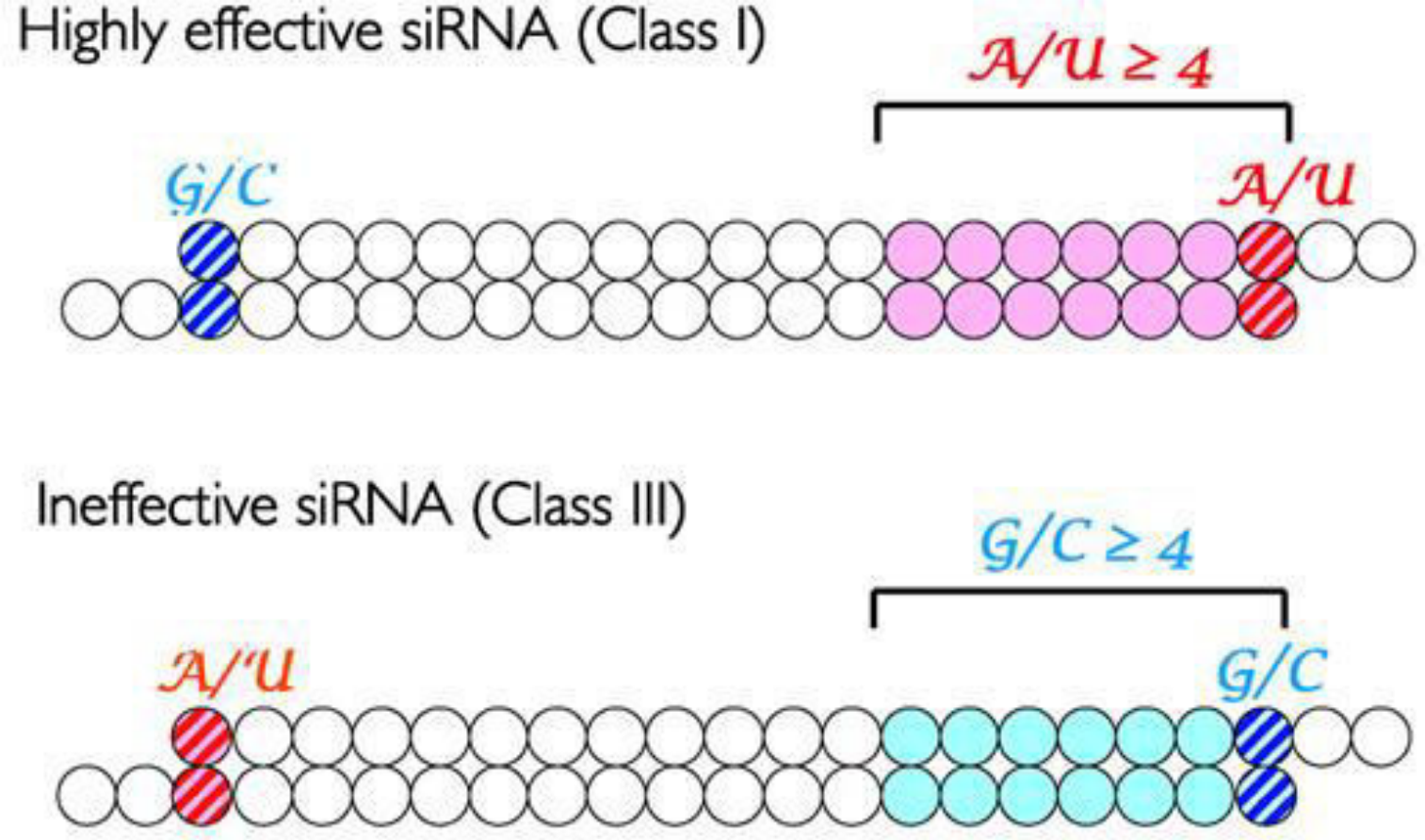
Design of siRNA

Different Design of siRNA



Design of siRNA

Design of siRNA



Design of siRNA

Design of siRNA

- 21-23 nt dsRNA,
- Perfect complimentary to target mRNA
- Targeting 3’ UTR works better than 5’ UTR

Applications of siRNA

Applications of siRNA

Applications of siRNA

- Determining protein function
- Easier than preparing knockout
- Used for partial knockdowns

Applications of siRNA

Applications of siRNA

- Applications for cancer prevention, infections and developmental defects

Applications of siRNA

siRNA

- Exogenously delivered
- 21-23mer dsRNA
- Acts through RISC
- Induces homologous target cleavage
- Perfect sequence match
- Results in target degradation

miRNA

- Endogenously produced
- 21-23mer dsRNA
- Acts through RISC
- Induces homologous target cleavage
- Imperfect sequence match
- Results in translation arrest

Applications of siRNA

Applications of siRNA

- Determining protein function
- Easier than a knockout
- Used for partial knockdowns

Genetics and Genomics

Transcriptomics

Transcriptomics

Transcriptome

- The population of mRNAs expressed by a genome at any given time
- The complete collection of transcribed elements of the genome

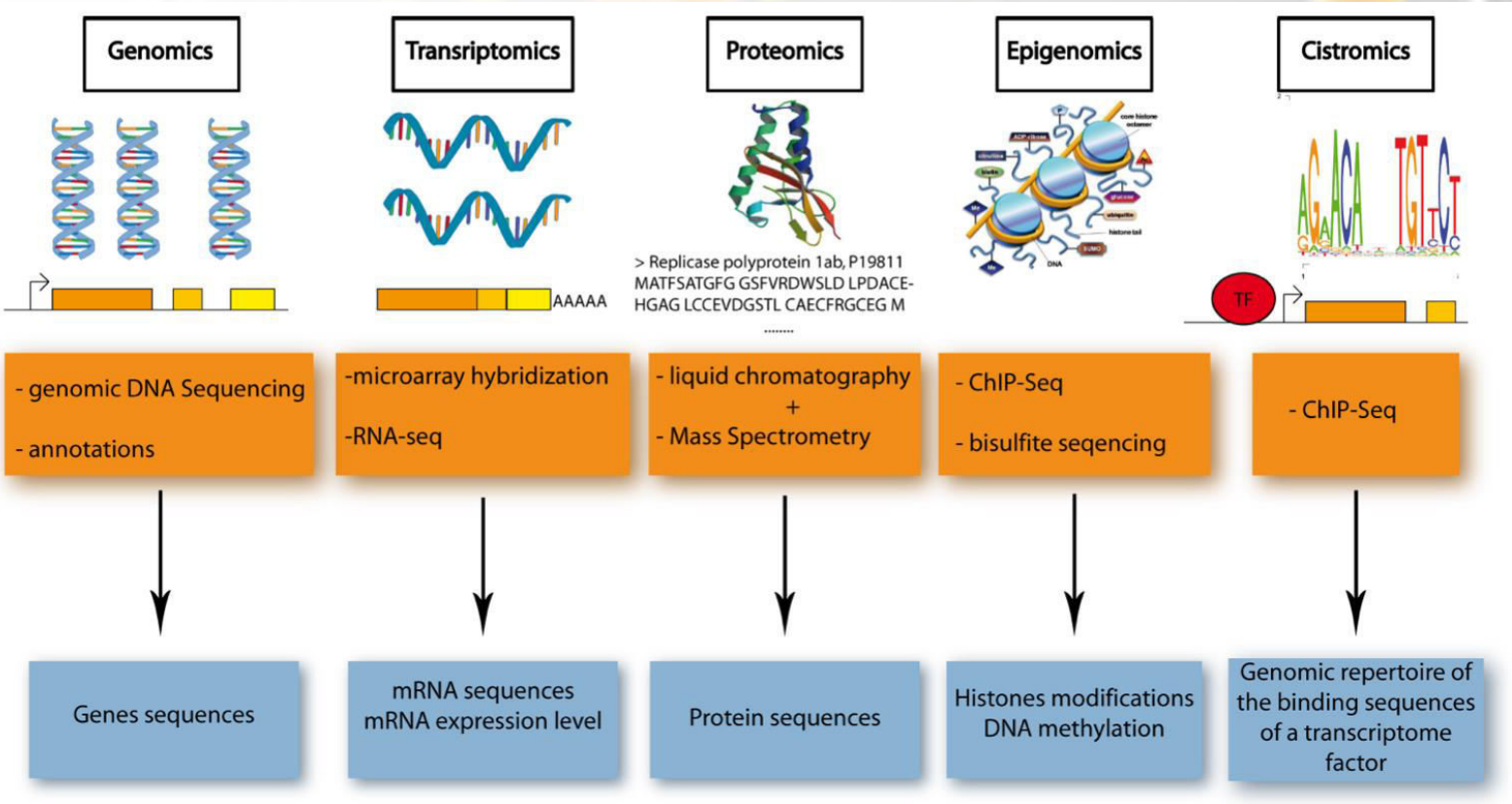
Transcriptomics

Transcriptomics

- The study of characteristics and regulation of the functional RNA transcript population of a cell/s or organism at a specific time

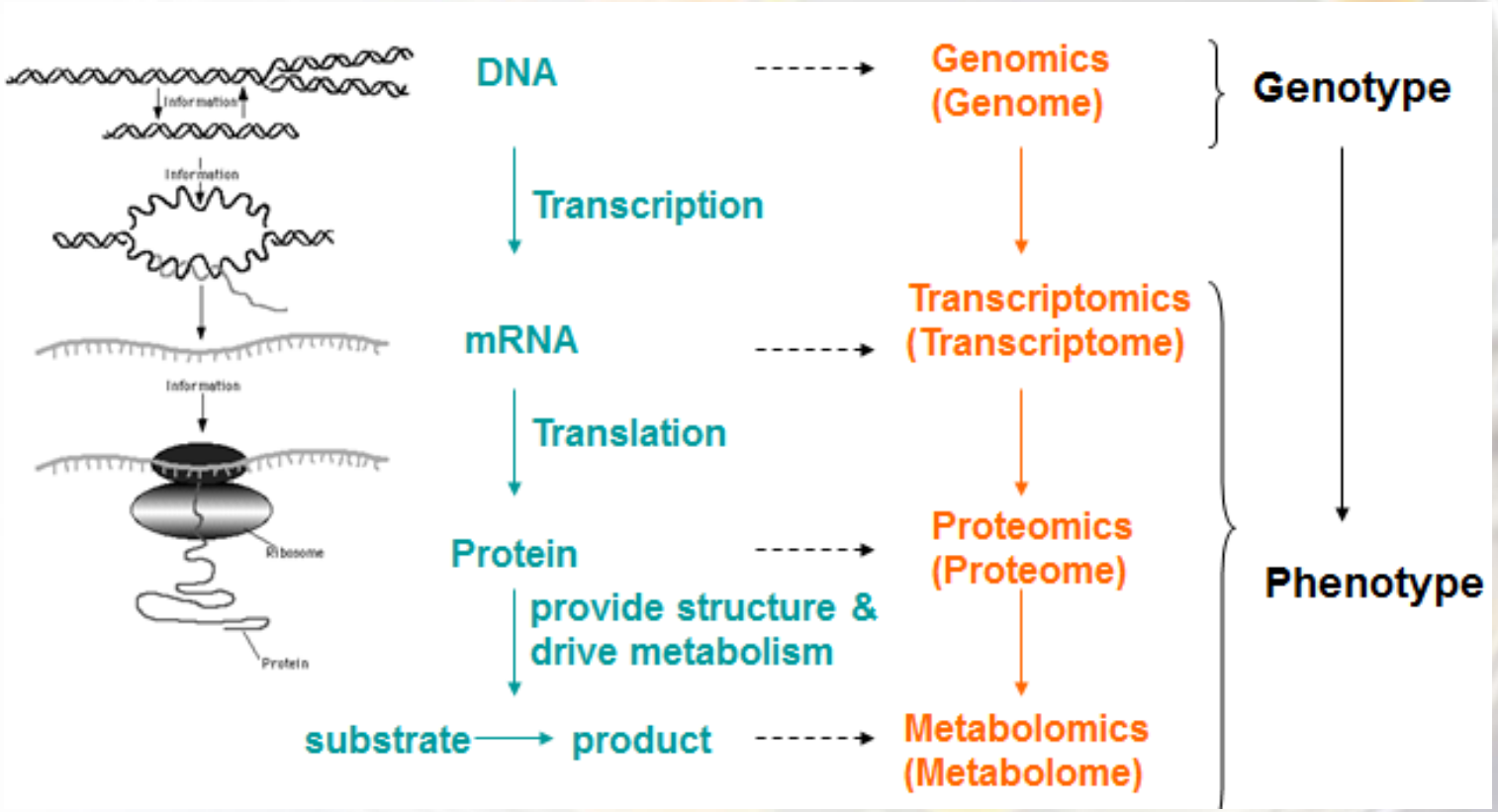
Transcriptomics

Transcriptomics



Transcriptomics

Transcriptomics



Transcriptomics

Transcriptomics

- Percentage of the genetic code that is transcribed into RNA molecules (depends on development, environment, time of the day, tissue)

Transcriptomics

Transcriptomics

- Collection of all gene transcripts present in a given cell/tissue at a given time

Transcriptomics

Transcriptomics

- Genes and pathways involved in biological processes
- Genes with similar expression may be functionally related and under the same genetic control mechanism

Transcriptomics

Transcriptomics

- Help elucidating the function of unknown genes based on their spatial and temporal expression
- Identifies marker genes for diagnosis of diseases

Sequencing Techniques

Sanger's Method

- Frederick Sanger (MRC Centre Cambridge, UK)
- DNA sequencing using **chain-terminating inhibitors** (1977)

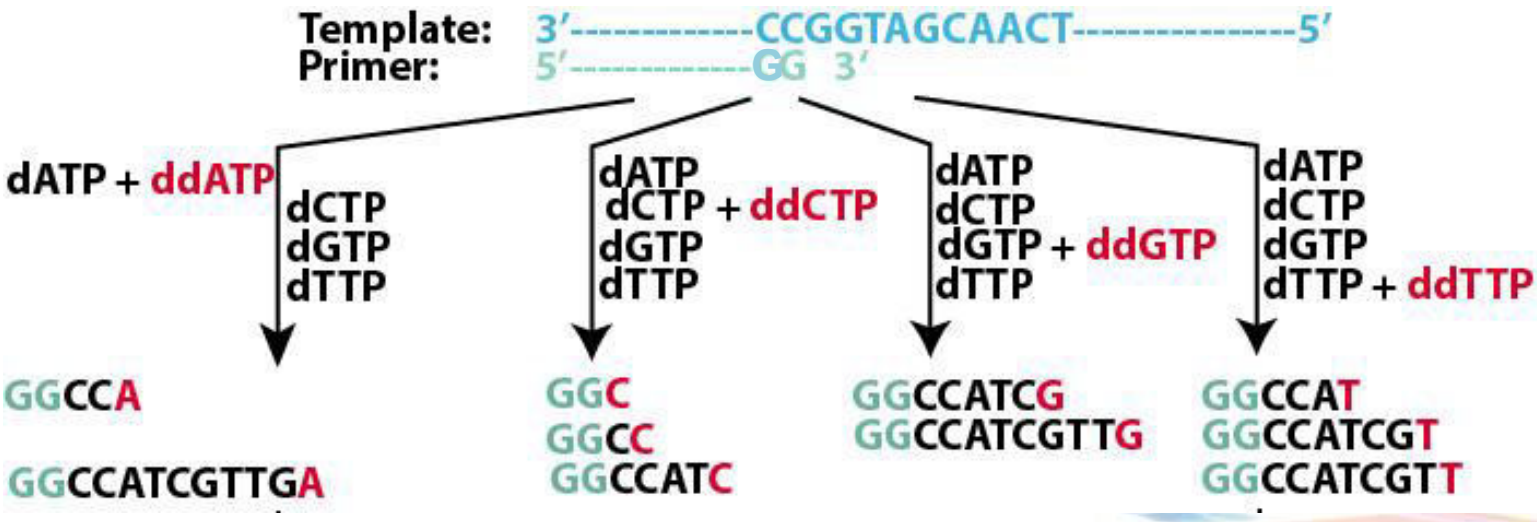
Sequencing Techniques

Fred Sanger

Nobel prize in Chemistry 1958 and 1980
Sanger Institute
University of Cambridge

Sequencing Techniques

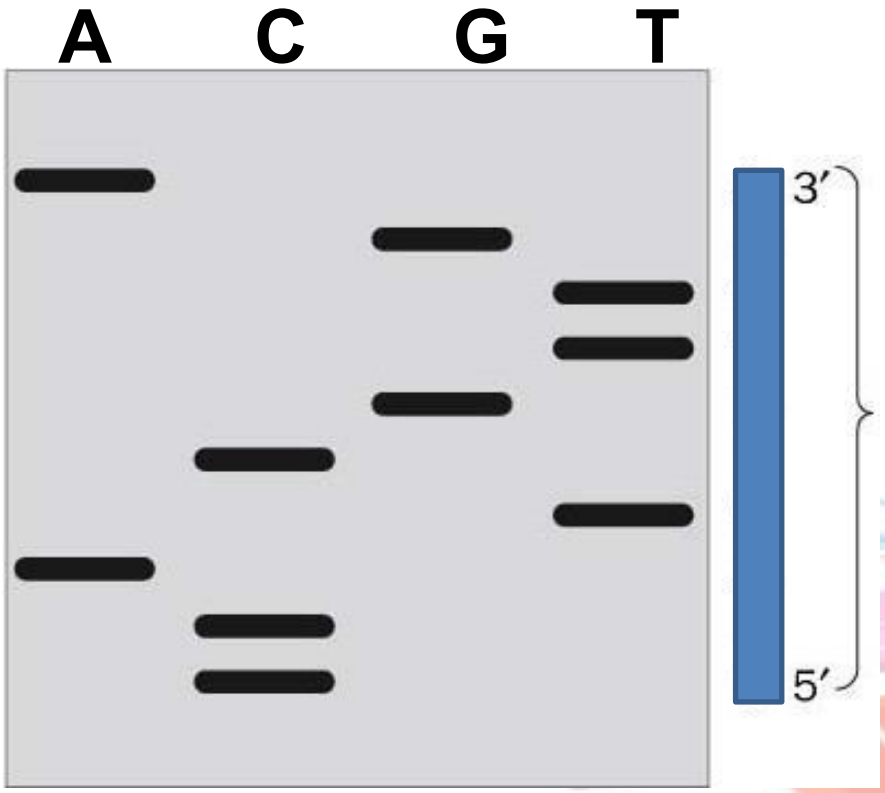
Sanger's Method



http://www.siumed.edu/~bbartholomew/course_material/methods_B.html

Sequencing Techniques

Sanger's Method



http://www.siumed.edu/~bbartholomew/course_material/methods_B.html

Sequencing Techniques

Maxam-Gilbert's Method

- Walter Gilbert and his student Allan Maxam at Harvard also developed sequencing methods, including one for "**DNA sequencing by chemical degradation**"

Sequencing Techniques

Allan Maxam

Walter Gilbert

Abdus Sallam

Sequencing Techniques

Maxam-Gilbert's Method

- Chemical modification followed by cleavage
- 5' end of the DNA to be sequenced is labelled (P^{32})
- Helps in screening

Sequencing Techniques

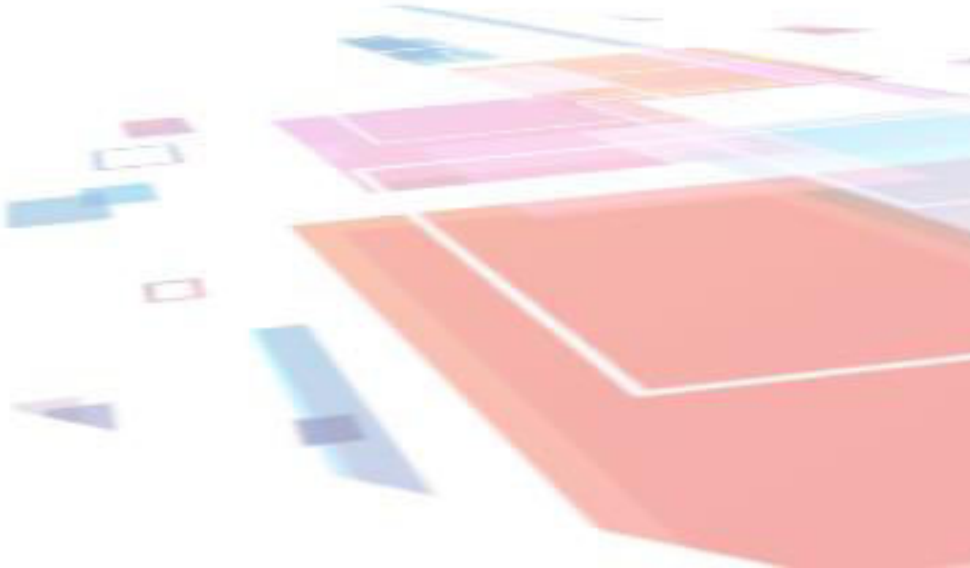
Base Modifications

- Purines (A+G) are depurinated by formic acid
- Gs methylated by dimethyl sulfate
- Pyrimidines (C+T) are hydrolized using hydrazine

Sequencing Techniques

Base Modifications

- NaCl retards hydrazine reaction with **T** in **C-only** reaction

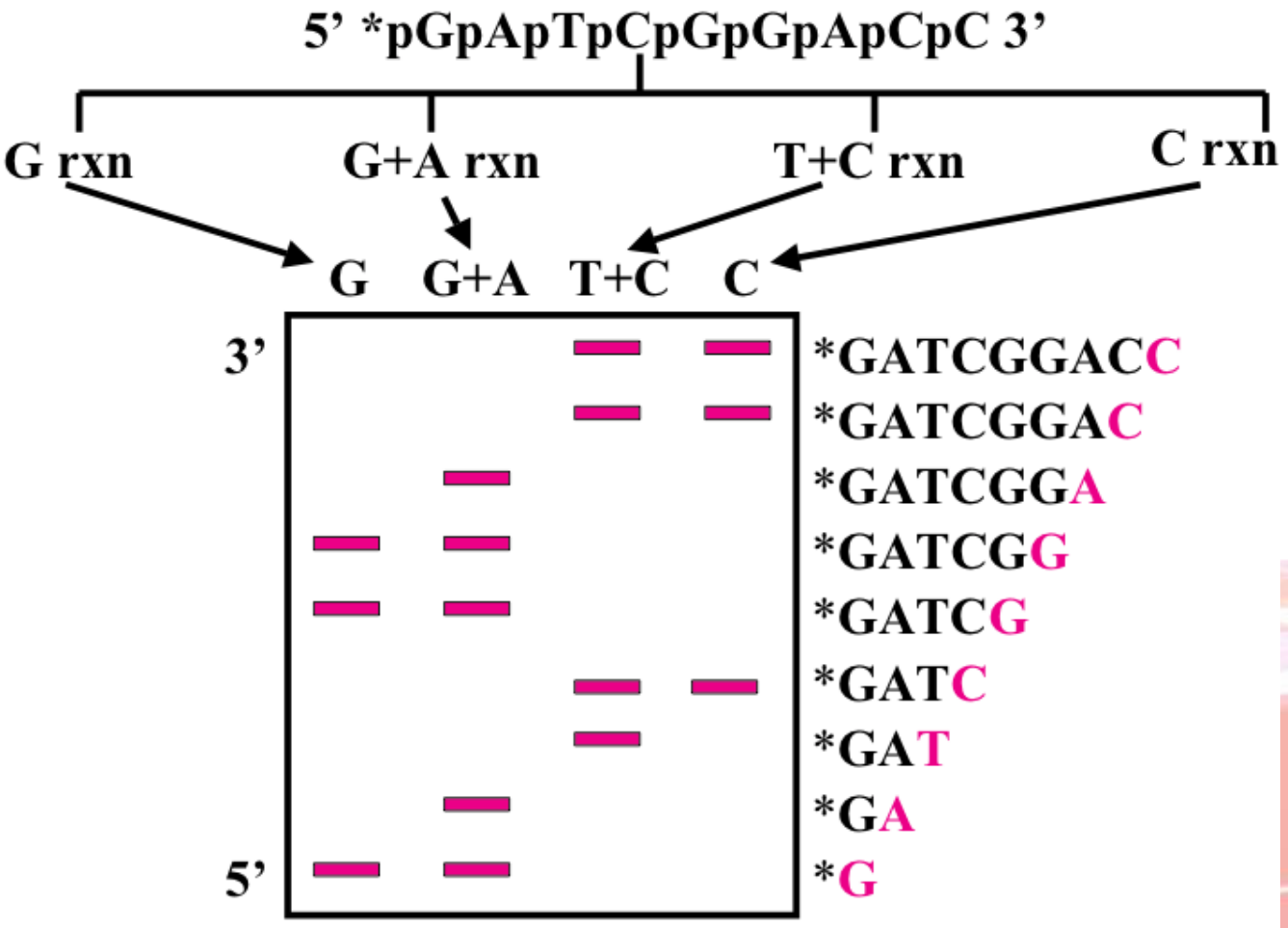


Sequencing Techniques

Maxam-Gilbert's Method

- One base modification per DNA molecule
- Cleavage is done where the base is modified

Sequencing Techniques



Sequencing Techniques

Sanger's vs Gilbert's

Gilbert method involved little complexity in terms of chemical processing therfore Sanger's method got more success

Modern sequencing technologies have adopted Sanger like approaches

Sequencing Techniques

Conclusions

- Different methods of sequencing had been developed over the pace of techniques
- Sanger’s method have got more success

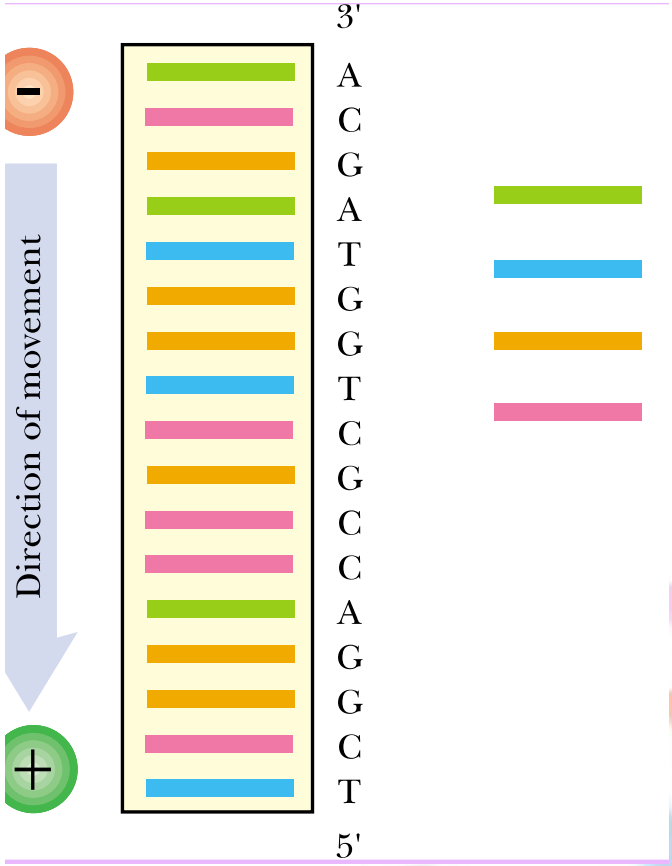
Automated Sequencing

Automated Sequencing

- Florescent dyes are used to label DNA
- Primers for each reaction (ddA, ddT, ddG, ddC) are tagged with separate dyes
- All reactions in same gel lane
- Band colors help in detection

Automated Sequencing

Florescent dye Sequencing



Automated Sequencing

Automated Sequencing

- A flourimeter and computer are hooked up to the gel and they detect and record the dye attached to the fragments as they come off the gel

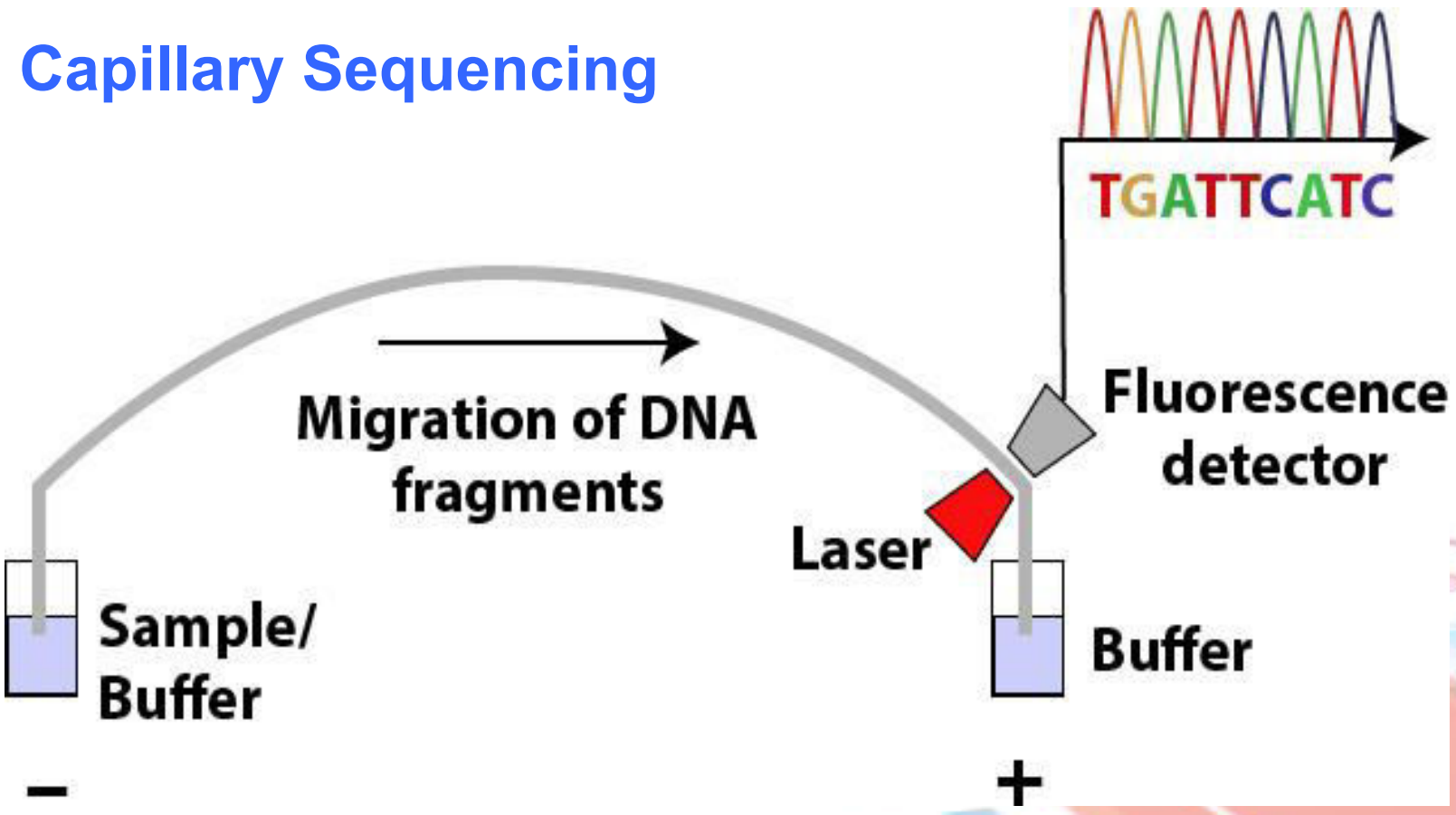
Automated Sequencing

Capillary Sequencing

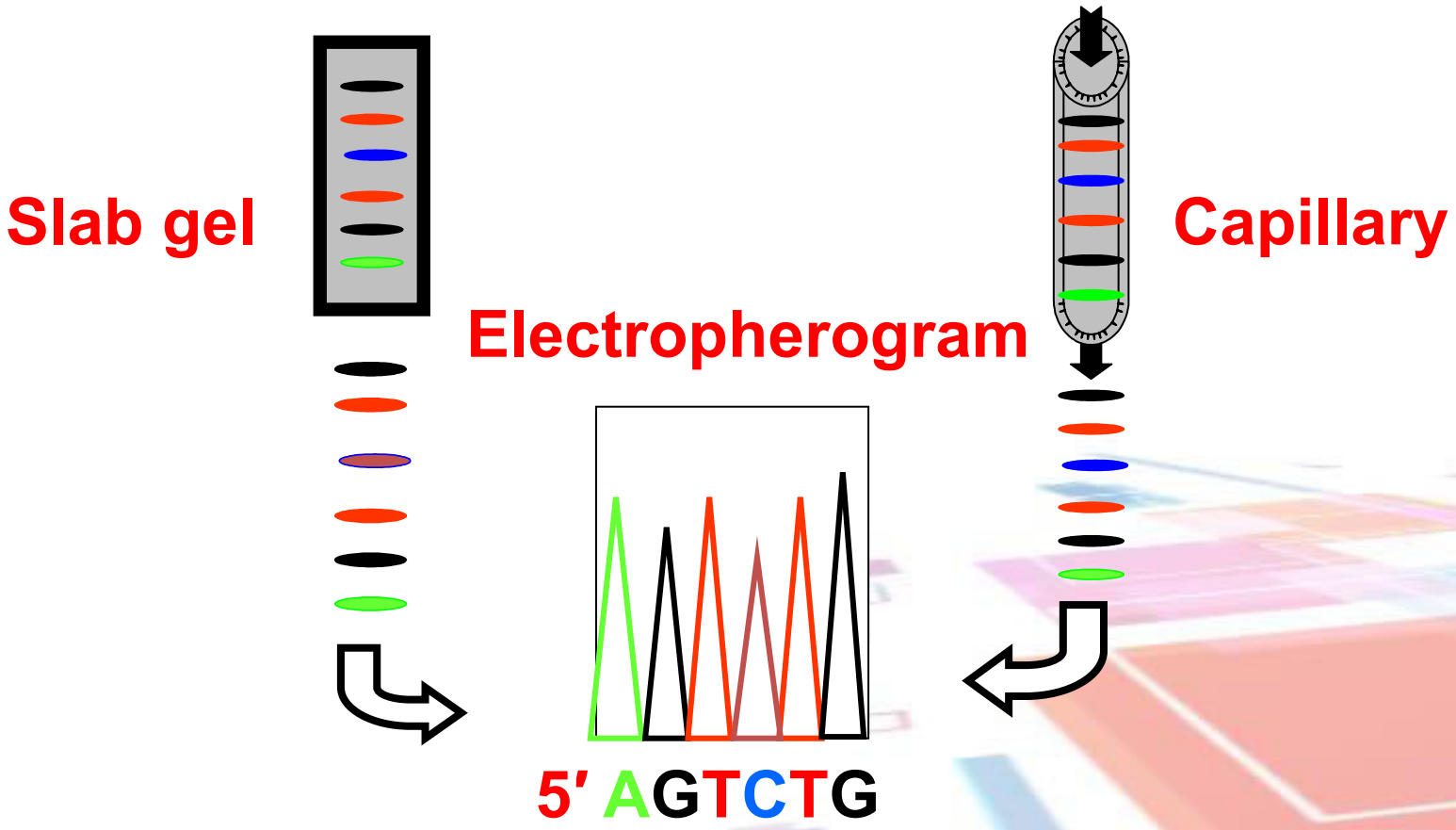
- Samples run through capillary instruments
- Sequence analysis software analyzes sequence in text and electropherogram form
- Peak patterns reflect mutations or sequence changes

Automated Sequencing

Capillary Sequencing

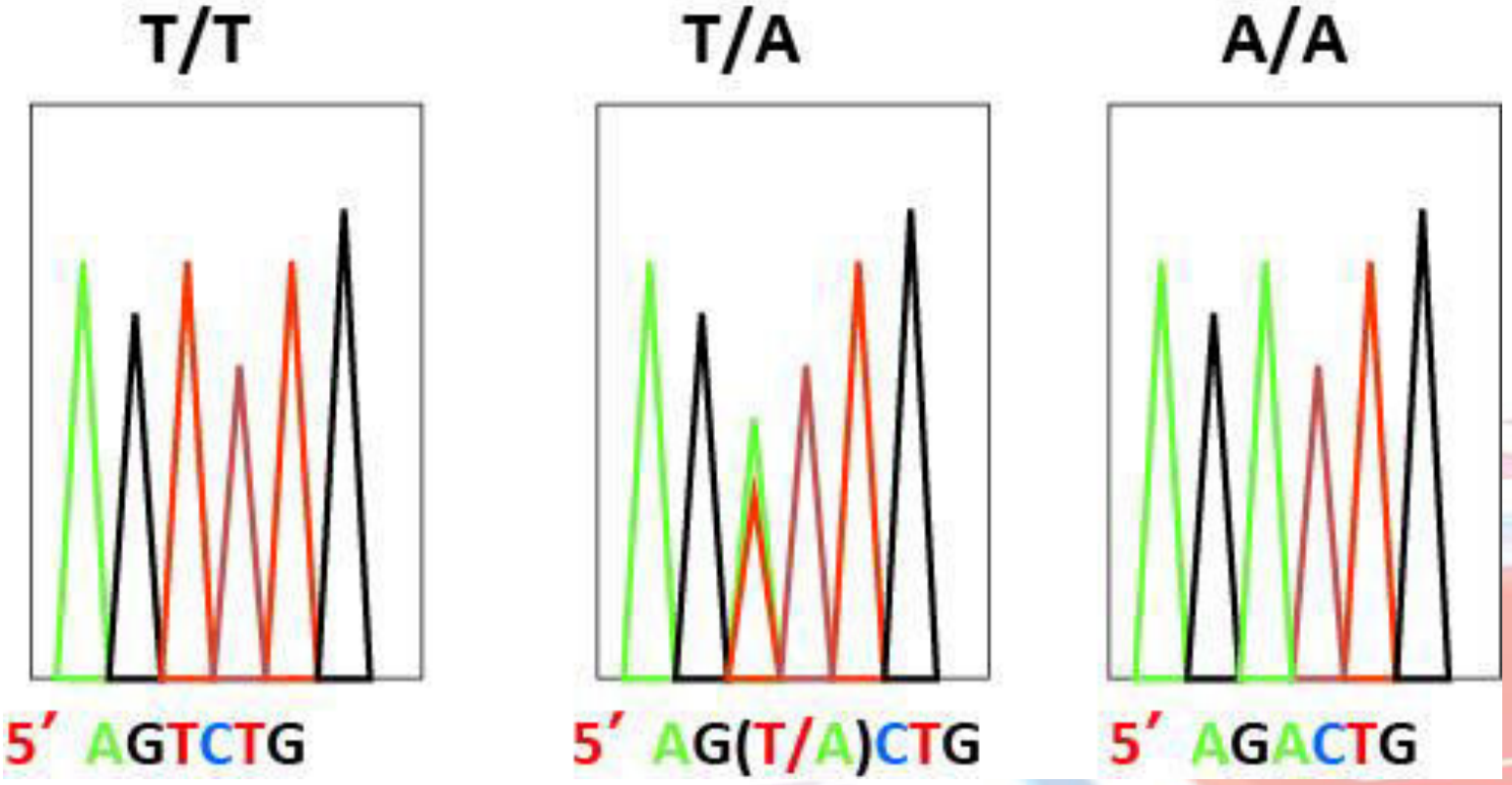


Automated Sequencing



Automated Sequencing

Electropherogram



Automated Sequencing

Conclusions

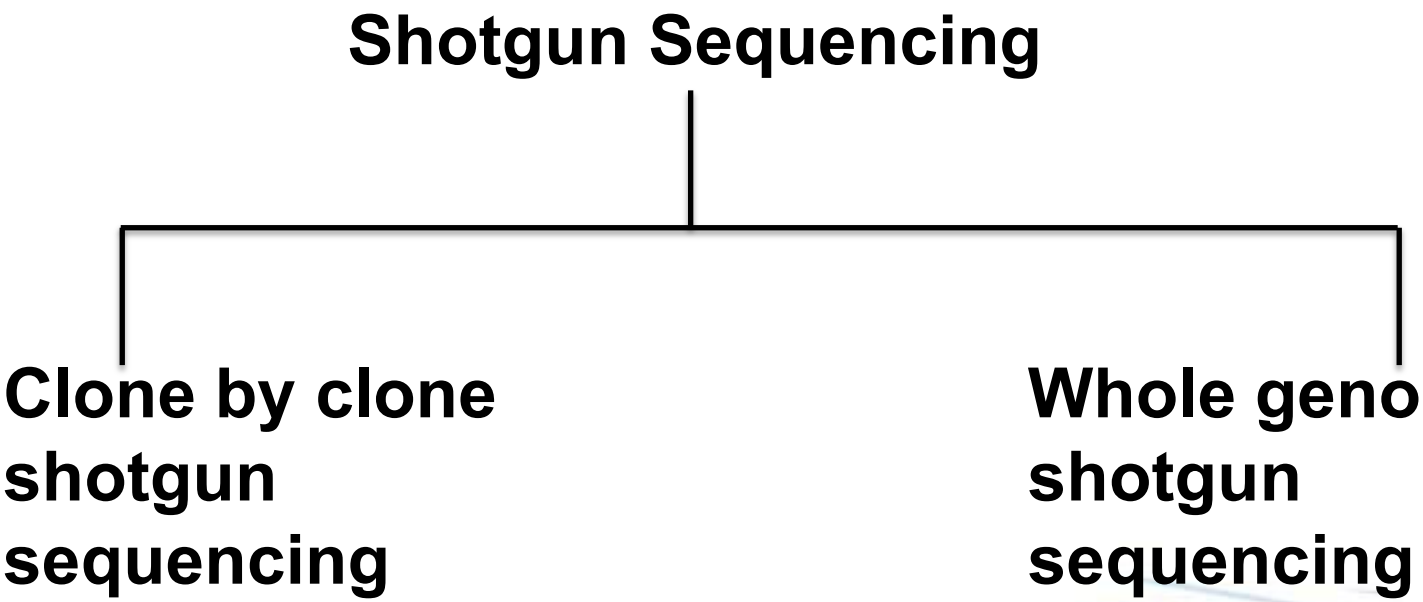
- Use of flourescent dyes and an automated system improved the effeciency of sequencing

Shotgun Sequencing

Shotgun sequencing

- The DNA is broken up randomly into numerous small segments, which are sequenced
- Named by analogy with the rapidly expanding, quasi-random firing pattern of a shotgun

Shotgun Sequencing



Shotgun Sequencing

Clone by clone shotgun sequencing

- Also known as hierarchical shotgun sequencing or map based shotgun sequencing

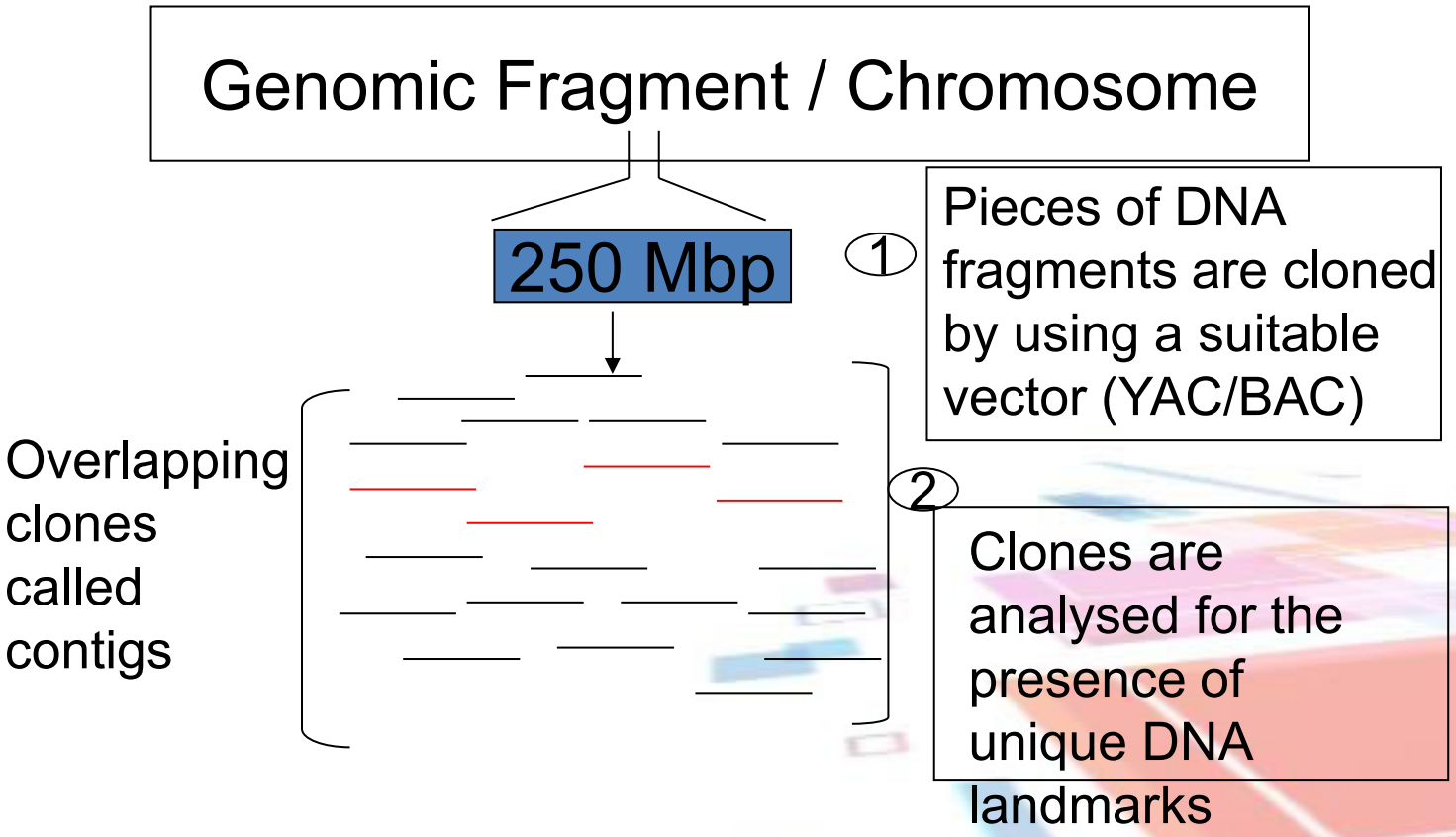
Shotgun Sequencing

Clone by clone shotgun sequencing

- The target DNA is first analysed by clone based physical mapping methods
- Individual clones that together span the region of interest are selected and subjected to shotgun sequencing

Shotgun Sequencing

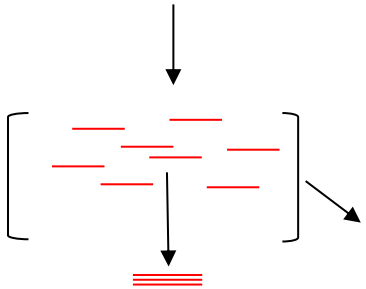
Clone-by-clone shotgun sequencing



Shotgun Sequencing

Clone-by-clone shotgun sequencing

③ Minimaly overlapping clones are then selected for shotgun sequencing



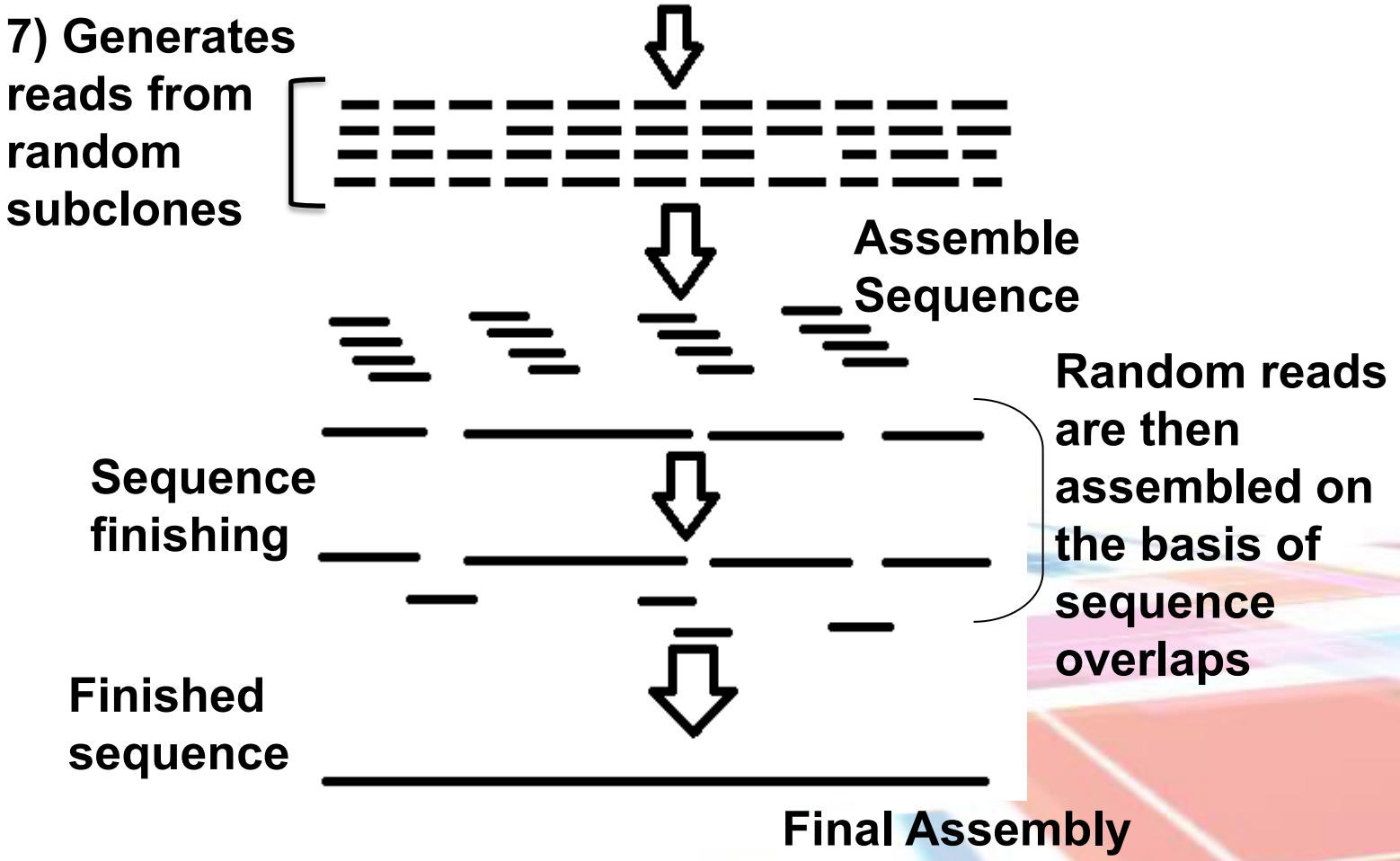
Minimal tiling path

④ Purified BAC/YAC DNA

⑤ Physically fragment the DNA

⑥ Subclone random fragments

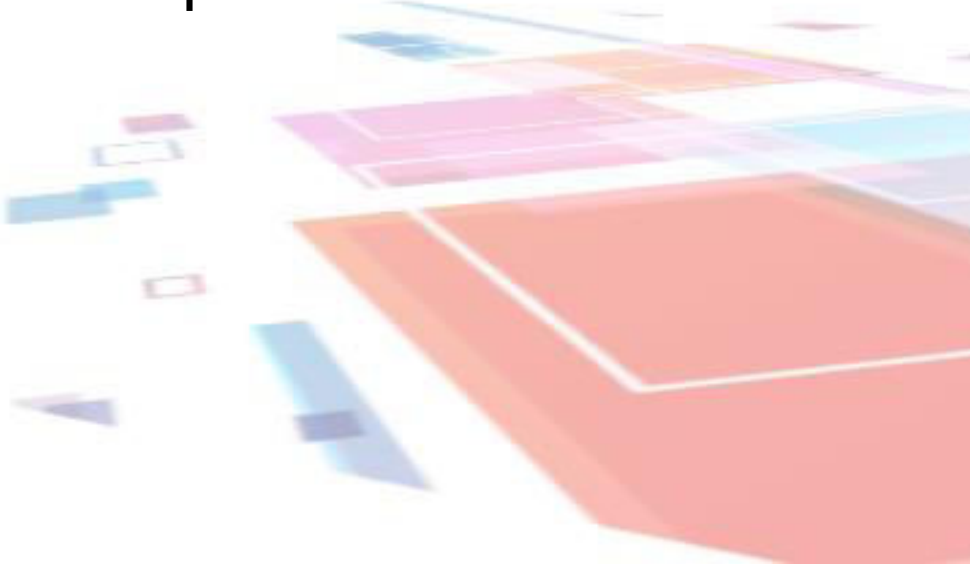
Shotgun Sequencing



Shotgun Sequencing

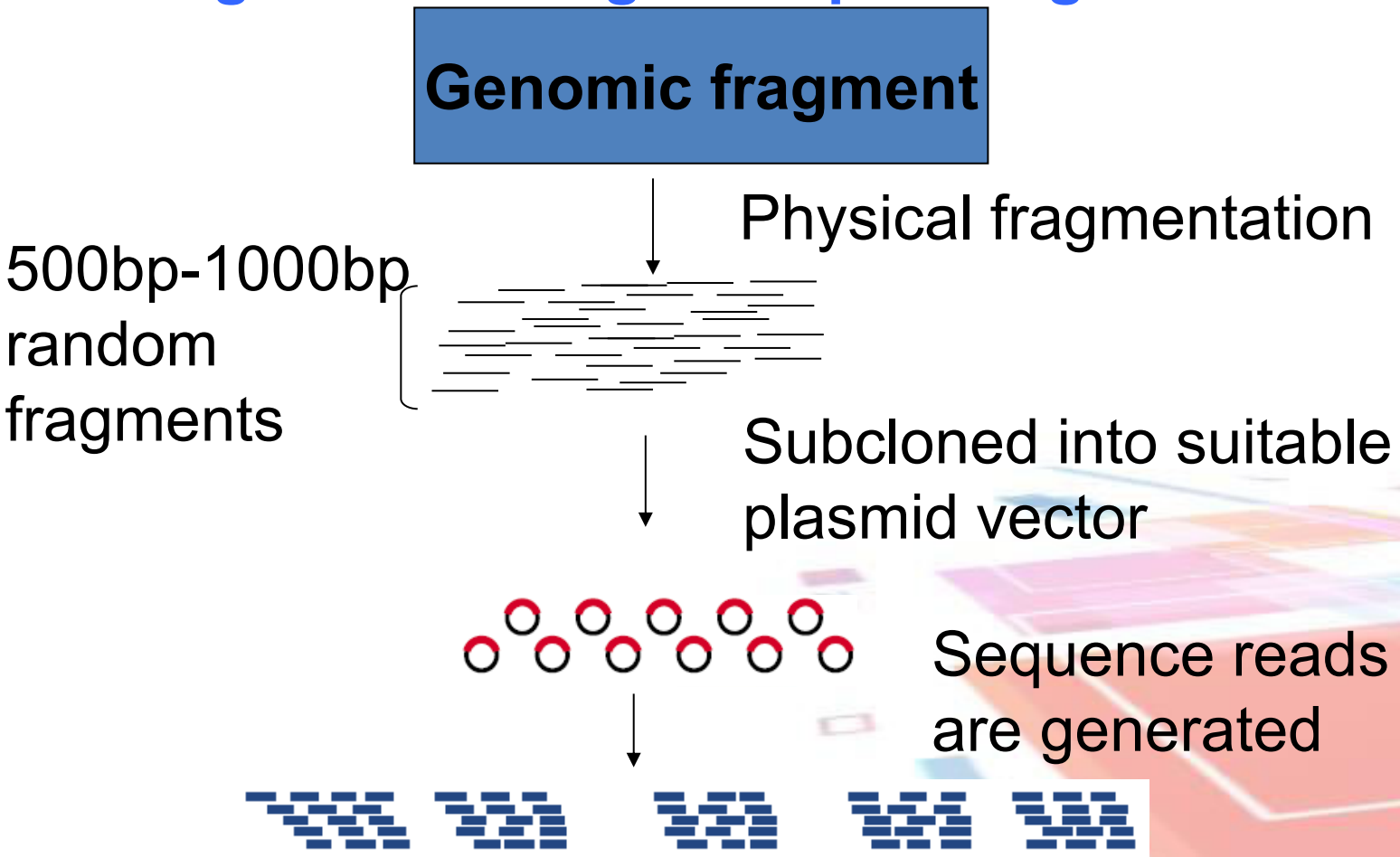
Whole Genome Shotgun Sequencing (WGS)

- Whole genome is broken down, cloned into vectors and then sequenced

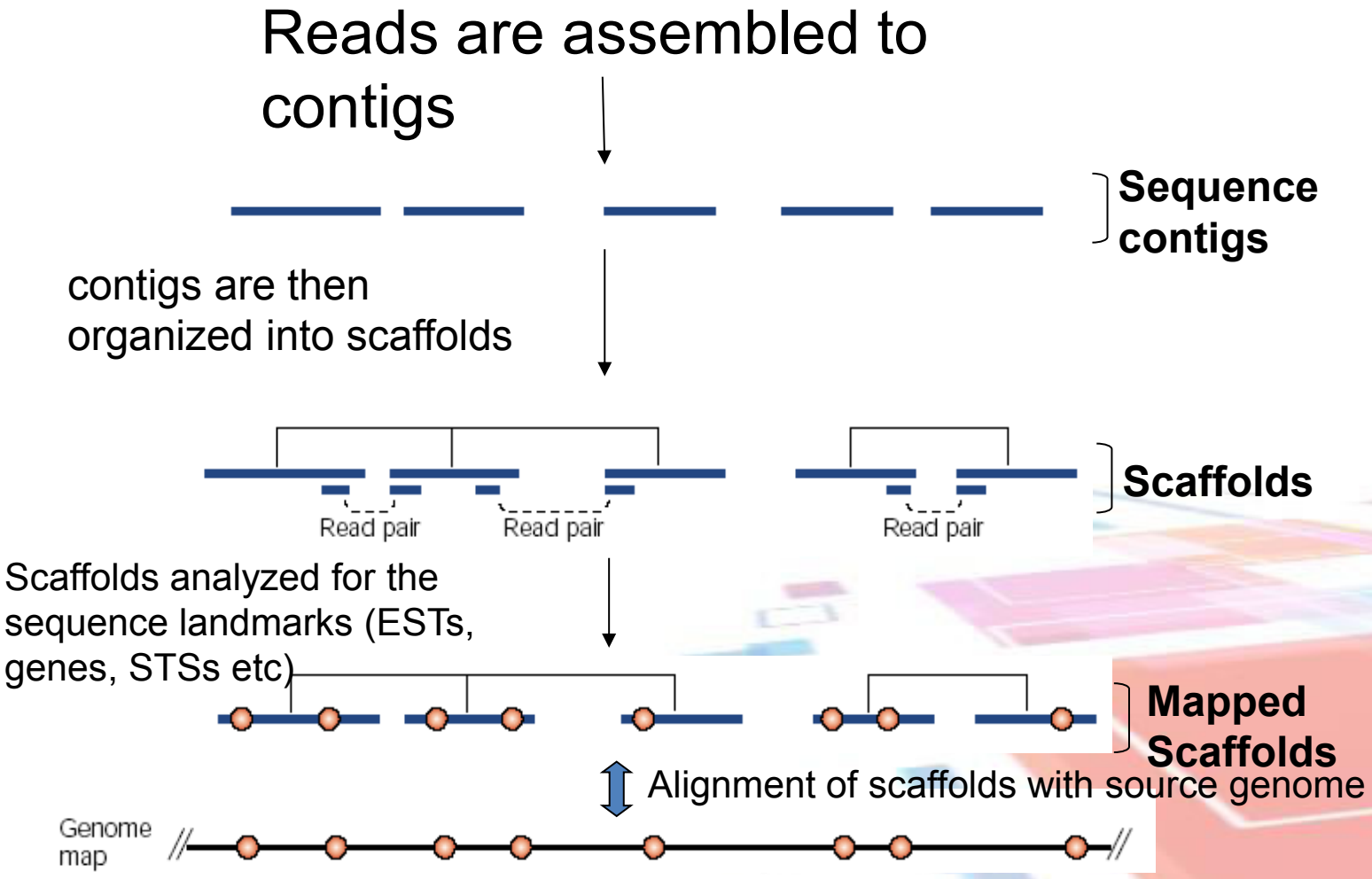


Shotgun Sequencing

Whole genome shotgun sequencing

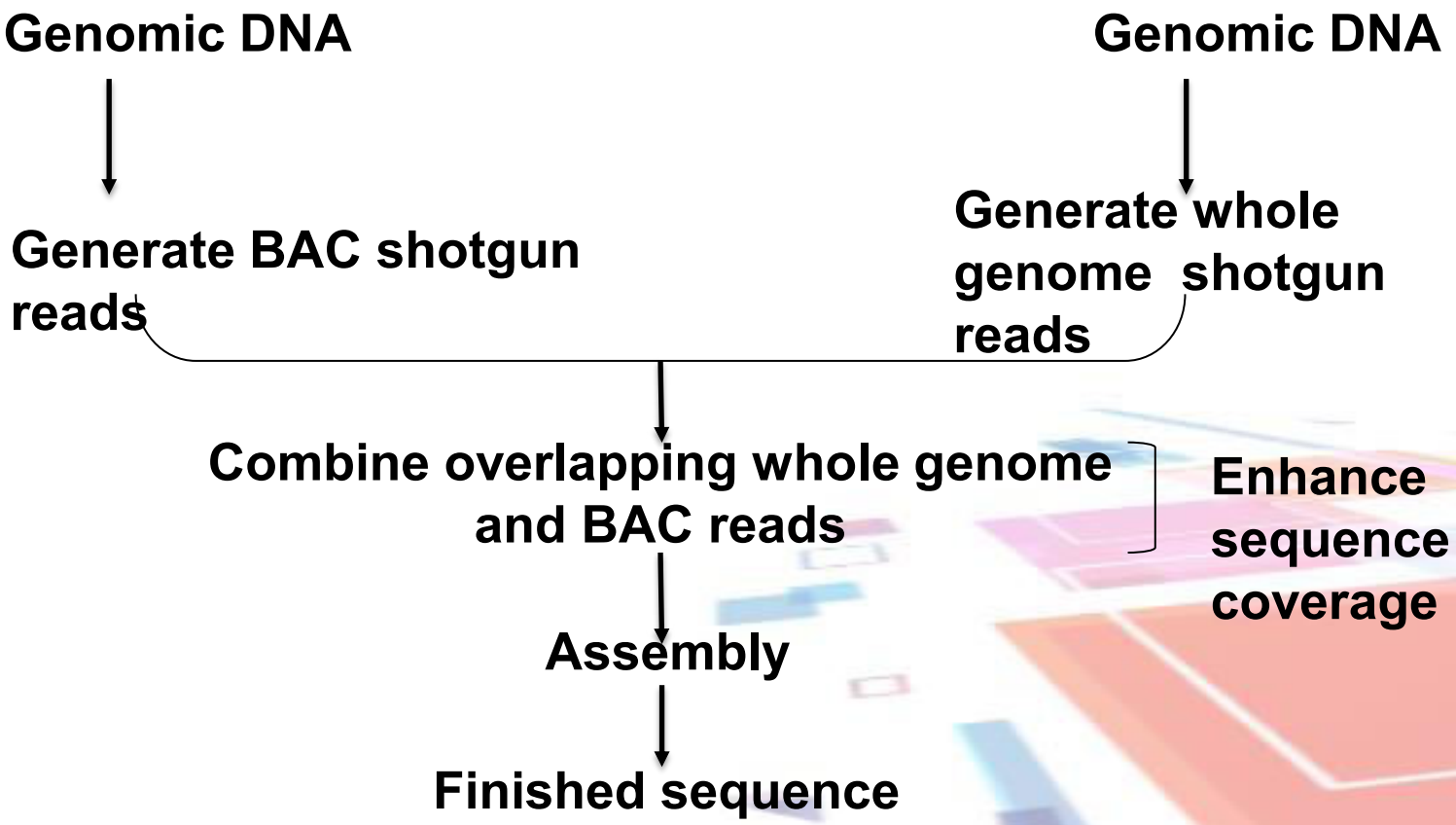


Shotgun Sequencing



Shotgun Sequencing

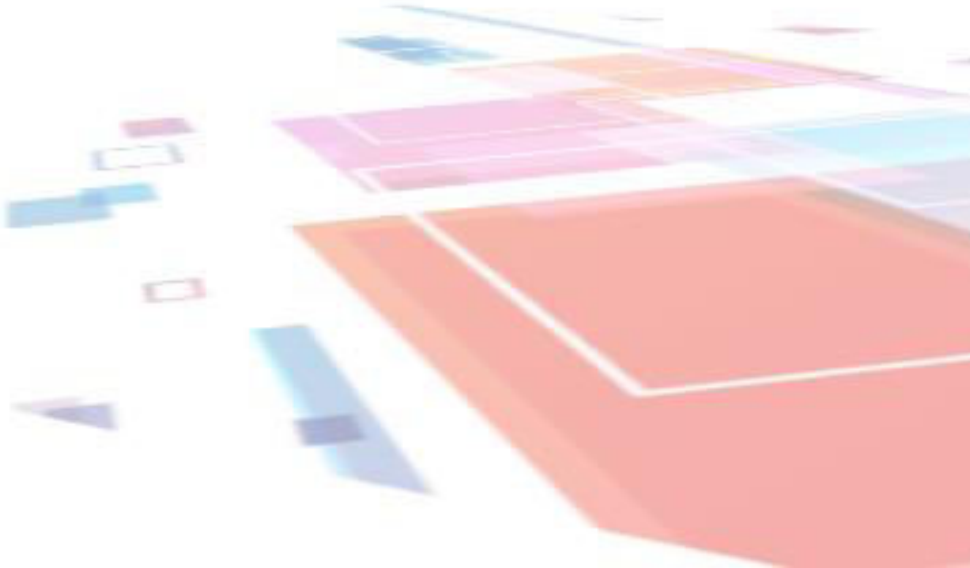
Hybrid shotgun sequencing



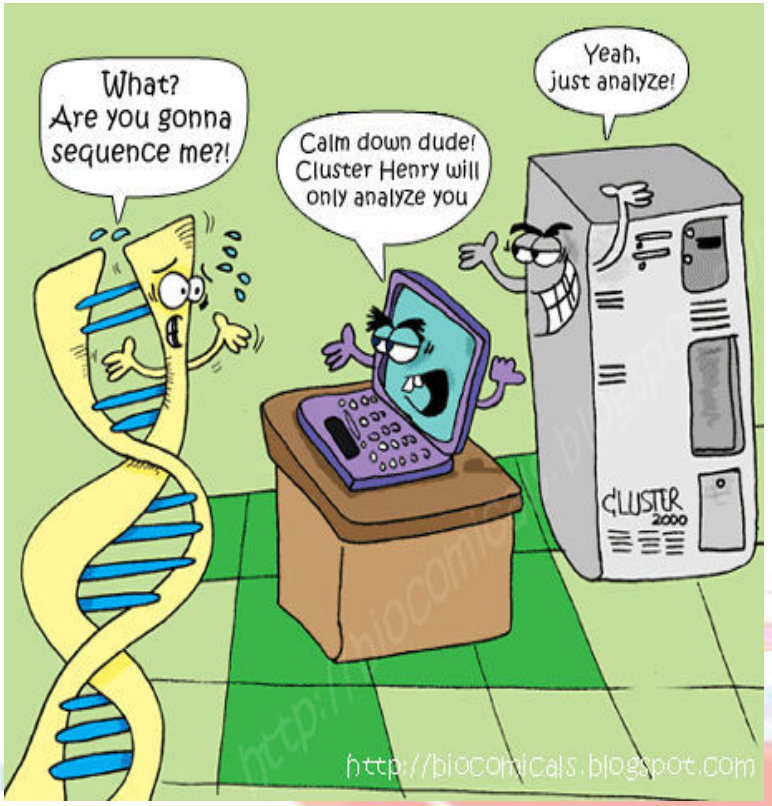
Shotgun Sequencing

Shotgun sequencing

- The DNA is broken up randomly into numerous small segments, which are sequenced



Next-generation sequencing (NGS)



Next-generation sequencing (NGS)

Historical Perspective

- The Human Genome sequencing was accomplished by sanger's method
- Sanger sequencing had been a gold standard for nucleic acid sequencing for about 25 years

Next-generation sequencing (NGS)

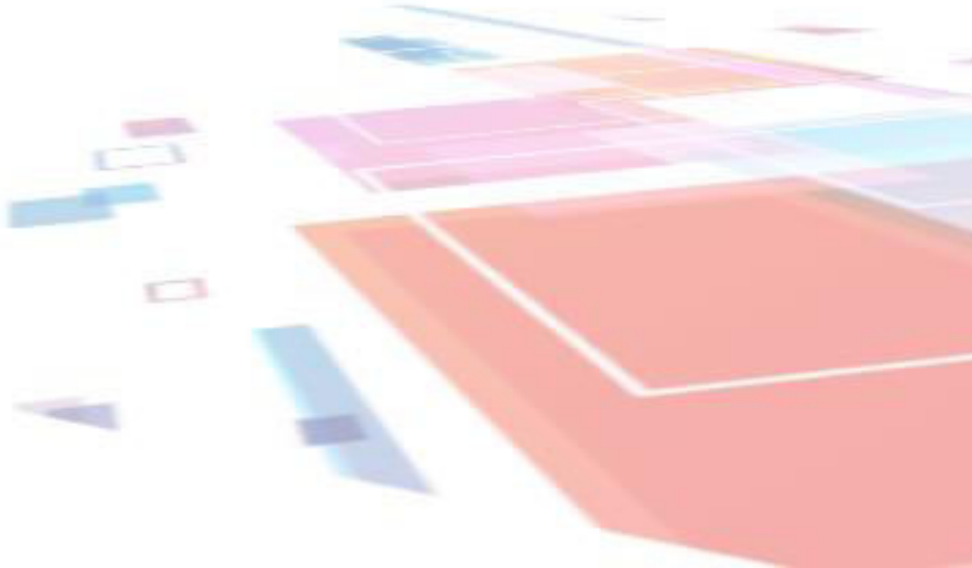
Historical Perspective

- The Lynx Therapeutics (now Solexa) published and marketed "Massively parallel signature sequencing", or MPSS, in 2000

Next-generation sequencing (NGS)

Historical Perspective

- It led to the development of Next Generation Sequencing Techniques



Next-generation sequencing (NGS)

Evolution of sequencing technologies

1953 : Discovery of DNA structure by Watson and Crick

1973 : First sequence of 24 bp published
1977 : Sanger sequencing method published
1980 : Nobel Prize Wally Gilbert and Fred Sanger
1982 : Genbank started 1983 : Development of PCR
1987 : 1st automated sequencer : Applied Biosystems Prism 373

Next-generation sequencing (NGS)

Evolution of sequencing technologies

- 1996 : Capillary sequencer : ABI 310
- 1998 : Genome of Caenorhabditis elegans sequenced
- 2000 : Human genome sequenced
- 2005 : 1st 454 Life Sciences Next Generation Sequencing system : GS 20 System 2006 : 1st Solexa Next Generation Sequencer : Genome Analyzer
- 2007 : 1st Applied Biosystems Next Generation Sequencer : SOLiD

Next-generation sequencing (NGS)

Evolution of sequencing technologies

2009 : 1st Helicos single molecule sequencer : Helicos Genetic Analyser System

2011 : 1st Ion Torrent Next Generation Sequencer : PGM
1st Pacific Biosciences single molecule sequencer : PacBio RS Systems

2012 : Oxford Nanopore Technologies demonstrates ultra long single molecule reads

Next-generation sequencing (NGS)

- **N**ext **G**eneration **S**equencing (**NGS**), **high-throughput** or **Massively parallel signature sequencing** is the catch-all term used to describe a number of different modern sequencing platforms

Next-generation sequencing (NGS)

NGS Technologies

- Illumina (Solexa) sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

Next-generation sequencing (NGS)

NGS

- NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced in unison
- High-**throughput** sequencing

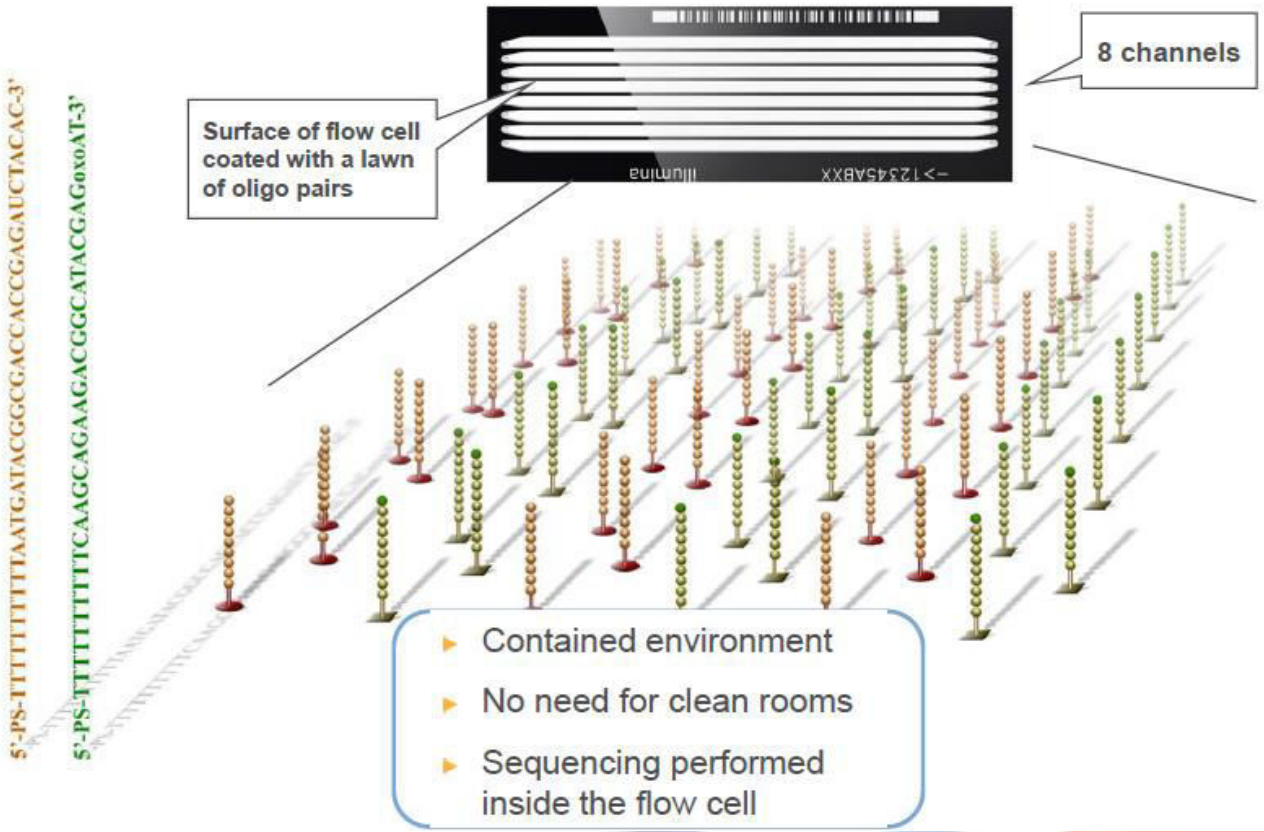
Next-generation sequencing (NGS)

NGS

- High-throughput sequencing (HTS) produces thousands or millions of sequences concurrently
- > 500,000 sequencing-by-synthesis operations run in parallel

Next-generation sequencing (NGS)

The flow cell design



Next-generation sequencing (NGS)

Generations of sequencing

- First generation
  - Sanger sequencing
- Second/Next generation
  - 454, Illumina, SOLiD

Next-generation sequencing (NGS)

Generations of sequencing

- Third generation/ Next Next Generation
  - Helicose, PacBio, Ion Torrent, Oxford Nanopore

Next-generation sequencing (NGS)

GS FLX 454
(ROCHE)

HiSeq 2000
(ILLUMINA)

5500xl SOLiD
(ABI)

GS Junior

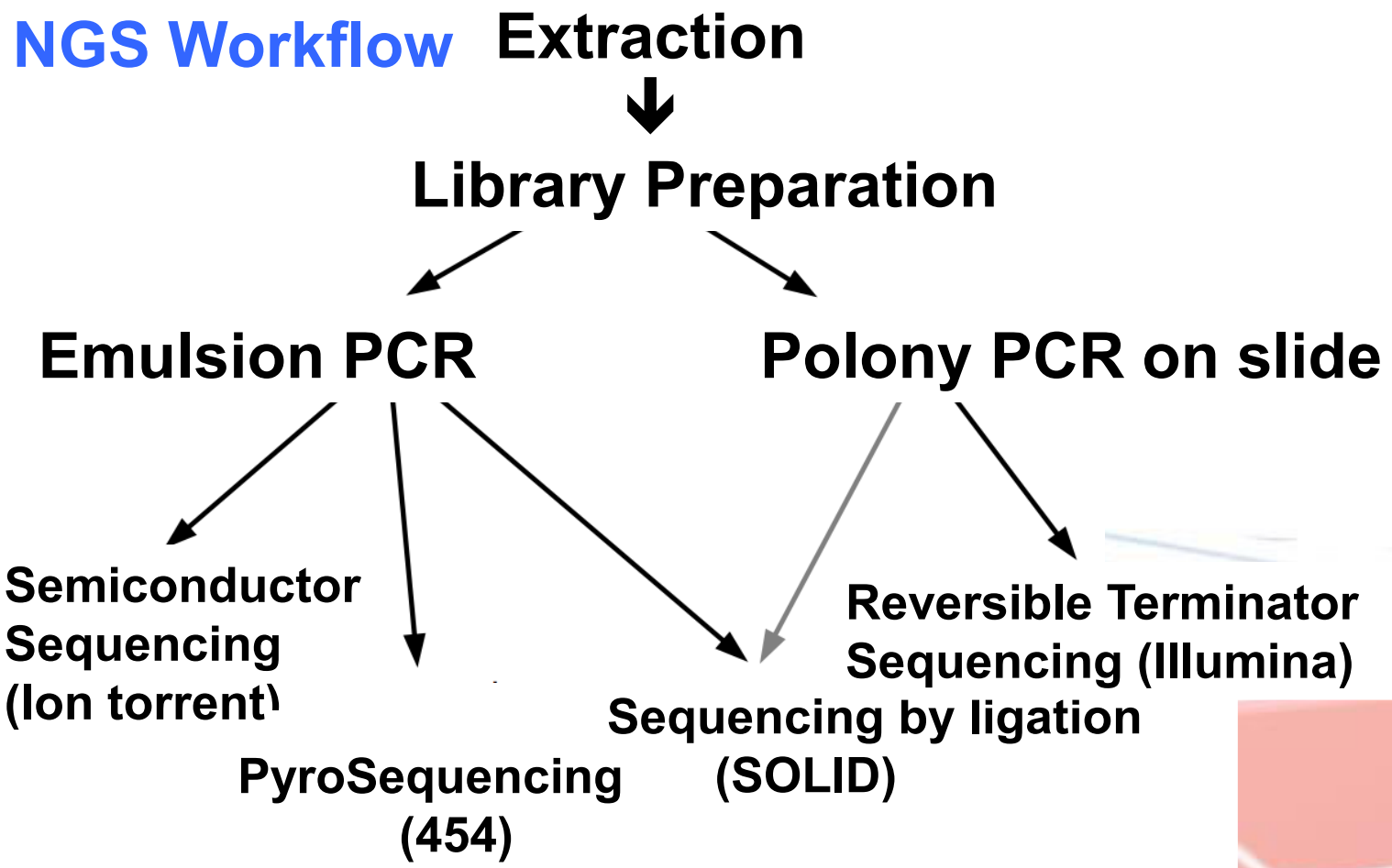
Ion TORRENT

Next-generation sequencing (NGS)

Advantage

These recent technologies allow us to sequence DNA and RNA much more **quickly and cheaply** than the previously used Sanger sequencing

Next-generation sequencing (NGS)



Next-generation sequencing (NGS)

Conclusions

- NGS technologies sequence millions of segments in parallel
- Sequence DNA much more **quickly and cheaply** than the previously used Sanger sequencing

Resources (for this presentation)

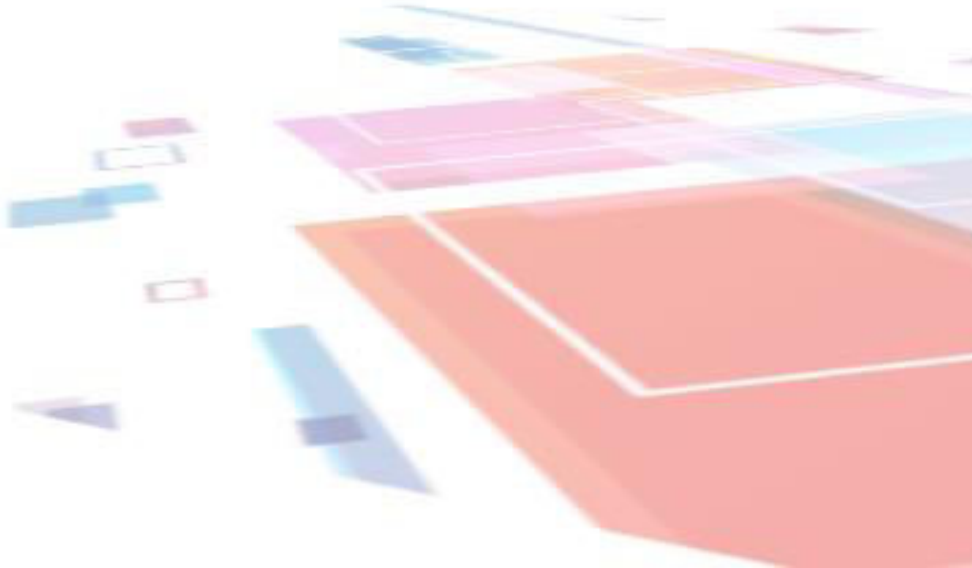
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



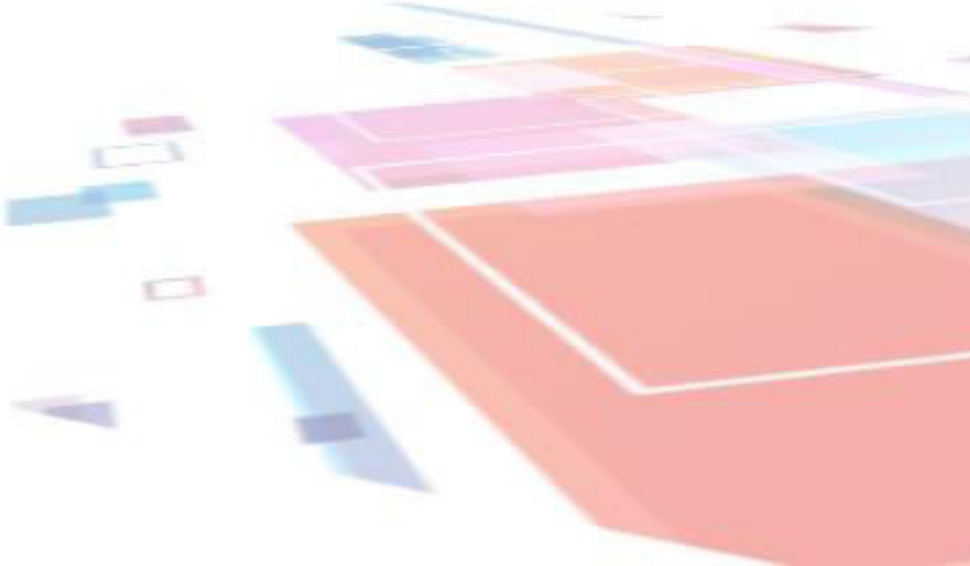
Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

NGS Platforms

NGS Technologies

- Roche
  - 454
- Illumina
- Applied Biosystems
- Ion torrent



NGS Platforms

454

- Founded June 2000
- By Jonathan M. Rothberg
- Subsidiary of Roche since 2007
- No more support from Roche by 2016
- Branford Connecticut
- http://my454.com/

NGS Platforms

Roche

454 SEQUENCING

NGS Platforms

GS Junior System

Roche

454 SEQUENCING

GS FLX+ System

NGS Platforms

	GS Junior	GS FLX Titanium XL+
Read Size	400 bp	700 bp
Throughput	35 Mb	700 Mb
Reads / run	100,000	1 million
Accuracy	99 %	99,997 %
Run Time	10 hours	23 hours

NGS Platforms

Illumina

- Public (NASDAC)
- Founded 1998
- San Diego, California
- Revenue: 512.38 M (31st December, 2014)
- http://illumina.com/

NGS Platforms

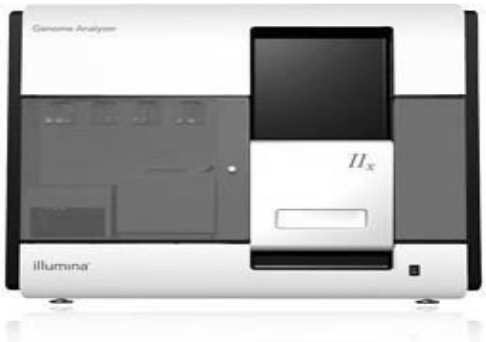
illumina®

NGS Platforms

HiSeq

HiScanSQ

Genome Analyzer IIx



MiSeq

NGS Platforms

	HiSeq	HiScanSQ	Genome Analyzer IIx	MiSeq
Read Size	100 bp	100 bp	150 bp	250 bp
Throughput	600 Gb	150 Gb	95 Gb	6 Gb
Reads / run	3 billion	750 million	320 million	1.2 million
Accuracy	99,9 %	99,9 %	99,9 %	99,9 %
Run Time	11 days	8 days	14 days	20-35 hr

NGS Platforms

Applied Biosystems Inc (ABI)

- Public (NASDAC)
- Founded 1981
- California
- Merged with invitrogen into Life Technologies which was purchased by Thermofisher
- http://AppliedBiosystems.com/

NGS Platforms

applied biosystems
by Life technologies

NGS Platforms

5500xl SOLiD Sequencer

5500 SOLiD Sequencer

applied biosystems
by life technologies

NGS Platforms

	5500xl	5500	5500 W
Read Size	75 bp	75 bp	75 bp
Throughput	180 Gb	90 Gb	360 Gb
Reads / run	3 million	1.5 million	6 million
Accuracy	99,99 %	99,99 %	99,99 %
Run Time	7-12 days	7-12 days	7-12 days

NGS Platforms

Ion Torrent

- Life Technologies
- Between 2nd and 3rd generation sequencing technologies
- California
- http://ioncommunity.lifetechnologies.com/

NGS Platforms

ion torrent

by life technologies™

Proton

PGM
(Personal Genome Machine)

NGS Platforms

	PGM	Proton
Read Size	200 bp	200 bp
Throughput	20Mb-1 GB	10-100 Gb
Reads / run	1.1 million	660 million
Accuracy	99%	99%
Run Time	4-5 hours	4-5 hours

NGS Platforms

Conclusions

- Roche
  - 454
- Illumina
- Applied Biosystems
- Ion torrent

Resources (for this presentation)

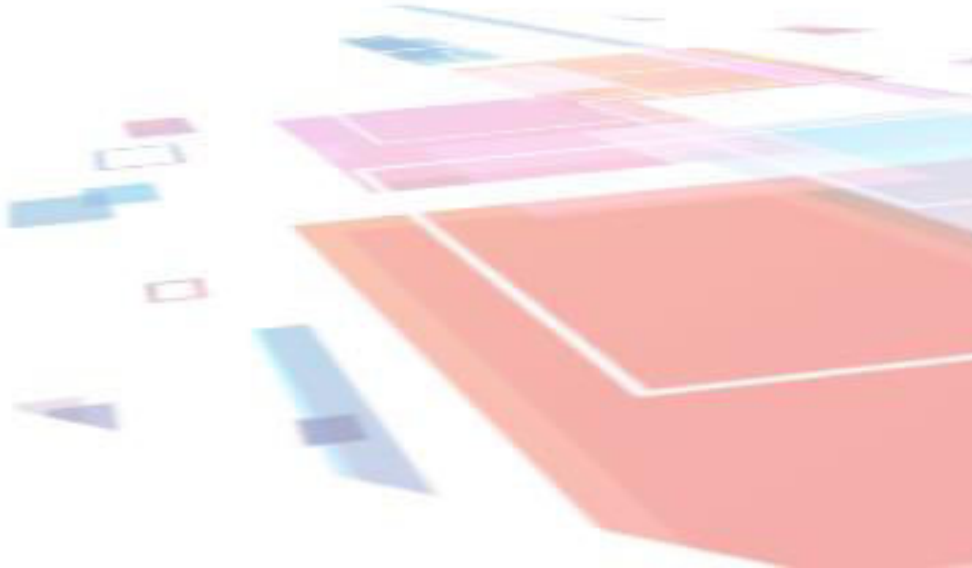
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews

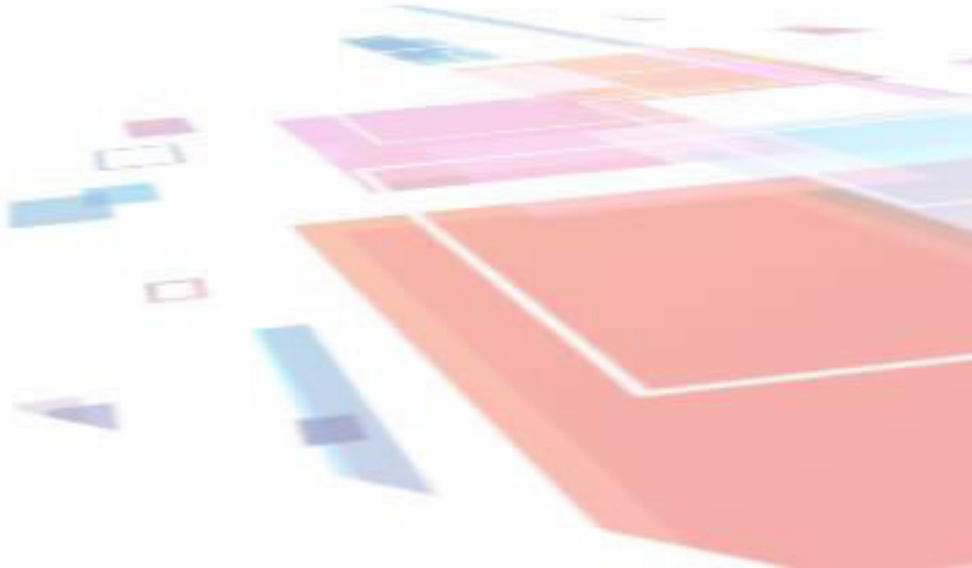


Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

454 Pyrosequencing

Sequencing in 454 machines is performed by pyrosequencing



454 Pyrosequencing

Principle

- Based on the generation of light signal through release of pyrophosphate (PPi) on nucleotide addition

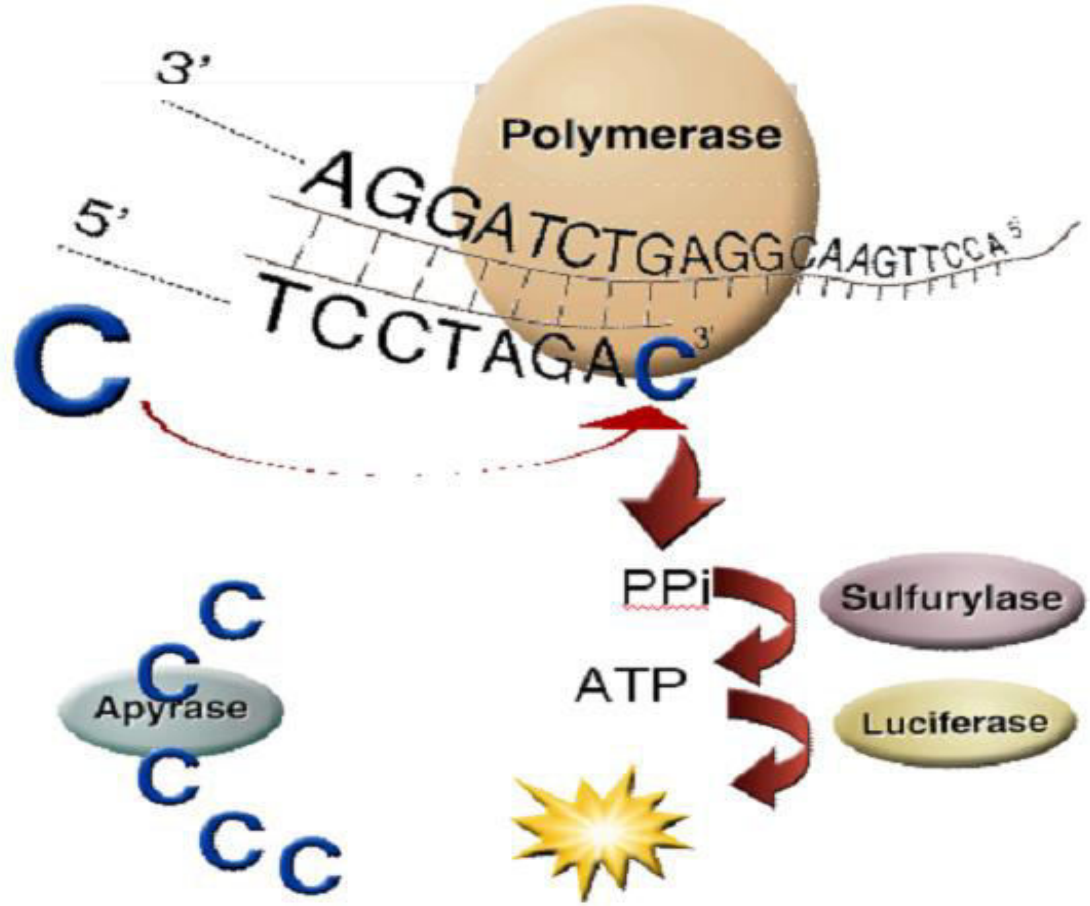
$$DNA_n + dNTP \rightarrow DNA_{n+1} + PP_i$$

454 Pyrosequencing

Principle

- PPi is used to generate ATP from adenosine phosphosulfate (APS)
  - APS + PP_I → ATP
- ATP and luciferase generate light by conversion of luciferin to oxyluciferin

454 Pyrosequencing



454 Pyrosequencing

Principle

- Only one of the four nucleotides will generate a light signal that will be recorded on a pyrogram

454 Pyrosequencing

loading

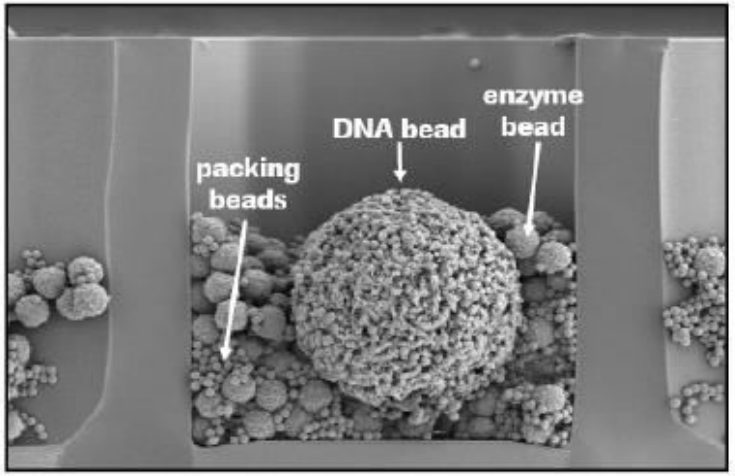
- After amplification, the emulsion shell is broken and the clonally amplified beads are ready for loading onto the fibre-optic PicoTiterDevice for sequencing

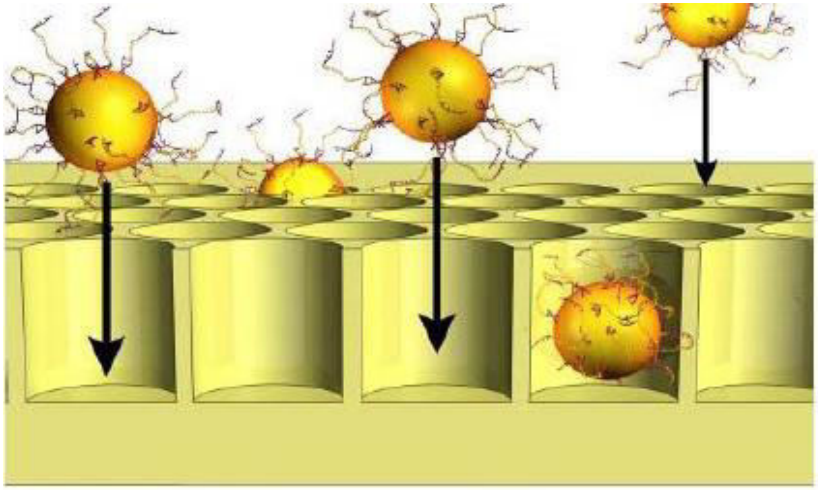
454 Pyrosequencing

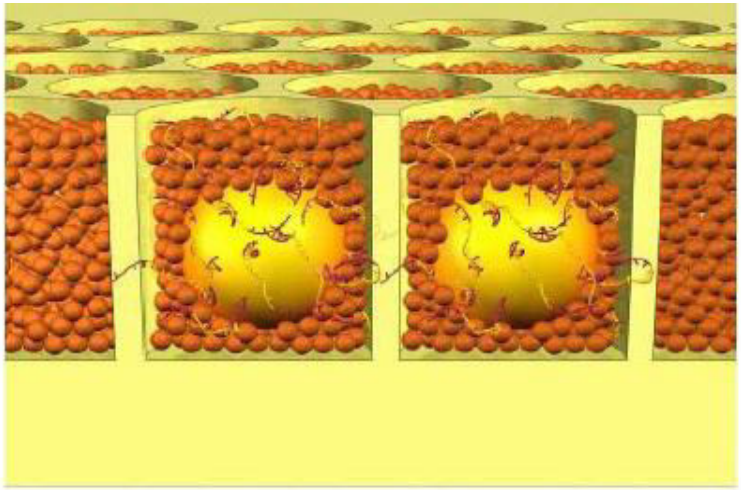
Sequencing

- The PicoTiterPlate is loaded with one fragment carrying bead per well and smaller beads with the necessary enzymes
- Each well is just big enough to hold a single bead

454 Pyrosequencing







454 Pyrosequencing

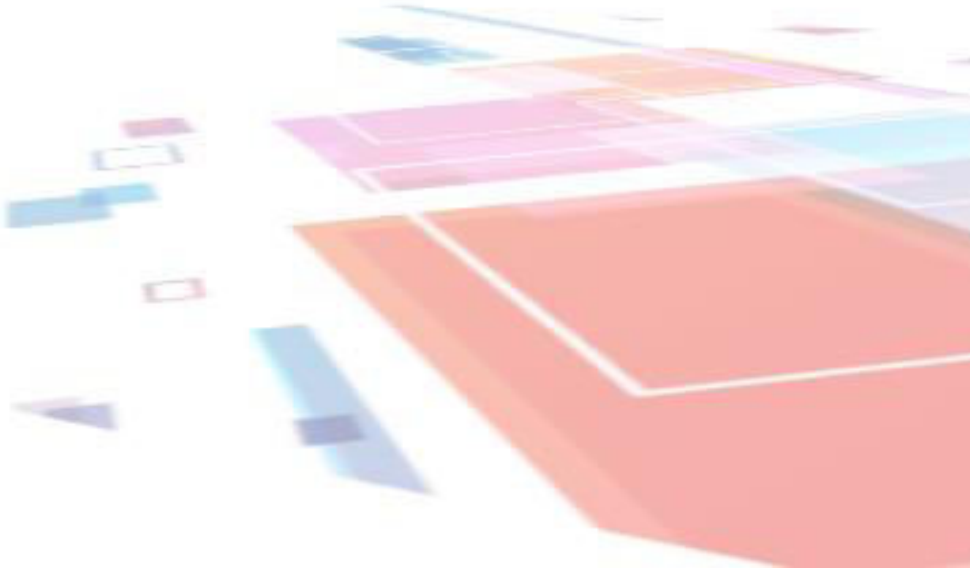
Sequencing

- The plate is repeatedly washed with each of the four dNTPs, plus other necessary reagents, in a repeating cycle
- The plate is coupled to a fiber optic chip

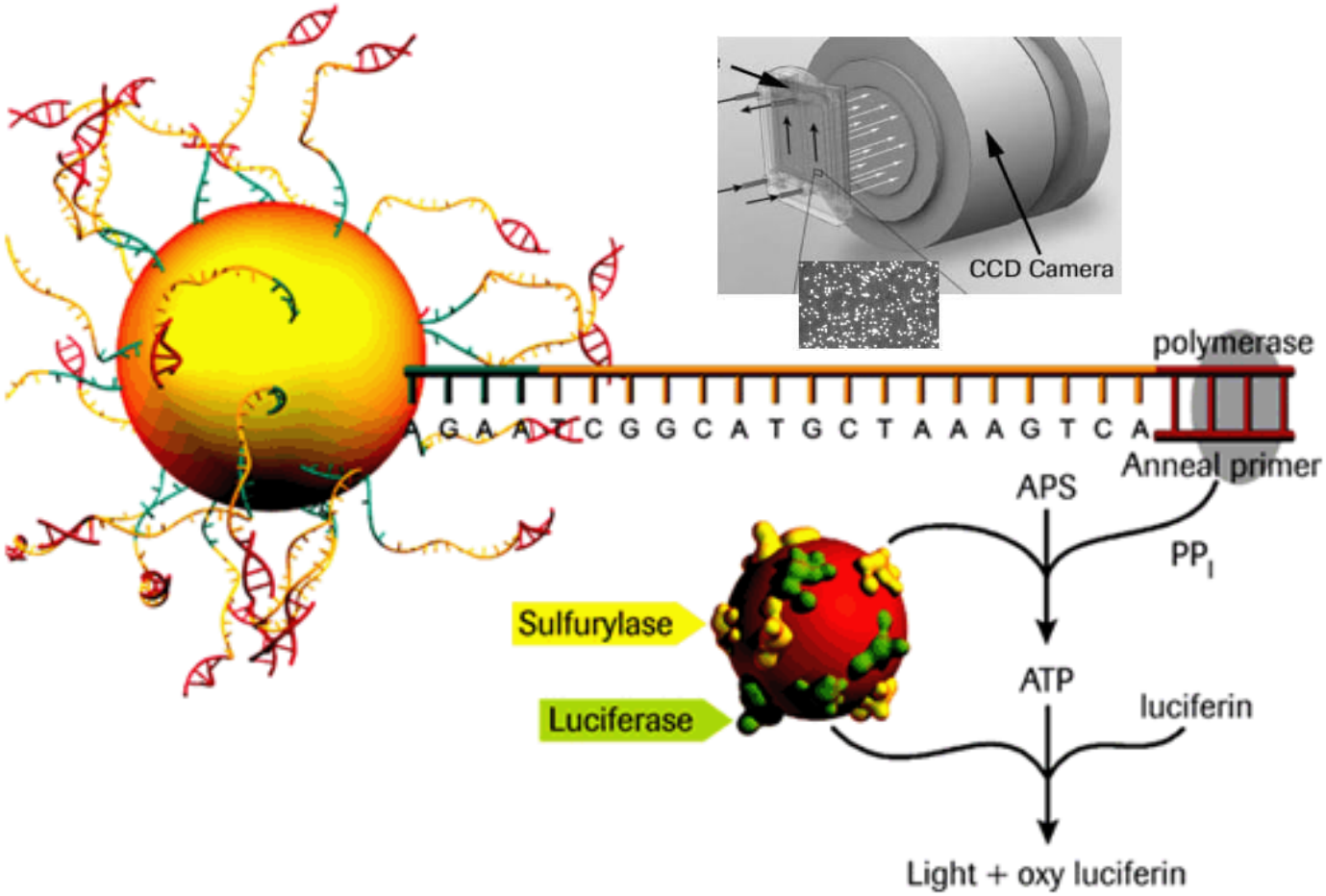
454 Pyrosequencing

Sequencing

- CCD camera records light flashes from each well

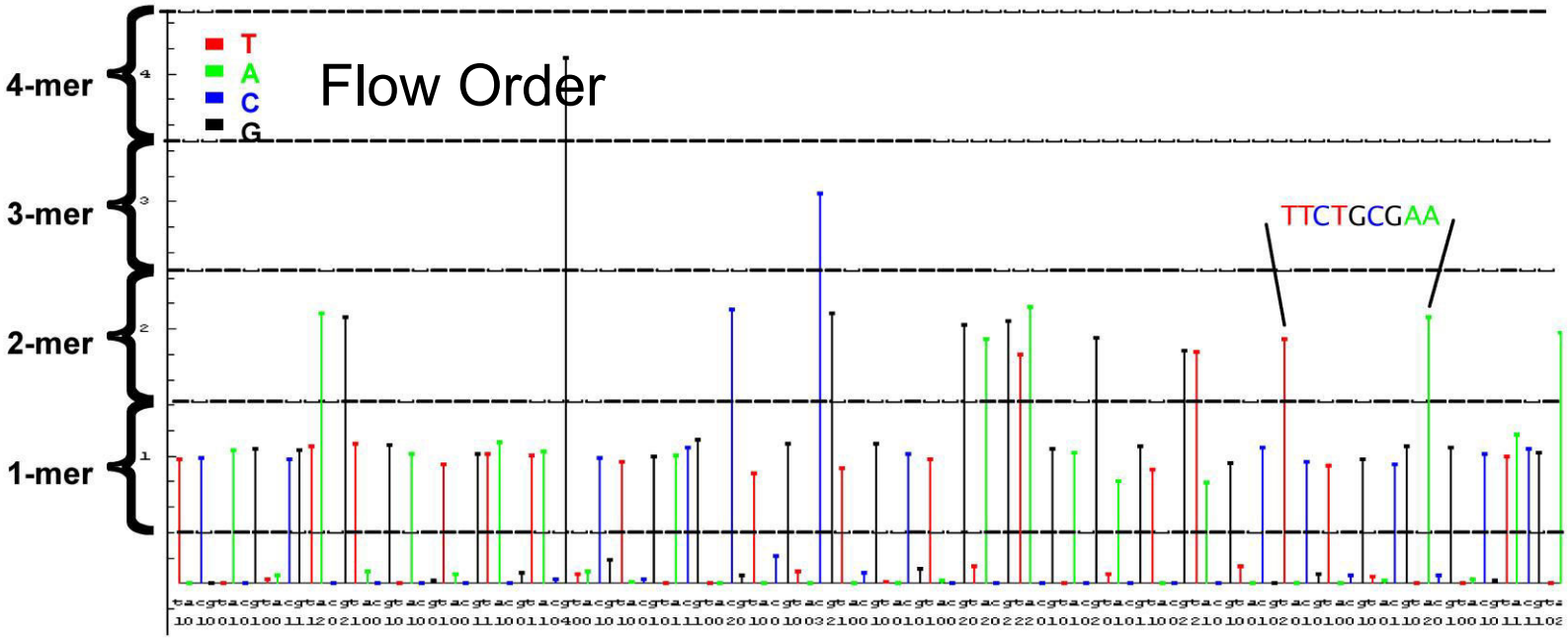


454 Pyrosequencing



454 Pyrosequencing

Flowgram

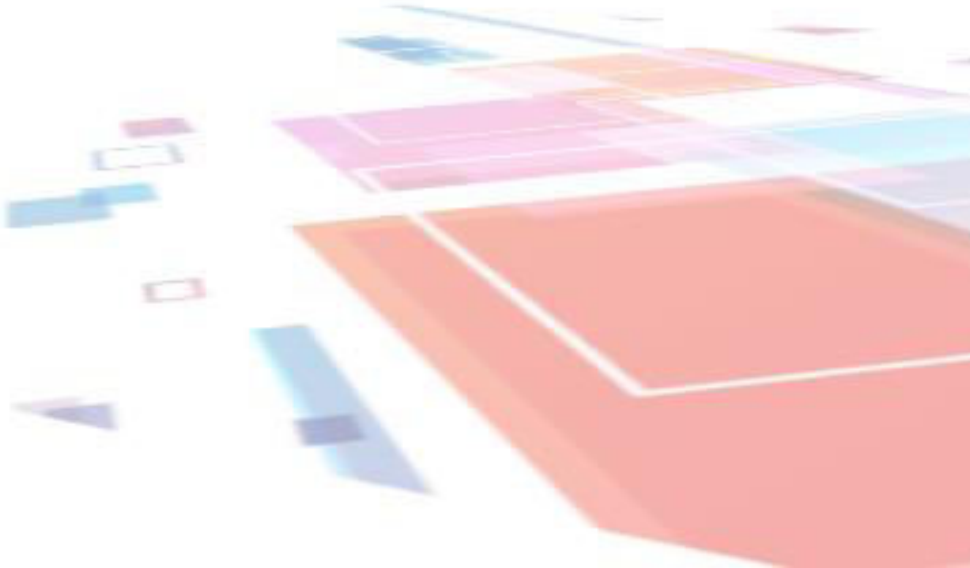


Key sequence = TCAG for signal calibration

454 Pyrosequencing

Conclusions

- Sequencing in 454 machines is performed by pyrosequencing



Resources (for this presentation)

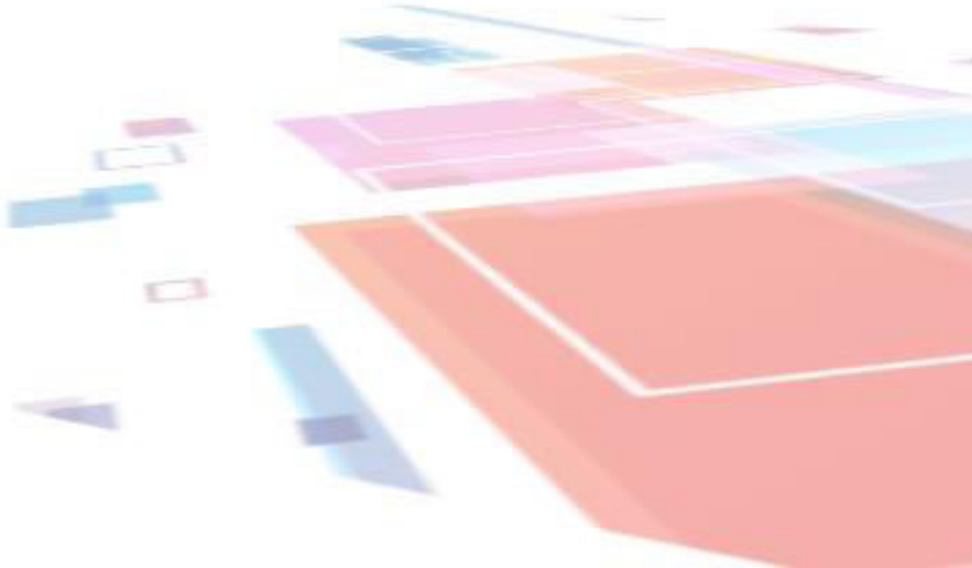
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

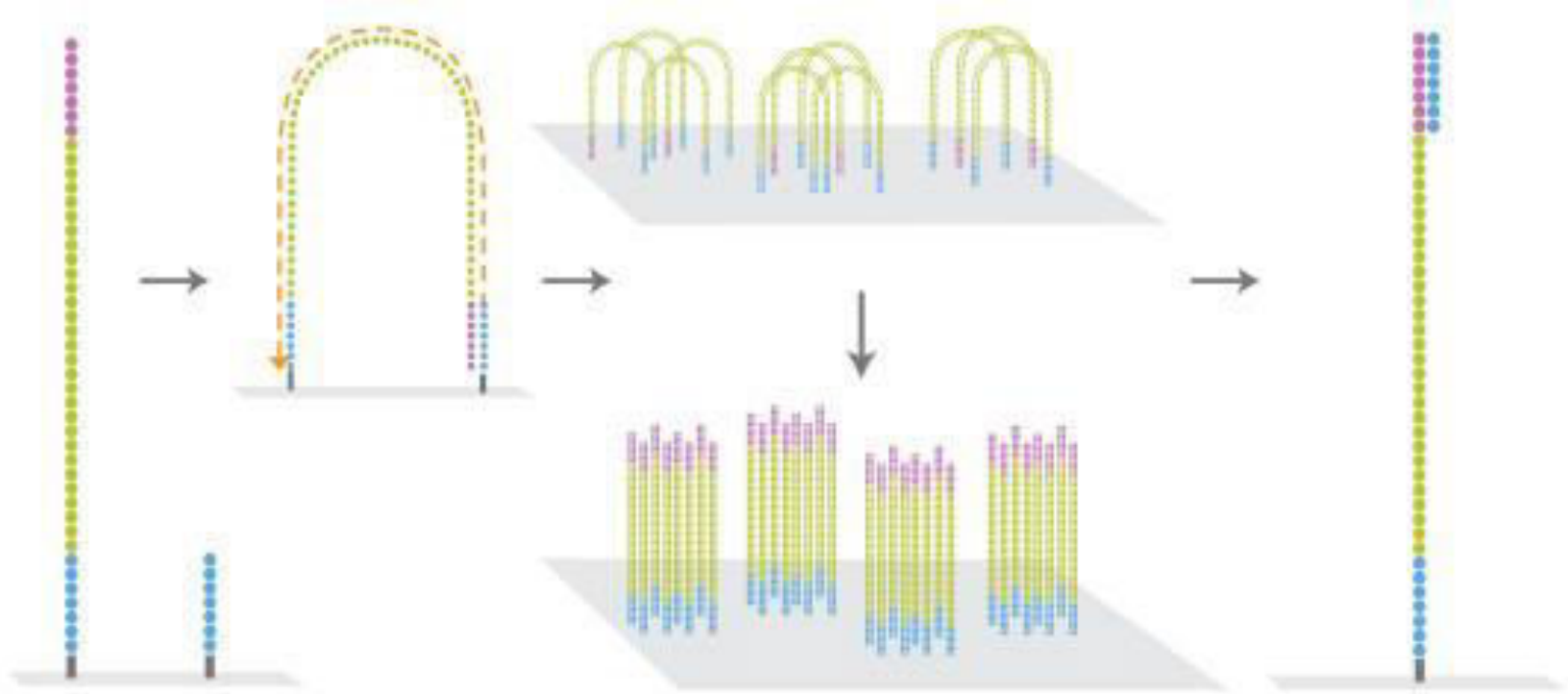
- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Illumina Sequencing

Illumina Sequencing

- After the amplification step, a flow cell with more than 40 million clusters is produced
- Each cluster with ~1000 clonal copies of a single template molecule

Illumina Sequencing

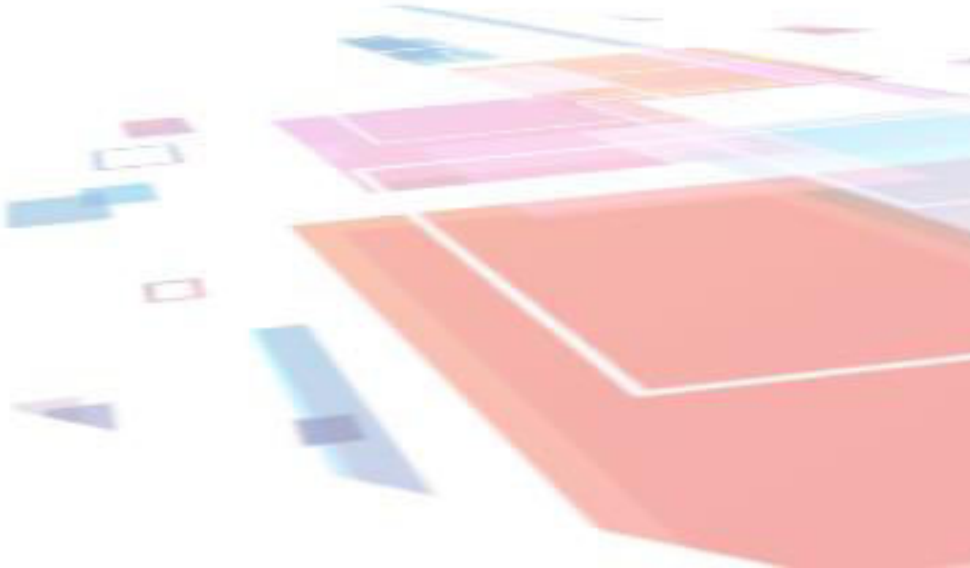


Illumina Sequencing

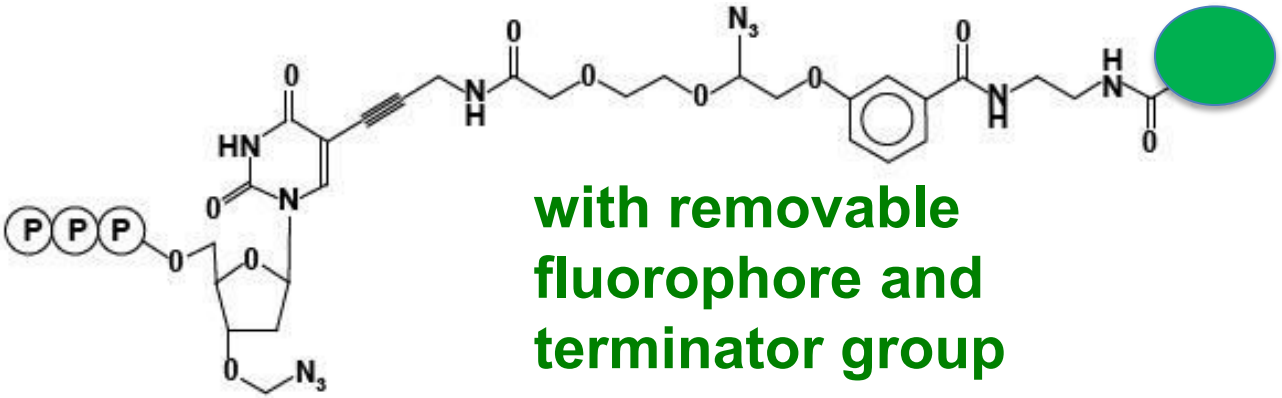
Sequencing

Illumina uses

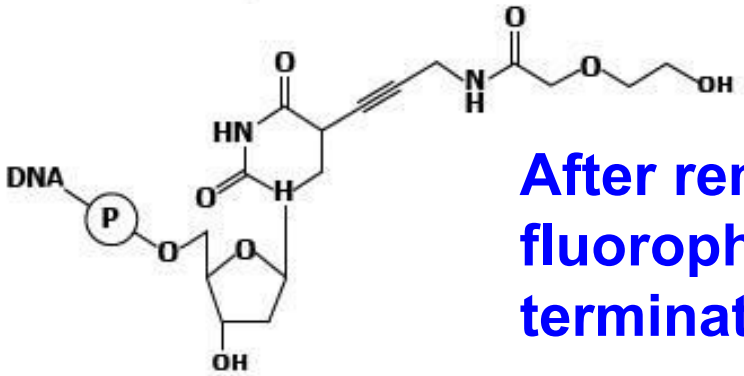
reversible terminators with flourophores



Illumina Sequencing



with removable fluorophore and terminator group



After removing fluorophore and terminator group

Illumina Sequencing

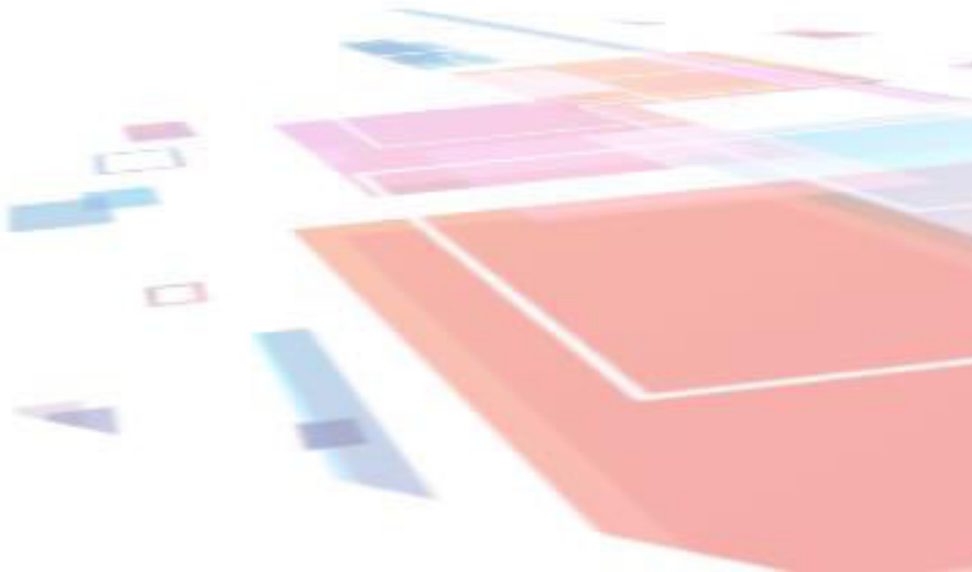
Sequencing

- All four labelled terminator nucleotides, primers and polymerase flow through each lane
- Base is incorporated
- Fluorescent signal
- Imaging

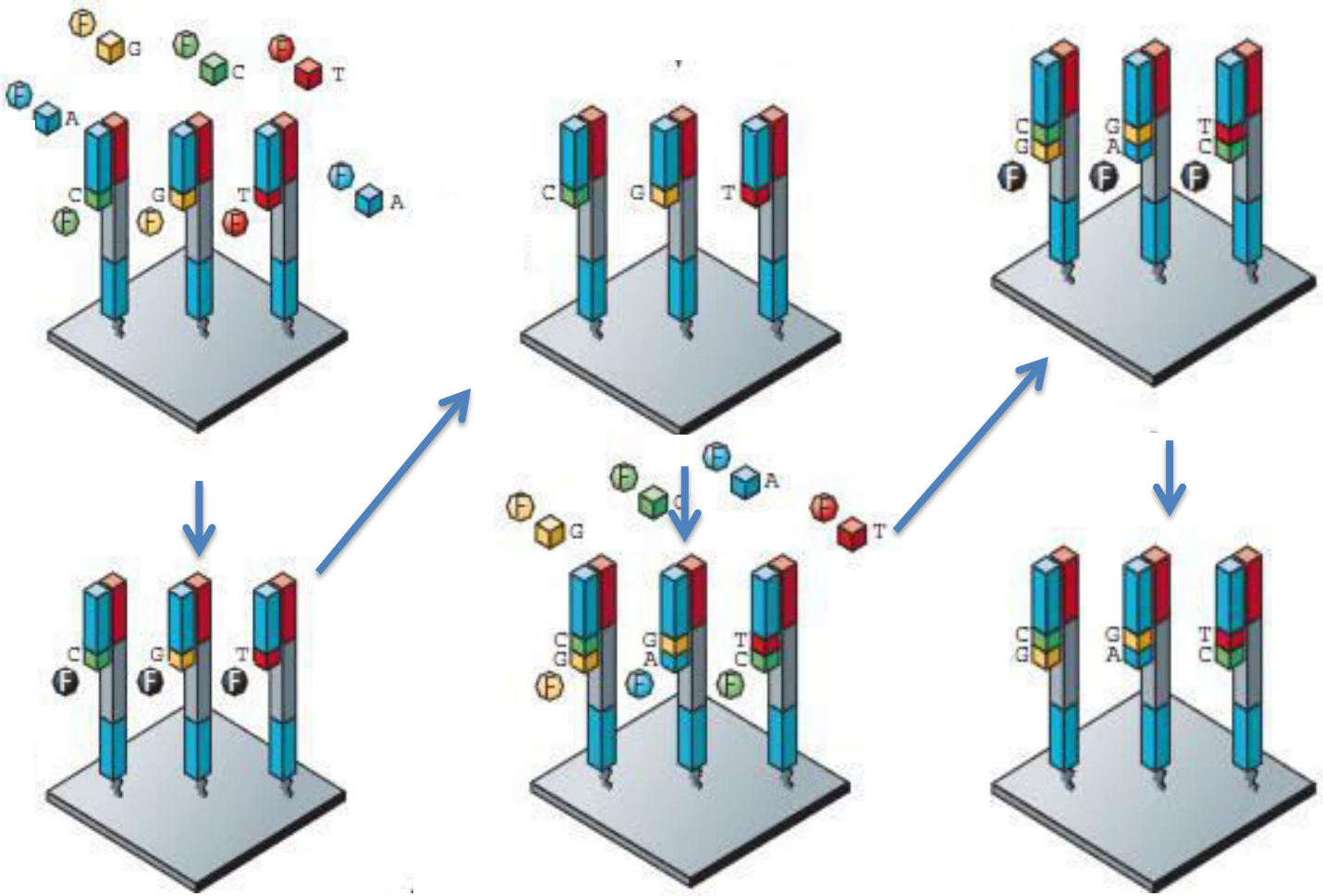
Illumina Sequencing

Sequencing

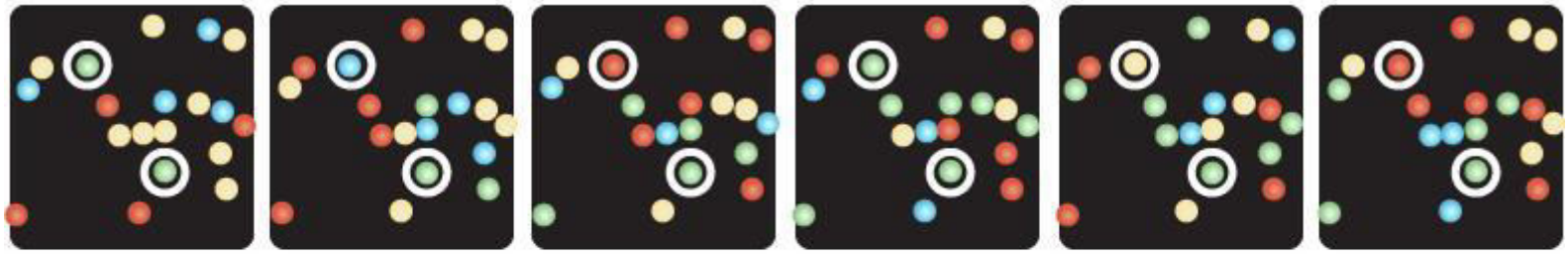
- Remove flourophore and block
- Repeat

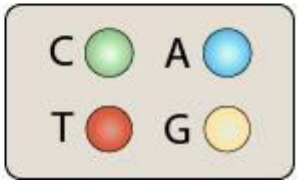


Illumina Sequencing



Illumina Sequencing





Top: CATCGT
Bottom: CCCCCC

Illumina Sequencing

Conclusions

- Illumina performs sequencing by synthesis (SBS) using reversible terminators with flourophores

Resources (for this presentation)

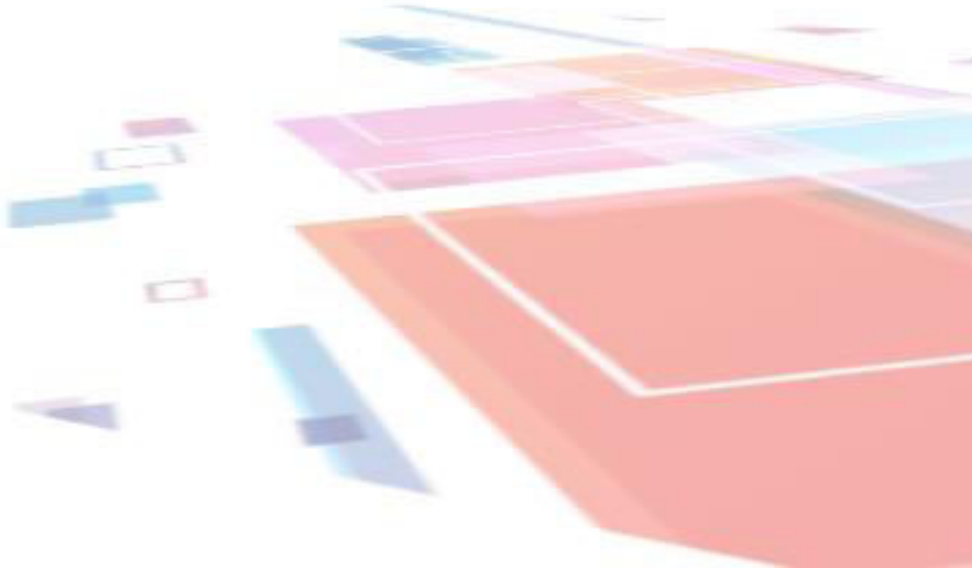
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Ion Semiconductor Sequencing

NGS Issues

- Alternating phases of nucleotide incorporation, signal detection and dephasing
- PCR biases
  - Duplicates, PCR errors

Ion Semiconductor Sequencing

Between 2nd and 3rd generation

Still use 'wash-and-scan'

- Ion Torrent
  - Non-optical sequencing
  - Emulsion PCR

Ion Semiconductor Sequencing

Between 2nd and 3rd generation

Still use 'wash-and-scan'

- Helicose Genetic Analysis System
  - Single molecule sequencing

Ion Semiconductor Sequencing

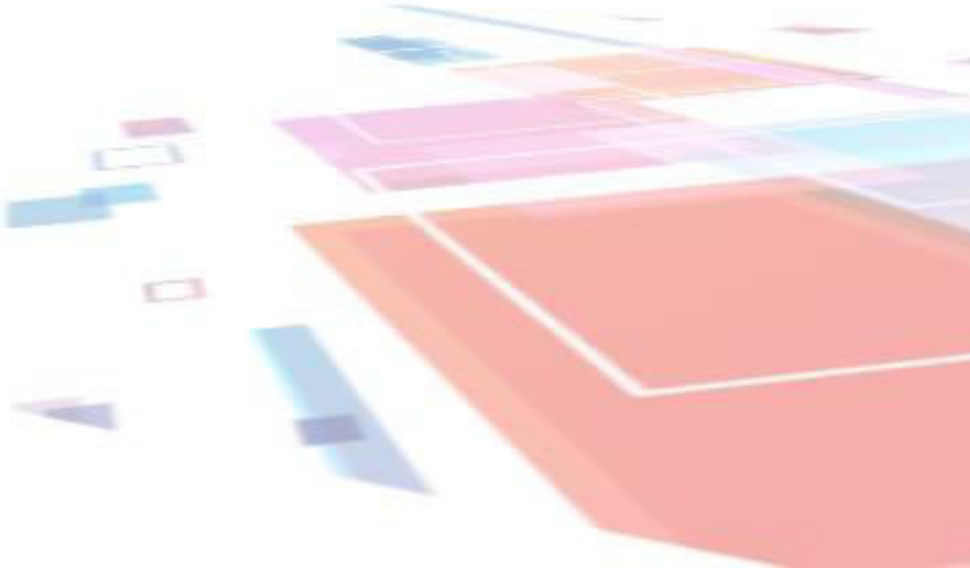
Semiconductor sequencing

Ion Semiconductor Sequencing

Semiconductor Chips

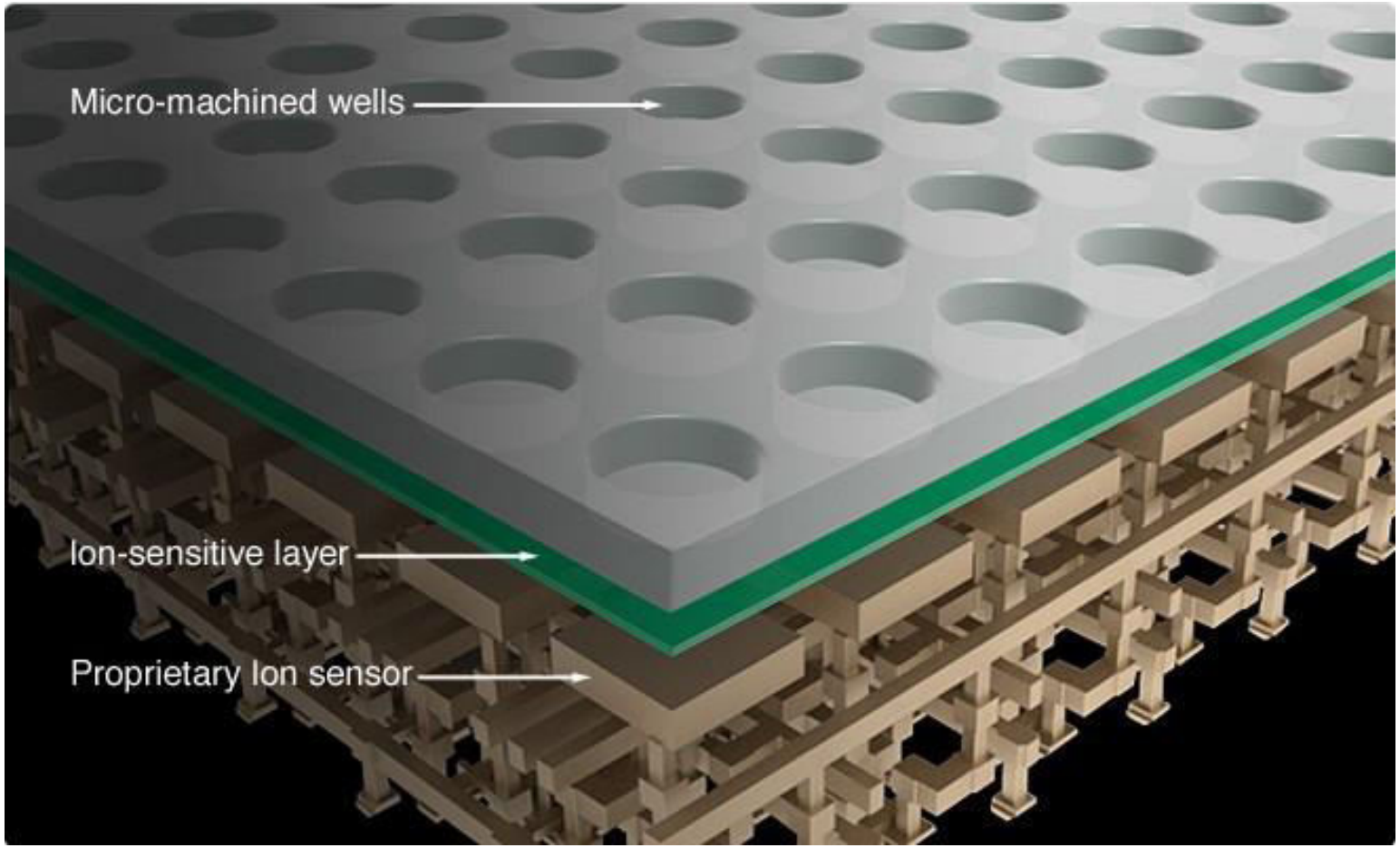
Three types of semiconductor chips:

- **314 – 20Mb**
- **316 - 200Mb**
- **318 – 1Gb**



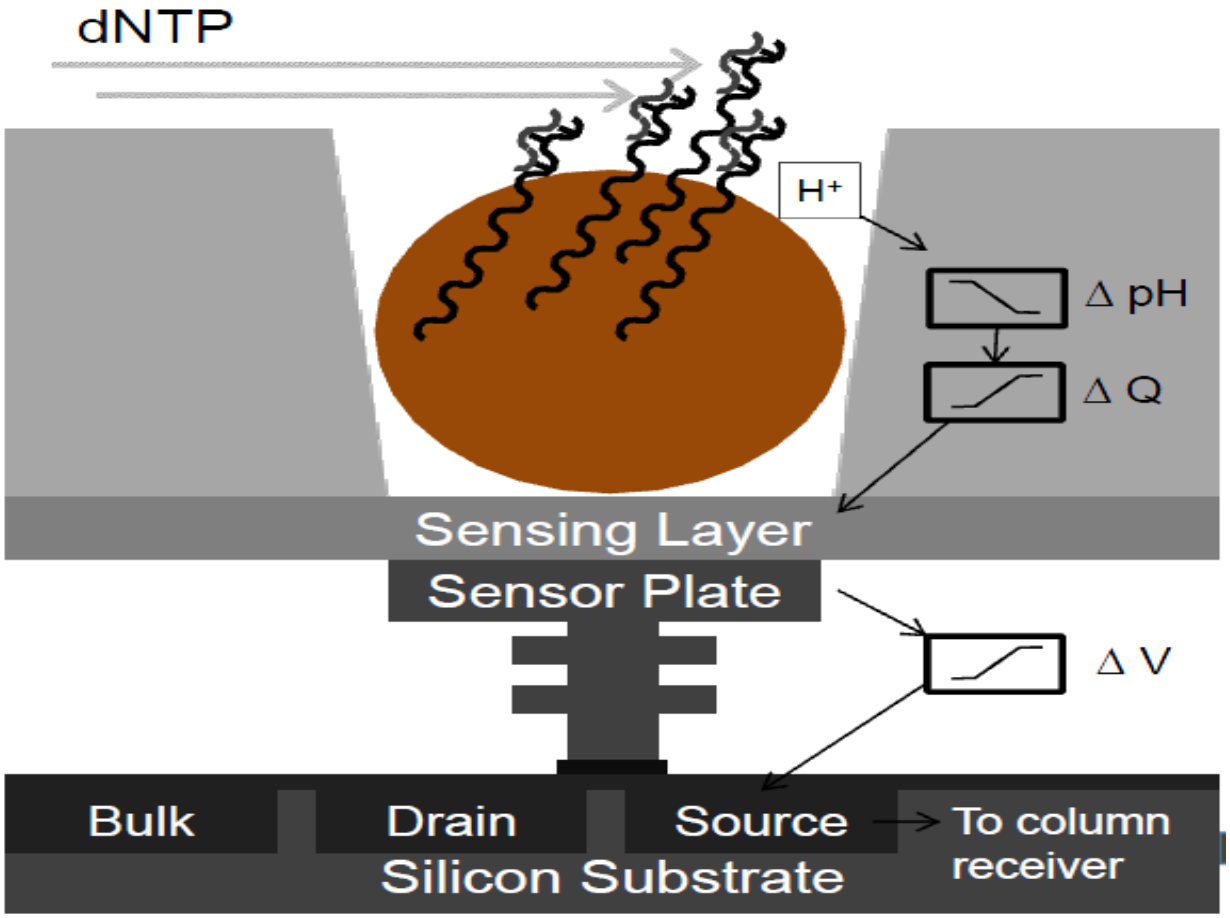
Ion Semiconductor Sequencing

Semiconductor sequencing



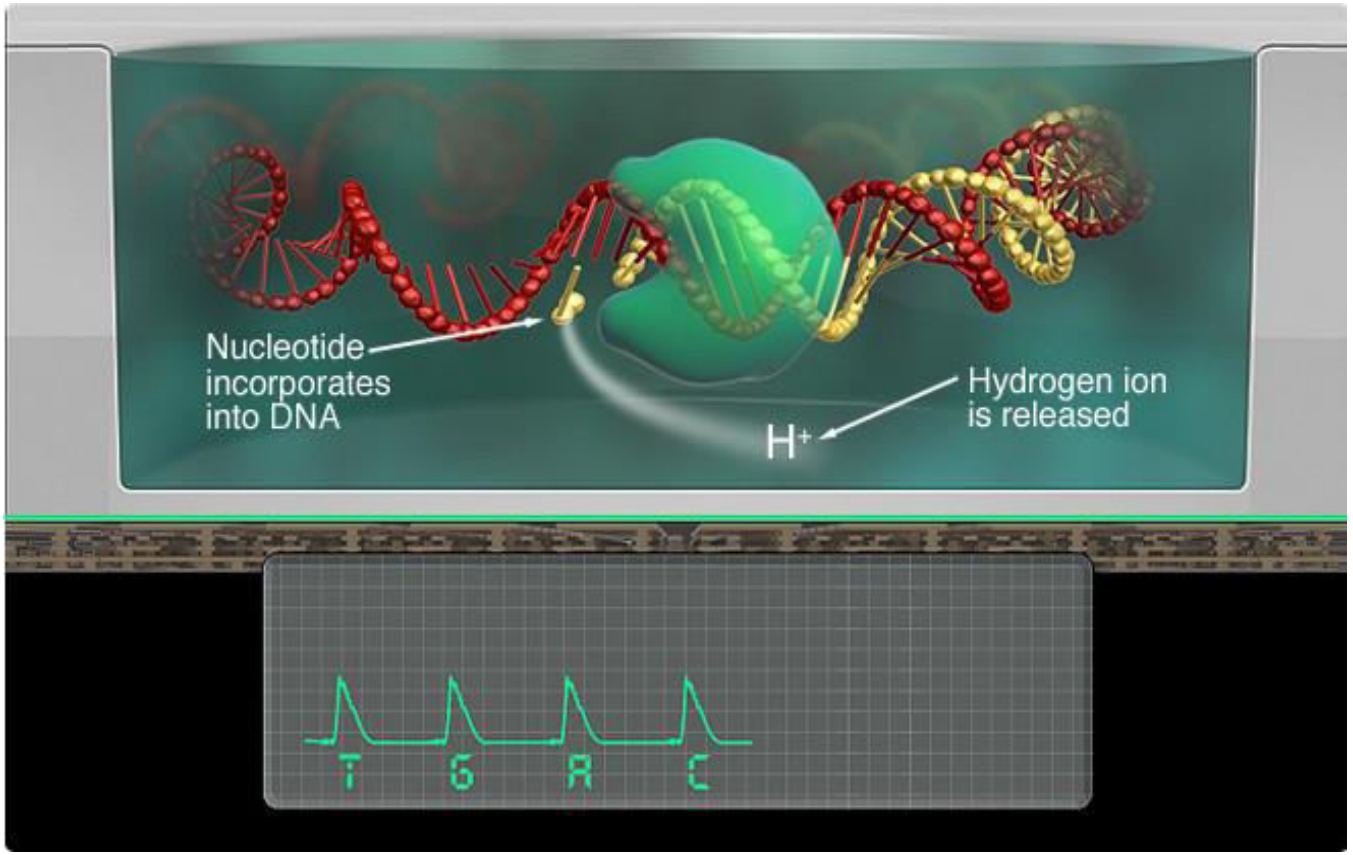
Ion Semiconductor Sequencing

Semiconductor sequencing



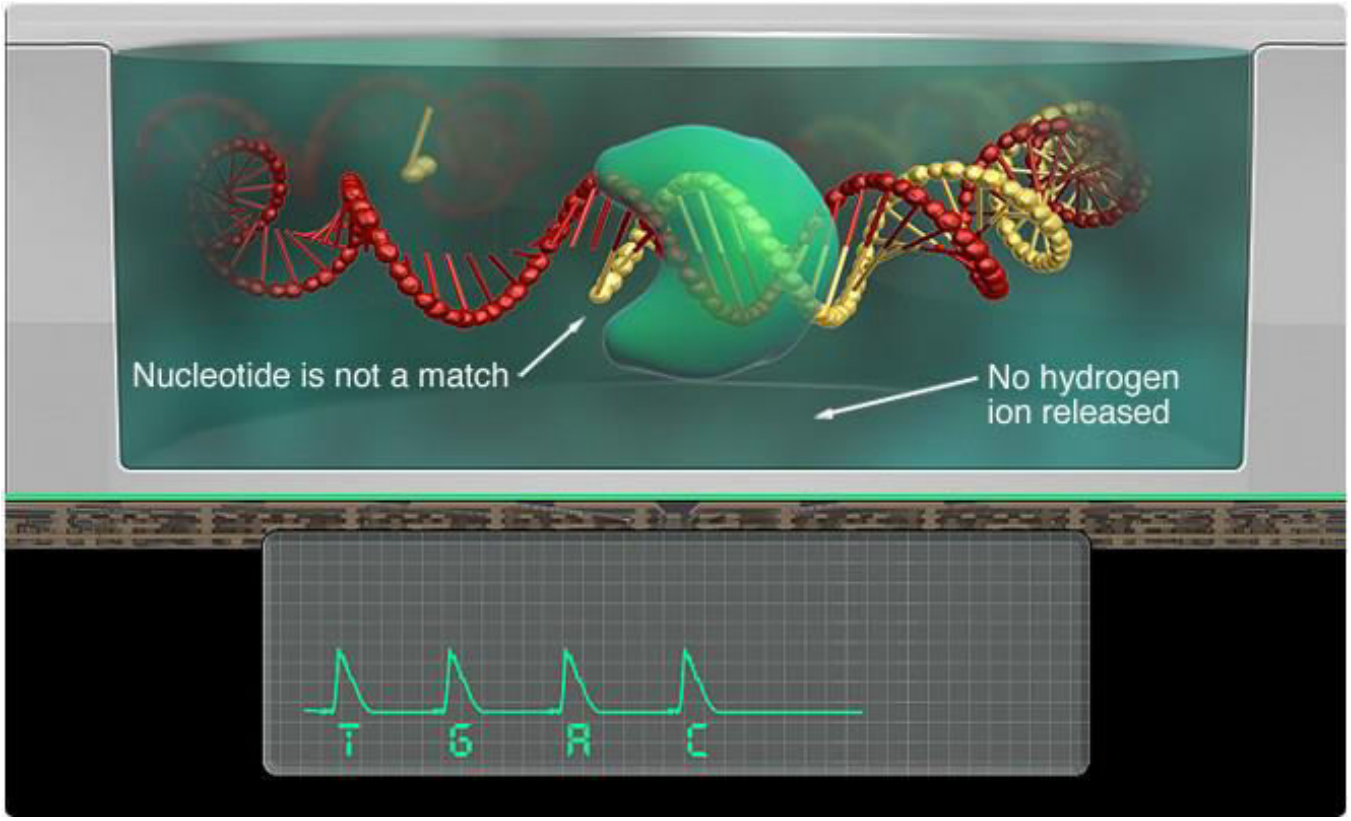
Ion Semiconductor Sequencing

Semiconductor sequencing



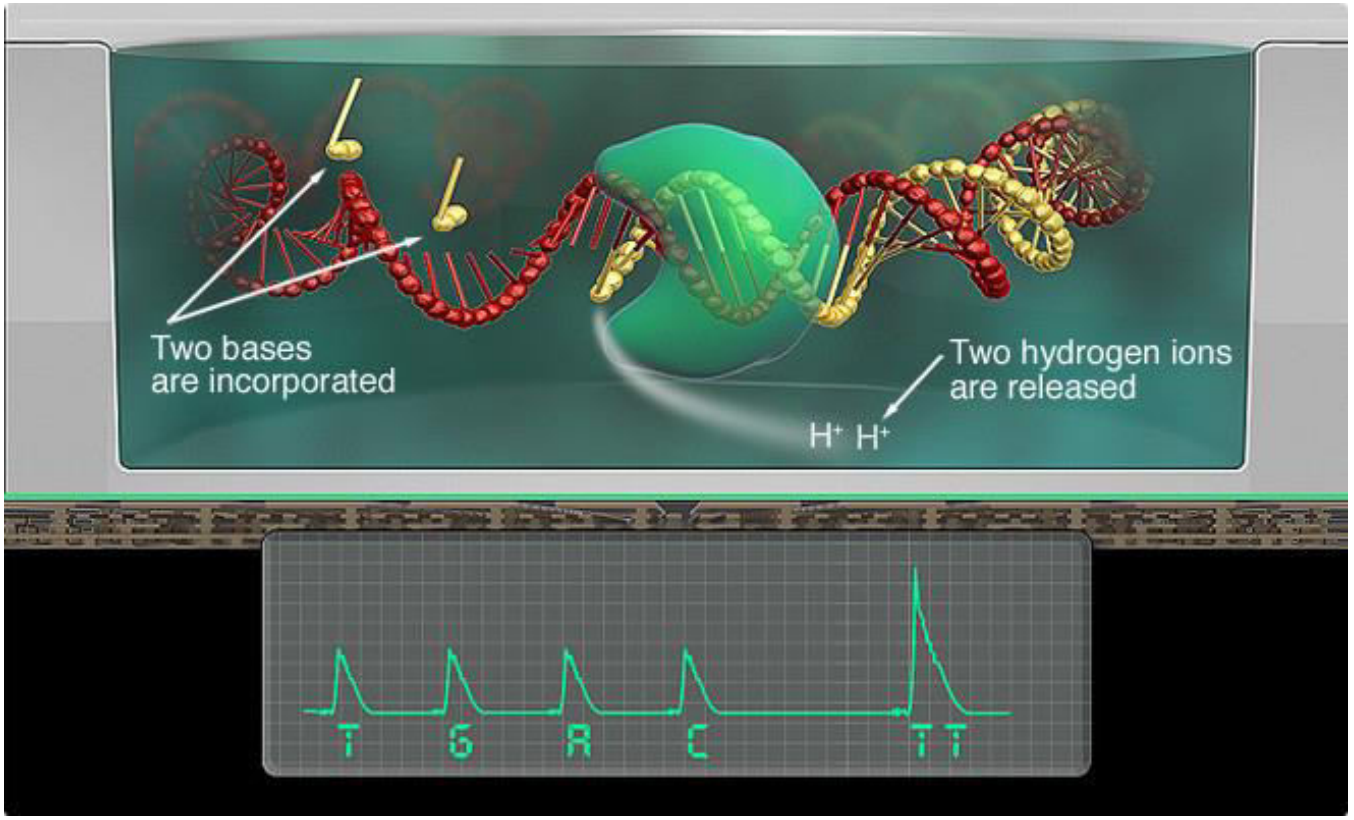
Ion Semiconductor Sequencing

Semiconductor sequencing



Ion Semiconductor Sequencing

Semiconductor sequencing



Resources (for this presentation)

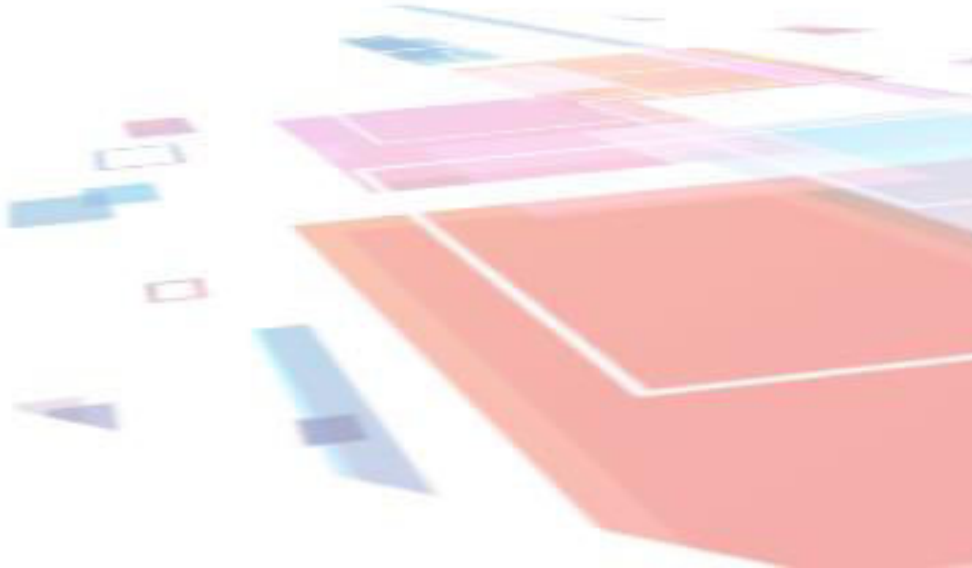
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

3rd Generation sequencing

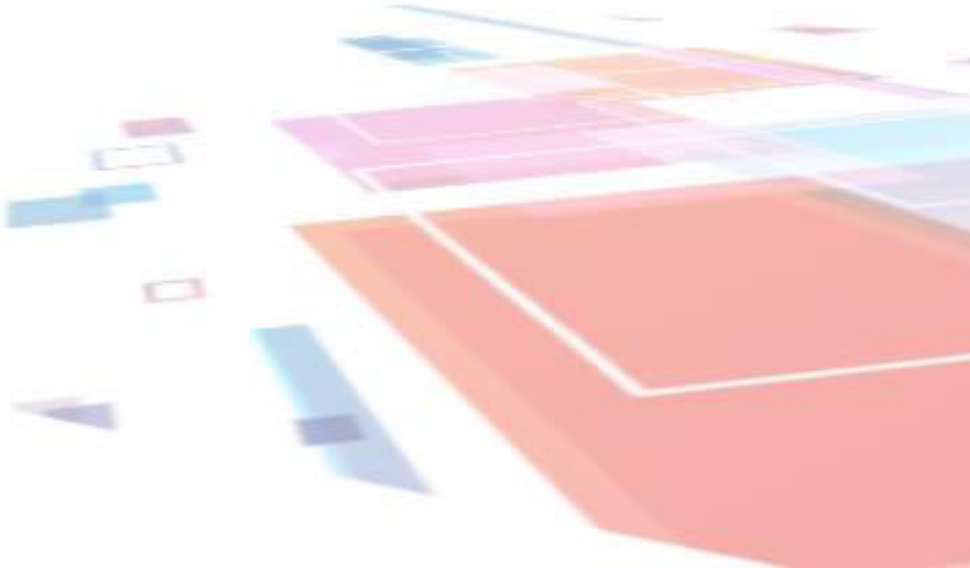
Introduction

- Single molecule sequencing
- No PCR biases
- Less input samples and reagents (no washing)
- Longer reads
- Better quality
- Higher throughput

3rd Generation sequencing

3GS Plateforms

- Helicos
- Pacific Biosciences
- Oxford nanopore



3rd Generation sequencing

Helicos

- A publically traded coy till June, 2012
- Cambridge Massachusetts
- Co founded by stanley Lapidus, Stephen Quake and Noubar Afeyan
- Delisted from NASDAQ in 2010

3rd Generation sequencing

Helicos Genetic Analysis System

Library preparation → Sequencing

3rd Generation sequencing

Helicos Genetic Analysis System

	Helicos
Read Length	35 bp
Throughput	35 Gb
Reads per run	600 M - 1000 M
Accuracy	97 %
Run Time	8 days

3rd Generation sequencing

Pacific Biosciences

- Pacbio RS
- Founded 2004
- Menlo Park, CA
- **SMRT** **S**ingle **M**olecule **R**eal **T**ime **S**equencing
- http://www.pacificbiosciences.com/

3rd Generation sequencing

Pacific Biosciences Pacbio RS

3rd Generation sequencing

Pacific Biosciences

	Pacbio RS
Read Length	3000 - 15,000 bp
Throughput	1 Gb
Reads per run	70,000
Accuracy	95 %
Run Time	30 minutes

3rd Generation sequencing

Oxford Nanopore

- Nanopore sequencing
- Spin out from Oxford University UK
- Since 2005
- had raised over £145 million in investment
- Products are under testing and evaluation phase

3rd Generation sequencing

Oxford Nanopore

MinION

GridION System

3rd Generation sequencing

Oxford Nanopore

	Nanopore
Read Length	48 kb ?
Throughput	? Gb
Reads per run	2000
Accuracy	75 %
Run Time	? minutes

3rd Generation sequencing

Advantages of single molecule sequencing

- Less sample preparation (no PCR)
- No amplification
- no PCR errors
- fewer contamination issues
- no GC-bias

3rd Generation sequencing

Advantages of single molecule sequencing

- analyze every sample (unPCRable / unclonable)
- analyze low quality DNA (museum, archeological, forensics samples)
- Absolute quantification
- Sequence RNA directly

Resources (for this presentation)

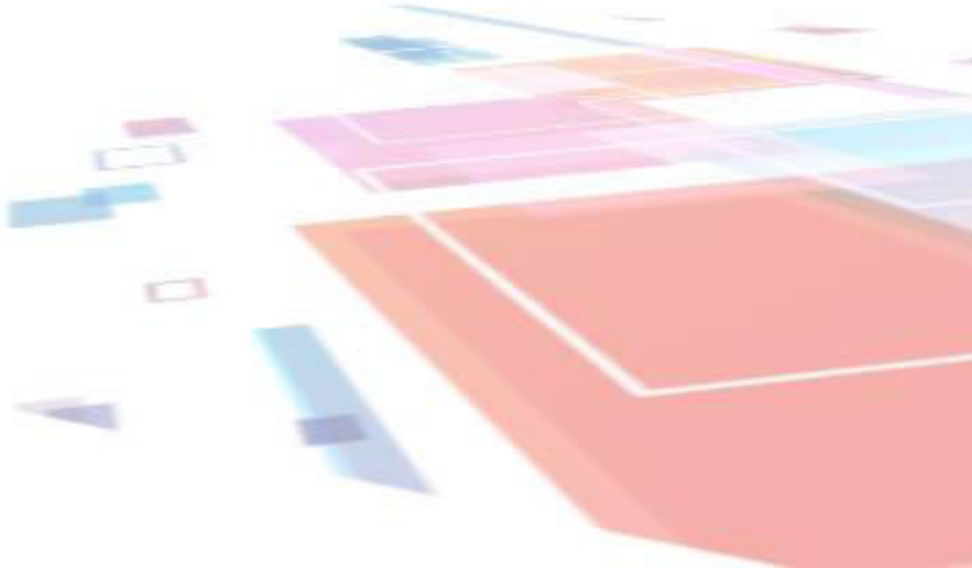
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Helicose sequencing

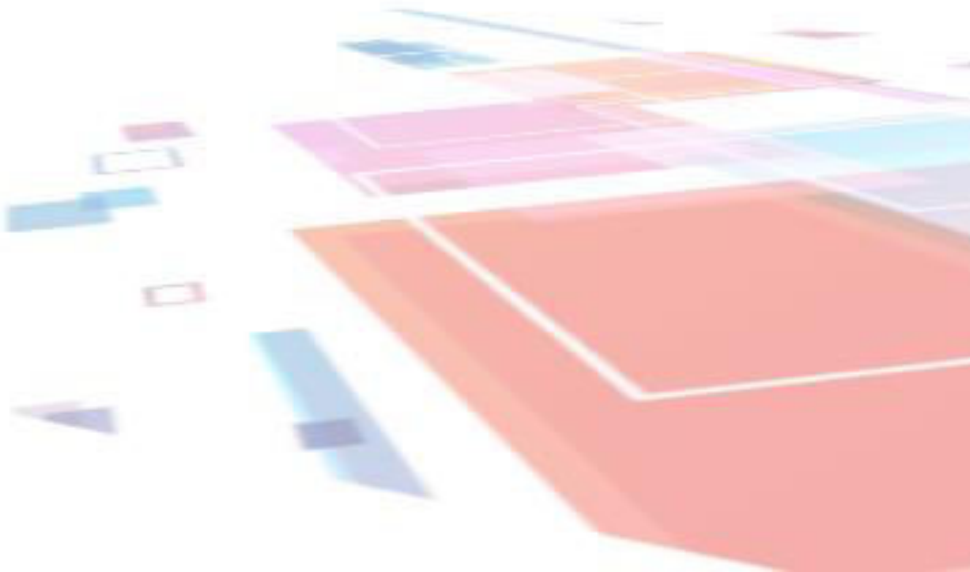
Helicos

- Images the extension of individual DNA molecules using a defined primer and individual fluorescently labeled nucleotides, which contain a "Virtual Terminator" preventing incorporation of multiple nucleotides per cycle

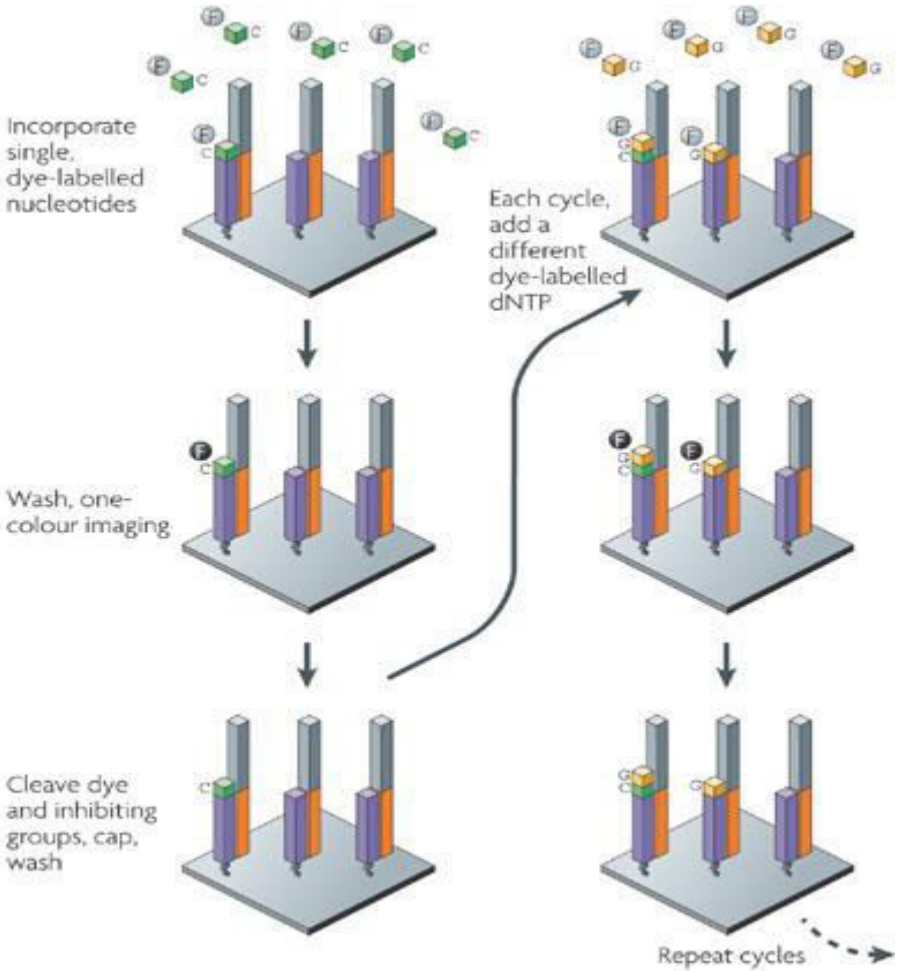
Helicose sequencing

Helicos

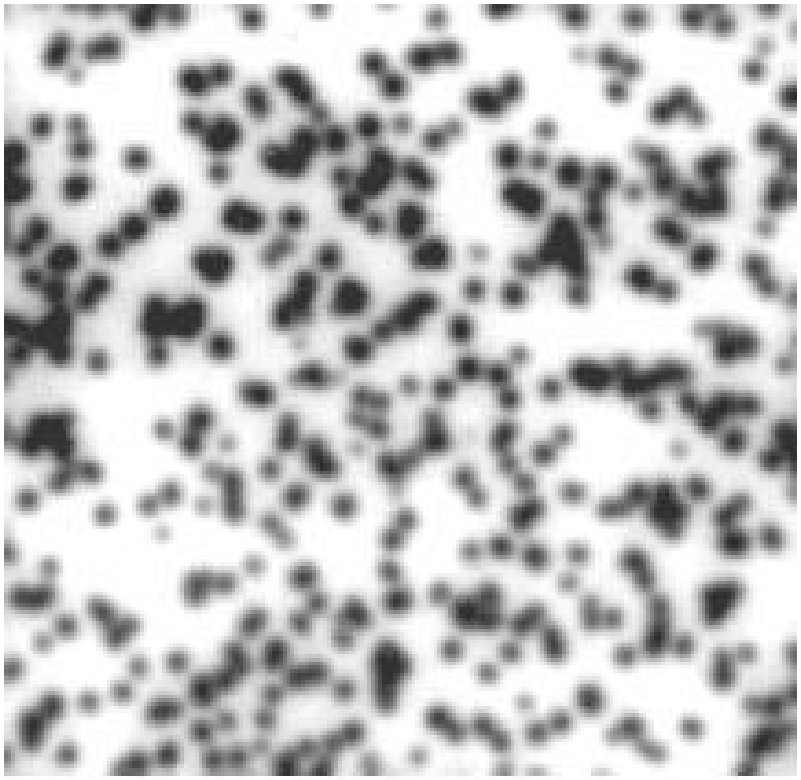
- The "Virtual Terminator" technology was developed by Dr. Suhaib Siddiqi



Helicos sequencing

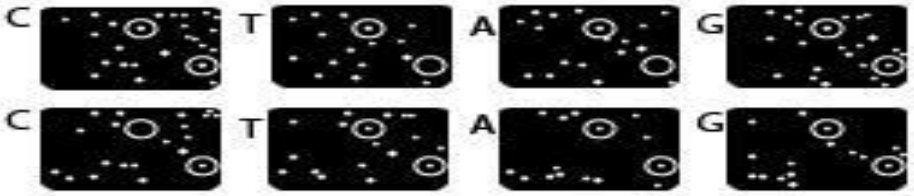


Helicos Sequencing



http://seq.molbiol.ru/

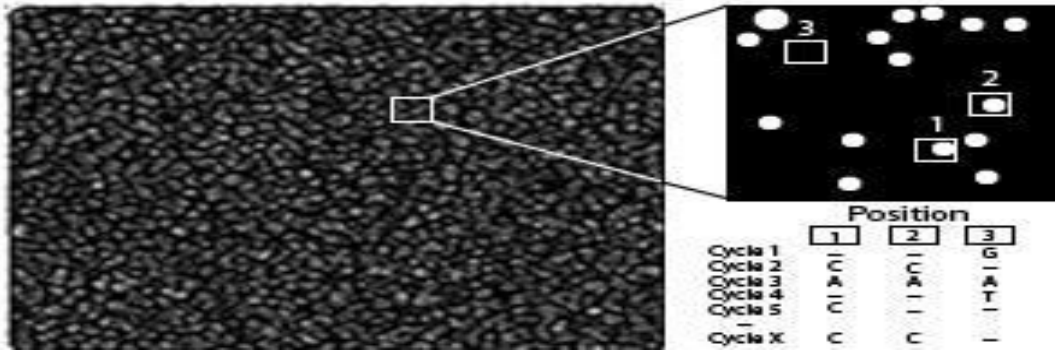
Helicos sequencing



Top: CTAGTG
Bottom: CAGCTA

Nucleotides flown sequentially

(Dark nucleotides: incorporation not detected)



Helicos sequencing

Conclusions

- Helicos was an effort towards single molecule sequencing using illumina like approach
- It was an important step towards third generation sequecing

Resources (for this presentation)

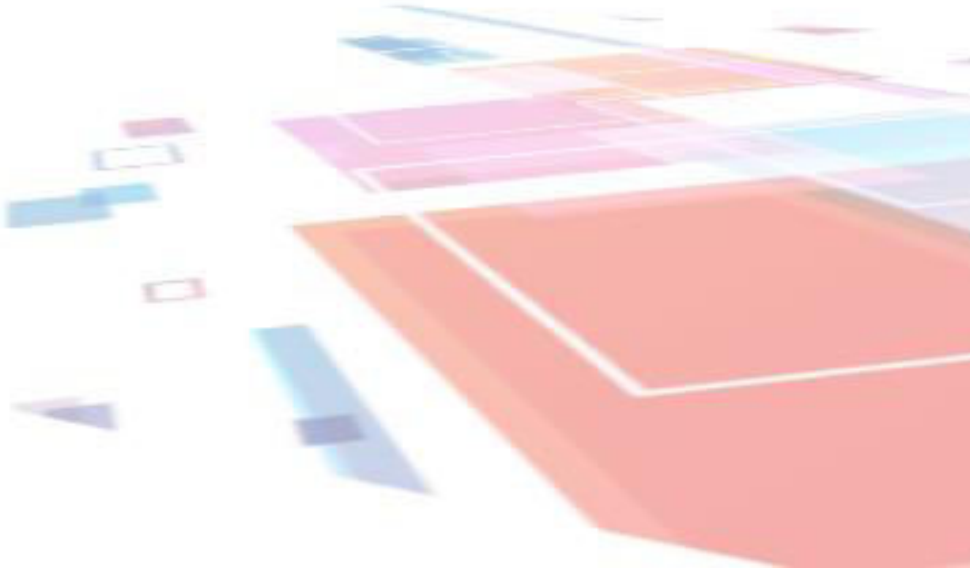
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Pacbio sequencing

SMRT sequencing

- Single Molecule Real Time Sequencing
- utilizes the zero-mode waveguide (ZMW), developed in the laboratories of Harold G. Craighead and Watt W. Webb at Cornell University

Pacbio sequencing

SMRT sequencing

- ZMW guides light energy into a volume that is small in all dimensions compared to the wavelength of the light
- Nanophotonic structure, a circular hole in an aluminum film on a silica substrate

Pacbio sequencing

SMRT sequencing

- With an active polymerase immobilized at the bottom of each ZMW, nucleotides diffuse into the ZMW chamber.
- A, C, G and T are labeled with a different fluorescent dye having a distinct emission spectrum

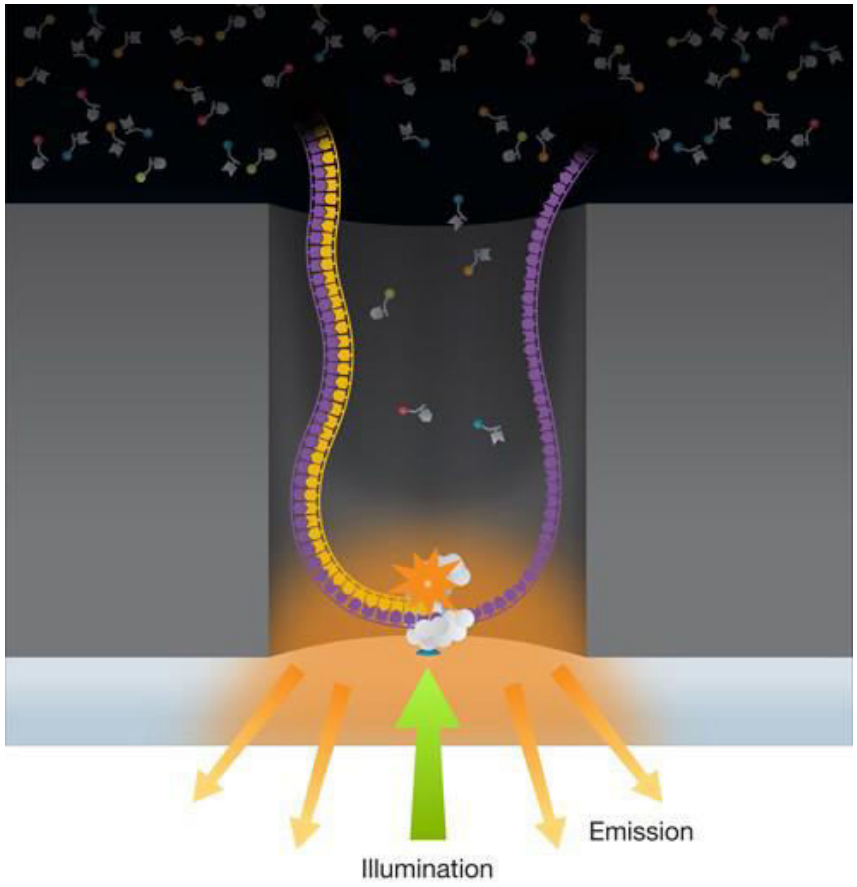
Pacbio sequencing

SMRT sequencing

- Nucleotides held by the polymerase prior to incorporation emit an extended signal that identifies the base being incorporated

Pacbio sequencing

SMRT sequencing



Pacbio sequencing

SMRT sequencing

- Single Molecule Real Time Sequencing
- utilizes the zero-mode waveguide (ZMW), developed in the laboratories of Harold G. Craighead and Watt W. Webb at Cornell University

Resources (for this presentation)

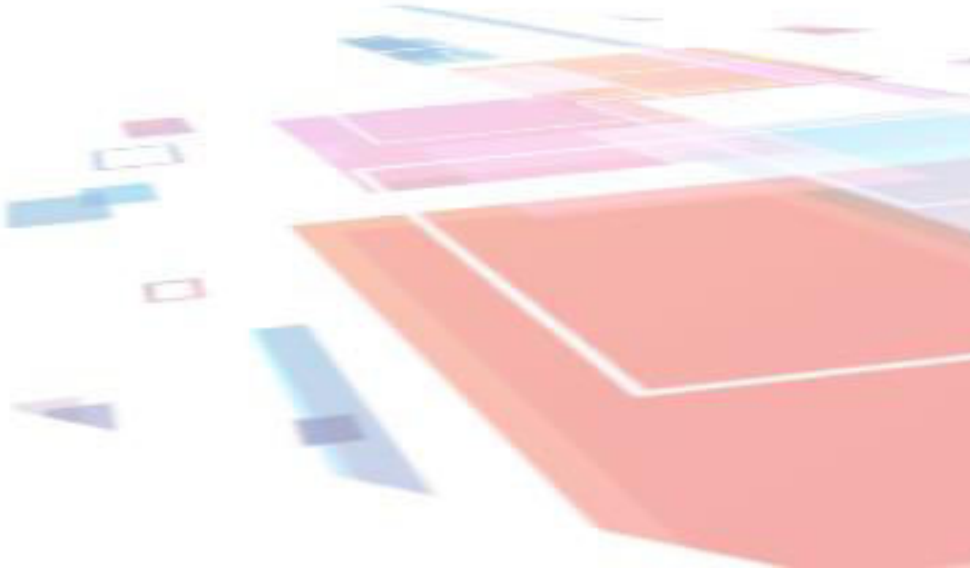
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Oxford Nanopore sequencing

Nanopore sequencing

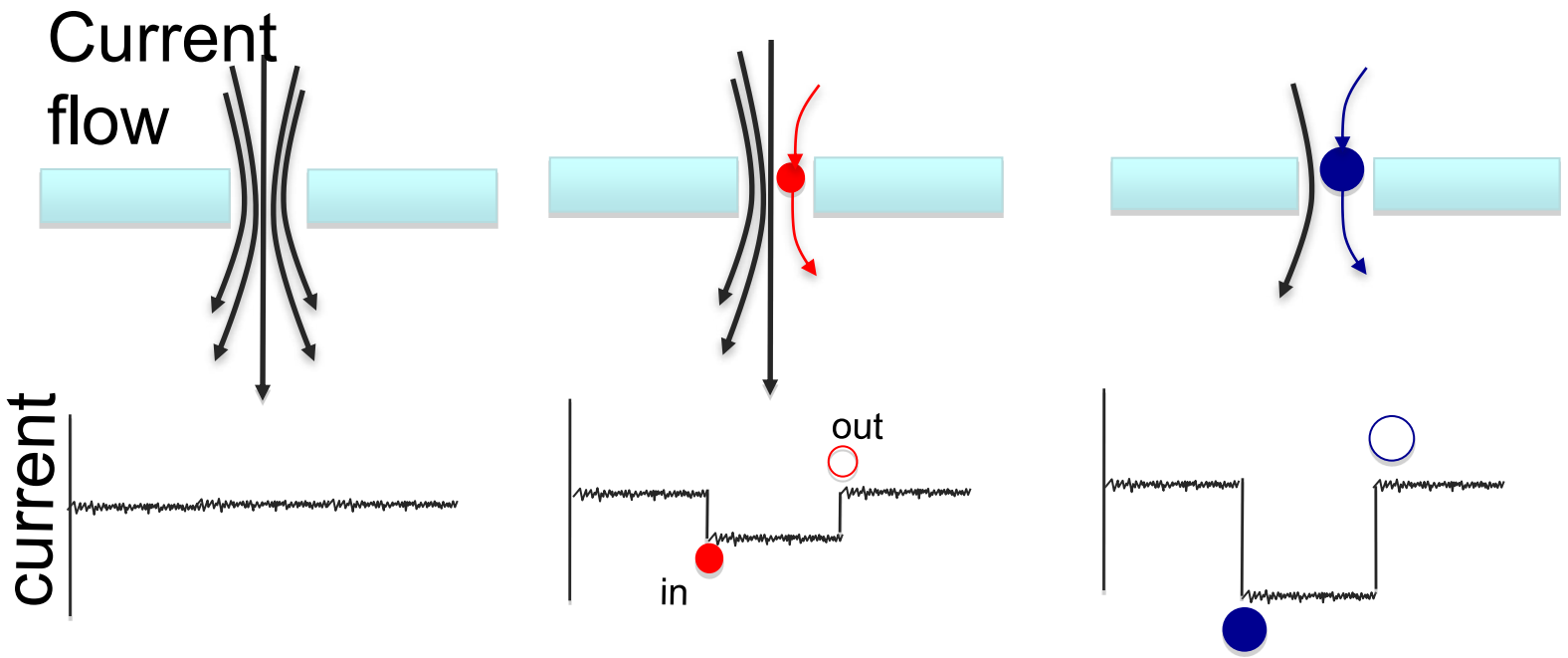
- Latest sequencing technology in development
- Size of USB drive
- May drive the next revolution in genomics

Oxford Nanopore sequencing

Nanopore sequencing

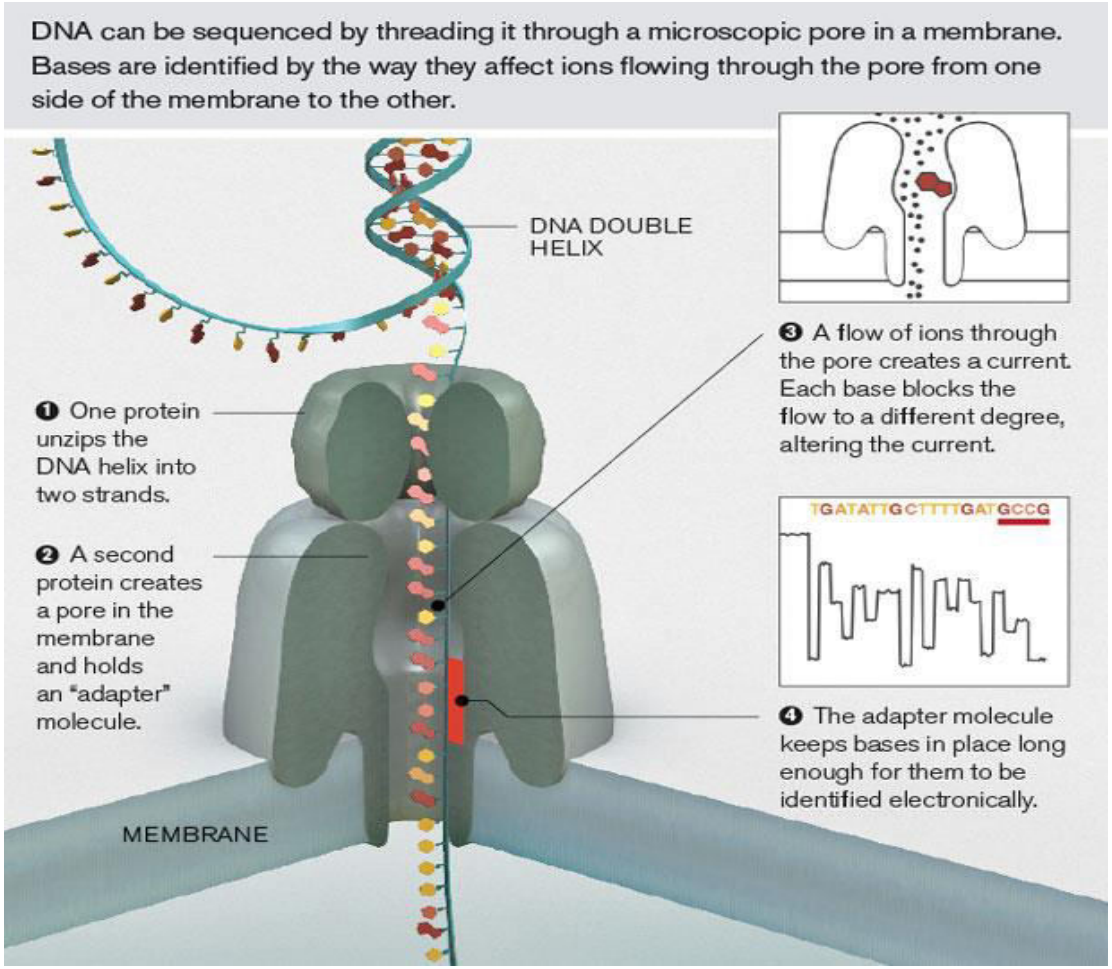
- Whole genome sequencing in 15 minutes for less than $1,000
- Expected to be available by the end of this year

Oxford Nanopore sequencing



Oxford Nanopore sequencing

Oxford Nanopore



Oxford Nanopore sequencing

Conclusions

- Based on sequencing via change in electric pulse caused by interaction of ions with nucleotides
- Expected to be available by the end of this year

Resources (for this presentation)

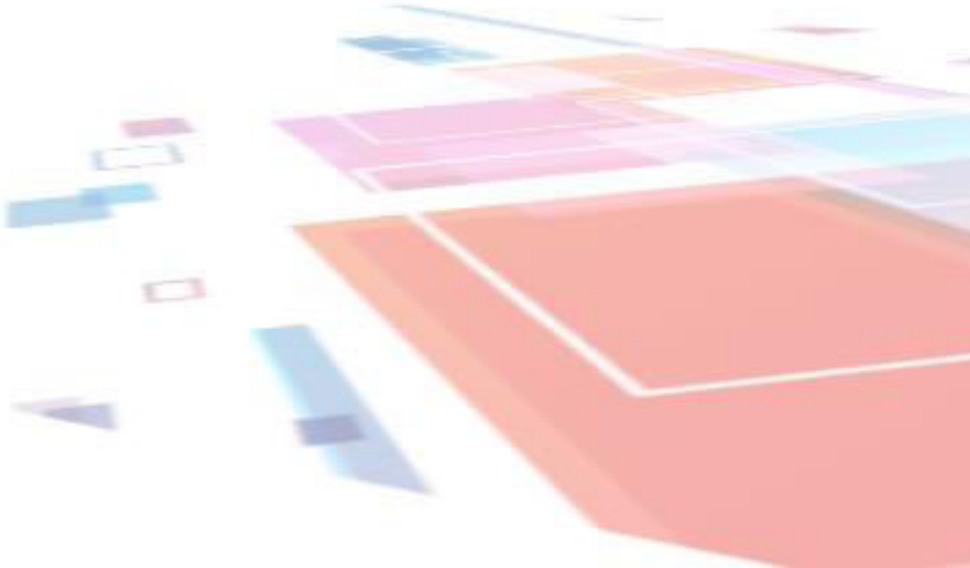
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://technology.illumina.com/technology/next-generation-sequencing.ilmn

Resources (for this presentation)

- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre

Resources (for this presentation)

- http://www.nature.com/subjects/next-generation-sequencing#research-and-reviews



Resources (for this presentation)

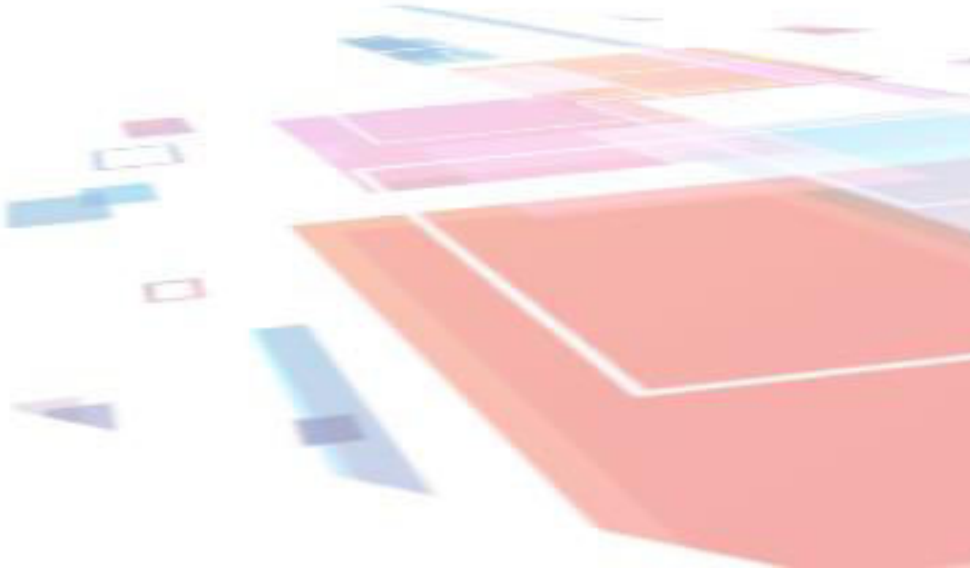
- https://www.lifetechnologies.com/pk/en/home/life-science/sequencing/next-generation-sequencing.html

Comparison of NGS Methods

Which Plate-form to use

Depends on

- Biological question
- Budget
- Analysis team's expertise



Comparison of NGS Methods

Method	Sanger	454	Illumina	SOLiD
Read length	400-900 bp	400-700 bp	100-250 bp	75 bp
Accuracy	99.9%	99.9%	98%	99.9%
Reads per run	N/A	0.1-1 million	up to 3 billion	1.2 to 1.4 billion
Time per run	20 min - 3 hours	10-24 hours	1- 10 days	1 - 2 weeks

Comparison of NGS Methods

	Sanger	454	Illumina	SOLiD
Cost/Mb	$2400	$10	$0.05 to $0.15	$0.13
Advantages	Long reads.	Long read size. Fast.	Potential for high sequence yield,	Low cost per base.
Drawbacks	impractical for larger sequences	expensive. Homopolymer errors.	Equipment expensive.	Slower than other methods.
Instrument Cost	3730XL $95K	$500 K	128K-MiSeq 650K-HiSeq	$495K

Comparison of NGS Methods

Method	Sanger	Illumina	Helicose	PacBio
Read length	400-900 bp	100-250 bp	35 bp	3000 bp
Accuracy	99.9%	98%	99%	99%
Reads per run	N/A	up to 3 billion	1 billion	1 billion
Time per run	20 min - 3 hours	1- 10 days	8 days	30 min

Comparison of NGS Methods

	Sanger	Illumina	Helicose	PacBio
Cost/Mb	$2400	$0.05 to $0.15	2$?	?
Advantages	Long reads.	Potential for high yeild	Long reads	Longest reads
Drawbacks	impractical for larger sequences	Equipment expensive.	Equipment expensive	Yield vs Quality
Instrument Cost	3730XL $95K	128K-MiSeq 650K-HiSeq		$695K

Comparison of NGS Methods

Conclusions

Different NGS Plate-forms have their own strengths and weaknesses, which one to choose from depends on

- Biological question
- Budget
- Analysis team's expertise

Sources for this part

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. doi:10.1186/1471-2164-13-341

Lin Liu, Yinhu Li, Siliang Li, et al., “Comparison of Next-Generation Sequencing Systems,” Journal of Biomedicine and Biotechnology, vol. 2012, Article ID 251364, 11 pages, 2012. doi:10.1155/2012/251364

Defining Genes

Finding genes

After assembly, we have a draft genome in the shape of completely or partially assembled [pseudo]chromosomes or contigs

Next comes the big question, how many genes are there?

Defining Genes

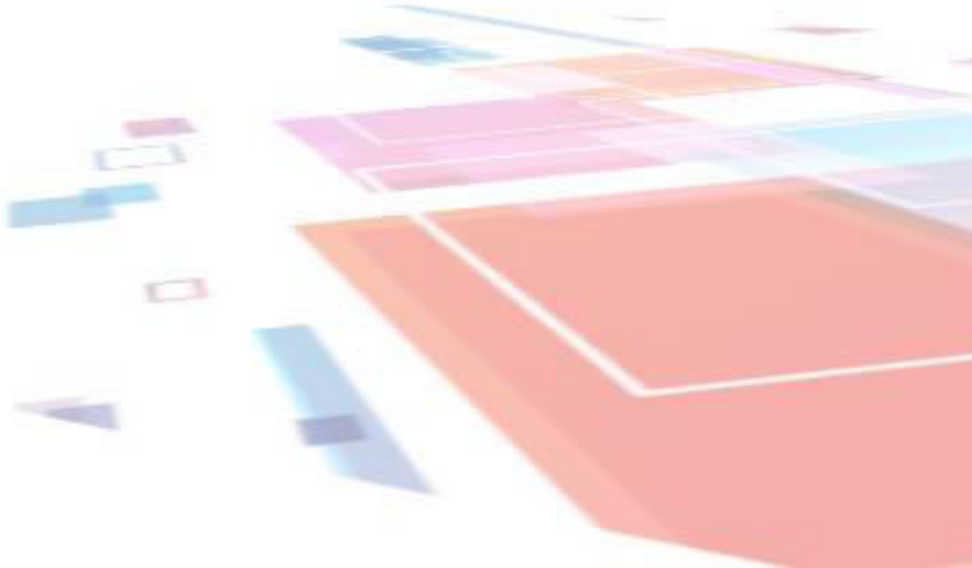
Issues

Defining a gene is problematic also because small genes can be difficult to detect, one gene can code for several protein products

Defining Genes

Issues

some genes code only for RNA, two genes can overlap, and there are many other complications



Defining Genes

GENE is a piece of DNA, a discrete unit of genetic information, which encodes RNA or protein (polypeptide) molecule performing some function in the cell (alone or together with other RNAs and/or proteins)

Defining Genes

Historical viewpoint

A tangled knot;

One gene – one function

One gene – one enzyme

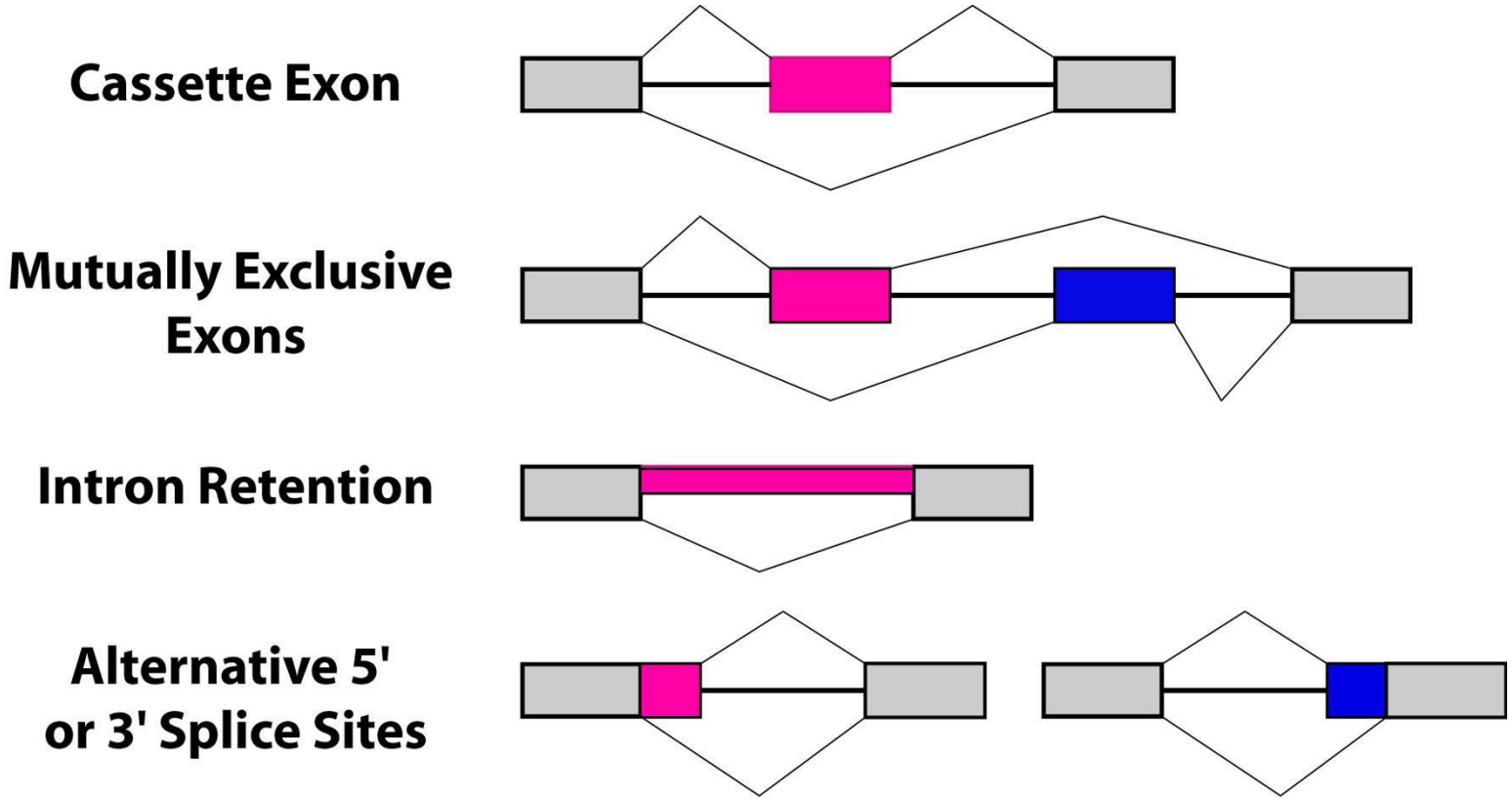
One gene – one polypeptide

One gene – A number of polypeptides

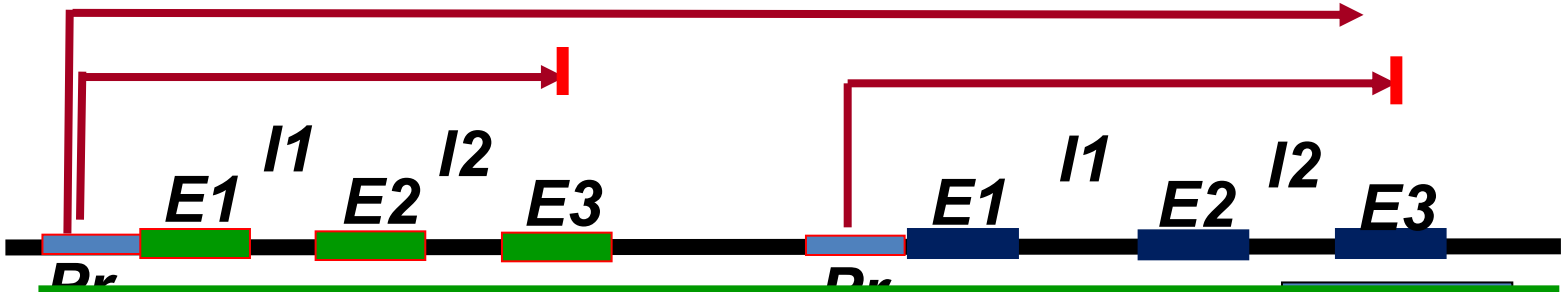
Number of genes – one polypeptide

Defining Genes

Alternative splicing



Defining Genes



At least, 4%-5% of the tandem gene pairs in the human genome are estimated to be transcribed into a single mRNA...

Splicing

Mat-mRNA

Translation

Chimera Protein

Defining Genes

Conclusions

Gene is a DNA sequence that's transcribed to produce one or more functional product(s)

Defining gene is a difficult task due to issues with complicated gene structure and function

Finding Genes

Approaches

Predictions are derived from different computational methods

Two famous approaches;

- ***"Ab initio"* gene finding**
- **Comparative Approach**

Finding Genes

"ab initio" gene finding

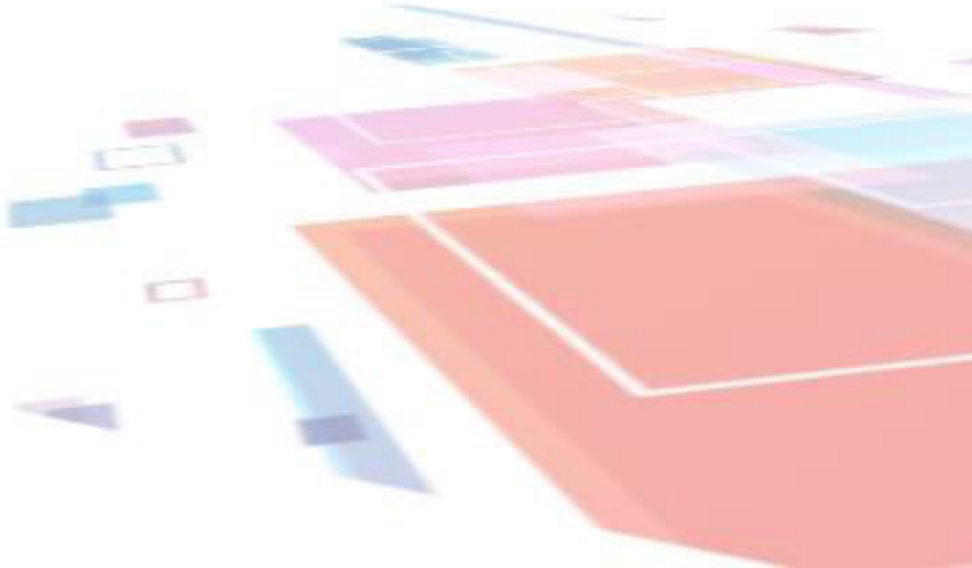
Detect genes by looking for distinct patterns that define where a gene begins and ends

- *ab initio* gene finding tends to overestimate gene numbers by counting any segment that looks like a gene

Finding Genes

Comparative gene finding

Look for genes by comparing segments of sequence with those of known genes



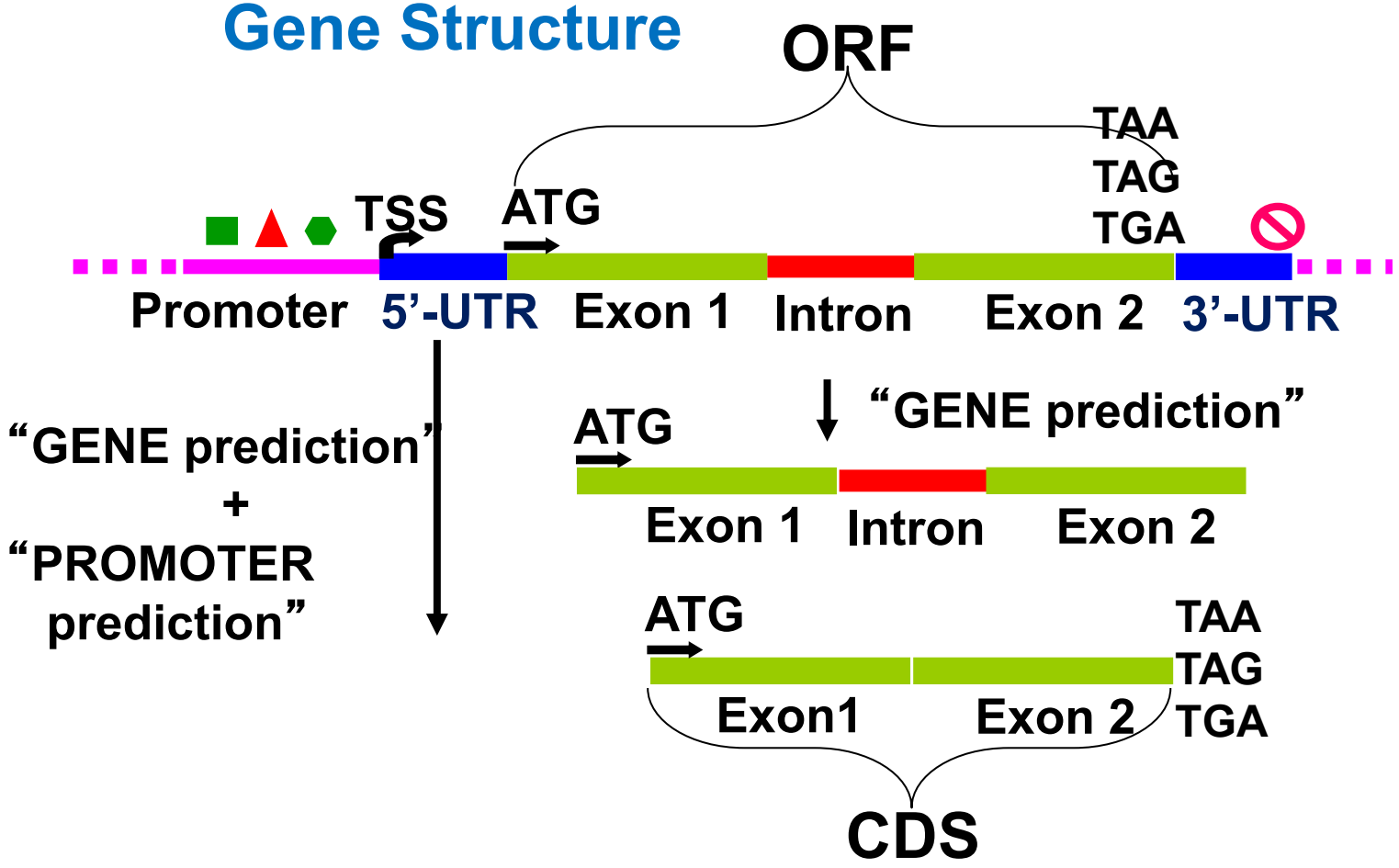
Finding Genes

Comparative gene finding

- Tends to underestimate since limited to recognize only genes similar to what have been seen before

Finding Genes

Gene Structure



Finding Genes

Issues

- Only about 2-3% of human DNA encodes functional genes
- Genes are interspersed among long stretches of non-coding DNA
- Repeats of unknown function occupy 40% and more of a genome
- Great variation of genes' lengths
- We have to distinguish pseudo-genes and true gene duplication(s)

Finding Genes

Issues

- We have to take into attention exon-intron structure of genes in the most of known eukaryotic genomes
  - 99% of yeast genes are intronless
- Non-consensus splice sites
  - other than **GT-----AG**
- Where is the true first 5' exon?
- cDNA data is incomplete and confusing
- Alternative splicing
- **TIC** gene
- Alternative Promoters

Finding Genes

Conclusions

There are two famous approaches for gene predictions

- ***"Ab initio"* gene finding**
- **Comparative Approach**

Finding eukaryotic genes is complicated

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

M Kanehisa

Oxford University Press, 2000

Gene Prediction

Gene prediction Involves;

- Prediction of coding regions
- Prediction of translation starts of gene
- Prediction of splice junctions

(eukaryotic genomes)

Gene Prediction

Coding Sequence or CDS

(**coding DNA sequence**)

- The portion of a gene's DNA or RNA, composed of exons, that codes for protein

Gene Prediction

Coding Sequence or CDS

(**coding DNA sequence**)

- The region is bounded nearer 5' end by a start codon and nearer 3' end by stop codon

Gene Prediction

Coding Sequence or CDS

(**coding DNA sequence**)

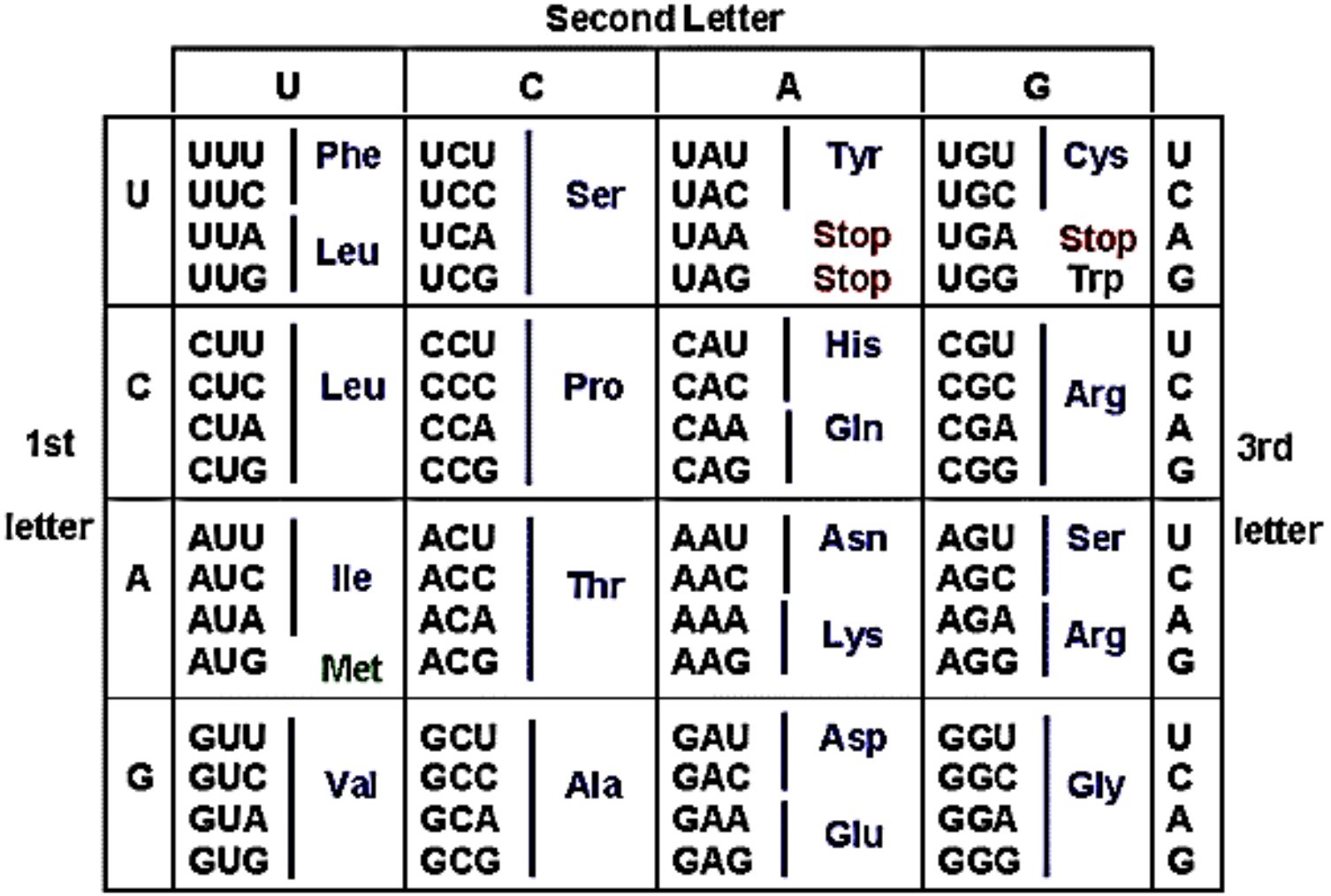
- Any full mRNA sequence (obtained from cDNA sequencing) will have a full coding sequence

Gene Prediction

ORF (Open Reading Frame)

It is the part of gene that has a potential to code for proteins comes in triples called **codons**, beginning and ending with a unique translation **start** (ATG) and one of three **stop** (TAA, TAG, TGA) codons

Gene Prediction



Gene Prediction

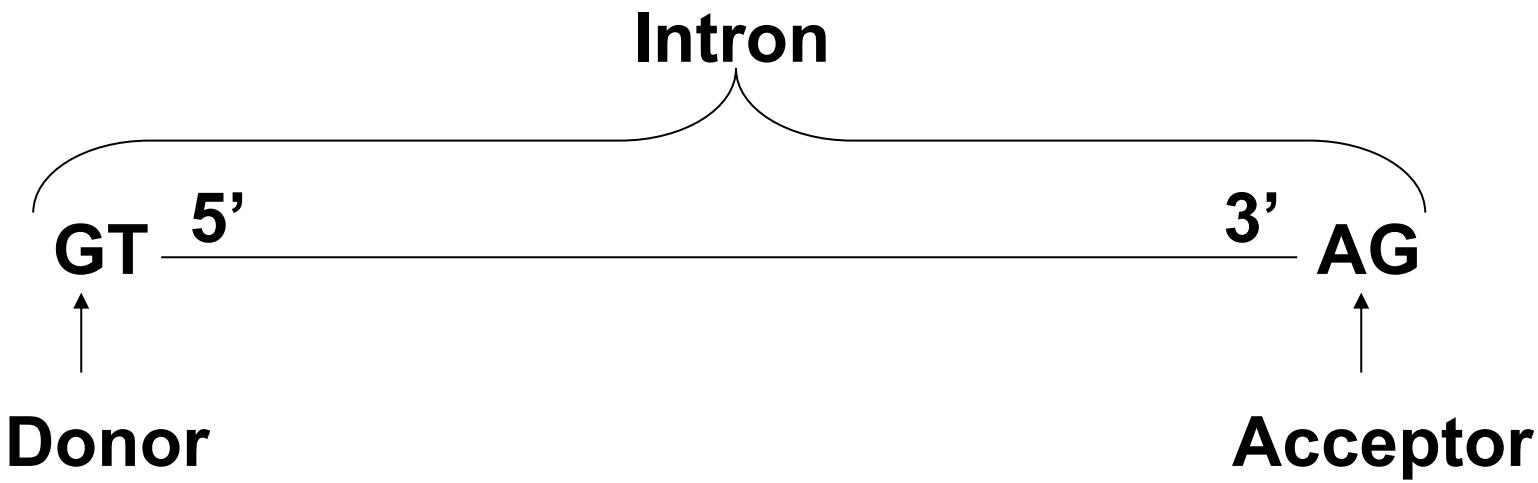
ORF vs CDS

- CDS is transcribed and is coding for something
- ORF is usually predicted based on DNA sequence and not proven to be transcribed

Gene Prediction

Splice Junction

- Donor site prediction
- Acceptor site prediction



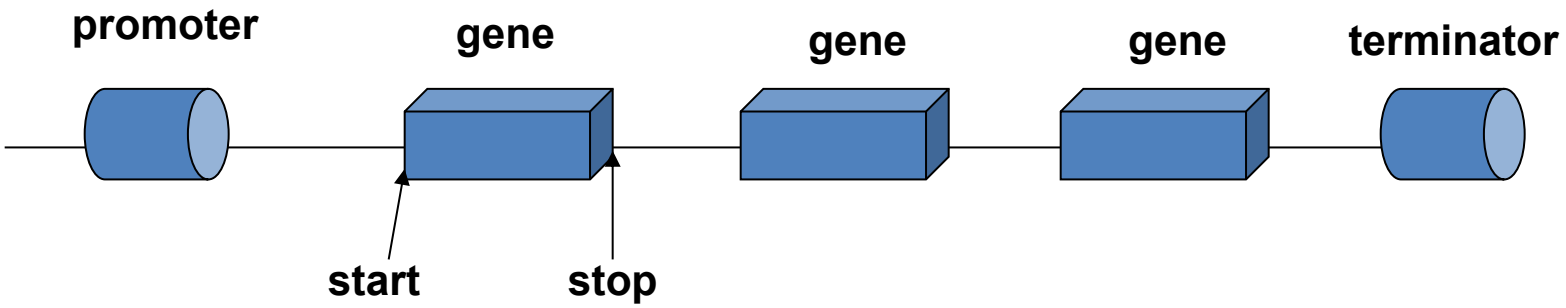
Gene Prediction

Splice Junction **Intron-Splicing Site-Exon**

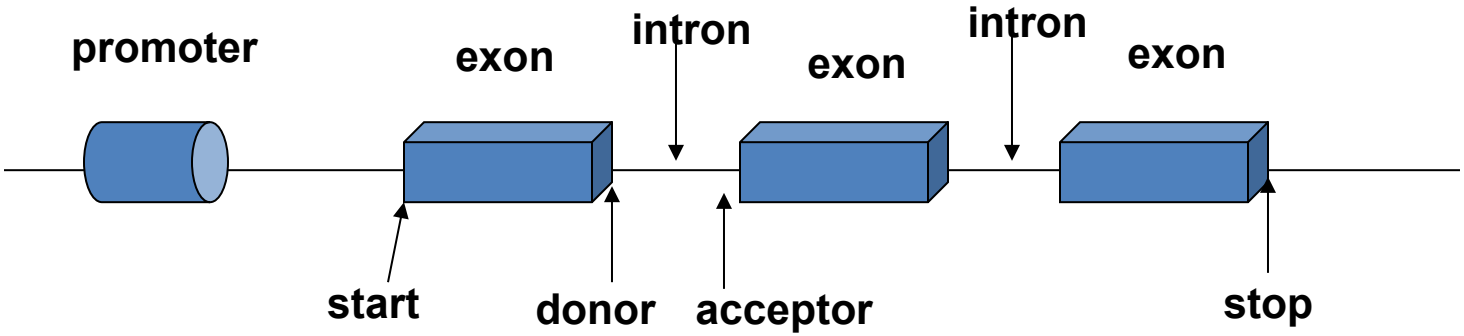
GTCTTTTTGAGATAGAAAACAAAAAAGAACTTAT
AGCTCCAAATATCTTTCTTTGGATAACAACTCAA
ATAATATTTGCTTTCTTGTGATTGTAGATGGTTTG
GGAACCGAAGTCAATTTGGACGGGATTGCTTAC
TACAATAATGTTATTAATGCTCTTCTAGAGAAAG
GTATTGAAGCATTCCTAATAAATTTGCTCCTAGTA
TAATTGTGGTTTATGCATAAGAACTTCTCCTGCC
AGGTATTCAGCCTTATGTAACTTTGTACCACTGG
GACCTTCCTTTGCATCTCCAGGAATCAATGAATG
GATGGTTAAATAGGAAGATTGTGTAAGTATTAAA
ATGCCAGGTGTTTAAATATGAAGTTACTGCATAT

Gene Prediction

Prokaryotic genes (Polycistron)



Eukaryotic genes (1 gene)



Gene Prediction

Information fusion

Involves;

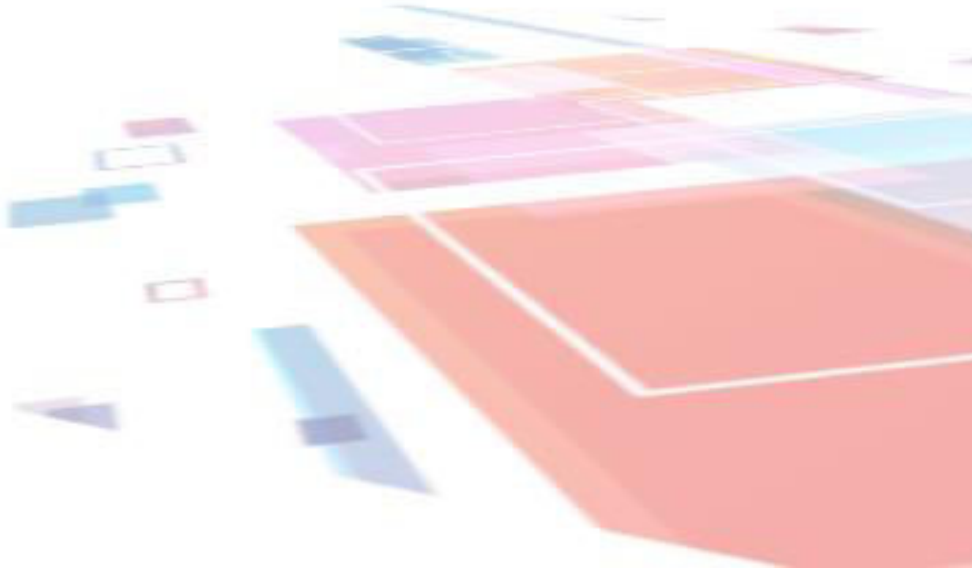
combining multiple pieces of information for the whole gene prediction:

- TSS(s), UTRs Alternative splicing variants, Gene product destination and function(s)

Gene Prediction

Signal vs Content

A small pattern within the genomic DNA is referred to as a **signal**, whereas a region of genomic DNA is a **content**



Gene Prediction

Signal

Splice sites, **starts** and **ends** of transcription or translation, **branch points**, **(TBS)** transcription factor binding sites, etc

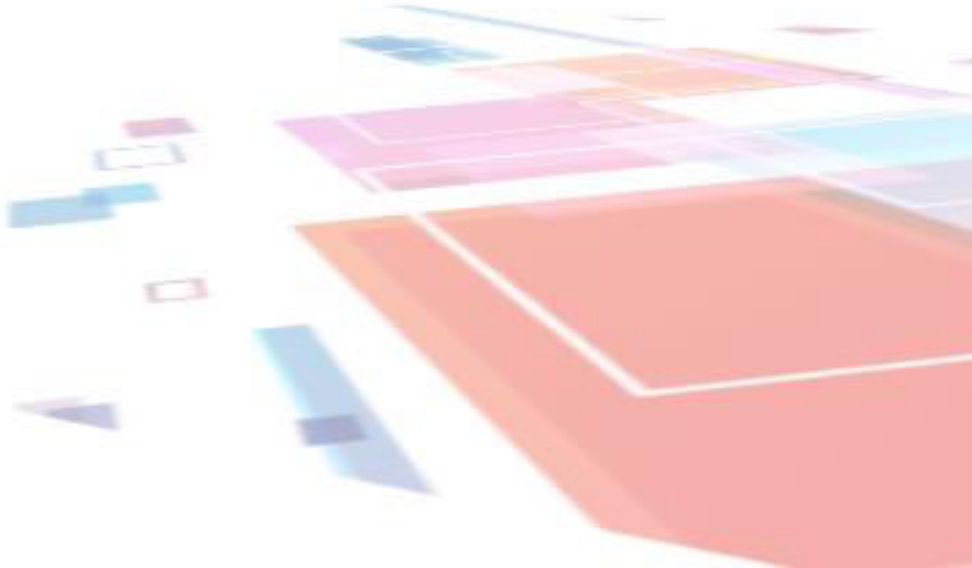
Content

exons, introns, UTRs, promoter regions

Gene Prediction

Conclusions

- Gene prediction involves prediction of coding regions, translation start site and splice junction



Gene Prediction

Conclusions

- CDS is the actual region of DNA, translated to form proteins while the ORF may contain introns

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

M Kanehisa

Oxford University Press, 2000

Genetics and Genomics

Genes Predictions - Prokaryotes

Genes Predictions in Prokaryotes

Genes Predictions

- Locating genes within genomic sequence.
- Defining initiation termination sites of genes.
- Extracting the coding region of each gene.
- Identifying function for coding region.

Genes Predictions in Prokaryotes

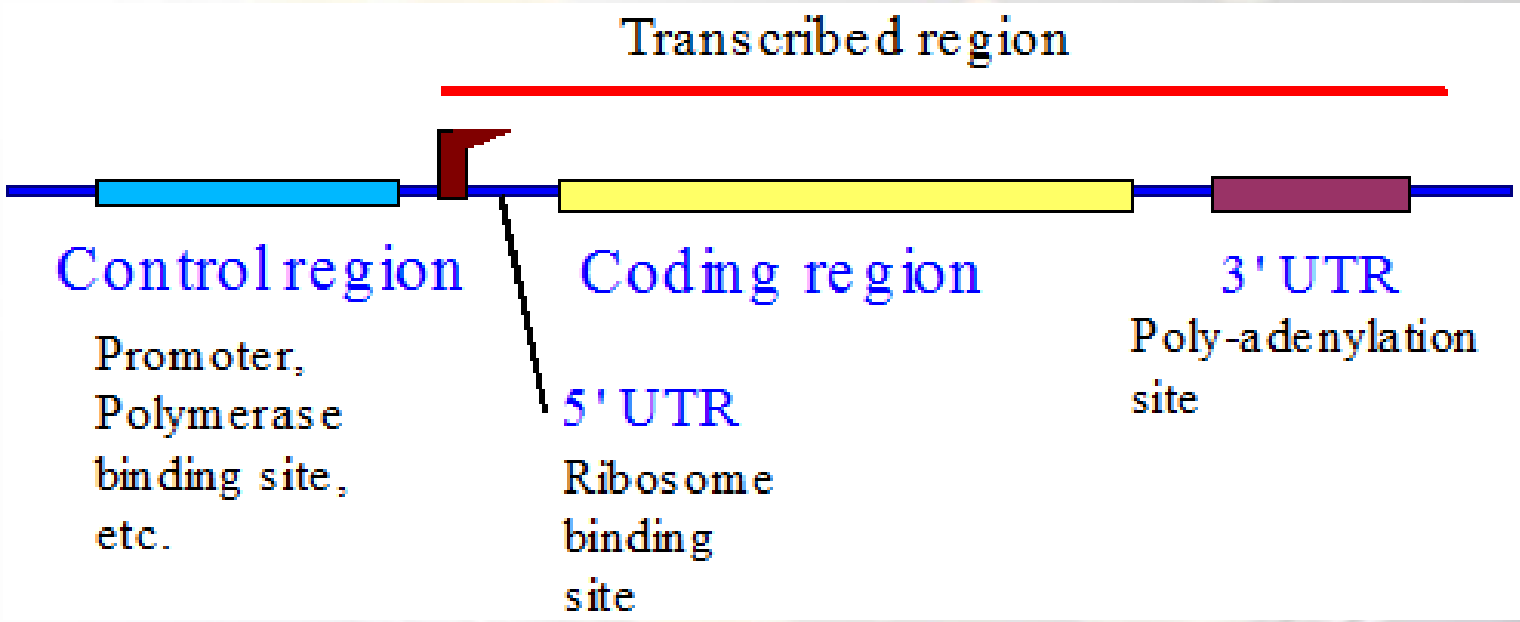
Genes Predictions - Gene

- A region of the genome that codes for a functional component such as an RNA or protein.

Genes Predictions in Prokaryotes

Genes Predictions - Gene

Identify coding regions computationally from raw genomic sequence data



Genes Predictions in Prokaryotes

Genes Predictions - Gene

- Translation utilizes a trinucleotide coding system: codons
- Translation begins at a start codon
- Translation ends at a stop codon

Genes Predictions in Prokaryotes

Genes Predictions - Gene

- Most organisms use ATG as a start codon
- A few bacteria also GTG and TTG
- Regardless of codon used, the first amino acid in every translated peptide chain is Methionine

Genes Predictions in Prokaryotes

Genes Predictions - Gene

- Almost all organisms use TAG, TGA and TAA as stop codons
- The major exception are the mycoplasmas

Genes Predictions in Prokaryotes

Genes Predictions - Gene

- In Bacteria and Archaea, the coding region is in one continuous sequence known as an open reading frame (ORF).

DNA:

ATG-GAA-GAG-CAC-CAA-GTC-CGA-TAG

Protein:

MET-GLU- GLU -HIS -GLN-VAL-ARG-Stop

Genes Predictions in Prokaryotes

Genes Predictions - Gene

- Translation region is contiguous in prokaryotes, gene finding focuses on identifying ORFs
- ORF-finder takes a syntactic approach to identifying putative coding regions

Genes Predictions in Prokaryotes

Genes Predictions – ORF Finder and Gene Glimmer

- ORF-finder is available from NCBI
- GLIMMER 2.0 is a more sophisticated program that attempts to model codon usage, average gene length and other features before identifying putative coding regions

Genes Predictions in Prokaryotes

Genes Predictions – ORF Finder Approach

- Identifies every stop codon in the genomic sequence.
- Scan upstream to the farthest, in-frame start codon.
- Locates ORFs that begin with ATG as well as GTG and TTG

Genes Predictions in Prokaryotes

Genes Predictions – ORF Finder

Advantages

- Can identify every possible ORF.
- Minimum length constraint ensures that many false positives are discarded prior to human review.

Disadvantages

- Does not eliminate overlapping ORFs.
- Even with a length constraint, there are often many false positives.
- Cannot take into account organism-specific idiosyncrasies

Genes Predictions in Prokaryotes

Genes Predictions – Glimmer Approach

- The model includes information about:
- Average length of coding region
- Codon usage bias (which codons are preferentially used)
- Evaluates the frequency of occurrence of higher order combinations of nucleotides

Genes Predictions in Prokaryotes

Genes Predictions – Glimmer Output

- For each ORF, GLIMMER assigns a likelihood score or probability that the ORF resembles a known gene.
- High scoring ORFs that overlap significantly with other high scoring ORFs are reported but highlighted.
- GLIMMER 2.0 is reported to be 98% accurate on prokaryotic genomes.

Genes Predictions in Prokaryotes

Genes Predictions – Glimmer

Advantages

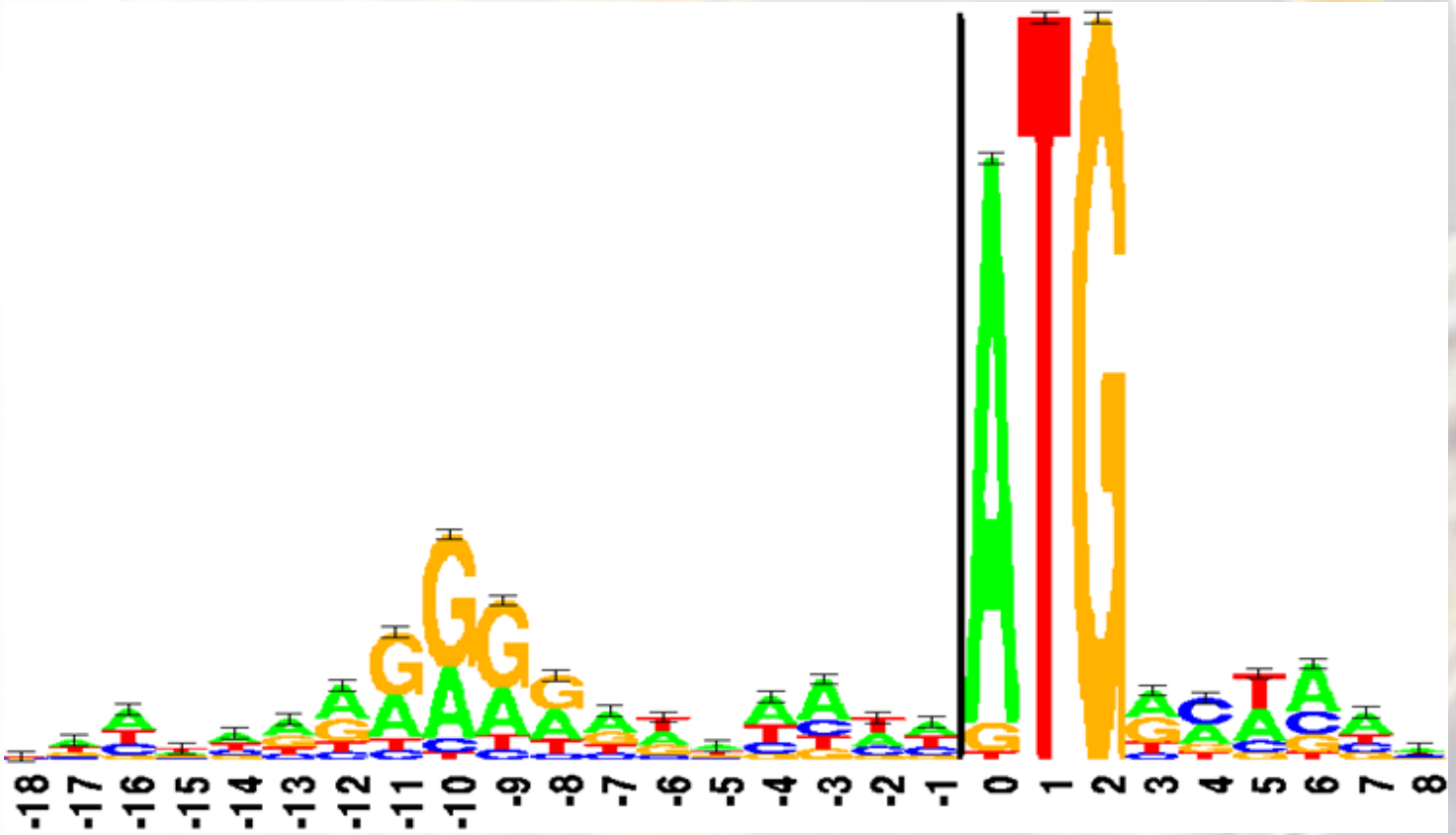
- Fewer false positives because ORFs are evaluated for likelihood of coding.
- Organism-specific because model is built on known genes.
- User can modify many parameters during search phase.

Disadvantages

- Requires approximately 500+ known genes for proper training.
- Genuine coding regions with unusual codon composition will be eliminated.
- Reported accuracy difficult to reproduce.

Genes Predictions in Prokaryotes

Genes Predictions – RBS



Genes Predictions in Prokaryotes

Genes Predictions - Gene

- ORF-finder
- Glimmer

Genetics and Genomics

Genes Predictions - Eukaryotes

Genes Predictions in Eukaryotes

Genes Predictions - Eukaryotes

- Tools for finding genes in eukaryotes
- Genie
- Fgenes
- Genscan

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genie

- The model includes information about:
- Average length of exons and introns.
- Compositional information about exons and introns.
- A neural-net derived model of splice junctions and consensus sequences around splice junctions.
- Splice junction information can be further improved by including results of homology searches.

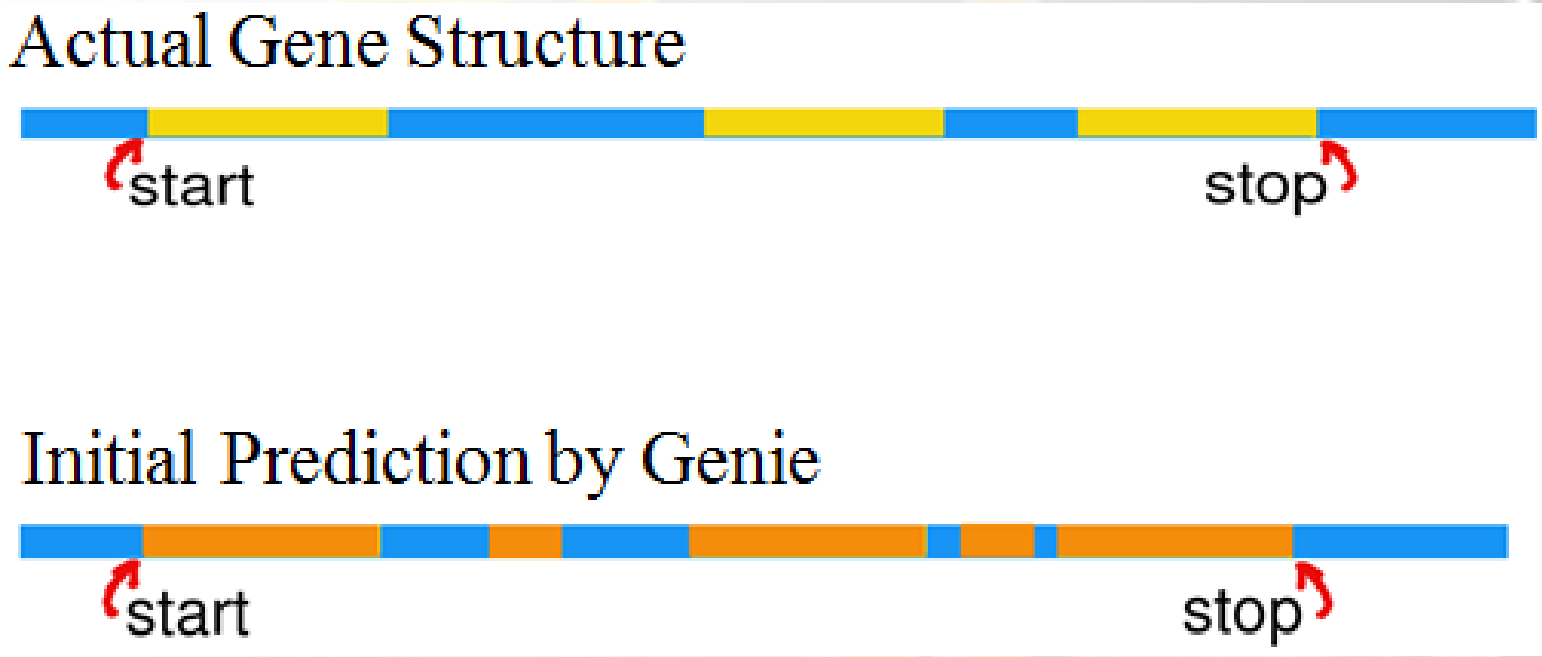
Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genie

- Genie is approximately 60-75% accurate on eukaryotic genomes.

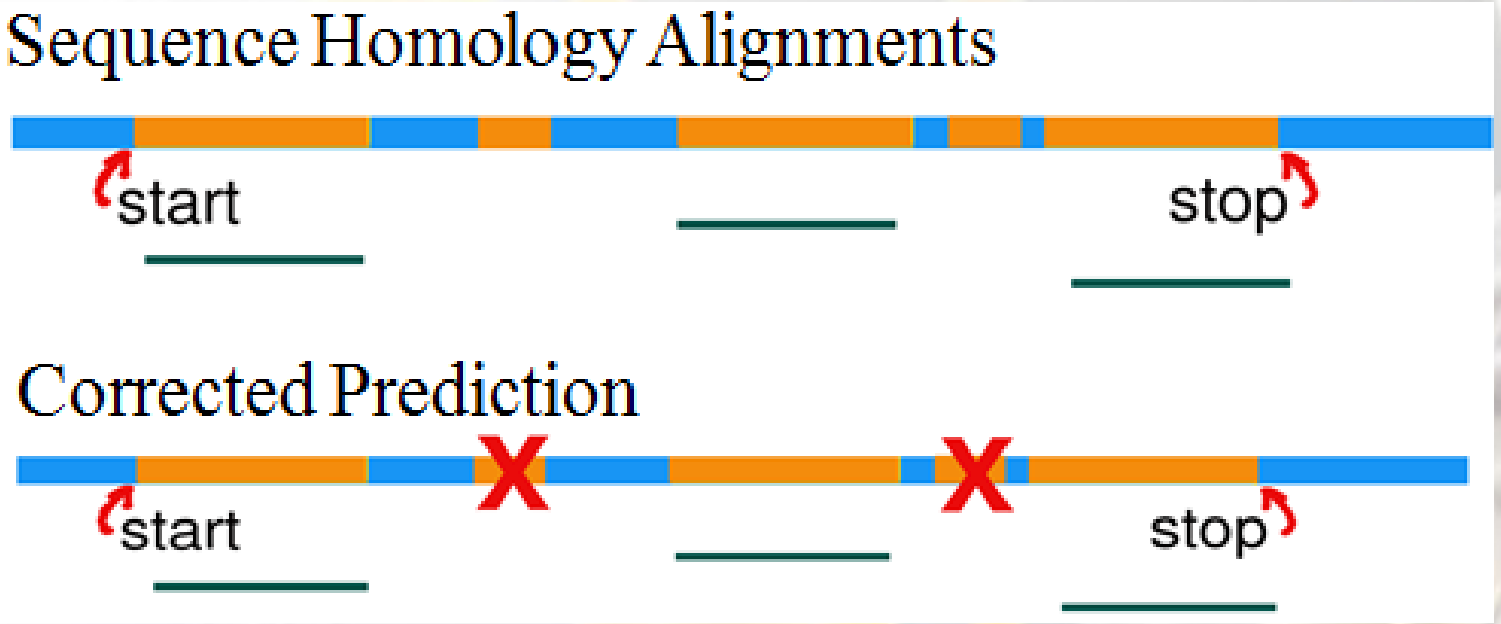
Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genie



Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genie



Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genie

Advantages

- Extra predicted exons can be eliminated based on evidence from homology searches.
- Likelihood scores provided for each predicted exon.

Disadvantages

- No organism-specific training is possible.
- Works best on mammalian genomes, not other eukaryotes.
- Reliance on homology evidence can result in oversight of novel genes unique to the organism of interest.

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Fgenes

- Identifies putative exons and introns.
- Scores each exon and intron based on composition.
- Uses dynamic programming to find the highest scoring path through these exons and introns.
- The best-scoring path is constrained by several factors including that exons must be in frame with each other and ordered sequentially.

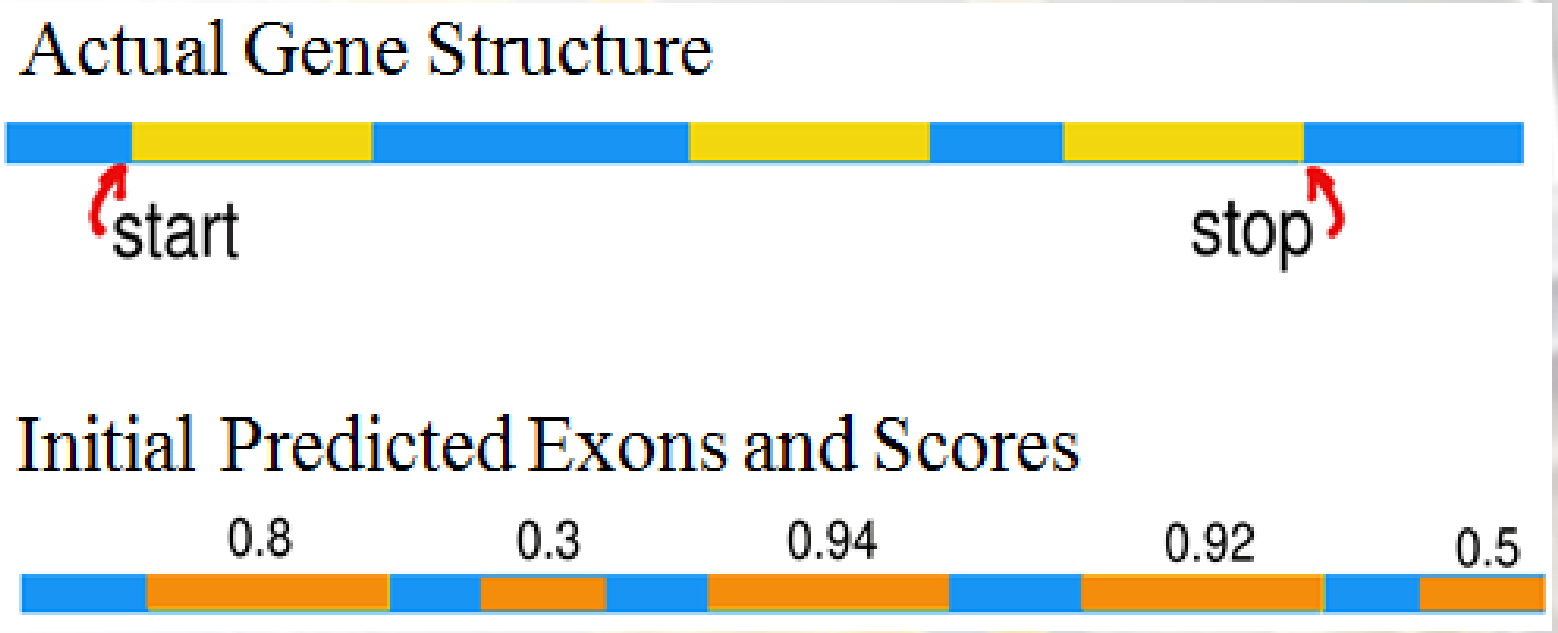
Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Fgenes

- Fgenes is about 70% accurate in most mammalian genomes.

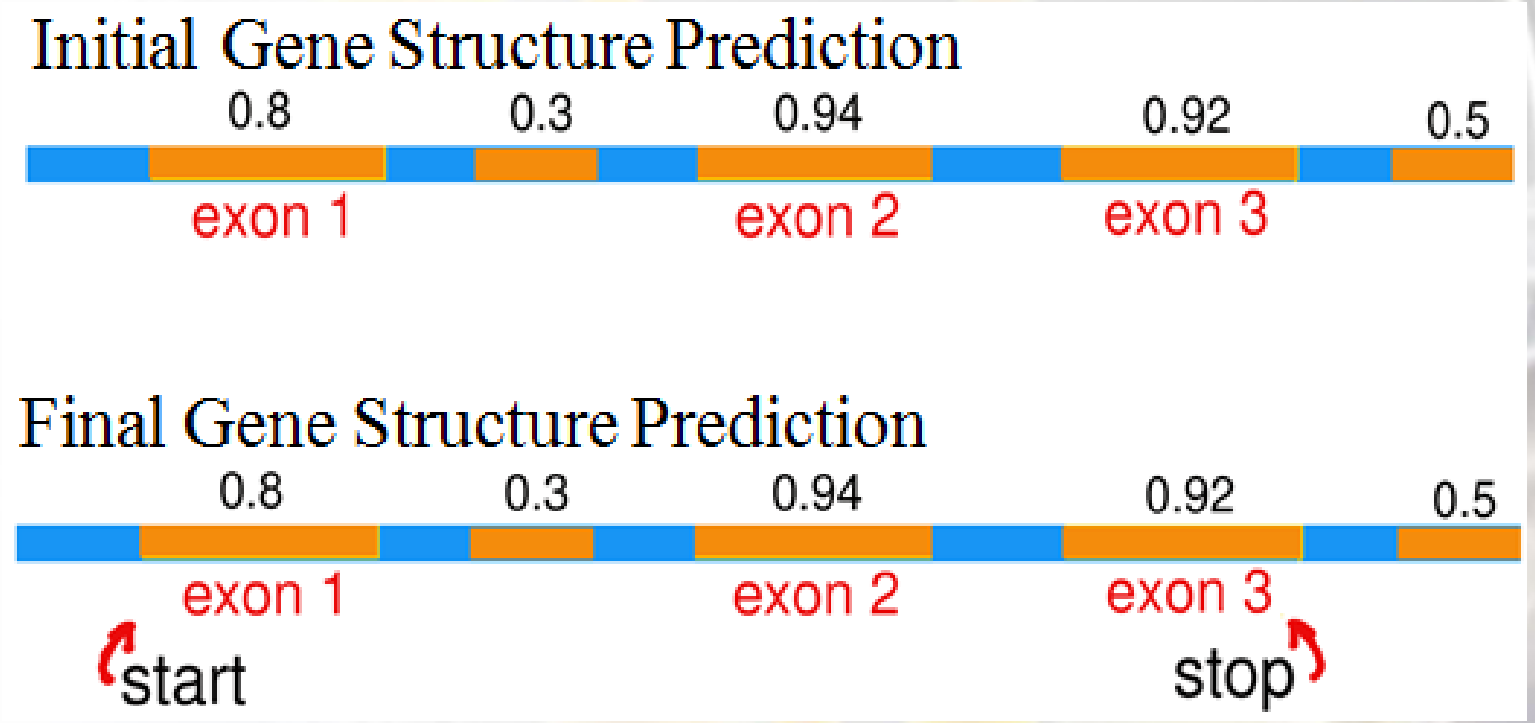
Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Fgenes



Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Fgenes



Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Fgenes

Advantages

- Alternative gene structures are reported.
- Also attempts to identify putative promoter and poly-A sites.

Disadvantages

- User cannot train models.
- Only human model-based version is available for unrestricted public use.

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genscan

- Models for different states (GHMMs)
- State 1 and 2: Exons and Introns
  - Length
  - Composition
- State 3: Splice junctions

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes - Genscan

- Weight matrix based array to identify consensus sequences
- Weight matrix to identify promoters, poly-A signals and other features.

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes – Genscan Output

- Gene structure
- Promoter site
- Translation initiation exon
- Internal exons
- Terminal exon (translation termination)

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes – Genscan Output

- Poly-adenylation site
- Genscan is 80% accurate on human sequences.

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes – Genscan

Advantages

- Most accurate of available tools.
- Excellent at identifying internal and terminal exons
- Provides some assistance in identifying putative promoters

Disadvantages

- User cannot train models nor weak parameters.
- Identification of initial exons is weaker than other kinds of exons.
- Promoter identification can be mis-leading.

Genes Predictions in Eukaryotes

Genes Predictions Eukaryotes – Genscan

- Eukaryotic gene structures can be quite complex.
- The best approaches to gene finding in eukaryotes combine different methods.

GO Functional Analysis

Introduction

After we get a list of genes, it is important to identify interesting Biological patterns

These functions are represented and classified as GO (Gene Ontology) terms

GO Functional Analysis

GO Terms

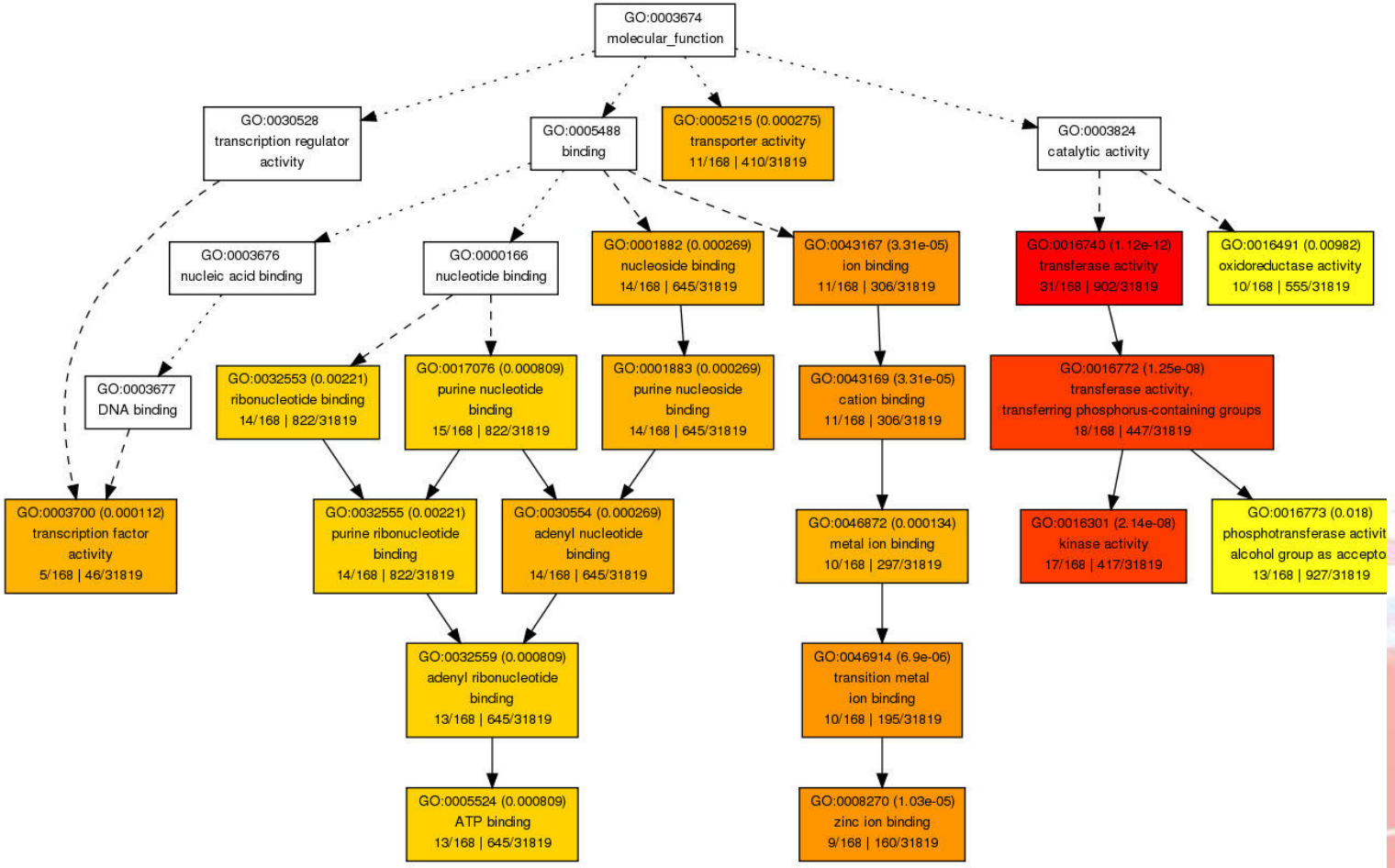
GO (Gene Ontology) terms are implemented as Directed Acyclic Graphs (DAG)

Three main types

Biological Process (BP)
Cellular Components (CC)

Molecule Function (MF)

GO Functional Analysis



GO Functional Analysis

DAVID Tool

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) (Huang et al., 2009).

http://david.abcc.ncifcrf.gov/

GO Functional Analysis

DAVID Tool

David provides a comprehensive set of functional annotation tools to understand biological meaning behind large list of genes.

http://david.abcc.ncifcrf.gov/

DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

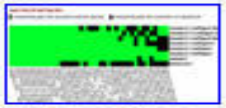
Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2014

The Database for Annotation, Visualization and Integrated Discovery (**DAVID**) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more



Screen Shot 1

Screen Shot 2

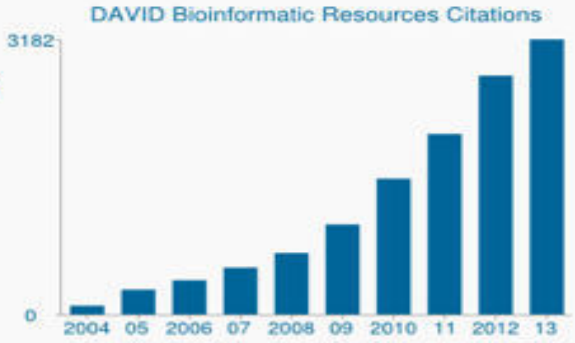
Screen Shot 3

Search

What's Important in DAVID?

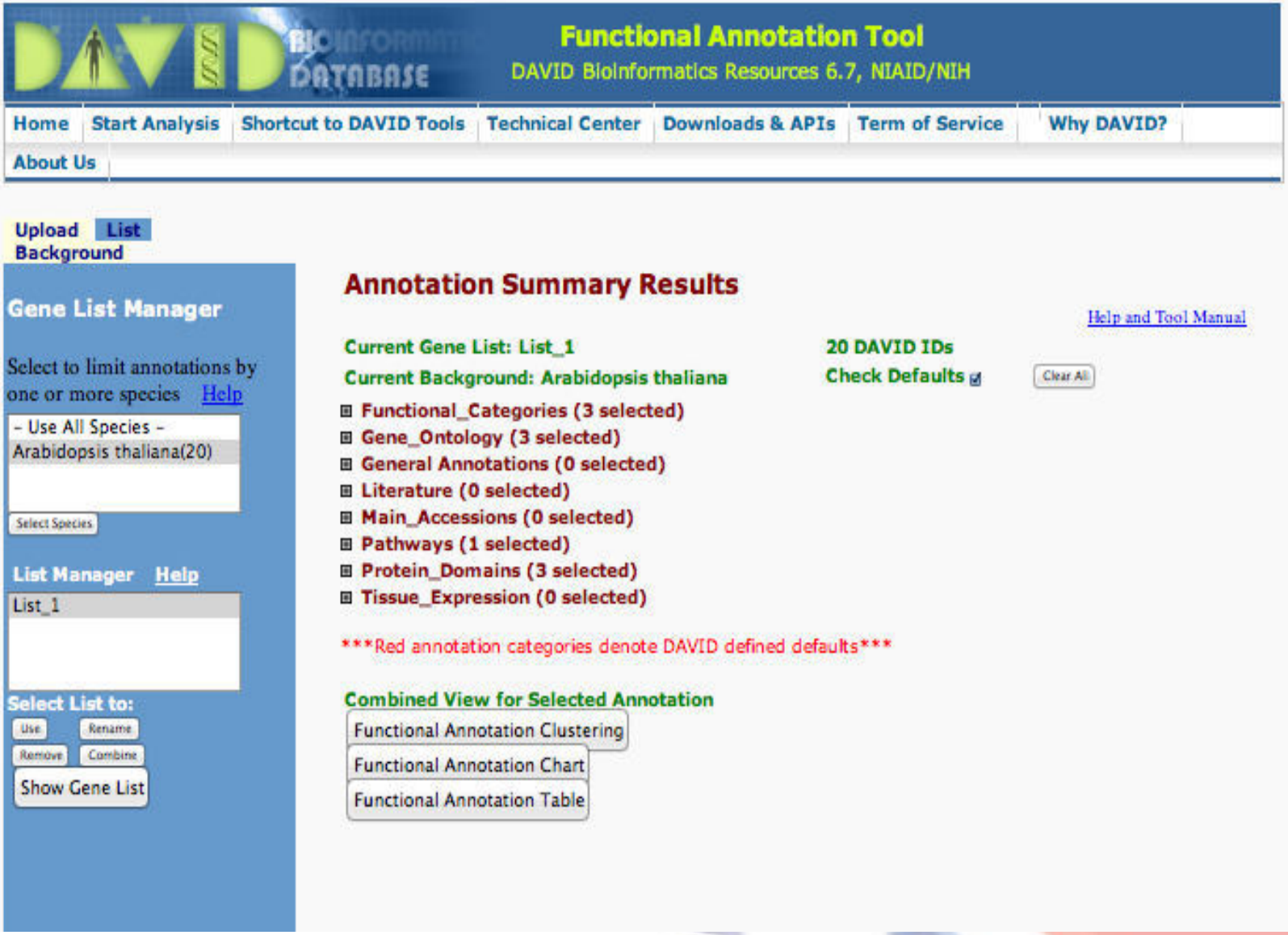
- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID



- > 10,000 Citations
- Daily Usage: ~1200 gene lists/sublists from ~400 unique researchers.
- Total Usage: ~800,000 gene lists/sublists from >5,000 research institutes world-wide

Functional Annotation Tool
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Upload Gene List
Demolist 1
Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
AT1G51370
AT1G50920
AT1G36960
AT1G44020
Clear
Or
B:Choose From a File
Choose File
no file selected
Multi-List File
Step 2: Select Identifier
TAIR_ID
Step 3: List Type
Gene List
Background
Step 4: Submit List
Submit List
Functional Annotation Tool
Submit your gene list to start the tool!
Tell us how you like the tool
Read technical notes of the tool
Contact us for questions
Key Concepts:
The DAVID Gene Concept
DAVID 6.7 is designed around the "DAVID Gene Concept", a graph theory evidence-based method to agglomerate species-specific gene/protein identifiers from a variety of public genomic resources including NCBI, PIR and Uniprot/SwissProt. The DAVID Gene Concept method groups tens of million of identifiers from over 65,000 species into 1.5 million unique protein/gene records. More
Term/Gene Co-Occurrence Probability
Ranking functional categories based on co-occurrence with sets of genes in a gene list can rapidly aid in unraveling new biological processes associated with cellular functions and pathways. DAVID 6.7 allows investigators to sort gene categories from dozens of annotation systems. Sorting can be based either the number of genes within each category or by the EASE-score. More
Gene Similarity Search
Any given gene is associating with a set of annotation terms. If genes share similar set of those terms, they are most likely involved in similar biological mechanisms. The algorithm tries to group those related genes based on the agreement of sharing similar annotation terms by Kappa statistics. More
Term Similarity Search
Typically, a biological process/term is done by a corporation of a set of genes. If two or more biological processes are done by similar set of genes, the processes might be related in the biological network somehow. This search function is to identify the related biological processes/terms by quantitatively measuring the degree of the agreement how terms share the similar participating genes. More
Integrated Solutions
Functional Annotation
Numerous Data Sources
Co-occurrence Probability
Use Homolog Annotation
Dynamic Pathway Maps
Disease Associations
Numerous public sources of protein and gene annotation have been parsed and integrat
into DAVID 6.7. DAVID 6.7 contains information on over 1.5 million genes from more tha
65,000 species. A list of protein or gene identifiers can be uploaded all at once to extract
and summarize functional annotation associated with group of genes or with each
individual gene. Data can be displayed in chart or table format or downloaded to the use
s hard drive.


Functional Annotation Tool
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Gene List Manager
Select to limit annotations by one or more species
Help
- Use All Species -
Arabidopsis thaliana(20)
Select Species
List Manager
Help
List_1
Select List to:
Use
Rename
Remove
Combine
Show Gene List
Annotation Summary Results
Help and Tool Manual
Current Gene List: List_1
20 DAVID IDs
Current Background: Arabidopsis thaliana
Check Defaults
Clear All
Functional_Categories (3 selected)
Gene_Ontology (3 selected)
General Annotations (0 selected)
Literature (0 selected)
Main_Accessions (0 selected)
Pathways (1 selected)
Protein_Domains (3 selected)
Tissue_Expression (0 selected)
Red annotation categories denote DAVID defined defaults
Combined View for Selected Annotation
Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

BP-FAT 954 genes, 98 terms

198 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_FAT	response to organic substance	RT		142	15.0	4.4E-29	1.5E-26
	GOTERM_BP_FAT	response to abiotic stimulus	RT		121	12.8	4.3E-18	7.3E-16
	GOTERM_BP_FAT	regulation of transcription	RT		119	12.6	1.1E-3	1.8E-2
	GOTERM_BP_FAT	defense response	RT		114	12.1	5.3E-20	1.1E-17
	GOTERM_BP_FAT	response to endogenous stimulus	RT		97	10.3	7.7E-14	9.7E-12
	GOTERM_BP_FAT	transcription	RT		88	9.3	9.0E-5	2.1E-3
	GOTERM_BP_FAT	response to hormone stimulus	RT		80	8.5	6.3E-9	3.6E-7
	GOTERM_BP_FAT	oxidation reduction	RT		80	8.5	1.8E-4	3.6E-3
	GOTERM_BP_FAT	phosphate metabolic process	RT		72	7.6	2.8E-2	2.3E-1
	GOTERM_BP_FAT	phosphorus metabolic process	RT		72	7.6	2.8E-2	2.3E-1
	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT		71	7.5	1.1E-3	1.8E-2
	GOTERM_BP_FAT	regulation of RNA metabolic process	RT		71	7.5	1.3E-3	2.1E-2
	GOTERM_BP_FAT	phosphorylation	RT		65	6.9	5.5E-2	3.5E-1
	GOTERM_BP_FAT	protein amino acid phosphorylation	RT		63	6.7	1.7E-2	1.7E-1
	GOTERM_BP_FAT	response to carbohydrate stimulus	RT		62	6.6	8.0E-35	4.0E-32
	GOTERM_BP_FAT	intracellular signaling cascade	RT		60	6.3	6.4E-5	1.6E-3
	GOTERM_BP_FAT	response to chitin	RT		59	6.2	1.4E-44	1.4E-41
	GOTERM_BP_FAT	secondary metabolic process	RT		51	5.4	2.5E-10	2.3E-8
	GOTERM_BP_FAT	immune response	RT		50	5.3	6.7E-16	9.6E-14
	GOTERM_BP_FAT	response to osmotic stress	RT		47	5.0	3.1E-9	1.9E-7
	GOTERM_BP_FAT	innate immune response	RT		45	4.8	1.1E-13	1.3E-11
	GOTERM_BP_FAT	response to temperature stimulus	RT		45	4.8	6.5E-10	5.1E-8
	GOTERM_BP_FAT	response to salt stress	RT		45	4.8	2.4E-9	1.7E-7
	GOTERM_BP_FAT	response to wounding	RT		41	4.3	4.8E-21	1.2E-18
	GOTERM_BP_FAT	cellular response to hormone stimulus	RT		40	4.2	2.2E-5	6.0E-4
	GOTERM_BP_FAT	hormone-mediated signaling	RT		40	4.2	2.2E-5	6.0E-4

CC-FAT 954 genes, 15 terms

15 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_FAT	intrinsic to membrane	RT		145	15.3	7.5E-8	1.1E-5
	GOTERM_CC_FAT	plasma membrane	RT		117	12.4	1.6E-5	1.1E-3
	GOTERM_CC_FAT	integral to membrane	RT		97	10.3	2.2E-2	3.0E-1
	GOTERM_CC_FAT	extracellular region	RT		60	6.3	1.3E-2	2.4E-1
	GOTERM_CC_FAT	cell wall	RT		42	4.4	9.5E-5	4.5E-3
	GOTERM_CC_FAT	external encapsulating structure	RT		42	4.4	1.3E-4	4.6E-3
	GOTERM_CC_FAT	vacuole	RT		32	3.4	6.1E-2	4.7E-1
	GOTERM_CC_FAT	anchored to membrane	RT		26	2.8	2.5E-4	7.0E-3
	GOTERM_CC_FAT	apoplast	RT		23	2.4	2.1E-2	3.1E-1
	GOTERM_CC_FAT	plant-type cell wall	RT		16	1.7	6.5E-2	4.7E-1
	GOTERM_CC_FAT	intrinsic to plasma membrane	RT		9	1.0	3.3E-2	3.0E-1
	GOTERM_CC_FAT	anchored to plasma membrane	RT		7	0.7	3.1E-2	3.1E-1
	GOTERM_CC_FAT	exocyst	RT		6	0.6	3.5E-3	7.8E-2
	GOTERM_CC_FAT	cell cortex part	RT		6	0.6	2.2E-2	2.7E-1
	GOTERM_CC_FAT	cell cortex	RT		6	0.6	2.6E-2	2.9E-1

MF-FAT 954 genes, 62 terms

62 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_FAT	ion binding	RT		188	19.9	4.3E-6	5.7E-4
	GOTERM_MF_FAT	cation binding	RT		187	19.8	5.6E-6	5.9E-4
	GOTERM_MF_FAT	metal ion binding	RT		176	18.6	2.4E-5	1.8E-3
	GOTERM_MF_FAT	transition metal ion binding	RT		136	14.4	2.6E-3	8.2E-2
	GOTERM_MF_FAT	purine nucleotide binding	RT		127	13.4	4.2E-2	4.0E-1
	GOTERM_MF_FAT	purine nucleoside binding	RT		122	12.9	9.4E-3	1.9E-1
	GOTERM_MF_FAT	adenyl nucleotide binding	RT		122	12.9	9.4E-3	1.9E-1
	GOTERM_MF_FAT	nucleoside binding	RT		122	12.9	1.1E-2	1.9E-1
	GOTERM_MF_FAT	DNA binding	RT		116	12.3	1.7E-2	2.7E-1
	GOTERM_MF_FAT	transcription regulator activity	RT		113	12.0	6.6E-6	5.8E-4
	GOTERM_MF_FAT	transcription factor activity	RT		107	11.3	3.6E-7	1.9E-4
	GOTERM_MF_FAT	adenyl ribonucleotide binding	RT		107	11.3	7.5E-2	5.5E-1
	GOTERM_MF_FAT	ATP binding	RT		105	11.1	8.9E-2	6.0E-1
	GOTERM_MF_FAT	protein kinase activity	RT		63	6.7	3.0E-3	8.9E-2
	GOTERM_MF_FAT	iron ion binding	RT		56	5.9	2.0E-6	3.6E-4
	GOTERM_MF_FAT	protein serine/threonine kinase activity	RT		54	5.7	9.8E-3	1.9E-1
	GOTERM_MF_FAT	electron carrier activity	RT		48	5.1	4.9E-5	2.9E-3
	GOTERM_MF_FAT	calcium ion binding	RT		39	4.1	6.0E-7	1.6E-4
	GOTERM_MF_FAT	tetrapyrrole binding	RT		34	3.6	5.3E-5	2.8E-3

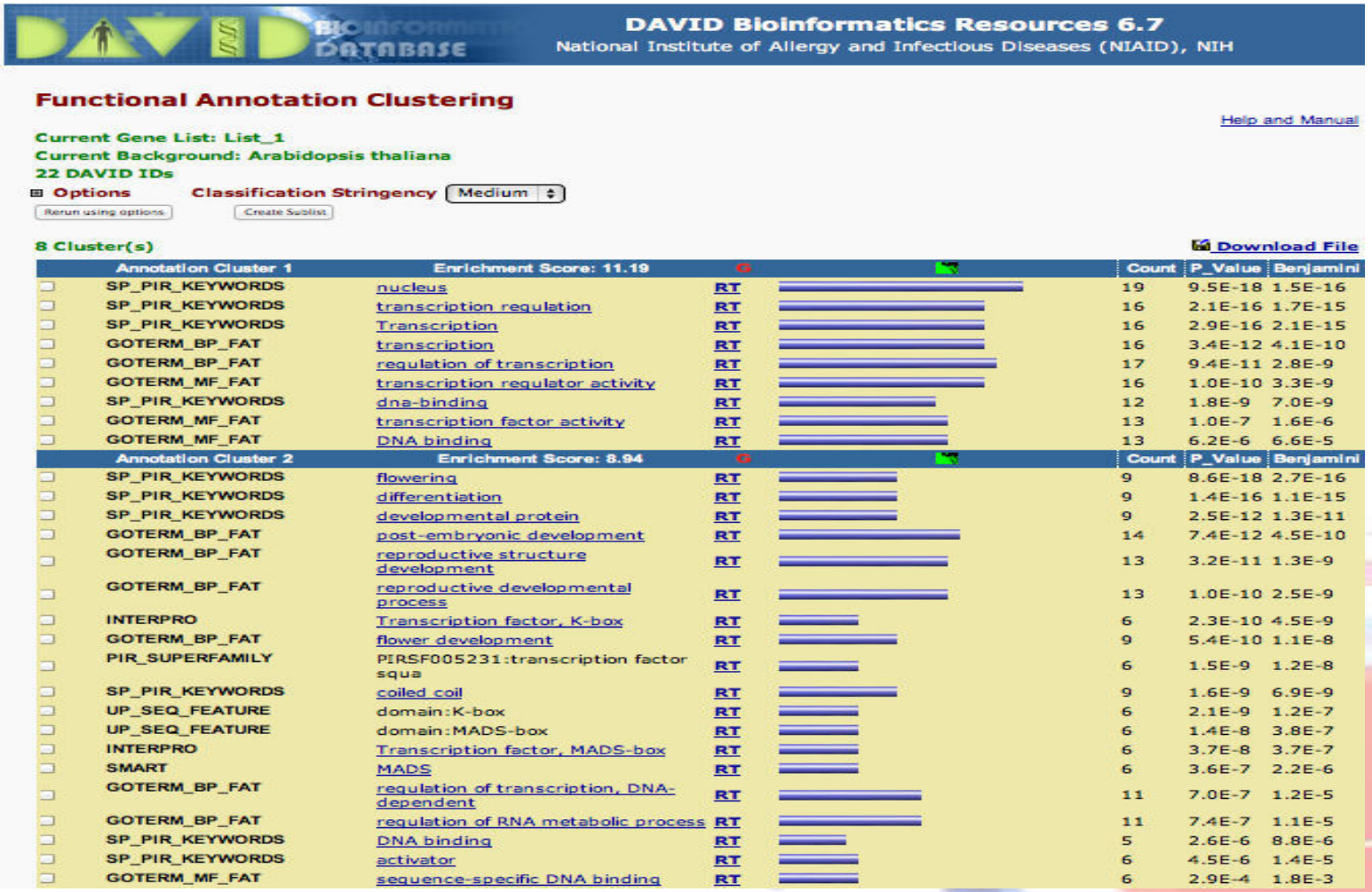
Associated Pathways

12 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Biosynthesis of plant hormones	RT		17	1.8	6.9E-2	3.3E-1
	KEGG_PATHWAY	Phenylalanine metabolism	RT		11	1.2	4.3E-4	8.9E-3
	KEGG_PATHWAY	Phenylpropanoid biosynthesis	RT		11	1.2	3.4E-3	3.4E-2
	KEGG_PATHWAY	Methane metabolism	RT		10	1.1	2.0E-3	3.1E-2
	KEGG_PATHWAY	alpha-Linolenic acid metabolism	RT		9	1.0	3.4E-6	1.1E-4
	KEGG_PATHWAY	Glucosinolate biosynthesis	RT		8	0.8	2.1E-6	1.3E-4
	KEGG_PATHWAY	Nitrogen metabolism	RT		7	0.7	3.3E-3	4.0E-2
	KEGG_PATHWAY	Glutathione metabolism	RT		6	0.6	3.6E-2	2.2E-1
	KEGG_PATHWAY	Tryptophan metabolism	RT		5	0.5	3.8E-2	2.1E-1
	KEGG_PATHWAY	Natural killer cell mediated cytotoxicity	RT		4	0.4	1.8E-2	1.5E-1
	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		4	0.4	3.3E-2	2.3E-1
	KEGG_PATHWAY	Ascorbate and aldarate metabolism	RT		4	0.4	9.7E-2	4.1E-1

Functional Clustering



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Arabidopsis thaliana
22 DAVID IDs

Options Classification Stringency: Medium

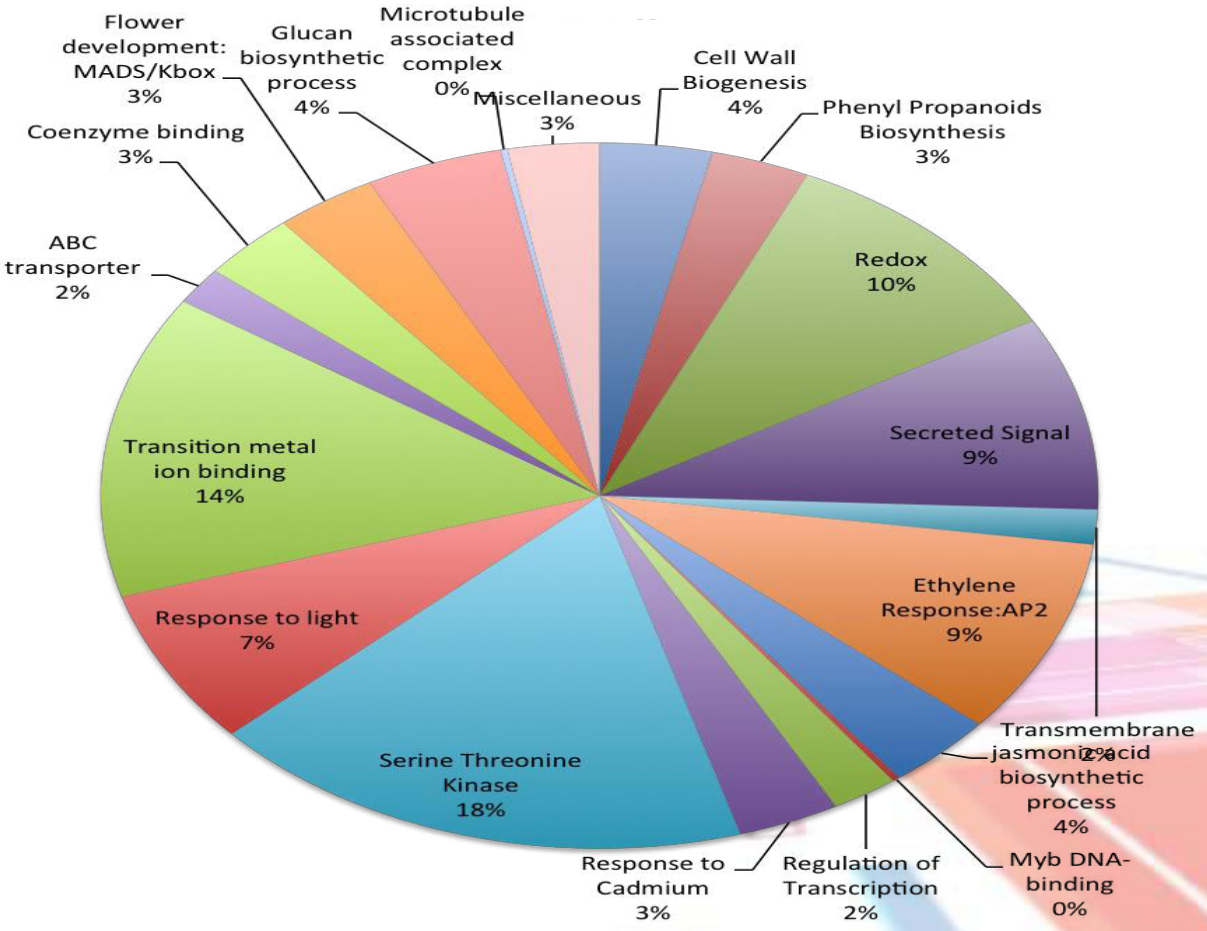
Rerun using options Create Sublist

8 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 11.19			Count	P_Value	Benjamini
	SP_PIR_KEYWORDS	nucleus	RT		19	9.5E-18	1.5E-16
	SP_PIR_KEYWORDS	transcription regulation	RT		16	2.1E-16	1.7E-15
	SP_PIR_KEYWORDS	Transcription	RT		16	2.9E-16	2.1E-15
	GOTERM_BP_FAT	transcription	RT		16	3.4E-12	4.1E-10
	GOTERM_BP_FAT	regulation of transcription	RT		17	9.4E-11	2.8E-9
	GOTERM_MF_FAT	transcription regulator activity	RT		16	1.0E-10	3.3E-9
	SP_PIR_KEYWORDS	dna-binding	RT		12	1.8E-9	7.0E-9
	GOTERM_MF_FAT	transcription factor activity	RT		13	1.0E-7	1.6E-6
	GOTERM_MF_FAT	DNA binding	RT		13	6.2E-6	6.6E-5
	Annotation Cluster 2	Enrichment Score: 8.94			Count	P_Value	Benjamini
	SP_PIR_KEYWORDS	flowering	RT		9	8.6E-18	2.7E-16
	SP_PIR_KEYWORDS	differentiation	RT		9	1.4E-16	1.1E-15
	SP_PIR_KEYWORDS	developmental protein	RT		9	2.5E-12	1.3E-11
	GOTERM_BP_FAT	post-embryonic development	RT		14	7.4E-12	4.5E-10
	GOTERM_BP_FAT	reproductive structure development	RT		13	3.2E-11	1.3E-9
	GOTERM_BP_FAT	reproductive developmental process	RT		13	1.0E-10	2.5E-9
	INTERPRO	Transcription factor, K-box	RT		6	2.3E-10	4.5E-9
	GOTERM_BP_FAT	flower development	RT		9	5.4E-10	1.1E-8
	PIR_SUPERFAMILY	PIRSF005231:transcription factor squa	RT		6	1.5E-9	1.2E-8
	SP_PIR_KEYWORDS	coiled coil	RT		9	1.6E-9	6.9E-9
	UP_SEQ_FEATURE	domain:K-box	RT		6	2.1E-9	1.2E-7
	UP_SEQ_FEATURE	domain:MADS-box	RT		6	1.4E-8	3.8E-7
	INTERPRO	Transcription factor, MADS-box	RT		6	3.7E-8	3.7E-7
	SMART	MADS	RT		6	3.6E-7	2.2E-6
	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT		11	7.0E-7	1.2E-5
	GOTERM_BP_FAT	regulation of RNA metabolic process	RT		11	7.4E-7	1.1E-5
	SP_PIR_KEYWORDS	DNA binding	RT		5	2.6E-6	8.8E-6
	SP_PIR_KEYWORDS	activator	RT		6	4.5E-6	1.4E-5
	GOTERM_MF_FAT	sequence-specific DNA binding	RT		6	2.9E-4	1.8E-3

Functional Clustering



GO Functional Analysis

Conclusions

After we get a list of genes, it is important to identify interesting Biological patterns

These functions are represented and classified as GO (Gene Ontology) terms

References

- Trevino, V., Falciani, F., and Barrera-Saldana, H.A. 2007. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med 13: 527-541.
- Kato, H., Saito, K., and Kimura, T. 2005. A perspective on DNA microarray technology in food and nutritional science. Curr Opin Clin Nutr Metab Care 8: 516-522.

Gene Network

Introduction

A network is a topology that connects its components based upon some relationship with each other.

- Networks are implemented as graphs.

.

Gene Network

Graphs

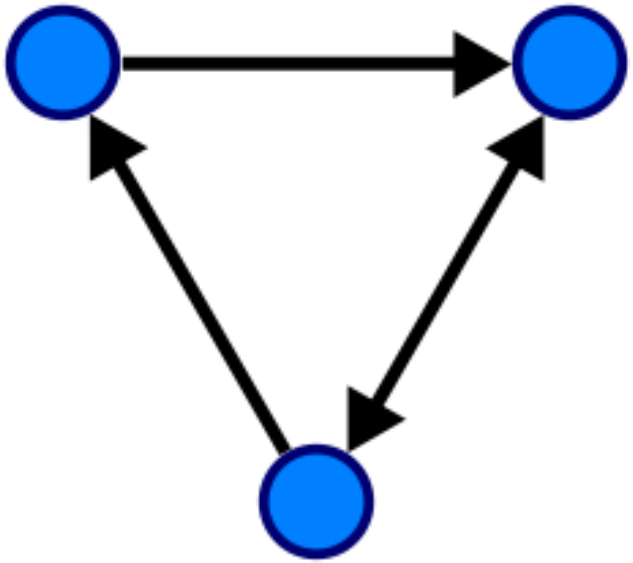
Graphs are abstract representation of a set of data objects which are somehow linked with each other.

- Directed graphs have an orientation or order as opposed to undirected graphs

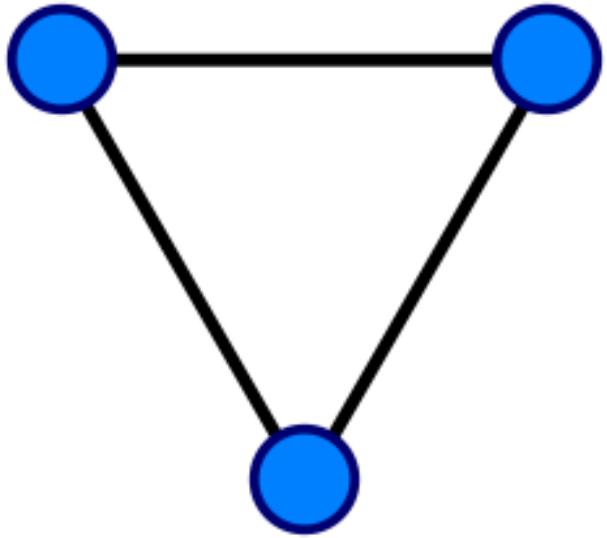
Gene Network

Graph Types

Directed Graph



Undirected/ Bidirectional Graph



Gene Network

Gene Regulatory Networks

- Genetic Regulatory Networks (GRNs) are the networks of genes based upon their regulatory interactions with each other
- Generally implemented as Directed Graphs
- Two Genes are connected if the expression of one gene modulates the effect of other gene by either activation or inhibition

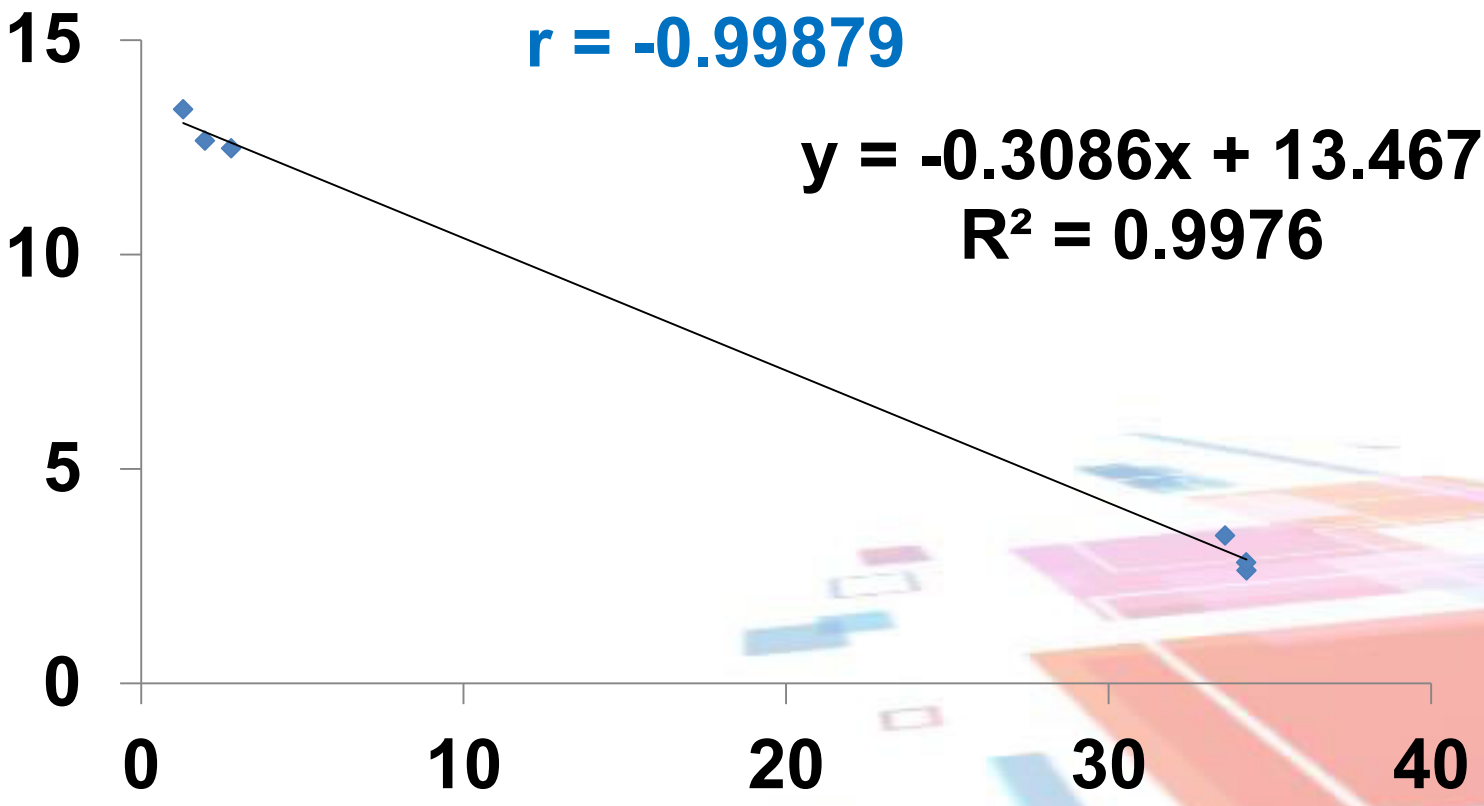
Gene Network

Modelling Gene Networks

- Correlation (coexpression) Networks
- Bayesian Networks
- Differential Equations
- Boolian Networks
- Gaussian Models

Gene Network

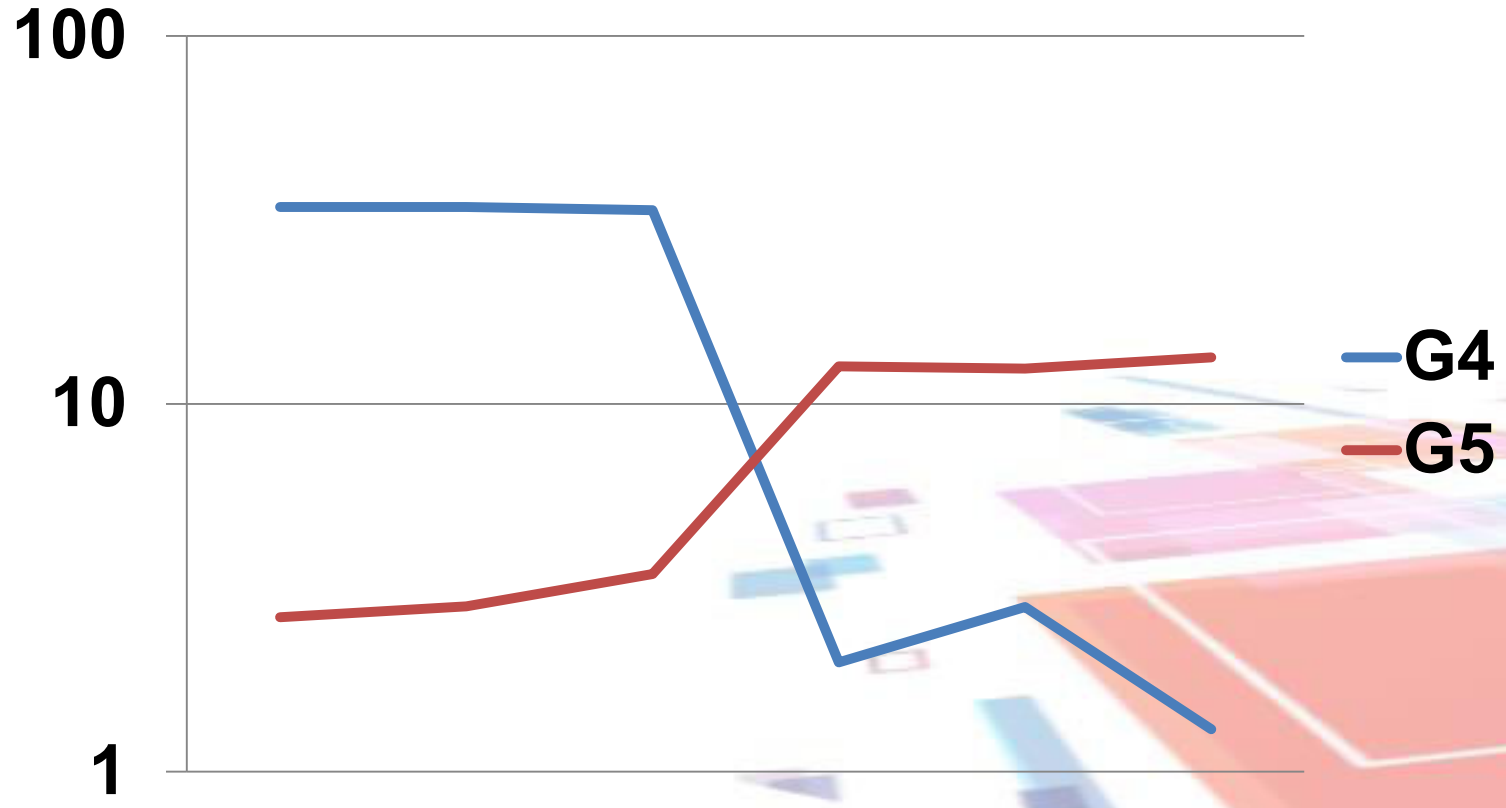
Correlation Networks (G4 vs G5)



Gene Network

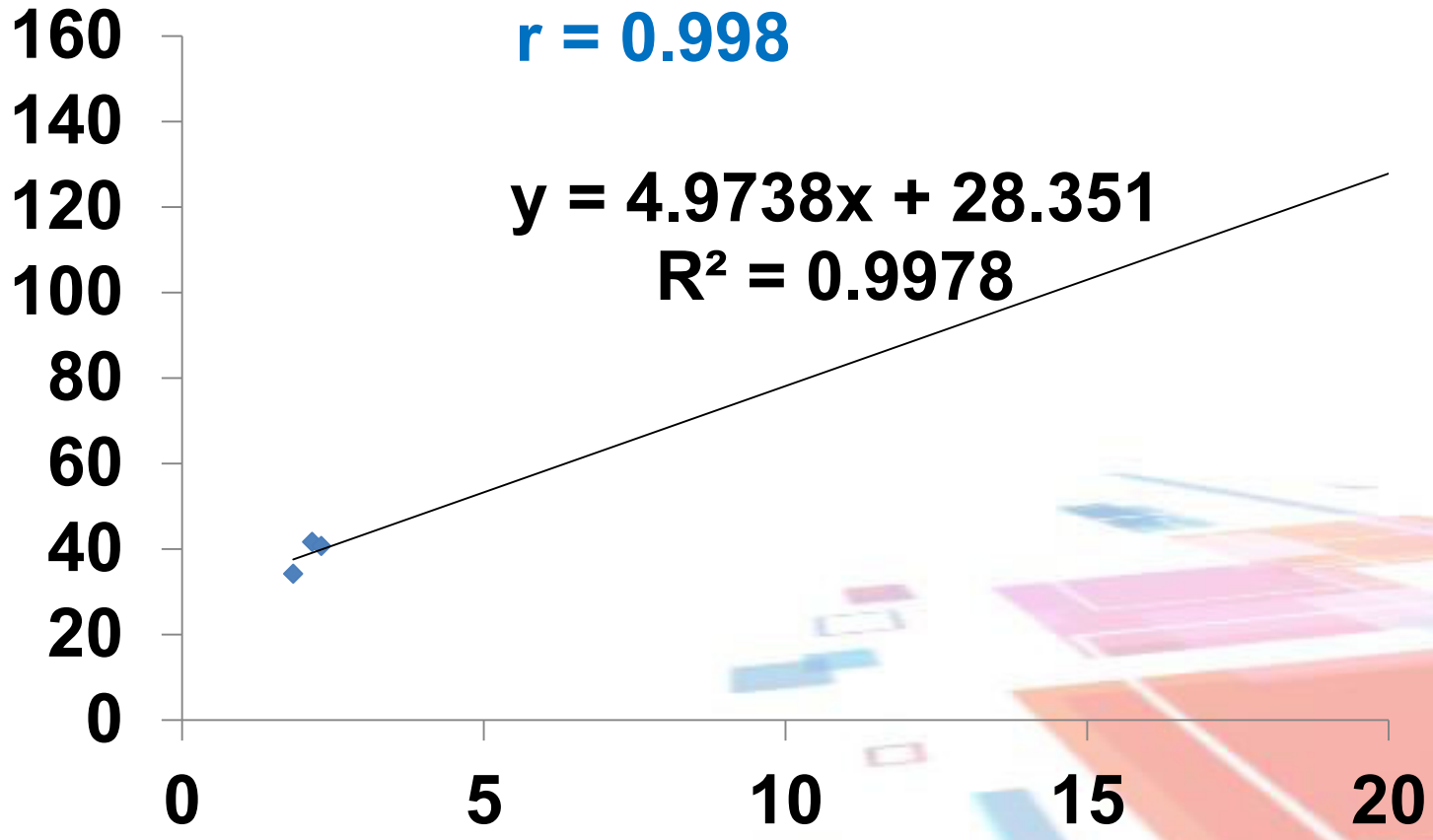
Correlation Networks (G4 vs G5)

r = -0.99879



Gene Network

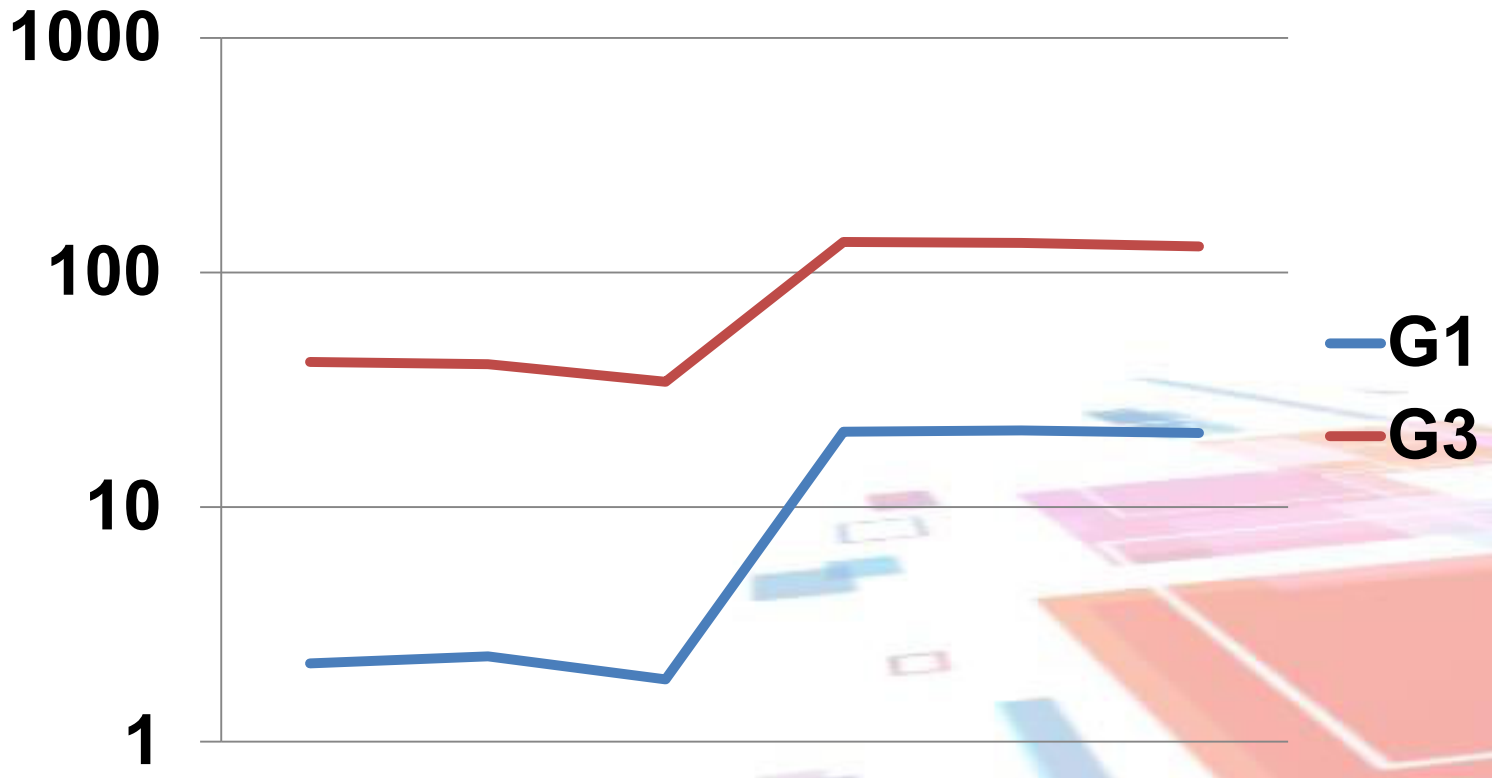
Correlation Networks (G1 vs G3)



Gene Network

Correlation Networks (G1 vs G3)

r = 0.998



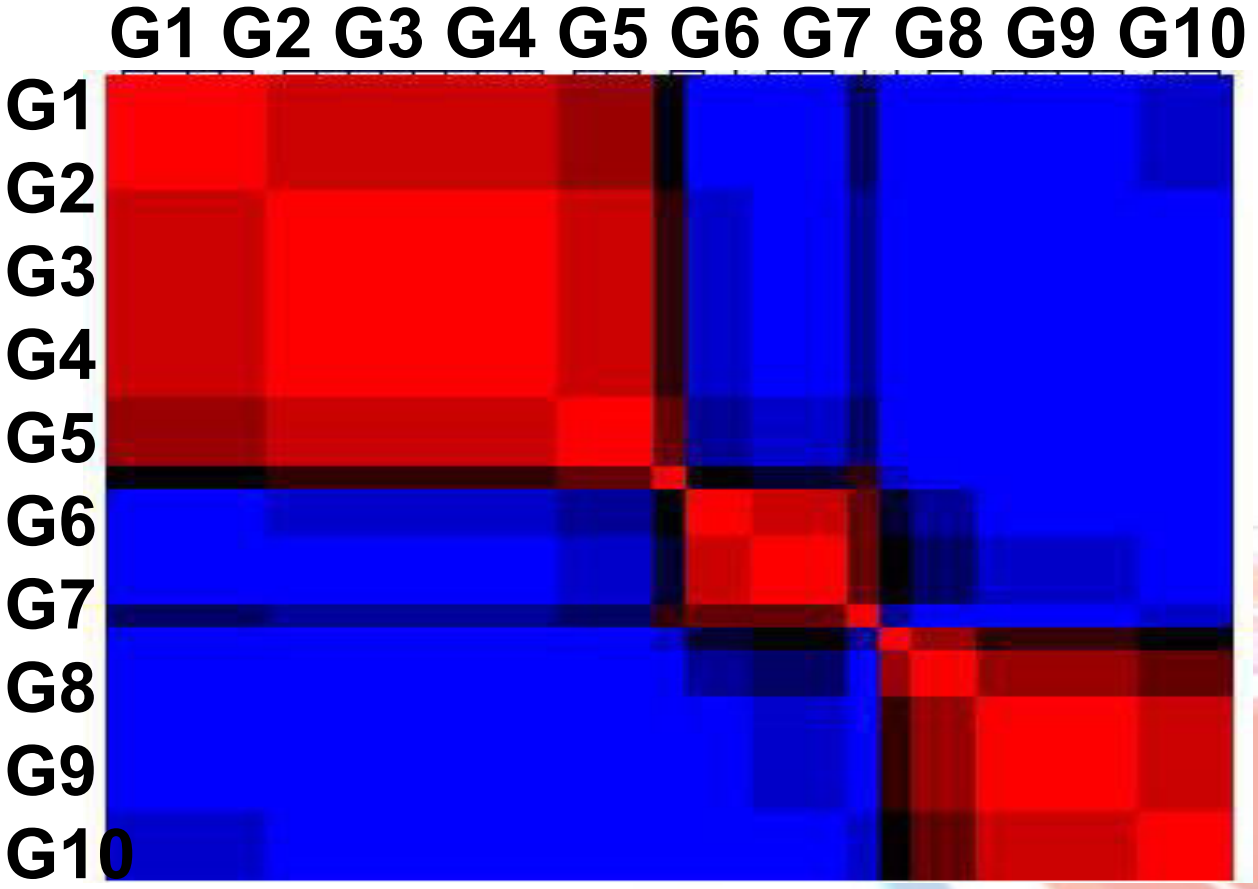
Gene Network

Correlation Networks

	G1	*G2*	*G3*	*G4*	*G5*
G1	1				
G2	0.573	1			
G3	0.998	0.540	1		
G4	-0.998	-0.579	-0.996	1	
G5	0.995	0.606	0.990	-0.998	1

Gene Network

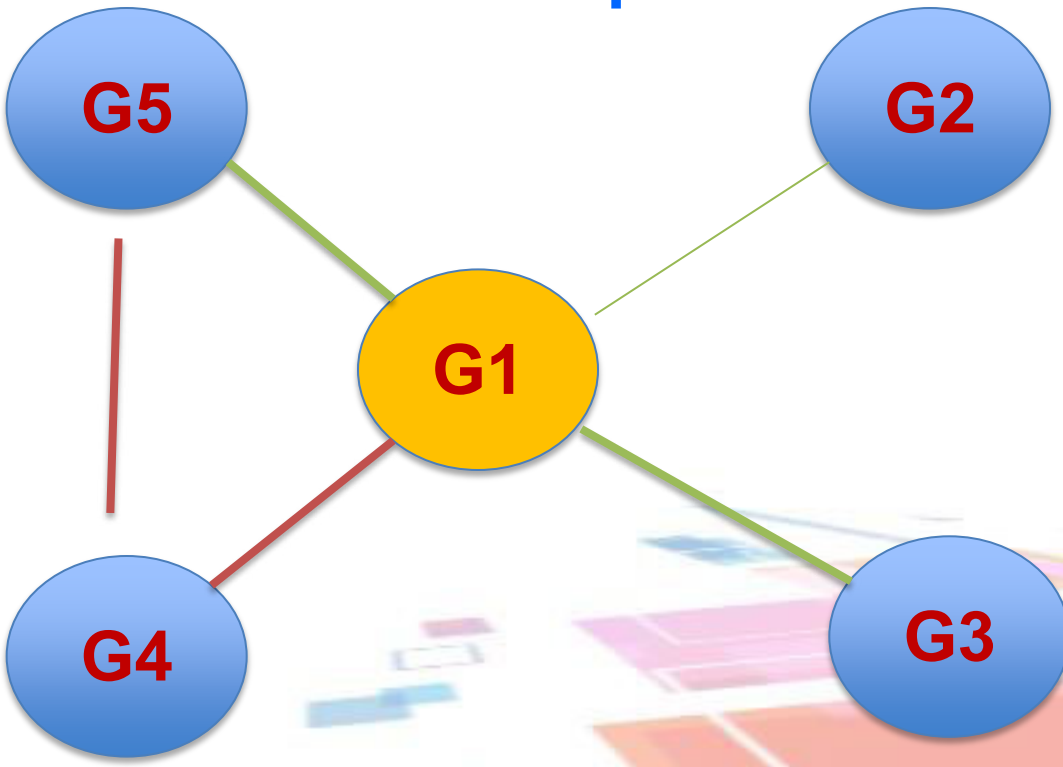
Correlation Network as heat Map



Gene Network

Correlation Network as heat Map

	G1
G1	1
G2	0.573
G3	0.998
G4	-0.998
G5	0.995



Positive correlation Negative correlation

Gene Network

Conclusions

Gene networks are abstract representation of relationships of genes based on gene to gene correlations

Genes are nodes while correlations are presented as edges

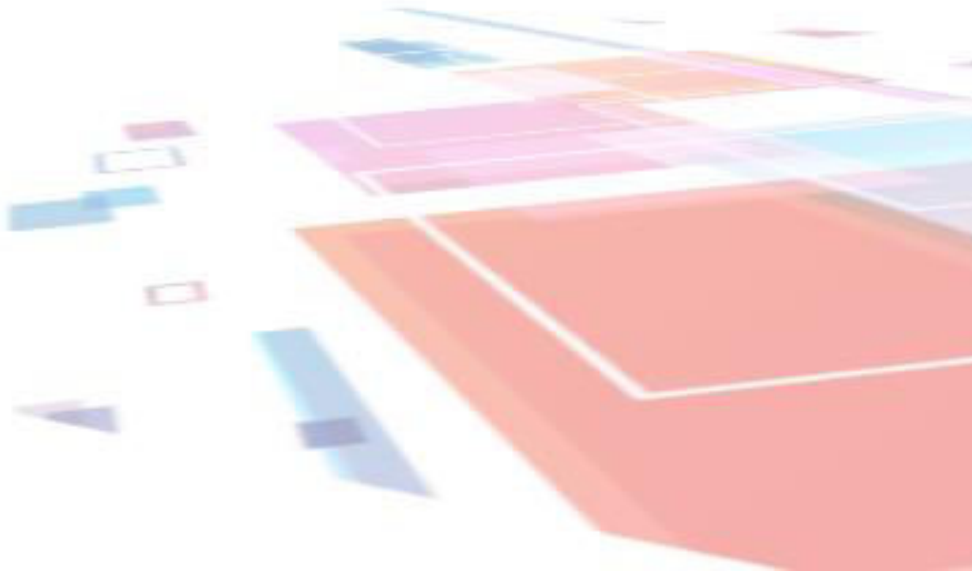
References

- Trevino, V., Falciani, F., and Barrera-Saldana, H.A. 2007. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med 13: 527-541.
- Kato, H., Saito, K., and Kimura, T. 2005. A perspective on DNA microarray technology in food and nutritional science. Curr Opin Clin Nutr Metab Care 8: 516-522.

ORF Prediction

ORF (Open Reading Frame)

Gene finding, specially in prokaryotes starts form searching for an open reading frames (ORF)



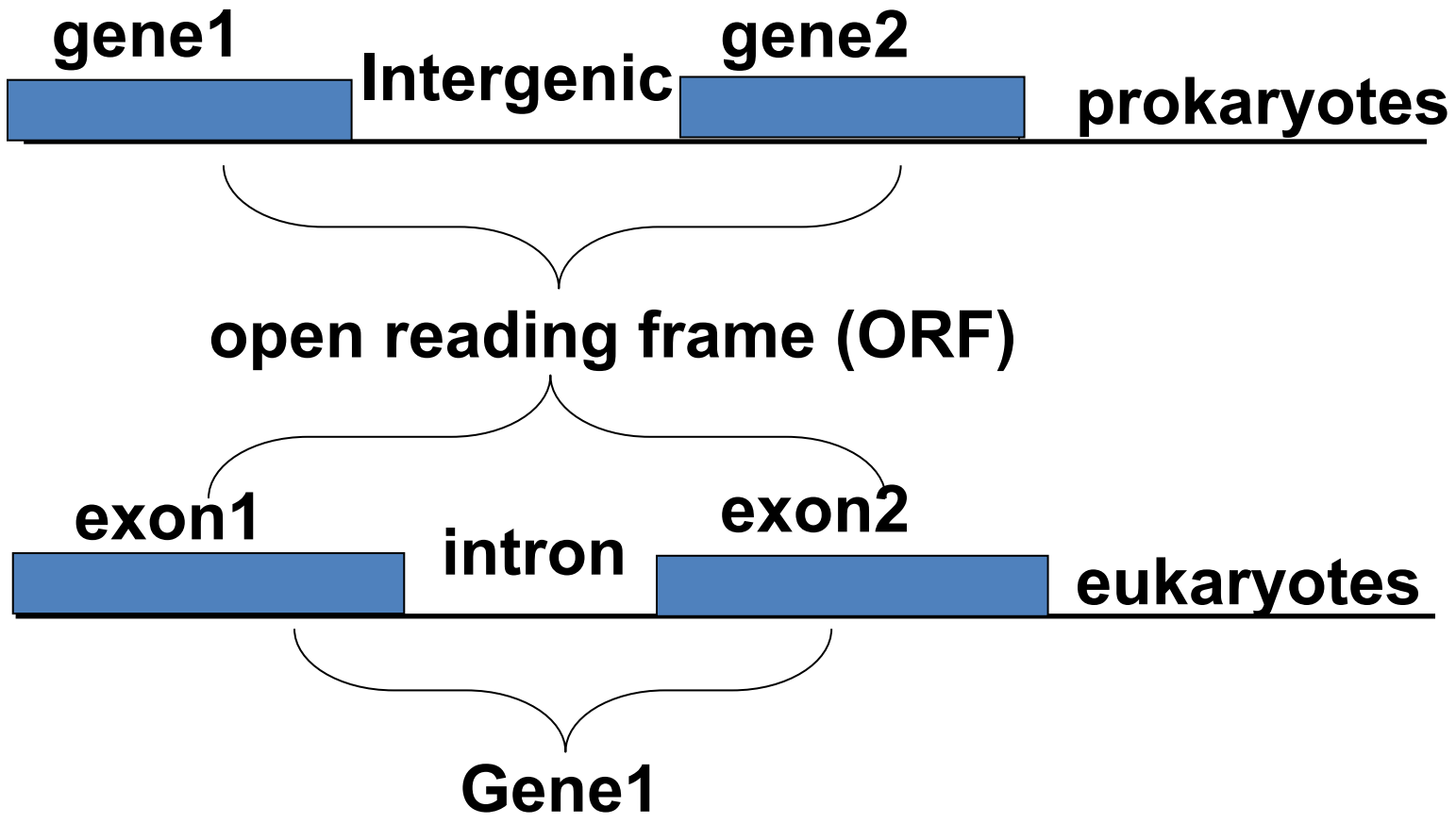
ORF Prediction

ORF (Open Reading Frame)

- An ORF is a sequence of DNA that starts with start codon “ATG” (not always) and ends with any of the three termination codons (TAA, TAG, TGA)

ORF Prediction

Coding regions: exons, genes



ORF Prediction

ORF and gene finding

ORF provide an important evidence in gene finding

- Generally longer ORFs are preferred
- However presence of ORF not necessarily means the region is translated to a functional product

ORF Prediction

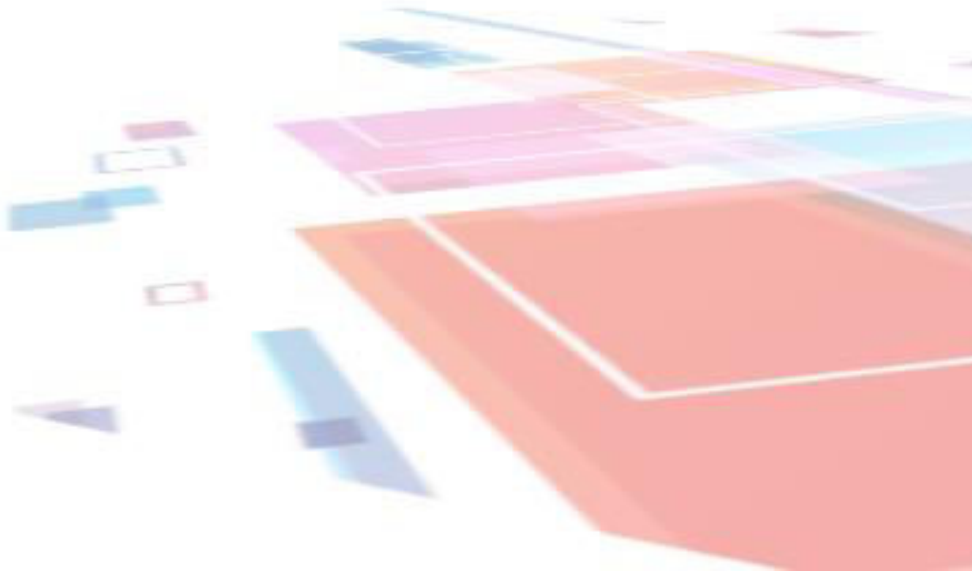
Reading Frames

Depending on the start point, there are six possible ways of translating any nucleotide sequence into amino acid sequence according to the genetic code

ORF Prediction

Reading Frames

- Three on forward strand and three on complementary strand



ORF Prediction

Six Frame translation

ORF Prediction

Conclusions

- An ORF is a sequence of DNA that starts with start codon “ATG” (not always) and ends with any of the three termination codons (TAA, TAG, TGA)

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

M Kanehisa

Oxford University Press, 2000

ORF Finders

ORF Finding

Long ORF may be a gene

- Expected 64/3 ~ 21 codons before we see a stop codon
- Genes are longer than this
- We might scan for ORF longer than a threshold

ORF Finders

Codon usage and likelihood ratio

An ORF is more ‘reliable’ if it has ‘likely’ codons

- We can do sliding window calculations to ORF having ‘likely’ codon usage

ORF Finders

Codon usage and likelihood ratio

An ORF is more ‘reliable’ if it has ‘likely’ codons

- However average vertebrate exon length (130 nucleotides) is too small for reliable peaks

ORF Finders

Codon usage and likelihood ratio

An improvement may be;

- In-Frame hexamer count

  i.e frequencies of pairs of consecutive codons

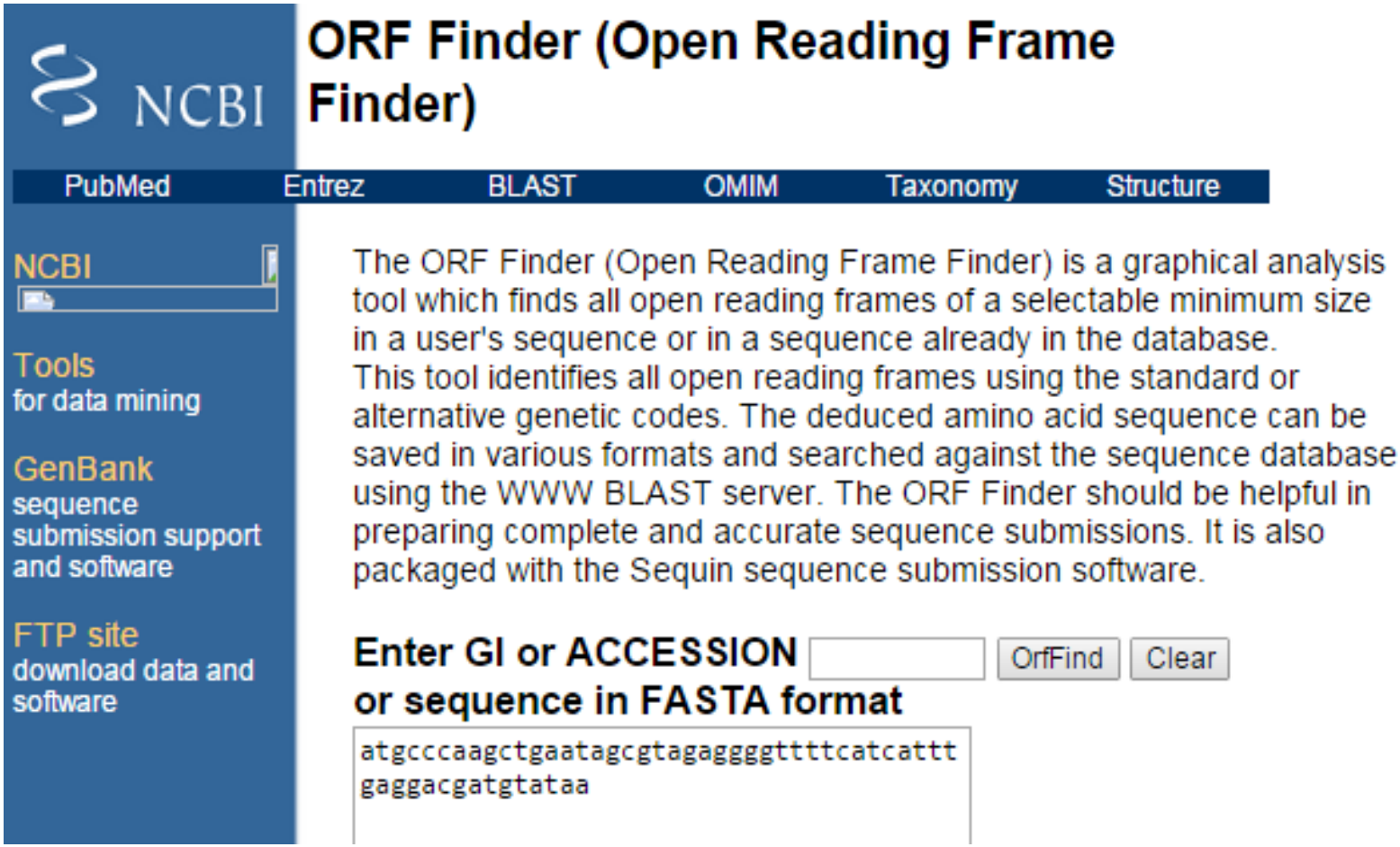
ORF Finders

ORF Finders

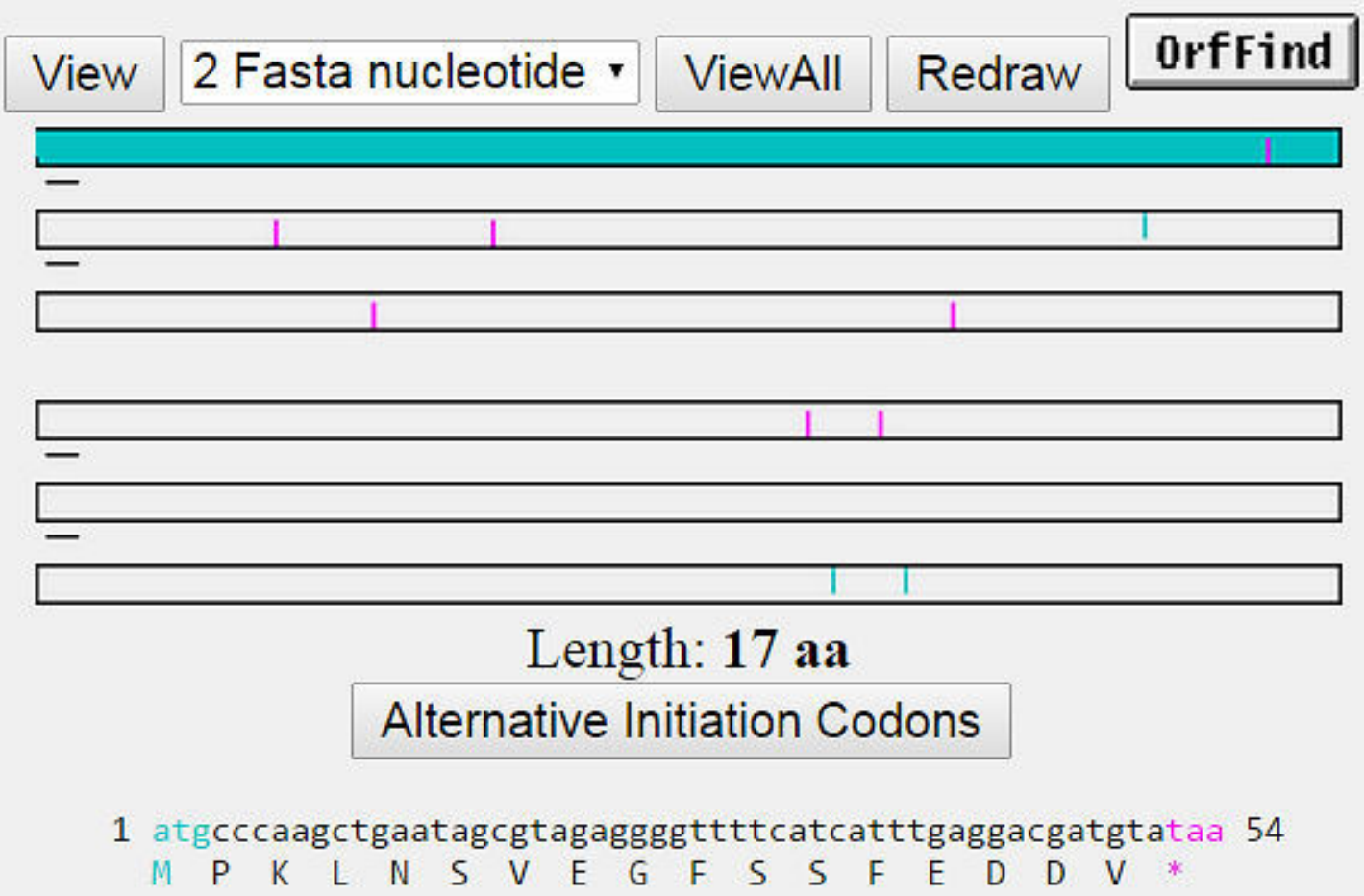
Tools are mainly based on pattern finding algorithms

- **NCBI's ORF Finder**
- **ORF Investigator**
- **OrfPredictor**

ORF Finders


NCBI
ORF Finder (Open Reading Frame Finder)
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
NCBI
Tools
for data mining
GenBank
sequence submission support and software
FTP site
download data and software
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.
Enter GI or ACCESSION
OrfFind
Clear
or sequence in FASTA format
atgcccaagctgaatagcgtagaggggttttcatcattt
gaggacgatgtataa

ORF Finders


View
2 Fasta nucleotide
ViewAll
Redraw
OrfFind
Length: 17 aa
Alternative Initiation Codons
1 atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa 54
M P K L N S V E G F S S F E D D V *

ORF Finders

Conclusions

ORF provide an important evidence in gene finding

- Generally longer ORFs are preferred
- However presence of ORF not necessarily means the region is translated to a functional product

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

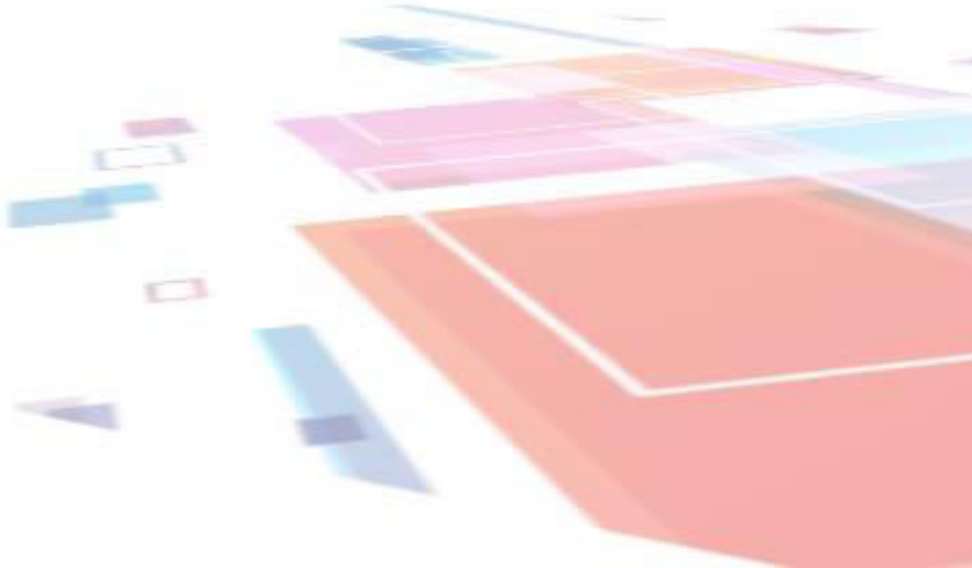
M Kanehisa

Oxford University Press, 2000

TSS Prediction

Translation Start Site (TSS)

Translation starts with ATG that codes for methionine in a polypeptide



TSS Prediction

Translation Start Site (TSS)

GCCATGGCGA …

ACGATGCTGT …

GACATGGTAC …

AGGATGGGCT …

TSS Prediction

Assumption

Certain nucleotides prefer to be around TSS than others

- The “biased” nucleotide distribution is information is a basis for translation start prediction

TSS Prediction

Coding Potential

Hexamer frequencies in coding versus non-coding regions may provide important insights Frequency of X(A,G,C,T) at position i is

$$F_i(X) = \Sigma \log(C_i(X)/N_i(X))$$

TSS Prediction

Frequency table (Biased Nucleotide distribution)

ATG

A	17.36	19.01	17.36	48.76	28.93	15.70	21.49	23.14	19.83	21.49	25.62	15.70
C	16.53	28.93	57.85	5.79	39.67	50.41	22.31	38.84	23.97	27.27	31.40	38.02
G	46.28	29.75	19.01	42.98	14.88	26.45	42.98	23.97	33.88	32.23	25.62	29.75
T	19.83	22.31	5.79	2.48	16.53	7.44	13.22	14.05	22.31	19.01	17.36	16.53

-4 -3 -2 -1 ATG +3 +4 +5 +6

TSS Prediction

Example

Which one is more probable to be a Translation Start?

CACC ATA GC

TCGA ATG TT

TSS Prediction

Solution

We can use frequency table and the scoring function as under;

$$S_i = \sum \log (F_i (X)/0.25)$$

We can call it

Information Content (IC)

TSS Prediction

Frequency table (Biased Nucleotide distribution)

ATG

A	17.36,	19.01,	17.36,	48.76,	28.93,	15.70,	21.49,	23.14,	19.83,	21.49,	25.62,	15.70,
C	16.53,	28.93,	57.85,	5.79,	39.67,	50.41,	22.31,	38.84,	23.97,	27.27,	31.40,	38.02,
G	46.28,	29.75,	19.01,	42.98,	14.88,	26.45,	42.98,	23.97,	33.88,	32.23,	25.62,	29.75,
T	19.83,	22.31,	5.79,	2.48,	16.53,	7.44,	13.22,	14.05,	22.31,	19.01,	17.36,	16.53}

-4 -3 -2 -1 +3 +4 +5 +6

C A

TSS Prediction

CACC ATA GC

log (58/0.25) + log (49/0.25) + log (40/0.25) + log (50/0.25) + log (43/0.25) + log (49/0.25)

=13.69

TCGA ATG TT

log (6/0.25) + log (6/0.25) + log (15/0.25) + log (7/0.25) + log (13/0.25) + log (14/0.25)

= 9.44

TSS Prediction

Algorithm

- Build a mathematical model, based on collected translation start sequence
- For each candidate translation start sequence, apply the model and get a score
- If the score is larger than zero, predict it is a “translation start”; the higher score, the higher the probability the prediction is true

TSS Prediction

Conclusions

- TSS prediction can be an important step in gene prediction
- TSS can be predicted while using the frequency of neighboring nucleotides

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

M Kanehisa

Oxford University Press, 2000

Prediction of splice junctions

Splice Junctions

Donor site

- Coding region | GT

Acceptor

- YAG | coding region
- **Canonical form**
  - GT-AG: 99.24%

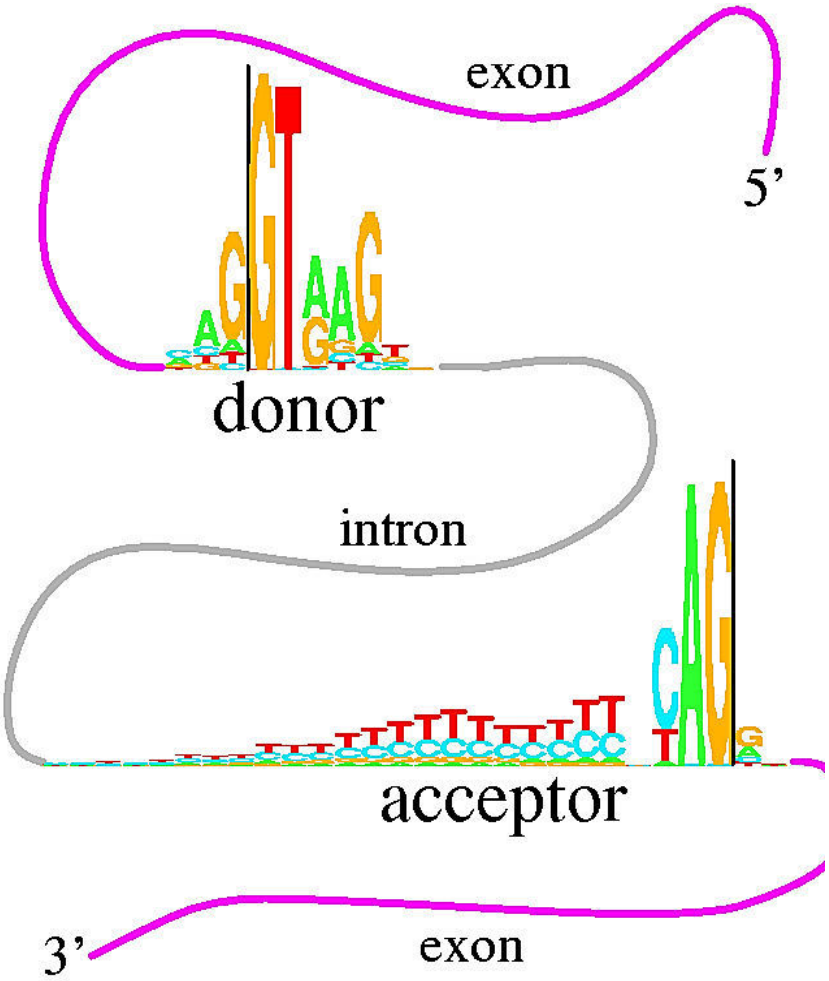
Prediction of splice junctions

Splice Junctions

Like TSS, the flanks of splice junctions show “biased” distributions of nucleotides in certain positions

- These biased distributions of nucleotides are the basis for prediction of splice junctions

Prediction of splice junctions



Prediction of splice junctions

Sequence LOGOS

A visual representation of a position-specific distribution

Easy for nucleotides, but we need colour to depict up to 20 amino acid proportions.

Prediction of splice junctions

Sequence LOGOS

Overall height at position is proportional to the information content

Prediction of splice junctions

Sequence LOGOS

- Proportions of each nucleotide/amino acid are in relation to their observed frequency, with most frequent on top, next most frequent below

Prediction of splice junctions

Non Canonical Splice Junctions

In addition to canonical **GT-AG (99.24%)**;

GC-AG: 0.69%

AT-AC: 0.05%

Others: 0.02%

Prediction of splice junctions

Information Content (IC)

$S_i = \Sigma \log (F_i (X)/0.25)$

If every nucleotide has 0.25 frequency in a position, then the position's information content is ZERO

Prediction of splice junctions

Information Content (IC)

$S_i = \Sigma \log (F_i (X)/0.25)$

Use “information content as a criterion for determining the length of flanks

Prediction of splice junctions

Accepter site prediction

	-6	-5	-4	-3	-2	-1	1
A	12.7	9.5	26.2	6.3	100	000	21.4
C	40.5	36.5	33.3	68.2	000	000	2.0
G	2.4	6.3	13.5	000	000	100	62.7
T/U	44.5	47.6	27.0	25.2	000	000	7.90

Multiple positions have high information content

Prediction of splice junctions

Donor site prediction

	-3	-2	-1	1	2	3	4
A	34.0	60.4	9.2	000	000	52.6	71.3
C	36.3	12.9	3.3	000	000	2.8	7.6
G	18.3	12.5	80.3	100	000	41.9	11.8
T/U	11.4	14.2	7.3	000	100	2.5	9.3

Multiple positions have high information content

Prediction of splice junctions

Algorithm

- Mathematical model: F_i (X): frequency of X (A, C, G, T) in position I
- Score a segment as a candidate donor/acceptor site by

  $\Sigma \log (F_i (X)/0.25)$

Prediction of splice junctions

Algorithm

- For each candidate sequence, apply the model and get a score
- If the score if larger than zero, predict it is “donor/acceptor”; the higher score, the higher the probability the prediction is true

Prediction of splice junctions

Conclusions

Like TSS, the flanks of splice junctions show “biased” distributions of nucleotides in certain positions

- These biased distributions of nucleotides can be used for prediction of splice junctions

References

Biological Sequence Analysis

R Durbin, S Eddy, A Krogh and G Mitchison

Cambridge University Press, 1998.

Bioinformatics The machine learning approach

P Baldi and S Brunak

The MIT Press, 1998

Post-Genome Informatics

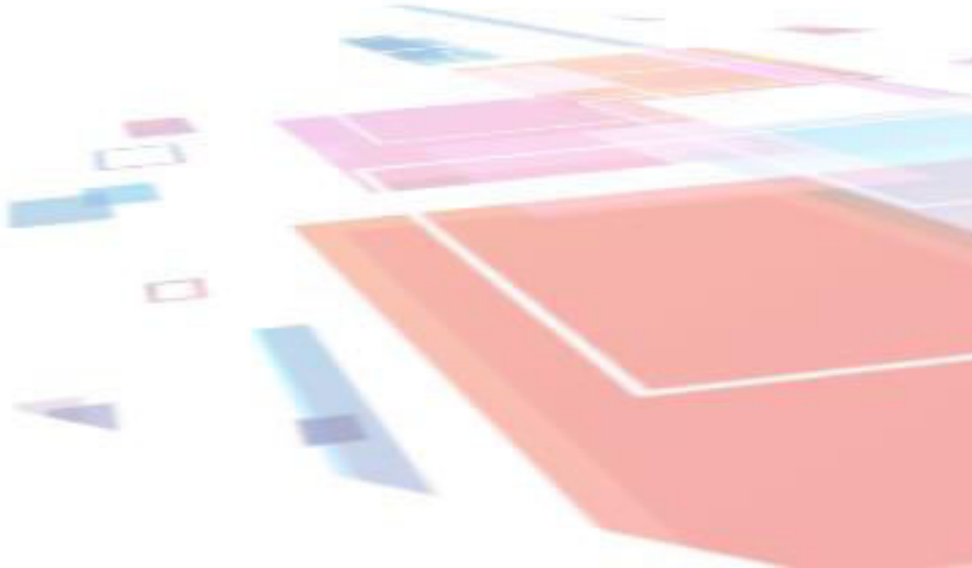
M Kanehisa

Oxford University Press, 2000

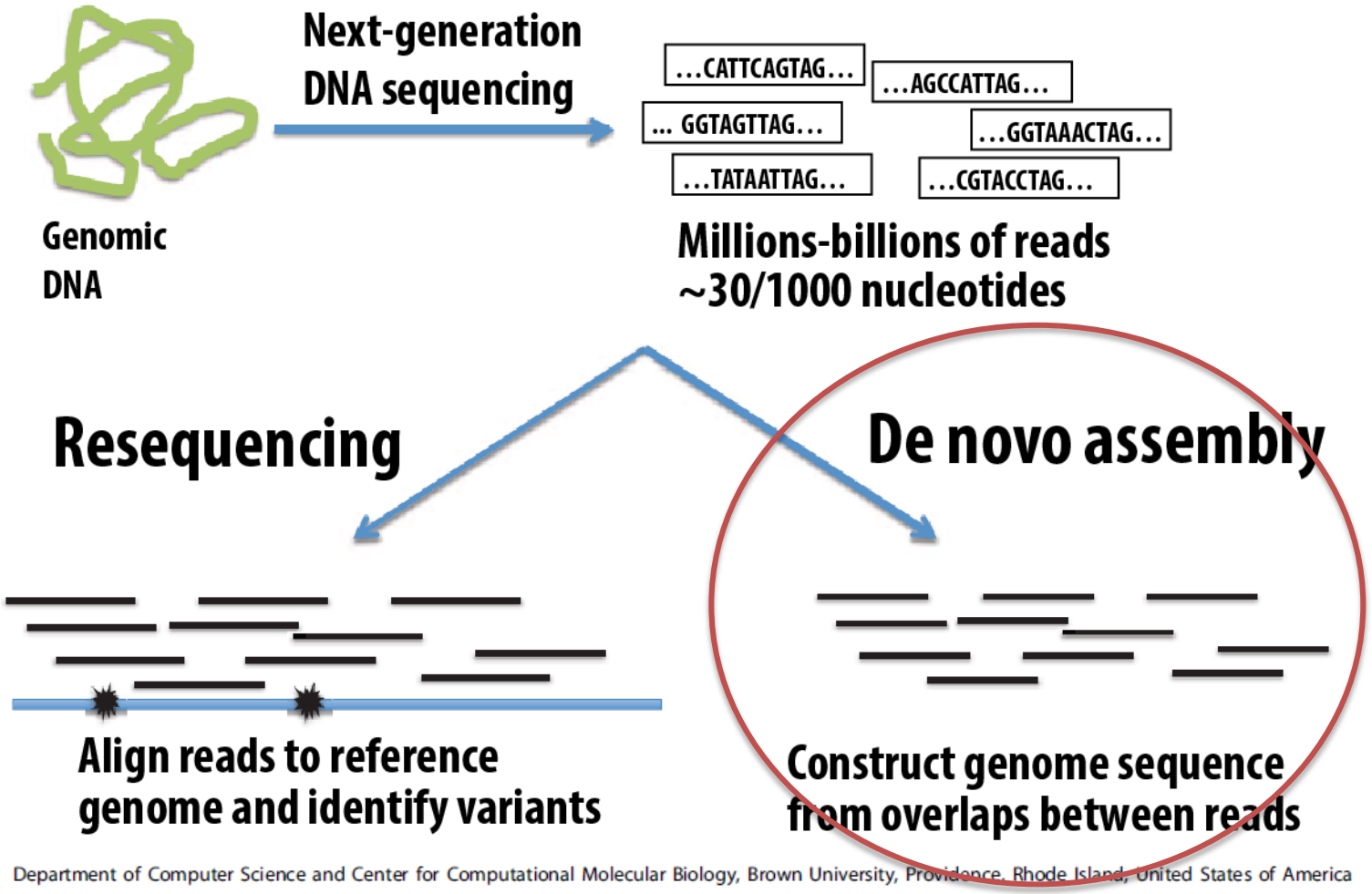
Genome Assembly

Outline

- Introduction
- Vocabulary terms
- Overview
- Conclusions



Genome Assembly



Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America

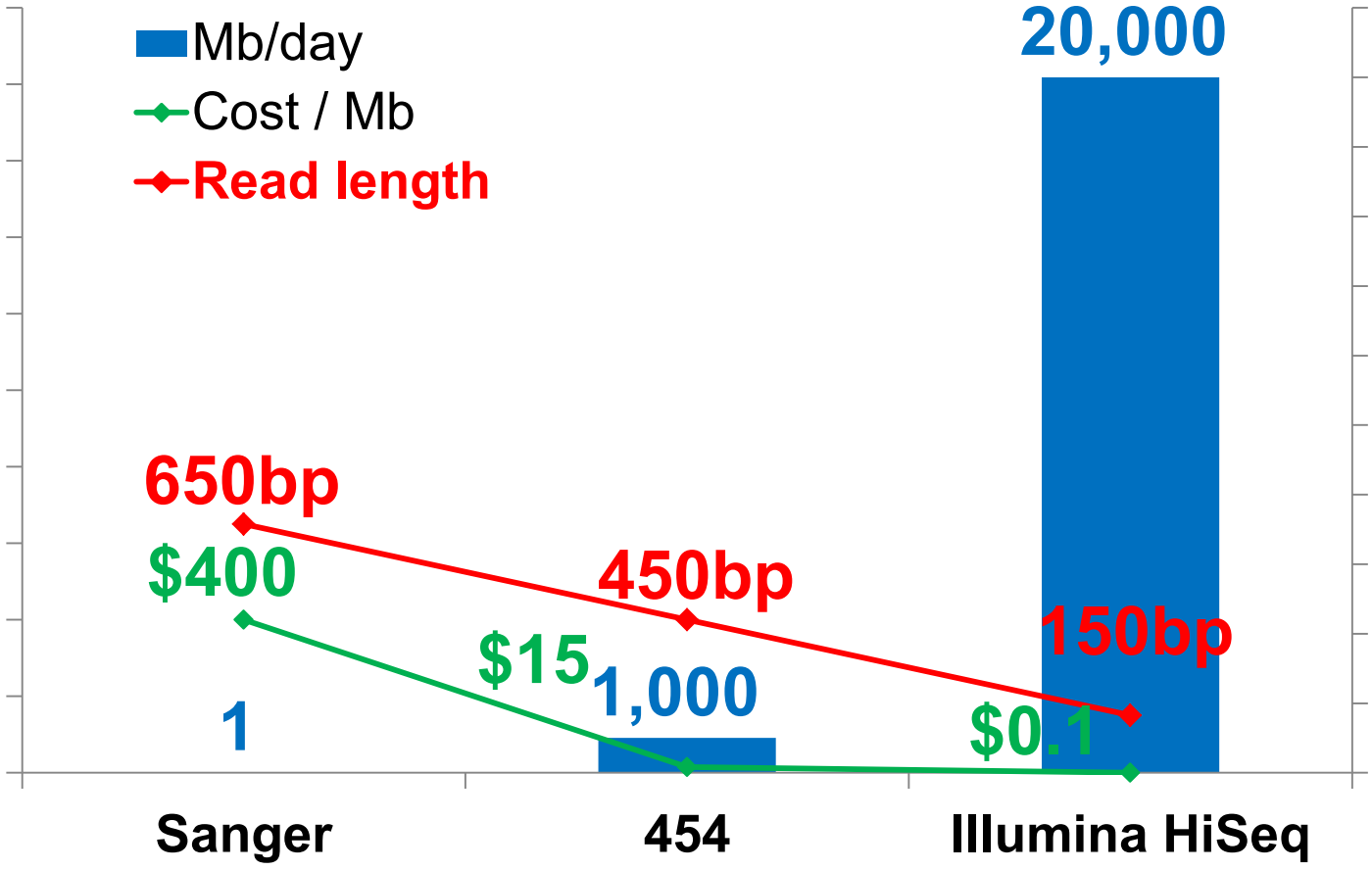
Genome Assembly

Genome Assembly

The process of reconstruction of original chromosomes based on overlaps between short sequence reads

- It groups reads into **contigs** and Contigs to **Scaffolds**

Genome Assembly



Genome Assembly

Vocabulary

- **Contig:** A consensus sequence of DNA that has been assembled from overlapping DNA fragments
- **Scaffolds:** One or more contigs linked together by unknown sequence

Genome Assembly

Contig 1 Contig 2

ATGCTANNNNNNNNNNNNNNNNNNNNAGCTA---

Genome Assembly

Vocabulary

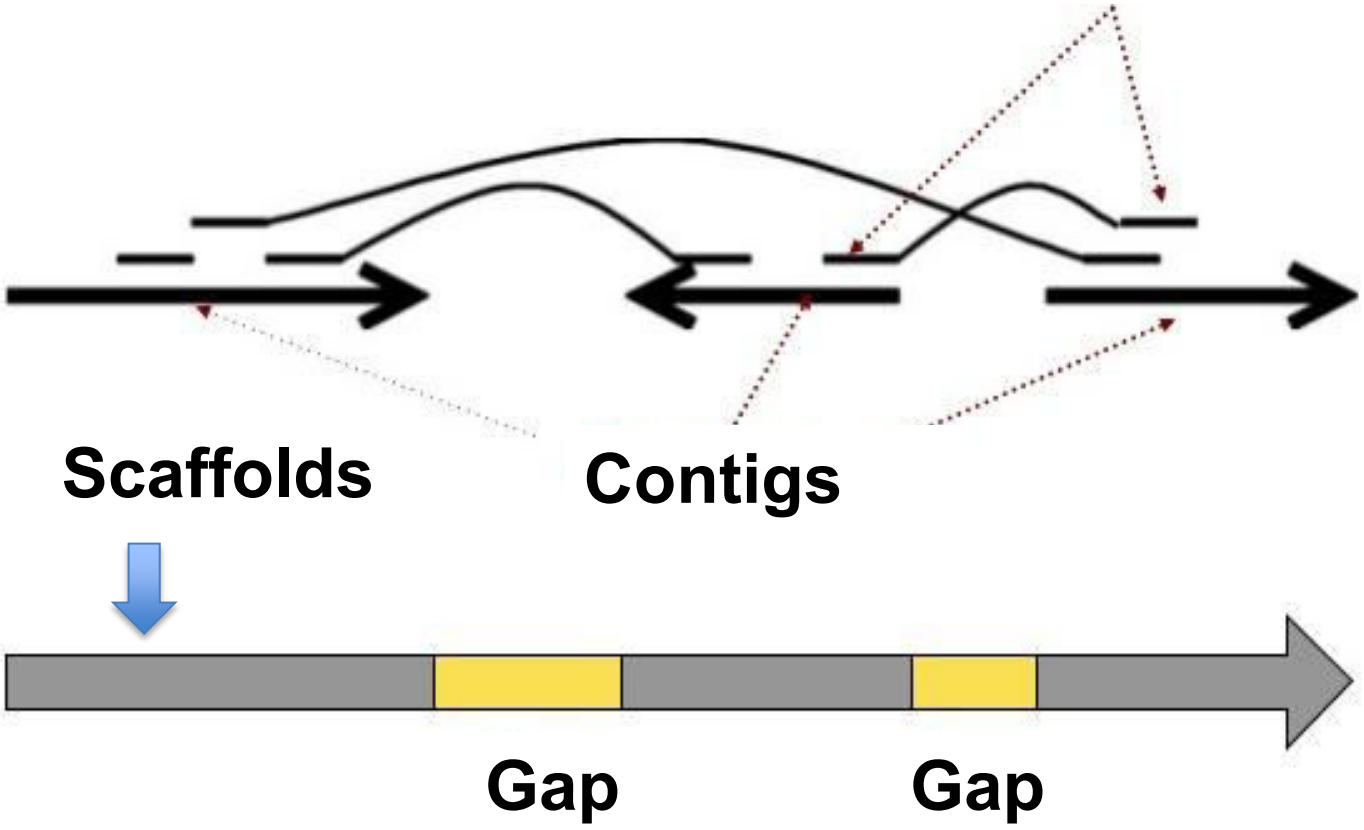
Captured Gap:

A gap within a scaffold

- The order and orientation of the contigs spanning the gap is known

Genome Assembly

Paired-end Reads



Genome Assembly

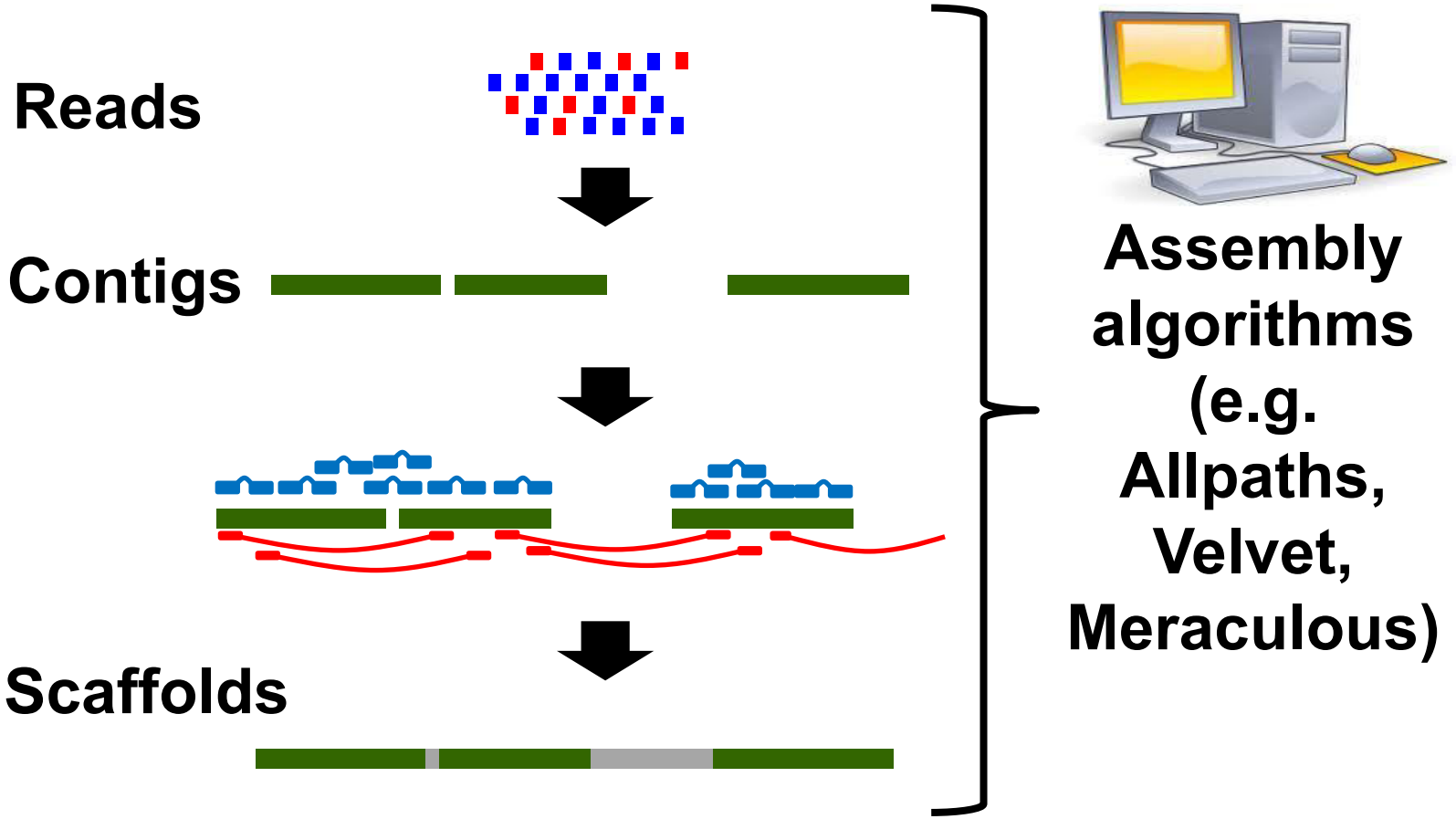
Vocabulary

Mis-assemblies:

Include regions where a genome is incorrectly re-arranged

- Or the places where large chunks of DNA sequence are simply deleted and the surrounding sequences just crunched together

Genome Assembly



MGM Workshop Assembly Tutorial Alicia Clum, DOE Joint Genome Institute, Walnut Creek, CA May 14, 2012

Genome Assembly

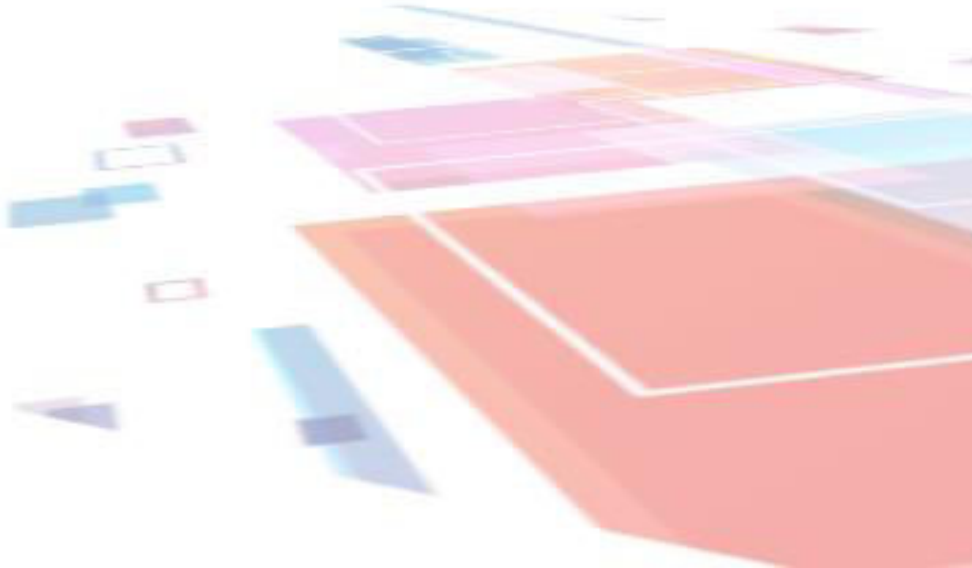
Conclusions

- The process of reconstruction of original chromosomes based on overlaps between short sequence reads
- Reads are assembled into contigs which are then assembled into scaffolds

Genome Assembly Overview

Outline

- Sample preparation
- Sequencing
- Assembly
- Finishing
- Conclusions



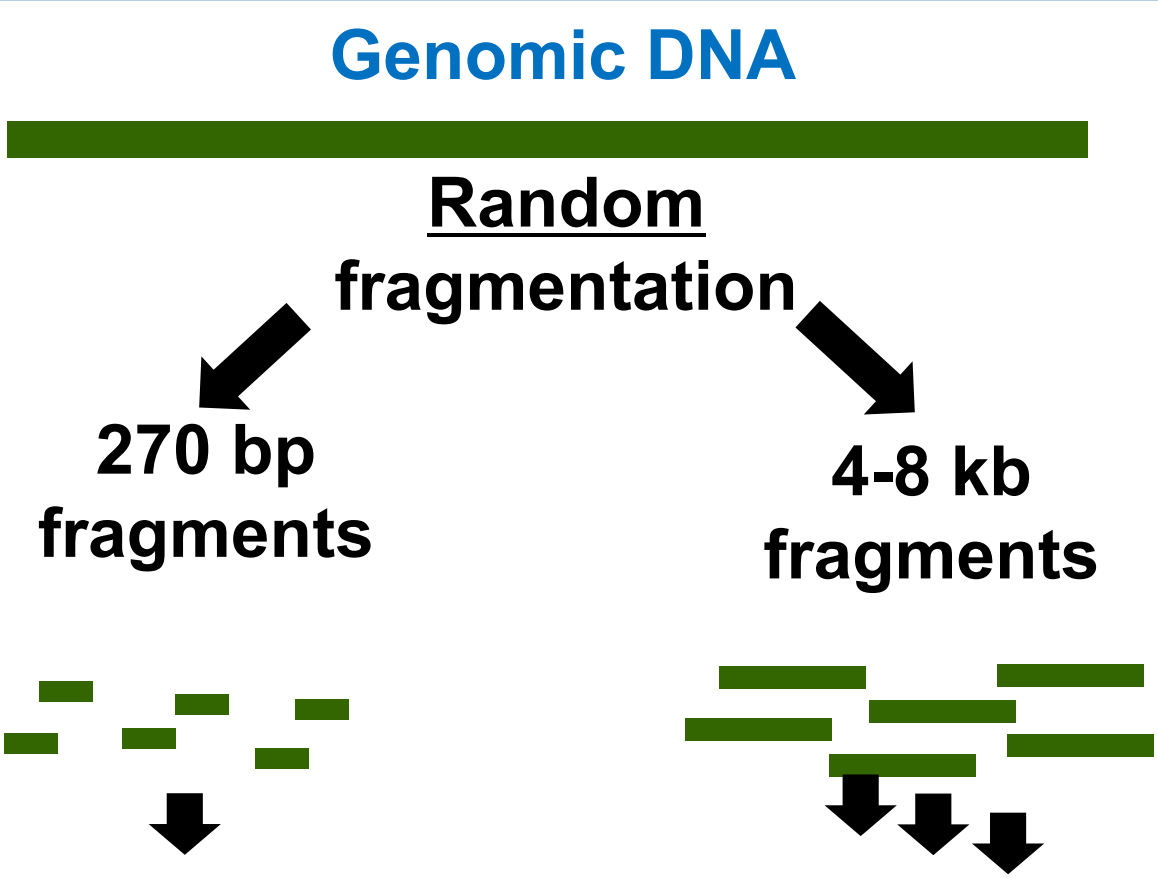
Genome Assembly Overview

Sample preparation

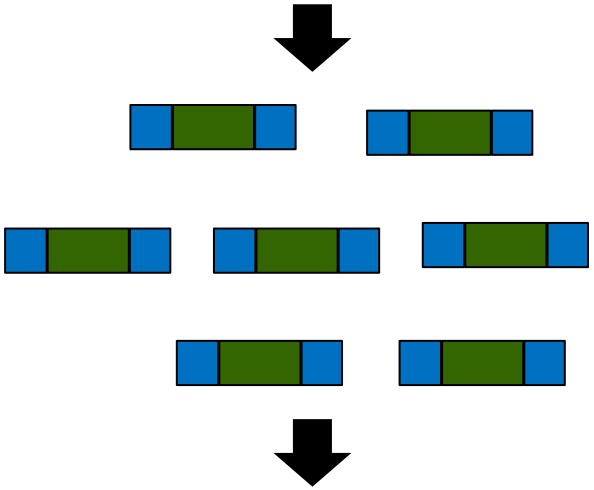
DNA is collected from the biological sample followed by library preparations

Schatz et al. Genome Biology 2012 13:243

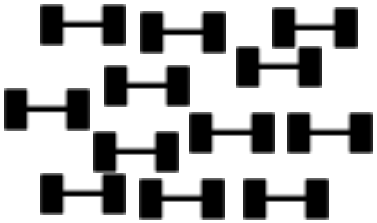
Genome Assembly Overview

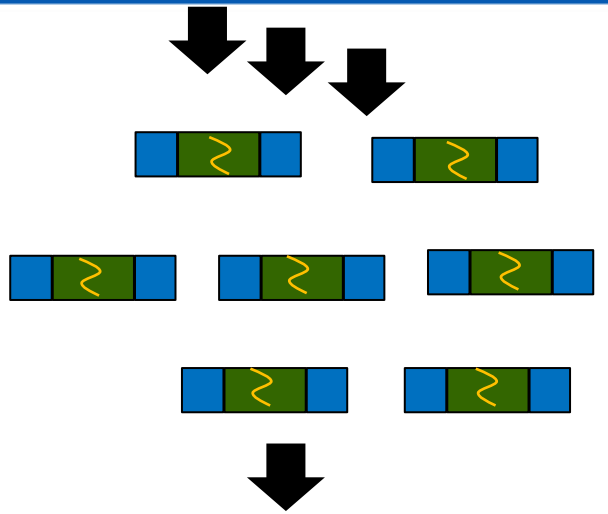


Genome Assembly Overview

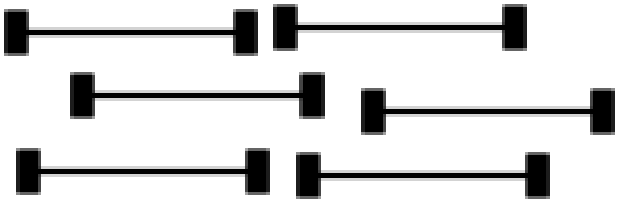


Paired-end short insert reads (10's millions)





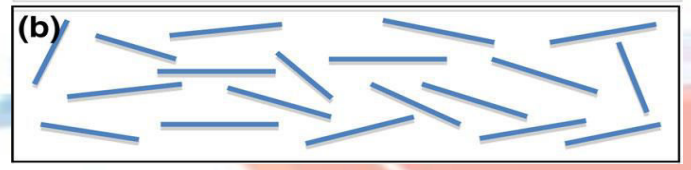
Paired-end long insert reads (10's millions)



Genome Assembly Overview

Sequencing

The output from the sequencer consists of many billions short, unordered DNA fragments (strings of ATGCs) from random positions in genome



Schatz et al. Genome Biology 2012 13:243

Genome Assembly Overview

Assembly

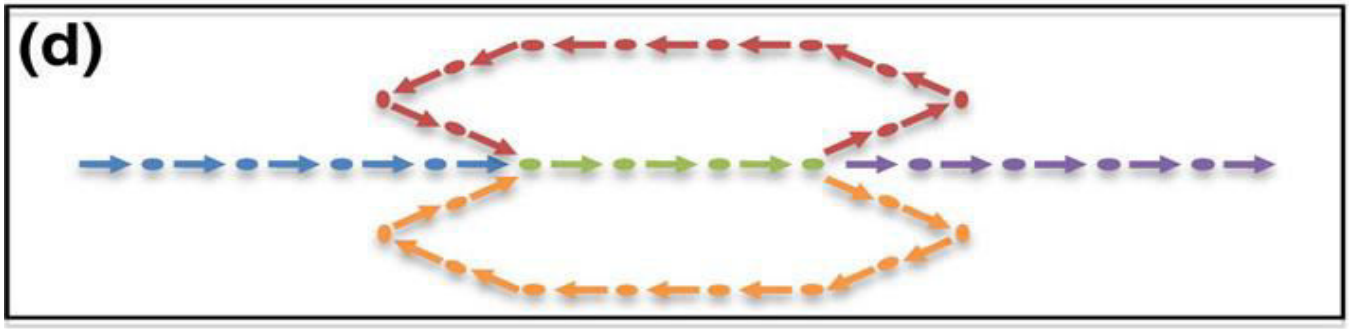
The short fragments are compared with each other to discover how they overlap

Schatz et al. Genome Biology 2012 13:243

Genome Assembly Overview

Assembly

The overlap relationships are captured in a large assembly graph shown as nodes representing *k*-mers or reads, with edges drawn between overlapping *k*-mers or reads

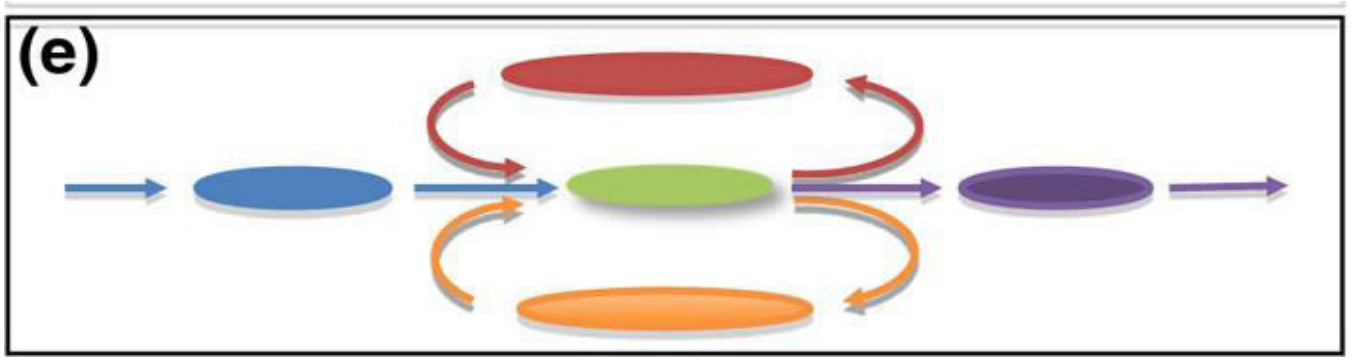


Schatz et al. Genome Biology 2012 13:243

Genome Assembly Overview

Refinement and simplification

The assembly graph is refined to correct errors and simplify into the initial set of contigs, shown as large ovals connected by edges

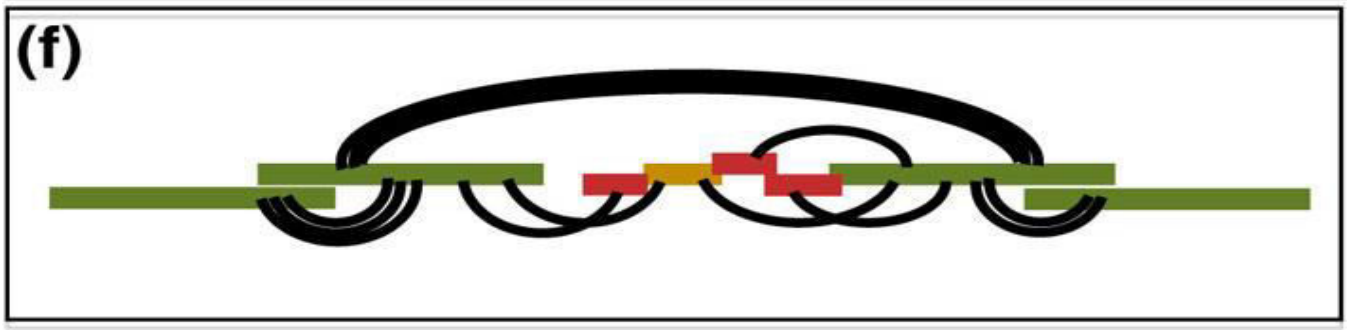


Schatz et al. Genome Biology 2012 13:243

Genome Assembly Overview

Refinement and simplification

Finally, mates, markers and other long-range information are used to order and orient the initial contigs into large scaffolds, as shown as thin black lines connecting the initial contigs



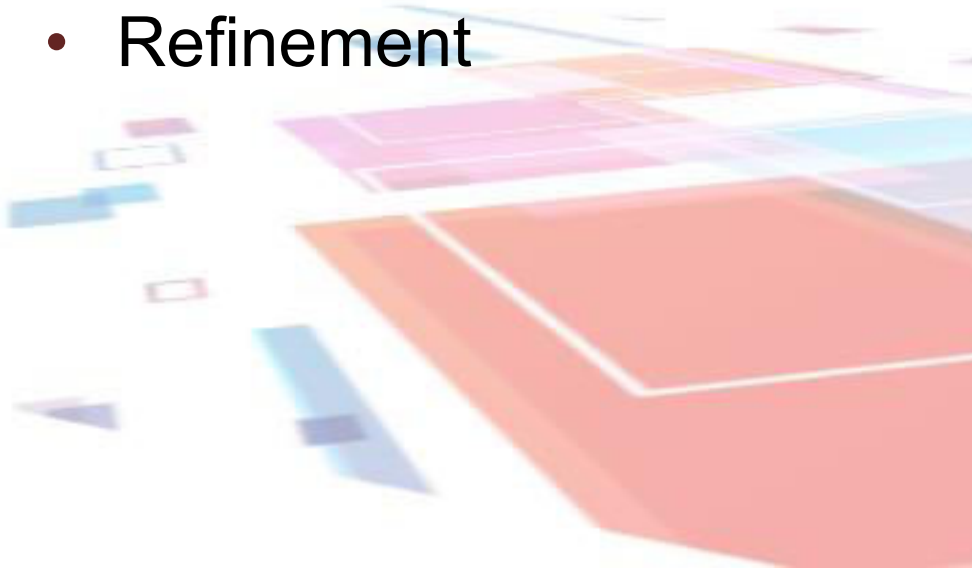
Schatz et al. Genome Biology 2012 13:243

Genome Assembly Overview

Conclusions

Genome assembly is performed as follows;

- Sample preparation
- Sequencing
- Assembly
- Refinement



Genome Assembly Planning

Factors

Following factors may be important in planning an Assembly project;

- **Size of Genome**
- Repeat content
- Related species Assembled
- Strategy for performing Assembly

Genome Assembly Planning

Size of the genome

Size of the Genome can be estimated from the ploidy **(N)** of the organism and the DNA content per cell

- Haploid/Monoploid
- Diploid
- Tri/Tetra/-------
- Polyploid

Genome Assembly Planning

Size of the genome

- This will affect:
  - How many reads will be required to attain sufficient coverage (typically **10x to 100x**)

Genome Assembly Planning

Size of the genome

- This will affect:
  - How many reads will be required to attain sufficient coverage (typically **10x to 100x**)

Genome Assembly Planning

Coverage

The average number of times any given base in the genome is sequenced

It can be derived by dividing the total length of acquired sequences by the genome length

Genome Assembly Planning

Coverage

$$C = NL / G$$

N: Total number of reads

L: Length of a Read

G: Genome size

Represented as **X**

Genome Assembly Planning

Coverage

Example

Illumina produced 1 million reads of 100 bp in length for a species whose genome size is 1 MB. **What is coverage?**

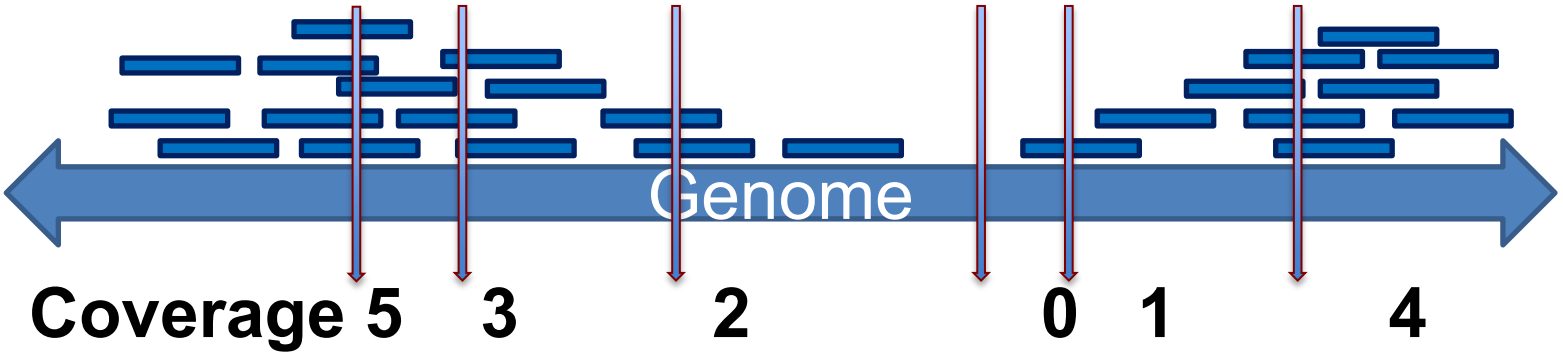
C = NL/G

1000000x100/1000000

= **100 X**

Genome Assembly Planning

Uneven Coverage



Genome Assembly Planning

Effect of genome size

- This will help decide which sequencing technology to use and what computational resources will be needed

Genome Assembly Planning

Conclusions

- Genome size is an important factor in Genome Sequencing.
- This will help decide which sequencing technology and computational resources will be needed

Effect of Repeats

Factors

Following factors may be important in planning an Assembly project;

- Size of Genome
- **Repeat content**
- Related species Assembled
- Strategy for performing Assembly

Effect of Repeats

Normal Sequence

Incorrect Assembly with an Orphan contig (Red)

Salzberg S L , and Yorke J A
Bioinformatics 2005;21:4320-4321

Effect of Repeats

Repeat content

A major proportion of eukaryotic genome is made up of repeats.

Repeats are the most common source of Mis-assemblies

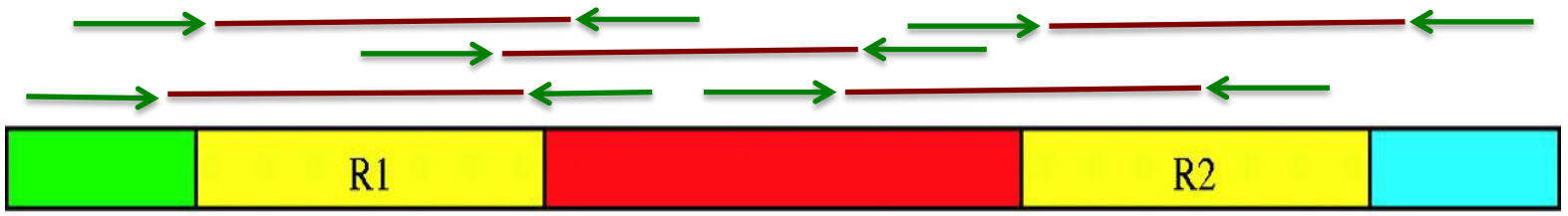
Effect of Repeats

Repeat solution

If sequencing technology produces reads longer than repeat size, impact is much smaller but that's not mostly the choice

- Most common solution is to generate mate pairs with spacing greater than largest known repeat

Effect of Repeats



Salzberg S L , and Yorke J A
Bioinformatics 2005;21:4320-4321

Effect of Repeats

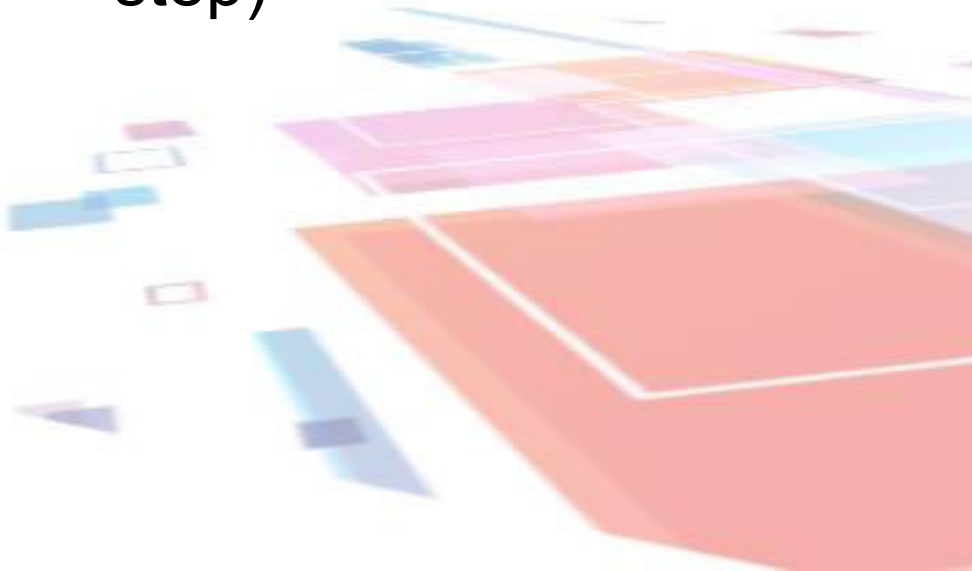
Double-barreled Shotgun Assembly

- DNA is randomly sheared into fragments (inserts)
- The ends of each insert are sequenced, resulting in two reads per fragment (double-barreled sequencing)

Effect of Repeats

Double-barreled Shotgun Assembly

- The original DNA sequence is reconstructed from the reads (assembly step)

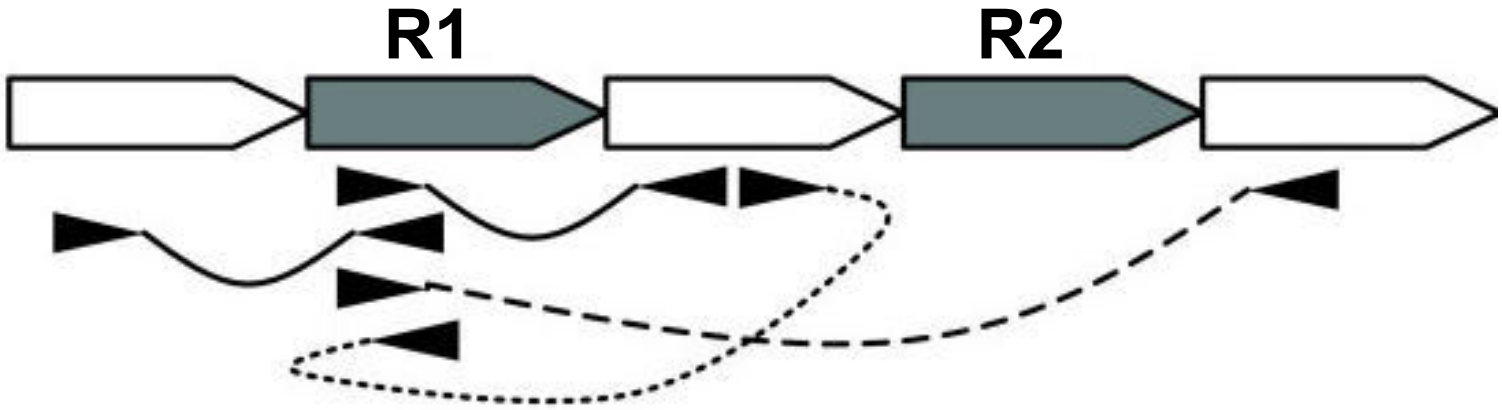


Effect of Repeats

Advantage of Double-barreled Shotgun Assembly

- It is unlikely that both reads (mate pairs) are coming from a repeat
- A read in a unique portion of DNA can help find the true repeat where it's mate is residing

Effect of Repeats



Effect of Repeats

Conclusions

- Most of the eukaryotic genomes possess repeats
- Repeats pose special challenges to assembly
- Using mate pair libraries can help fix these issues

Related Species Assembled

Factors

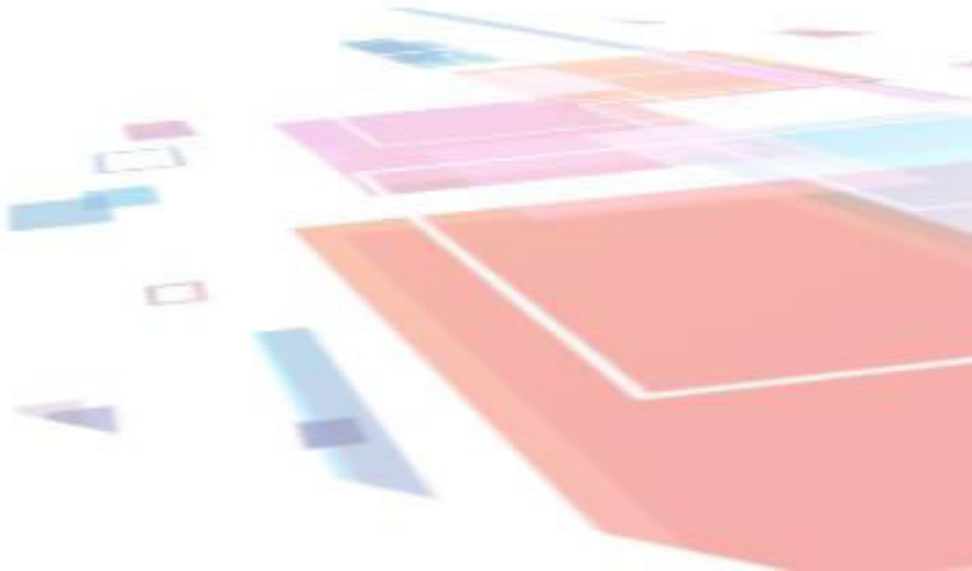
Following factors may be important in planning an Assembly project;

- Size of Genome
- Repeat content
- **Related species Assembled**
- Strategies for performing Assembly

Related Species Assembled

Importance

Availability of a related specie’s good quality genome having large reliable scaffolds is always a great help



Related Species Assembled

Importance

Related species genomes helps in;

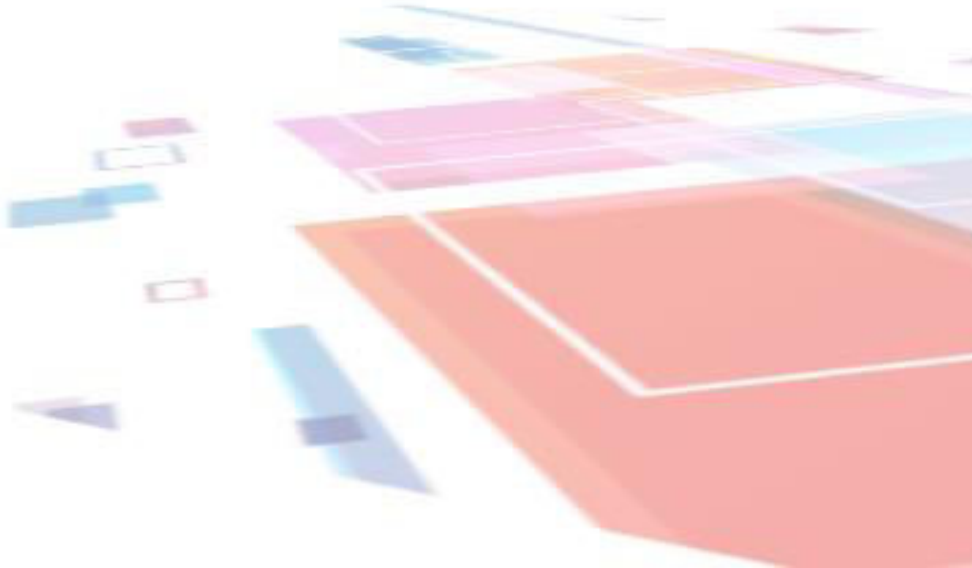
- Guiding the assembly of the target species
- Verifying the completeness of the assembly

Related Species Assembled

Importance

Related species genomes helps in;

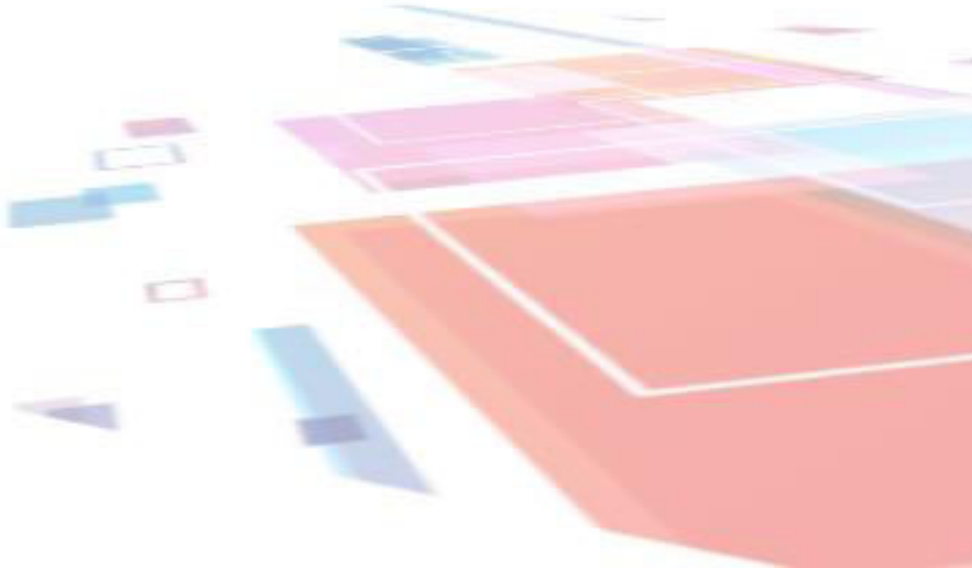
- Can themselves be improved in some cases



Related Species Assembled

Caution

- Can cause errors while architectures are different



Strategies for performing Assembly

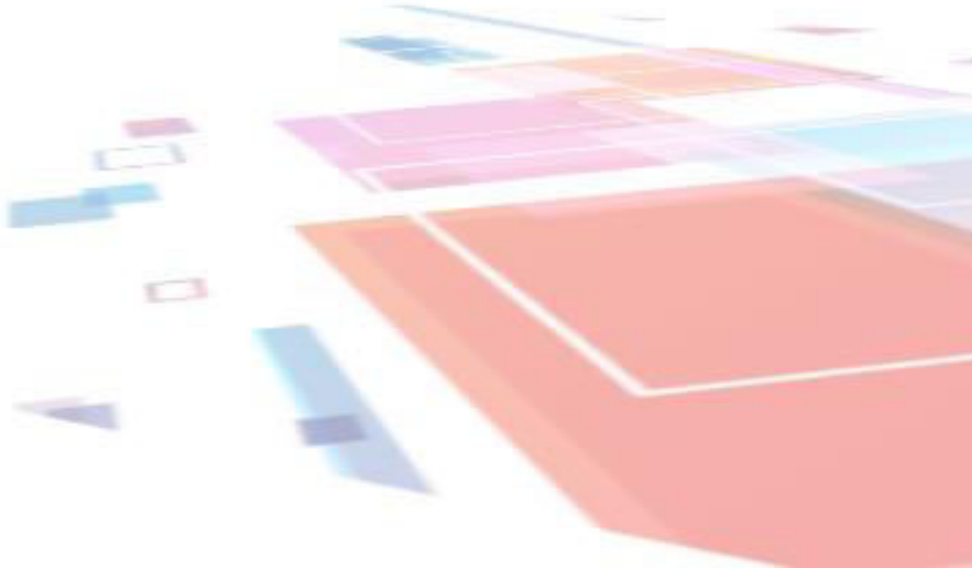
Factors

Following factors may be important in planning an Assembly project;

- Size of Genome
- Repeat content
- Related species Assembled
- **Strategies for performing Assembly**

Strategies for performing Assembly

- The sequencing approaches and assembly strategies are interdependent



Strategies for performing Assembly

	Library	Sequencing	Assembler
Bacterial Genome	Shotgun or mate-pair	>500nt **454** reads at 25x Or **Pacbio**	Newbler Celera or PBJelly
Vertebrate Genome	Paired end reads	100 nt Illumina reads (100x)	ALLPATHS-LG

Strategies for performing Assembly

Conclusions

- Availability of a related specie's good quality genome is always a great help in assembly
- The sequencing approaches and assembly strategies are interdependent

Greedy Graph Algorithm

Outline

- Introduction
- Steps
- Advantages
- Limitations
- Greedy Graph based Assemblers
- Conclusion

Greedy Graph Algorithm

Introduction

- Greedy Graph algorithms represent the simplest, most intuitive, solution to the assembly problem.

Greedy Graph Algorithm

Steps

Greedy Graph algorithms works as follows:

- Compare all reads or contigs in a pairwise fashion to identify overlapping sequences

Greedy Graph Algorithm

Steps

Greedy Graph algorithms works as follows:

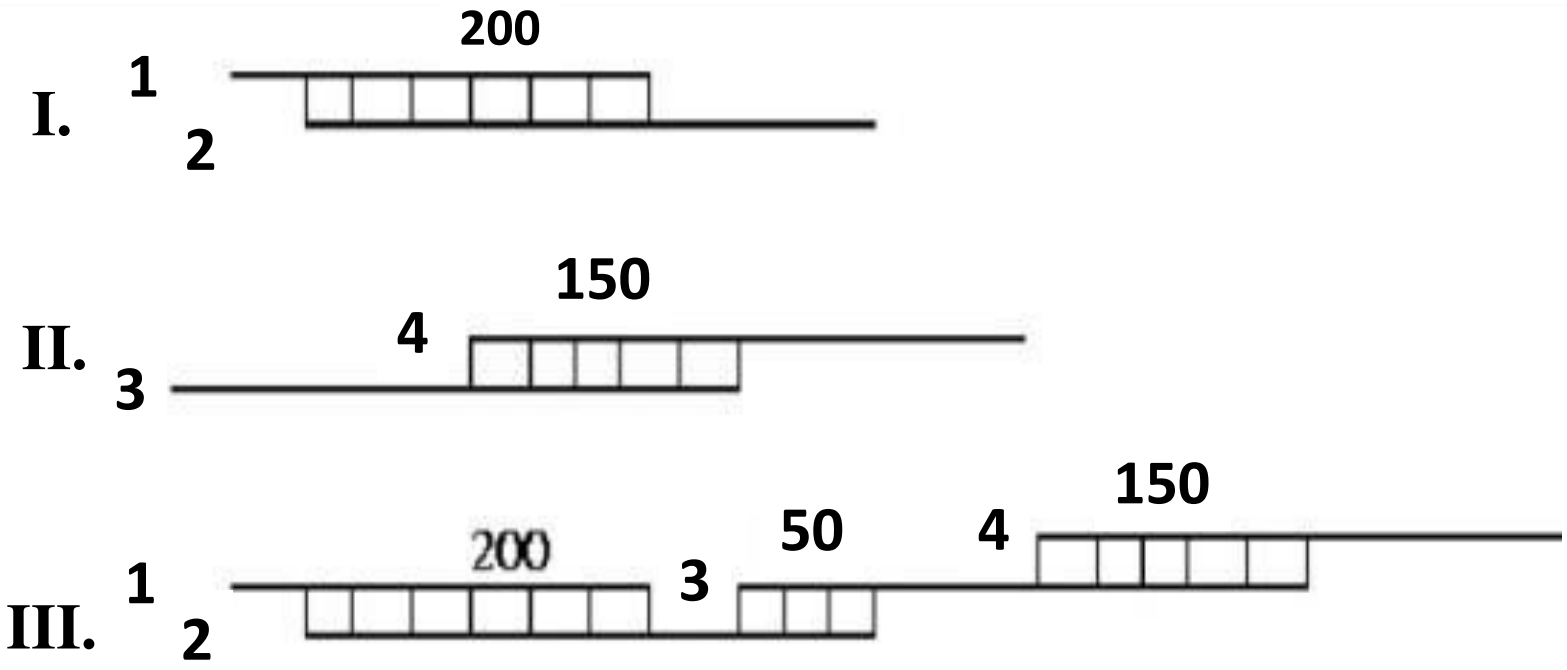
- Merge the sequences that overlap each other the best

Greedy Graph Algorithm

Steps

- Repeat step 2 until no more sequences can be merged, or the remaining overlaps conflict with existing contigs

Greedy Graph Algorithm



Greedy Graph Algorithm

- Best overlapping fragments are the one having the highest score
- The scoring function measures the number of matching bases in overlap

Greedy Graph Algorithm

Advantages:

- Suitable for small size genomes
- They drastically simplify the graph by considering only the high scoring edges
- May discard each overlap immediately after contig extension

Greedy Graph Algorithm

Limitations

- Greedy assemblers can detect false overlaps and high-scoring ones that are resulted from repetitive sequences.

Greedy Graph Algorithm

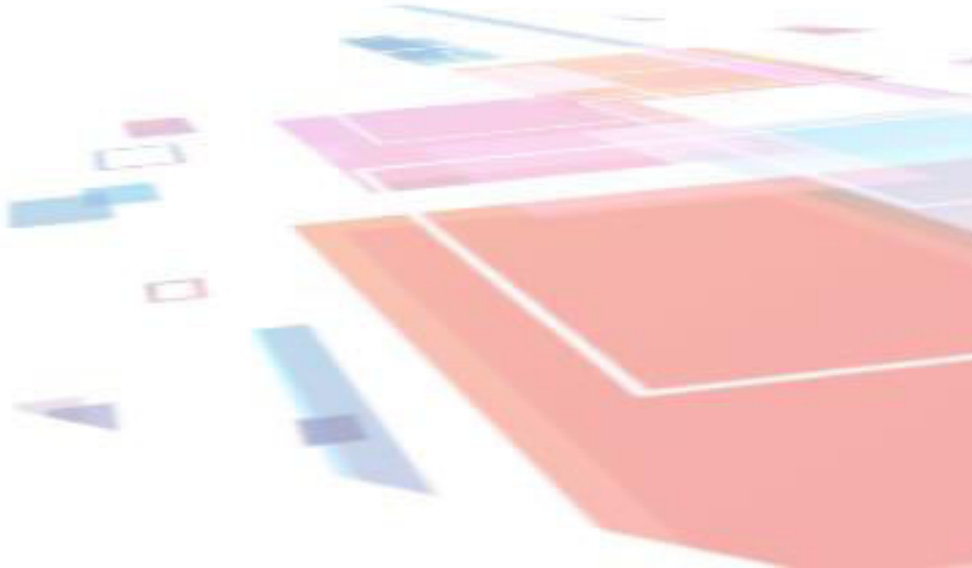
Limitations

- Graph traversal using greedy approach may cause algorithm to become stuck in local maxima, which produces a suboptimal solution for the assembly problem

Greedy Graph Algorithm

Limitations

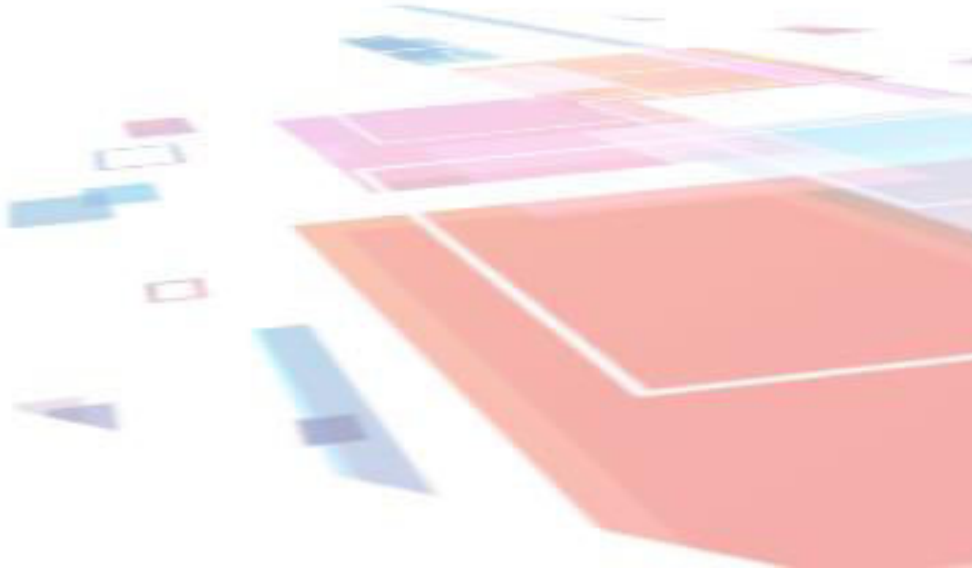
- The local maxima will increase the gaps between contigs in the assembly finishing process



Greedy Graph Algorithm

Greedy Graph based Assemblers:

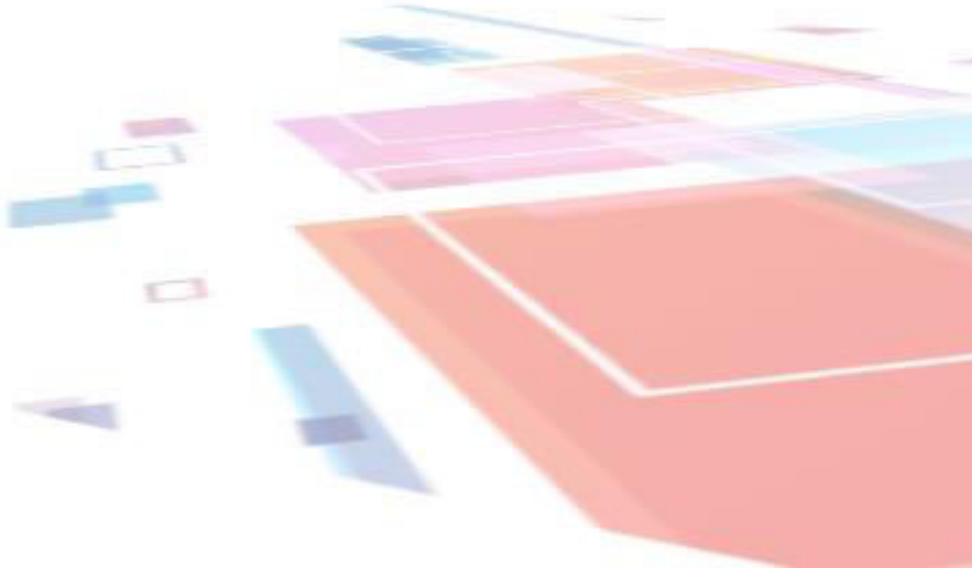
- SSAKE
- VCAKE
- SHARCGS



Greedy Graph Algorithm

Conclusion:

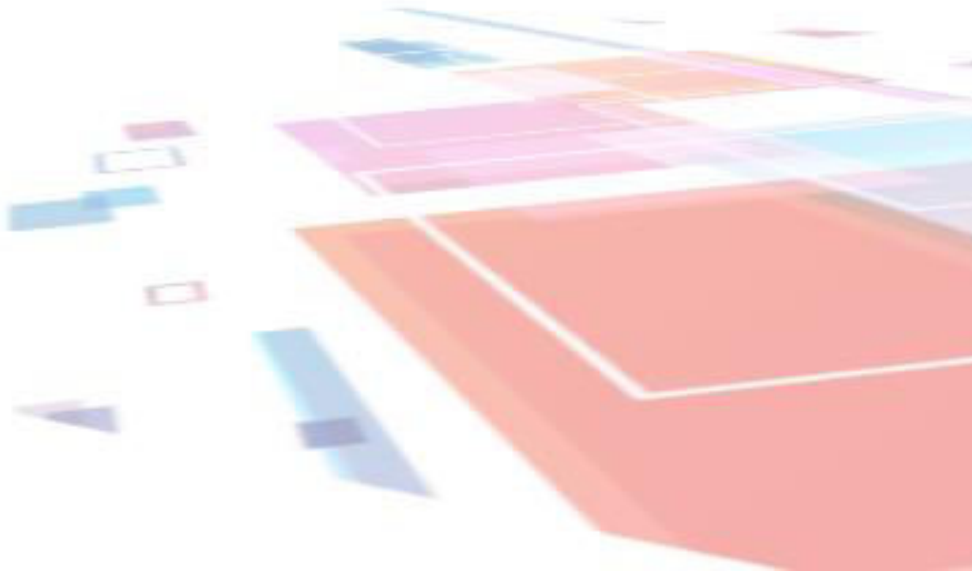
- Greedy graph approach is simple, easy and fast



Greedy Graph Algorithm

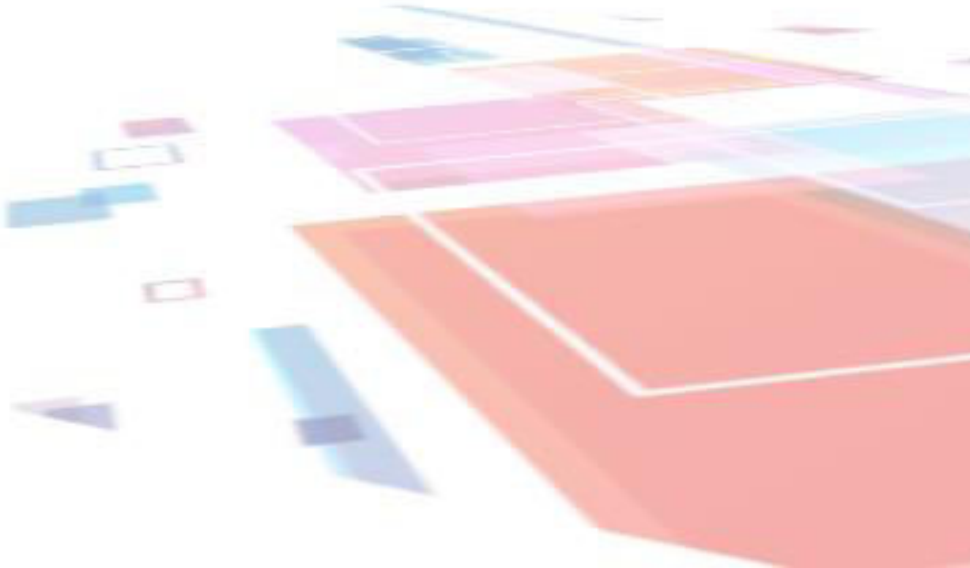
References:

- https://extras.csc.fi/biosciences/courses/ngs/denovo/denovo.pdf



Acknowledgements

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA



ALLPATHS-LG

Genome Assemblers

Variants of simple sequence alignment programs to piece together vast quantities of fragments generated by the sequencers

ALLPATHS-LG

Introduction

It’s a short read genome assembler from the Computational Research and Development group at the Broad Institute

ALLPATHS-LG

Goal

- High quality genome assembly from low cost data
- It can generate high quality genome assemblies using short reads (~100bp) by the new generation of sequencers

ALLPATHS-LG

Difference between ALLPATHS-LG and Traditional Assemblers

- ALLPATHS-LG assemblies are not necessarily linear, but instead are presented in the form of a graph

ALLPATHS-LG

Linearized Graph Assemblies

- An assembly consisting of contigs and scaffolds with embedded ambiguity codes

ALLPATHS-LG

ALLPATHS-LG Sequencing Model

Libraries, insert types*	**Fragment size, bp**	**Read length, bases**	**Sequence Coverage**
Fragment	180	≥100	45
Short jump	3,000	≥100 preferable	45
Long jump	6,000	≥100 preferable	5
Fosmid Jump	40,000	≥26	1

ALLPATHS-LG

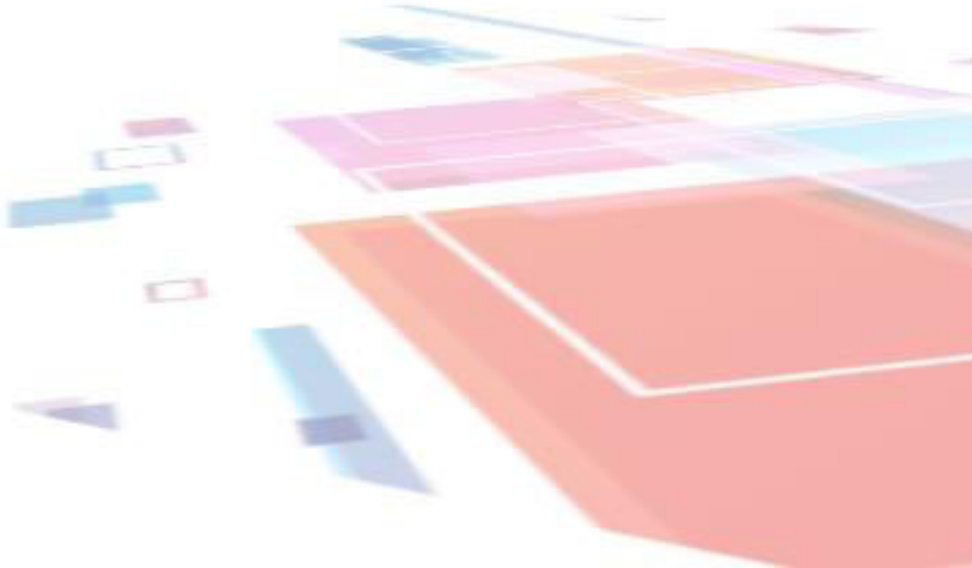
Computational Requirements

- 64 bit linux
- Memory requirements
  - About 160 bytes per genome base implying

ALLPATHS-LG

Advantages

- Relatively fast runtime
- Can use long reads only for small genomes



ALLPATHS-LG

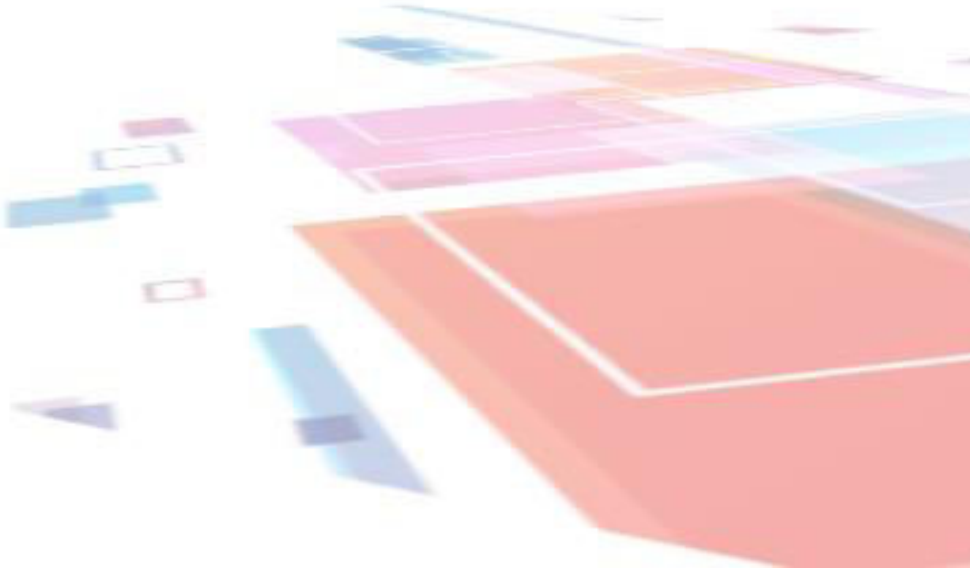
Conclusion

- It’s a short read genome assembler from the Computational Research and Development group at the Broad Institute

ALLPATHS-LG

References

- http://www.broadinstitute.org/software/allpaths-lg/blog/
- Assembly Tutorial by Michael Schatz



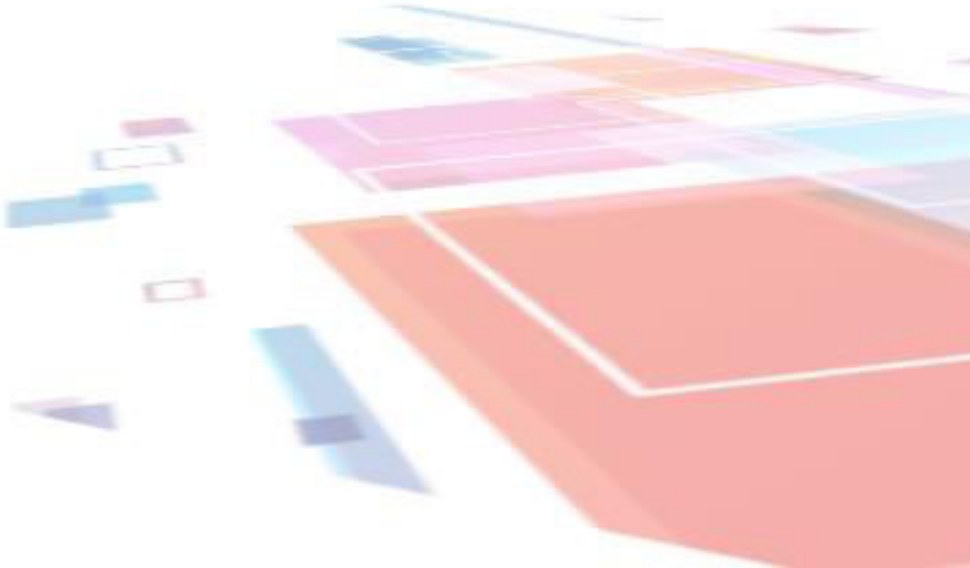
ALLPATHS-LG

References

- Gnerre, Sante, et al. "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." *Proceedings of the National Academy of Sciences* 108.4 (2011): 1513-1518.

Acknowledgements

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA



Lander Waterman Curve

Outline

- Introduction
- Assumptions
- Lander Waterman Model
- Variables used in calculation
- Calculations
- Results
- Conclusion

Lander Waterman Curve

Introduction

- Lander/Waterman suggested in their 1988 paper that the number of times a base is sequenced follows a Poisson distribution

Lander Waterman Curve

Assumptions

There are two key assumptions they made;

- Reads will be randomly distributed in the genome

Lander Waterman Curve

Assumptions

There are two key assumptions they made;

- The ability to detect an overlap between two truly overlapping reads does not vary from clone to clone

Lander Waterman Curve

Lander Waterman Model

- If overlap length was larger than a cutoff (T), then the two reads should be merged into a contig
- This process is iterated until no reads or contigs can be merge

Lander Waterman Curve

Variables used in Calculations

Variables used in calculation are;

G = haploid genome length in bp (base pairs)

L = sequenced read length in bp

Lander Waterman Curve

Variables used in Calculations

Variables used in calculation are;

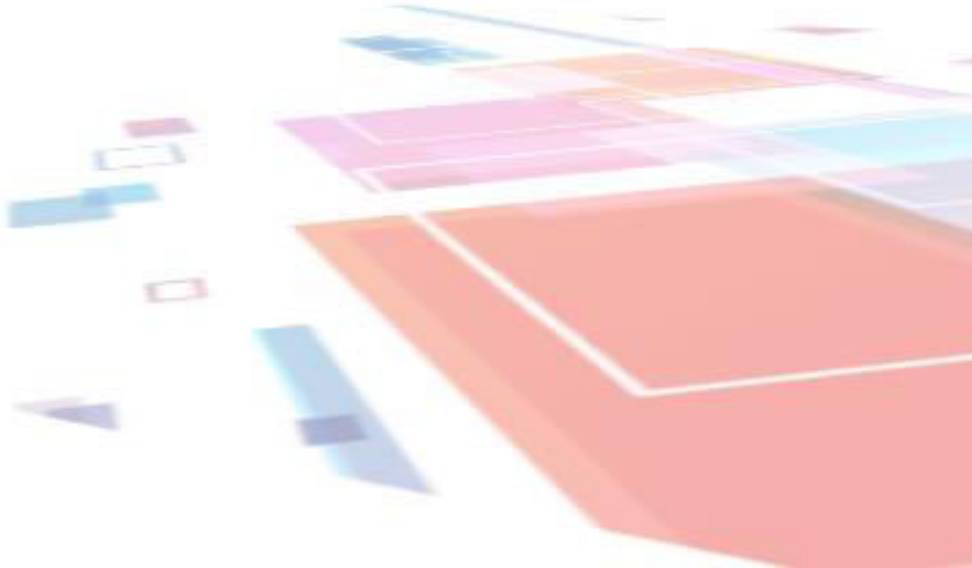
N = Number of reads sequenced

T = amount of overlap needed for detection in bp

Lander Waterman Curve

Calculations

Coverage can be defined as the average number of times any given base in the genome is sequenced



Lander Waterman Curve

Coverage

It can be derived by dividing the total length of acquired sequences by the genome length

$$\mathbf{C} = \frac{\boldsymbol{LN}}{\boldsymbol{G}}$$

Lander Waterman Curve

Gap Length

Let e^{-C} be the probability that any base is not sequenced

- **Total gap length**

  = % genome not sequenced * total genome in bp

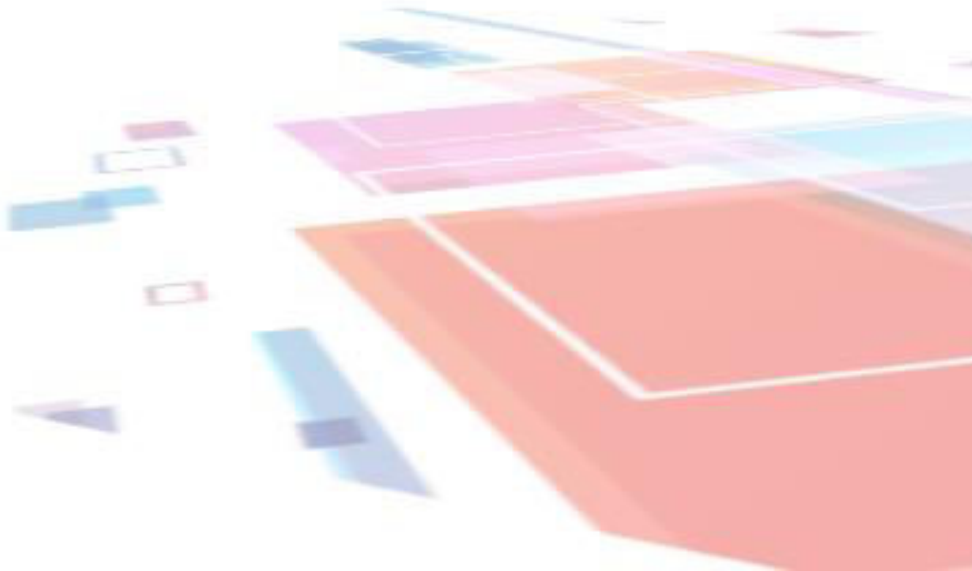
  = $e^{-C} * G$

Lander Waterman Curve

Number of Gaps

Number of gaps = % genome not sequenced) * # of reads sequenced

$$= N * e^{-C}$$



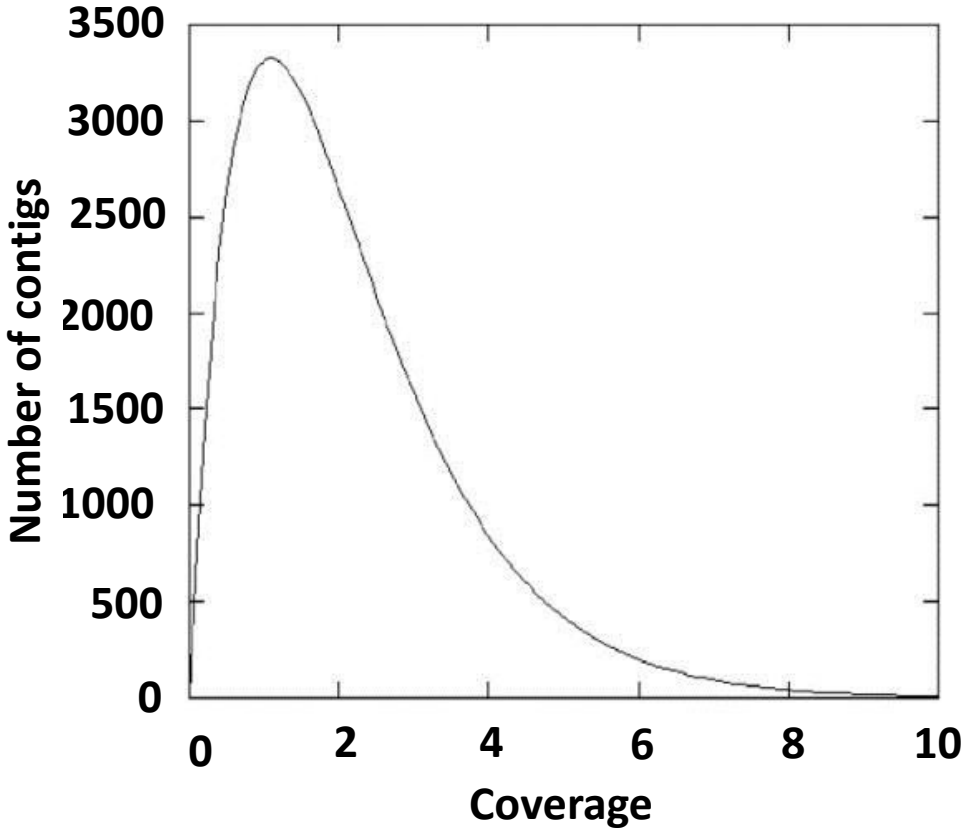
Lander Waterman Curve

Results

- Resulting contig number can be calculated.
- Calculate the sequencing reads needed to cover a genome.

Lander Waterman Curve

Lander-Waterman estimation of number of contigs w.r.t. genome coverage



Lander Waterman Curve

Conclusion

- The original paper relates to clone fingerprinting for physical mapping, but the even so the theory is readily applicable to whole genome shotgun projects.

Lander Waterman Curve

References

- Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, *2*(3), 231-239.

Lander Waterman Curve

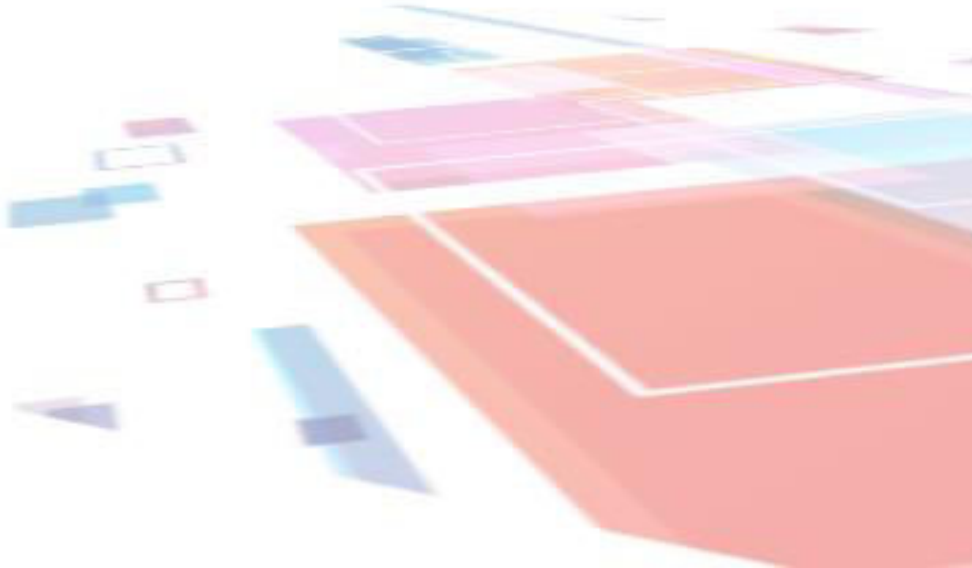
References

- Lander-Waterman Statistics for Shotgun Sequencing by Prof. Tesler
- Phylogenomics by Rob DeSalle and Jefferey A. Rosenfeld (Chapter 7: Genomic Sequencing and Annotation

Lander Waterman Curve

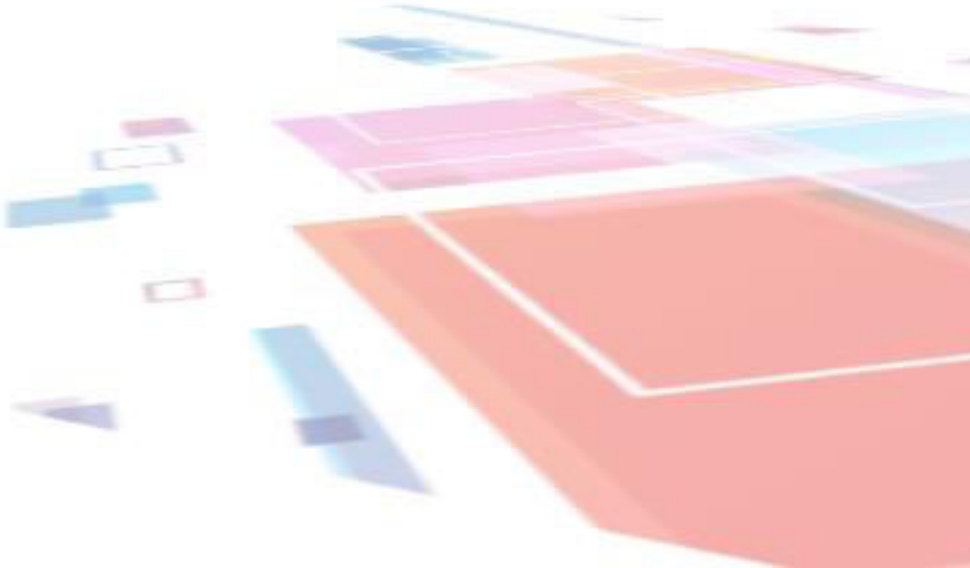
References

- http://www.cbcb.umd.edu/research/assembly_primer



Acknowledgements

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA



Assembly Pipeline

Strategy

Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- Repeat resolution

Assembly Pipeline

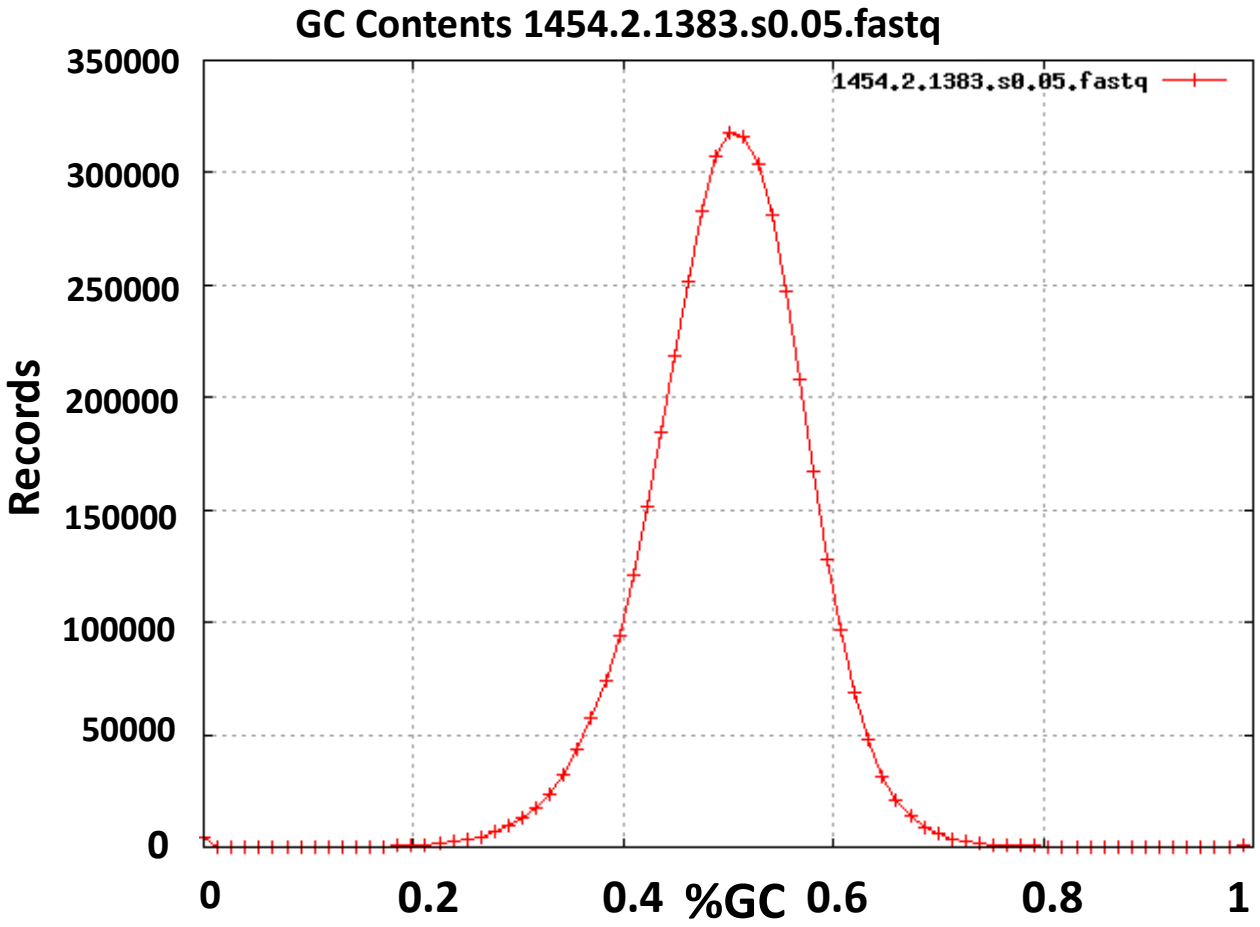
Quality screening

Screening is done while satisfying the following matrices;

- Genome properties (GC)
- Library quality
- Sequencing Run Quality
- Read Quality

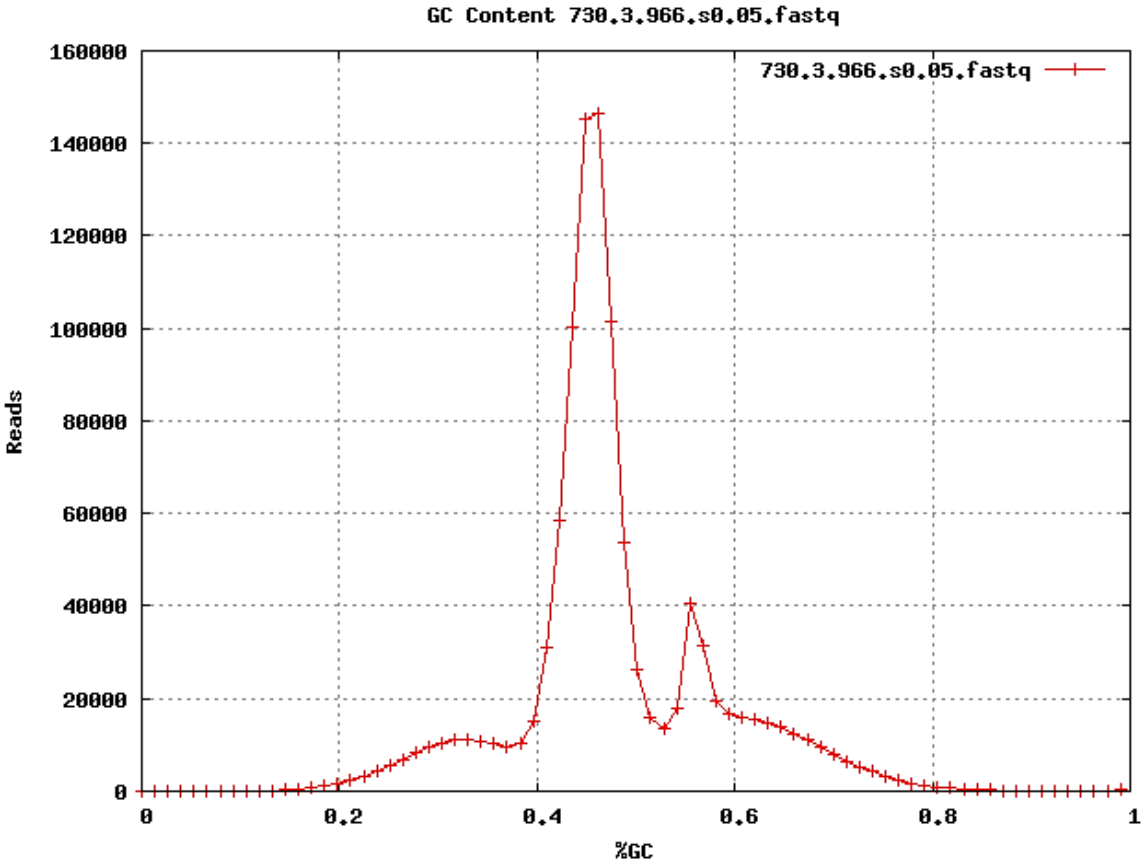
Assembly Pipeline

Genome Properties



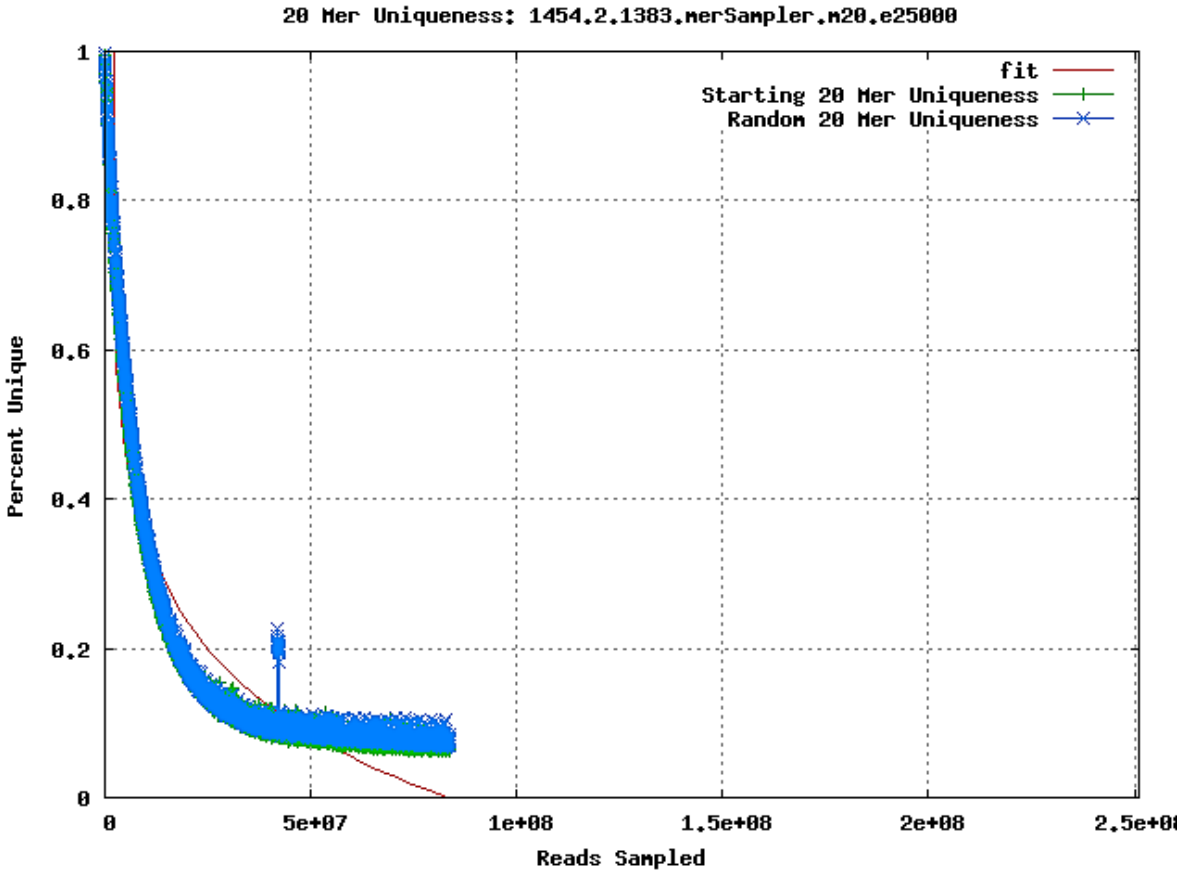
Assembly Pipeline

Genome Properties



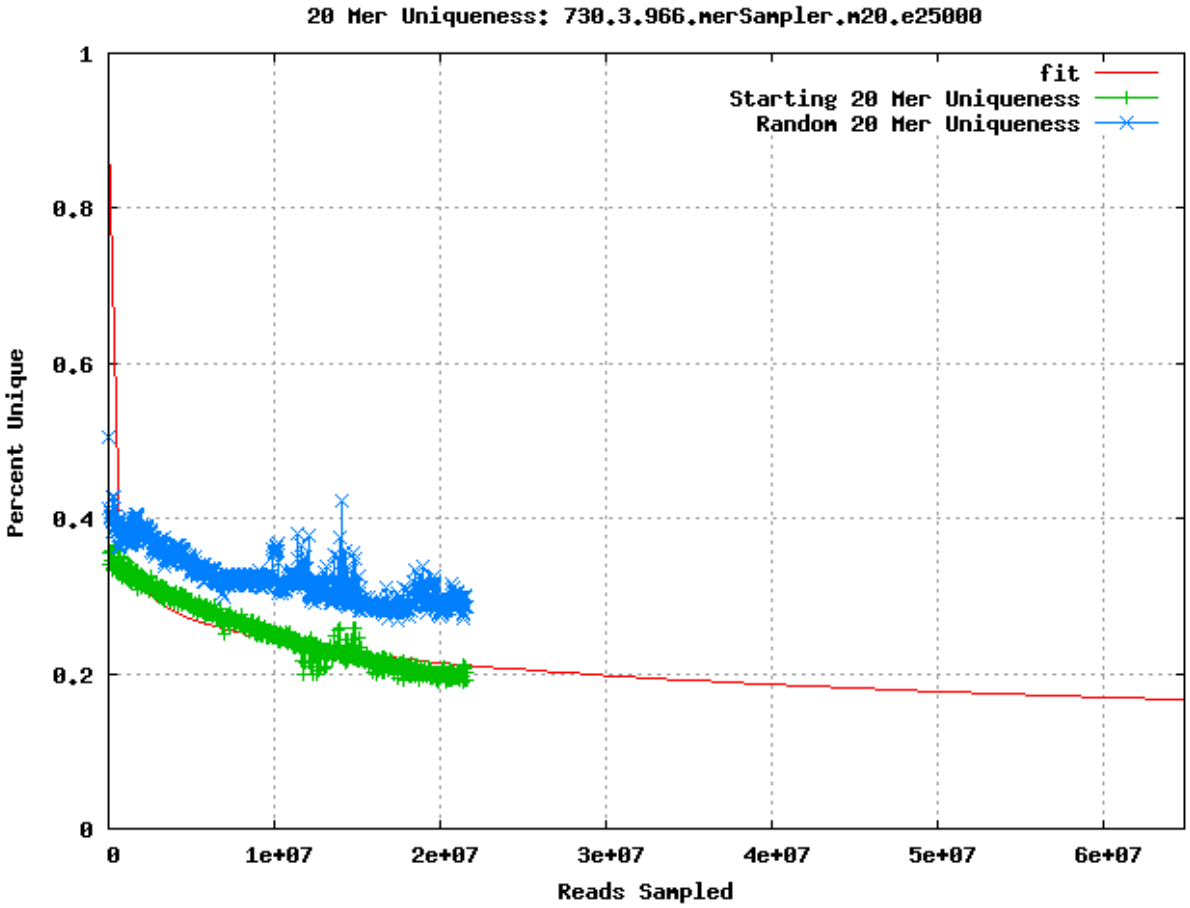
Assembly Pipeline

Library quality



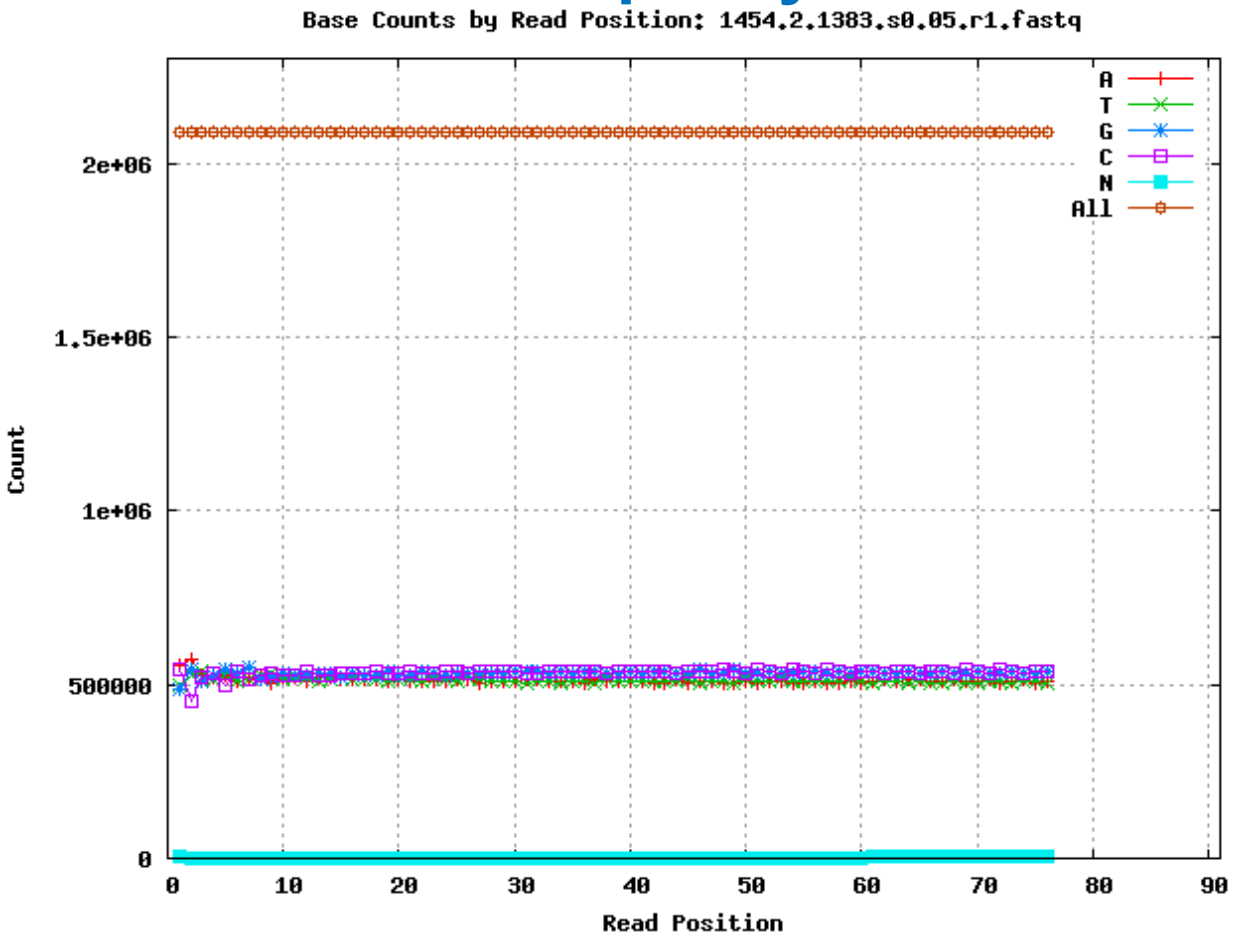
Assembly Pipeline

Library quality



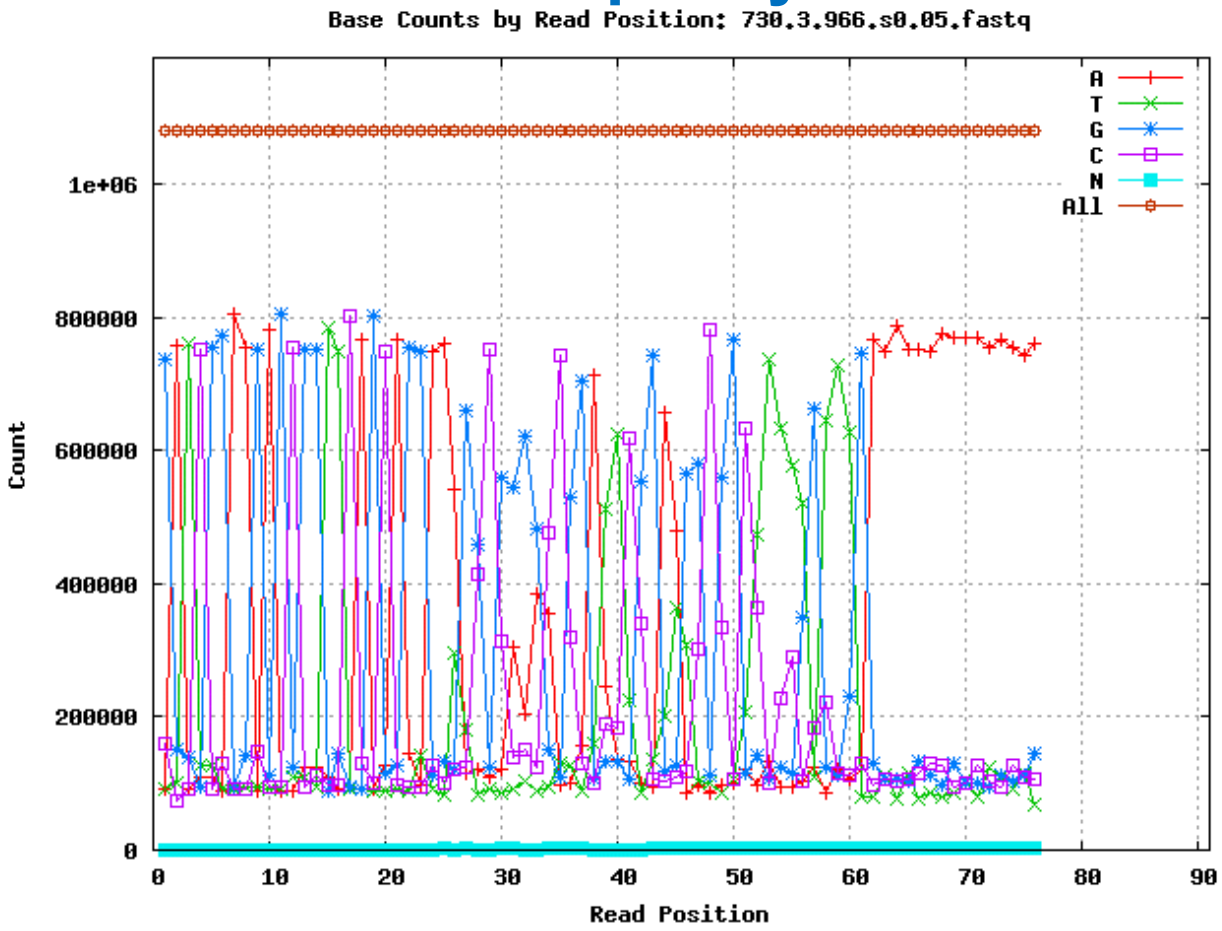
Assembly Pipeline

Run quality



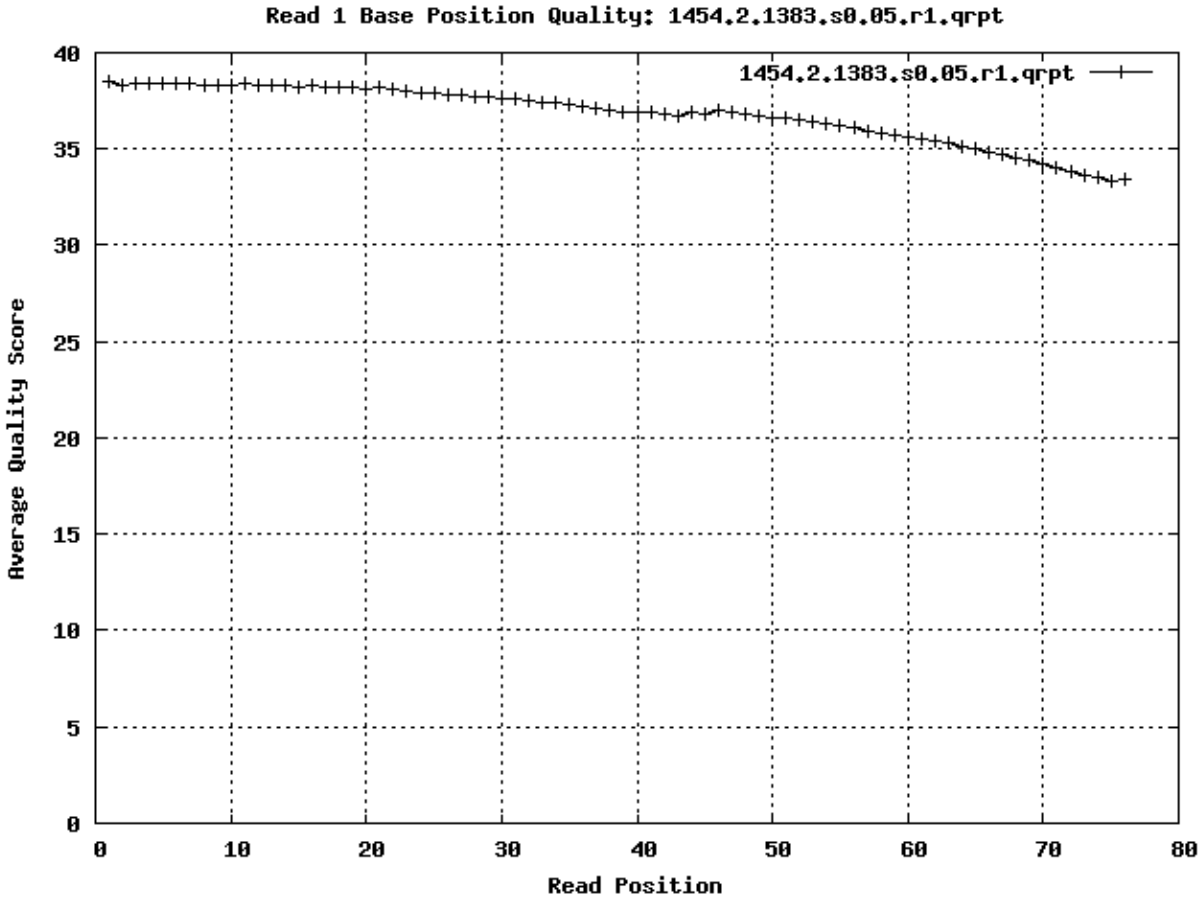
Assembly Pipeline

Run quality



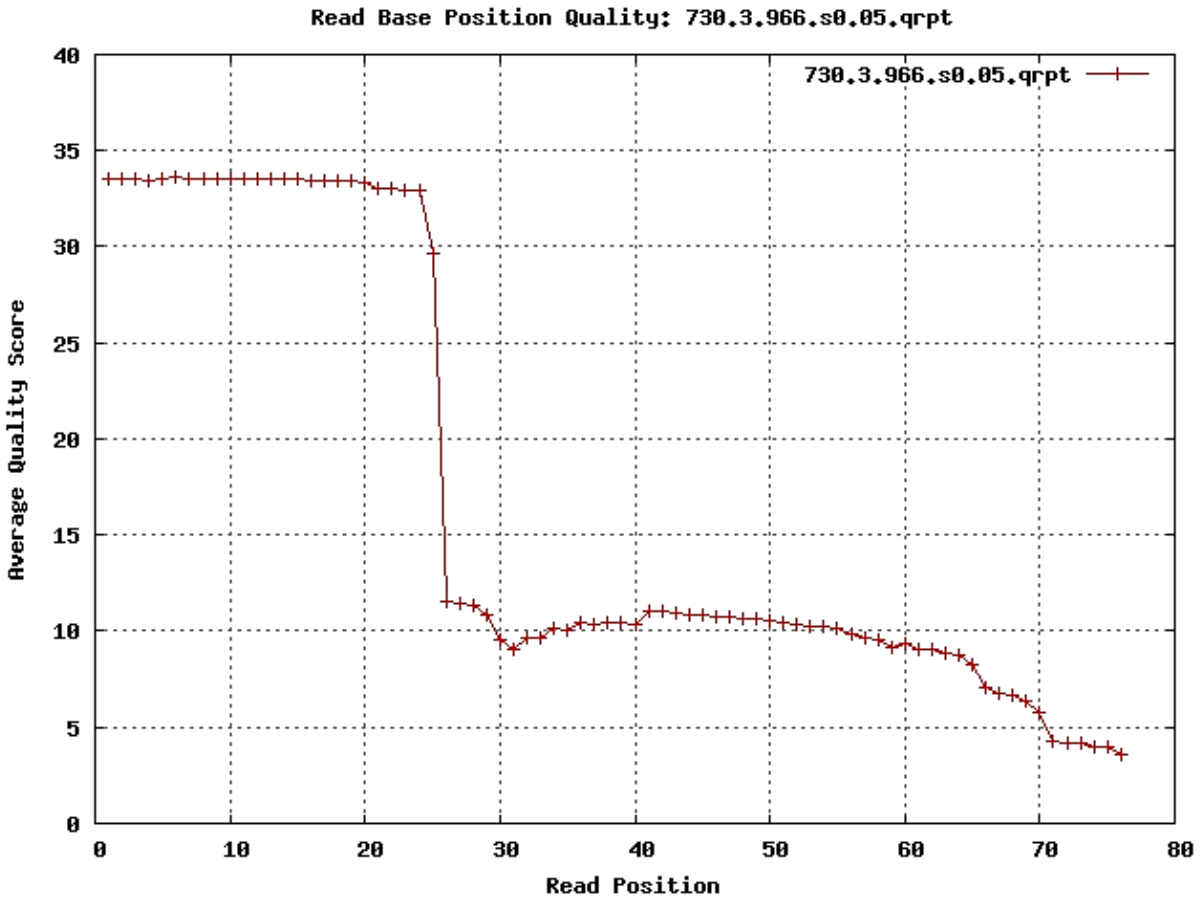
Assembly Pipeline

Read Quality



Assembly Pipeline

Read Quality



Acknowledgements

For this slide deck I “borrowed” figures and slides from many publications, Web pages and presentations by

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA
- Thank you!

Assembly Pipeline

Strategy

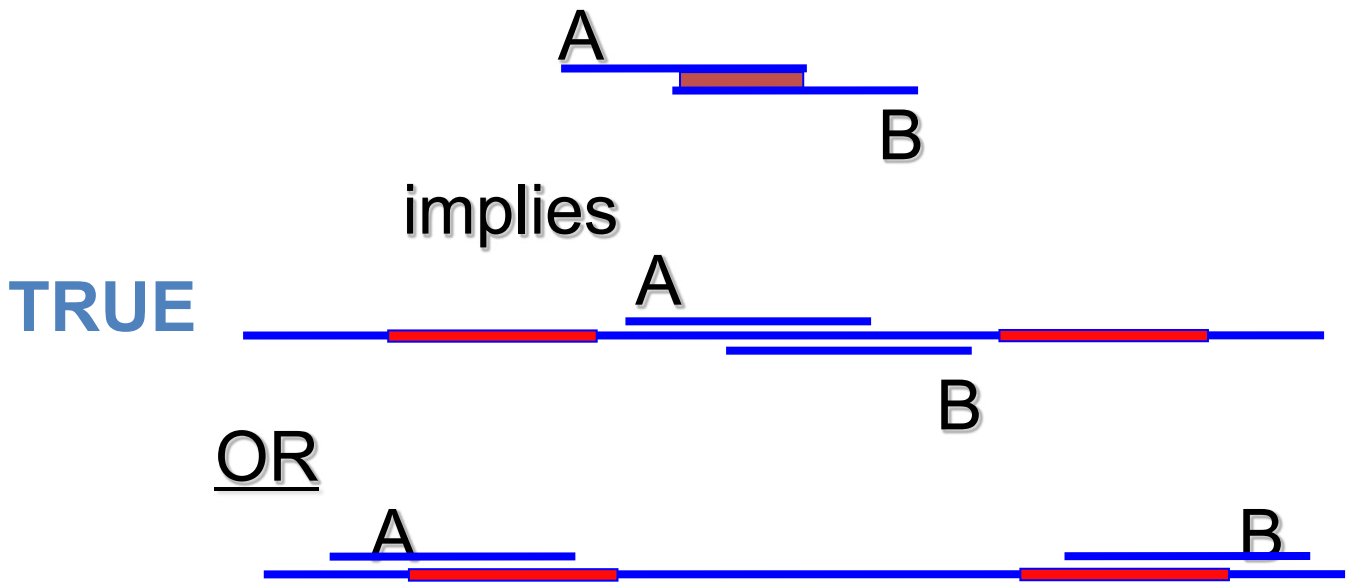
Overall process can be divided into following major phases;

- Quality Screening
- **Overlapping**
- Unitiging
- Scaffolding
- Repeat resolution

Assembly Pipeline

Overlapping

Find all overlaps ≥ 40bp allowing 6% mismatch.



REPEAT-INDUCED

Assembly Pipeline

Strategy

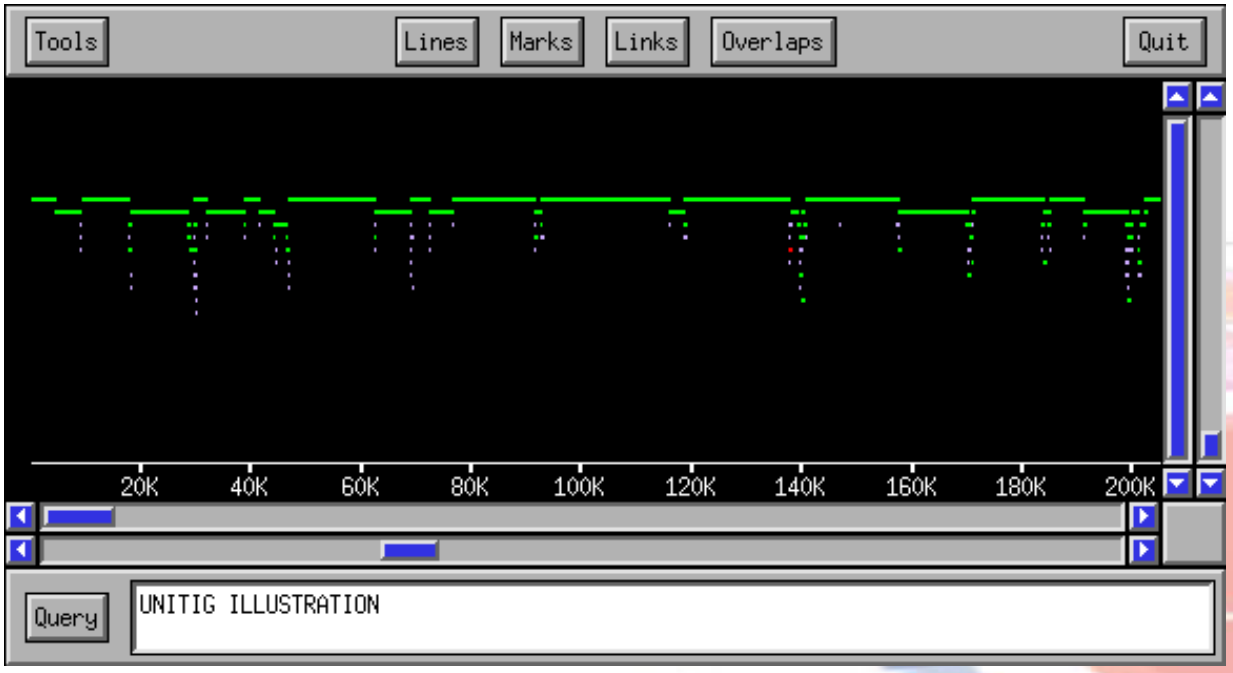
Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- **Unitiging**
- Scaffolding
- Repeat resolution

Assembly Pipeline

Compute all overlap consistent sub-assemblies

Unitigs (Uniquely Assembled Contig)



Assembly Pipeline

Unitigs (Uniquely Assembled Contig)

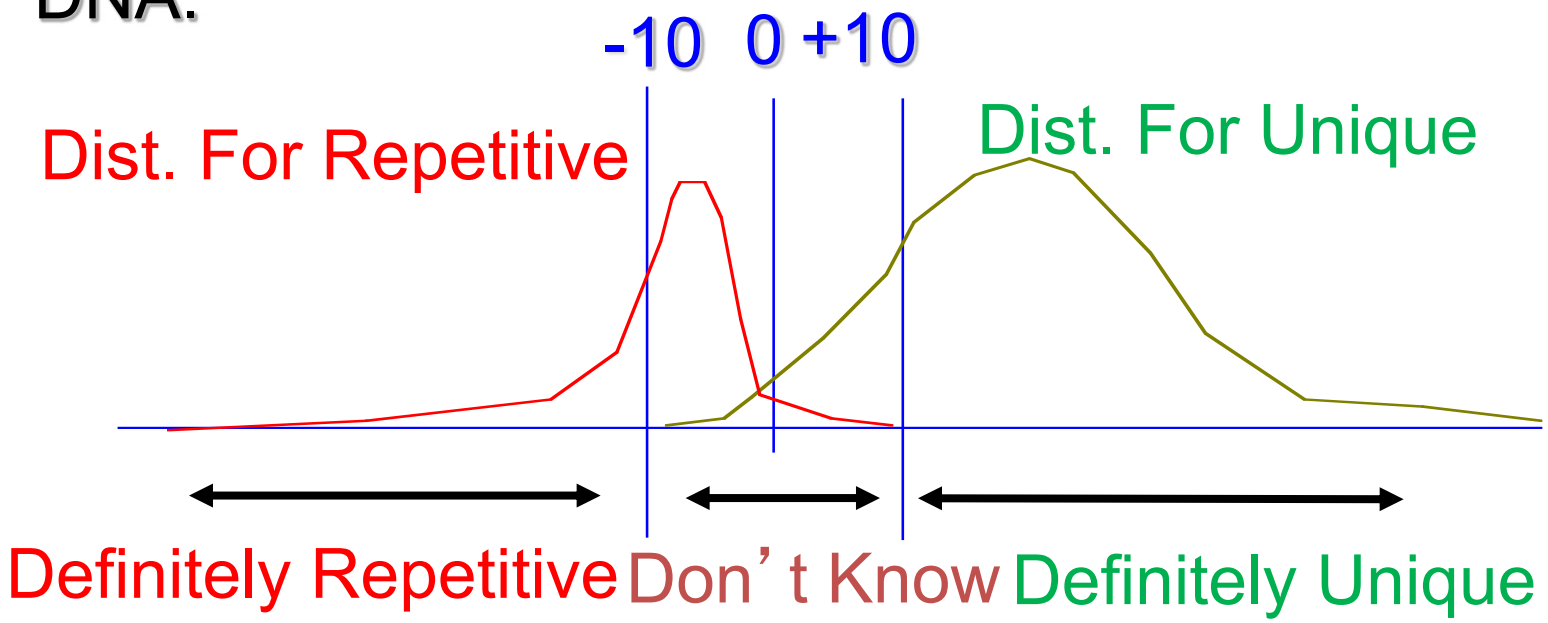
Unique DNA unitig

Repetitive DNA unitig

Assembly Pipeline

Unitigs (Uniquely Assembled Contig)

Discriminator Statistic is log-odds ratio of probability unitig is unique DNA versus 2-copy DNA.



Assembly Pipeline

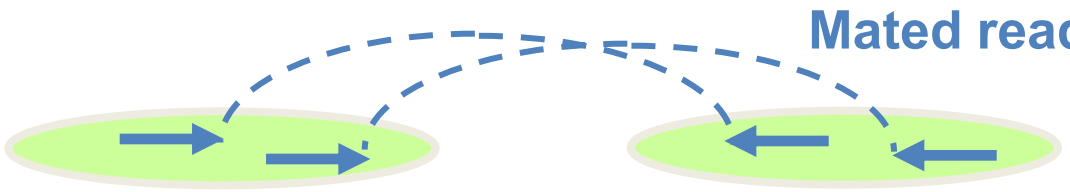
Strategy

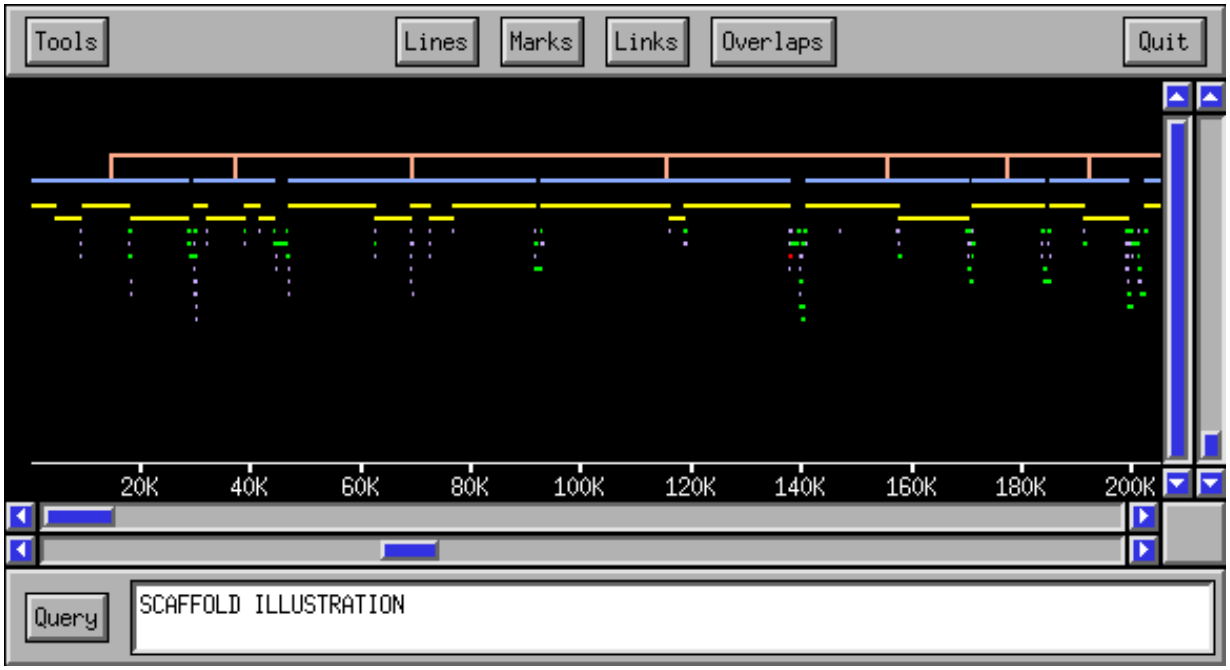
Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- Unitiging
- **Scaffolding**
- Repeat resolution

Assembly Pipeline

Scaffold U-unitigs with confirmed pairs





Assembly Pipeline

Strategy

Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- **Repeat resolution**

Assembly Pipeline

Repeat Resolution

pre-assembly

find fragments that belong to repeats

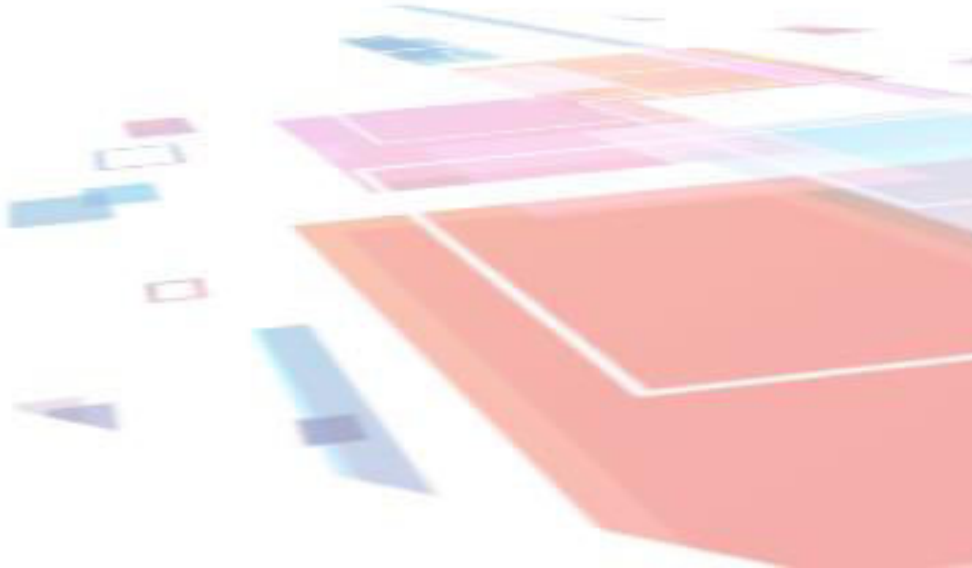
- statistically (most existing assemblers)
- repeat database (*RepeatMasker*)

Assembly Pipeline

Repeat Resolution

during assembly

- detect "tangles" indicative of repeats (Pevzner, Tang, Waterman 2001)



Assembly Pipeline

Repeat Resolution

post-assembly: find repetitive regions and potential mis-assemblies.

- *Reputer, RepeatMasker*
- "unhappy" mate-pairs (too close, too far, mis-oriented)

Assembly Pipeline

Conclusions

Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- Repeat resolution

Acknowledgements

For this slide deck I “borrowed” figures and slides from many publications, Web pages and presentations by

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA
- Thank you!

Quality of Assembled Genome

Factors

The qualities of genome assemblies are evaluated by;

- Percentage of sequences assembled
- Accuracy of contigs and scaffolds
- Repeat resolution
- Presence of expected genes

Quality of Assembled Genome

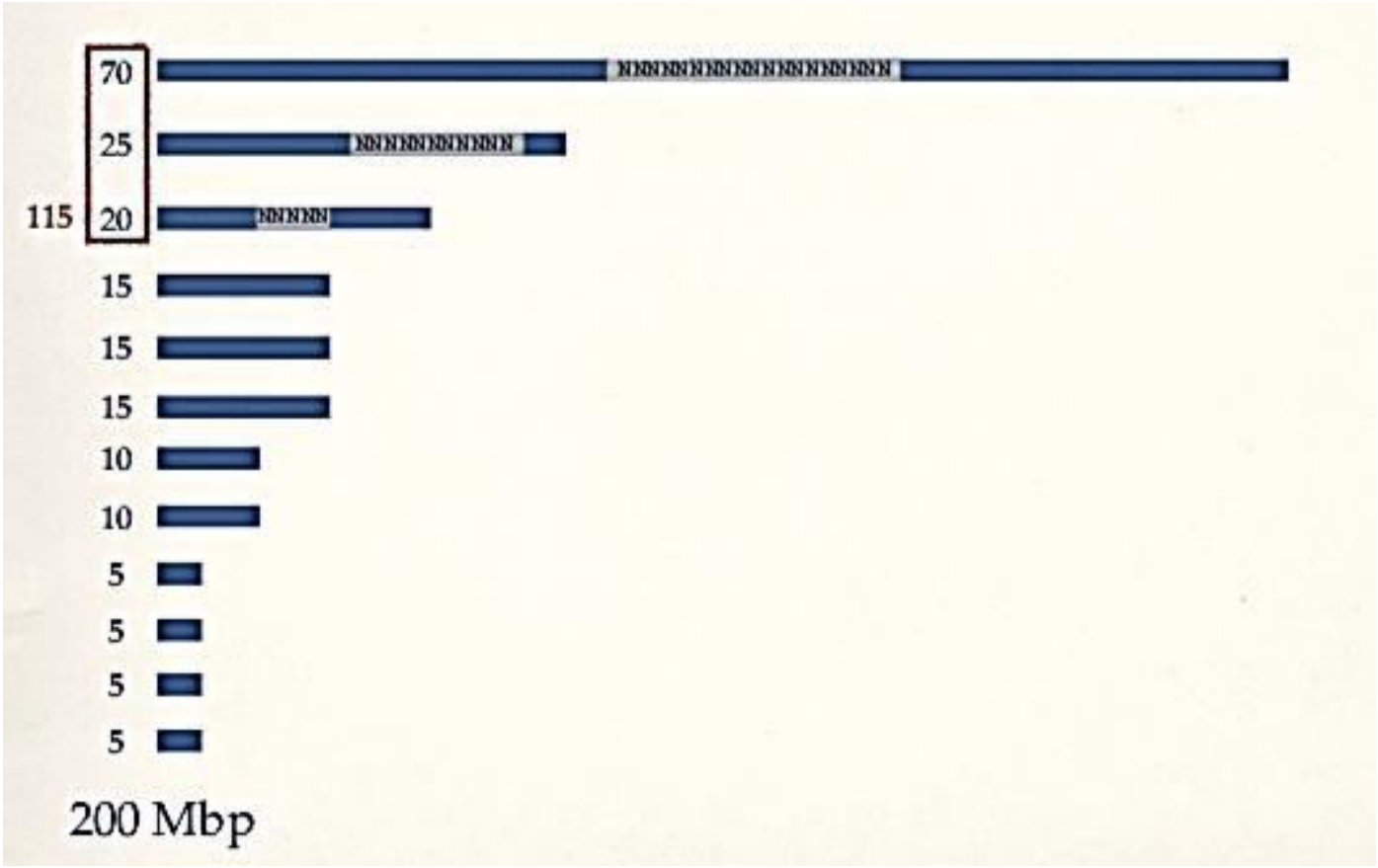
N50

Most common measure of assembly quality;

- N50 = length of the shortest contig in a set making up 50% of the assembly length

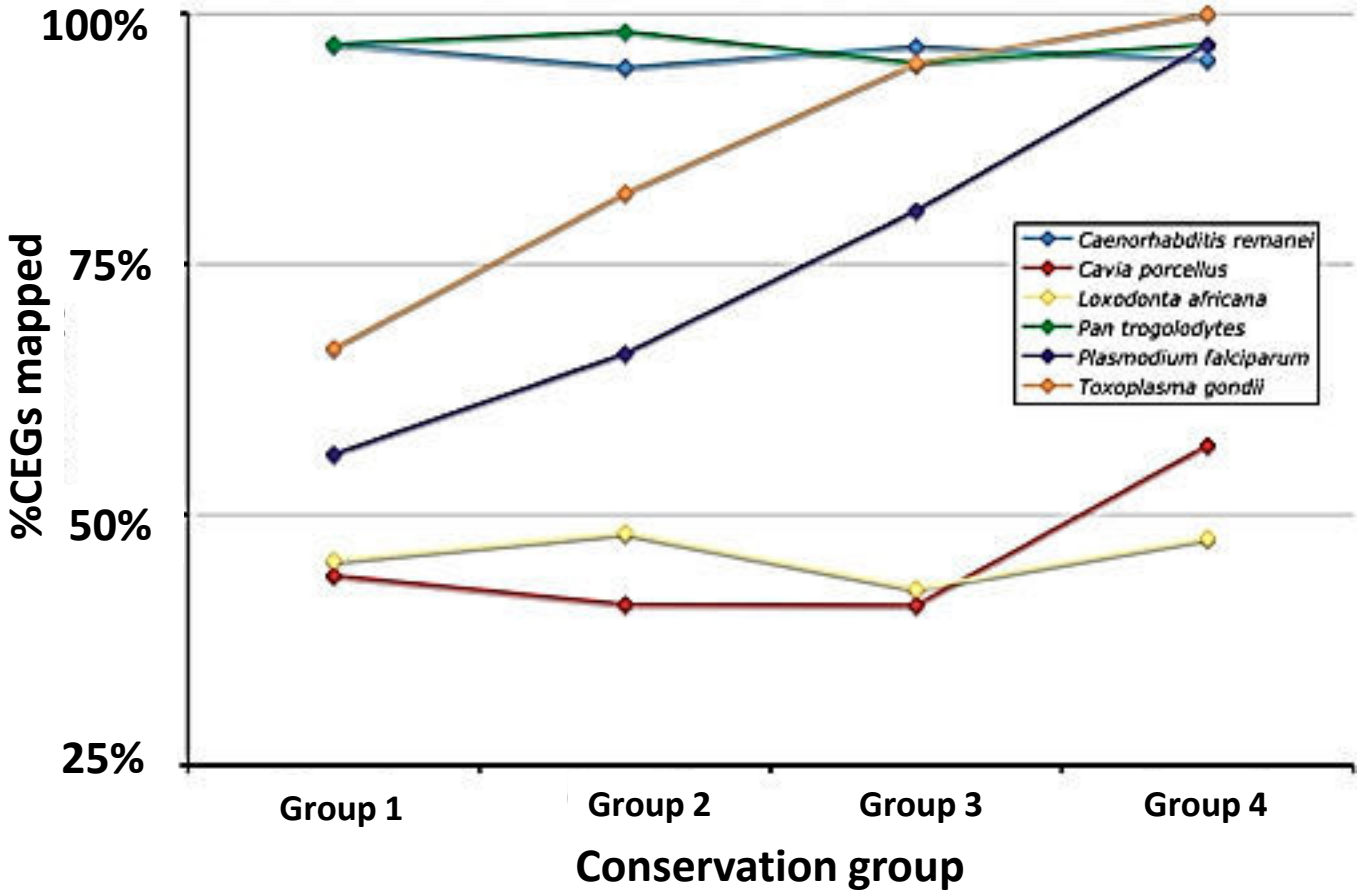
Quality of Assembled Genome

N50



Quality of Assembled Genome

Conserved eukaryotic gene sets



Quality of Assembled Genome

Best Genomes are not perfect

Caenorhabditis elegans

- Genome published 1998
- 2004: last N removed
- 1998–2013: genome sequence changes
  - 558 insertions
  - 230 deletions
  - 614 substitutions

Quality of Assembled Genome

Assembly tips

- Try different values of key parameters like k-mer size for DBG assemblers, and evaluate the output (some assemblers can do this automatically)

Quality of Assembled Genome

Assembly tips

- Try different subsets of data as libraries might be of poor quality
- Duplicates and homopolymers might also effect
- Try different assemblers as there is no ‘best assembler’ **(CAGE)**

Quality of Assembled Genome

CAGE (Genome Assembly Gold-standard Evaluation)

(U. of Maryland and Johns Hopkins)

- Select datasets associated with known high-quality genomes

Quality of Assembled Genome

CAGE (Genome Assembly Gold-standard Evaluation)

- Run a set of open source assemblers with parameter sweeps on these datasets

Quality of Assembled Genome

CAGE (Genome Assembly Gold-standard Evaluation)

- Compare the results, publish in scholarly Journals with complete documentation of parameters

Quality of Assembled Genome

CAGE (Genome Assembly Gold-standard Evaluation)

- Run a set of open source assemblers with parameter sweeps on these datasets

Quality of Assembled Genome

CAGE (Genome Assembly Gold-standard Evaluation)

- Compare the results, publish in scholarly Journals with complete documentation of parameters

Quality of Assembled Genome

Conclusions

The qualities of genome assemblies are evaluated by;

- Percentage of sequences assembled
- Accuracy of contigs and scaffolds
- Repeat resolution
- Presence of expected genes

Acknowledgements

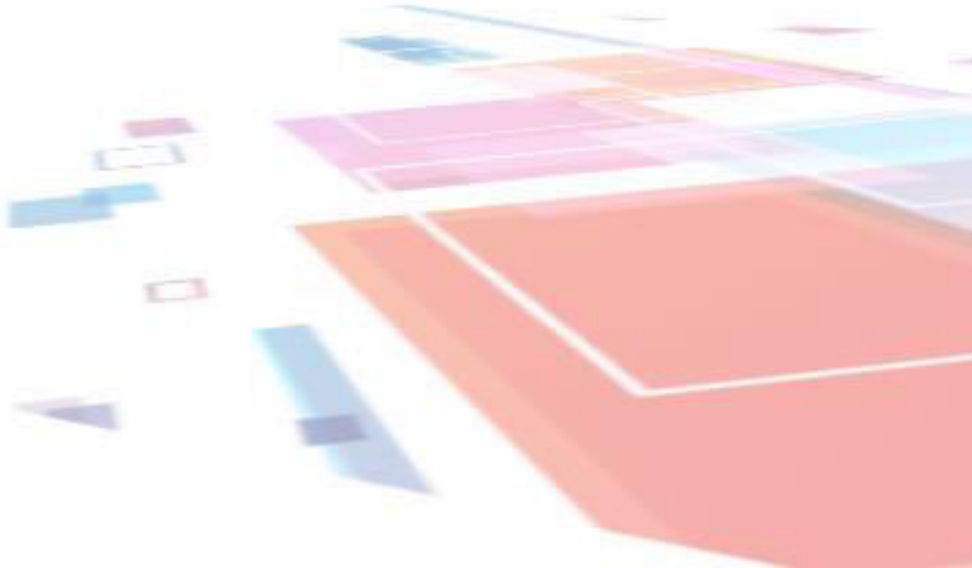
For this slide deck I “borrowed” figures and slides from many publications, Web pages and presentations by

- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA
- Thank you!

Graphs

Outline

- Introduction
- Types of graphs
- Graph Algorithms for Assembly
- Conclusion

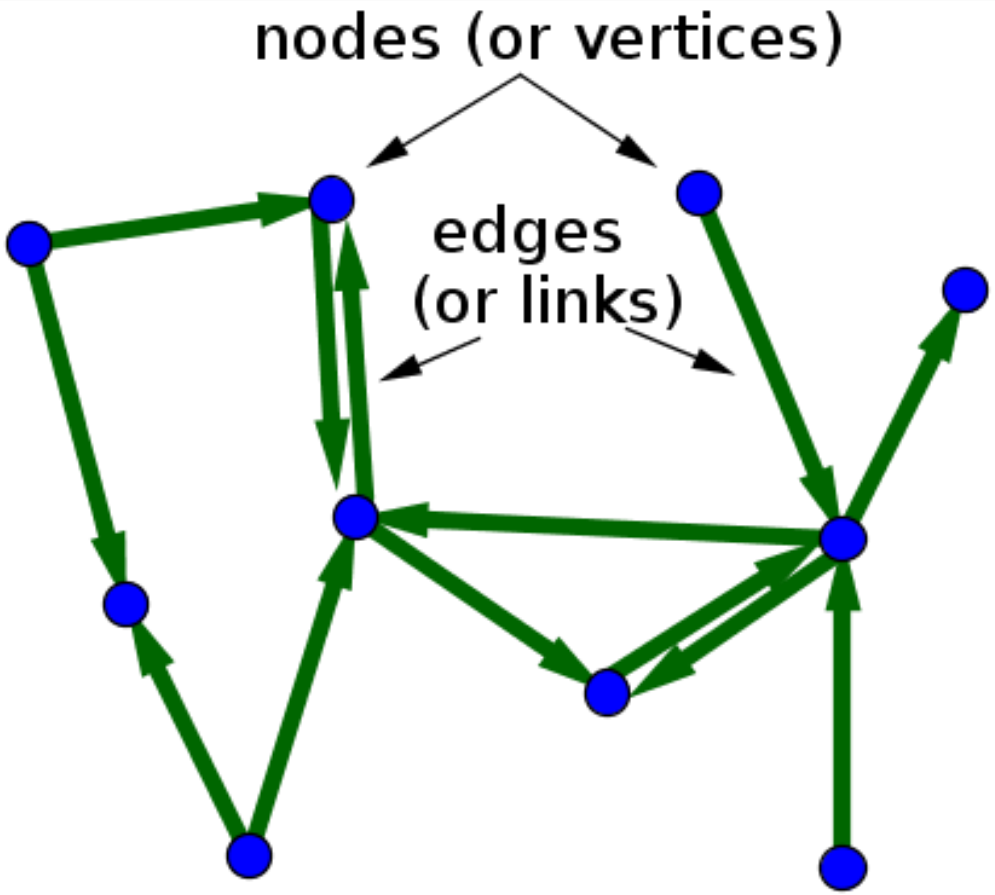


Graphs

Introduction

- Graph is a set of node plus set of edges between the nodes
- Nodes and edges may also be called vertices and arc, respectively

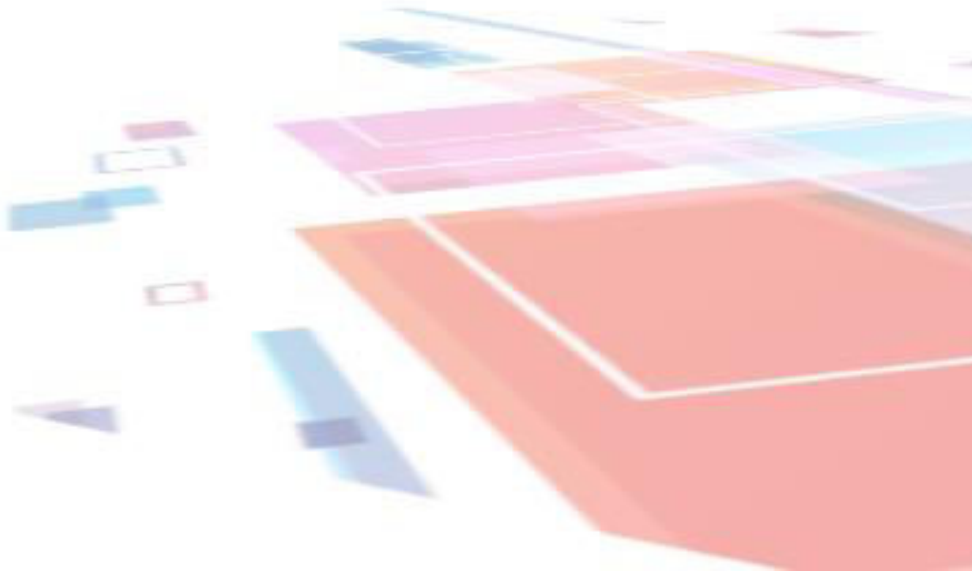
Graphs



Graphs

Types of Graphs

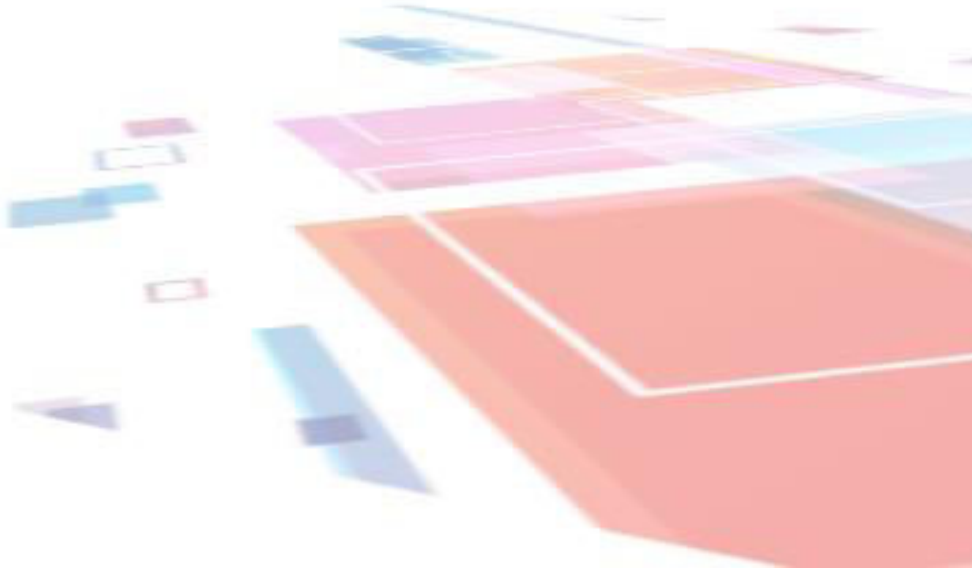
- Directed Graphs
- Undirected Graphs



Graphs

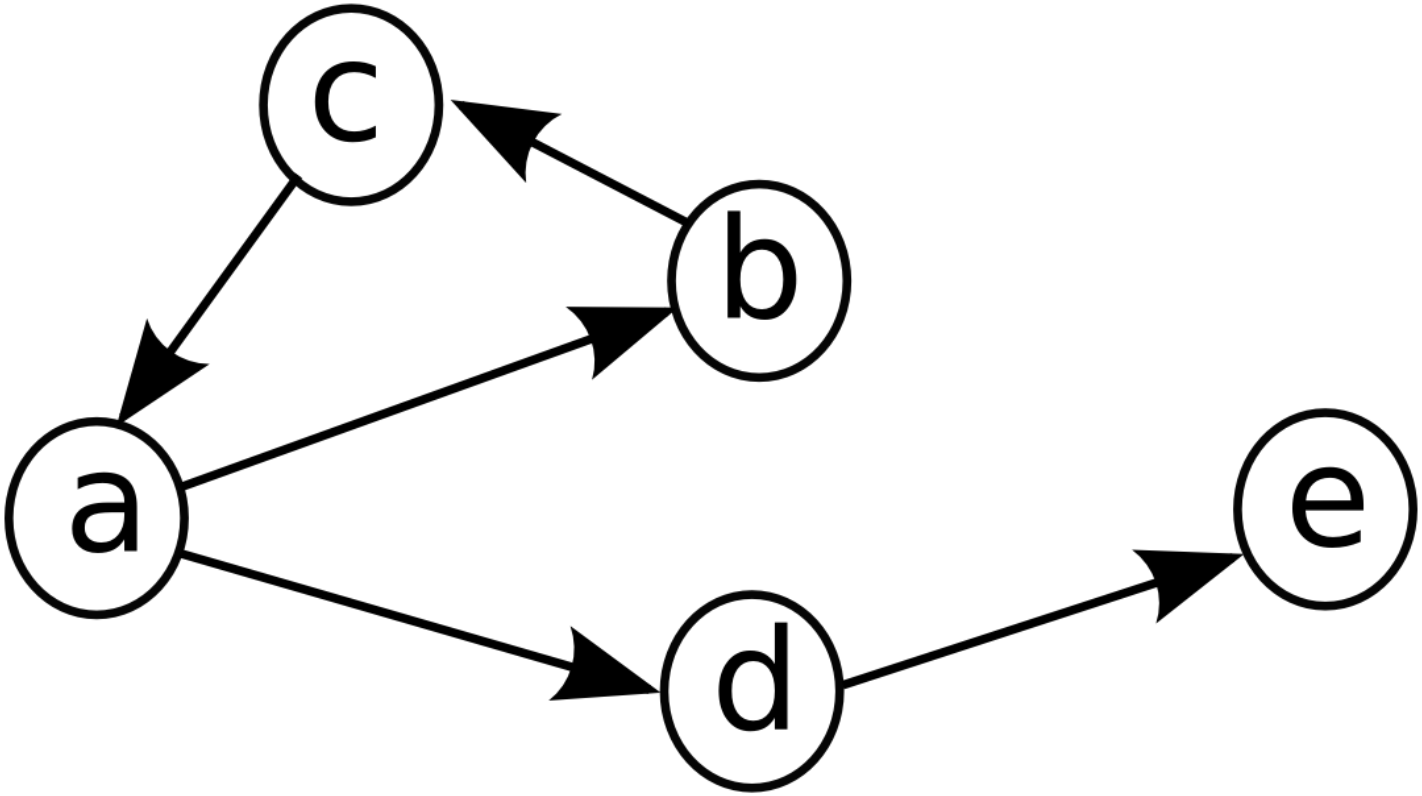
Directed Graphs

- If the edges may only be traversed in one direction, the graph is known as directed graph



Graphs

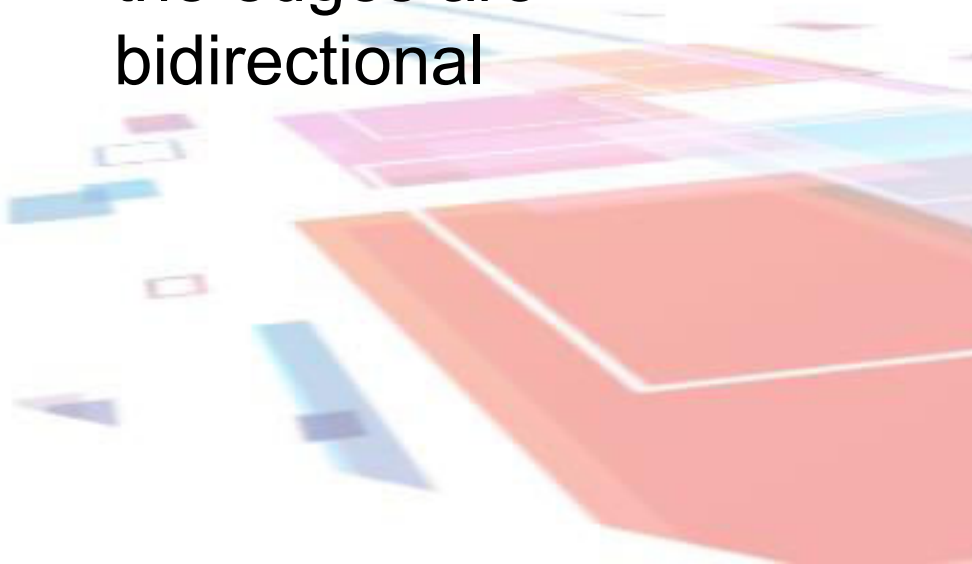
Directed Graphs



Graphs

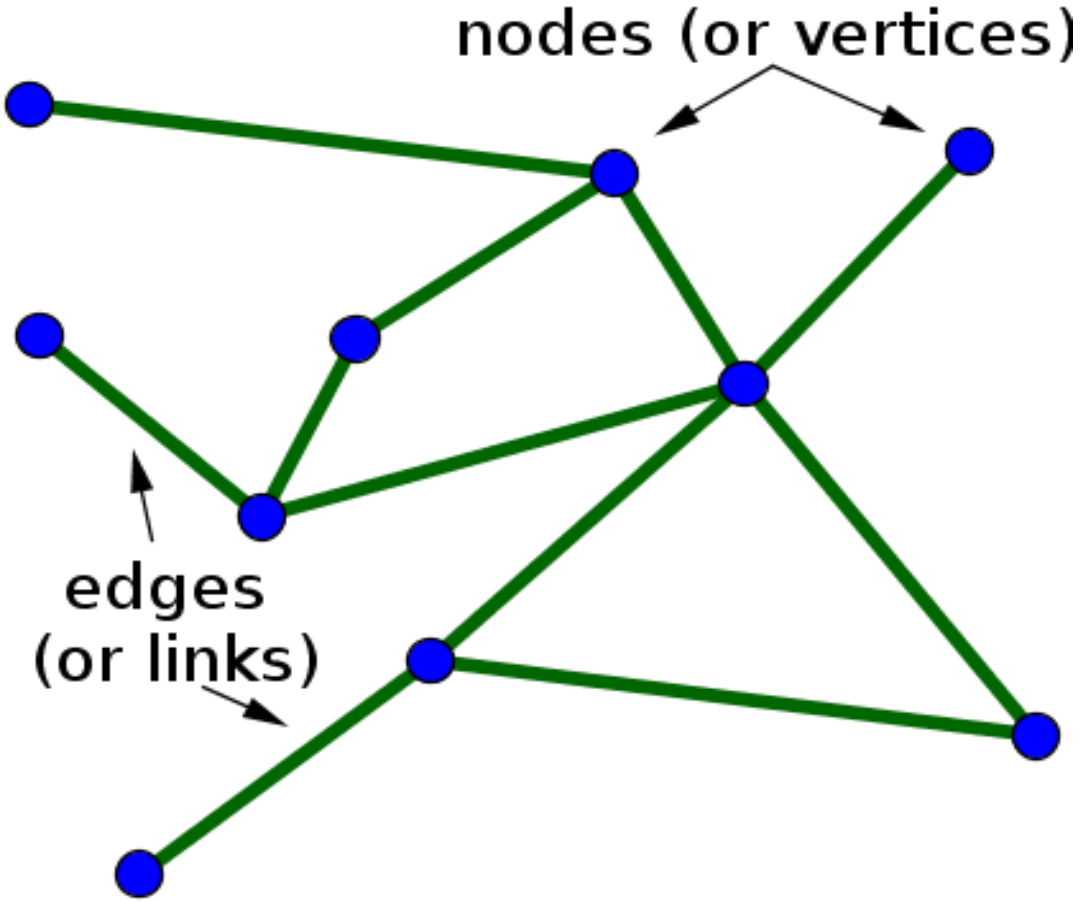
Undirected Graphs

- Undirected graph is set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional



Graphs

Undirected Graphs



Graphs

Graph Algorithms for Assembly

NGS assemblers have been organized into three categories, based on graphs:

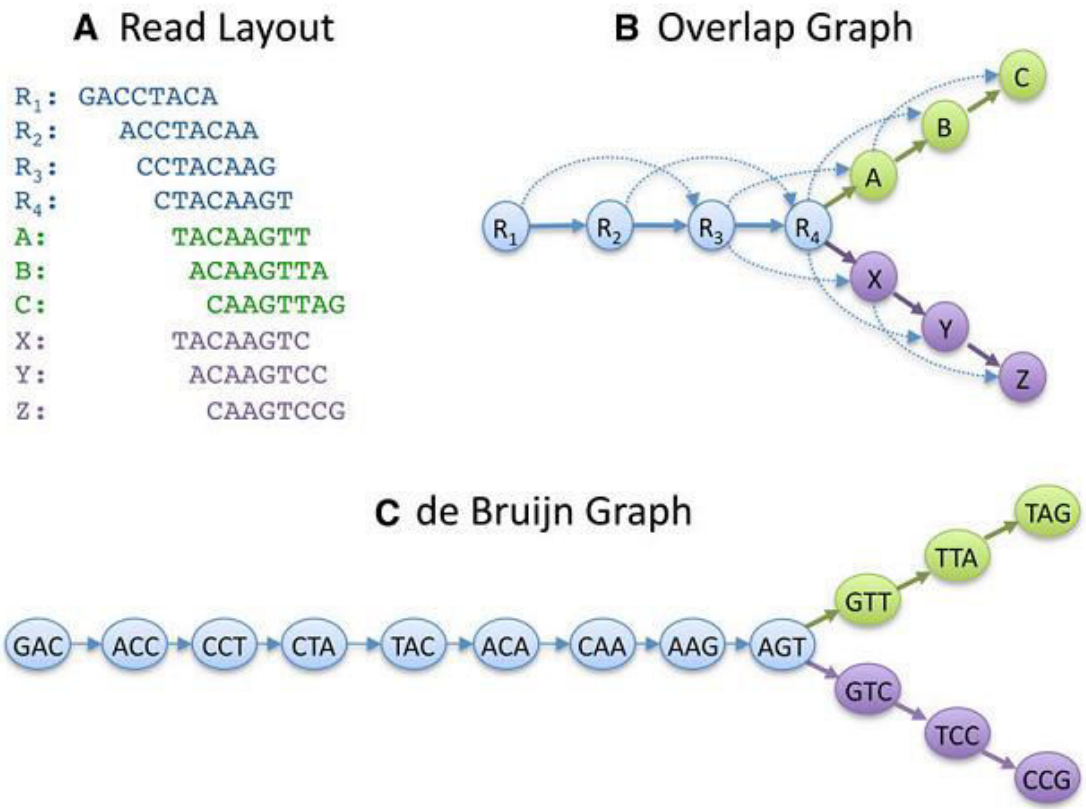
- **Overlap-Layout-Consensus (OLC) methods** rely on an overlap graph

Graphs

Graph Algorithms for Assembly

- The **de Bruijn Graph (DBG) methods** use some form of K-mer graph
- The **greedy graph algorithms** may use OLC or DBG

Graphs



http://www.biologicalproceduresonline.com/content/11/1/52/figure/F2?highres=y

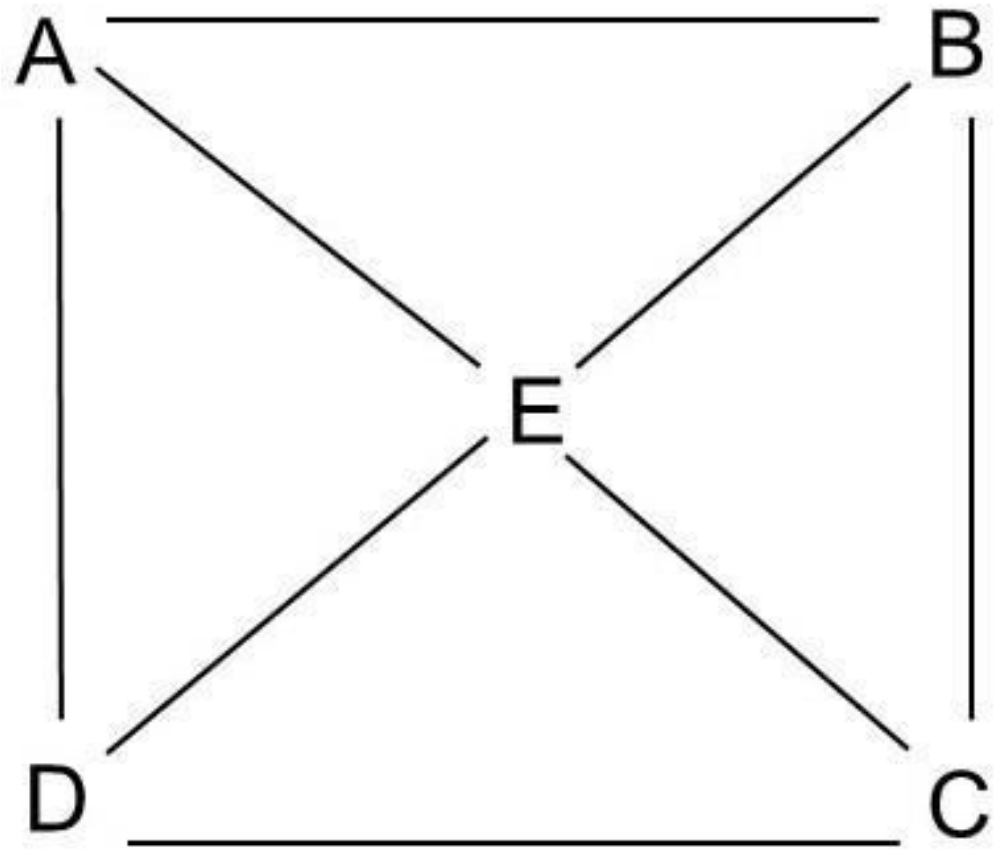
Graphs

Hamiltonian Path

- The Hamiltonian path in a graph is a path that includes each vertex of the graph once and only once
- At the end, of course, the circuit must return to the starting vertex

http://www4.ncsu.edu/~uzgeorge/HamiltonCircuits7-19and22.pdf

Graphs



http://www4.ncsu.edu/~uzgeorge/HamiltonCircuits7-19and22.pdf

Graphs

Eulerian Path

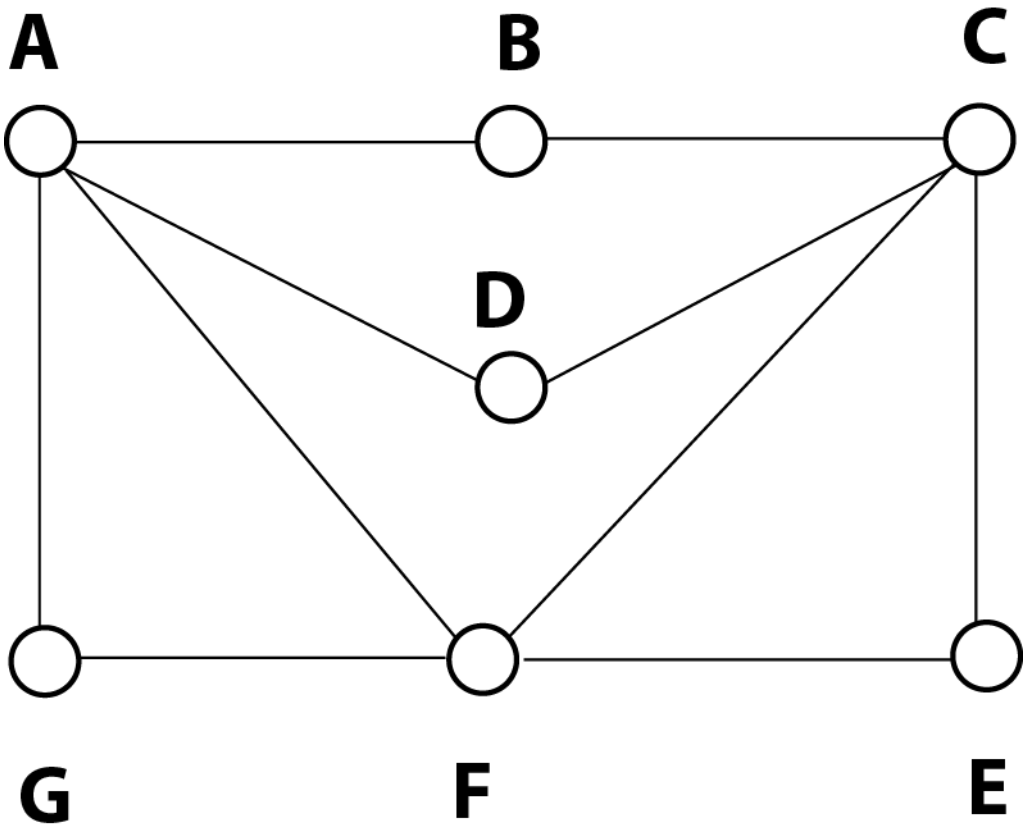
- An Eulerian path is a path that visits every edge of the graph once and only once.
- It can end on a vertex different from the one on which it began.

Graphs

Eulerian Circuit

- An Eulerian path which begins and ends on the same vertex.
- It starts and ends on the same vertex.

Graphs



Graphs

Conclusions

- Graph is a set of node plus set of edges between the nodes
- NGS assemblers have been organized into three categories, based on graphs

Overlap Layout Consensus

Outline

- Introduction
- Steps
- Advantages
- Limitations
- OLC based Assemblers
- Conclusion

Overlap Layout Consensus

Introduction

- The OLC strategy is arguably the most successful assembly strategy in practical setting
- 3 steps
  - Overlap
  - Layout
  - Consensus

Overlap Layout Consensus

Overlap stage

- Reads are compared to each other to construct a list of pairwise overlaps
- This information is then used to construct an overlap graph

Overlap Layout Consensus

Overlap stage

Read1 - TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

Read2 - GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

overlap

TTTGGTGCTCTTCGAAAAGGGATCTTCGAGAGAGATCTCGCGATAAGGTTG

GAGAGAGATCTCGCGATAAGGTTGAAGTAGAAAAATGTGTGTGGTGAA

http://www.homolog.us/Tutorials/index.php?p=1.3&s=1

Overlap Layout Consensus

Overlap Graph

- Each node corresponds to reads
- An edge connects two nodes if an overlap was identified between the corresponding reads

Overlap Layout Consensus

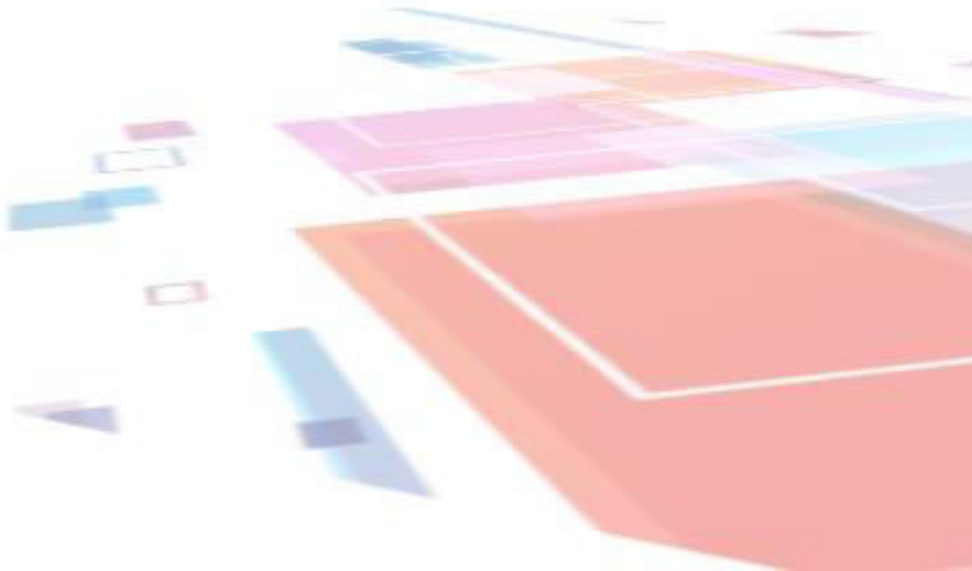
Layout stage

- During this stage, overlap graph is analyzed in order to identify paths through the graph that correspond to the genome being assembled

Overlap Layout Consensus

Layout stage

- The goal is to find single shortest Hamiltonian path that visits each node in graph exactly once

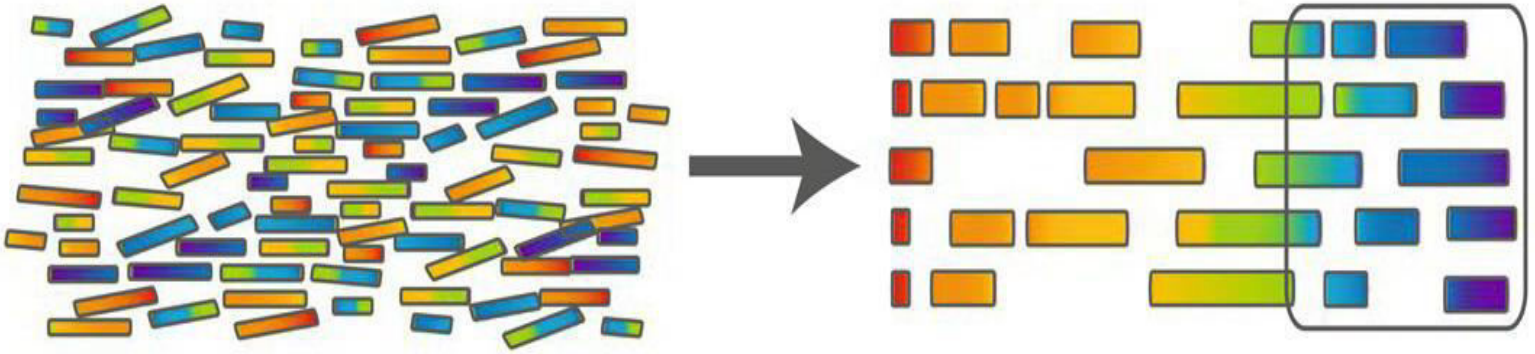


Overlap Layout Consensus

Consensus stage

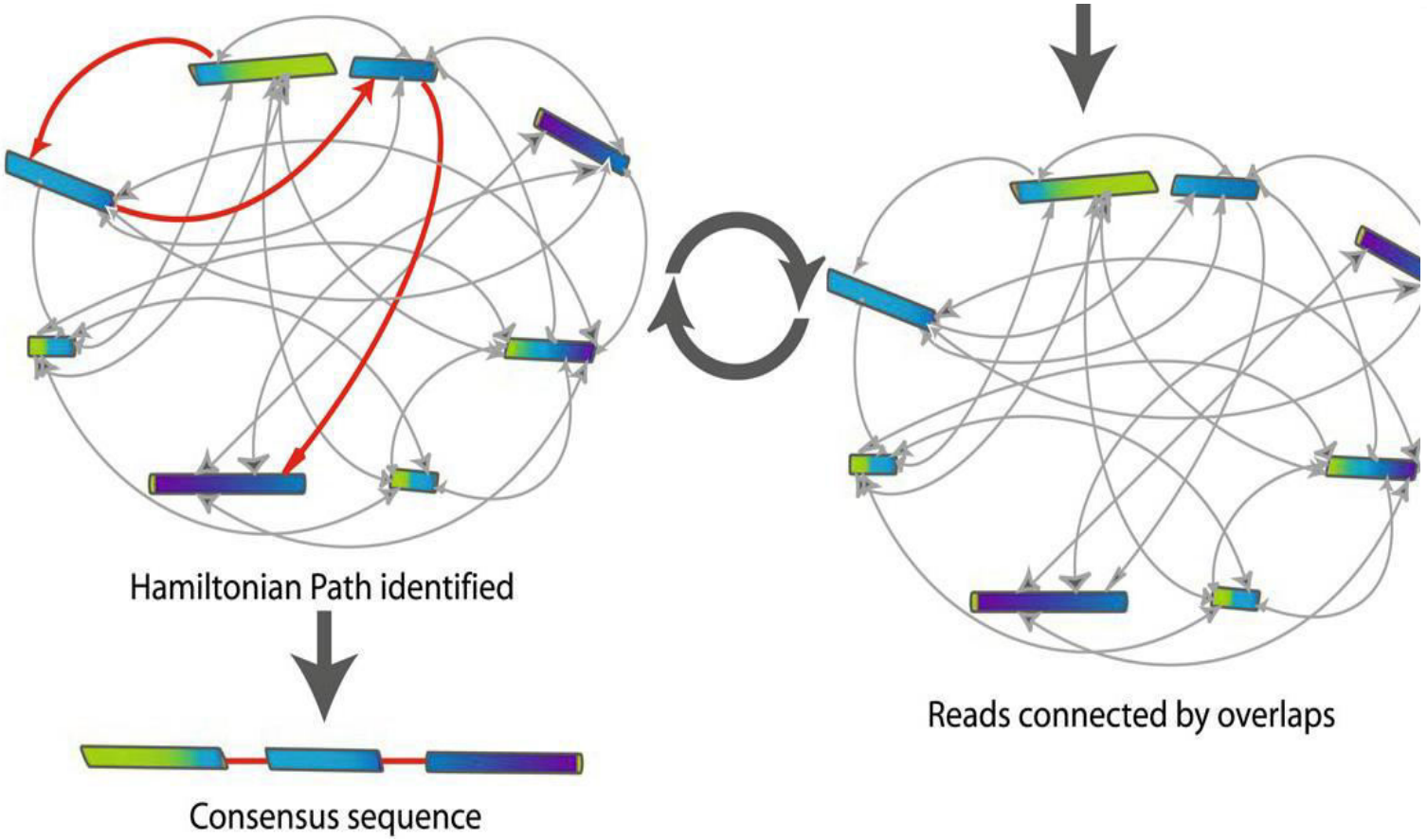
- Multiple sequence alignment of all reads covering the genome and the sequence of the genome is inferred through the consensus of the aligned reads

Overlap Layout Consensus



http://www.biologicalproceduresonline.com/content/11/1/52/figure/F2?highres=y

Overlap Layout Consensus



http://www.biologicalproceduresonline.com/content/11/1/52/figure/F2?highres=y

Overlap Layout Consensus

Advantages

- Suitable for large size genomes
- Works for both short and long reads
- OLC approach is capable of handling NGS data

Overlap Layout Consensus

Advantages

- Generate correct order of contigs
- OLC recognizes repeats but may collapse then into one
  - ABBBCBBD => ABCBD

Overlap Layout Consensus

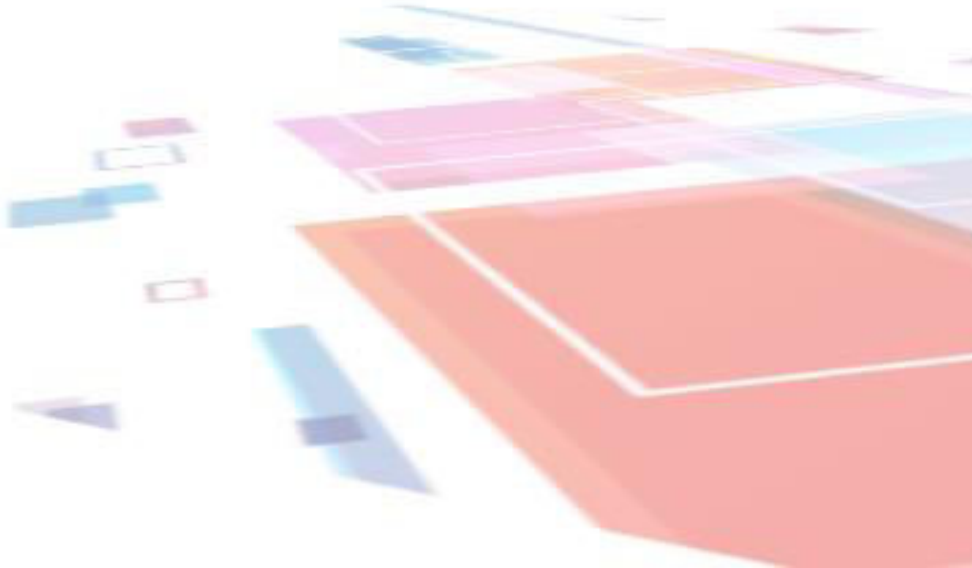
Limitations

- The processing cost of the overlap phase is very high
  - time consuming phase to determine the overlap of every pair of reads in the data set

Overlap Layout Consensus

Limitations

- No efficient algorithm to find the Hamiltonian path



Overlap Layout Consensus

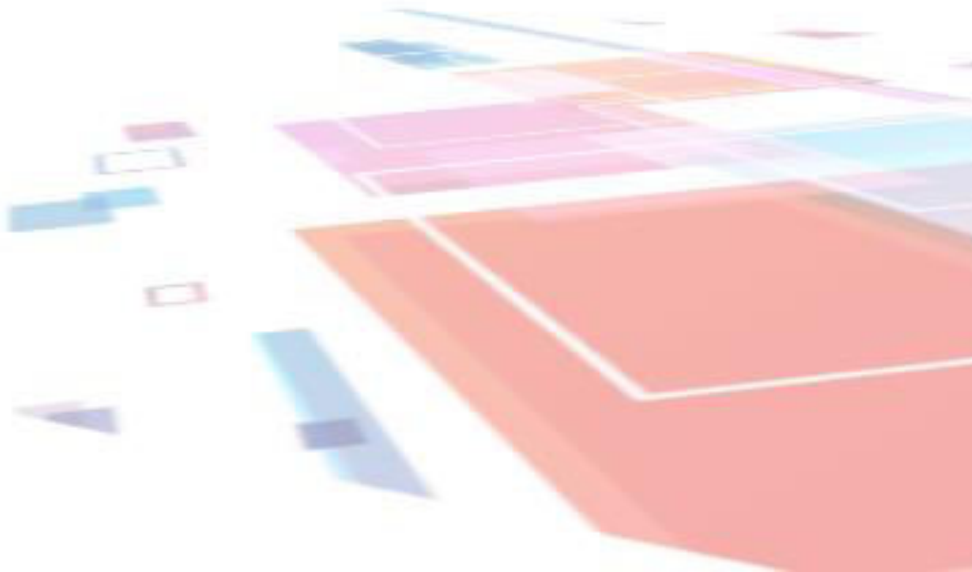
OLC Based Assemblers

- Arachne
- Celera Assembler (CABOG)
- Newbler
- Minimus
- Edena
- CAP
- PCAP

Overlap Layout Consensus

Conclusion

- The OLC strategy is arguably the most successful assembly strategy in practical setting.



De Bruijn Graph

Outline

- Introduction
- Steps
- Problems in Graph Construction
- Advantages
- Limitations
- DBG based Assemblers
- Conclusion

De Bruijn Graph

Introduction

- This approach for genome assembly is most widely applied to short reads from the Solexa and SOLiD Platforms

De Bruijn Graph

Steps

- Generate overlapping substrings of length k from the reads
- These substrings are called **Kmers**

De Bruijn Graph

Steps

- Generate De Bruijn graph
  - Nodes represent all possible fixed length strings or **Kmers**

De Bruijn Graph

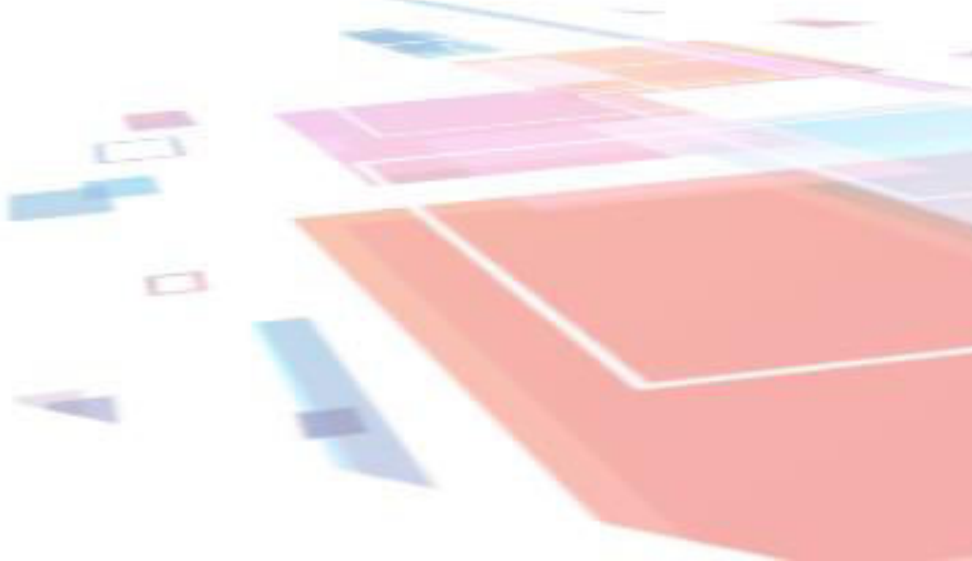
Steps

- Edges represent overlap of k-1 nucleotides
- Two nodes are linked with an edge if they share a k-1 mer

De Bruijn Graph

Steps

- Connect one Kmer to another if the two k-mers completely overlap except for one nucleotide at each end

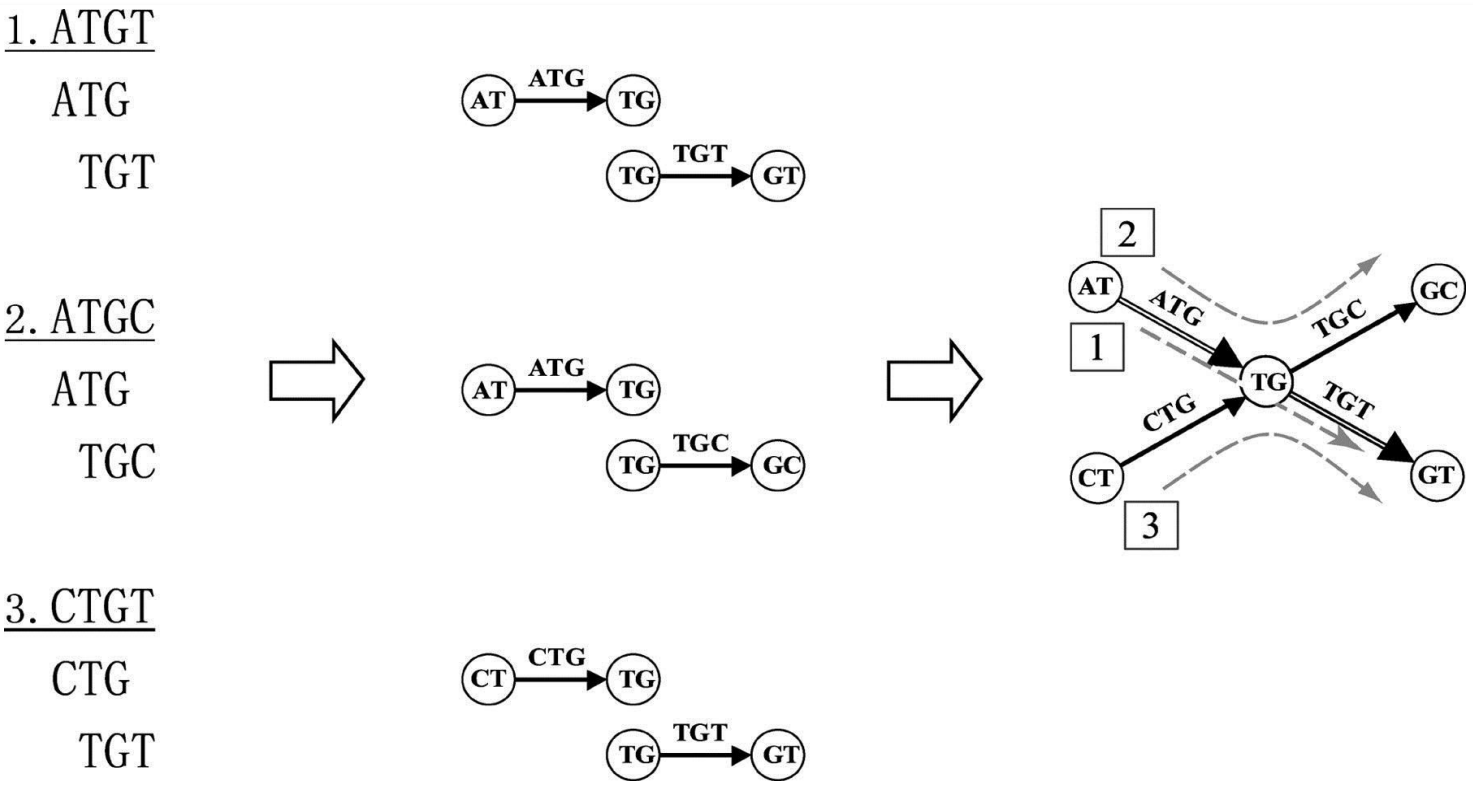


De Bruijn Graph

Steps

- Third, look for a Euler cycle, representing a candidate genome because it visits every edge of graph exactly once

De Bruijn Graph



Yu Zhang, and Michael S. Waterman PNAS 2005;102:1285-1290

De Bruijn Graph

DBG Based Assemblers

- EULER(-SR)
- Velvet
- ALLPATH(S2)
- ABySS
- SOAPdenovo

De Bruijn Graph

Advantages

- Works for short reads
- No pairwise overlap computation between reads which speeds up the process
- Efficient algorithms exist to find a Eulerian path in the graph

De Bruijn Graph

Conclusion

- This approach for genome assembly is most widely applied to short reads from the Solexa and SOLiD Platforms.

De Bruijn Graph Example

Deconstruction

- The first step of the De Bruijn assembler is to deconstruct the sequencing reads into its constitutive “kmers”.

De Bruijn Graph Example

Deconstruction

- The first step of the De Bruijn assembler is to deconstruct the sequencing reads into its constitutive “kmers”.

De Bruijn Graph Example

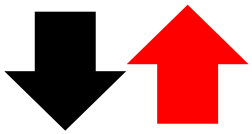
“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity,.... “

Dickens, Charles. A Tale of Two Cities. 1859. London: Chapman Hall
Example Adopted from J. Leipzig 2010

De Bruijn Graph Example

Itwasthebestoftimesitwastheworstoftimesitwas

theageofwisdomitwastheageoffoolishness…

Generate random 'reads'  **How do we assemble?**

fincreduli geoffoolis Itwasthebe Itwasthebe geofwisdom
itwastheep epochofinc timesitwas stheepocho nessitwast
wastheageo theepochof stheepocho hofincredu estoftimes
eoffoolish lishnessit hofbeliefi pochofincr itwasthewo twastheage
toftimesit domitwasth ochofbelie eepochofbe eepochofbe
astheworst chofincred theageofwi iefitwasth ssitwasthe
astheepoch efitwasthe wisdomitwa ageoffooli twasthewor

…etc. to 10's of millions of reads

De Bruijn Graph Example

Traditional Method

- All-vs-all assemblers fail due to immense computational
- A **million (10^6)** reads requires a **trillion (10^{12})** pairwise alignments

De Bruijn Graph Example

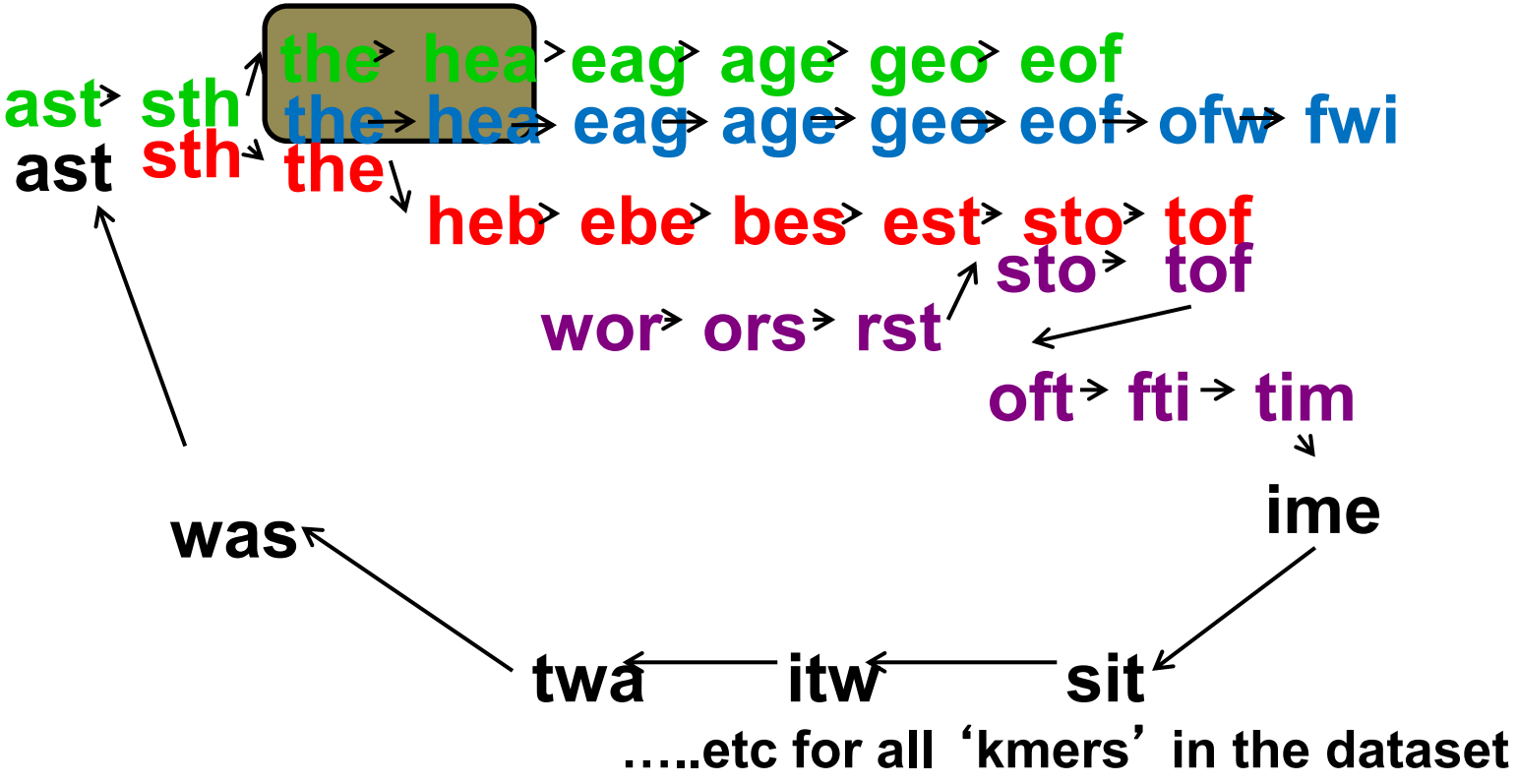
Step 1: Convert reads into "Kmers"

Reads:	theageofwi	sthebestof	astheageof	worstoftim
Kmers : (k=3)	the	sth	ast	wor
	hea	the	sth	ors
	eag	heb	the	rst
	age	ebe	hea	sto
	geo	bes	eag	tof
	eof	est	age	oft
	ofw	sto	geo	fti
	fwi	tof	eof	tim

…..etc for all reads in the dataset

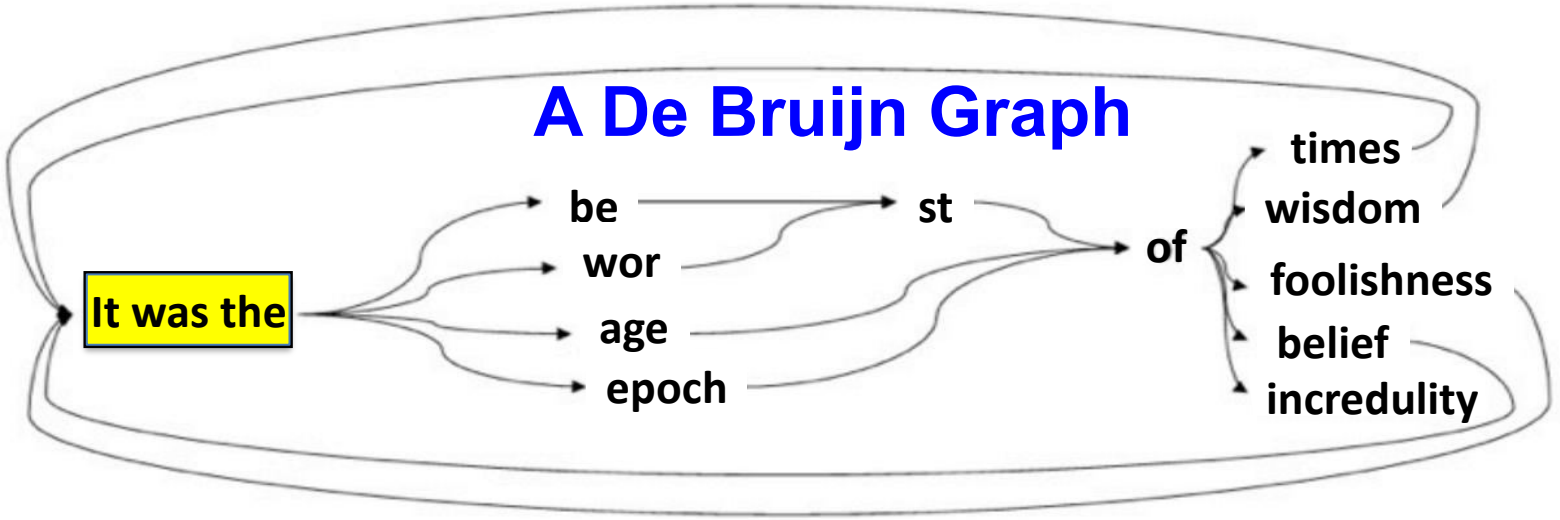
De Bruijn Graph Example

Step 2: Build a De-Bruijn graph from the kmers



De Bruijn Graph Example

Step 3: Simplify the graph as much as possible



De Bruijn Graph Example

The final assembly (k=3)

wor times itwasthe foolishness st wisdom

incredulity age epoch be of belief

A better assembly (k=20)

Itwasthebestoftimesitwastheworstoftimesitwasthe

ageofwisdomitwastheageoffoolis…

De Bruijn Graph Example

The final assembly (k=3)

wor times itwasthe foolishness st wisdom

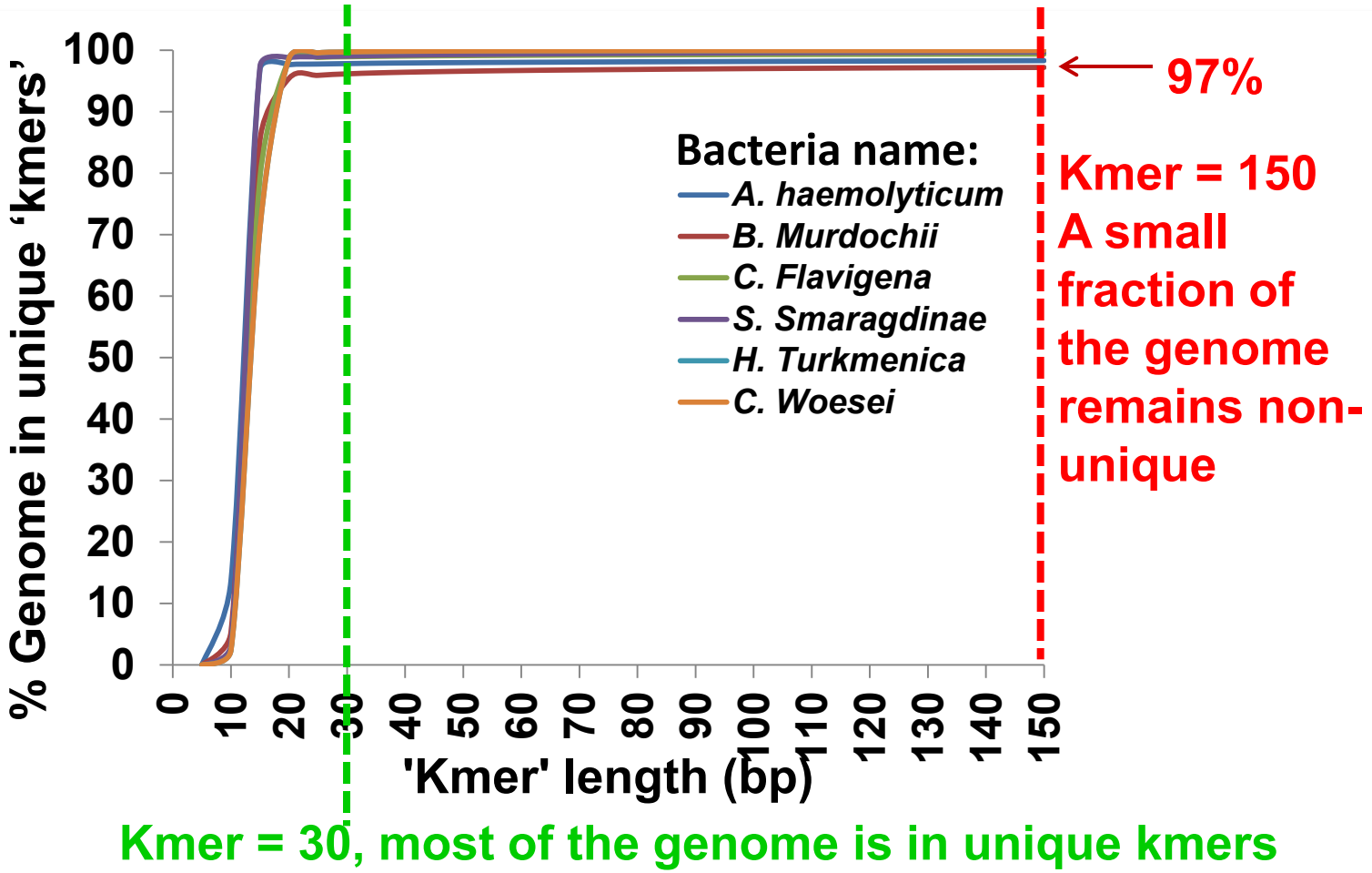
incredulity age epoch be of belief

A better assembly (k=20)

Itwasthebestoftimesitwastheworstoftimesitwasthe

ageofwisdomitwastheageoffoolis…

De Bruijn Graph Example



De Bruijn Graph Example

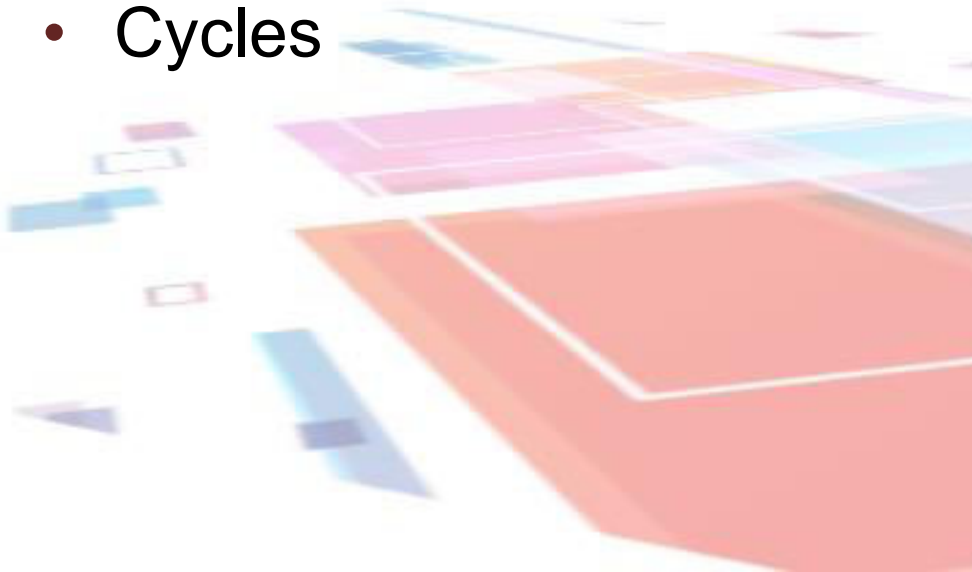
Conclusions

- Using an apropriate size Kmer and storing information stored in millions of reads is an important aspact of de Bruijn graph Algorithm

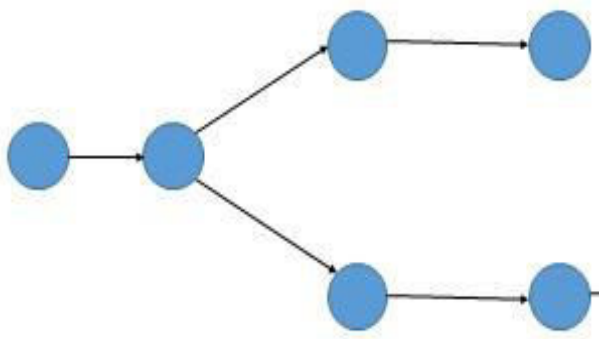
Assembly Issues

Problems in Graph Construction

- Spurs
- Bubbles
- Converging and diverging paths
- Cycles



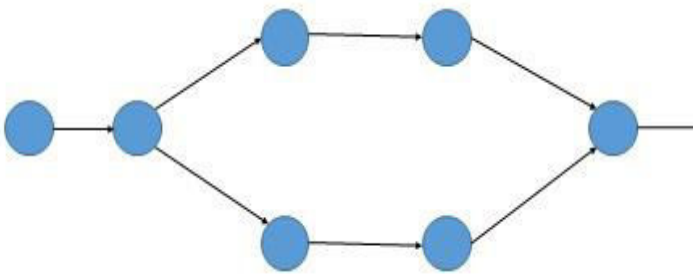
Assembly Issues



Spurs

- Short dead-end branches (divergences) of the main path
- Possible causes include sequencing errors toward one end of a read, and low coverage

Assembly Issues

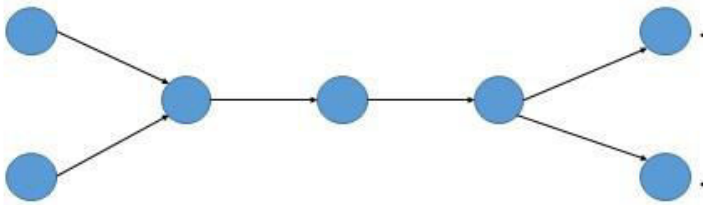


Bubbles

- Divergence of a path into two branches that afterwards join together again into one path
- Possible causes include sequencing errors toward the middle of a read, and by polymorphisms in the target

Assembly Issues

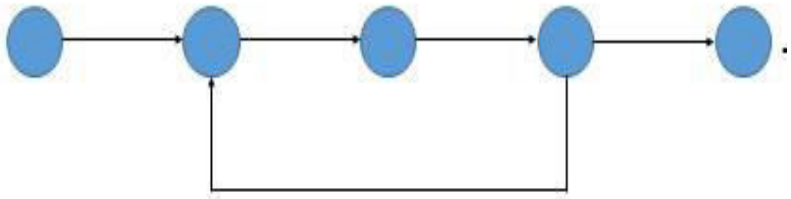
Convergence



Converging and Diverging Paths

- Inverse definition than for the bubbles, two paths converge into one, that later diverges again into two separate paths
- Possible causes are repeats in the target genome

Assembly Issues



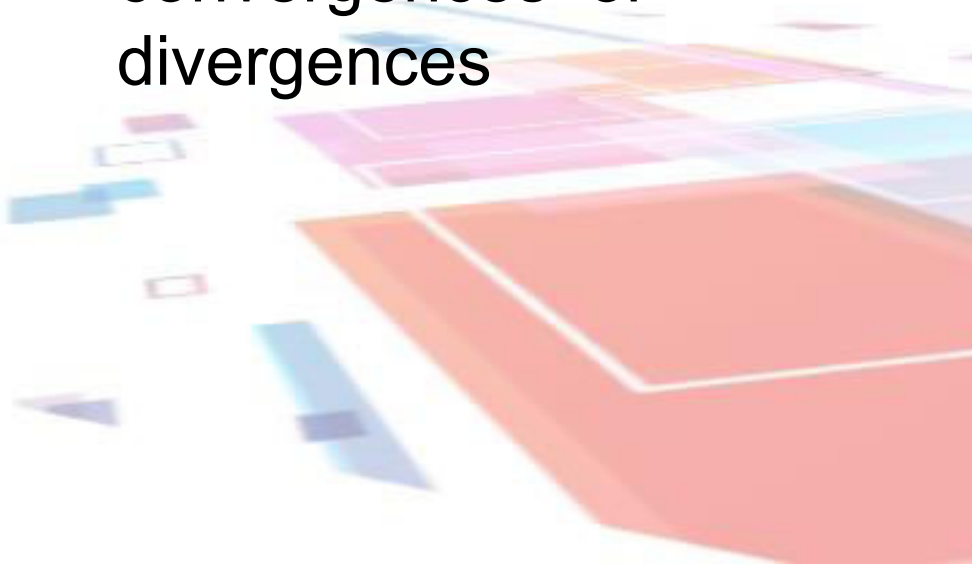
Cycles

- Paths that converge on themselves.
- Possible causes are repeats in the target genome.

Assembly Issues

Quality Control

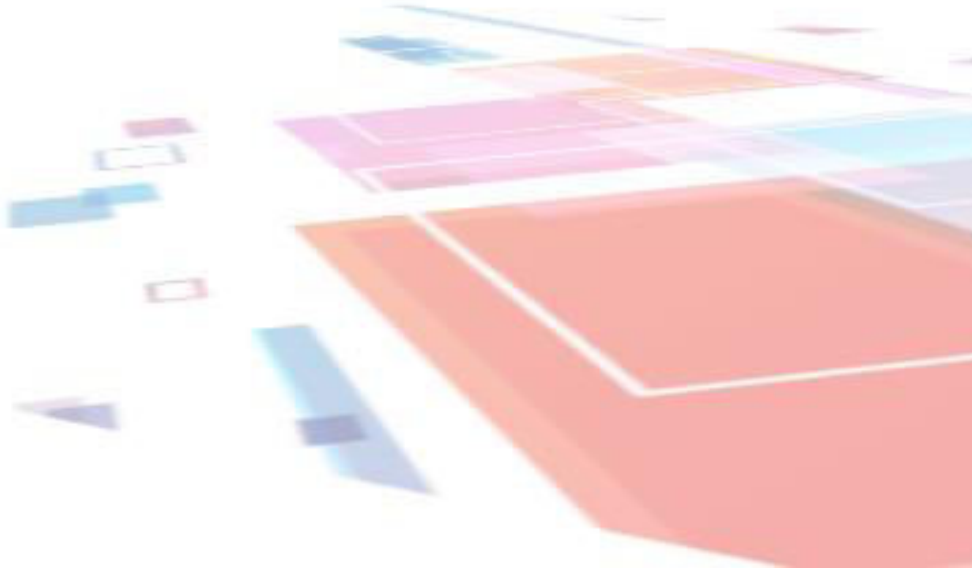
- Filter the graph of erroneous occurrences (i.e., bubbles, spurs, cycles) convergences or divergences



Assembly Issues

Quality Control

- Nodes that are unambiguously connected by an edge are merged together



Assembly Issues

Limitations

- Loss of information due to splitting into k mers
- Exponential number of Eulerian paths because of number of repeats in genome

De Bruijn Graph Example

Limitations

- Assembler has to find most probable path
- Increase in memory consumption and runtime

De Bruijn Graph Example

K mer size

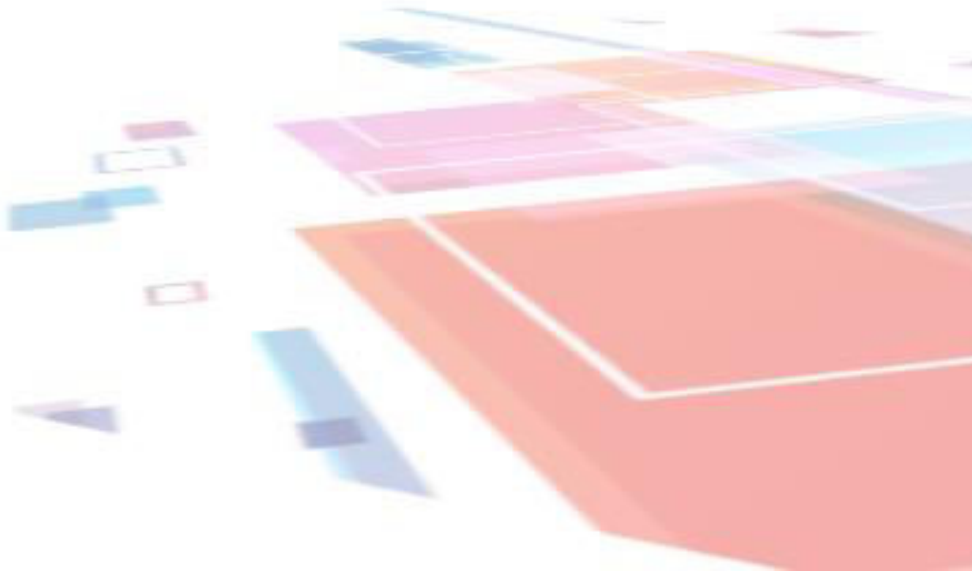
Larger K-mers;

- It is easy to convert the de Bruijn graph into an unique sequence
- the retention of more information about short repeats

De Bruijn Graph Example

K mer Size

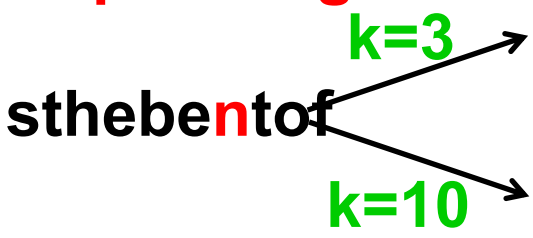
- **Smaller** Kmer value results in the loss of useful information about short repeats



Assembly Issues

Why not always use longest ‘k’ possible?

Sequencing errors:



sth the ebe ent tof
heb ben nto

Mostly unaffected kmers

sthebentof 100% wrong kmer

Assembly Issues

Conclusions

- Selecting appropriate Kmer is an important step
- Repeats might collapse these assemblies

Credit

References

- Sovic, I., Skala, K., & Sikic, M. (2013, May). Approaches to DNA de novo assembly. In *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on* (pp. 351-359). IEEE.

Credit

References

- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Credit

References

- Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, *10*(4), 354-366.
- http://gcat.davidson.edu/phast/debruijn.html

Proteomics Introduction

BIO 601

Lesson-1

Proteomics is the study of all the proteins found in an organism

Lesson-1

Introduction to Proteomic

- The terms “proteomics” and “proteome” were coined by Marc Wilkins and colleagues in 1994 and mirror the terms “genomics” and “genome.

fppt.com

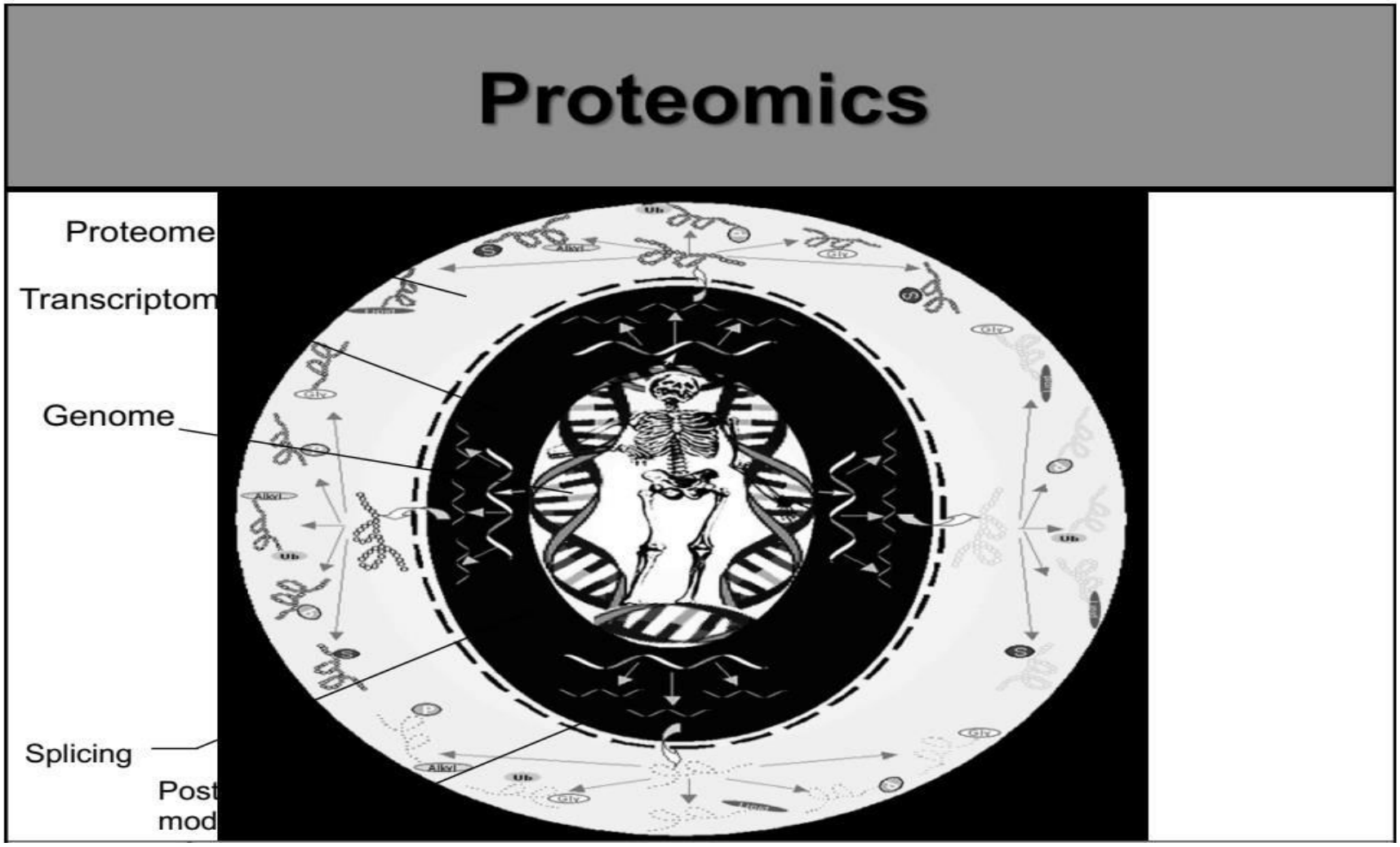
GENOMICS

- Genome: The entire sequence of an organism's hereditary information, including both coding and non-coding regions, encoded in DNA is known as "genome".

- Studying genome of an organism by employing sequencing and genome mapping is known as "genomics".

fppt.com

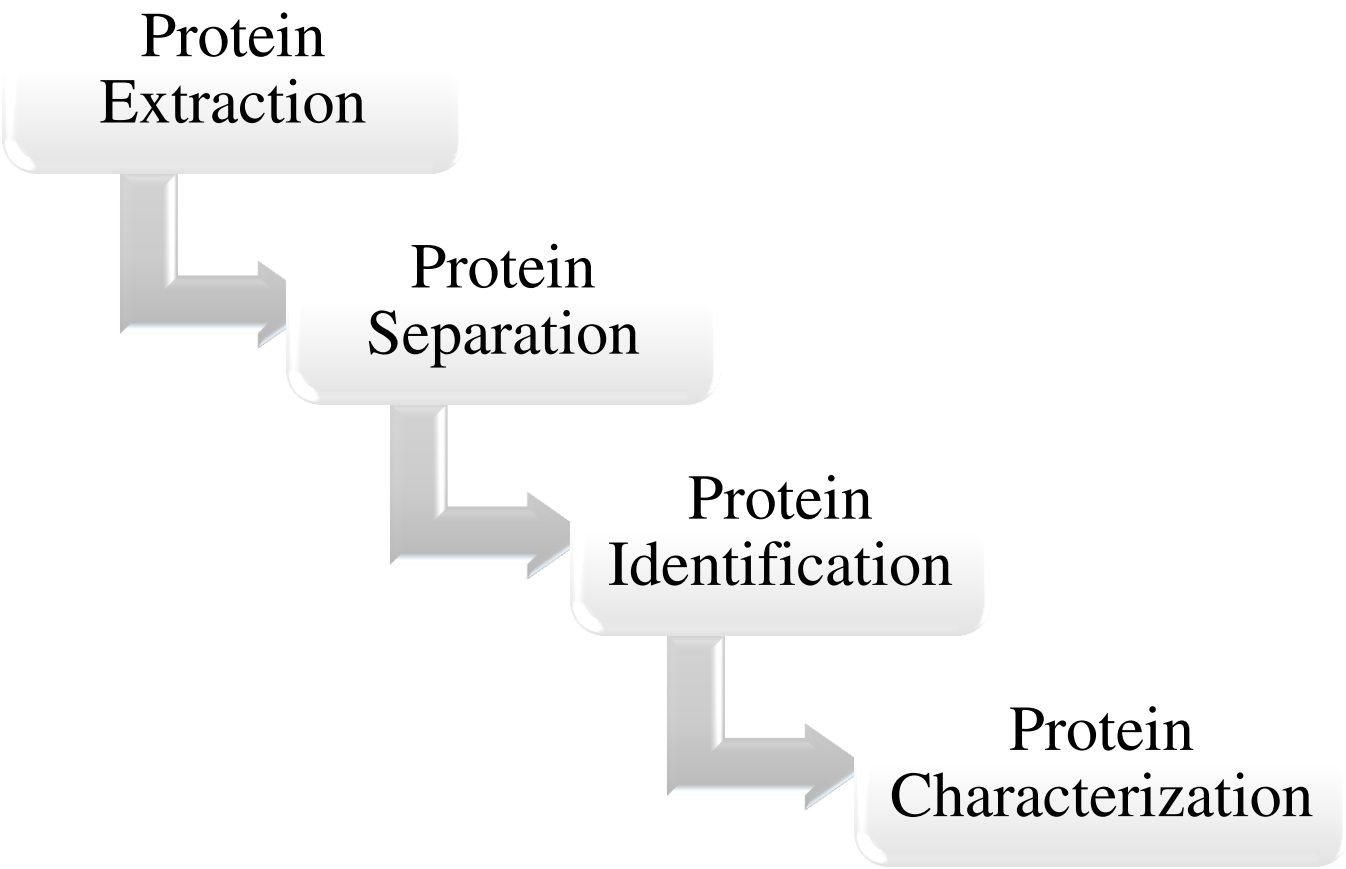
Introduction to Proteomics



fppt.com

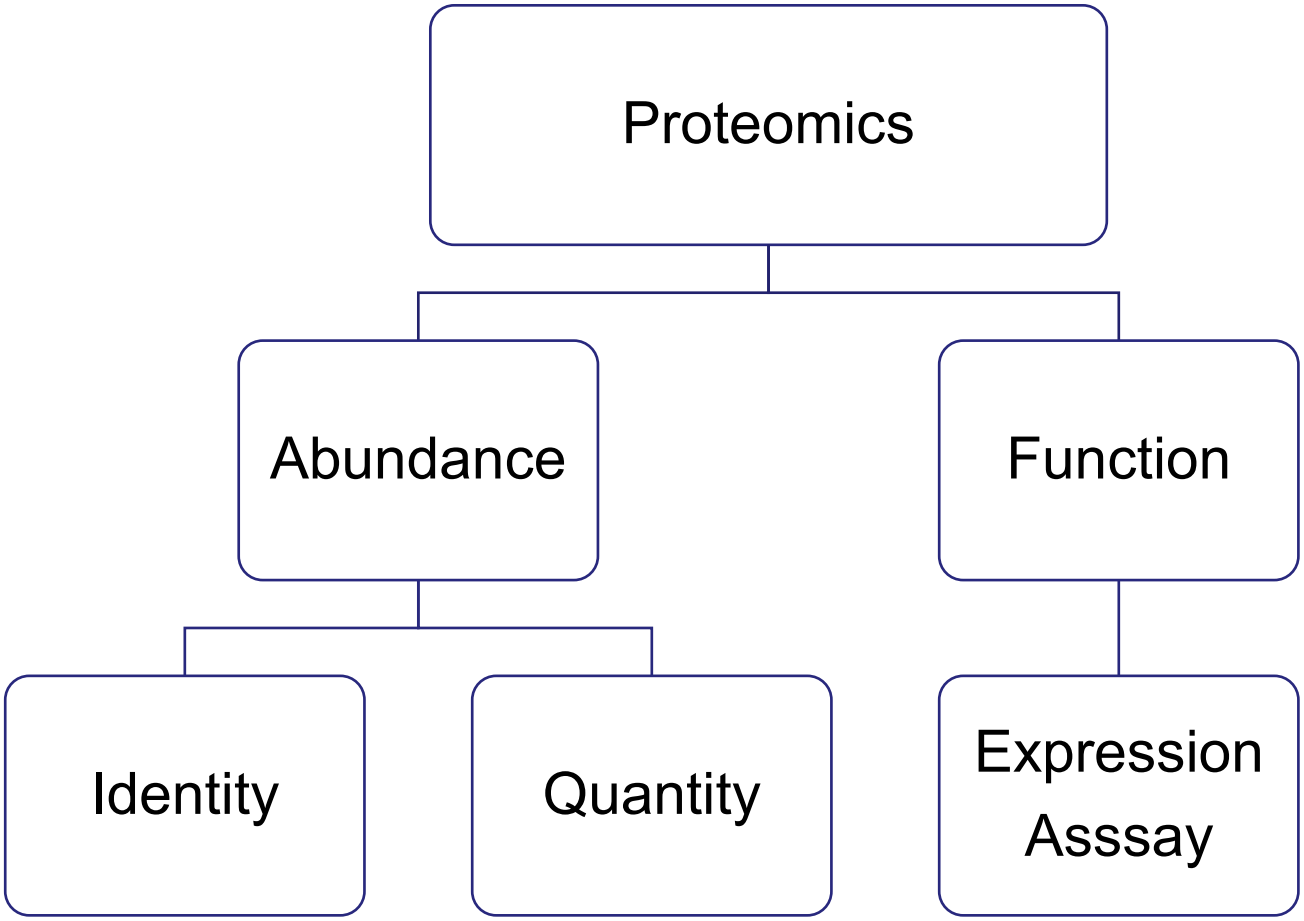
Proteomics Introduction

Steps involved in Proteome Analysis

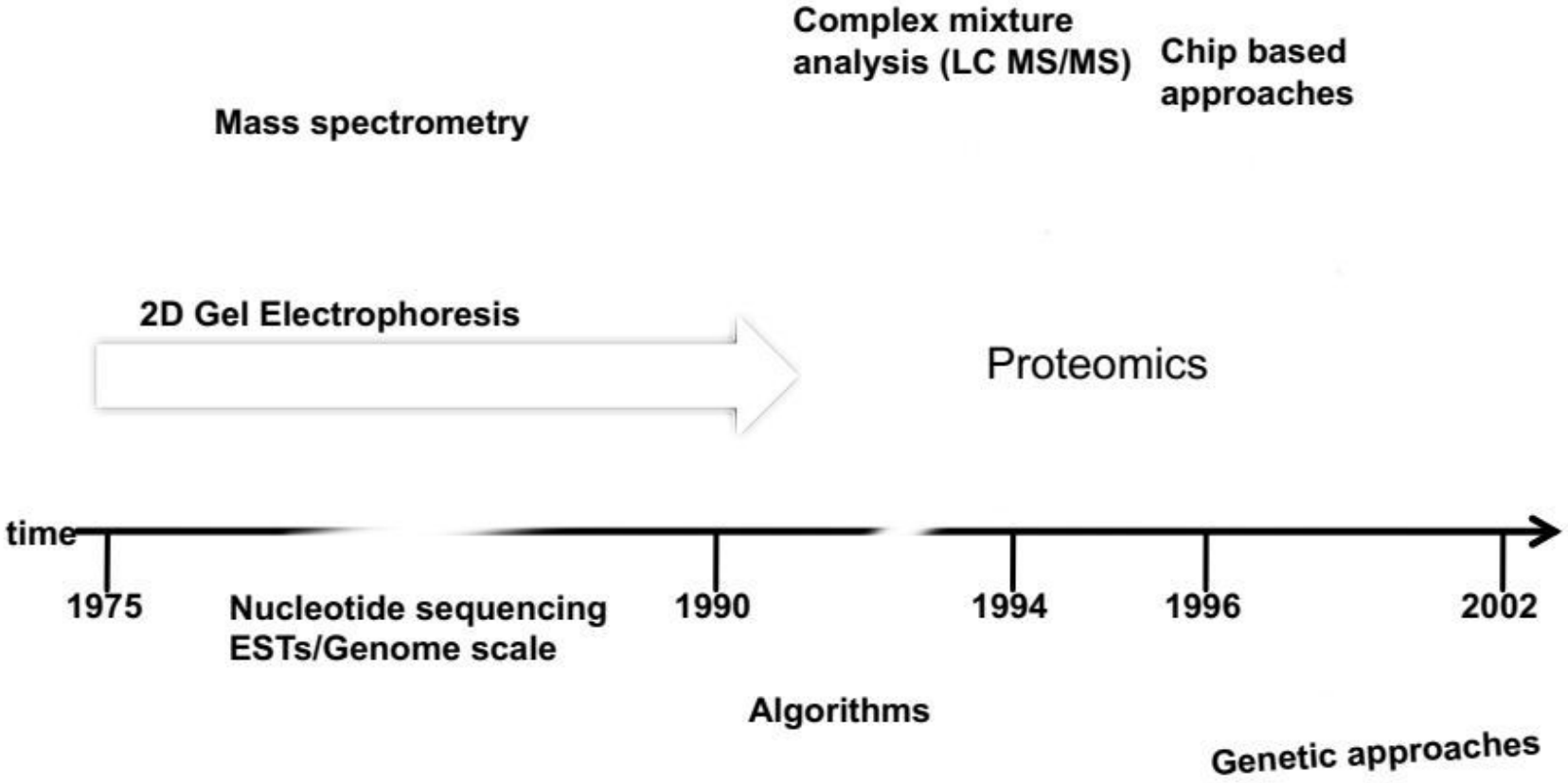


Proteomics Introduction

Function and Abundance based Proteomics



Emergence of Proteomics



Nature genetics 2003, 33, 311-323

Proteomics Introduction

Developmental Phases of proteomics

- Protein chemistry to proteomics
- Diversification of proteomics technologies
- Advent of novel technology platforms

Comparative Proteomics

Introduction

- For a new Genome, predicted genes are translated
- The collection of protein sequences encoded by the genome makes up the proteome of that individual

Comparative Proteomics

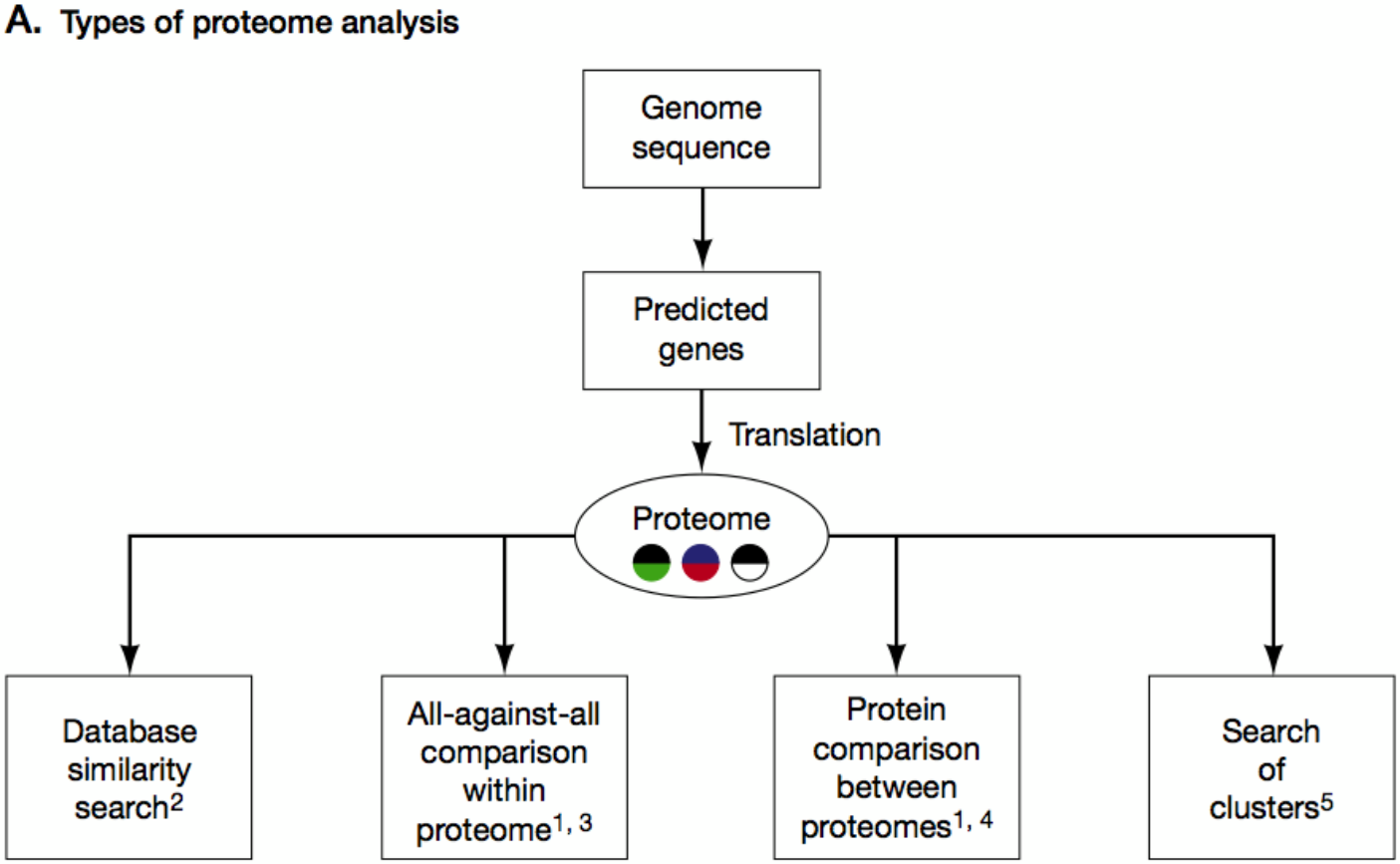


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Comparative Proteomics

All against all, self comparison

- Comparison of all proteins with each other within proteome
- Identify unique proteins from the ones having paralogs
- Identify Gene families

Comparative Proteomics

B. Examples of database hits resulting from domain structure of proteins[6]

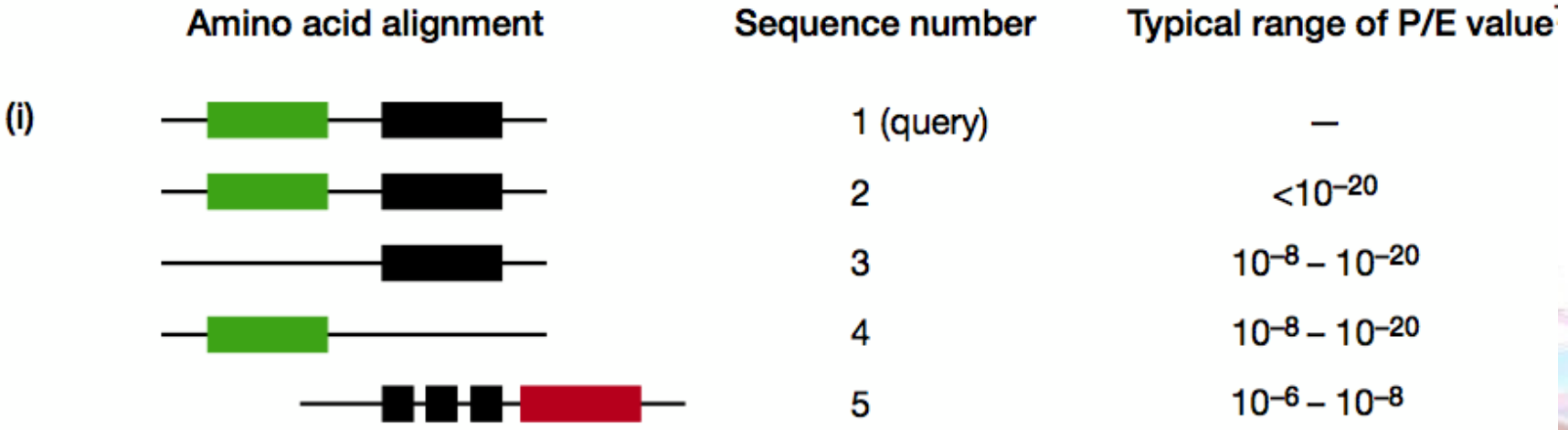


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Comparative Proteomics

Cluster Analysis

- To sort out the relationships of all related proteins
- Clustering classify the proteins based on some objective criteria e.g
  - E value cut off
  - Distance in alignment

Comparative Proteomics

Clustering by subgraph

- Each sequence is a vertex
- Significant alignment score is an edge
- Trimming by removing weak edges (High P/E)

Comparative Proteomics

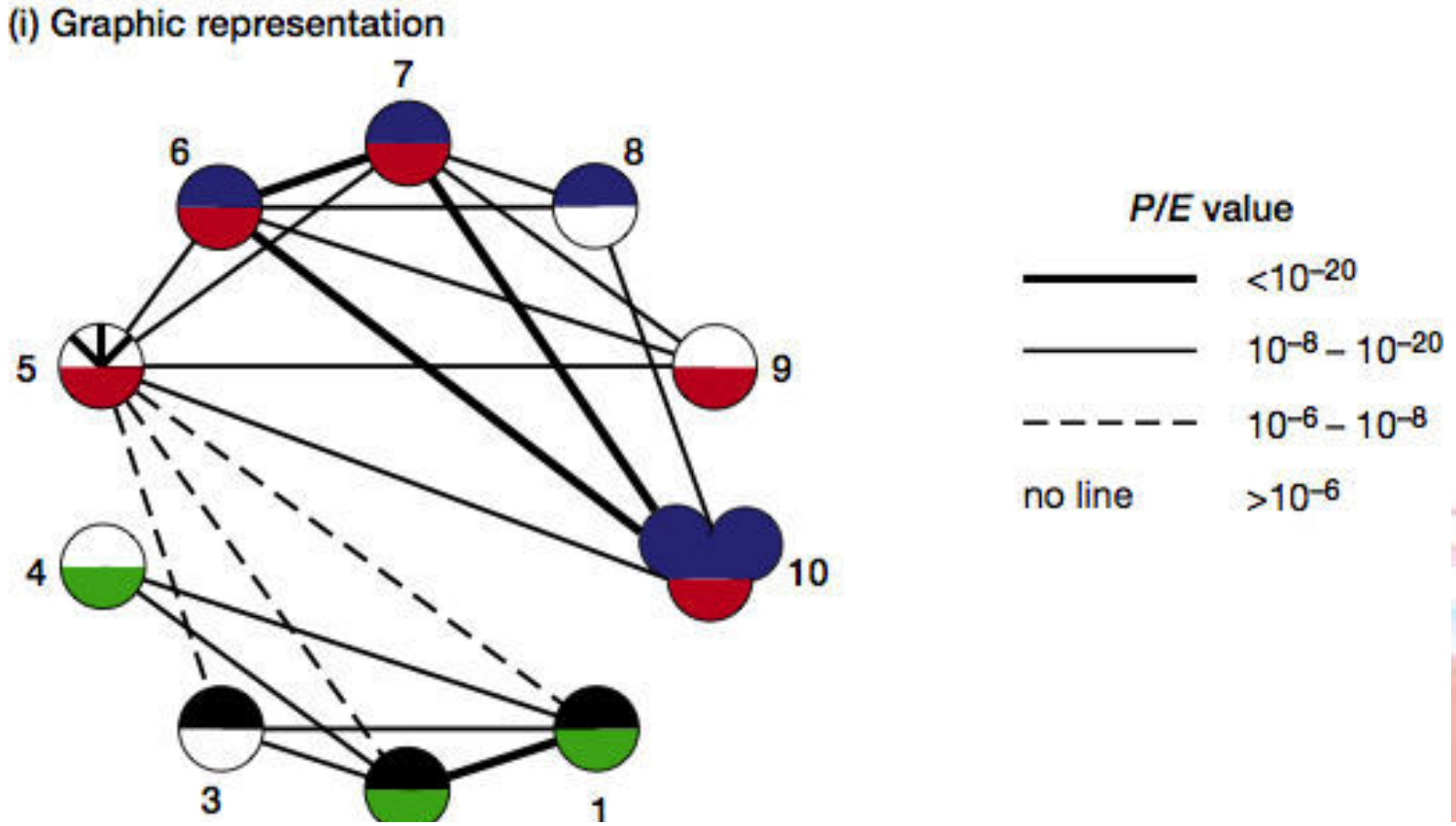


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Comparative Proteomics

Clustering by Linkage

- Each sequence is a vertex
- Significant alignment score is an edge
- Trimming by removing weak edges (High P/E)
- Or remove > e^{-6}
- Remaining subgraph should share 2/3rd of edges

Comparative Proteomics

Single Linkage

- A group of sequences in all-against-all comparison is subjected to MSA
- Create distance matrix
- Neigbour joining is then used to do clustering

Comparative Proteomics

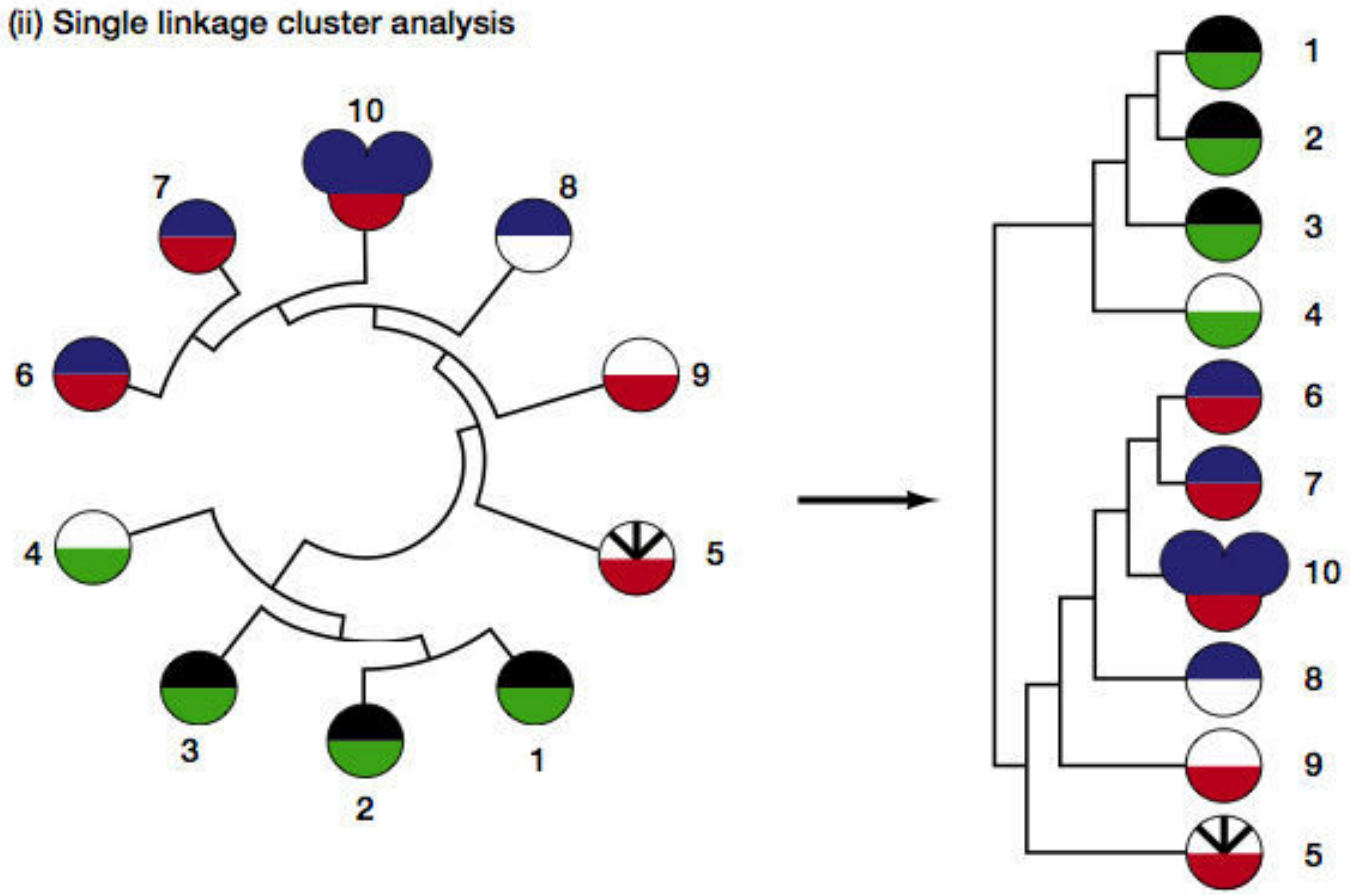


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Comparative Proteomics

Core Proteome

- All-against-all comparison provides an indication of Gene/Protein families
- Unique set of proteins is core proteome

Comparative Proteomics

Table 10.4. Numbers of gene families and duplicated genes in model organisms (Rubin et al. 2000)

Organism	Total number of genes	Number of gene families[a]	Number of duplicated genes[b]
Hemophilus influenzae (bacteria)	1709	1425[c]	284
Saccharomyces cerevisiae (yeast)	6241	4383	1858
Caenorhabditis elegans (worm)	18,424	9453	8971
Drosophila melanogaster (fly)	13, 600	8065	5536

[a] The number of clustered groups in the all-against-all analysis using the algorithm described in the text. This number represents the core proteome of the organism.

[b] Count of number of duplicated genes within the protein family clusters.

[c] 178 families have paralogs.

Comparative Proteomics

Conclusions

- Genome is translated into proteome
- Self comparison of proteome yields gene families and duplications
- Unique set of proteins is core proteome

Between-Proteome comparisons

Introduction

- Each proteome is used as a query in a database similarity search against another proteome or a set of proteomes
- If proteome is not available, EST database may be searched

Between-Proteome comparisons

Significance

- Helps finding orthologs, Gene families and domains
- Proteins with highly significant alignment score are likely to be orthologs
- Mostly the proteins related to core biological functions

Between-Proteome comparisons

Finding true orthologs

- Method 1
  - Reciprocal Hits
  - E < 0.01
  - 60% coverage
- Keep matched pairs with a very conservative P value 10^{-10} to 10^{-100}

Between-Proteome comparisons

Clusters of Orthologous Group (COG)

- Orthologs are assumed to be derived from common ancestor
- Each orthologs might also have paralogs
- Orthologs are clustered to form **COG**

Between-Proteome comparisons

Table 10.5. *Numbers of closely related yeast and worm sequences*

Cut-off *P* value	$< 10^{-10}$	$< 10^{-20}$	$< 10^{-50}$	$< 10^{-100}$
Total number of sequence groups	1171	984	552	236
Number of groups with more than two members	560	442	230	79
Number and percent of all yeast proteins (6217) represented in groups	2697 (40)	1848 (30)	888 (14)	330 (5)
Number and percent of all worm proteins represented in groups	3653 (19)	2497 (13)	1094 (6)	370 (2)

Adapted, with permission, from Chervitz et al. 1998 (copyright AAAS).

Between-Proteome comparisons

Proteomes to EST databases

- Expressed Sequence Tags (EST) are cDNA copies of cell's mRNA sequences
- For organisms whose genome sequence is not available

Between-Proteome comparisons

Proteomes to EST databases

- EST are single DNA reads
- Mostly 3’ biased
- May be Incomplete being dependent on gene expression
- **TBLASTN** is frequently used

Between-Proteome comparisons

Family and Domain Analysis

- Proteins are organized into domains that represent modules of structure or function
- Domain comparison reveals their biological roles

Comparative Proteomics

D. Domain identification[9]

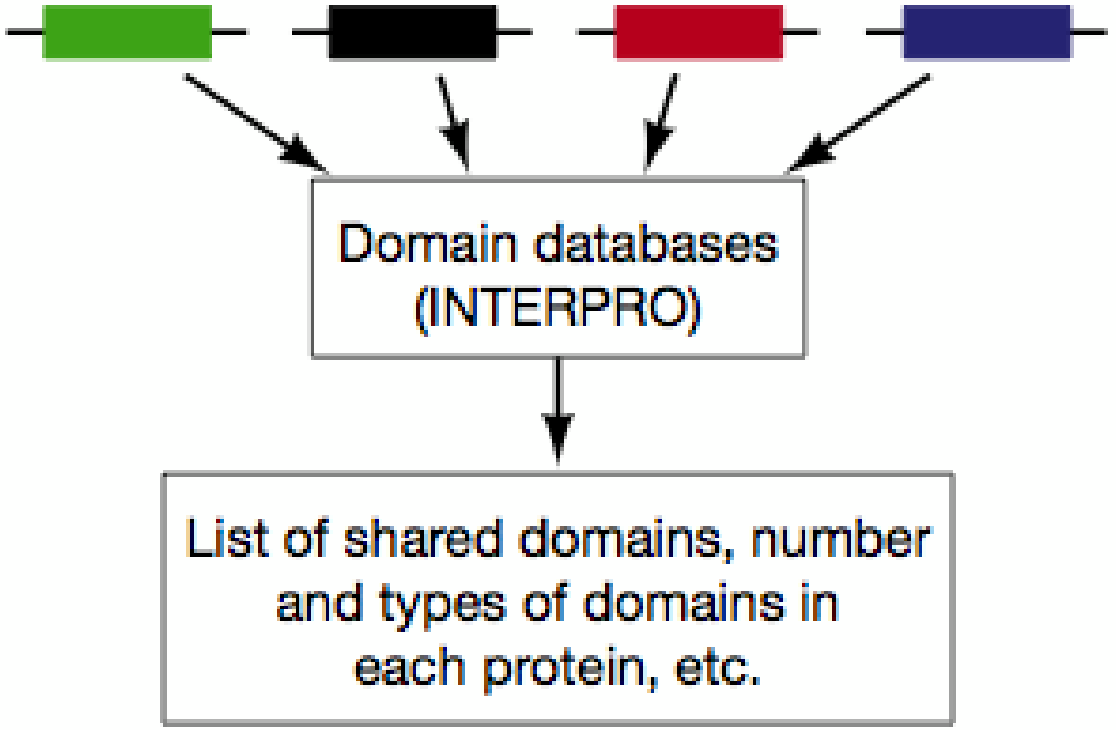


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Between-Proteome comparisons

Conclusions

- Proteome comparison helps finding orthologs, gene families and protein domains
- Domain comparison reveals their biological roles

Genetics and Genomics

Proteomics, Types and Techniques

Proteomics, Types and Techniques

Study of Proteomics

- Proteomics is the study of all the proteins found in an organism
- Different Types

Proteomics, Types and Techniques

Proteomics: Proteome Complexity

GENOME	PROTEOME
▪ 4 nucleotides	▪ 20 amino acids
▪ Double helix	▪ Each protein has unique 3D shape
▪ Same in all cells	▪ Differs with cell type

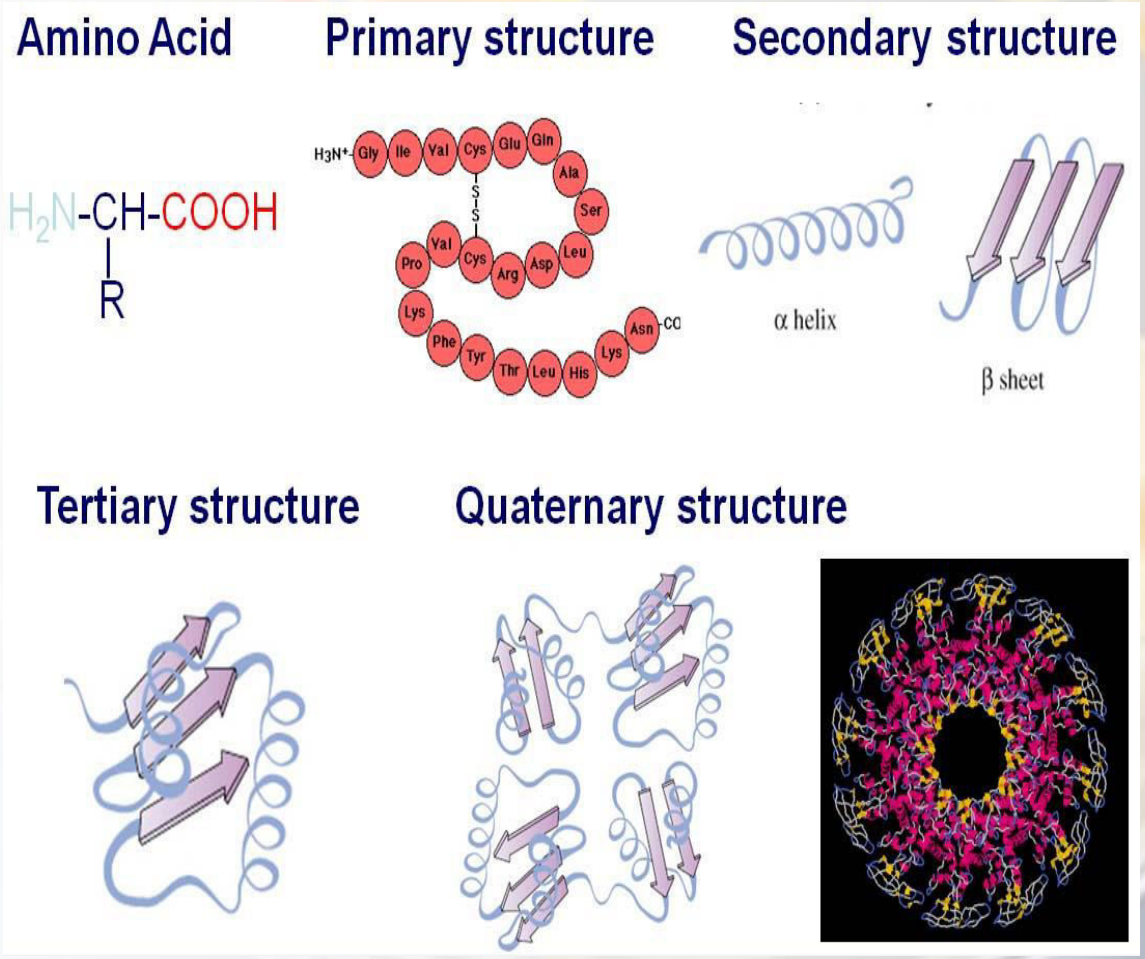
Proteomics, Types and Techniques

Structural Proteomics

- Structural proteomics deals with mapping out the 3-D structure and nature of protein and/or proteins complexes

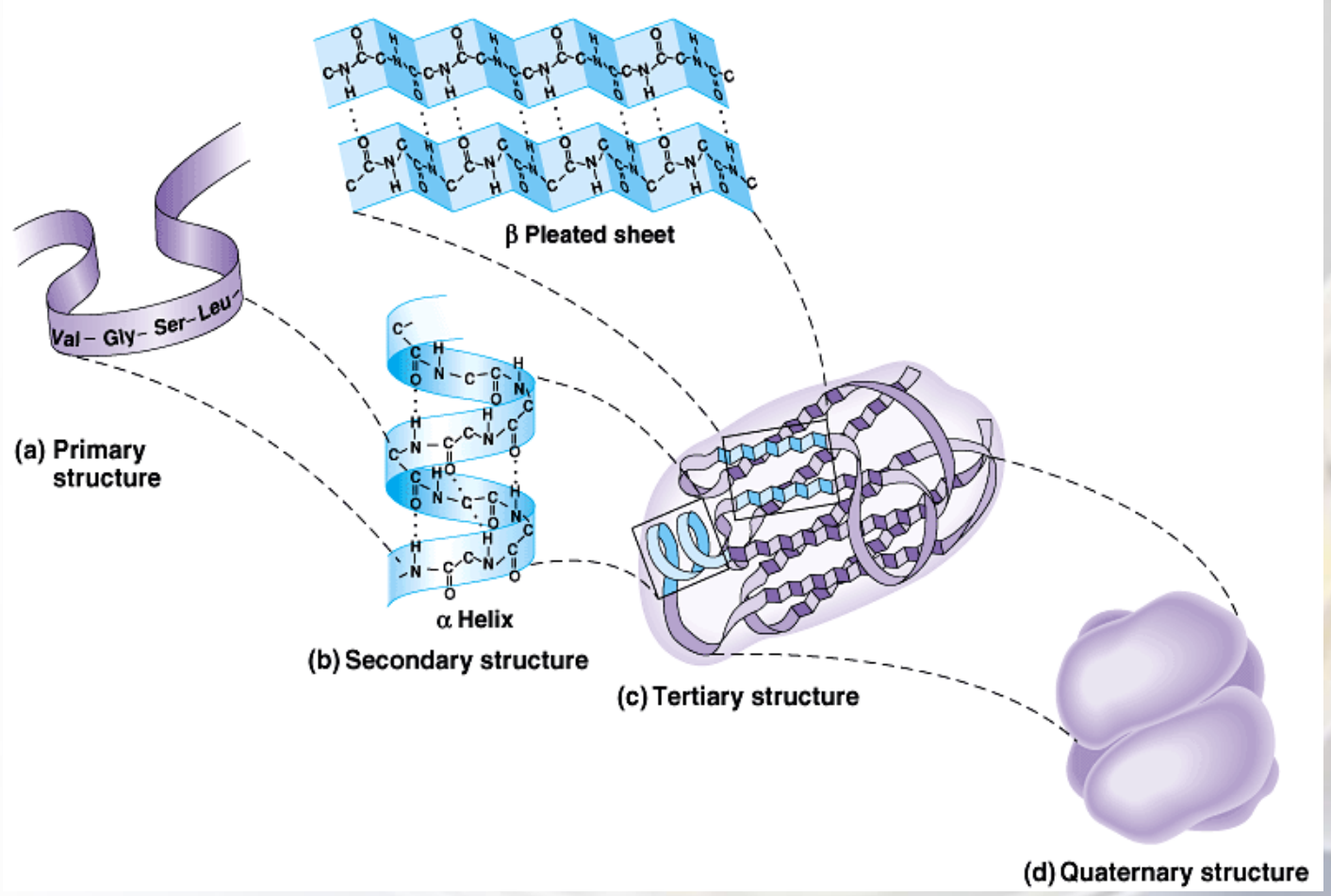
Proteomics, Types and Techniques

Proteomics: Protein Structures



Proteomics, Types and Techniques

Proteomics: Protein Structures



Proteomics, Types and Techniques

Functional Proteomics

- Posttranslational modifications
- Protein-protein, protein-ligand interactions
- Sequence-structure-function relationships

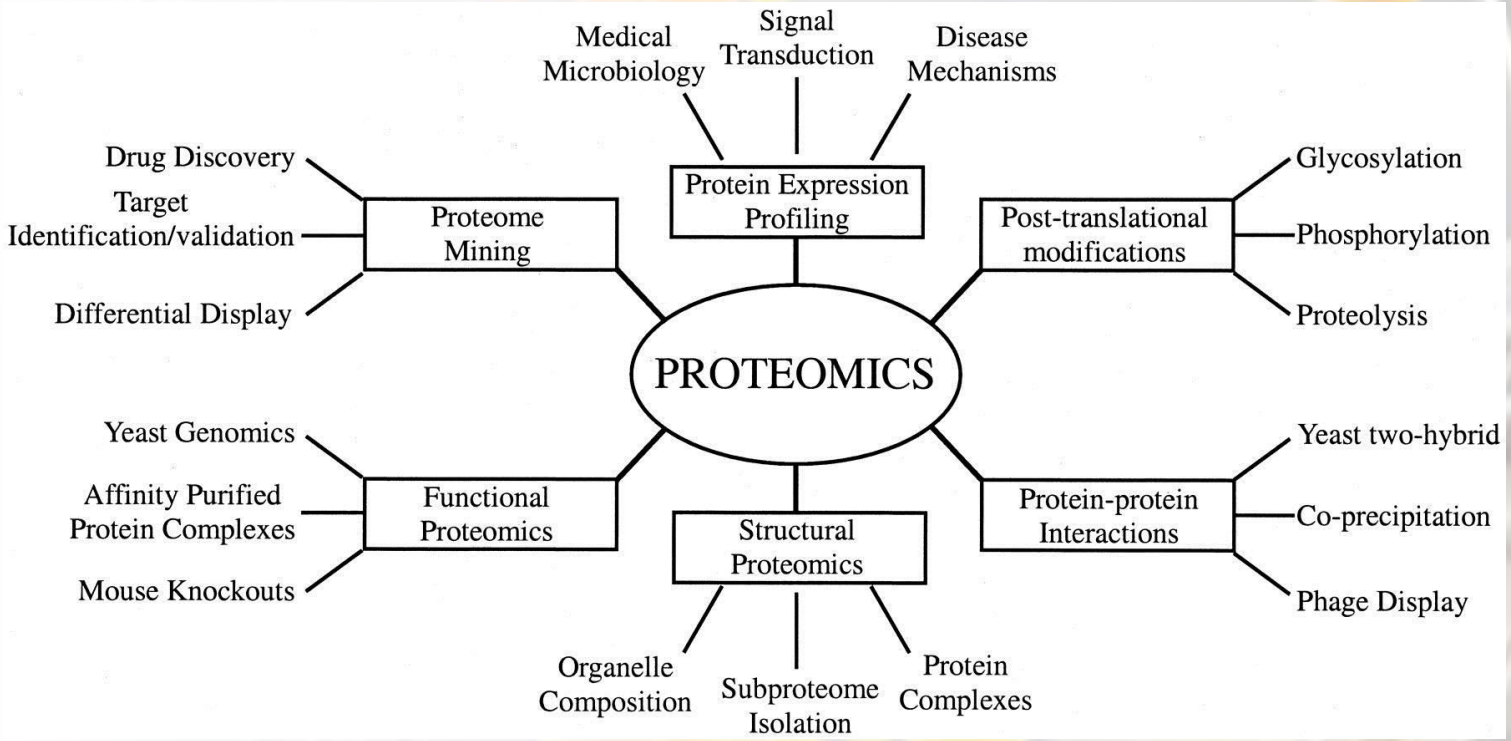
Proteomics, Types and Techniques

Expression Proteomics

- Expression proteomics refers to the quantitative study of protein expression between samples differing by some variable.
- Analysis of differential protein expression

Proteomics, Types and Techniques

Expression Proteomics



Proteomics, Types and Techniques

Study of Proteomics

- Protein profiling
- Predicting protein structure
- Protein networks

Proteomics, Types and Techniques

Protein Profiling

- Determination of the proteins that make up a given proteome

Proteomics, Types and Techniques

Protein Profiling

- Proteomes vary by cell type
- Proteomes vary by stage of cell development
- Some proteins abundant, others very rare

Proteomics, Types and Techniques

Protein Profiling Techniques

- Two-dimensional gel electrophoresis
- Chemical protein sequencing
- Protein sequencing by mass spectrometry

Proteomics, Types and Techniques

Proteins Structure Determination Techniques

- X-ray crystallography reliable but slow, not all protein crystallize
- Computer structure-prediction programs not reliable for all proteins

Proteomics, Types and Techniques

Proteomics - Conclusion

- Proteomics is the study of all the proteins found in an organism
- Different Types

Genetics and Genomics

Human Proteome - Characteristics

Human Proteome - Characteristics

Proteomics

- Proteomics is the study of proteins that are generated from the genetic code of an organism.

Human Proteome - Characteristics

Proteomics

- Proteomics differs from genomics in that chromosomes for a genome are consistent throughout a multicellular organism, protein output varies from cell to cell.

Human Proteome - Characteristics

Human Proteome is Larger than Human Genome

- The proteome is larger than the genome due to alternative splicing.
- Humans ~ 250,000 proteins. Another estimate 1,000,000 individual proteins

Human Proteome - Characteristics

Human Proteome

- Less than 2 % of human diseases result from a single gene defect
- 98 % complex diseases like cancer are reflected in a modified protein network

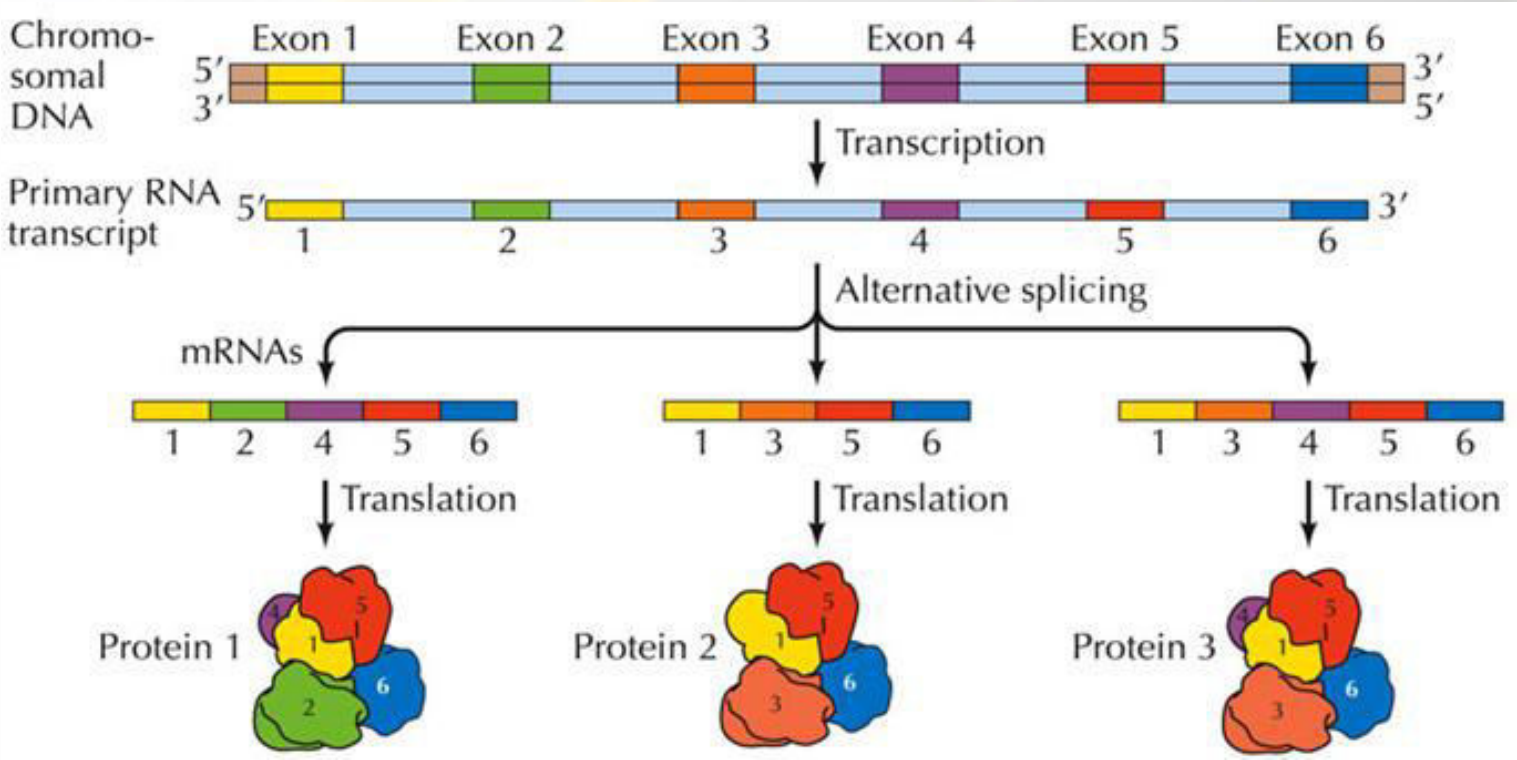
Human Proteome - Characteristics

Multiple Functions

- One protein or peptide may have multiple functions.
- Complex regulation of protein function.
- Proteins modifications

Human Proteome - Characteristics

Proteomics – More Number of Proteins than Genes



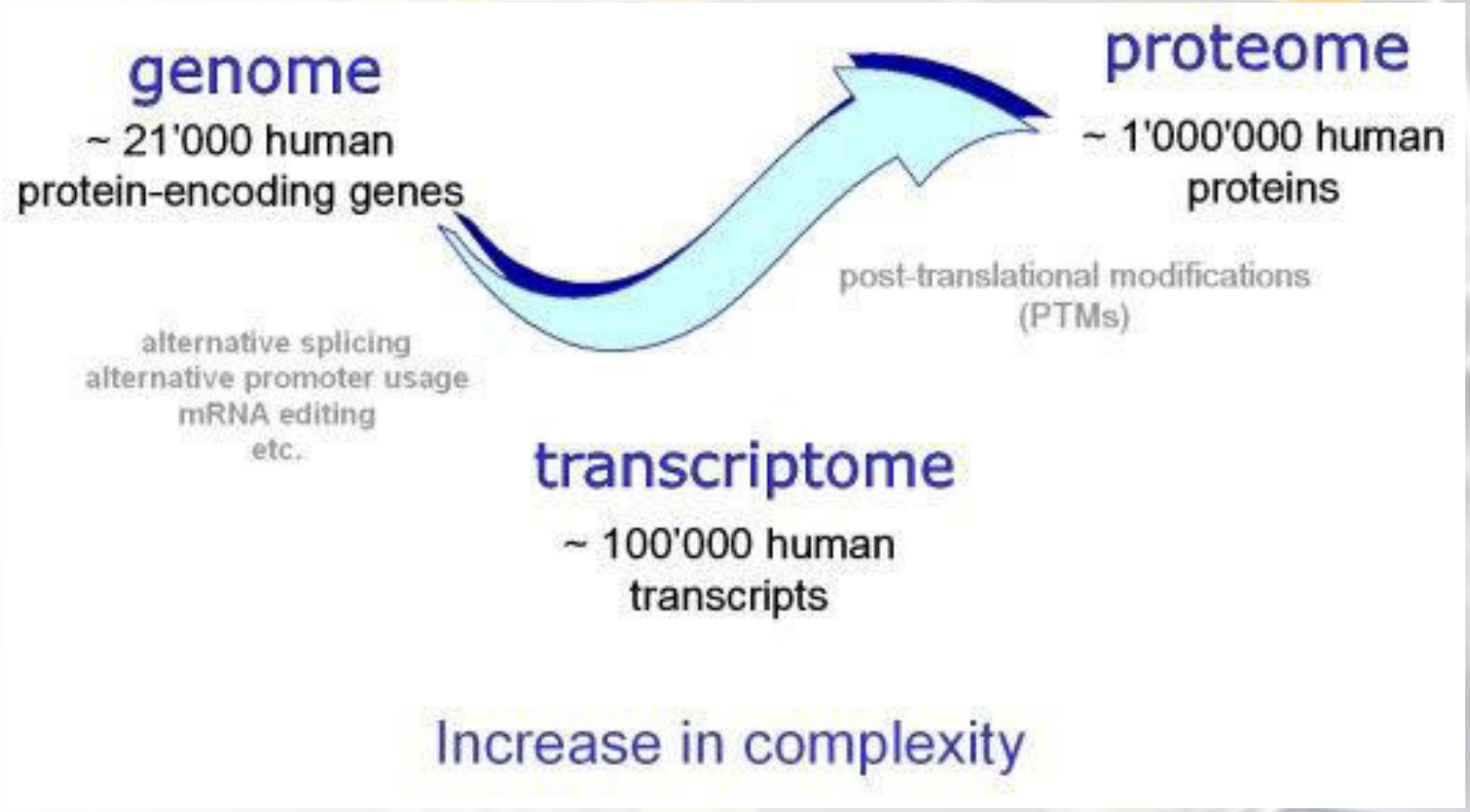
Human Proteome - Characteristics

More Complex Proteome than Invertebrates

- Proteome more complex than those of invertebrates.

Human Proteome - Characteristics

Human Genome and Proteome



Human Proteome - Characteristics

Proteome Complexity increases from Yeast to Humans

- Complexity of proteome increase from yeast to humans
- More genes
- Genes shuffling increases

Human Proteome - Characteristics

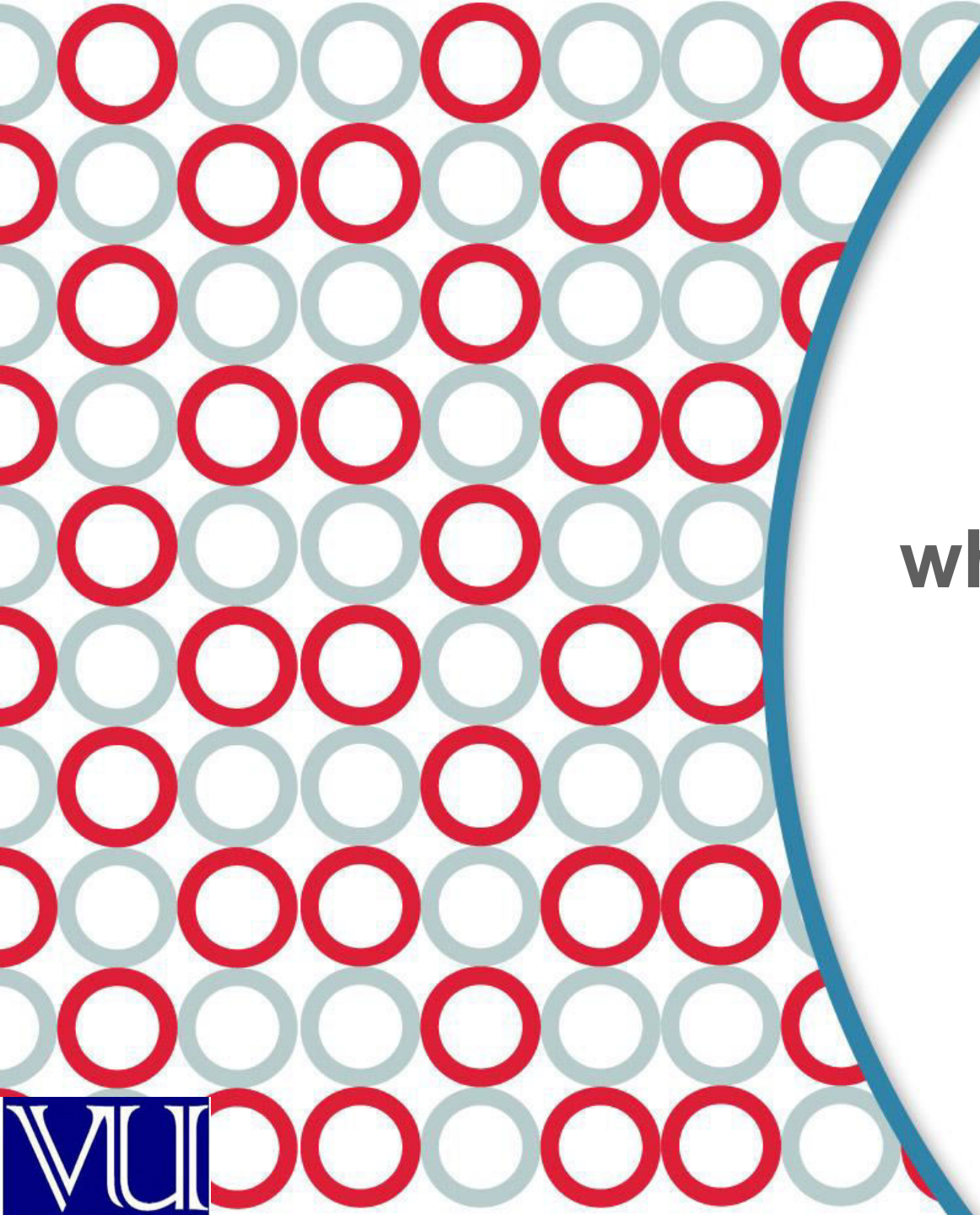
Proteome – More Chemical Modifications

- Alternative RNA Splicing –
- Humans exhibit significantly more chemical modification of proteins

Human Proteome - Characteristics

Human Genome and Human Proteome

- Human Genes ~ 21,000 - 23,000
- Human Transcripts ~ 100,000
- Human Proteins ~ 250,000- 1,000,000



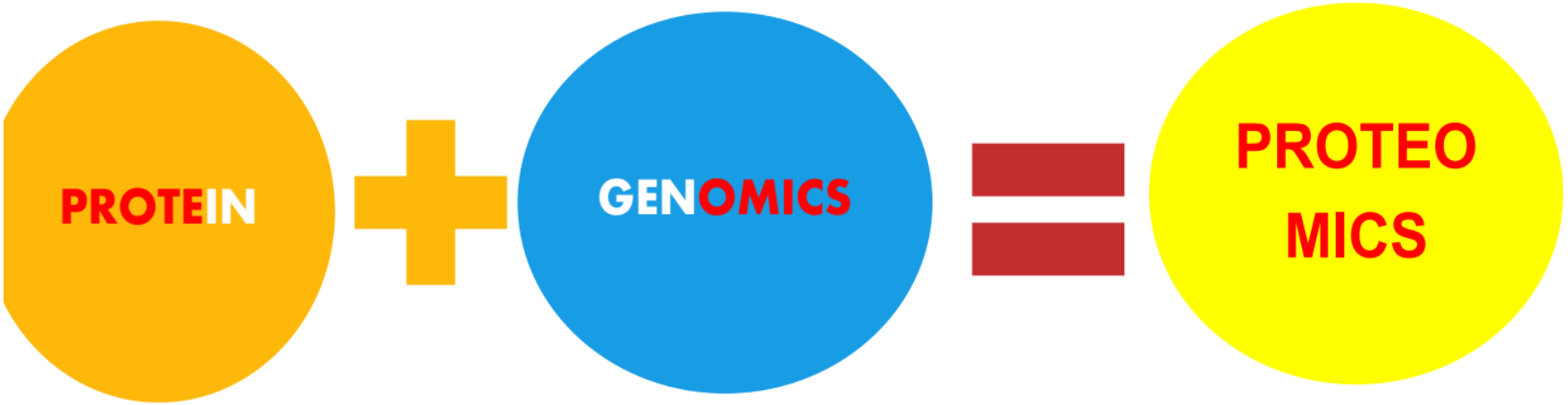
why Proteomics

Lecture-2

What do mean by proteomics?

- Proteomics is the large-scale study of proteins, usually by biochemical methods. The word proteomics has been associated
- traditionally with displaying a large number of proteins from a given
- cell line or organism on two -dimensional polyacrylamide gels

fppt.com



fppt.com

Scope of Proteomics

The identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system.

Or - A complete description of proteins expressed in any given cell at any given time

fppt.com

Why should we study Proteomics?

- directly contributes to drug development
- Verification of a gene product by proteomic methods
- Modifications of the proteins
- Protein expression level d

fppt.com

WHY PROTEOMICS?

Many types of information cannot be obtained from the study of genes alone. For example, proteins, not genes, are responsible for the phenotypes of cells. It is impossible to elucidate mechanisms of disease, aging, and effects of the environment solely by studying the genome.

fppt.com

Advantages of study of proteomics

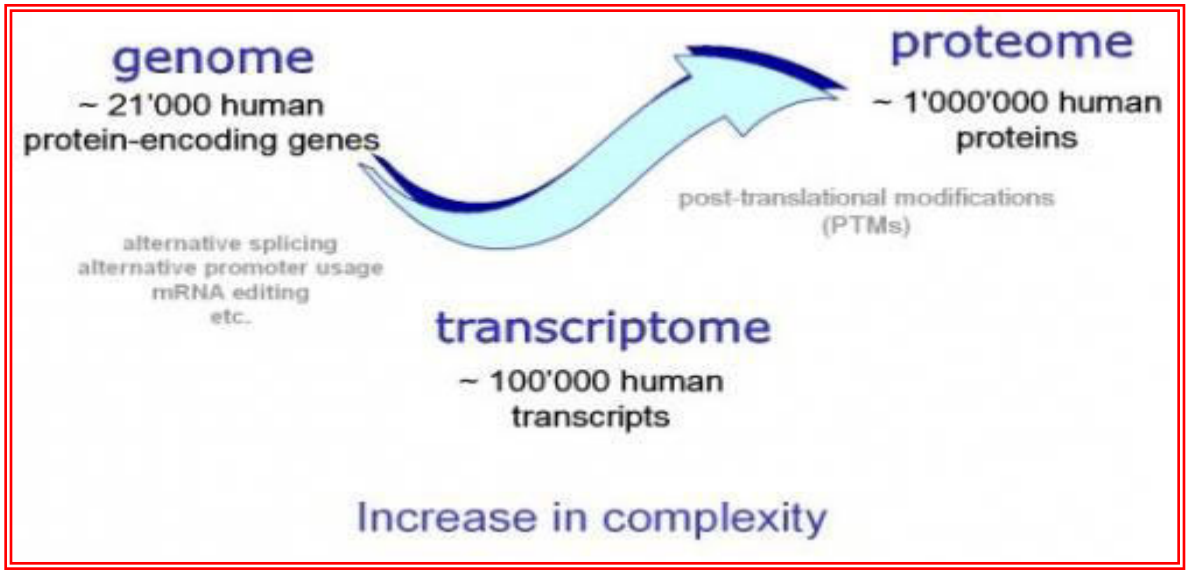
- Shows that genetic alterations are not the reason for all types of diseases
- Helps in determining the proper treatment of diseases
- With the help of three dimensional analysis of proteins we have found that HIV protease is the enzyme which is responsible for AIDS.
- One of the most important use of proteomics in diagnosis is the identification of biomarkers. The study of drugs in proteomics is called pharmacoproteomics.

fppt.com

Proteomics aims

- Genomics integrated strategies
- Study of post-translational modifications
- Identification of novel protein targets for drugs
- Analysis of tumor tissues
- Comparison between normal and diseased tissues
- Comparison between diseased and pharmacologically treated tissues

fppt.com


genome
~ 21'000 human protein-encoding genes
proteome
~ 1'000'000 human proteins
post-translational modifications (PTMs)
alternative splicing
alternative promoter usage
mRNA editing
etc.
transcriptome
~ 100'000 human transcripts
Increase in complexity

fppt.com

Limitations of Genomics Challenge of Proteomics

- co-translational modifications
  - differential mRNA splicing
- post-translational modifications (PTMs)
  - C-terminal GPI anchor
  - phosphorylation
  - sulfation
  - glycosylation
  - N-myristoylation
  - hydroxylation
  - N-methylation
  - carboxymethylation
  - signal peptidase site.....

fppt.com

Proteomics Introducción

BIO 601

Lecture-1

Biological Importance of proteins

- Cellular Function
- Structural importance
- Enzymes
- Harmones
- Transport protein
- Messenger protein
- Antibodies
- Plasma protein
- Protein in diet

fppt.com

Pharmaceutical Importance of proteins

- Proteins as pharmaceuticals
- Proteins applications
- Whey proteins health effects
- Iron chelate Protein
- Zinc chelate Protein
- Tumor markers

fppt.com

Cellular Function

- Structural support
- Biochemical reactions of cell
- Role of proteins in plasma membrane

fppt.com

Structural Support

- Microfilaments:

Distribution of Organelles

Formation of mitotic Spindle

- Actin Filaments : Cell movements
- Intermediate Filaments
- Architectural support inside cells

fppt.com

Example of some Important Harmones

Testosterone

Estrogen

Growth Hormone (GH)

Follicle Stimulating Hormone (FSH)

Thyroid Hormone (TH)

Melanocyte Stimulating Hormone(MSH)

Anti Diuretic Hormone (ADH)

Prolactin

are generally small proteins, many hormones. e.g.

Oxytocin Occurs females and stimulates uterine

Transport Proteins

- These are often Globular type of proteins
- Generally tightly packed with polar side group on the outside of molecule. e.g.

Serum albumin

Myoglobin

Hemoglobin

fppt.com

Enzymes

There are three types of enzymes

- Metabolic enzymes
- Digestive enzymes
- Food enzymes

fppt.com

Messenger Proteins

- These are generally quiet small proteins many are harmones e.g
- Oxytocin:
- Occurs in female and stimulate uterine contraction
- VESSOPRESIN ;
- major function as anti diuretic
- each of these has 9 amino acids
- Insulin : regulate the glucose metabolism

fppt.com

Antibodies

- These are protein components of immune system
- Have main function to bind antigens.

fppt.com

Plasma Proteins

- Albumin

Made mainly in liver.

Helps to keep the blood from leaking out of blood vessels.

Help to carry medicines and other substances.

Important for tissue growth and healing.

- Globulin

Made up of different proteins i.e. alpha, beta and gamma types.

Have a role in immunity.

Determines chances of developing an infection.

fppt.com

Proteins in Diet

Sources of Proteins

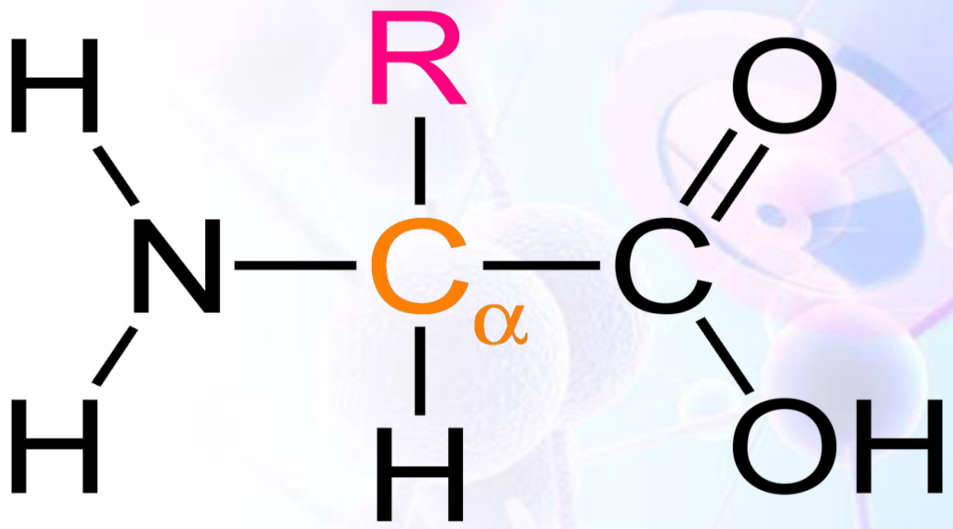
- Meat
- Beef
- Poultry
- Fish
- Egg
- Beans
- Milk

fppt.com

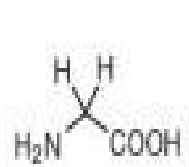
Chemical composition of proteins

- Proteins are polymers of amino acids.
- They range in size from small to very large.
- All the proteins are made up of Twenty different types of amino acids. So these amino acids are called standard amino acids.

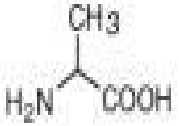
Chemical composition of proteins



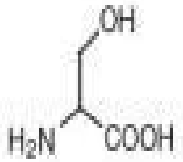
Chemical composition of proteins



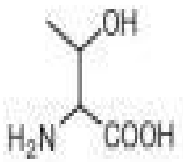
Glycine (Gly, G)



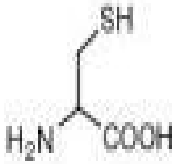
Alanine (Ala, A)



Serine (Ser, S)



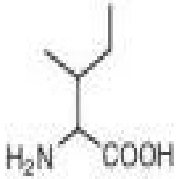
Threonine (Thr, T)



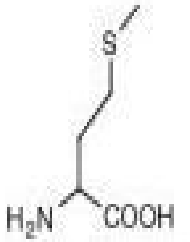
Cysteine (Cys, C)

Valine (Val, V)

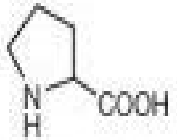
Leucine (Leu, L)



Isoleucine (Ile, I)

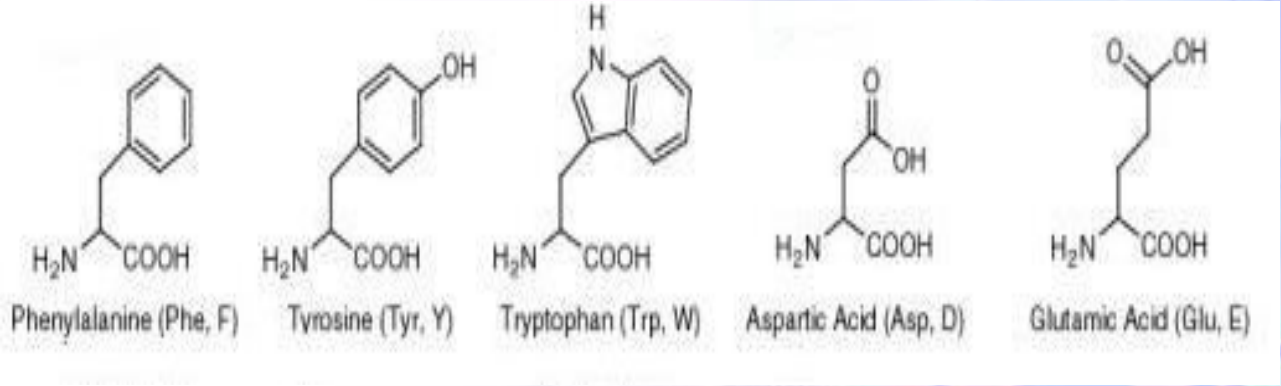


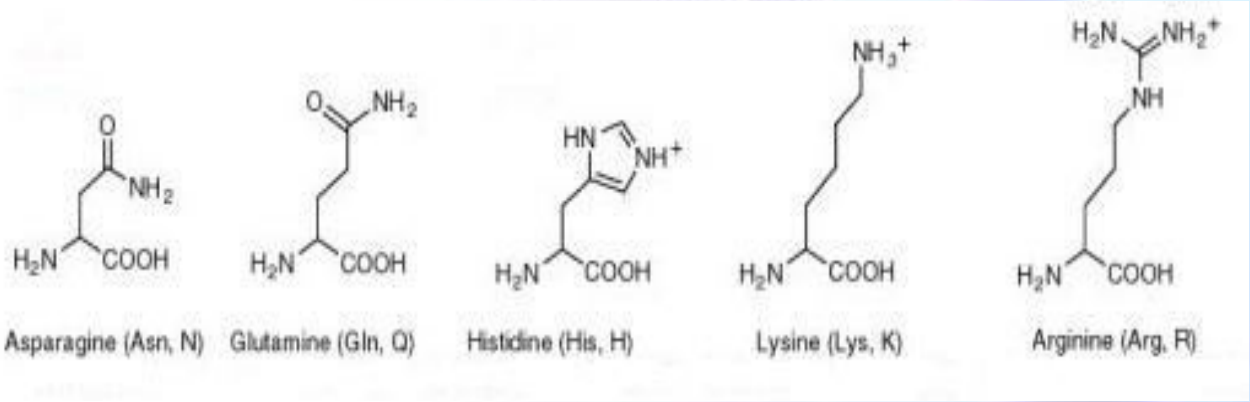
Methionine (Met, M)



Proline (Pro, P)

Chemical composition of proteins

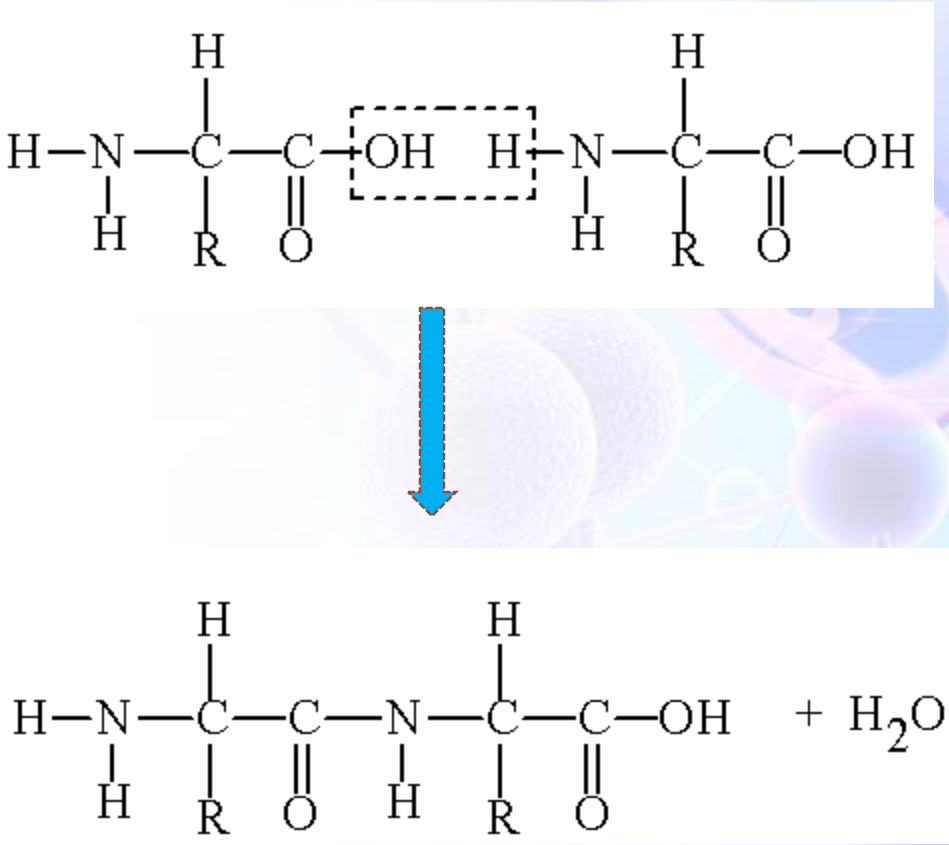




Chemical composition of proteins

- In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond.

Chemical composition of proteins



Chemical composition of proteins

- Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides.

Chemical composition of proteins

Primary structure of proteins

- Primary structure or covalent structure of protein refers to the amino acid sequence of its polypeptide chain.
- Each type of protein has a unique amino acid sequence.

Primary structure of proteins

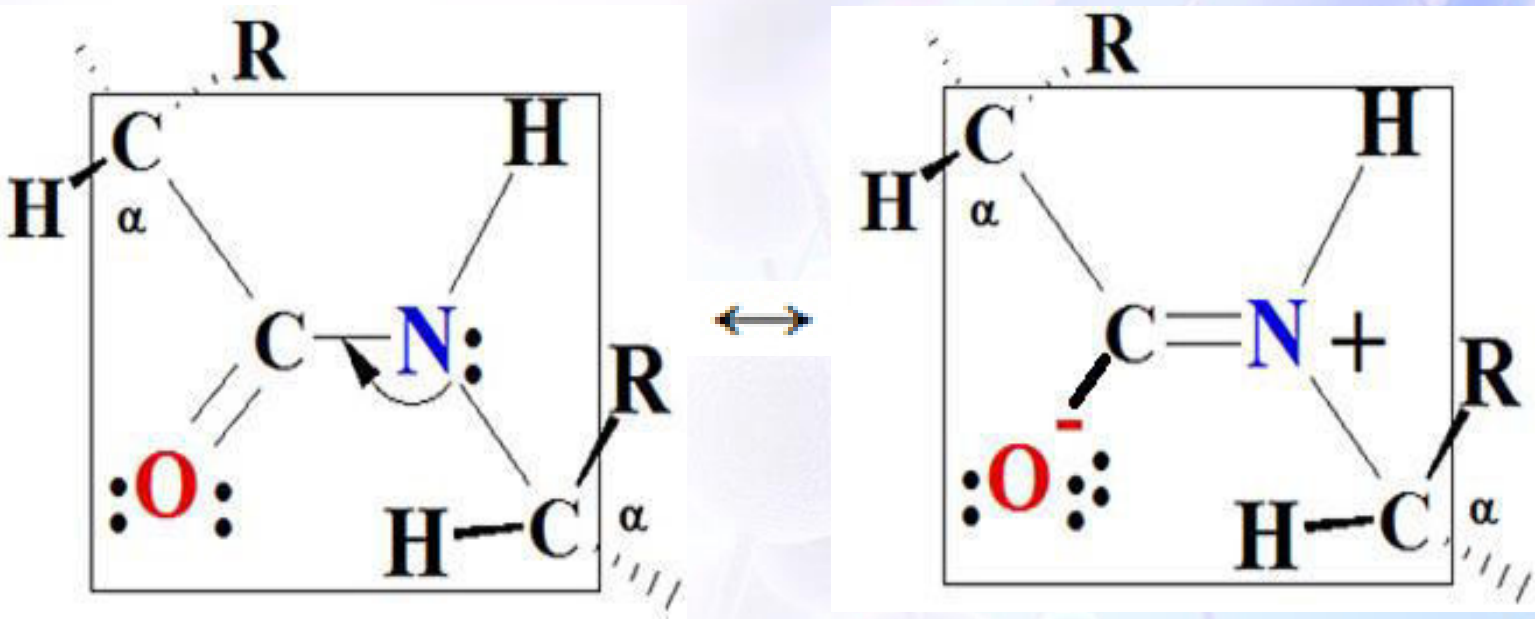
Peptide Bond Is Rigid and Planar

- Linus Pauling and Robert Corey carefully analyzed the peptide bond.
- Their findings laid the foundation for our present understanding of protein structure.

Primary structure of proteins

- They demonstrated that the peptide C - N bond is somewhat shorter than the C - N bond in a simple amine.

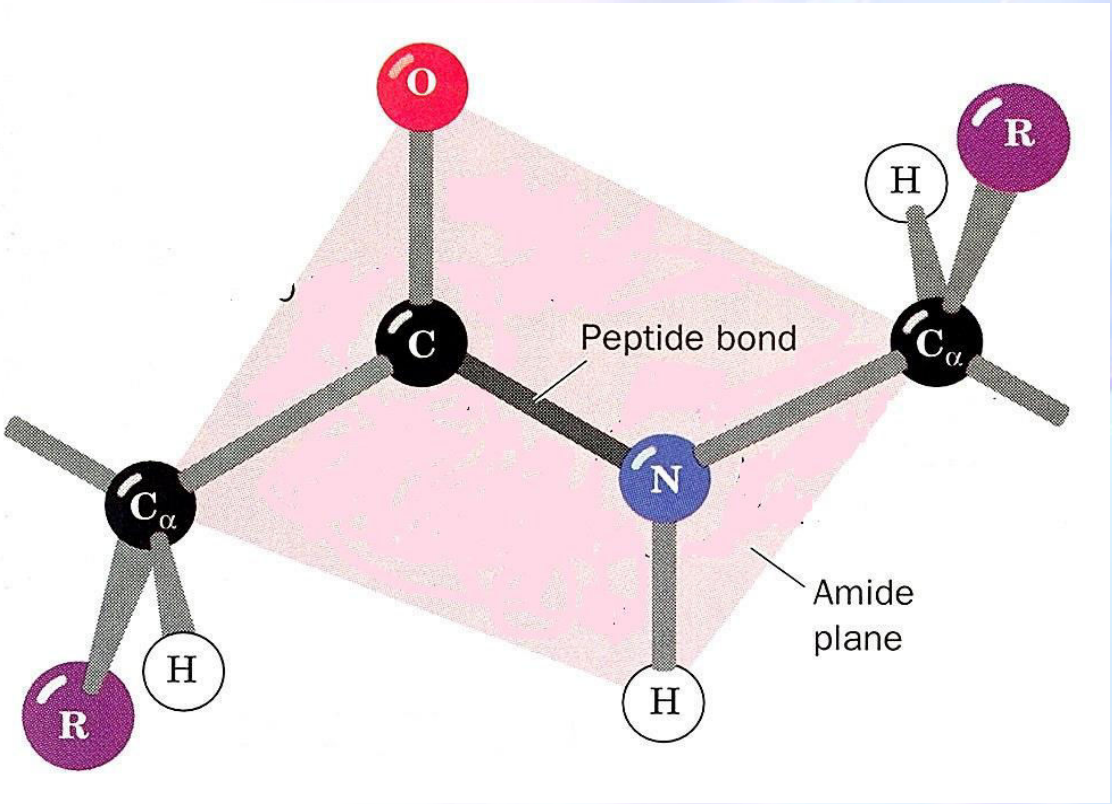
Primary structure of proteins



Primary structure of proteins

- The six atoms of the peptide group are co-planar i.e., lie in a single plane, with the oxygen atom of the carbonyl group and the hydrogen atom of the amide nitrogen trans to each other.

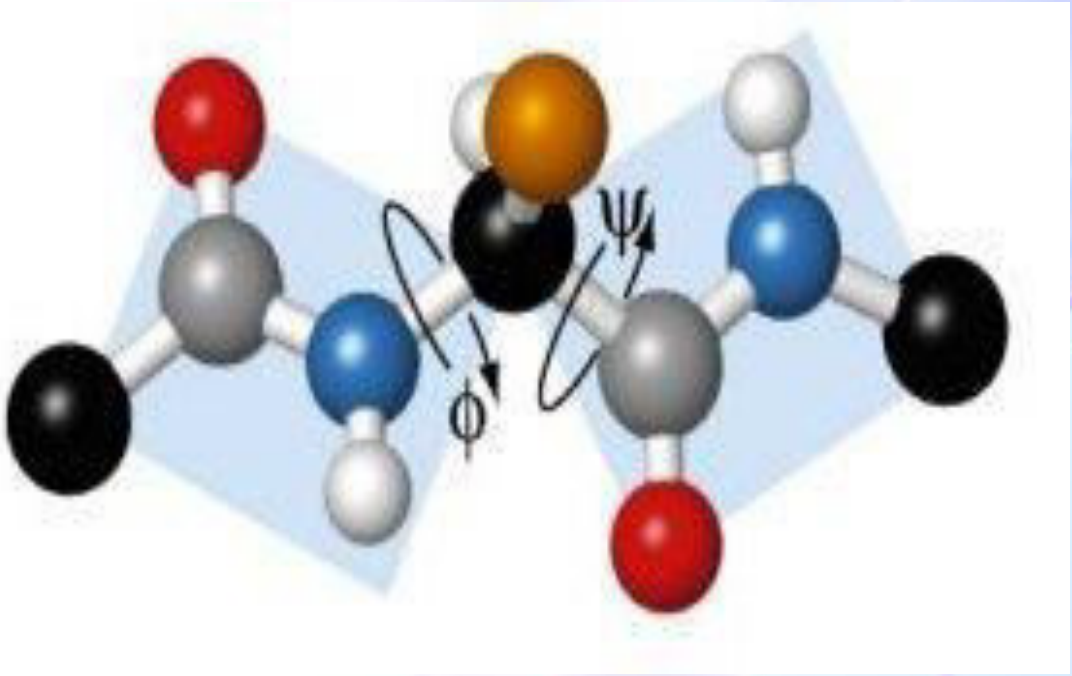
Primary structure of proteins



Primary structure of proteins

- Pauling and Corey concluded that the peptide C - N bonds are unable to rotate freely because of their partial double-bond character.
- Rotation is permitted about the N - αC and the αC - C bonds.

Primary structure of proteins



Primary structure of proteins

- The bond angles resulting from rotations at C are labelled ϕ (phi) for the N - αC bond and ψ(psi) for the αC - C bond.
- In principle, ϕ and ψ can have any value between +180 & -180.

Secondary structure of proteins

- Secondary structure of proteins refers to the local conformation of some part of a polypeptide.

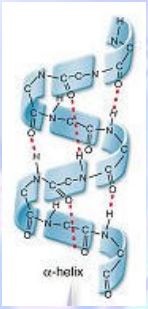
Secondary structure of proteins

Secondary structure of proteins

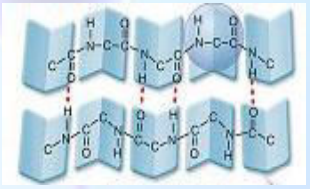
- A few types of secondary structures are particularly stable and occur widely in proteins.

Secondary structure of proteins

- The most prominent are:-
  - α-helix



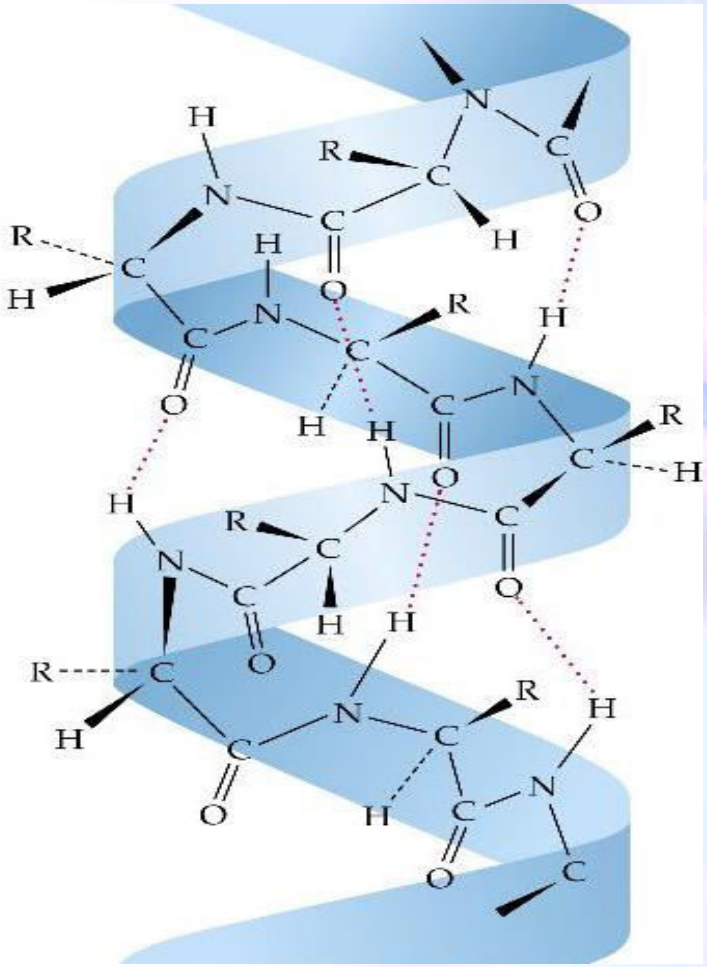
  - β- conformations.



α- Helix

- The simplest arrangement which a polypeptide chain could assume with its rigid peptide bonds is a helical structure, which Pauling and Corey called the **α-helix**.

α- Helix



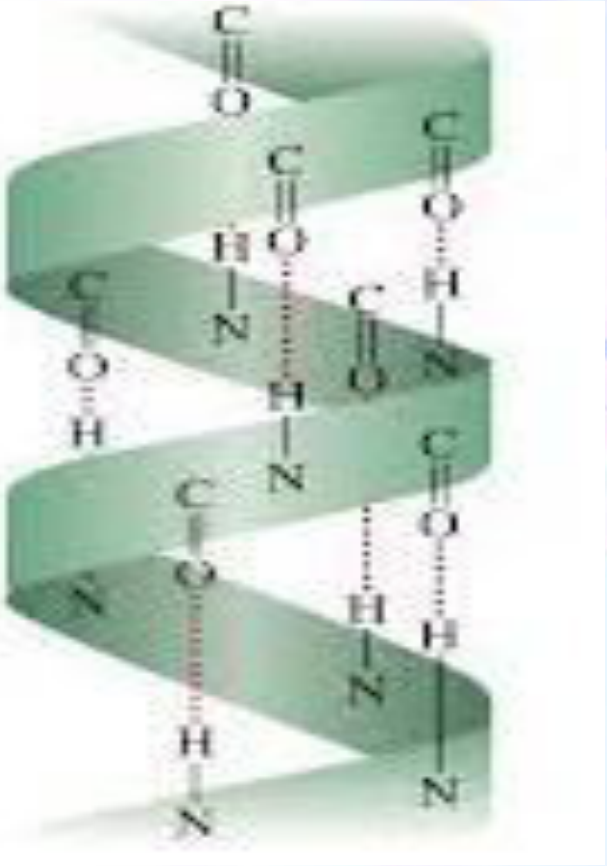
α- Helix

- The helical twist of the *α*-helix found in all proteins is right-handed.
- The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis.

α- Helix

- The amino acid residues in an helix have conformations with psi = –45 to –50 and phi = – 60.
- An helix makes optimal use of internal hydrogen bonds.

α- Helix



α- Helix

- About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure.

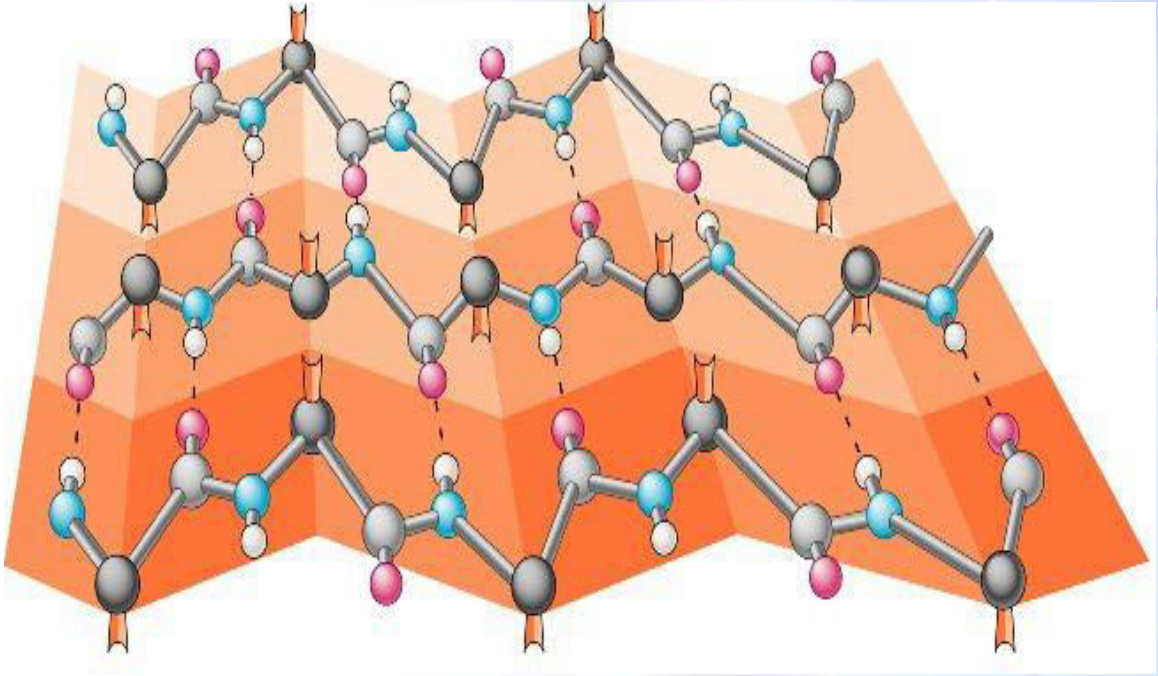
β- Pleated Sheets

- Pauling and Corey predicted a second type of secondary structure which they called **β-sheets.**
- This is a more extended conformation of polypeptide chains.

β- Pleated Sheets

- The backbone of the polypeptide chain is extended into a zigzag structure.
- The zigzag polypeptide chains are arranged side by side to form a structure resembling a series of pleats.

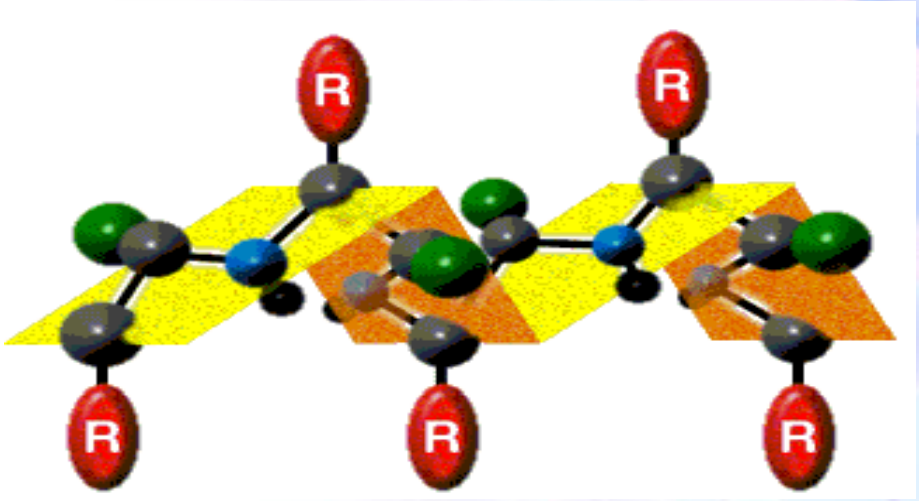
β- Pleated Sheets



β- Pleated Sheets

- The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions.

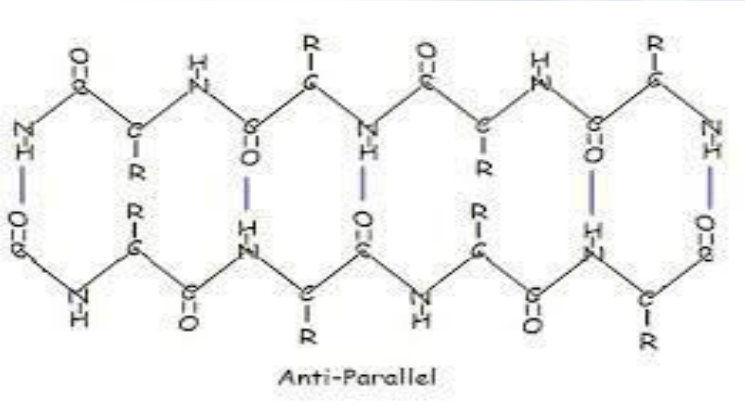
β- Pleated Sheets



β- Pleated Sheets

- Hydrogen bonds are formed between adjacent segments of polypeptide chain.
- The adjacent polypeptide chains in a sheet can be either parallel or antiparallel.

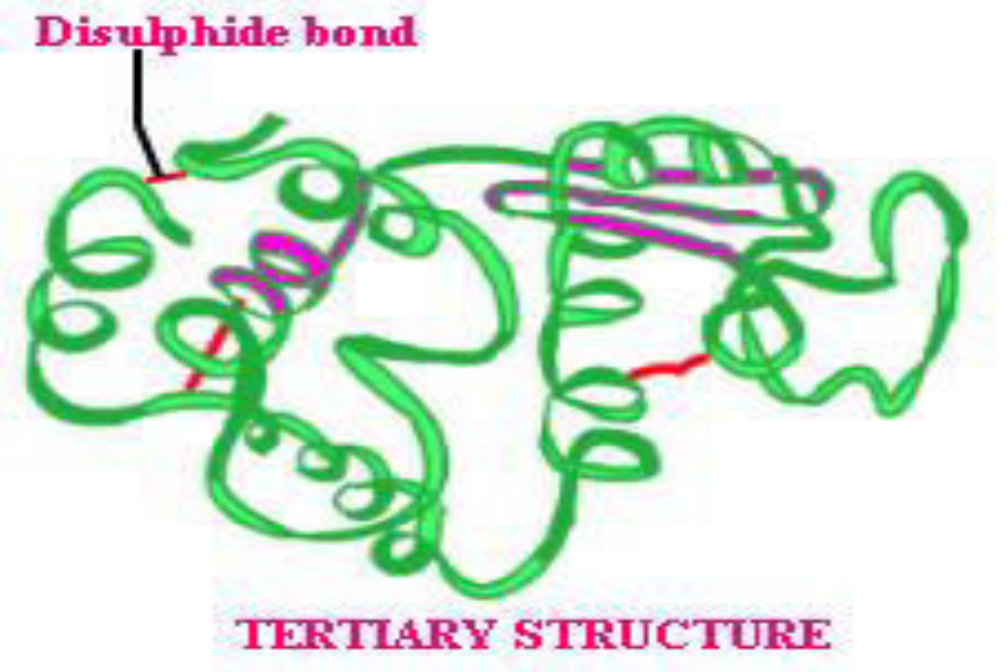
β- Pleated Sheets



Tertiary Structure of Proteins

- The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's **tertiary structure.**

Tertiary Structure of Proteins



Tertiary Structure of Proteins

- It includes longer-range aspects of amino acid sequence.
- Amino acids that are far apart in the polypeptide chain may interact within the completely folded structure of a protein.

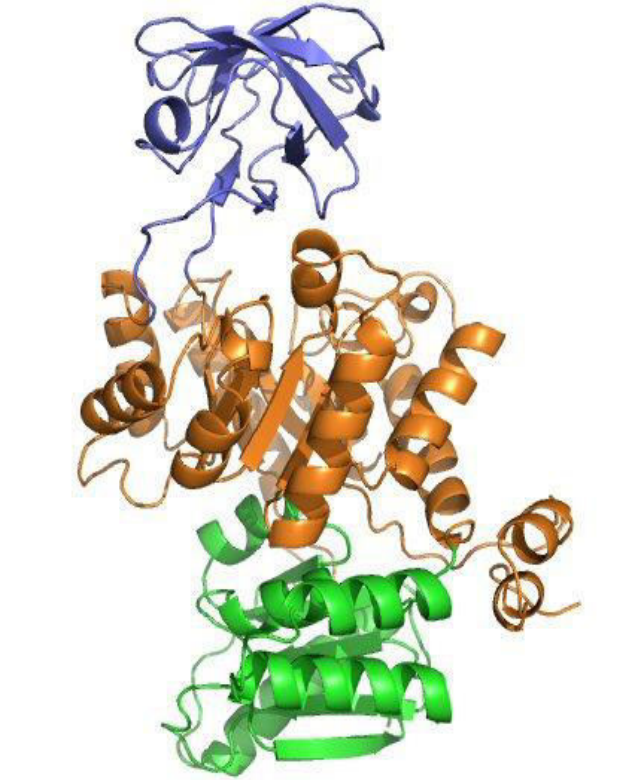
Tertiary Structure of Proteins

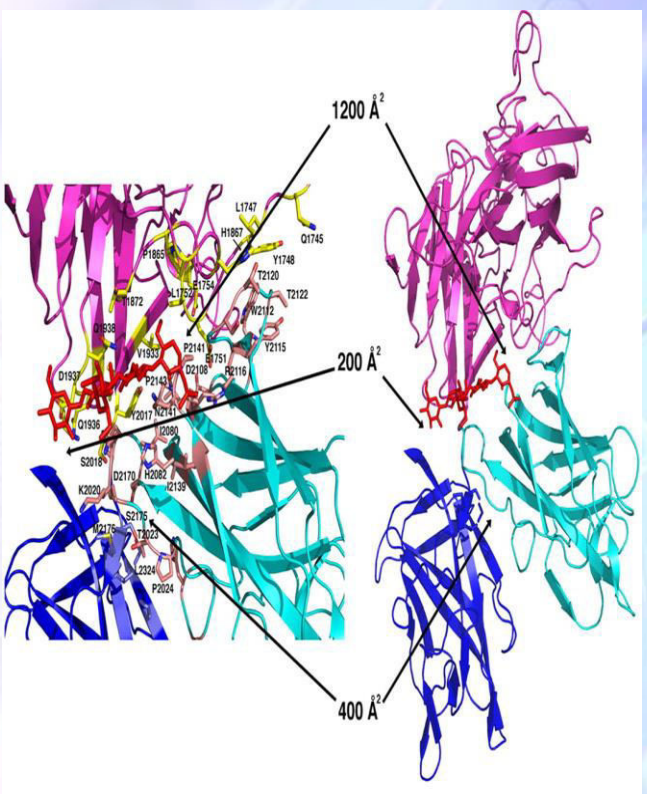
- Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak interactions (and sometimes by covalent bonds) between the segments.

Tertiary Structure of Proteins

- Large polypeptide chains usually fold into two or more globular clusters known as **domains,** which often give these proteins a bi- or multilobal appearance.

Tertiary Structure of Proteins





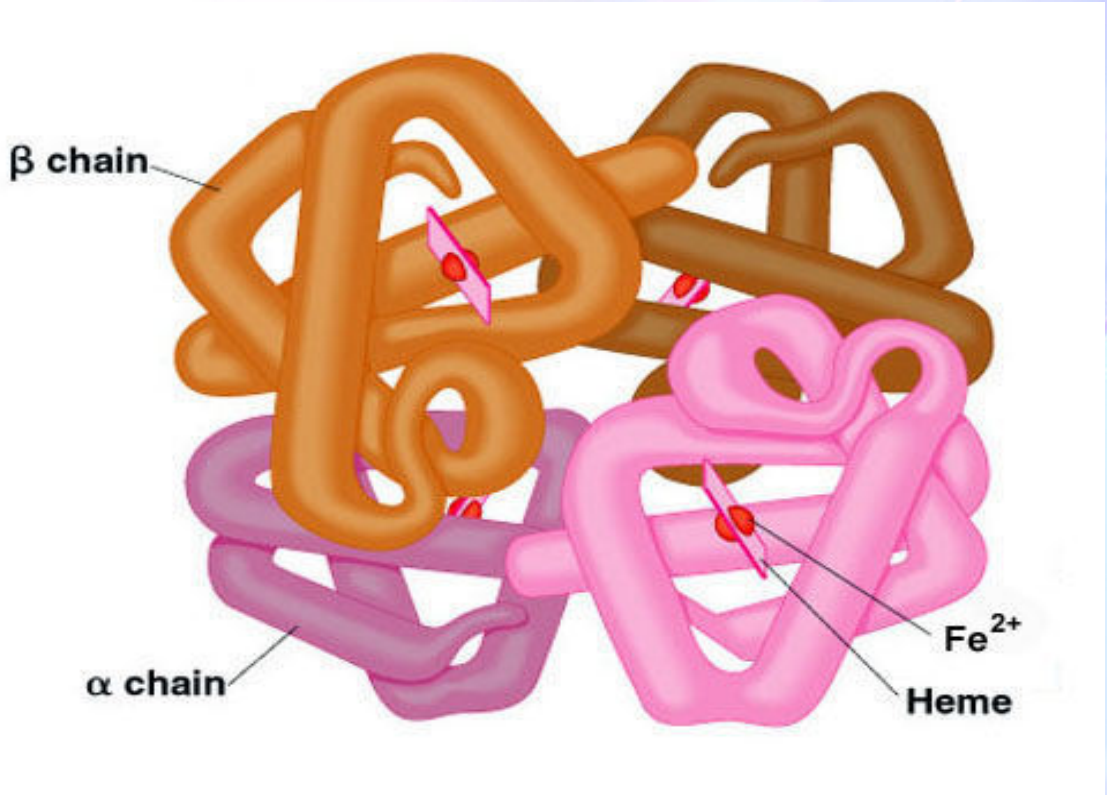
Quaternary Structure of Proteins

- Some proteins contain two or more separate polypeptide chains or subunits, which may be identical or different.
- The spatial arrangement of these subunits is known as a protein's quaternary structure.

Quaternary Structure of Proteins

- A multi-subunit protein is also referred to as a multimer.
- A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer.

Quaternary Structure of Proteins



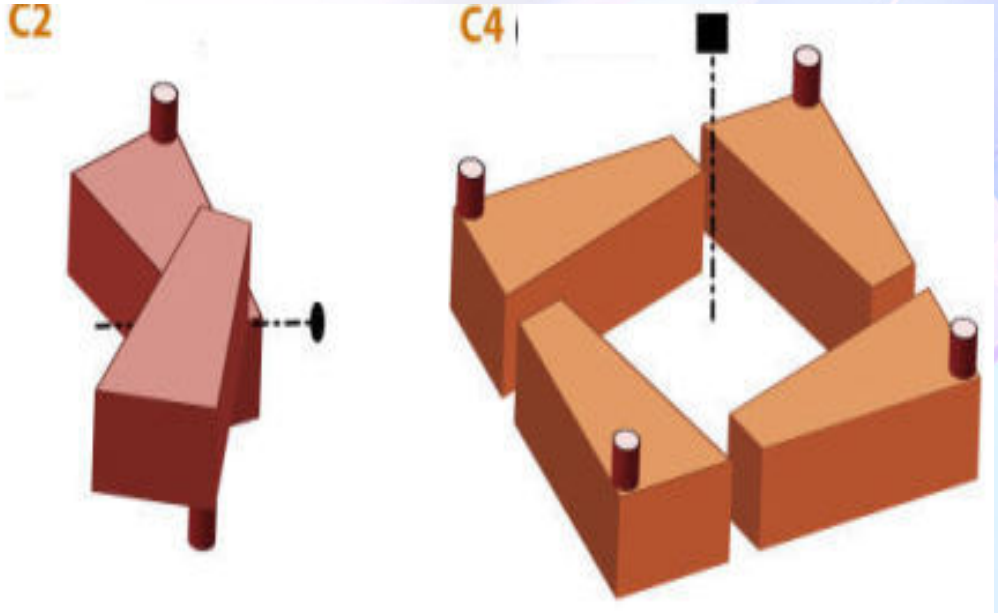
Quaternary Structure of Proteins

- Identical subunits of multimeric proteins are generally arranged in a symmetric patterns.
- Oligomers can have either rotational symmetry or helical symmetry.

Quaternary Structure of Proteins

- There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis.

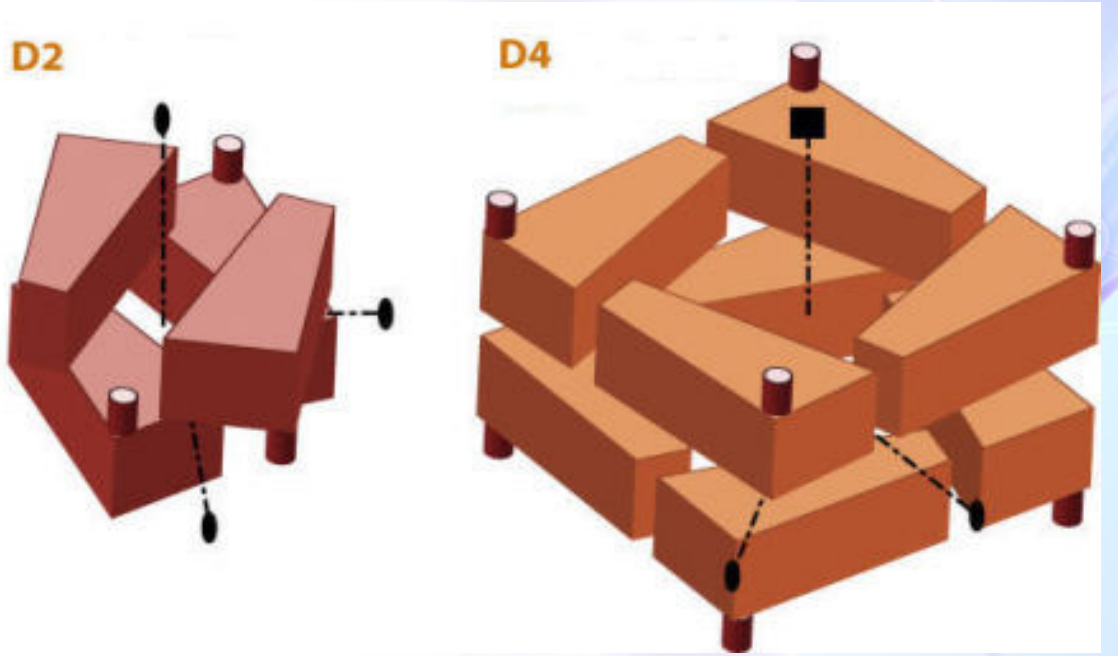
Quaternary Structure of Proteins



Quaternary Structure of Proteins

- A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present.

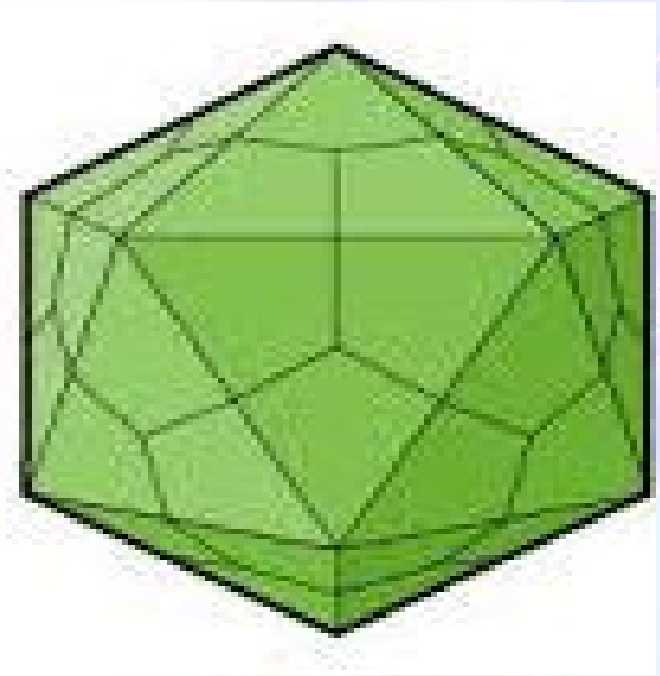
Quaternary Structure of Proteins



Quaternary Structure of Proteins

- More complex rotational symmetries include icosahedral symmetry.
- An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces.

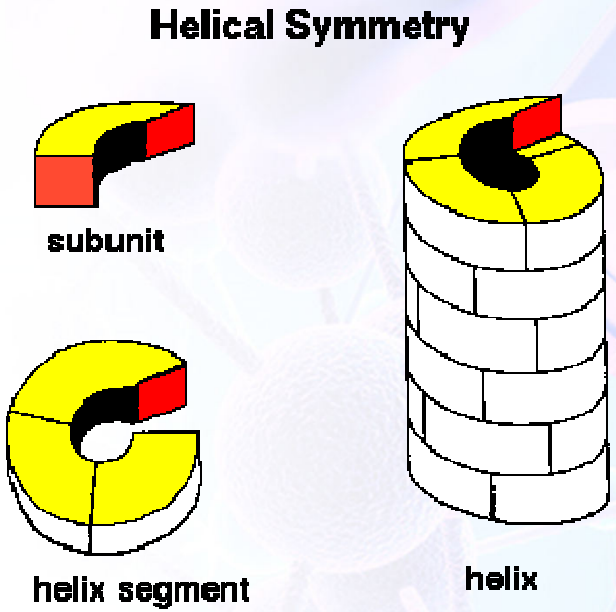
Quaternary Structure of Proteins



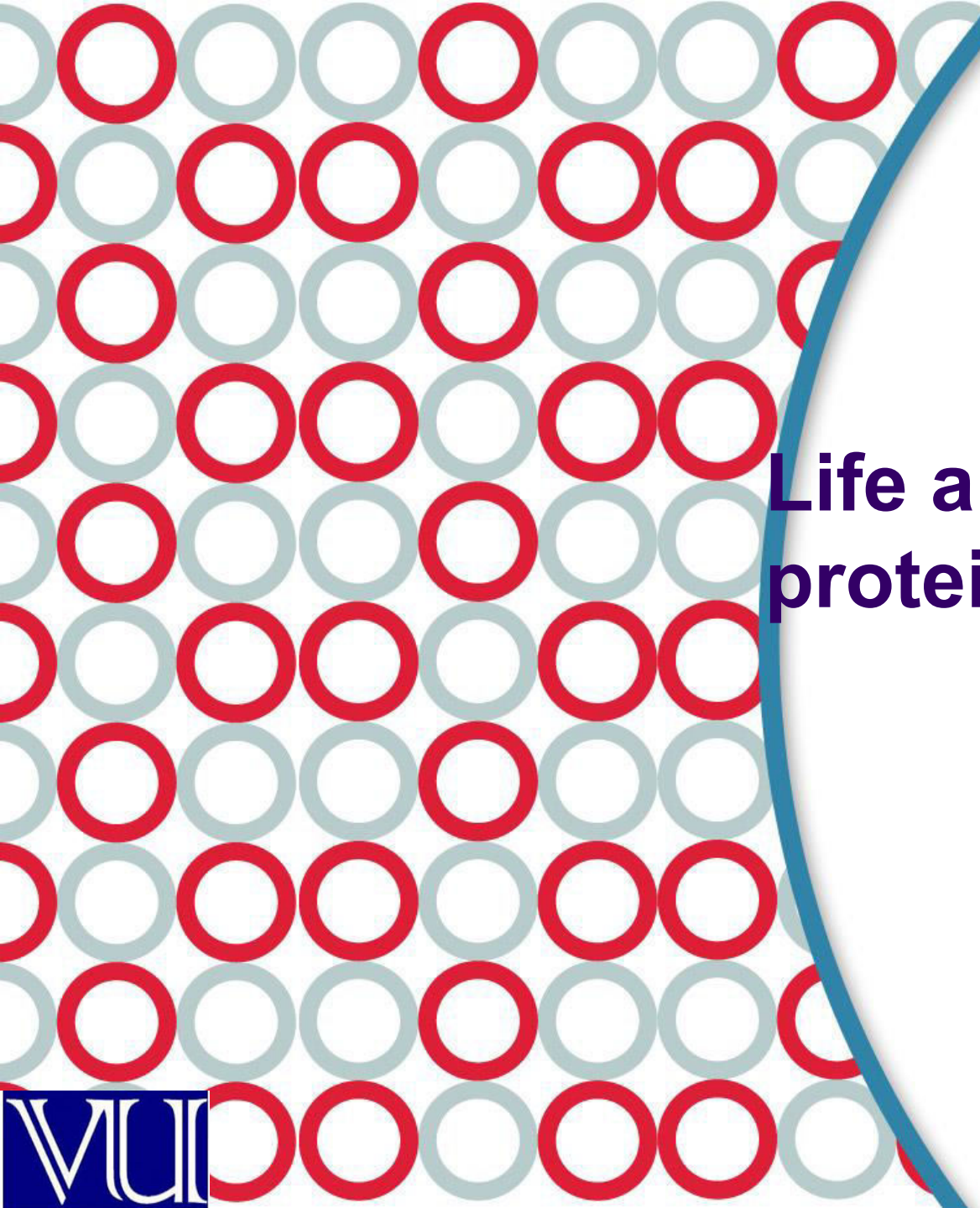
Quaternary Structure of Proteins

- The other major type of symmetry found in oligomers is helical symmetry.

Quaternary Structure of Proteins



(adapted from Voet and Voet, 1990)



Life and death of a protein

Lecture-1

Life and Death of a Protein

Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes.

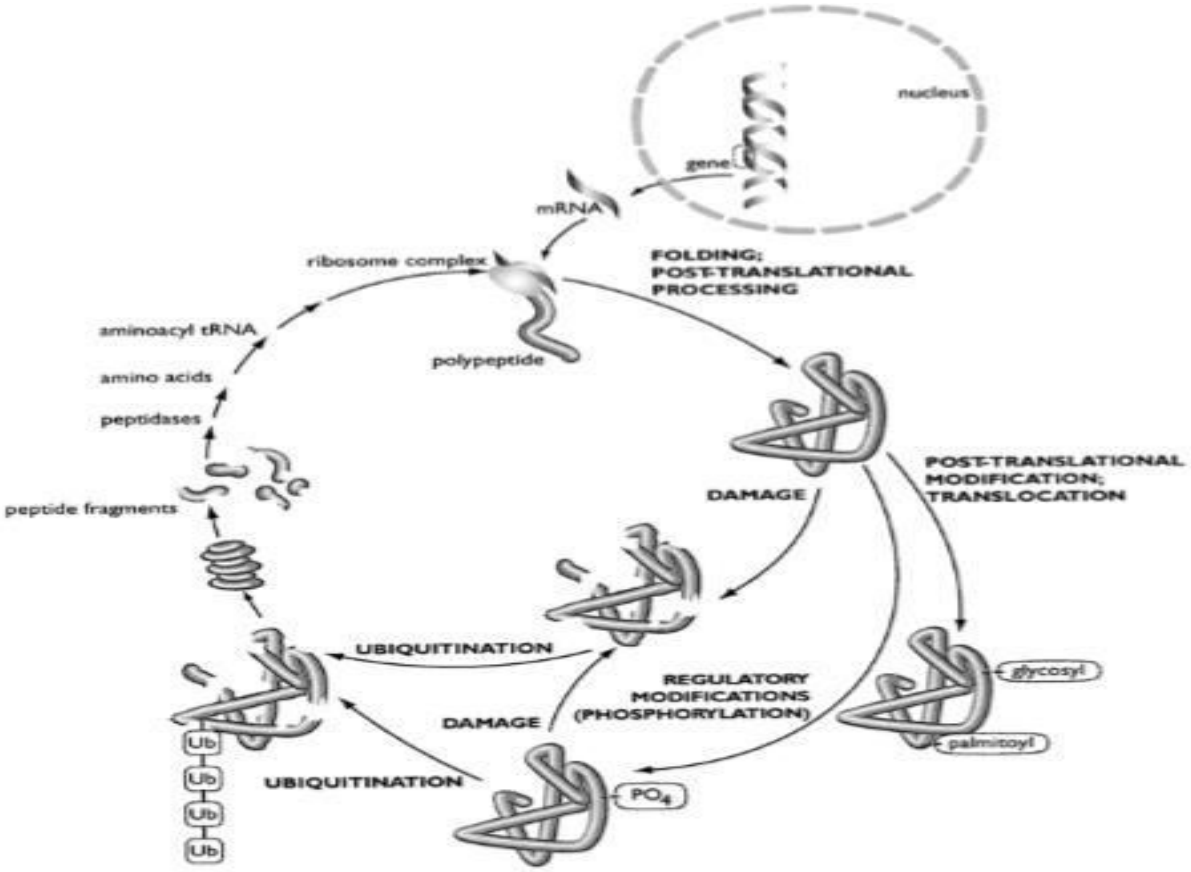
In most cases, the initial polypeptide-translation product undergoes some type of modification before it assumes its functional role in a living system.

These changes are broadly termed “posttranslational modifications” and encompass a wide variety of reversible and irreversible chemical reactions.

Approximately 200 different types of posttranslational modifications have been reported. Some of these are summarized in Fig. 1, which depicts the life cycle of a prototypical protein.

fppt.com

Life Cycle of the Cell



fppt.com

Modifications during Protein Cycle

Modifications those occur early in the life of the protein

- Carboxylation of glutamate residues
- Removal of the N-terminal methionine
- Glycosylation
- Addition of Prosthetic groups
- Formation of multisubunit complexes
- Prenylation of cysteine residues assists anchoring of proteins in or on membranes.

These more or less “permanent” modifications and transport ultimately result in the delivery of functional proteins to specific locations in cells.

fppt.com

Cont…

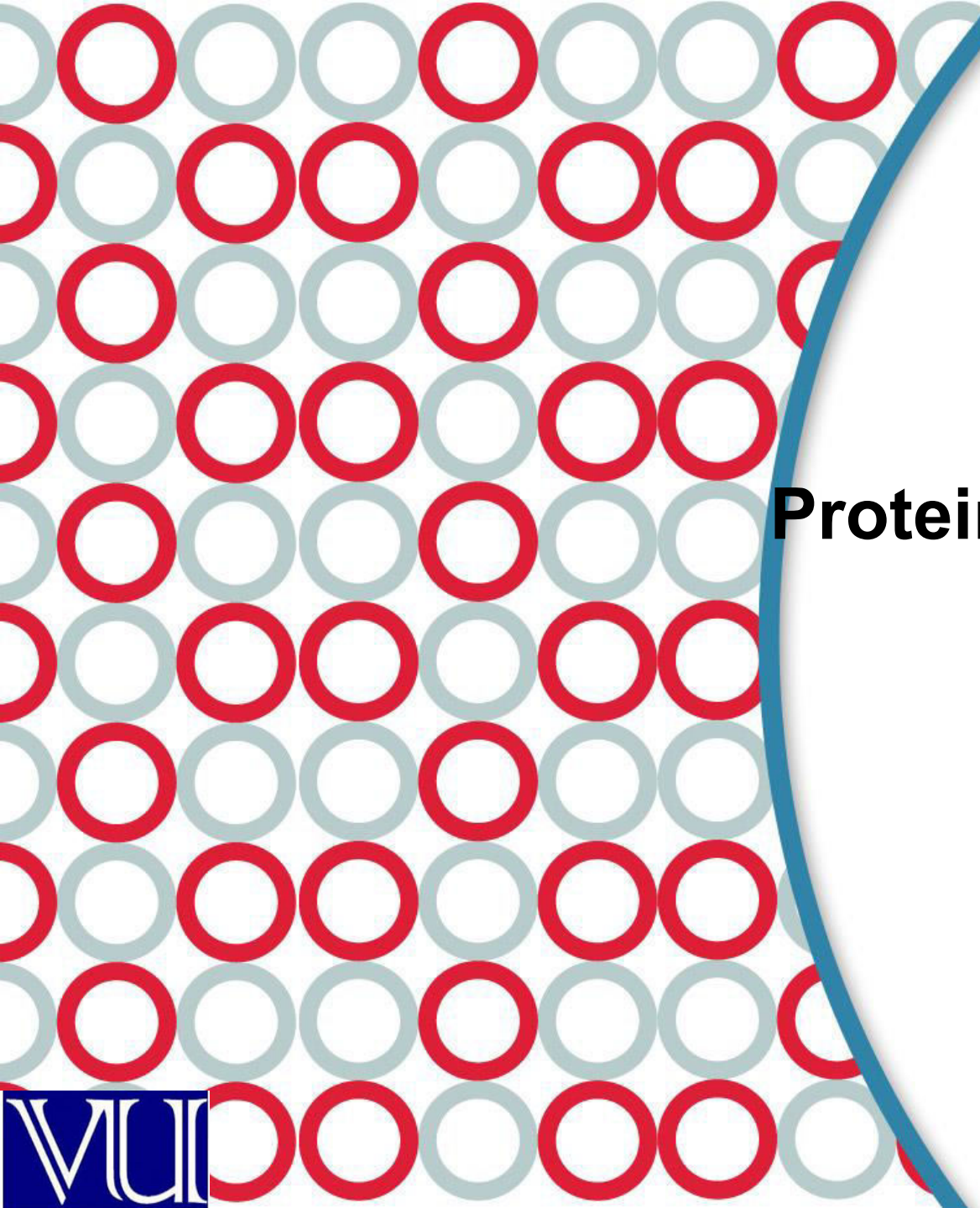
- The activities of many proteins are then controlled by posttranslational modifications.
- The most prominent and best-understood of these is phosphorylation of serine, threonine, or tyrosine residues.
- Phosphorylation may activate or inactivate enzymes, alter proteinprotein interactions and associations, change protein structures, and target proteins for degradation.
- Protein phosphorylation regulates protein function in diverse contexts and appears to be a key switch for rapid on-off control of signaling cascades, cell-cycle control, and other key cellular functions.

fppt.com

Degradation of Proteins

- Protein modifications appear to be critical to initiating processes that ultimately degrade proteins.
- Phosphorylation of some proteins is rapidly followed by conjugation with ubiquitin, which leads to degradation by the 26S proteasomal complex.
- There evidently are other stimuli for protein ubiquitination and turnover, including oxidative damage and other protein modifications.
- Proteins also undergo degradation by lysosomal enzymes.
- Any protein may be present in many forms at any one time in a cell.
- Collectively, the proteome of a cell comprises all of these many forms of all expressed proteins. This certainly makes the proteome bewilderingly complex.

fppt.com



Proteins as Modular Stru

BIO 601

Lecture-1

Proteins as Modular Structures

Segments of amino acid sequences can be considered as functional building blocks or modules.

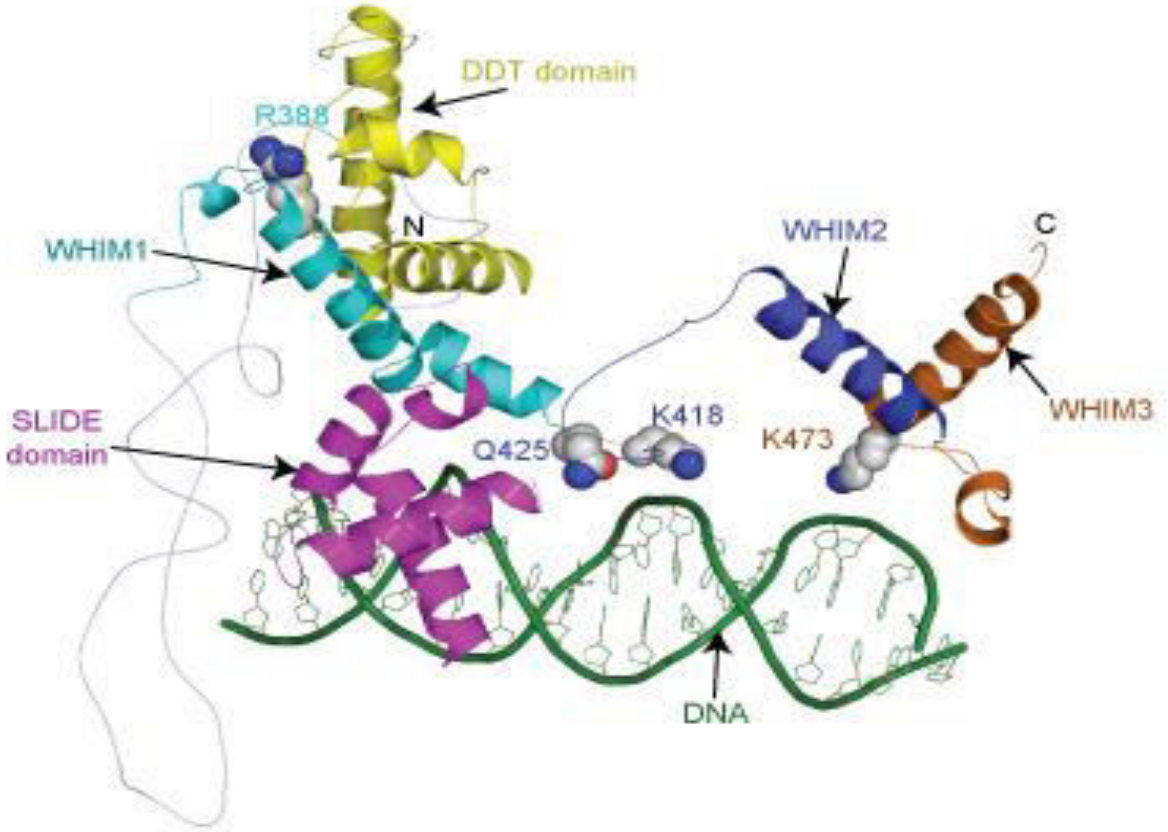
The modular units in proteins that confer specific properties and functions are referred to as “motifs” or “domains”.

Motifs and domains are recognizable sequences that confer similar properties or functions when they occur in a variety of proteins.

In some cases, amino acid sequences within motifs and domains are highly conserved and do not vary from protein to protein.

In other cases, some key amino acids occur in a reproducible relationship to each other in a sequence, even though various substitutions in other amino acids occur

fppt.com


DDT domain
R388
N
WHIM1
WHIM2
C
SLIDE
domain
K418
Q425
K473
WHIM3
DNA

fppt.com

Proteins as Modular Structures

Longer amino acid sequences often form domains, which confer specific properties or functions on a protein.

Some domain structures refer simply to sequences that confer a bulk physical property to a segment of the polypeptide, such as transmembrane domains, which simply form helices that span a lipid bilayer membrane.

Other domain structures provide hydrogen bonding or other contacts for key enzyme substrates or prosthetic groups.

In many cases, domains are made up of combinations of units of secondary structure, such as helix-loop-helix domains.

fppt.com

Genetics and Genomics

Type of Proteins/ Families

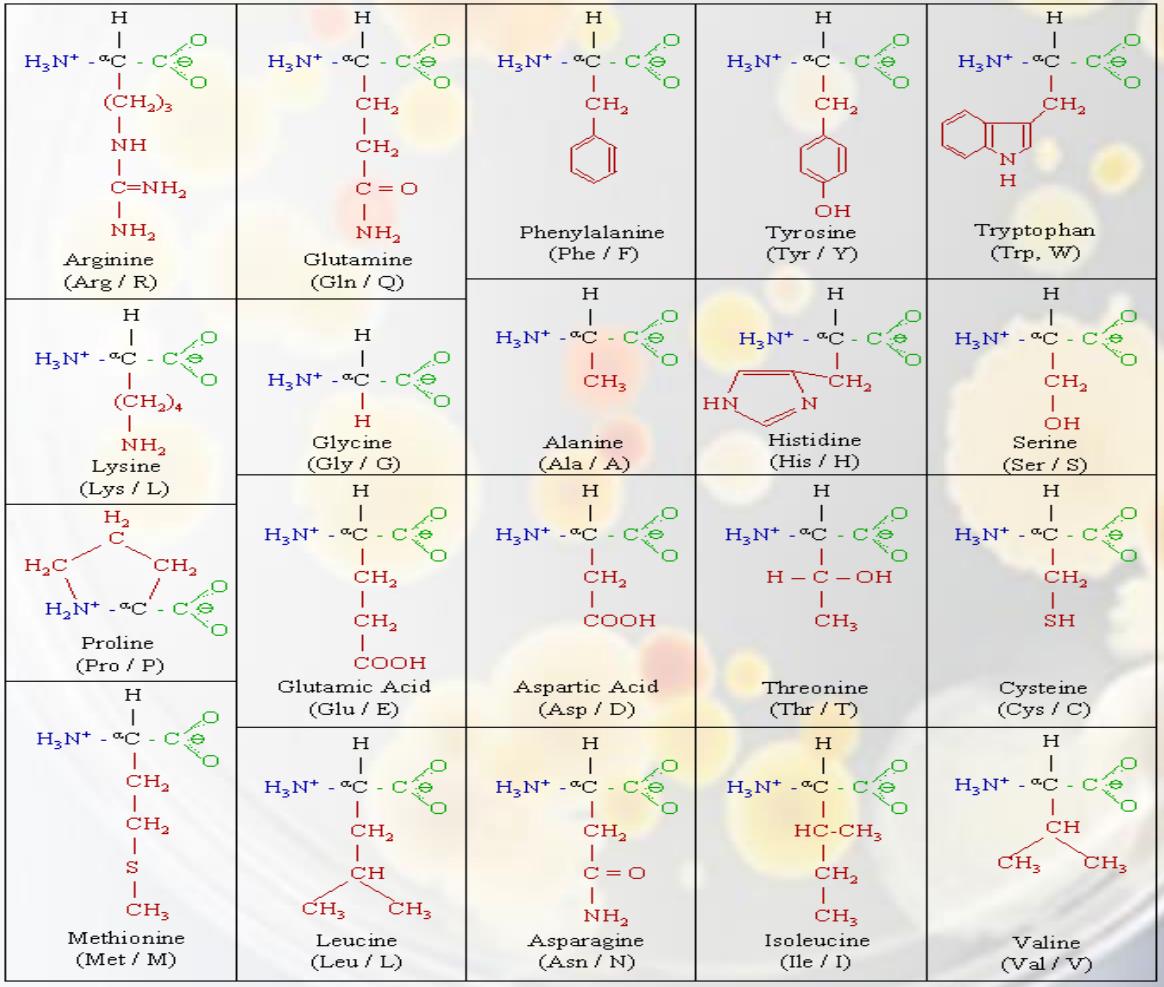
Type of Proteins/Families

Protein Structure

- Amino acids
  - 20 amino acids
  - Hydrophobic / hydrophylic
  - Charged / neutral
- Functions
  - Enzymes
  - Structure protein
  - Channel
  - Other functions

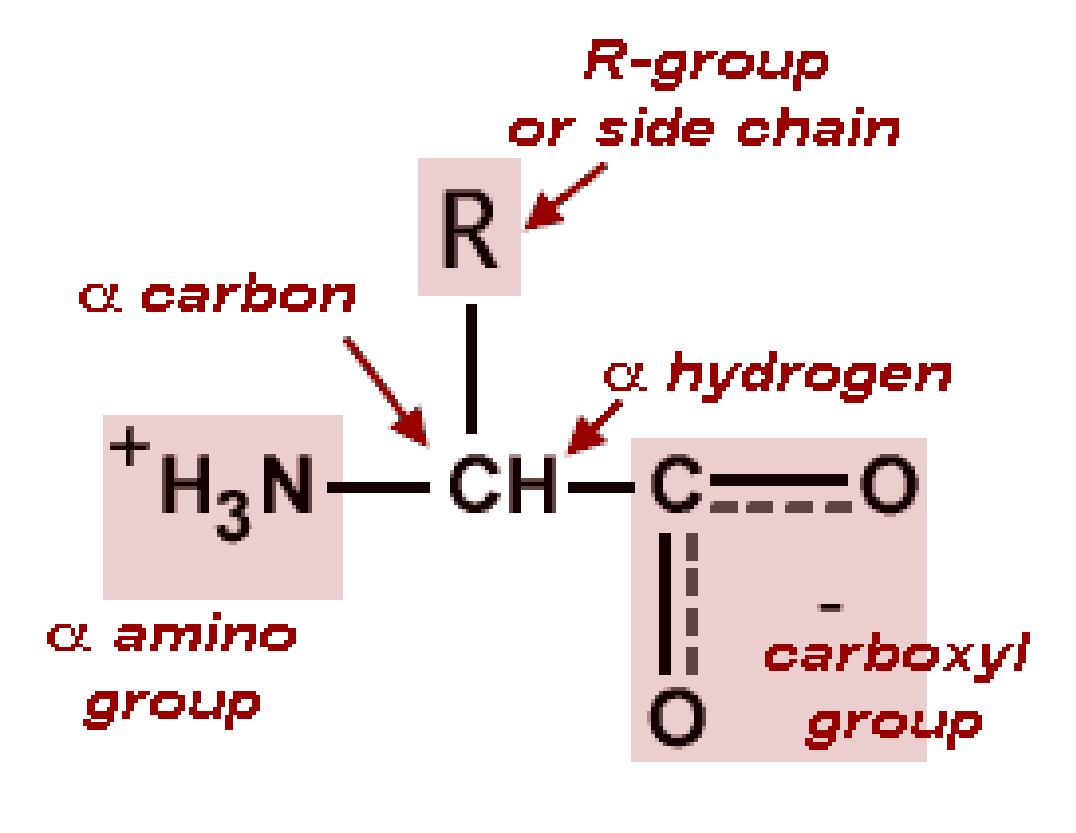
Type of Proteins/Families

Amino Acids: Building Blocks of Proteins



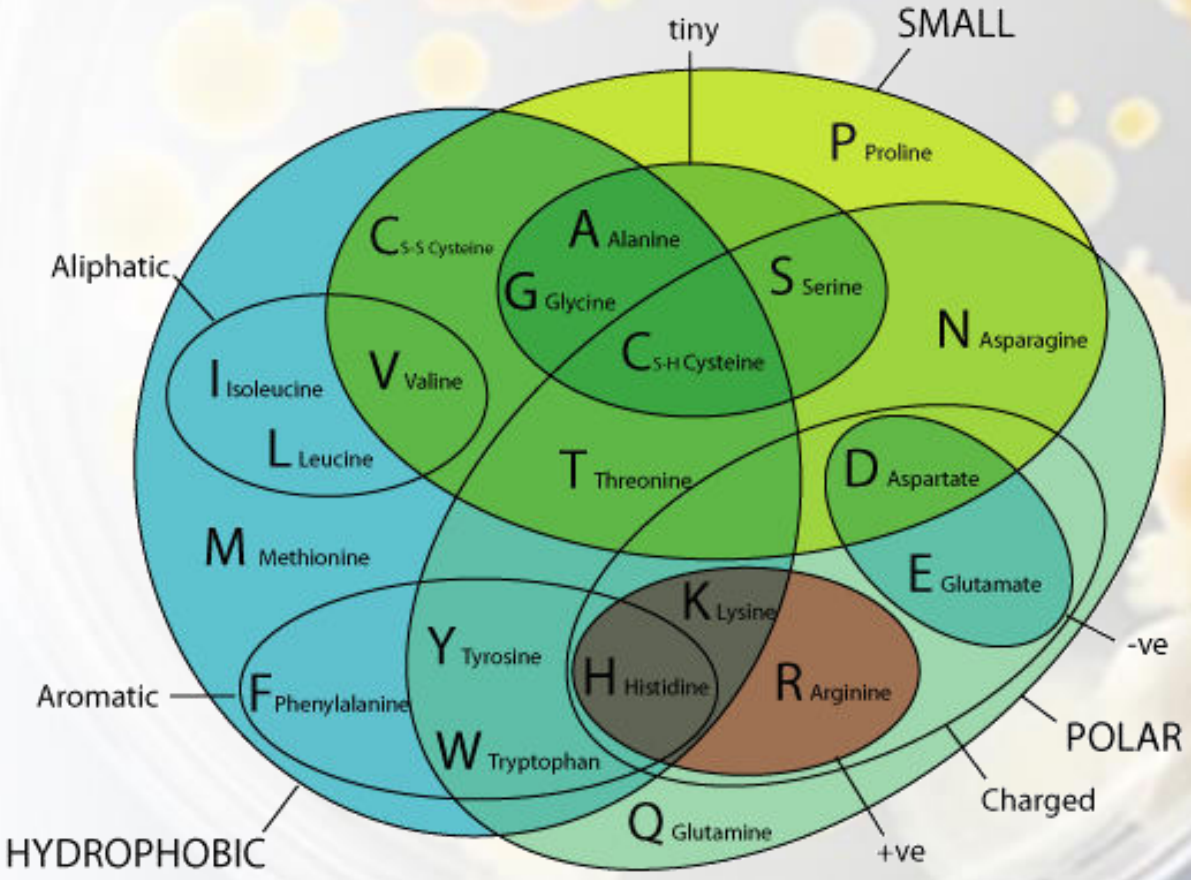
Type of Proteins/Families

Protein Structure: Amino Acid



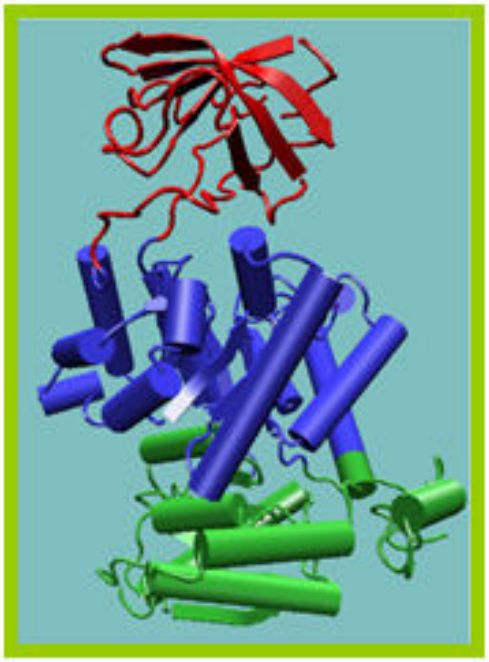
Type of Proteins/Families

Protein Structure: Types of Amino Acids



Type of Proteins/Families

Protein Domains: Domains are units of compact structure, function and evolution and folding



SKSHSEAGSAFIQTQQLHAAMADTFLEHMCRLDIDSAP
ITARNTGIICTIGPASRSVETLKEMIKSGMNVARMNFS
HGTHEYHAETIKNVRTATESFASDPILYRPVAVALDTK
GPEIRTGLIKGSGTAEVELKKGATLKITLDNAYMAACD
ENILWLDYKNICKVVEVGSKVYVDDGLISLQVKQKGPD
FLVTEVENGGFLGSKKGVNLPGAAVDLPAVSEKDIQDL
KFGVDEDVDMVFASFIRKAADVHEVRKILGEKGKNIKI
ISKIENHEGVRRFDEILEASDGIMVARGDLGIEIPAEK
VFLAQKMIIGRCNRAGKPVICATQMLESMIKKPRPTRA
EGSDVANAVLDGADCIMLSGETAKGDYPLEAVRMQHLI
AREAEAAMFHRKLFEELARSSSHSTDLMEAMAMGSVEA
SYKCLAAALIVLTESGRSAHQVARYRPRAPIIAVTRNH
QTARQAHLYRGIFPVVCKDPVQEAWAEDVDLRVNLAMN
VGKAAGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP

Type of Proteins/Families

Proteins

- Different method of categorizing proteins
- Three major categories
- Fibrous proteins
- Globular proteins
- Complexes with multiple components including proteins

Type of Proteins/Families

Fibrous Proteins – Cytoskeletal Proteins

- Actin
- Coronin
- Dystrophin
- Keratin
- Tubulin

Type of Proteins/Families

Fibrous Proteins – Extracellular Matrix Proteins

- Collagen
- Elastin
- Fibronectin

Type of Proteins/Families

Globular Proteins - Major Types

- Plasma proteins
- Hemoproteins
- Cell adhesion
- Transmembrane transport proteins
- Hormones and growth factors
- Receptors

Type of Proteins/Families

Globular Proteins - Major Types

- DNA-binding protein
- Immune system proteins
- Nutrient storage/transport
- Chaperone proteins
- Enzymes

Type of Proteins/Families

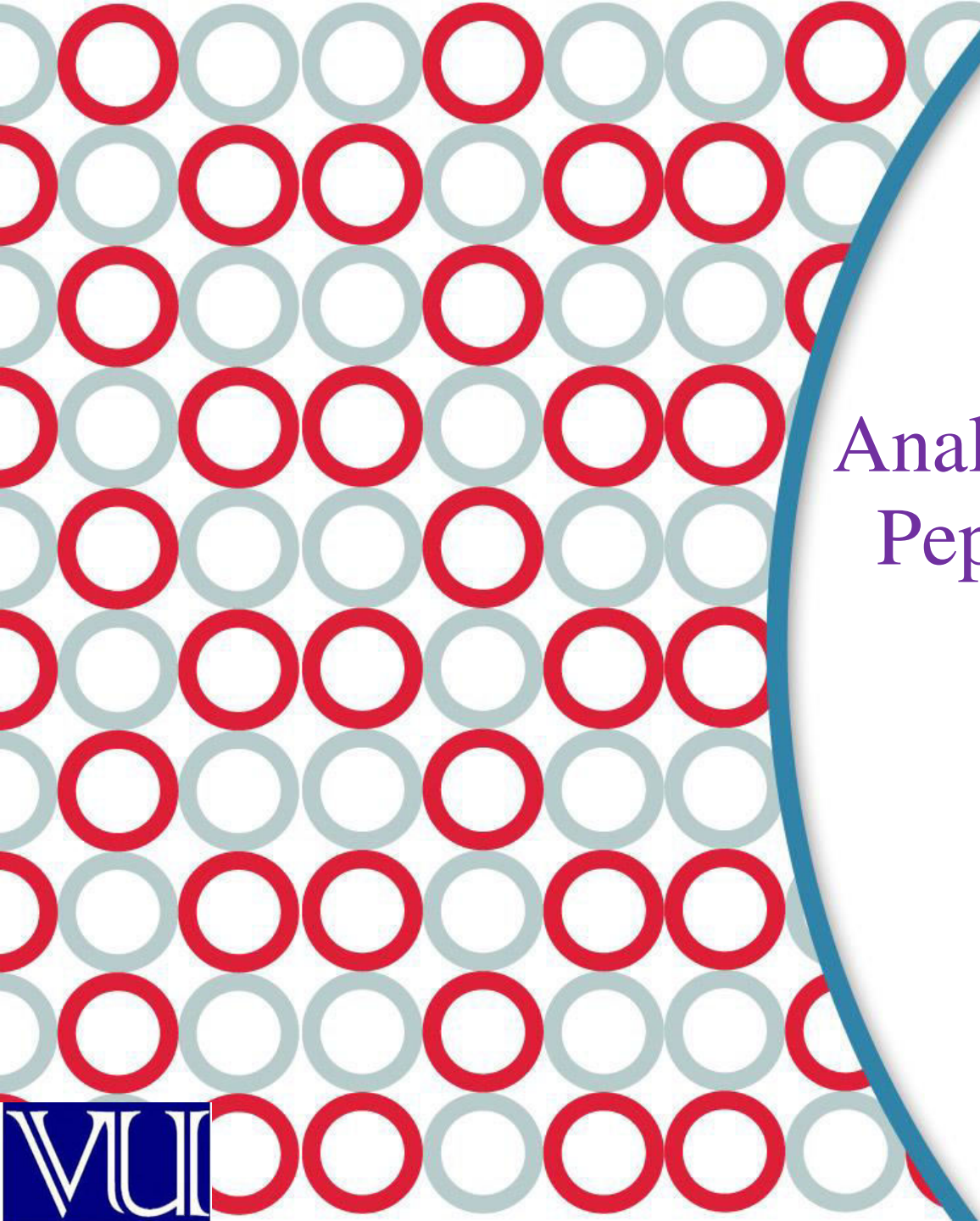
Complexes with multiple components including proteins

- Nucleosome
- Ribonucleoprotein (generic)
- Signal recognition particle
- Spliceosome

Type of Proteins/Families

Types of Proteins - Conclusion

- Three major categories
- Fibrous proteins
- Globular proteins
- Complexes with multiple components including proteins



Analytical Protein and Peptide Separations

BIO 601

Lecture-1

Analytical Protein and Peptide Separations

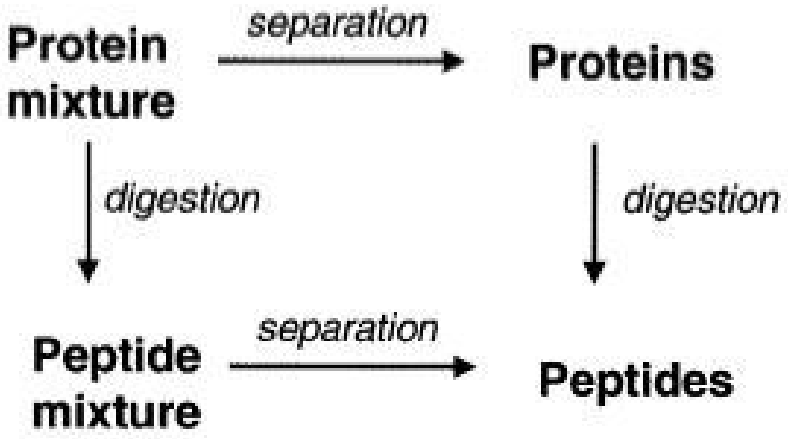
Overview

- At the stage of proteomic analysis, we must do two things

1. First, we must convert proteins to peptides. This is generally done with proteolytic enzymes.
2. Second, we must separate very complex mixtures of proteins or peptides into somewhat less complex mixtures

- This gives the MS instruments a better opportunity to obtain useful data on the

fppt.com


Protein mixture
separation
Proteins
digestion
digestion
Peptide mixture
separation
Peptides

fppt.com

Analytical Protein and Peptide Separations

Extracting Proteins from Biological Samples

Introduction

- Biological sample can be a piece of tissue, a plate of cultured cells, a flask of bacteria, a leaf, and so on.

- The sample then is usually pulverized, homogenized, sonicated, or otherwise disrupted to yield a soup that contains cells, subcellular components, and other biological debris in an aqueous buffer or suspension.

- For proteomic analysis, the objective here is to recover as much of the protein as possible from this soup with as little contamination by other biomaterials (e.g., lipids, cellulose, nucleic acid, etc.) as possible.

- Protein is generally extracted with least contamination with the aid of

1. Detergents
2. Reductants
3. Denaturing agents
4. Enzymes

1. Detergents

- Detergents e.g SDS, 3- ([3-cholamidopropyl] dimethylammonio)-1propane sulfonate (CHAPS), cholate, Tween can be used for protein extraction.

- These help to solubilize membrane proteins and aid their separation from lipids

2. Reductants

- Reductants e.g. e.g., dithiothreitol [DTT], mercaptoethanol, thiourea are used.

- These reduce disulfide bonds or prevent protein oxidation

3. Denaturing agents

- Examples of denaturing agents are urea and acids etc.

- These agents disrupt protein protein interactions, secondary and tertiary structures by altering solution ionic strength and pH

4. Enzymes

- Enzymes like DNAse and RNAse are used.
- Enzymes digest contaminating nucleic acids, carbohydrates, and lipids present in the soup.

Analytical Protein and Peptide Separations

Protein Separations Before Digestion

- The three principal separation approaches used with intact proteins are;

1. 1D-SDS-PAGE
2. 2D-SDS-PAGE
3. preparative isoelectric focusing (IEF).

- The protein mixture may be separated into a relatively small number of fractions as in 1D-SDS-PAGE and preparative IEF.

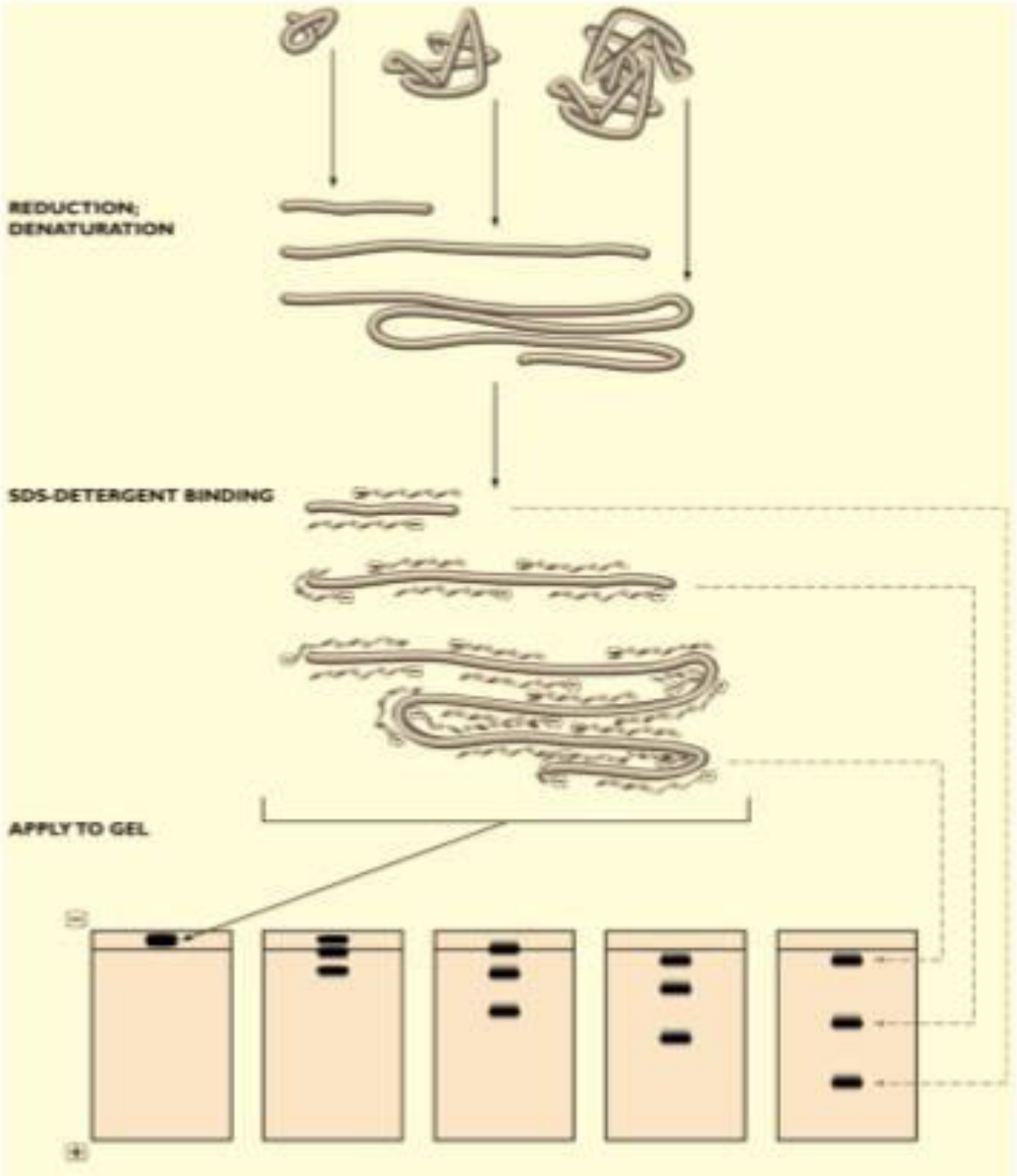
- Or it can be separated into many fractions as in the many spots in 2D-SDS-PAGE.

- The fractions then are taken for proteolytic digestion followed either by further separation of the peptide fragments or direct MS analysis of the peptides.

One-Dimensional SDS-PAGE

Principle

- It is a single most widely used analytical separation in all of protein chemistry.
- protein sample is dissolved in a loading buffer that usually contains a thiol reductant (mercaptoethanol or DTT) and SDS.
- The separation method is based on the binding of SDS to the protein, which imparts negative charge to the protein in proportion to molecular weight.
- When the gel is subjected to high voltage, the protein-SDS complexes migrate through the cross-linked polyacrylamide gel.
- Rate of migration is based on the ability of proteins to penetrate the pore matrix of the gel.
- The proteins thus are resolved into bands in order of molecular weight.


REDUCTION;
DENATURATION
SDS-DETERGENT BINDING
APPLY TO GEL
−
+

Extent of cross-linking in 1D-SDS PAGE

- In 1D SDS-PAGE the extent of cross linking in gels varies from 5-15%.
- Lower degrees of cross-linking allow easier passage of larger proteins through the gel.
- A sample containing low molecular-weight proteins is better resolved on a more highly cross-linked gel.

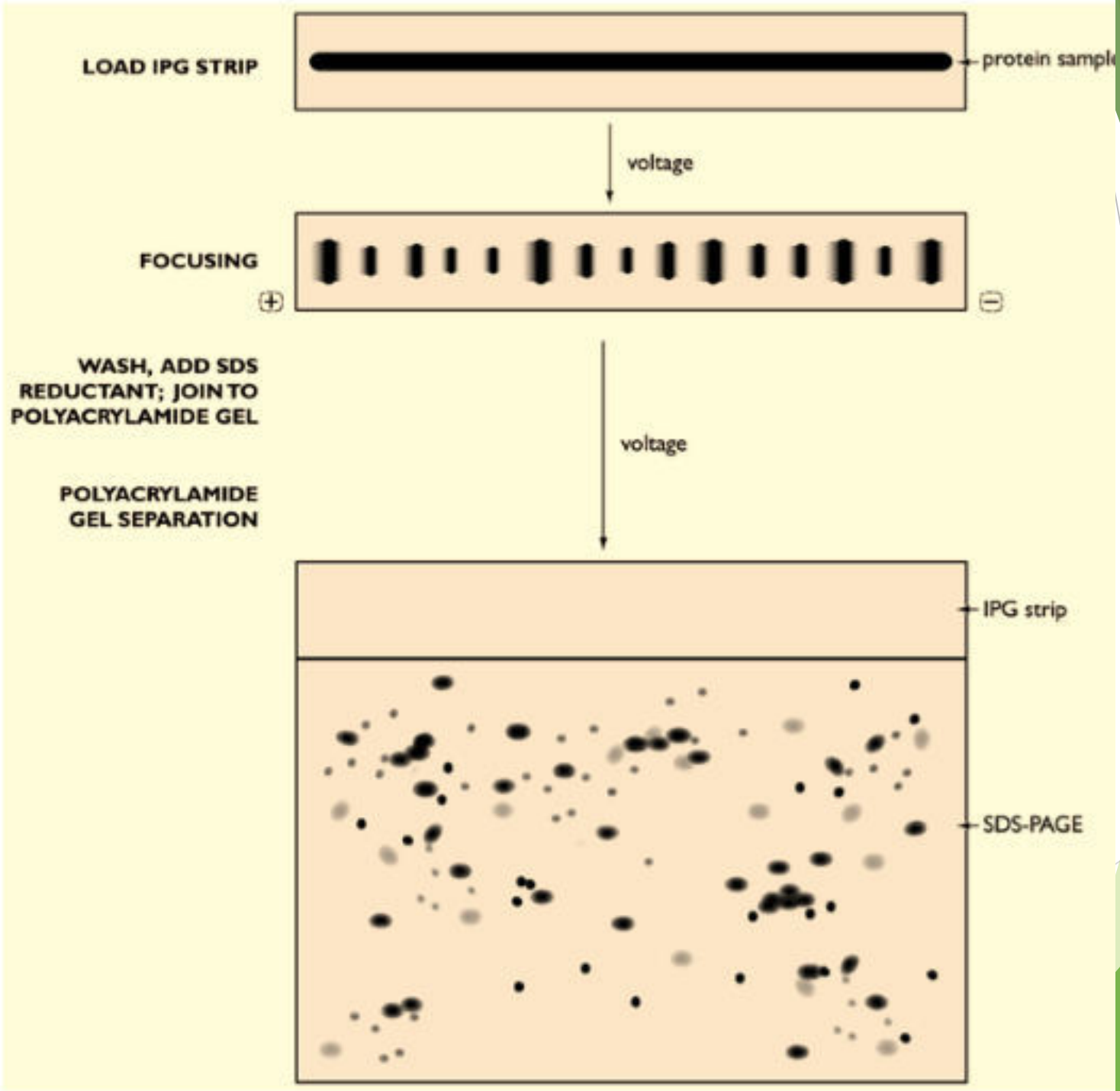
Two-Dimensional SDS-PAGE

Principle

- It is a single best method for resolving highly complex protein mixtures.

- It involves two different types of separations.

1. In the first, the proteins are resolved on the basis of isoelectric point by IEF.
2. In the second, focused proteins then are further resolved by electrophoresis on a polyacrylamide gel


LOAD IPG STRIP
protein sample
voltage
FOCUSING
WASH, ADD SDS REDUCTANT; JOIN TO POLYACRYLAMIDE GEL
voltage
POLYACRYLAMIDE GEL SEPARATION
IPG strip
SDS-PAGE

In past 2D SDS-PAGE was difficult to use because of

- The relative technical difficulty of performing the IEF step.
- Getting the delicate tube gel containing the focused proteins set up to efficiently transfer the proteins in the SDS-PAGE slab gel was a technical challenge.

New designed 2D SDS-PAGE System

- It uses immobilized pH gradient (IPG) strips and relatively foolproof hardware to facilitate the transfer of proteins from the IPG strip into the SDS-PAGE slab gel.
- In this, use of narrow pH ranges facilitates the separation of proteins with highly similar isoelectric points.
- The strip is hydrated with a buffer and the protein is slowly loaded into the strip under voltage.
- Then the voltage is increased to achieve focusing.

- After the focusing step, the strip is treated with a buffer that contains a thiol reductant and SDS and then is joined to the SDS-PAGE slab gel. In this respect, the IPG strip containing the focused proteins acts as a “stacking” gel in 1D-SDS-PAGE.

- The proteins then are resolved on the SDS-PAGE slab gel in the same manner as for 1D-SDS-PAGE.

- Proteins separated by 2D gels are visualized by conventional staining techniques, including silver, Coomassie, and amido black stains.

Problems of 2D SDS-PAGE

Problems 1

- The first is the difficulty of performing completely reproducible 2D-SDS-PAGE analyses. This problem becomes important when one wishes to use 2D-SDS-PAGE to compare two samples by comparing the images of the stained gels. Differences in protein migration in either dimension could be mistaken for differences in levels of certain proteins between the two samples.

Problem 2

- A second problem with 2D-SDS-PAGE is the relative incompatibility of some proteins with the first-dimension IEF step. Many large, hydrophobic proteins simply do not behave well in this type of analysis. Marginal solubility leads to protein precipitation and aggregation, which leads to “smearing” of proteins within the IPG strip, rather than clean focusing into discrete bands. When these proteins are subsequently run in the second (SDS-PAGE) dimension, these proteins appear as streaks across a molecular-weight region

Problem 3

- A third problem with 2D-SDS-PAGE is the relatively small dynamic range of protein staining as a detection technique. Spot densities reflect about a 100-fold range of protein concentrations, at best. This means that staining of 2D-gels allows the visualization of abundant proteins, whereas less abundant proteins frequently cannot be detected.

Isoelectric Focusing (IEF)

Introduction

- The *isoelectric point* is the pH at which the net charge of the protein molecule is neutral.
- Different proteins have different isoelectric points.
- Isoelectric point is found by drawing the sample through a stable pH gradient.
- The range of the gradient determines the resolution of the separation.

- This technique is analogous to the first step in 2D-SDS-PAGE.

- In IEF generation of a pH gradient is achieved with soluble ampholytes, which are polycarboxylic acid compounds that generate a stable pH gradient when voltage is applied across the focusing cell.

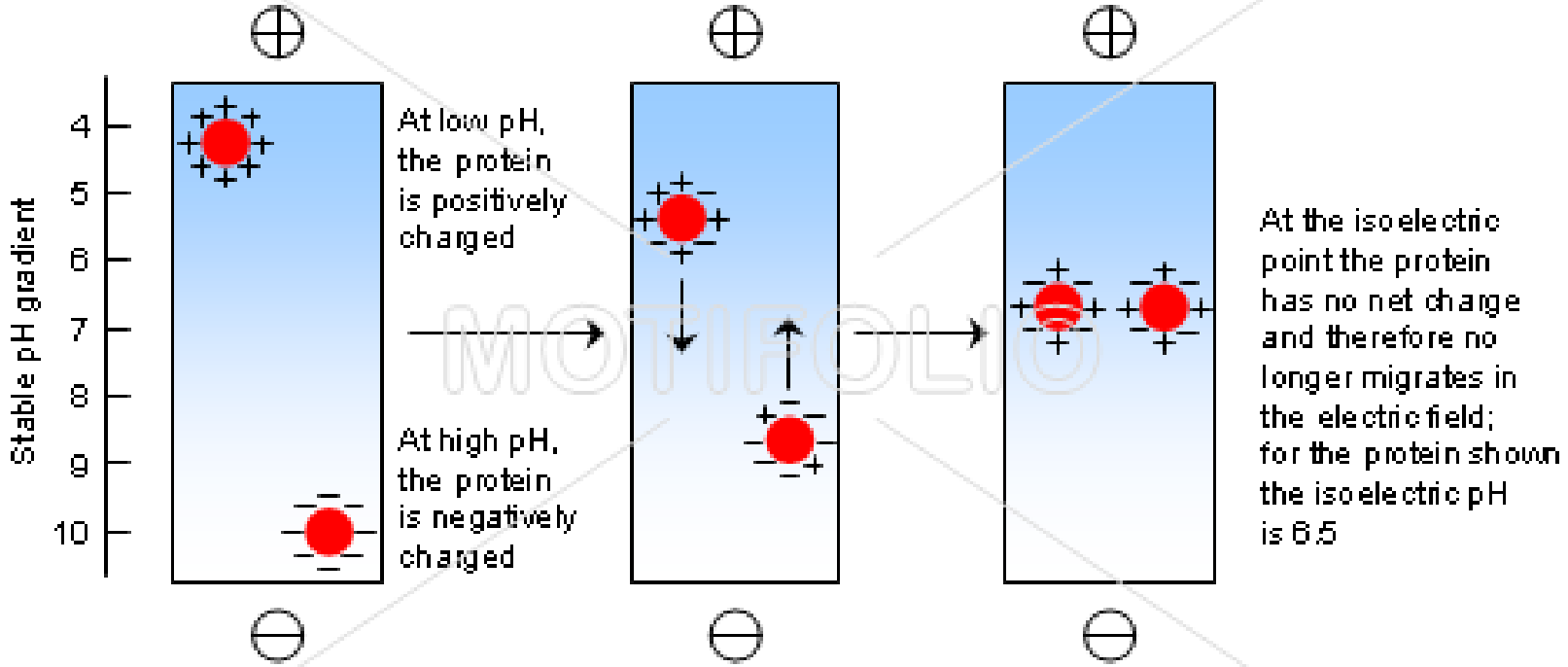
- The protein sample then is added, voltage again is applied, and the proteins then are separated by isoelectric point.

- In commercially available apparatus, such as the BioRad Rotofor™ cell, the focusing cell is divided by permeable membranes into a series of chambers.

- After the focusing step, the chambers are quickly and simultaneously emptied by a vacuum sipper that draws the contents of each section of the cell into a separate tube.

- With this type of apparatus, the entire protein mixture is separated into 12–20 fractions

Separation of protein molecules by isoelectric focusing



Advantages of solution phase IEF

- It has relatively large sample capacity (milligrams to grams of total protein per run).
- It has relative ease of working with samples in solutions as opposed to gels.
- The ampholytes can be removed from the fractionated samples by dialysis or gel filtration prior to further processing of the proteins.
- Recovery of proteins from solution-phase IEF typically exceeds 85–90%.
- Detergents and chaotropic agents can be used to maintain solubility of hydrophobic proteins.

High-Performance Liquid Chromatography

Introduction

- High-performance liquid chromatography (HPLC; formerly referred to as high-pressure liquid chromatography), is a technique in analytical chemisry used to separate, identify, and quantify each component in a mixture.

- HPLC is a chromatographic technique that can separate a mixture of compounds. It is used in biochemistry and analytical chemistry to identify, quantify and purify the individual components of a mixture.

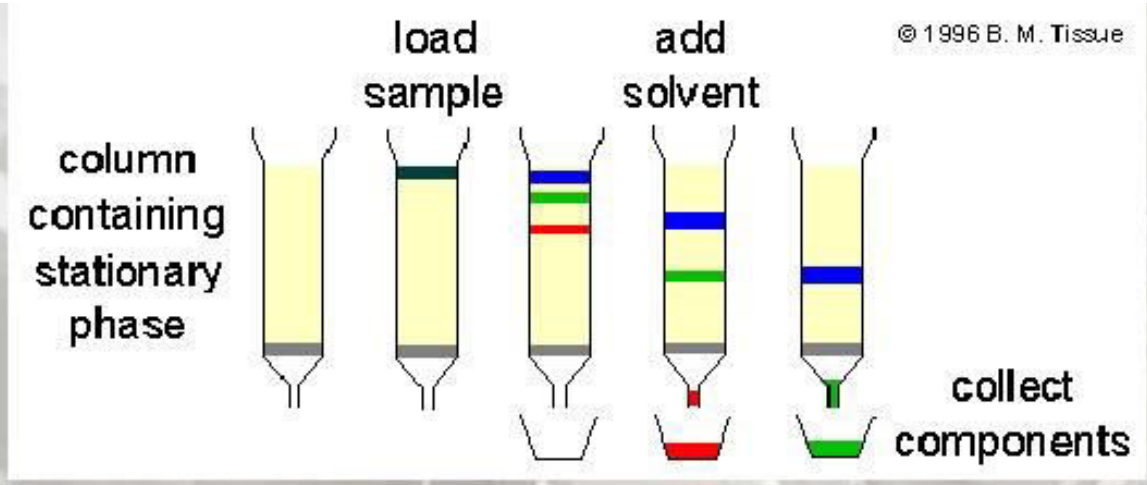
- Although HPLC of intact proteins has not become a widely used technique for analytical proteomics, it is nevertheless highly applicable as an initial step to fractionate protein mixtures.

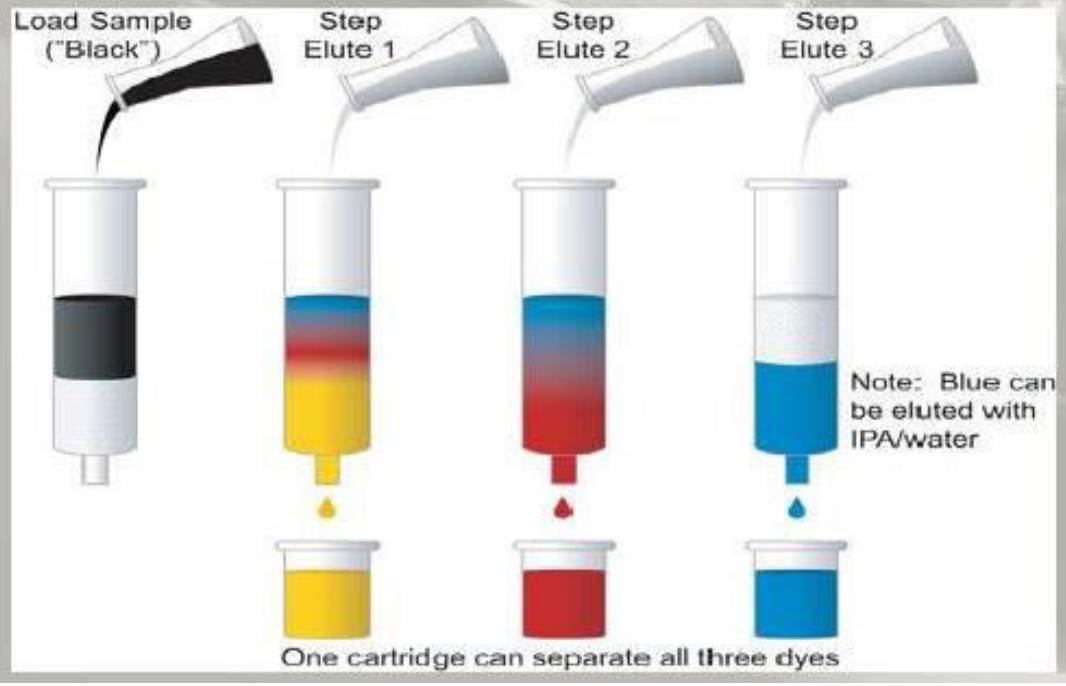
- HPLC would appear to be about as useful as preparative IEF for resolving protein mixtures into fractions.

Principle

- HPLC relies on pumps to pass a pressurized liquid solvent containing the sample mixture through a column filled with a solid adsorbent material.

- Each component in the sample interacts slightly differently with the adsorbent material, causing different flow rates for the different components and leading to the separation of the components as they flow out the column


load
sample
add
solvent
© 1996 B. M. Tissue
column
containing
stationary
phase
collect
components


Load Sample
("Black")
Step
Elute 1
Step
Elute 2
Step
Elute 3
Note: Blue can
be eluted with
IPA/water
One cartridge can separate all three dyes

Advantages of HPLC

- The advantage of HPLC is the diversity of separation modes available. Indeed, tandem HPLC separations combine two different types of chromatography. For example, strong cation exchange, followed by RP (Reverse phase), would apply two completely different separation modes.

Protein Separations After Digestion

Introduction

- In this approach, the proteins in the sample are first digested into a mixture of peptides, then the peptides are separated prior to analysis.
- The extreme application of this approach would be to digest a complete cell or tissue extract to peptides and then perform MS analysis on the mixture.
- Indeed, this sort of analysis has been done with considerable success.

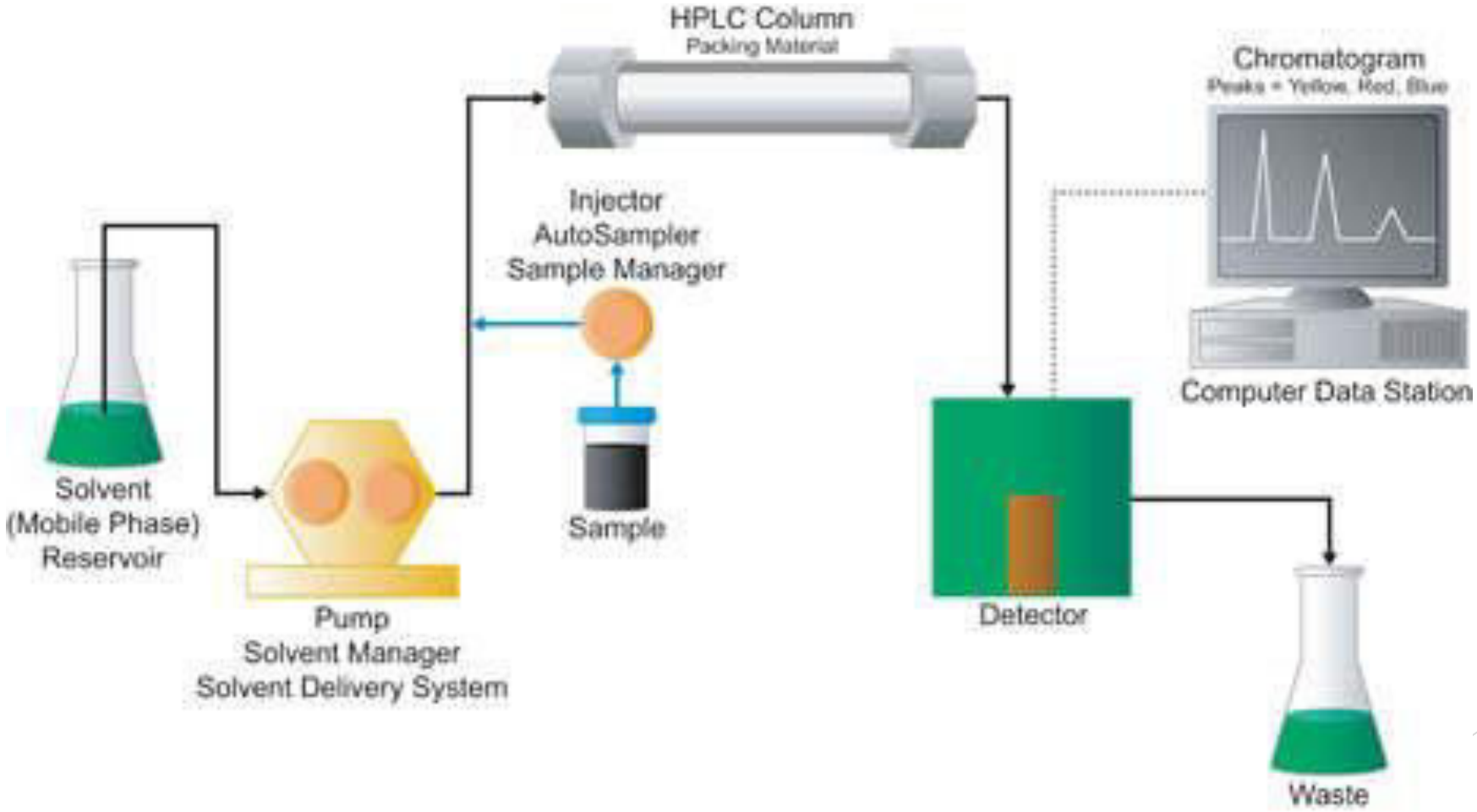
Use of HPLC for peptide separation

- The use of microcapillary HPLC with special control adaptations and automated MS instrument control allowed the acquisition of MS data on hundreds or thousands of peptides in a single run.

- The primary rationale for this approach is that it permits one to convert a very heterogeneous mixture of proteins to a more homogeneous mixture of peptides, which can be more easily analyzed.

- If one does elect this approach, the number of available methods to separate peptide mixtures is far more limited.

Instrumentation of HPLC



Use of other methods for peptide separation

- One-dimensional- and 2D-SDS-PAGE are out, as they are not practically useful in resolving peptides from digests, which typically display a much more limited range of pI and molecular weight.

- Although preparative IEF can be performed on peptide mixtures, but it may be of limited utility for resolving peptide mixtures.

- However, little has been done to evaluate preparative IEF as a tool for peptide separations and it cannot be ruled out.

Tandem LC Approaches for Peptide Analysis

Introduction

- Certainly the most widely used approach to analysis of peptide mixtures is HPLC.
- The diversity of stationary phases and separation modes gives HPLC considerable resolving power.
- The use of combined separation modes in series is referred to as “tandem HPLC.”
- The idea behind tandem LC is that the combination of dissimilar separation modes allows a greater resolution of peptides in a mixture.

Major HPLC separation modes and the characteristics that dictate separation are

- RP: hydrophobicity
- Strong cation exchange: net positive charge
- Strong anion exchange: net negative charge
- Size exclusion: peptide size/molecular weight
- Affinity: interaction with specific functional groups

Capillary Electrophoresis (CE)

- Capillary electrophoresis (CE) operates on the same general principal as IEF.
- Proteins placed in an electric field will migrate to a point in a pH gradient where they display an overall neutral change.
- The performance of the analysis in a microcapillary tube provides greatly enhanced resolution over the preparative IEF techniques discussed earlier.
- CE offers the greatest resolution of all peptide analytical techniques and can be coupled directly to MS instruments.
- CE thus has great potential as a technique for analytical proteomics.

- The utility of CE is limited at the present time by the lack of commercially available, robust, and reliable CE-MS instrumentation for analytical proteomics.

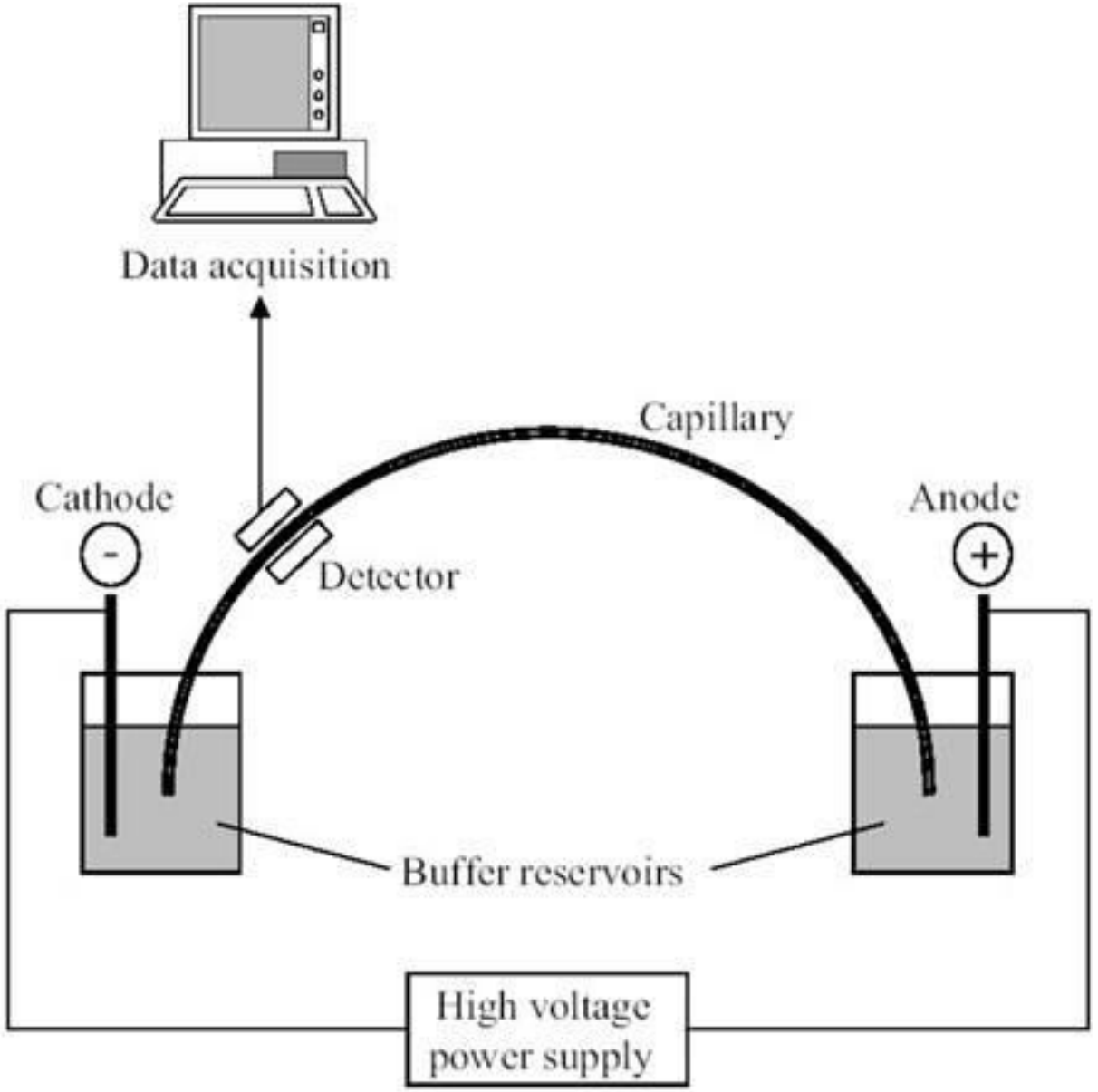
- Development of instrumentation for this purpose is continuing and CE-MS may become a very useful tool in proteomics analysis in the near future.

Working principle

- Capillary tube is placed between two buffer reservoir, and an electric field is applied, separation depends on electrophoretic mobility & electro-osmosis .

- Defined volume of analysate is introduced in to the capillary by replacing one buffer reservoir with sample vial.

- Electrophoretic separation is measured by detector.


Data acquisition
Capillary
Cathode
Anode
Detector
Buffer reservoirs
High voltage
power supply