Virtual University of Pakistan

Genomics and Proteomics

Bt504

Bt504@vu.edu.pk

Created By Team

Cluesbook

https://www.facebook.com/groups/VuHelpforum

WWW.CLUESBOOK.COM

Topic 1

Genetics and Genomics

Study of Omics

Genomics: all the genes

Pharmacogenomics choice of personalized medicine

Nutrigenomics choice of best diet

Toxicogenomics prediction of toxicity

Epigenomics: all epigenetic changes in genome

Transcriptomics: all the mRNAs

Proteomics : all the proteins

Interactomics : all interactions between all proteins

Metabolomics (or metabonomics): all metabolites

Topic 2

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.

Proteome and Proteomics

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

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

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.

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

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.

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

Topic 3

Structure of RNA

Each nucleotide in RNA contains a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, in general, adenine (A), cytosine (C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine and uracil are pyrimidines. A phosphategroup is attached to the 3' position of one ribose and the 5' position of the next. The phosphate groups have a negative charge each, making RNA a charged molecule (polyanion). The bases form hydrogen bonds between cytosine and guanine, between adenine and uracil and between guanine and uracil.However, other interactions are possible, such as a group of adenine bases binding to each other in a bulge, or the GNRA tetraloop that has a guanine–adenine base-pair.

Chemical structure of RNA

An important structural component of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. The presence of this functional group causes the helix to mostly take the A-form geometry, although in single strand dinucleotide contexts, RNA can rarely also adopt the B-form most commonly observed in DNA. The A-form geometry results in a very deep and narrow major groove and a shallow and wide minor groove. A second consequence of the presence of the 2'-hydroxyl group is that in conformationally flexible regions of an RNA molecule (that is, not involved in formation of a double helix), it can chemically attack the adjacent phosphodiester bond to cleave the backbone.

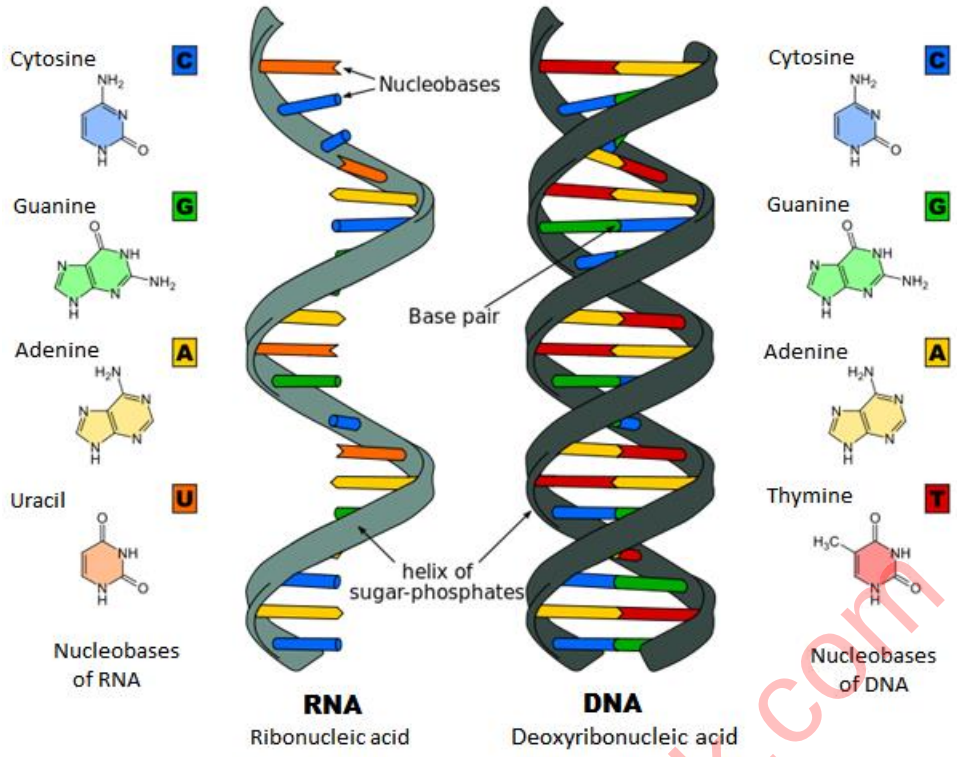
Secondary structure of a telomerase RNA.

RNA is transcribed with only four bases (adenine, cytosine, guanine and uracil), but these bases and attached sugars can be modified in numerous ways as the RNAs mature. Pseudouridine (Ψ), in which the linkage between uracil and ribose is changed from a C–N bond to a C–C bond, and ribothymidine (T) are found in various places (the most notable ones being in the TΨC loop of tRNA). Another notable modified base is hypoxanthine, a deaminated adenine base whose nucleoside is called inosine (I). Inosine plays a key role in the wobble hypothesis of the genetic code.

There are more than 100 other naturally occurring modified nucleosides. The greatest structural diversity of modifications can be found in tRNA,[while pseudouridine and nucleosides with 2'-O-methylribose often present in rRNA are the most common.The specific roles of many of these modifications in RNA are not fully understood. However, it is notable that, in ribosomal RNA, many of the post-transcriptional modifications occur in highly functional regions, such as the peptidyl transferase center and the subunit interface, implying that they are important for normal function.

The functional form of single-stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements that are hydrogen bonds within the molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and internal loops.Since RNA is charged, metal ions such as Mg2+ are needed to stabilise many secondary and tertiary structures.

The naturally occurring enantiomer of RNA is D-RNA composed of D-ribonucleotides. All chirality centers are located in the D-ribose. By the use of L-ribose or rather L-ribonucleotides, L-RNA can be synthesized. L-RNA is much more stable against degradation by RNase.Like other structured biopolymers such as proteins, one can define topology of a folded RNA molecule. This is often done based on arrangement of intra-chain contacts within a folded RNA, termed as circuit topology.



Topic 4

DNA Transcription

Transcription is the first step of gene expression, in which a particular segment of DNA is copied into RNA (especially mRNA) by the enzyme RNA polymerase. Both DNA and RNA are nucleic acids, which use base pairs of nucleotides as a complementary language. During transcription, a DNA sequence is read by an RNA polymerase, which produces a complementary, antiparallel RNA strand called a primary transcript.

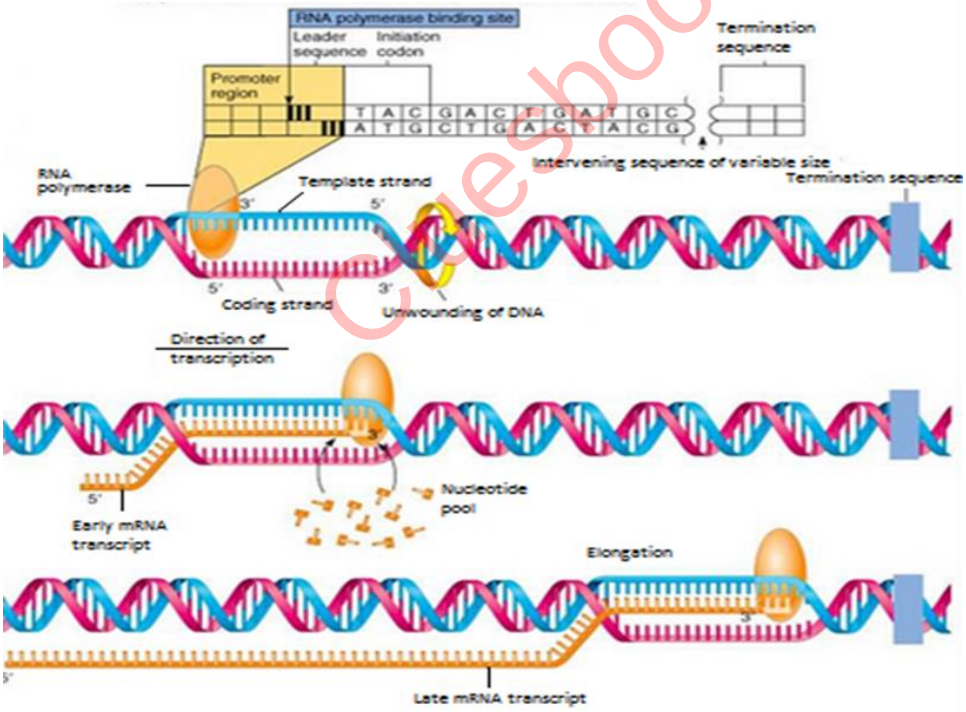
Transcription proceeds in the following general steps:

- RNA polymerase, together with one or more general transcription factors, binds to promoter DNA.
- RNA polymerase creates a transcription bubble, which separates the two strands of the DNA helix. This is done by breaking the hydrogen bonds between complementary DNA nucleotides.
- RNA polymerase adds RNA nucleotides (which are complementary to the nucleotides of one DNA strand).
- RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
- Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesized RNA strand.
- If the cell has a nucleus, the RNA may be further processed. This may include polyadenylation, capping, and splicing.
- The RNA may remain in the nucleus or exit to the cytoplasm through the nuclear pore complex.

The stretch of DNA transcribed into an RNA molecule is called a transcription unit and encodes at least one gene. If the gene encodes a protein, the transcription produces messenger RNA (mRNA); the mRNA, in turn, serves as a template for the protein's synthesis through translation. Alternatively, the transcribed gene may encode for non-coding RNAsuch as microRNA, ribosomal RNA (rRNA), transfer RNA (tRNA), or enzymatic RNA molecules called ribozymes.[1] Overall, RNA helps synthesize, regulate, and process proteins; it therefore plays a fundamental role in performing functions within a cell.

Transcription involves four steps:

- **Initiation.** The DNA molecule unwinds and separates to form a small open complex. RNA polymerase binds to the promoter of the template strand.
- **Elongation.** RNA polymerase moves along the template strand, synthesising an mRNA molecule. In prokaryotes RNA polymerase is a holoenzyme consisting of a number of subunits, including a sigma factor (transcription factor) that recognises the promoter. In eukaryotes there are three RNA polymerases: I, II and III. The process includes a proofreading mechanism.
- **Termination**. In prokaryotes there are two ways in which transcription is terminated. In Rho-dependent termination, a protein factor called "Rho" is responsible for disrupting the complex involving the template strand, RNA polymerase and RNA molecule. In Rho-independent termination, a loop forms at the end of the RNA molecule, causing it to detach itself. Termination in eukaryotes is more complicated, involving the addition of additional adenine nucleotides at the 3' of the RNA transcript (a process referred to as polyadenylation).
- **Processing.** After transcription the RNA molecule is processed in a number of ways: introns are removed and the exons are spliced together to form a mature mRNA molecule consisting of a single protein-coding sequence. RNA synthesis involves the normal base pairing rules, but the base thymine is replaced with the base uracil.



Topic 5

Protein Translation

Translation is the process of translating the sequence of a messenger RNA (mRNA) molecule to a sequence of amino acids during protein synthesis. The genetic code describes the relationship between the sequence of base pairs in a gene and the corresponding amino acid sequence that it encodes.

Translation involves four steps:

- **Initiation.** The DNA molecule unwinds and separates to form a small open complex. RNA polymerase binds to the promoter of the template strand.
- **Elongation.** RNA polymerase moves along the template strand, synthesising an mRNA molecule. In prokaryotes RNA polymerase is a holoenzyme consisting of a number of subunits, including a sigma factor (transcription factor) that recognises the promoter. In eukaryotes there are three RNA polymerases: I, II and III. The process includes a proofreading mechanism.
- **Termination**. In prokaryotes there are two ways in which transcription is terminated. In Rho-dependent termination, a protein factor called "Rho" is responsible for disrupting the complex involving the template strand, RNA polymerase and RNA molecule. In Rho-independent termination, a loop forms at the end of the RNA molecule, causing it to detach itself. Termination in eukaryotes is more complicated, involving the addition of additional adenine nucleotides at the 3' of the RNA transcript (a process referred to as polyadenylation).
- **Processing.** After transcription the RNA molecule is processed in a number of ways: introns are removed and the exons are spliced together to form a mature mRNA molecule consisting of a single protein-coding sequence. RNA synthesis involves the normal base pairing rules, but the base thymine is replaced with the base uracil.

Topic 6

What is Genomic

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.

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

To generate physical, genetic, and sequence maps of different genomes.

To sequence the genomes of model organisms.

develops new technologies for mapping/sequencing.

Genomics helps in different ways

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

Study of Genomics helps to understand

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

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

Genomics Sub-disciplines

- Metagenomics
- Microbial genomics

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

Topic 7

Genomes Anatomy/Organization

Genome Anatomy

Anatomy of different genomes differ from each other.

Eukaryotes and prokaryotes genomes differ very significantly.

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

~ 30 fold between genomes of different eukaryotes.

In humans ~ 23,000

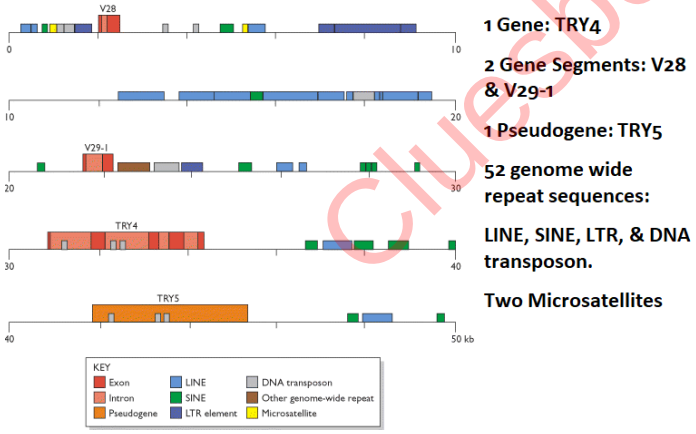
Bacterial genomes ~ 1,500 – 2,000 genes.

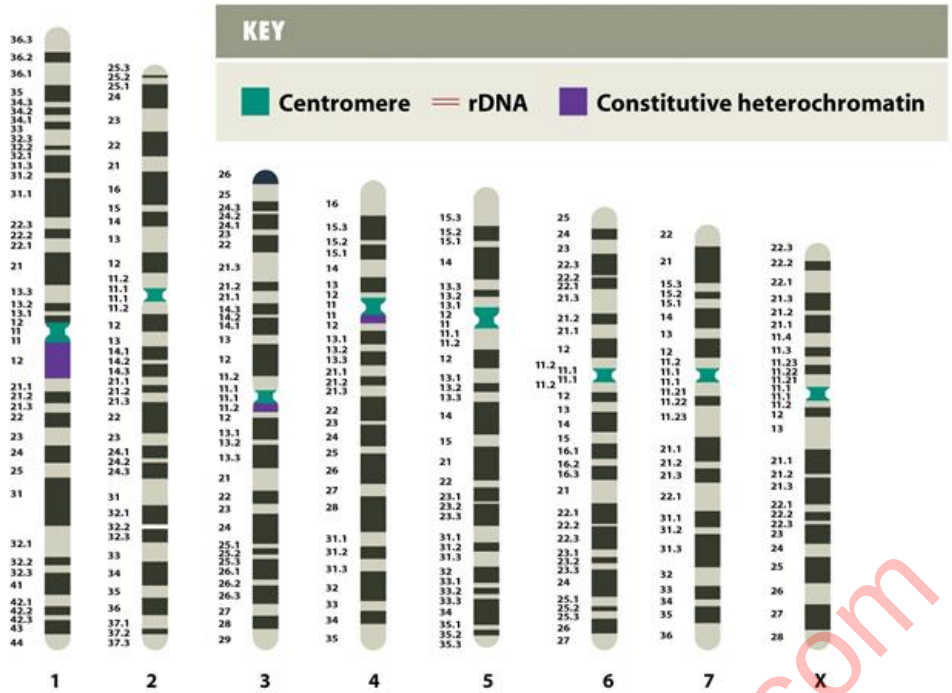
Genome Anatomy - Eukaryotes

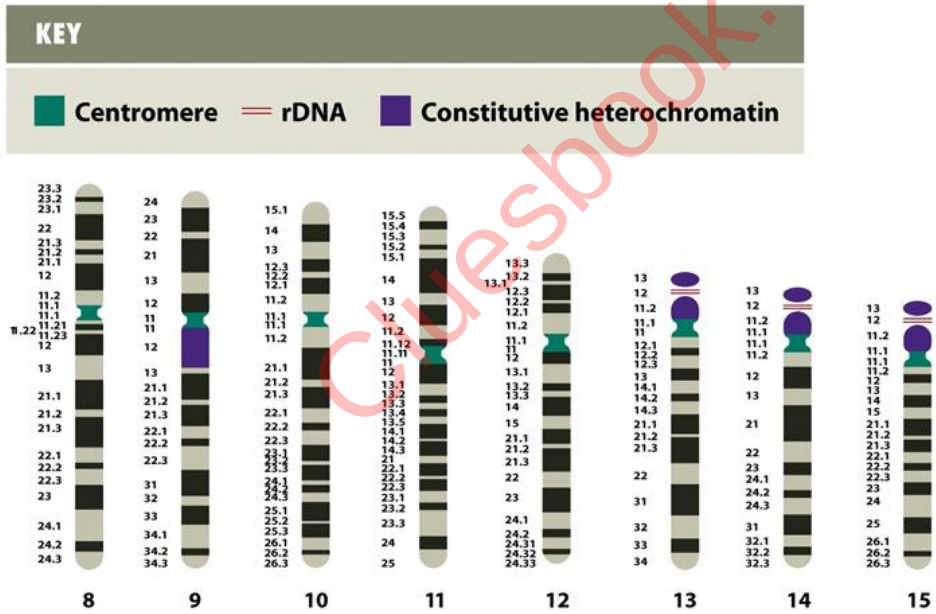
Eukaryotic genomes are full of simple repeats, numerous types of transposable elements and other sequences.

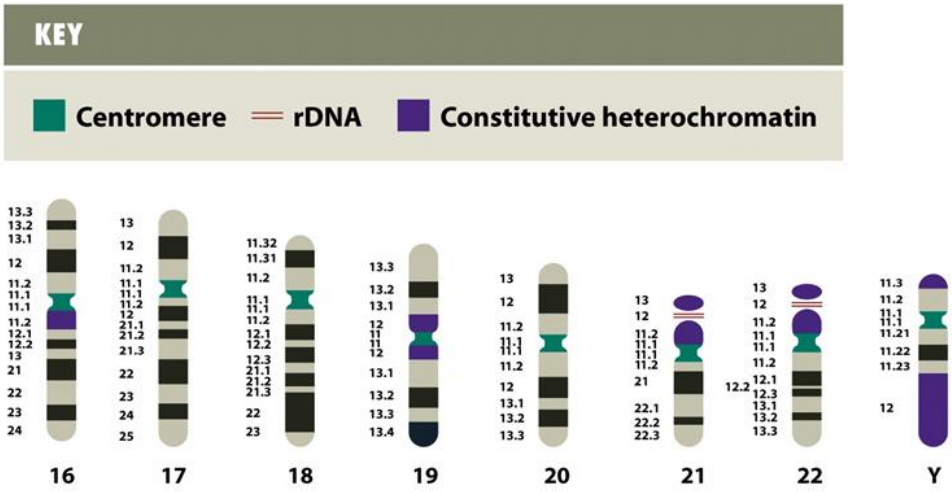
Genome Anatomy - Prokaryotes

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









Genome Organization in Prokaryotes

Species	DNA molecules	Genome organization 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

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

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

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

Genome Organization: Human, Yeast, Fruit Fly , Maize

Conclusion

•Anatomy of different genomes differ from each other.

•Eukaryotes and prokaryotes genomes differ very significantly.

Topic 8

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).

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

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).

mRNA is translated into Proteins

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

Three components

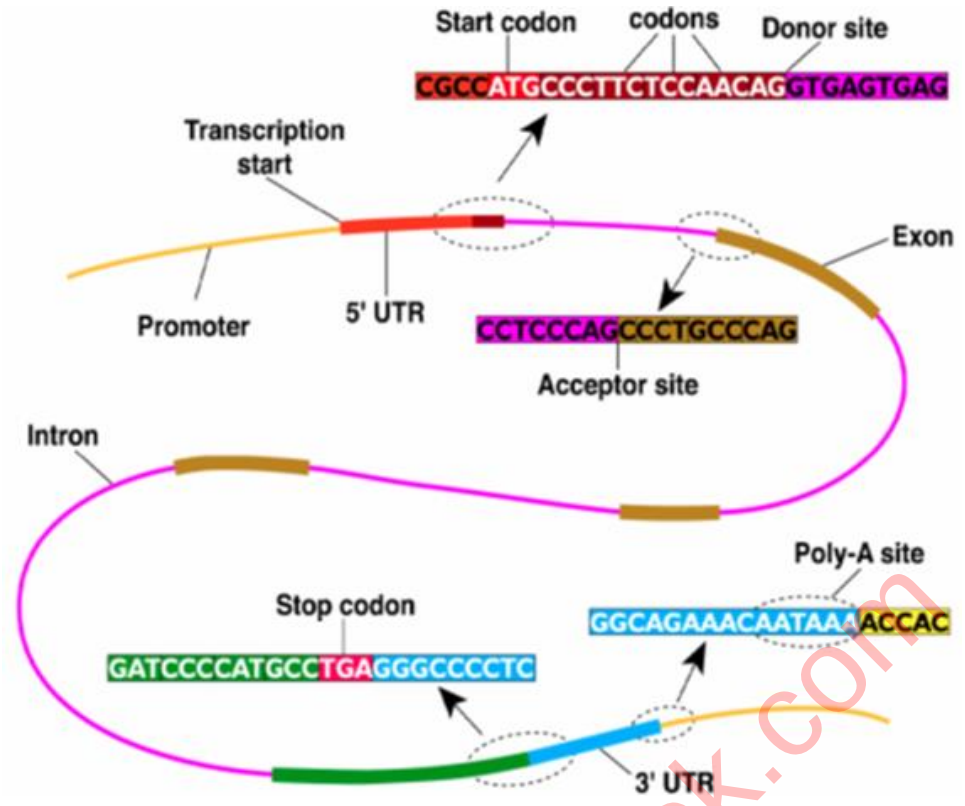
Open Reading Frame: From start codon (ATG) to stop (TGA, TAA, TAG)

Upstream region with binding site. (e.g. TATA box, GC box, CAAT box).

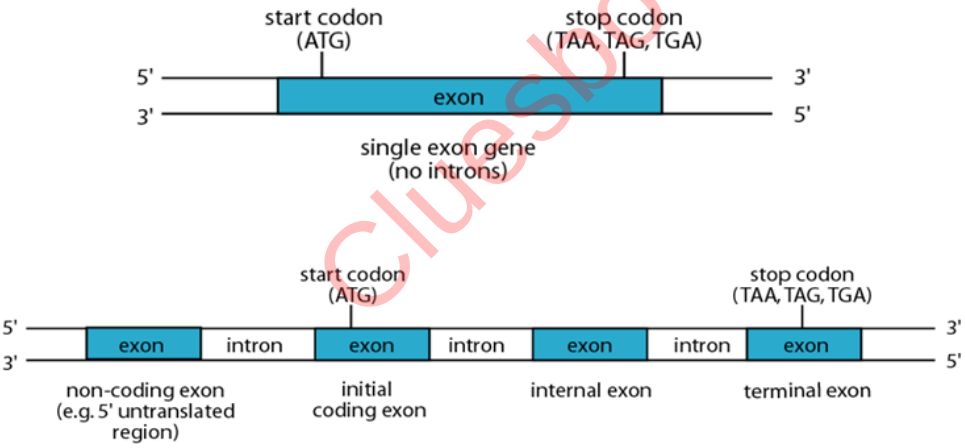
Poly-a tail

Gene Anatomy – Typical Prokaryotic Gene

Gene Anatomy – Typical Eukaryotic Gene



Single Exon Gene and Multiple Exons Gene



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

Topic 9

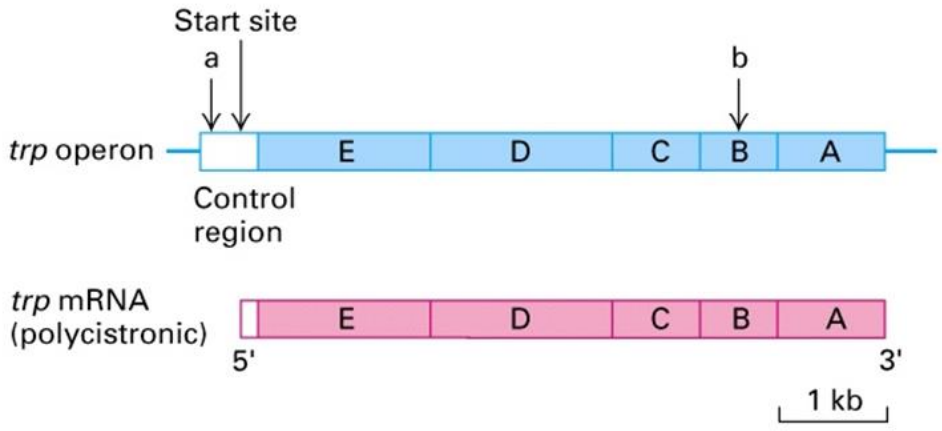
Prokaryotic Gene and Eukaryotic Gene

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

Polycistronic mRNA encodes several proteins



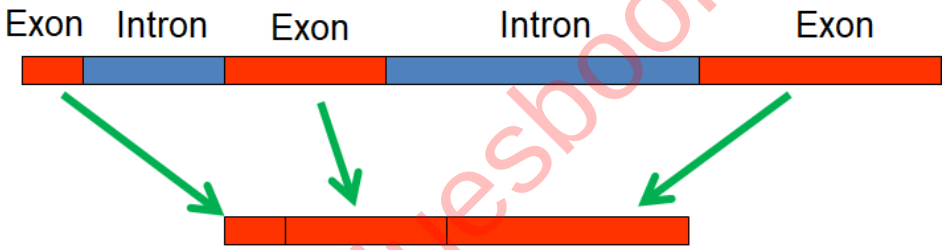
Bacterial Gene: Polycistronic mRNA encodes several proteins

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.

Splicing: Removal of introns from the mRNA molecule



Organize expression of genes' (function calls)

Promoter region (binding site), usually near coding region

Binding can block (inhibit) expression

Most have introns

Produce monocistronic mRNA

Large in size

Computational challenges

Identify binding sites

Correlate sequence to expression

Conclusion

Most of prokaryotic genes are without introns and in the are polycistronic.

Eukaryotic genes have introns and alternative splicing.

Topic 10

Genetics and Genomics

Prokaryotic and Eukaryotic Gene Expression - Difference

Bacterial genetics are different.

Prokaryote genes are grouped in operons

Prokaryotes have one type of RNA polymerase for all types of RNA

mRNA is not modified

The existence of introns in prokaryotes is extremely rare

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.

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

In prokaryotes, transcription of a gene and translation of the resulting mRNA occur simultaneously.

Many polysomes are found associated with an active gene.

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

Prokaryotes-operons simultaneously transcription and translation

Eukaryotes – no operons, RNA processing, chromatin remodeling

Topic 11

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

Related genes may be organized as several clusters at different locations

These are known as gene families e.g. globin genes family

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

Gene Family and Family members -

Peroxiredoxin family

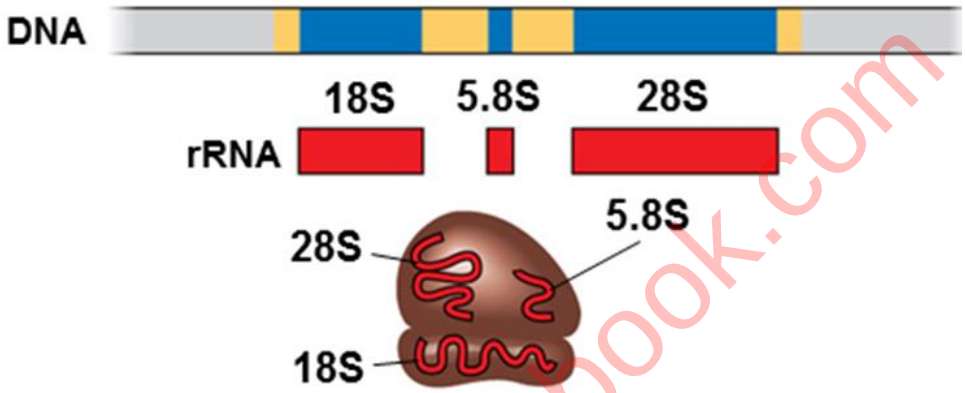
PRDX is the root symbol

Family members are *PRDX1*, *PRDX2*, *PRDX3*, *PRDX4*, *PRDX5* and *PRDX6*

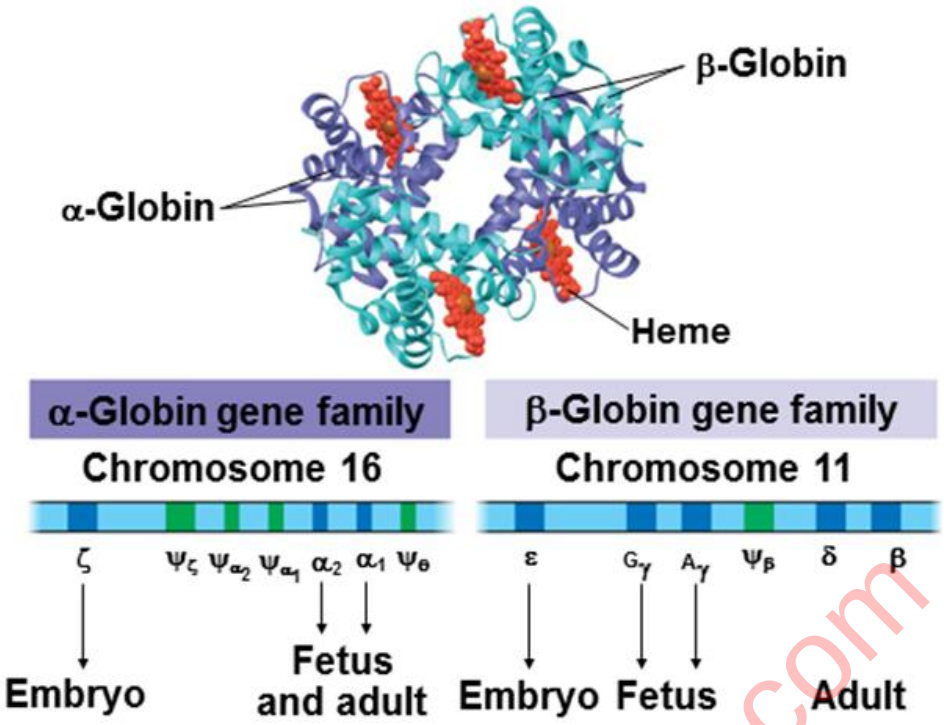
Genes Families

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

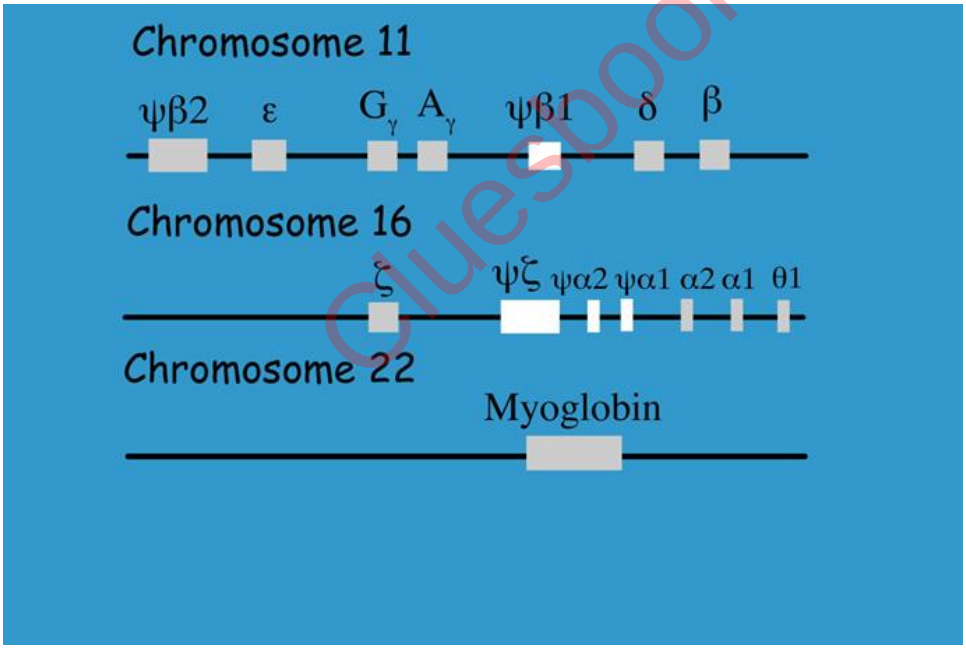
Ribosomal RNA Gene Family



Human Alpha & Beta Globin Gene Families



Genes Families on Different Chromosomes

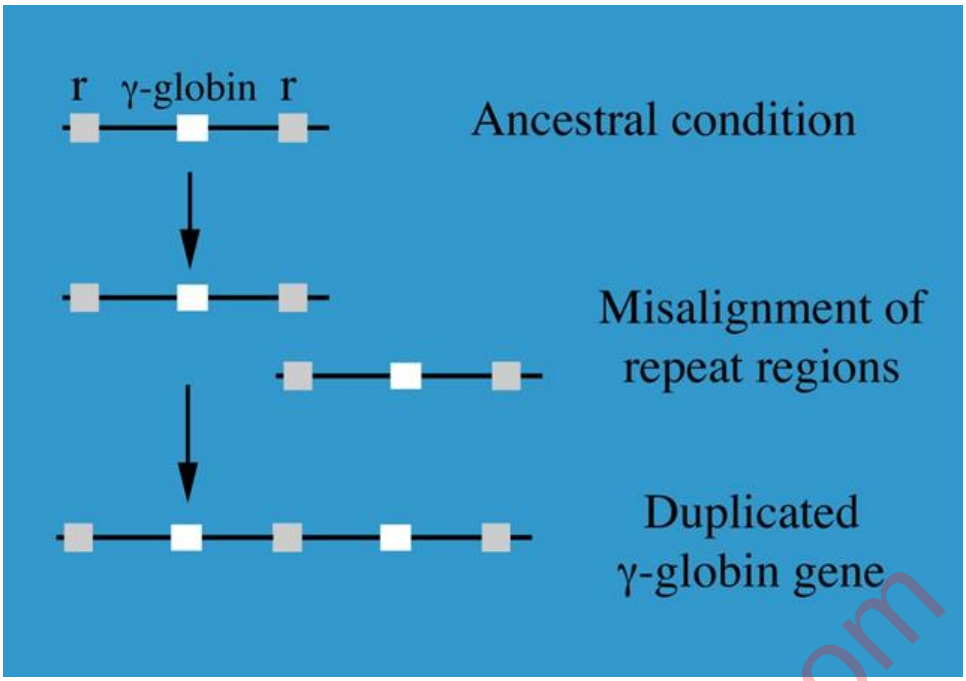


Genes Families

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: The misalignment – Globins



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

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 - Conclusion

Single individuals Genes

Gene families

Gene families in form of clusters

Gene families as Super families

Topic 12

Type of Proteins/Families

Protein Structure

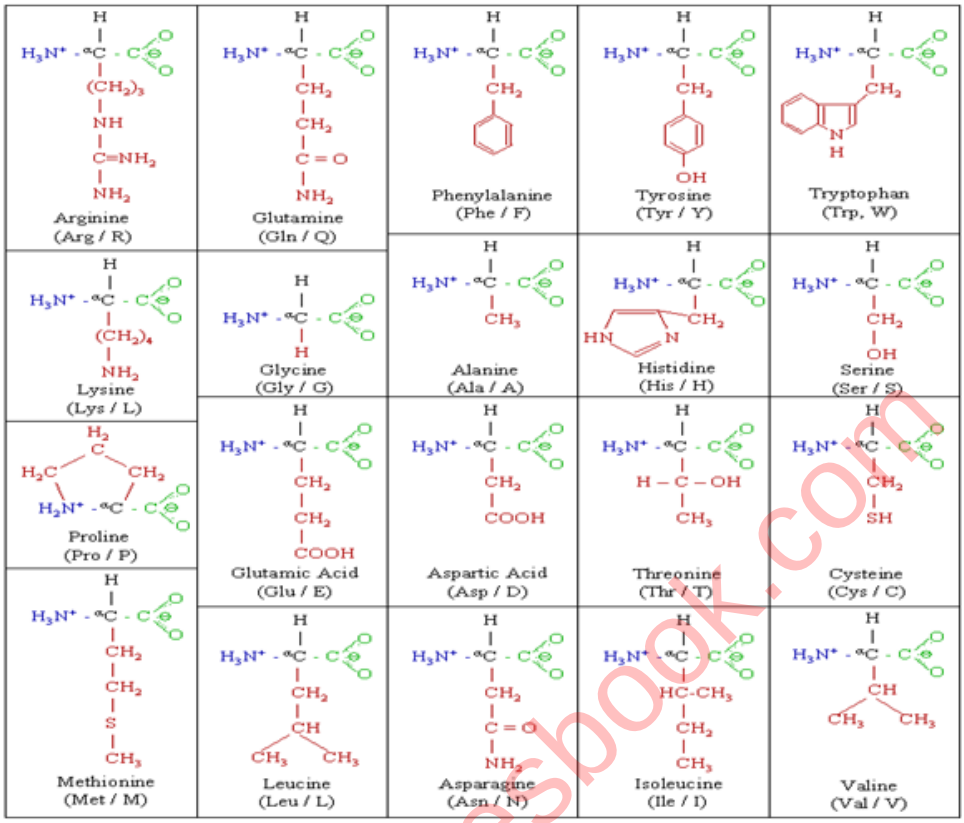
- Amino acids there are 20 amino acids that are esential for our body
- Hydrophobic / hydrophylic
- Charged / neutral

Functions

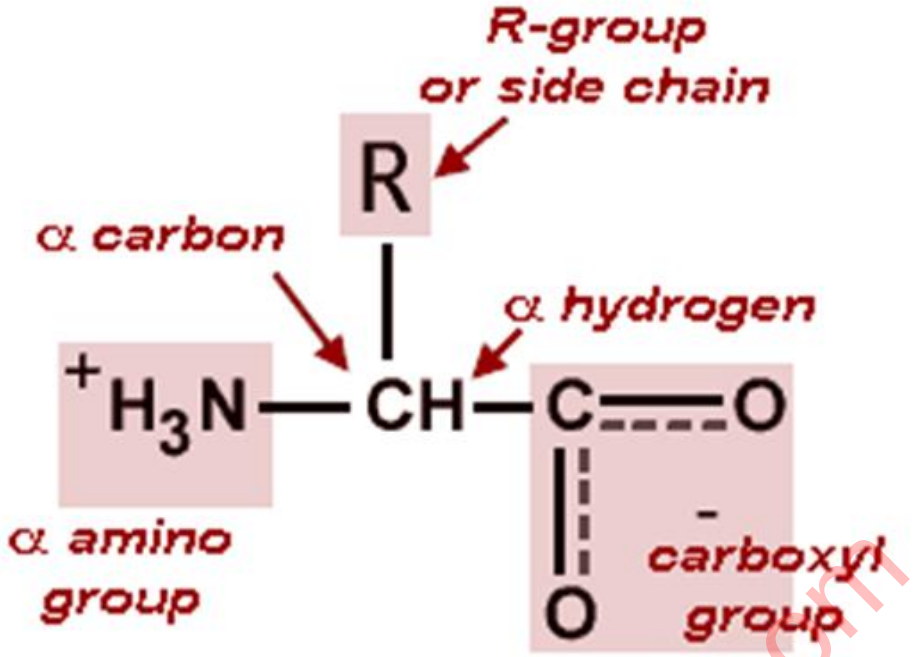
- Enzymes

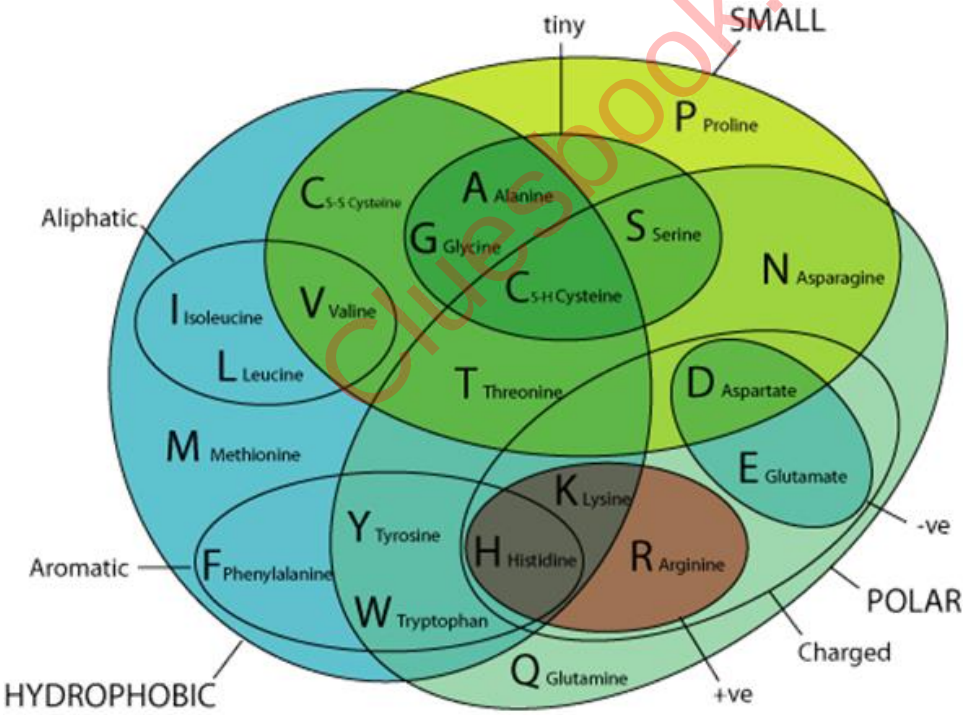
- Structure protein
- Channel
- Other functions

Amino Acids: Building Blocks of Proteins



Protein Structure: Amino Acid





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

Proteins

Different method of categorizing proteins

Three major categories

Fibrous proteins

Globular proteins

Complexes with multiple components including proteins

Fibrous Proteins – Cytoskeletal Proteins

Actin

Coronin

Dystrophin

Keratin

Tubulin

Collagen

Elastin

Fibronectin

Globular Proteins - Major Types

Plasma proteins

Hemoproteins

Cell adhesion

Transmembrane transport proteins

Hormones and growth factors

Receptors

DNA-binding protein

Immune system proteins

Nutrient storage/transport

Chaperone proteins

Enzymes

Complexes with multiple components including proteins

Nucleosome

Ribonucleoprotein (generic)

Signal recognition particle

Spliceosome

Types of Proteins - Conclusion

Three major categories

Fibrous proteins

Globular proteins

Complexes with multiple components including proteins

Topic 13

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

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

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

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

(Iossifov, et al. 2014)

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 Analysis

Following tasks;

Sequencing

Assembly

Repeat identification and masking out

Gene prediction

Looking for EST and cDNA sequences

Genome annotation

Expression analysis

Metabolic pathways and regulation studies

Functional genomics

Gene location/gene map identification

Comparative genomics

Identify clusters of functionally related genes

Evolutionary modeling

Self-comparison of proteome

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

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

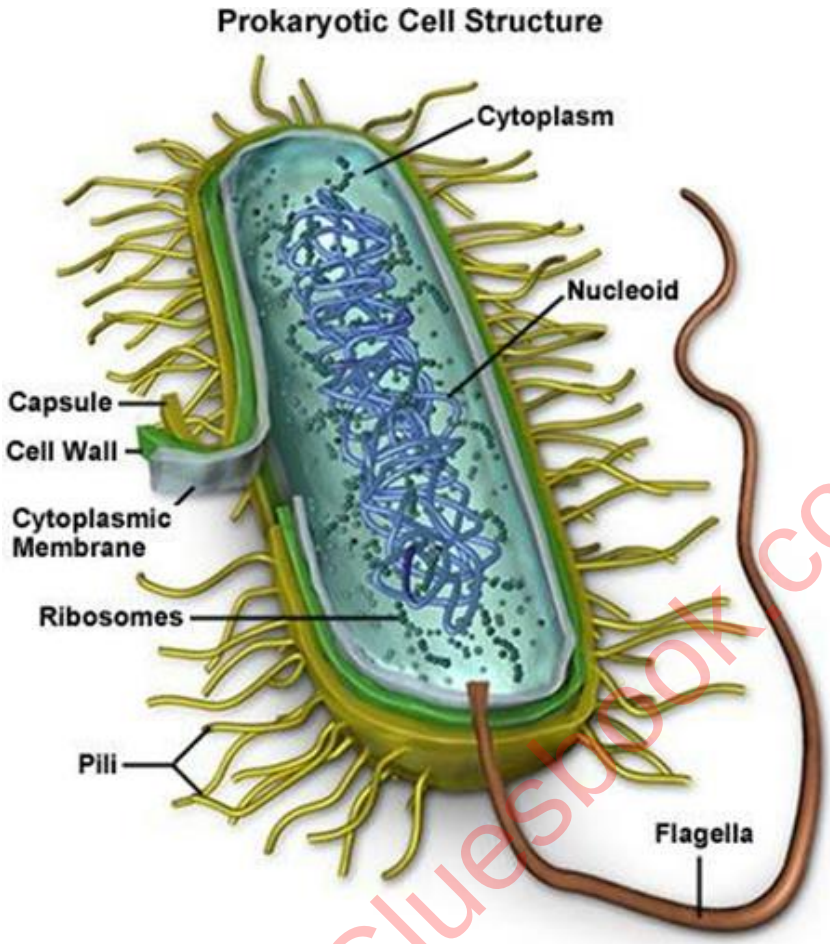
Topic 14

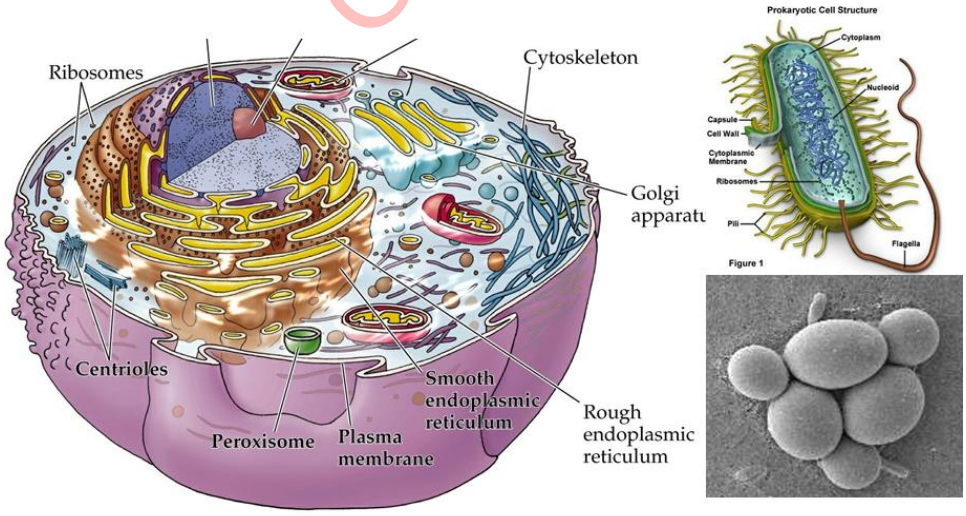
Prokaryotic Genome

Prokaryotes are the organisms whose Genetic material (DNA) is not enclosed in a nuclear membrane

No membrane bound organelles

Prokaryotic Genome





Mitochondria evolved from a bacterial endosymbiont

First prokaryotic Genome sequenced was that of *Hemophilus influenzae*

Paved the way for sequencing of many other organisms

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

important human pathogens

They were of phylogenetic interest

Sequences were annotated as they were sequenced

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

Conclusions

Prokaryotes are simple Genomes

Easy models to study Biochemistry and Molecular biology of life processes

Sequencing is done on economically important organisms

Topic 15

Eukaryotic Genome

Introduction

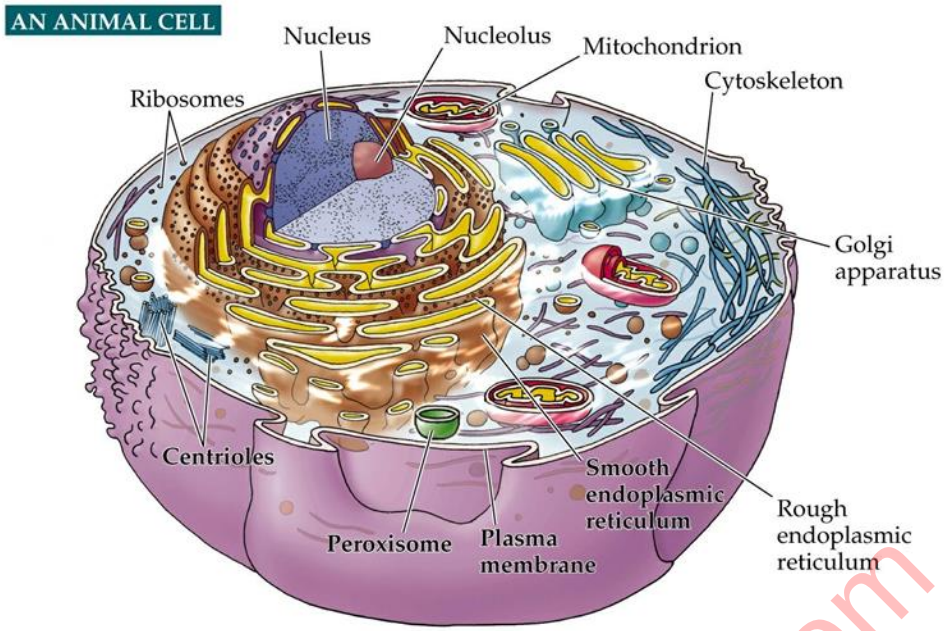
Eukaryotes have larger genomes

Have tandem repeats

Have introns in their protein-coding genes

Heterochromatin and euchromatin region

Eukaryotic Genome



Eukaryotic Genome

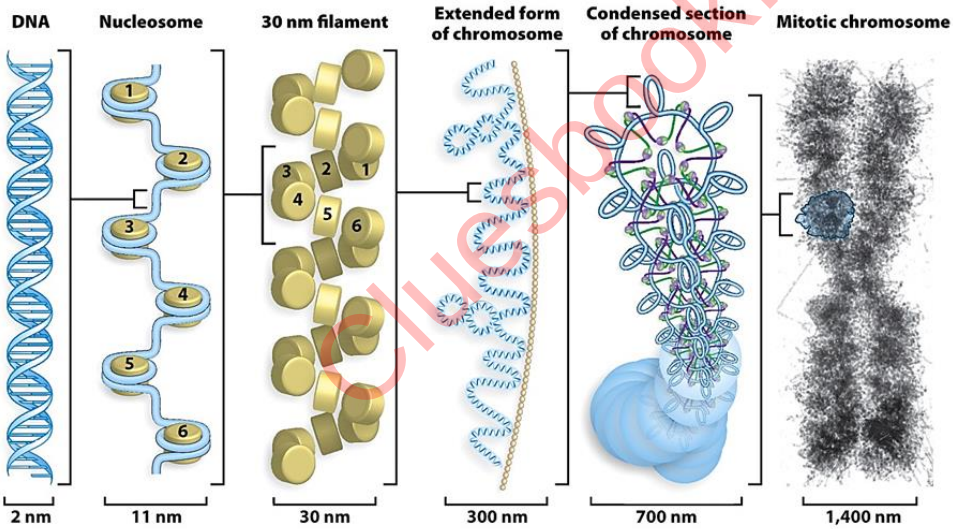


Figure 10-16
Molecular Biology: Principles and Practice

Eukaryotic Genome

Staining with dyes

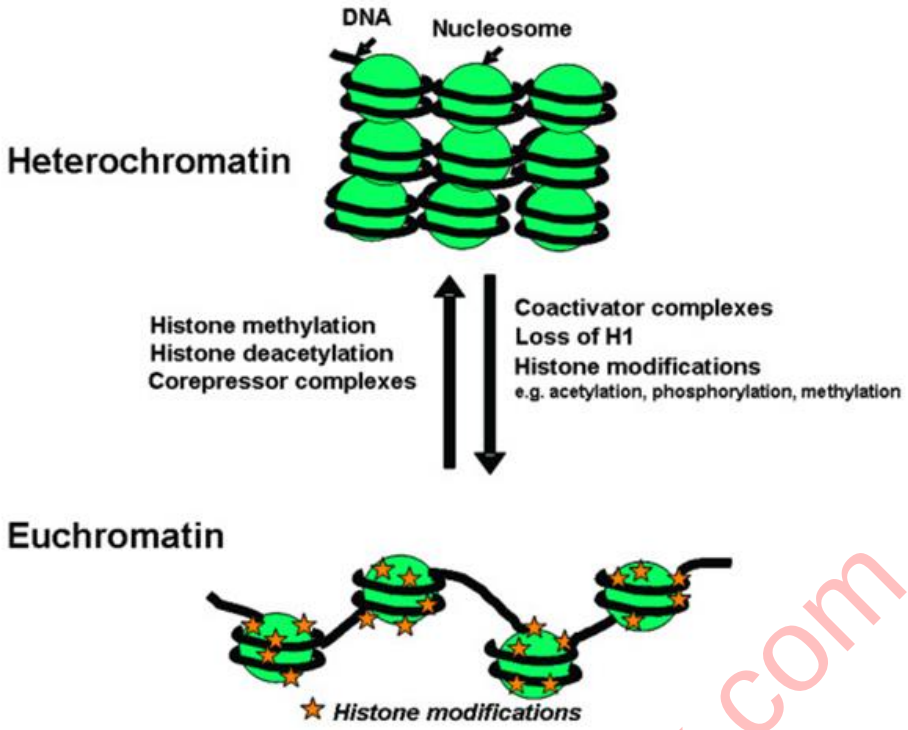
Dense heterochromatin

Light Euchromatin

Heterochromatic regions are poorly transcribed (expressed)

Euchromatic regions are highly expressed

Eukaryotic Genome



Eukaryotic Genome

Conclusions

Eukaryotes are distinguished by the presence of prominent nuclei

have larger genomes, tandem repeats and introns in their protein-coding genes

Topic 16

Epichromosomal Elements (EEs)

Introduction

Genome is the total collection of genetic material and is made up of

–Chromosomes

–Epichromosomal Elements

Prokaryotic EEs

Plasmids

- Self replicating
- Additional rings

Bacteriophages

- Host colonization

Transposons

- Parasitic DNA elements

Eukaryotic EEs

Organellar DNA

- Mitochondria
- Chloroplasts

Plasmids

- yeast

Transposons

- Viral genomes
- Retroviruses

membrane-bound organelles

- hundreds to thousands of them
- Site of respiration (mitochondria) or photosynthesis (chloroplast)
- mtDNA
- CpDNA

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

Organelle Genome

Circular

Double stranded

Supercoiled

No histones

Multiple copies

of mtDNA/cell or cpDNA/cell

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

Organelle Genome

Encode only a subset of genes required to elaborate a functional organelle

rRNAs, tRNAs, ribosomal proteins, membrane-associated respiratory or photosynthetic components

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.

Conclusions

Organelle genome is similar to prokaryotes

High copy number

mtDNA and cpDNA depends on Nuclear DNA

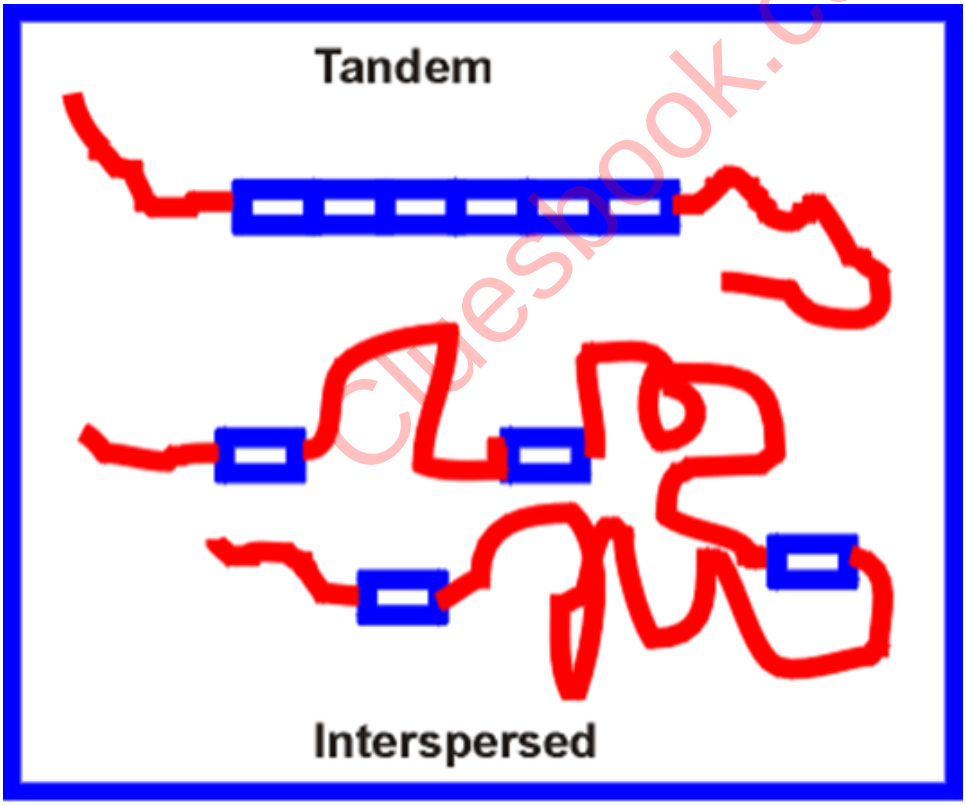
Topic 17

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



Satellite DNA

One to several thousand bp long

Tandem array of 100 million bases long

Near centromere and telomere

Minisatellite and Nicrosatellite

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

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

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

Have a potential to increase in number

Make up a large proportion of eukaryotic genome

Detectable until blend into genome due to mutations

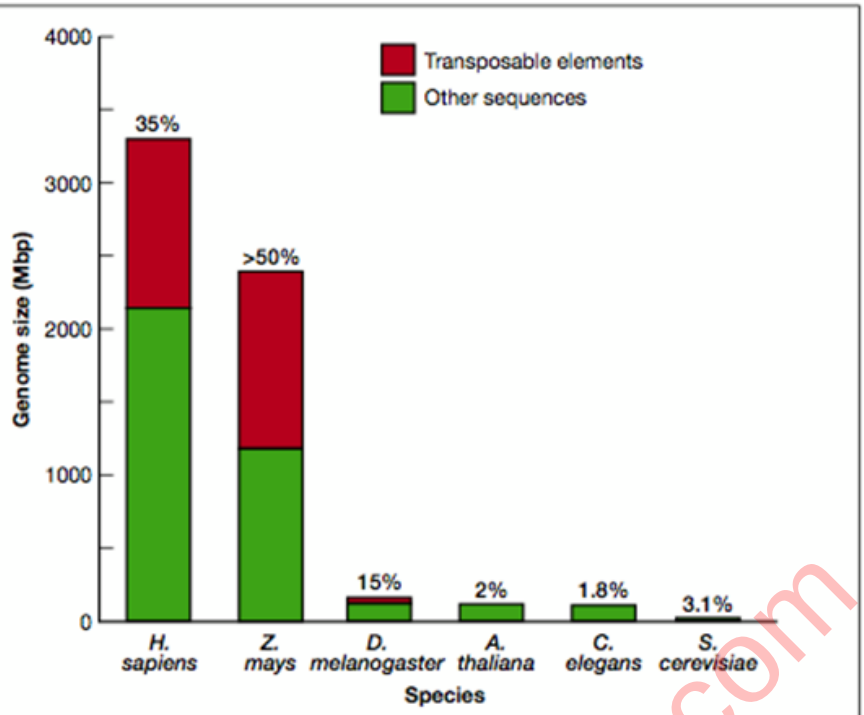


Figure 10.1. Percentages of representative genomes that are made up of transposable elements. The genomes include those of humans, maize, the fruit fly *Drosophila*, the model plant *Arabidopsis*, the nematode *C. elegans*, and budding yeast *S. cerevisiae*, respectively. (Adapted, with permission, from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

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

Topic 18

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

IS elements

usually very small (< 1 Kb to 2 Kb)

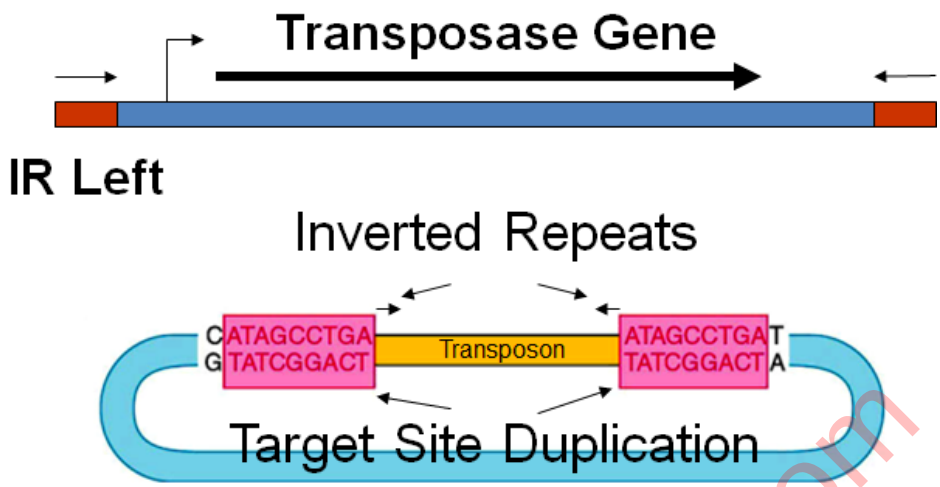
Flanked by inverted repeat sequences (IRs)

Encode at least one gene that provides their own transposition functions

Do not code for noticeable (phenotypic) traits

Can cause mutations by transposition into genes

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

Eukaryotic TEs

2 Classes; Class I and Class II

Class I

Encode reverse transcriptase

Use RNA-mediated mechanisms of transcription

Class I

Three classes

LTR (Long Terminal Repeats) retrotransposons

Retrotransposons

Retrovirus like Elements

SINES (Short Interspersed Nuclear Elements

80-300 bp

***Alu* repeats**

LINES (Long Interspersed Nuclear Elements)

6-8 KB long

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)

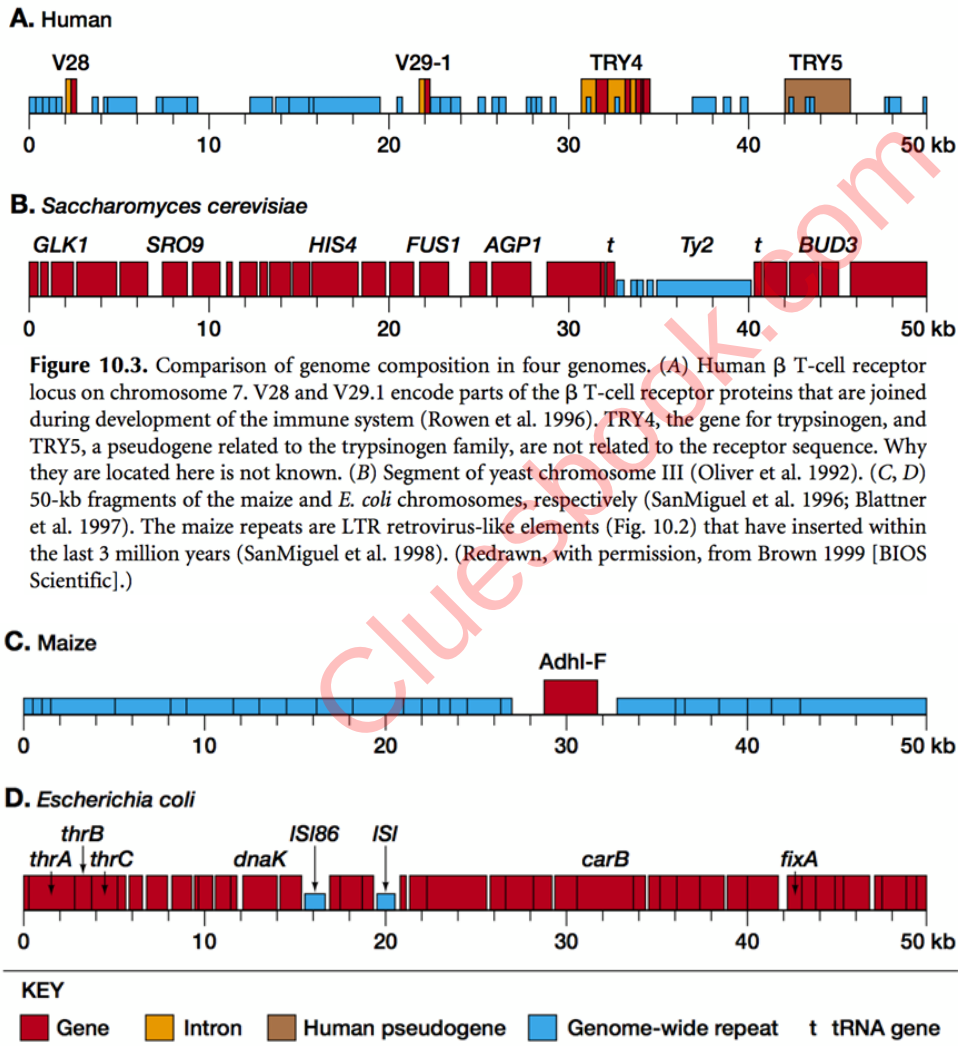


Figure 10.3. Comparison of genome composition in four genomes. (*A*) Human β T-cell receptor locus on chromosome 7. V28 and V29.1 encode parts of the β T-cell receptor proteins that are joined during development of the immune system (Rowen et al. 1996). TRY4, the gene for trypsinogen, and TRY5, a pseudogene 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].)

Conclusions

Transposable elements make up a significant part of eukaryotic genome

Move within and across genomes

Cause genome expansion

Topic 19

Eukaryotic Gene Structure

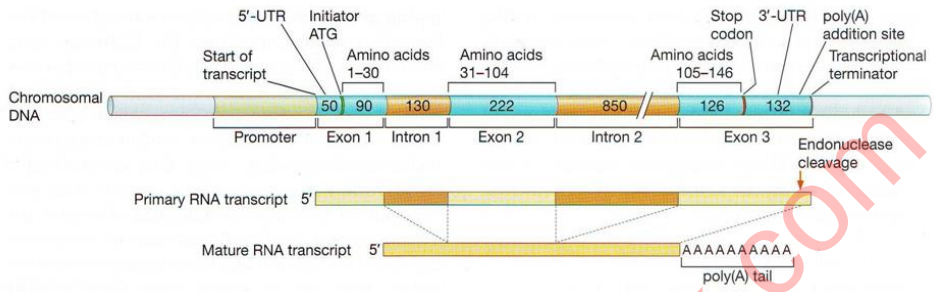
Eukaryotic genes

Eukaryotic genes are complicated

They possess exons which are protein coding regions and are interrupted with introns

Whole gene is transcribed as a large transcript

Introns are spliced leaving an ORF that is translated



Human β-globin gene

Introns start on the 5' ends as GT and end as AG towards 3' end **GT----------AG**

This trend is highly conserved

Introns proportion

A small proportion in yeast, only 239 introns in genome

Human genes have hundreds, often 95% of genes

Might have embedded genes

Intron origin

Intron-early

used to assemble first genes from ancient conserved introns

Intron-late

Broke up previously continuous genes by inserting into them

Number of Genes

Degree of compactness contributes to variation in gene density

Compact genomes have higher genome density

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

Pseudogenes

Non functional genes

Derived from functional genes through mutations following genome duplication

Processed pseudogenes lack introns and promoters

Gene families

Set of genes having similar sequences and functions

Gene families arise from gene duplication and subsequent divergence

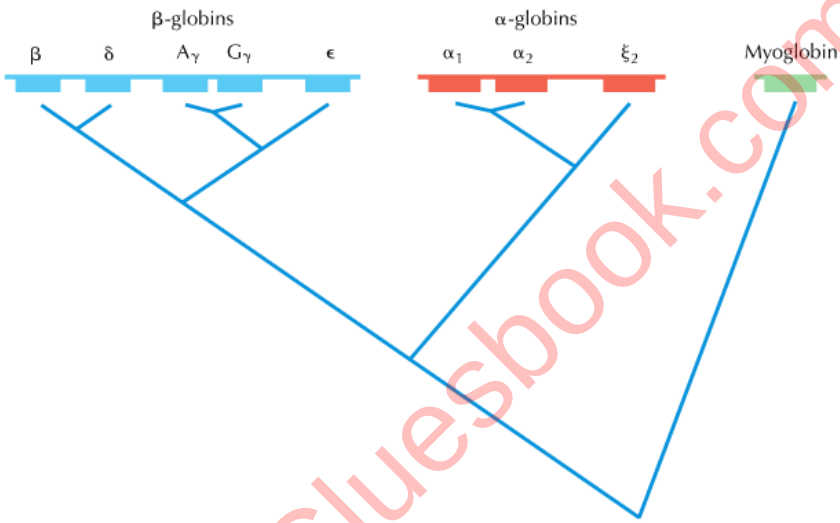


FIGURE 27.29. Gene duplications during the evolution of the human globin gene families. The initial split gave rise to two lineages, one leading to the modern gene for myoglobin and the other to the globin genes. Subsequently, the proto-α-globin and proto-β-globin lineages split following a duplication. 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

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

Topic 20

Comparative Genomics

Introduction

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

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

Orthologs

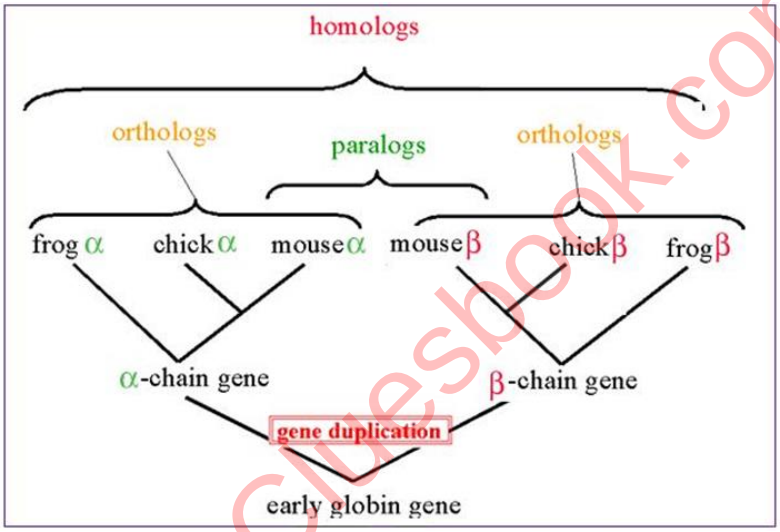
Genes in two organisms that are so similar that they must have the same function and evolutionary history are called orthologs (Fitch, 1970)

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



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

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

Drosophila and Human

Some human disease genes absent in *Drosophila*

A number of previously unknown counterparts to human cancer and neurological disorders genes

Conclusions

Comparative genomics reveals the relationship among different organisms

Fruit fly has more similarities with mammals

Topic 21

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.

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 relies on Molecular Biology, Biochemistry, Genetics and Bioinformatics tools

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

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 – what is included

Subtracted cDNA libraries

Differential display

Representational difference analysis

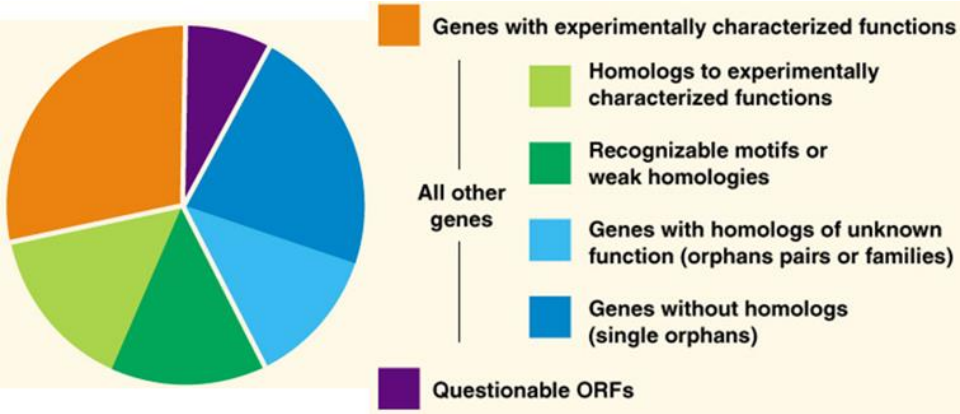
Suppression subtractive hybridization

cDNA Microarrays

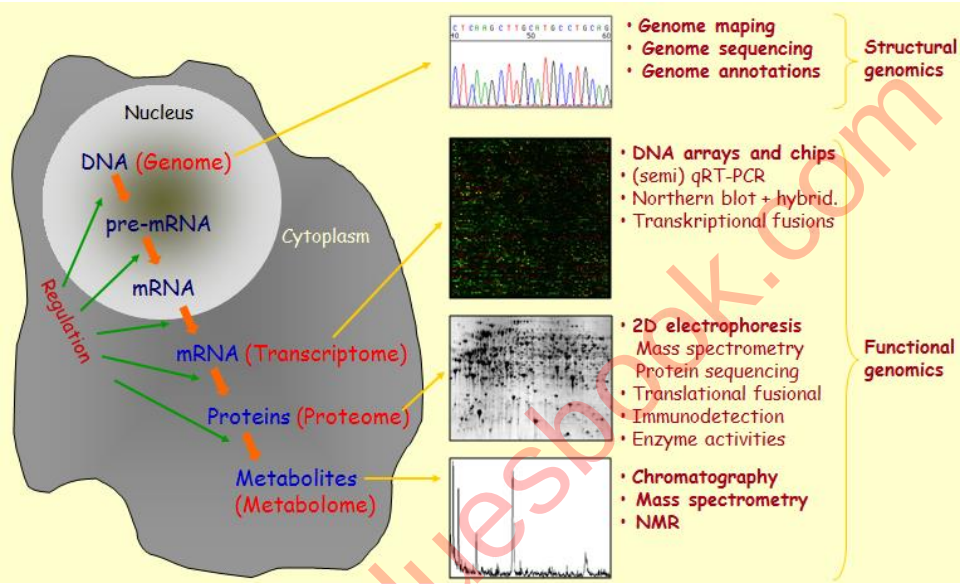
Serial analysis of gene expression

2-D Gel electrophoresis

Yeast Genome with Experimentally Characterized Functions



Functional Genomics



Functional Genomics - Conclusion

Function of gene products.

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

Considering phylogenetic relationships.

Topic 22

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.

Genetic Mapping

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

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 incudes

Distinct components of genomes

Abundance and complexity of mRNA

Genome sequences

Gene numbers

Coding and non-coding DNA

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

Structure of Complex Genomes

Introns in genes

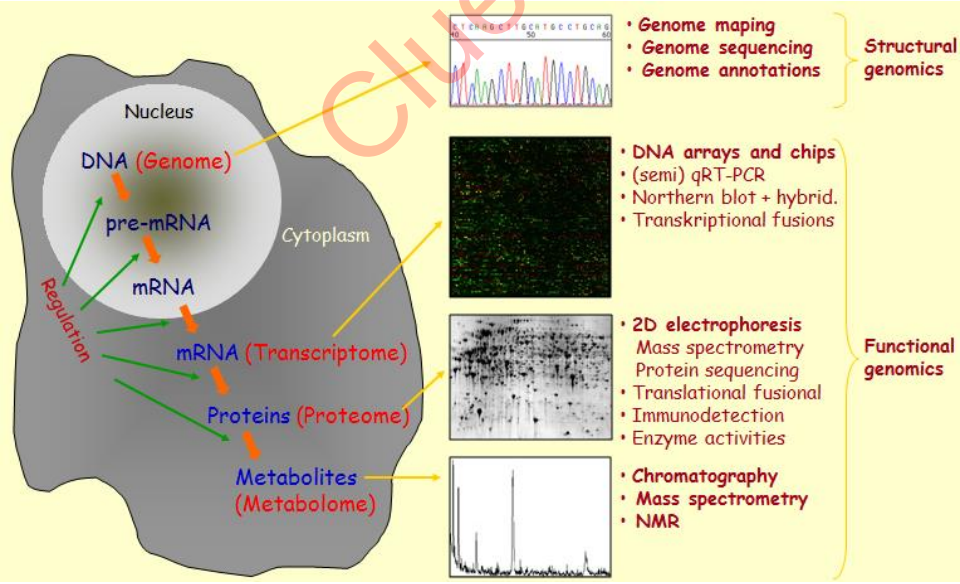
Regulatory elements of genes

Multiple copies of genes, including pseudogenes

Intergenic sequences

Interspersed repeats

Structural Genomics



Structural Genomics – Transposable Elements

Vast majority of TEs can be classified into four families:

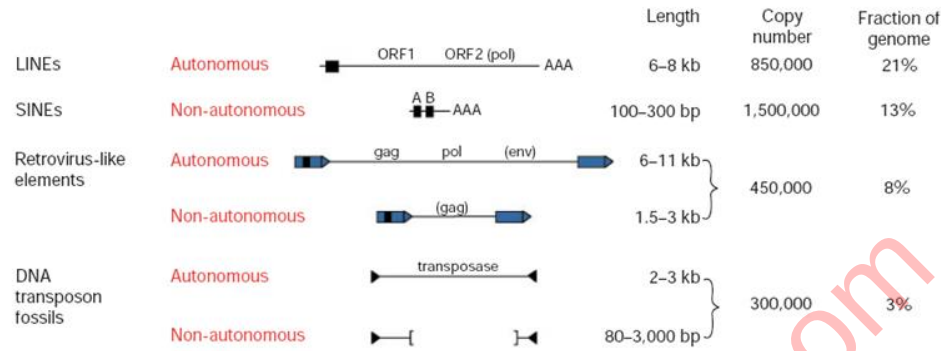
LINEs (Long Interspersed Nuclear Elements, autonomous)

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 – Repetitive DNA in Humans



Structural Genomics - Conclusion

To determine the nucleotides sequences of genomes.

To study genetic maps

To study physical maps.

Topic 23

Comparative Genomics

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

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

Homology

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

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.

Homologoues can be

Orthologues

Paralogues

Xenologues

Analogues

Orthologues

Orthologues are homologues that have evolved from a common ancestral gene by speciation.

They usually have similar functions.

Paralogues

Paralogues are homologues that are related or produced by duplication within a genome followed by subsequent divergence.

They often have different functions.

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.

Analogues

Analogues are non-homologues genes/proteins that have descended convergently from an unrelated ancestor.

Similar function, different sequence or structure.

Comparative Genomics – Evolutionary Trends

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

Comparative Genomics – Functional elements

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

Comparative Genomics – Ancestral Genome

Reconstruction of an ancestral genome for a group of organism.

Conclusion

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

Topic 24

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.

Deeper insights about genomic size.

Further complexities in the genomes of the organisms.

Variations between individuals or strains.

Human genome is 200 times larger than yeast but 200 times smaller than Amoeba.

Less than 2% of human genome is coding sequence

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.

Any two humans are more than 99% identical at genetic level.

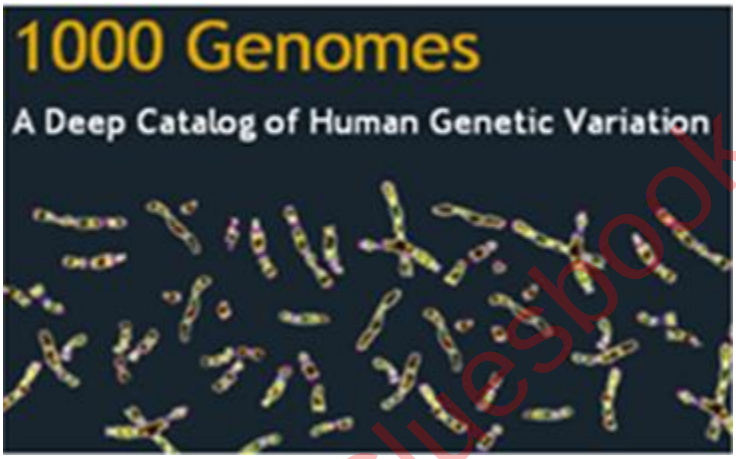
Genetic variations may explain individual differences in susceptibility to diseases, responses to drugs.

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



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.

Provide better understanding of very rare genetic diseases.

Understand contribution of common variants to common diseases like diabetes and heart diseases.

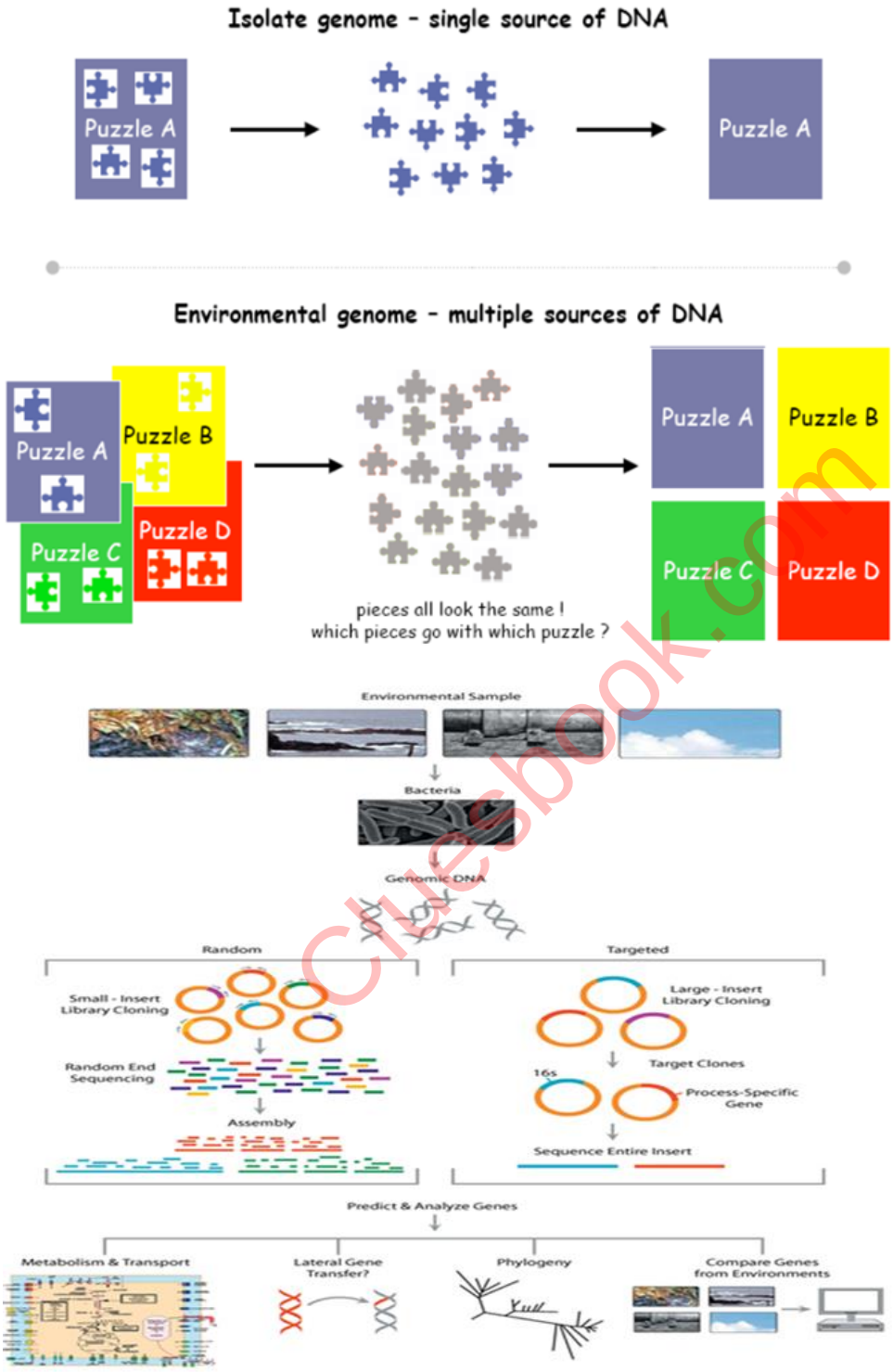
Identify SNP but also large differences like rearrangements, deletions or duplications

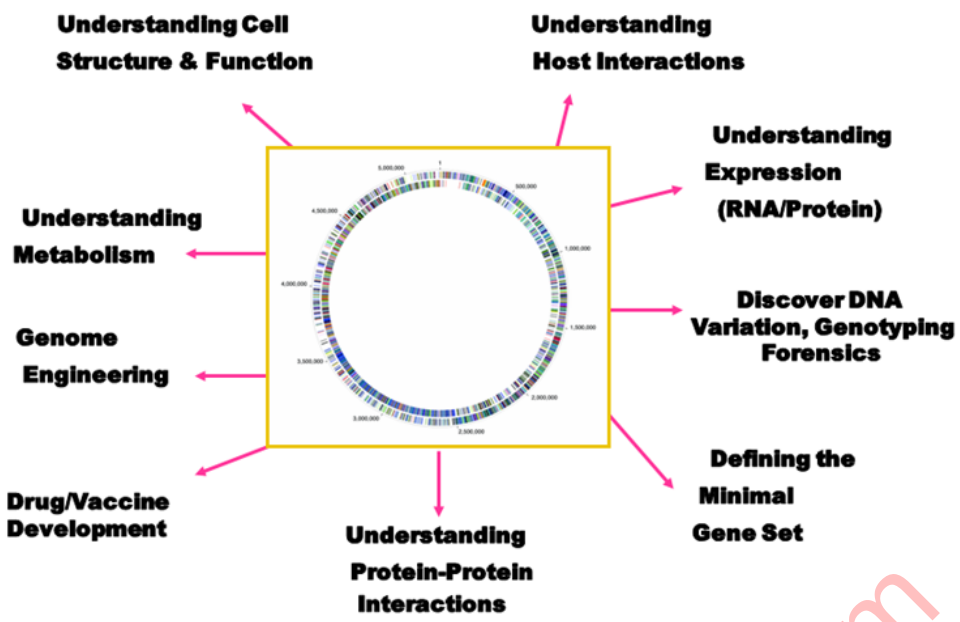
Topic 25

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.





Objectives of Metagenomics

Examining phylogenetic diversity using 16s rRNA

Diversity patterns of microorganisms for monitoring and predicting environmental conditions/change.

Examining genes/operons for desirable enzymes (cellulases, lipases, antibiotics, other natural products).

Exploited for industrial or medical applications.

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

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

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.

Examining metabolic pathways.

Directed approach towards designing culture media.

Examining genes that predominate in a given environment compared to others.

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.

Conclusion

Metagenome-environmental genome

Topic 26

Why Sequence Genomes

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

Genes regulation

DNA sequence

Genome organization

Chromosomal structure and organization

Noncoding DNA types, amount, distribution and functions.

Coordination of gene expression, protein synthesis, and post-translational events.

Interaction of proteins in complex molecular machines

Predicted vs experimentally determined gene function

Evolutionary conservation

Proteins structure and function.

Proteomes (total protein content and function) in organisms.

Correlation of SNPs with health and disease

Disease-susceptibility prediction based on gene sequence variation

Genes involved in complex traits and multigene diseases

Novel Diagnostics

Complex systems biology, developmental genetics.

To provide platform for microchips and DNA microarrays.

Gene expression - RNA

Complex systems biology, developmental genetics and genomics

Novel Therapeutics

Drug target discovery

Rational drug design

Molecular docking

Gene therapy

Stem cell therapy

Understanding Metabolism

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

Understanding mechanism of diseases

Inherited diseases

Infectious diseases

Pathogenic bacteria

Viruses

Conclusions

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

Topic 27

Major Techniques used for Genomes Characterization

Cloning

Hybridization

PCR amplification

Sequencing

Computational tool

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.

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

Genomes Characterization - Techniques

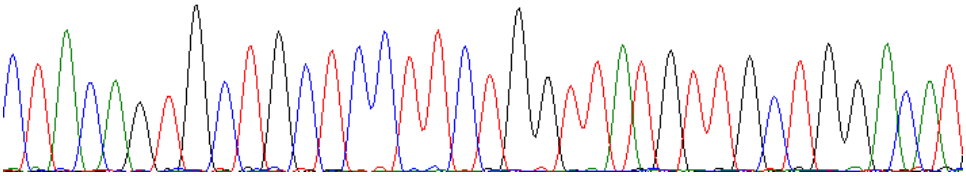
Genomes Characterization Techniques – PCR

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

Genomes Characterization - Techniques

Genomes Characterization Techniques – DNA Sequencing

C T A C A G T G C T G C T C C T T C T G G T T A T G T T G C T G G A C A T



One of the important technique used to characterize the genomes

To study structure and function of genomes.

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

Genomes Characterization Techniques – Conclusion

Different techniques used for genomes characterizations.

Topic 28

Steps of Genomes Analysis

Genome sequence assembled

Identify repetitive sequences – mask out

Gene prediction – train a model for each genome

Steps of Genomes Analysis

Look for EST and cDNA sequences

Genome annotation

Microarray analysis

Metabolic pathways and regulation

Protein 2D gel electrophoresis

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

Topic 29

Steps of Genomes Analysis

Genome sequence assembled

Identify repetitive sequences – mask out

Gene prediction – train a model for each genome

Genomes Analysis - Steps

Look for EST and cDNA sequences

Genome annotation

Microarray analysis

Metabolic pathways and regulation

Protein 2D gel electrophoresis

Functional genomics

Gene location/gene map

Self-comparison of proteome

Comparative genomics

Identify clusters of functionally related genes

Evolutionary modeling

Topic 30

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

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

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

Conclusion

•Although most eukaryotes have large size of genomes.

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

Topic 31

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.

Almost all plants viruses and some bacterial and animal viruses

Genomes are rather small (a few thousands nucleotides)

Often a circular genome

lambda = 48,502 bp

Replicative form of Viral Genomes

All ssRNA viruses produce dsRNA molecules

Many linear DNA molecules become circular

Viruses and Kingdoms

Many plants viruses contain ssRNA genomes.

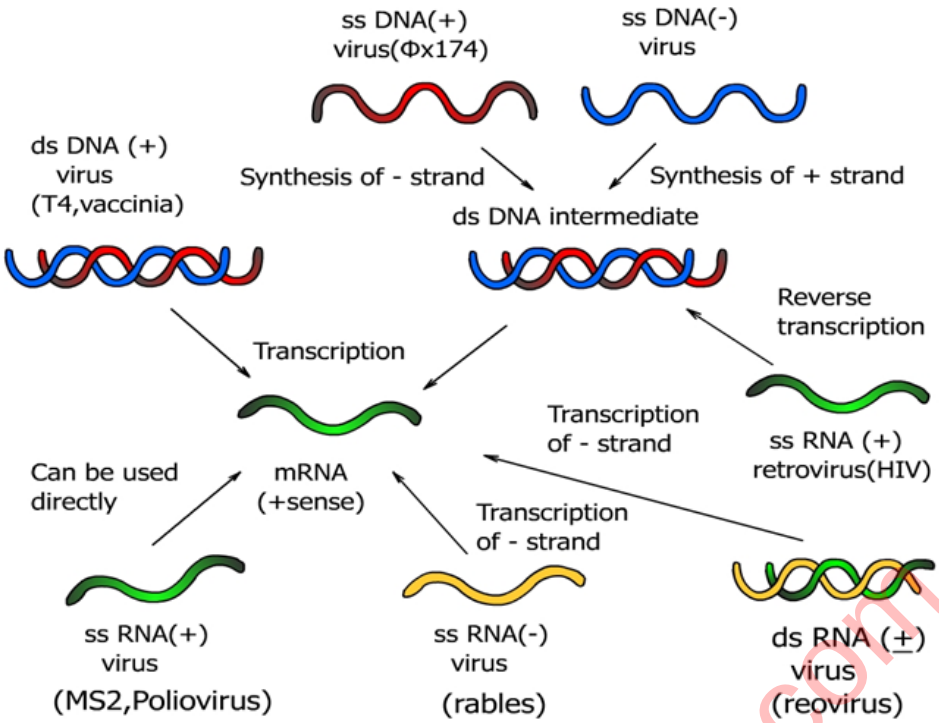
Many fungal viruses contain dsRNA genomes.

Many bacterial viruses contain dsDNA 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.

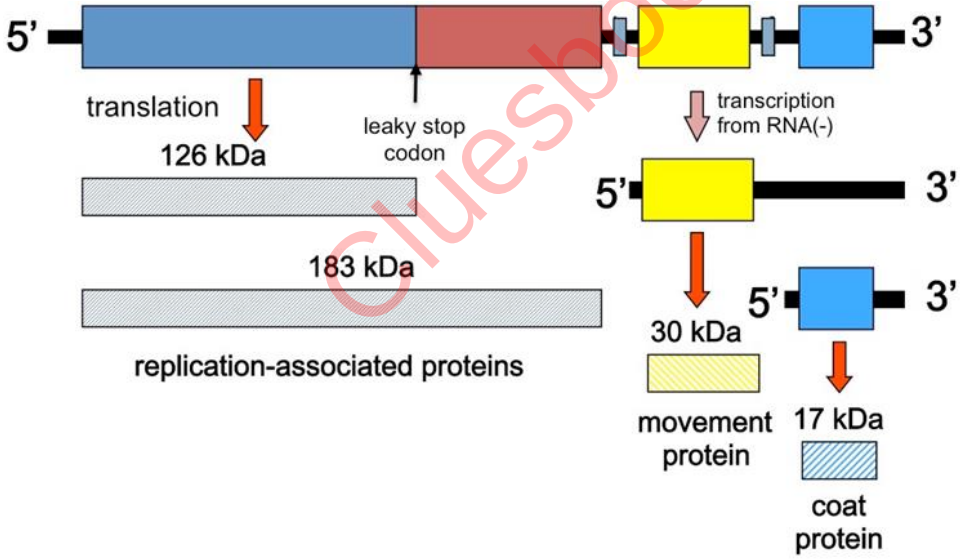
Virus	Host	Type of Nucleic Acid	Number of Genes
Parvovirus	Mammals	ssDNA	5
Phage fd	*E. coli*	ssDNA	10
Lambda	*E. coli*	dsDNA	36
T4	*E. coli*	dsDNA	>190
Qβ	*E. coli*	ssRNA	4
TMV	Many plants	ssRNA	6
Influenza virus	Mammals	ssRNA	12

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



Genome of Tobacco Mosaic Virus

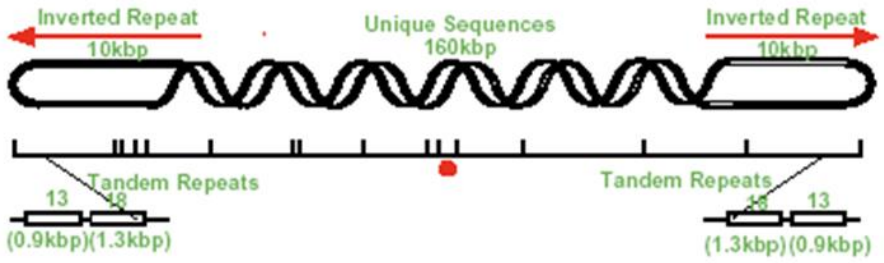
•Single, 6400 nucleotides RNA, 3 Essential Genes



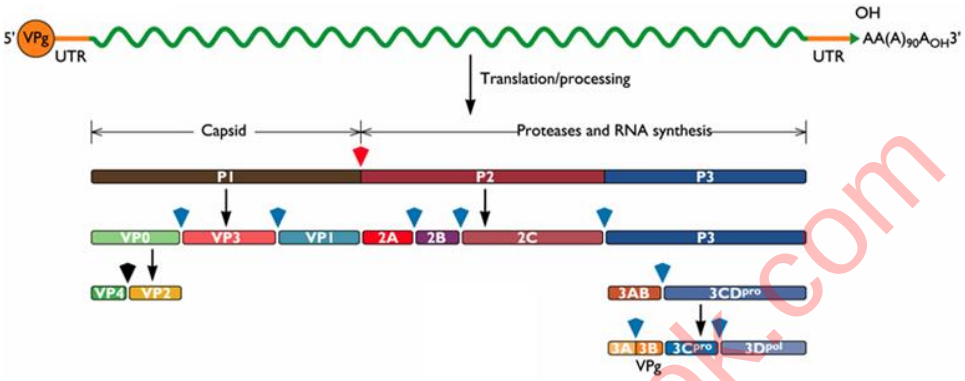
•Viral Genomes

Genome of Poxvirus –A typical large dsDNA Virus

•180 kb DNA, >100 Essential Genes



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.

Topic 32

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

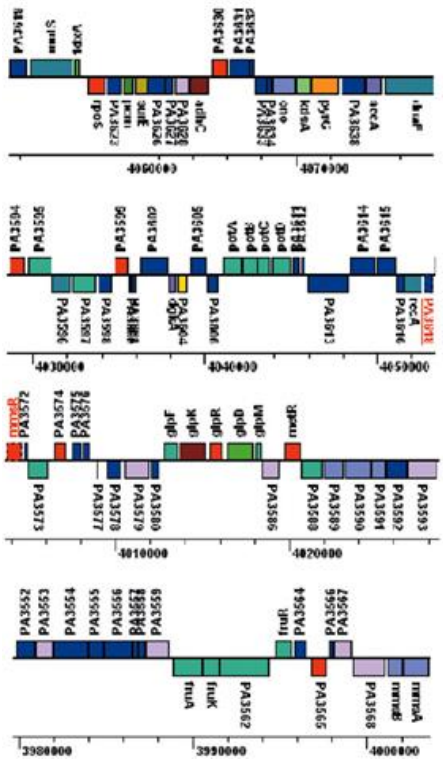
11599 coding sequences

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

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%)

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 - Conclusion

Small organisms carry high coding density.

Large variation in genome size between bacteria.

Topic 33

Yeast Genome

The nuclear genome consists of 16 chromosomes.

In addition, there is a mitochondrial genome and a plasmid, 2 micron circle.

The haploid yeast genome consists of ~ 12.1 Mb

Yeast genome was completely sequenced by 1996

Yeast Genome - Characteristics

Small and compact

Small intergenic sequences

Few transposable elements

Few introns

Limited RNA interference

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

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: Genome of Yeast Cell

The largest known regulatory sequences are spread over about 2,800bp

MUC1/FLO11

Yeast genes have names consisting of three letters and up to three numbers

GPD1, HSP12, PDC6

Usually they are meaningful

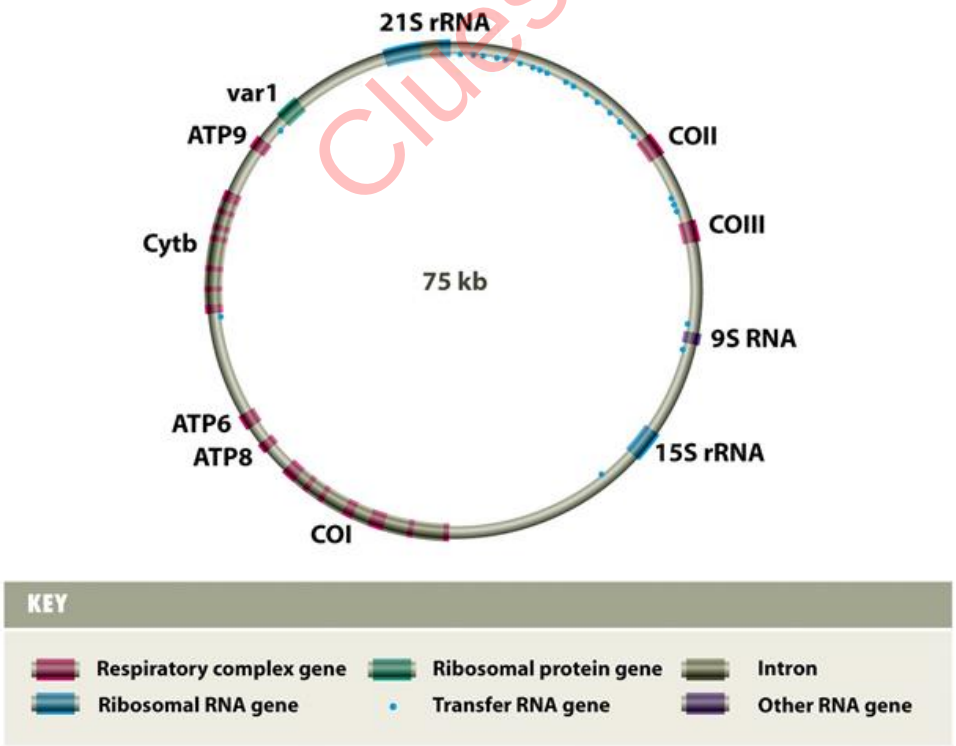
Yeast Genome – Genes Nomenclature

Wild type genes are written with capital letters in italics: *TPS1, RHO1, CDC28*

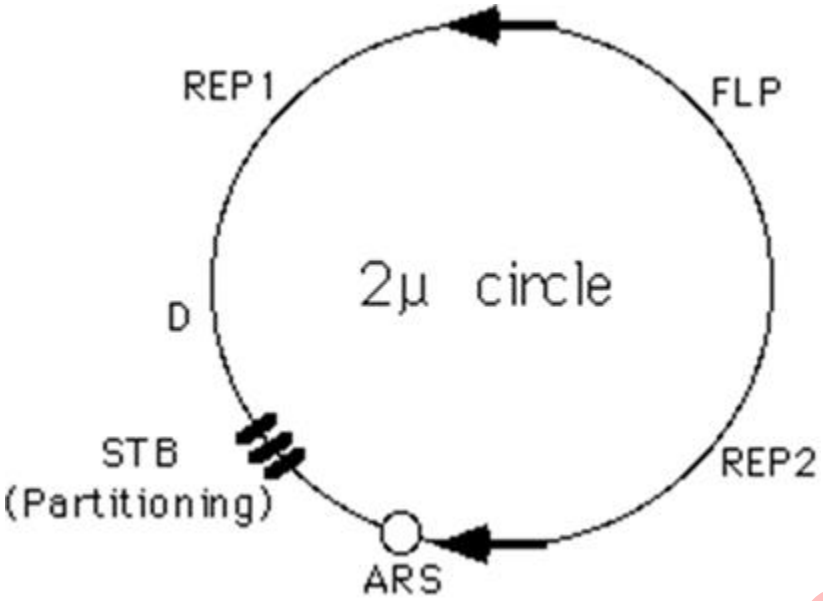
Recessive mutant genes are written with small letters in italics: *tps1, rho1, cdc28*

Three letters provides information about a function, mutant phenotype, or process related to that gene.

CDC - Cell Division Cycle ; ADE-ADEnine biosynthesis



Yeast – Mitochondrial DNA



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 nuclear genome has 16 chromosomes.

A mitochondrial genome.

A plasmid.

Topic 34

Topic 35

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 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

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

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

Humans and Mammals have low gene density

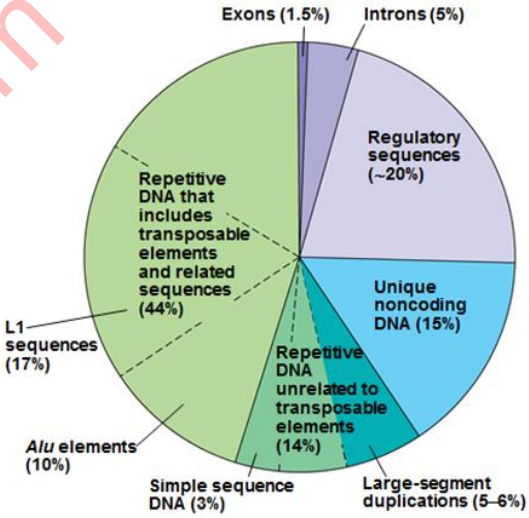
Humans and other mammals have the lowest gene density, or number of genes, in a given length of DNA.
Multicellular eukaryotes have many introns within genes and noncoding DNA between genes

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

Human Genome: Distribution of coding and non-coding DNA



Comparing Genomes

Significant similarity between genomes of "distant" species (Man – Yeast 23%)

Similarity increases for taxonomically close species.

Closely related species help us understand recent evolutionary events

Distantly related species help us understand ancient evolutionary events

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

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

Topic 36

Genetics and Genomics

Comparing distantly/closely 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.

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

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

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

Several genes are evolving faster in humans than chimpanzees

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

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

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

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.

Topic 37

Genome Mapping

Genome mapping

Different types

Genetic mapping

Physical 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.

Genetic techniques include cross-breeding experiments

Case of humans, the examination of family histories (pedigrees).

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.

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

Different types

Genetic mapping

Physical mapping

Topic 38

Genetic Mapping of Genomes

Linkage analysis is the basis of genetic mapping

Recombination fraction is a measure of the distance between two loci.

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

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 - 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 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

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

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

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.

Recombinations Frequencies (RF)

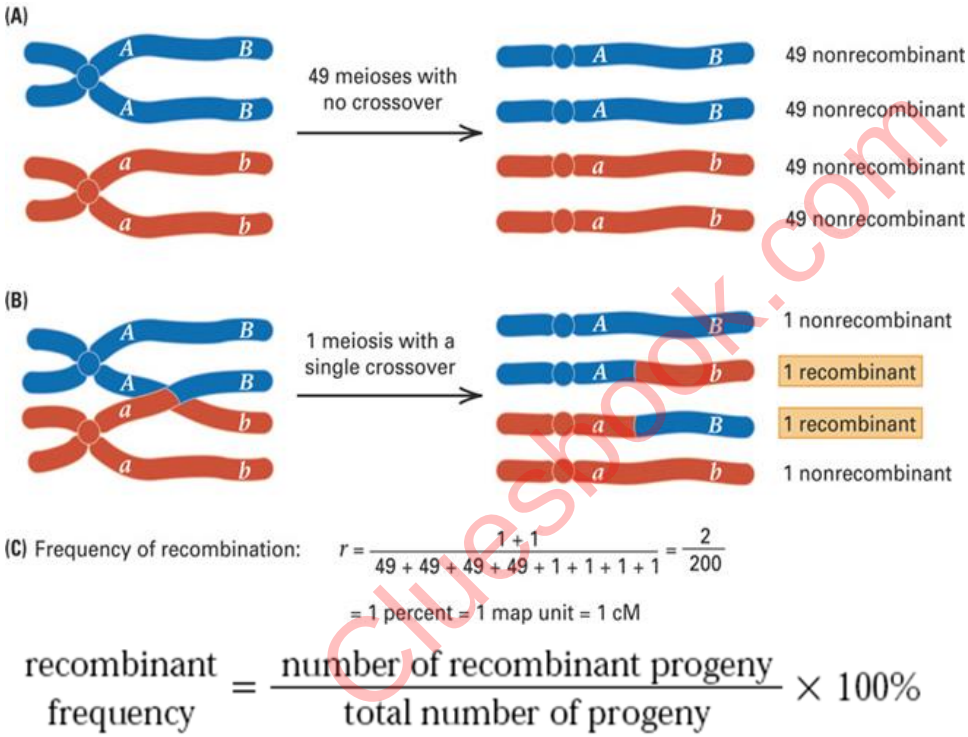
Two genes that undergo independent assortment have RF of 50 % and are located on non-homologous chromosomes

They are unlinked

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

They are linked

Genetic Mapping of Genomes – Recombinations Frequency Calculated as:



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

Genetic Mapping of Genomes

The LOD score is calculated as follows:

$$\text{LOD} = Z = \text{Log}_{10} \frac{\text{probability of birth sequence with a given linkage}}{\text{probability of birth sequence with no linkage}}$$

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 – Recombinations frequency which is calculated as follows:

Linkage analysis is the basis of genetic mapping

Recombination fraction is a measure of the distance between two loci.

Topic 39

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

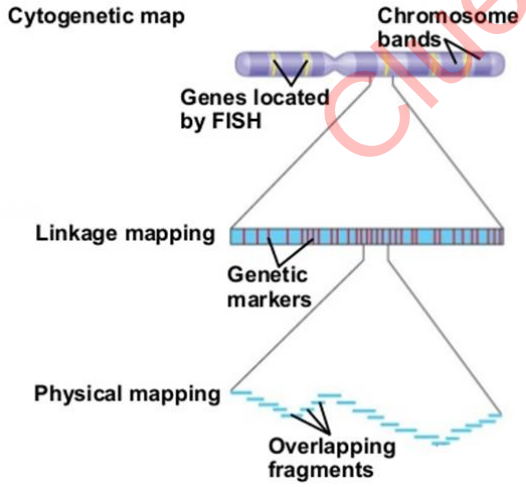
Physical Mapping of Genomes

Physical Mapping

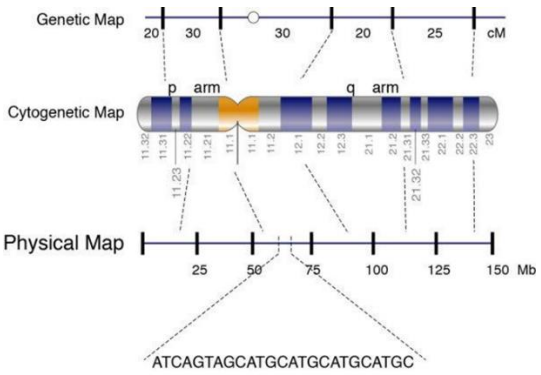
In humans, even the detailed genetic map lacks the required resolution.

Therefore, a physical map derived directly from genomic DNA rather than analysis of recombinants has been generated.

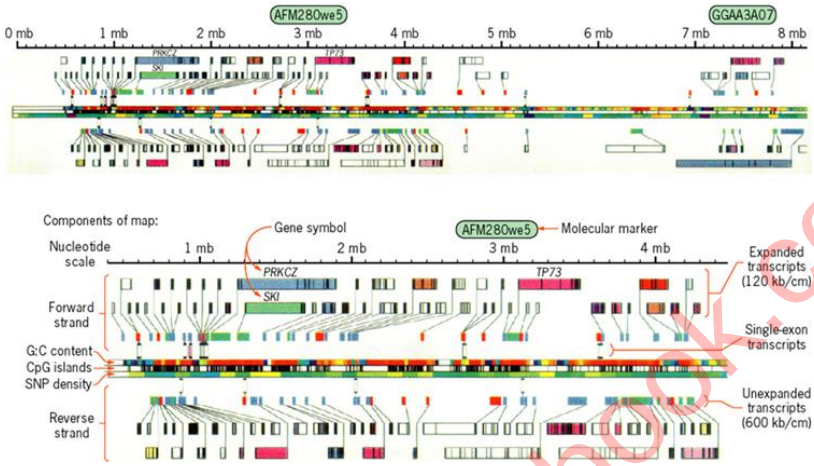
Cytogenetic Map, Genetic Map, Physical Map



Genetic Map and Physical Map



Physical Map



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.

Topic 40

Genetics and Genomics

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

Chromosomal banding pattern.

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

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

Visual study of chromosomes at microscopic level

Karyotype

Chromosome complement

– also applied to picture of chromosomes

Idiogram

Stylised form of karyotype

Classified according to position of centromere

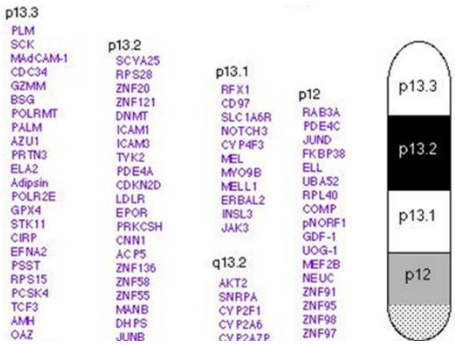
Central centromere - metacentric

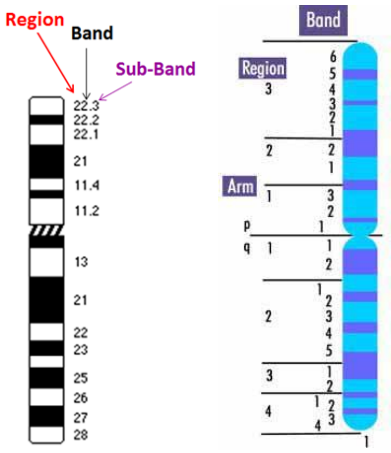
Sub-terminal centromere - acrocentric

have satellites which contain multiple copies of genes for ribosomal RNA on short arm

Intermediate centromere – submetacentric

Terminal centromere – Telocentric





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

Cytogenetic Mapping

Microscopic examination of stained chromosome reveals a pattern of bands.

Regions are designated based on their chromosomal position relative to the centromere

Topic 41

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.

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

Fluorescence microscopy used to find out where fluorescent probe is bound to the chromosomes.

Used in genetic counseling, medicine and species identification.

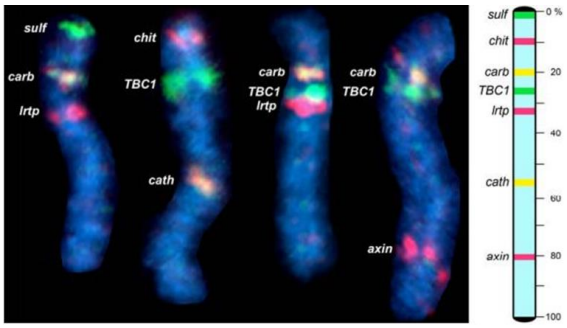
Individual metaphase chromosomes are probed in situ with specific fluorescently labeled DNA sequences, identifying homologous sequences in the chromosome

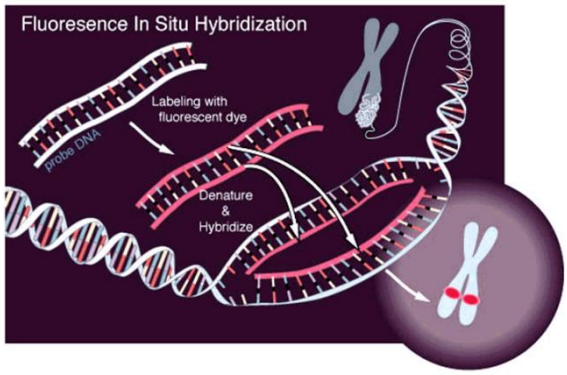
Different probes labeled with different fluorescent dyes may be used in same experiment.

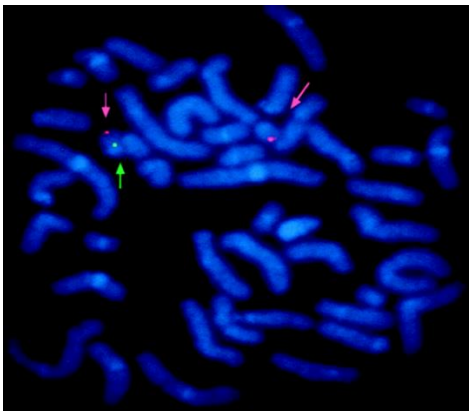
Fluorescence microscopy provides data for computer imaging analysis to determine binding site for each probe.

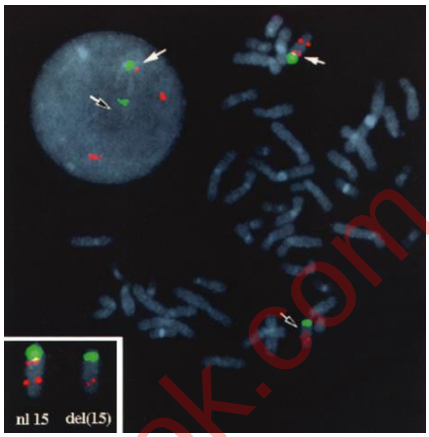
With a resolution of 2-5 Mb in metaphase chromosomes, FISH can localize markers to sub-regions of chromosomal bands.

Less condensed chromosomes may be resolved in the 5-700 kb range.









Advantages

Highly specific, Microdeletions /Microduplications

Disadvantages

500-600 probes needed to match the power of karyotyping

Topic 42

Genetics and Genomics

Restriction Mapping of Genomes

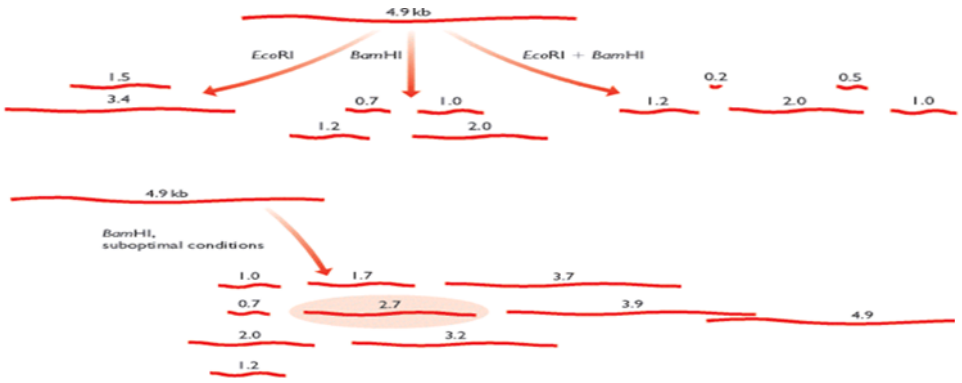
Restriction Mapping

A restriction map is a map of known restriction sites within a sequence of DNA.

Restriction mapping requires the use of restriction enzymes.

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

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.

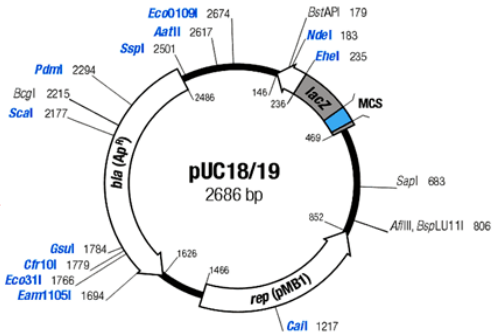


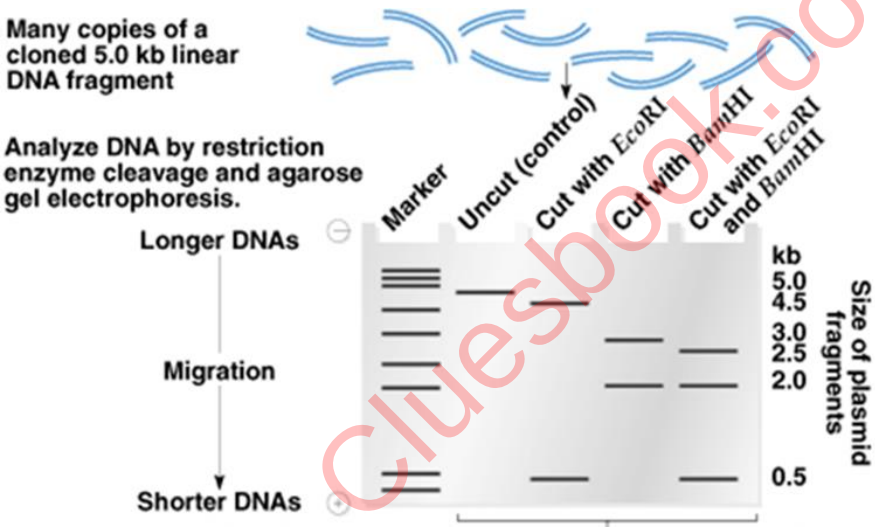
A map showing positions of restriction sites in a DNA sequence

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 is more applicable to small rather than large molecules.





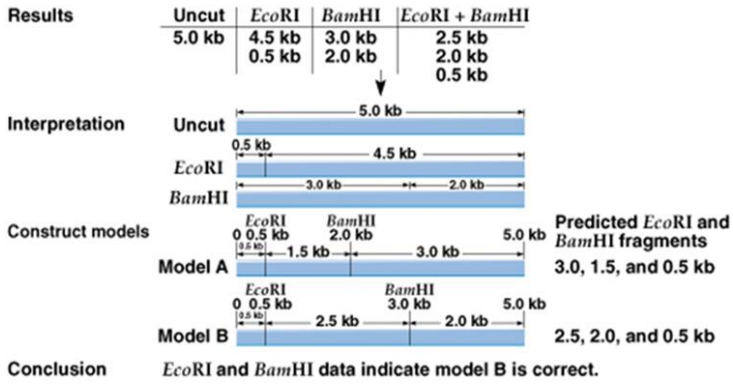
Restriction Mapping of Genomes

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

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.



Topic 43

Genetics and Genomics

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.

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

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

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.

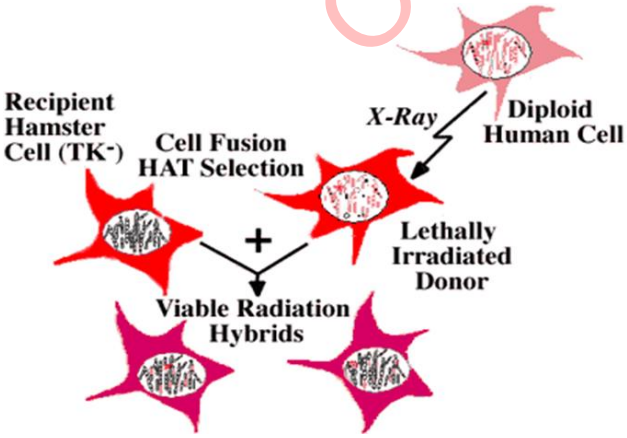
Radiation Hybrid Mapping: Methodology

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

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

Topic 44

Genetics and Genomics

Clone Contigs Mapping of Genomes

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

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

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

YAC clones are then assembled into a map.

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.

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

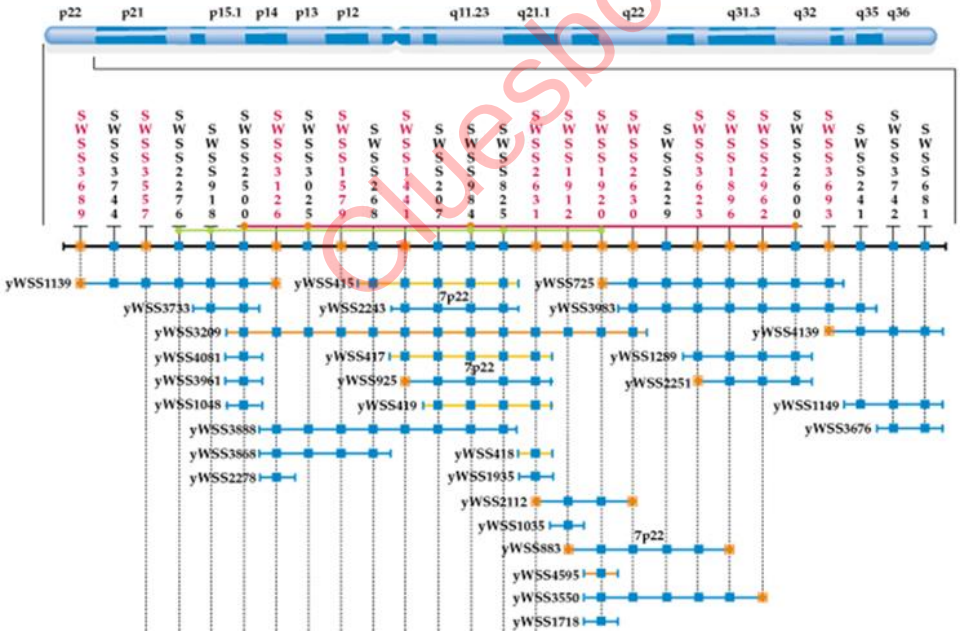
A complete library should yield a complete contig map that indicates the order in which the cloned fragments occur in the chromosome

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

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.

YAC Contigs Map of Chromosome 7



Clone Contigs Mapping of Genomes

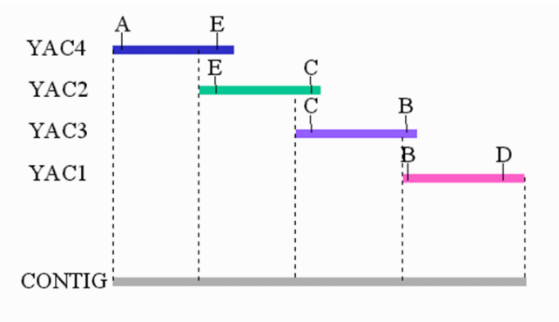
Clone Contigs Mapping of Genomes

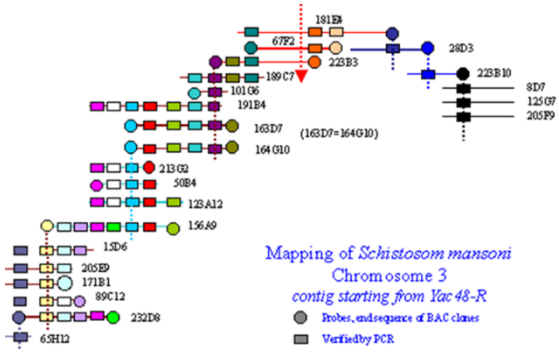
A contig is a set of partially overlapping clones, a contiguous set of clones. No gaps between them.

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: Yac48-R

Clone Contigs Mapping

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

Topic 45

Generating Sequence of Genomes

Dideoxy sequencing used to sequence the genomes

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

Two competing sequencing strategies used

Clone-by-clone

Whole-genome shotgun

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

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

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

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

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

Topic 46

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

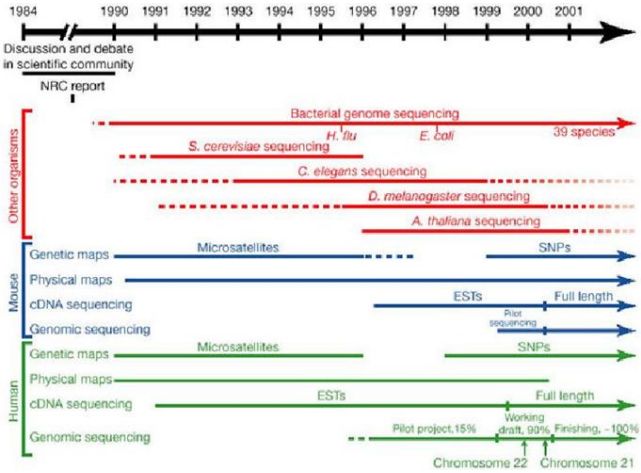
Human Genome Project

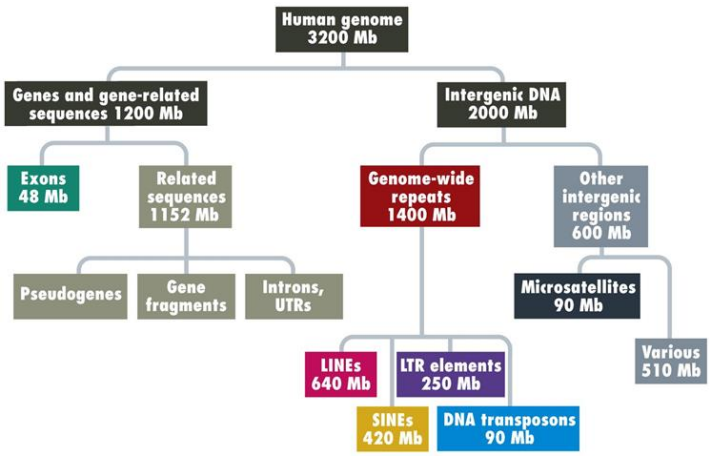
Human Genome Project - Objectives

Determine the nucleotide sequence of the estimated 3 billion base pairs

Analyze genetic variations in humans

Map and sequence the genomes of model organisms





Topic 47

Genetics and Genomics

HGP – Hierarchical Shotgun Sequencing

Map Construction

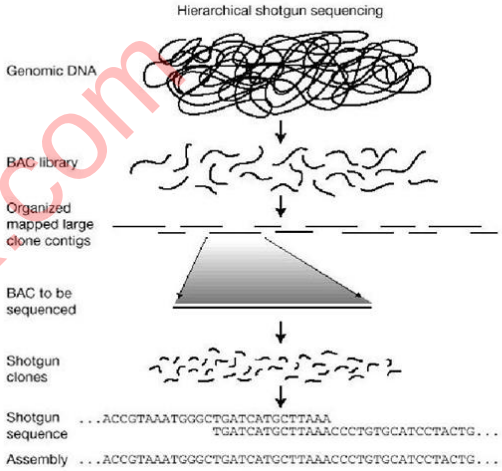
Clone selection

Sub-clone library construction

Random shotgun phase

Directed finishing phase and sequence authentication

HGP – Hierarchical Shotgun Sequencing



HGP–Hierarchical Shotgun Sequencing

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

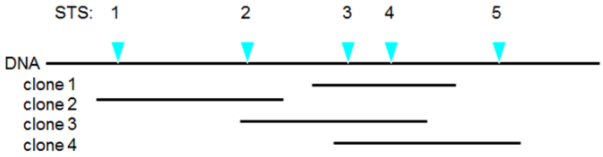
PCR or probe hybridization to screen STS

Restriction site fingerprint

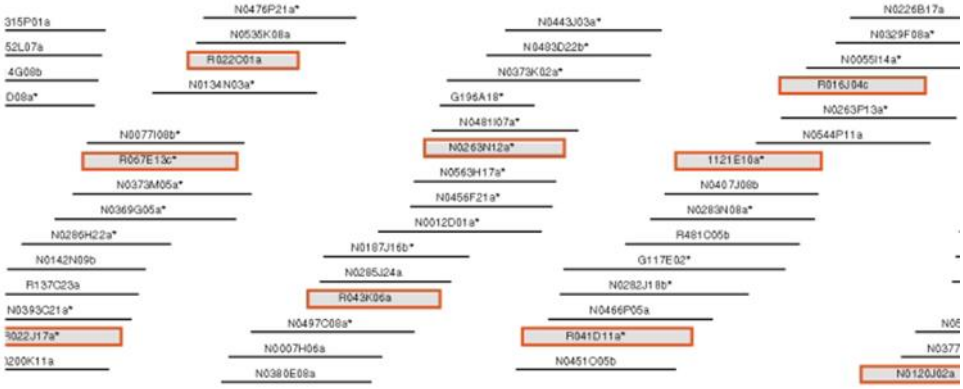
Most time consuming

1990-98 to generate physical maps for human

Hierarchical Shotgun Sequencing: Resolve Clone Relative Location

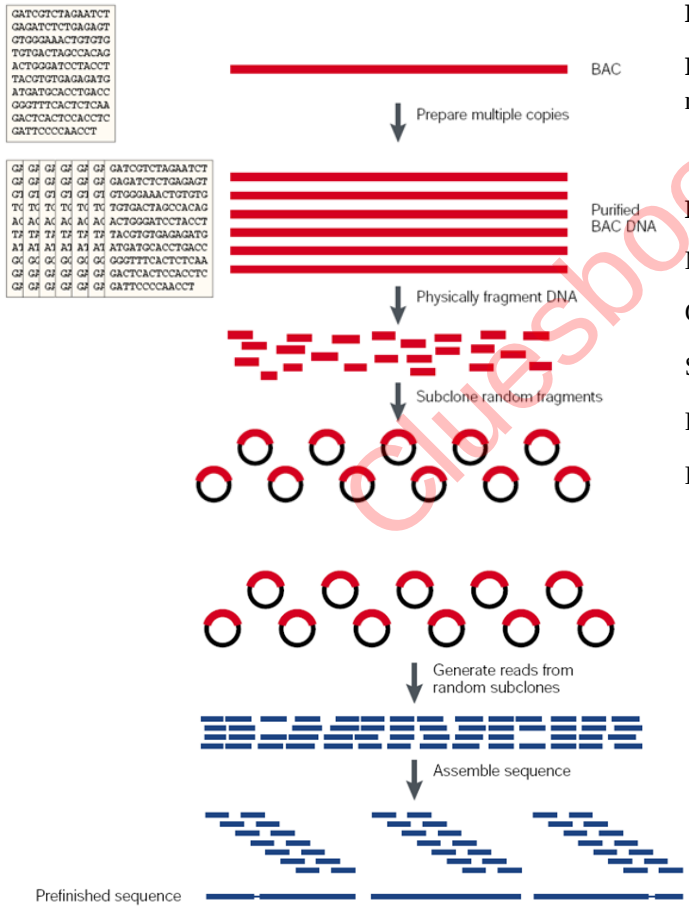


Hierarchical Shotgun Sequencing: Clone Selection: Based on clone map, select authentic clones to generate a minimum tiling path



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

Map Construction

Clone selection

Subclone library construction

Random shotgun phase

Directed finishing phase and sequence authentication

Topic 48

HGP – Whole Genome Shotgun Sequencing

Whole genome randomly digested three times

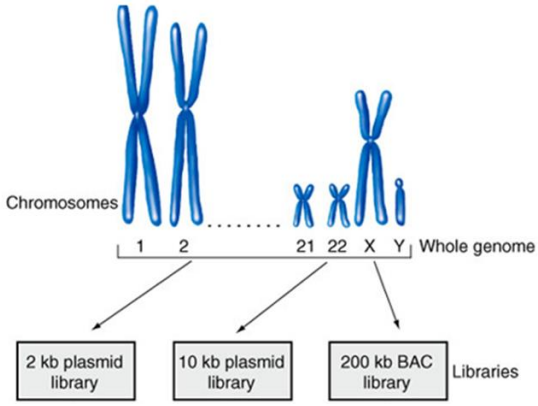
Plasmid library constructed with ~ 2kb inserts and ~10 kb inserts

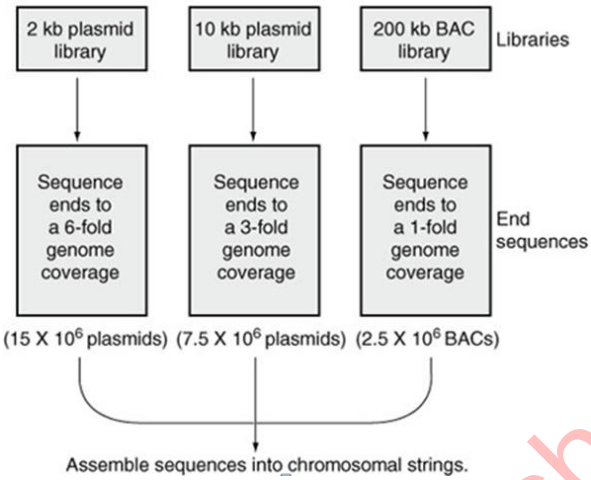
BAC library with ~ 200 kb inserts

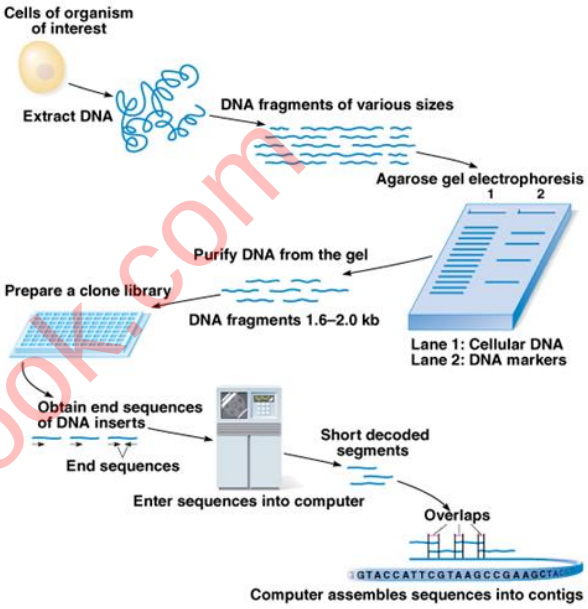
Computer program assembles sequences into chromosomes

No physical map construction

Reduces problems of repeat sequences

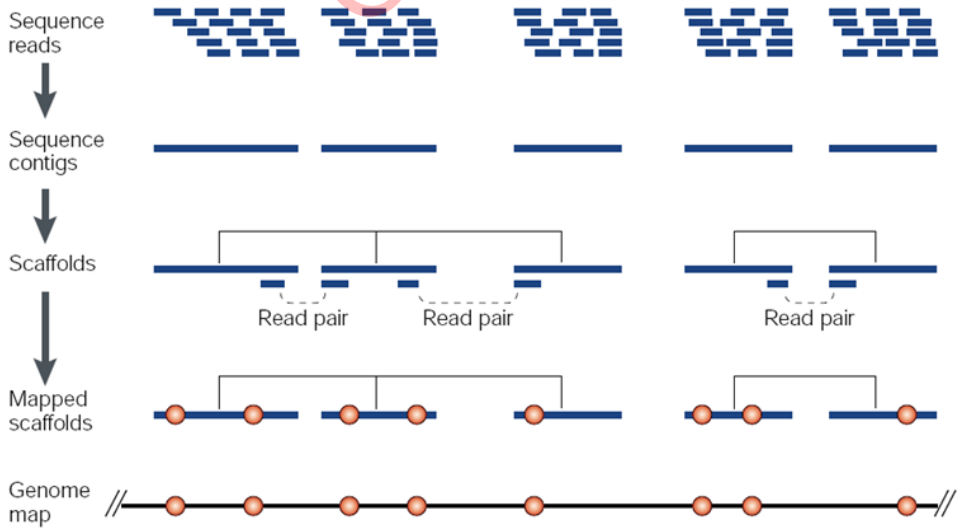


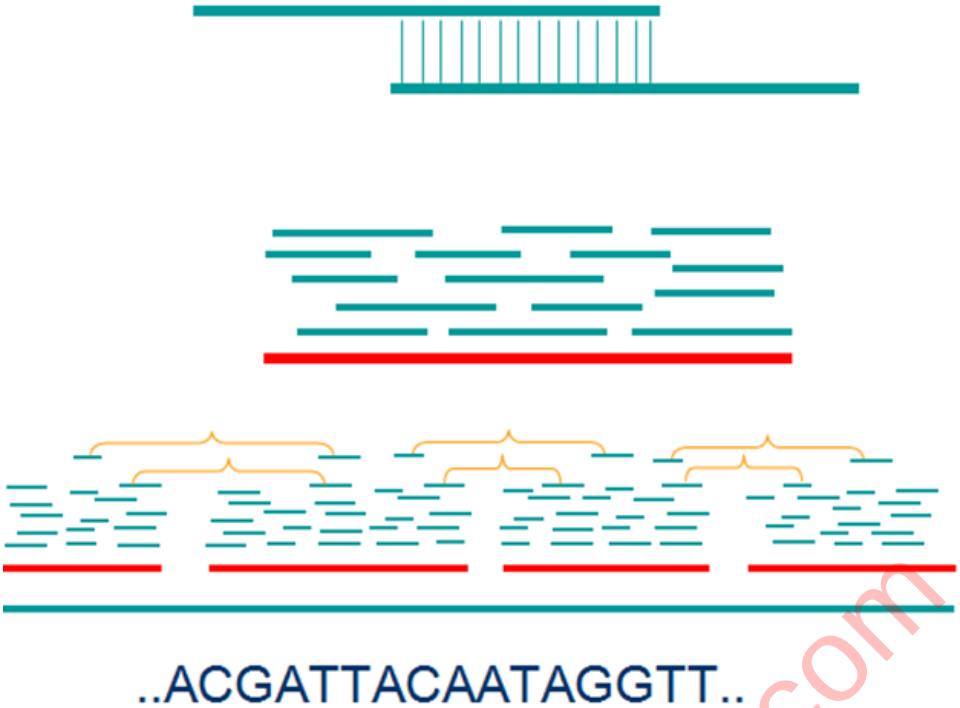




HGP – Whole Genome Shotgun Sequencing

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

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

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

By looking at these contigs and orientation the scaffold become complete

Mapped scaffolds were arranged to prepare chromosome maps

Topic 49

Genetics and Genomics

Human Genome - Characteristics (A)

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

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

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

Genes appear to be concentrated in random areas along the genome

Large areas of noncoding DNA between the genes

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

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

3,000 Mb

~ 23,000 genes

Exons ~ 1.5%

Introns ~ 3.5%

Repeats ~ 45%

Topic 50

Genetics and Genomics

Human Genome - Characteristics (B)

~ 1.5% of the genome codes for proteins.

Repeated sequences make up at least 45 % of the human genome

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

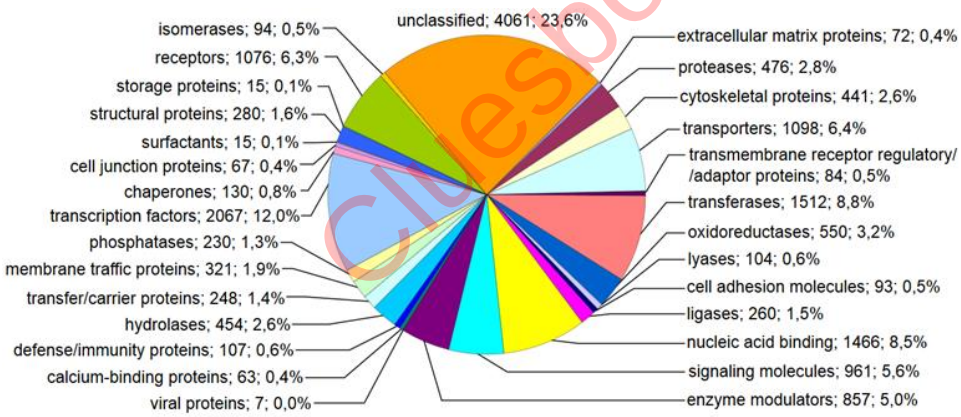
Human Genome-Characteristics (B)

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

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 - 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.

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

Mobile elements:

LTR retrotransposons (8.3% of total genome)

SINEs (13.1% of total genome) including Alu elements

LINEs (20.4% of total genome)

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

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

Male mutation rate about twice female

Most mutations occurs in males

Recombination rates much higher in distal regions of chromosomes

Topic 51

Genome Browser - UCSC

Genome Browsers

The genomes are so large that useful information is hard to find.

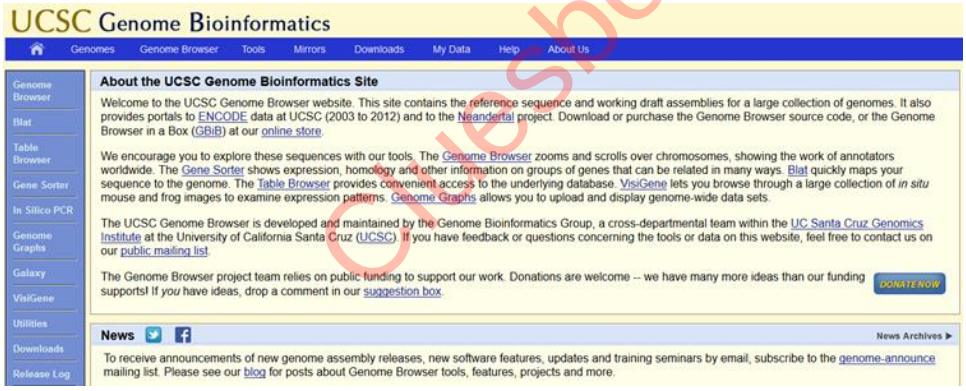
Researchers at UCSC decided to make a computational microscope to help scientists search the genomes.

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

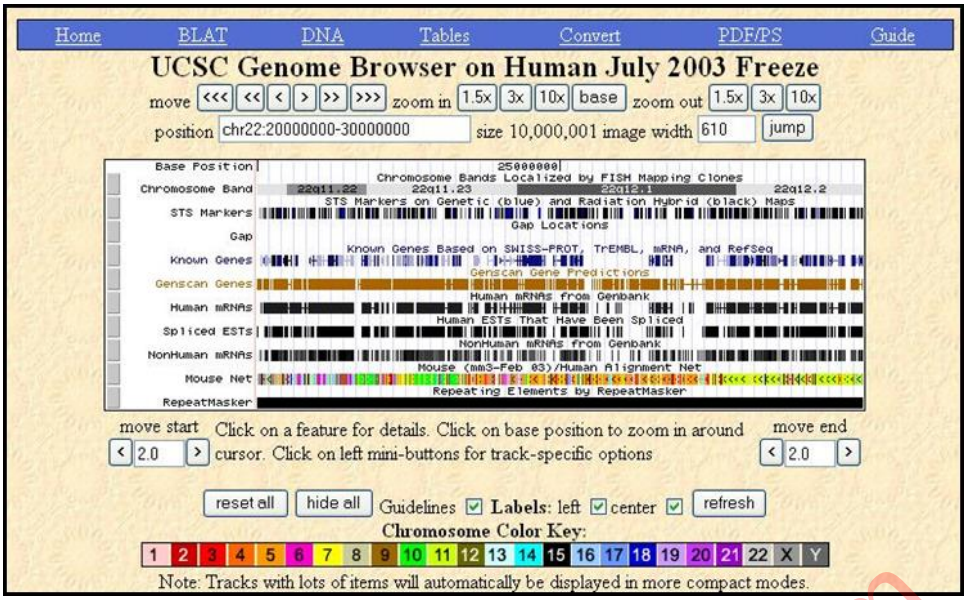
UCSC Genome Browser

http://genome.ucsc.edu

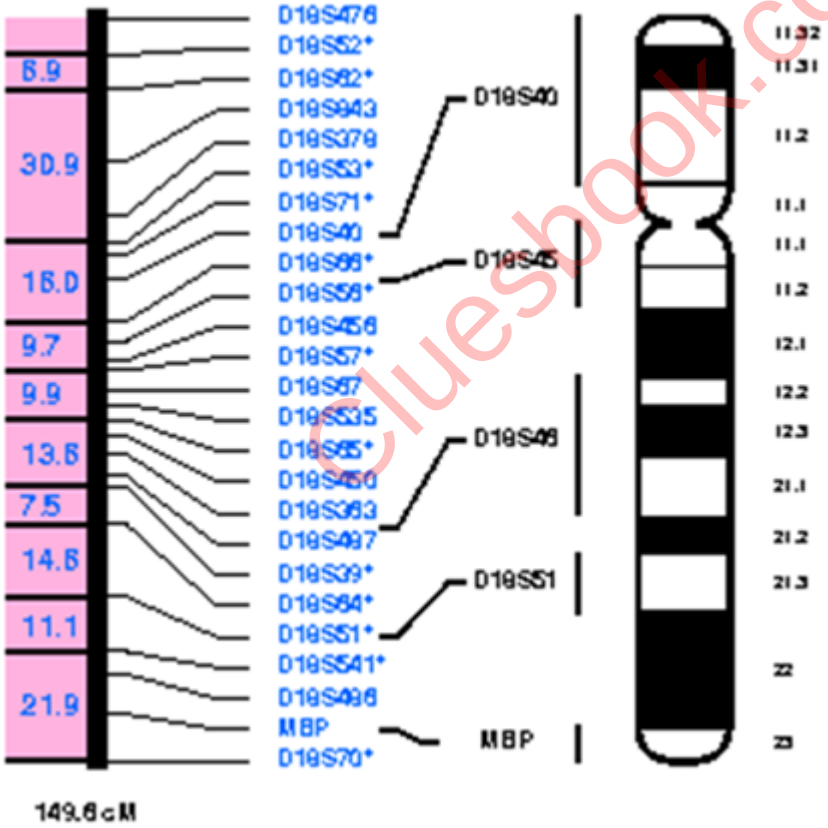
The UCSC Genome Browser - Homepage



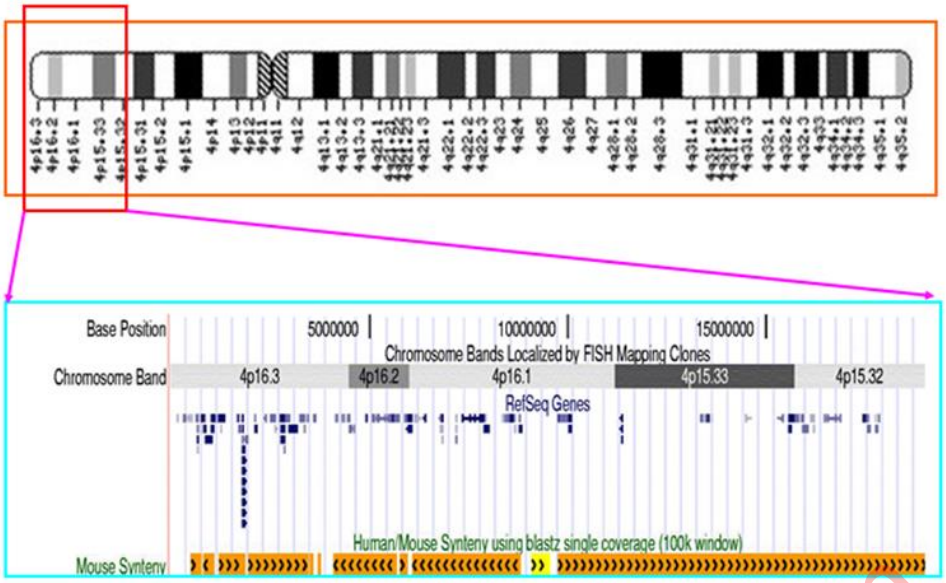
The UCSC Genome Browser



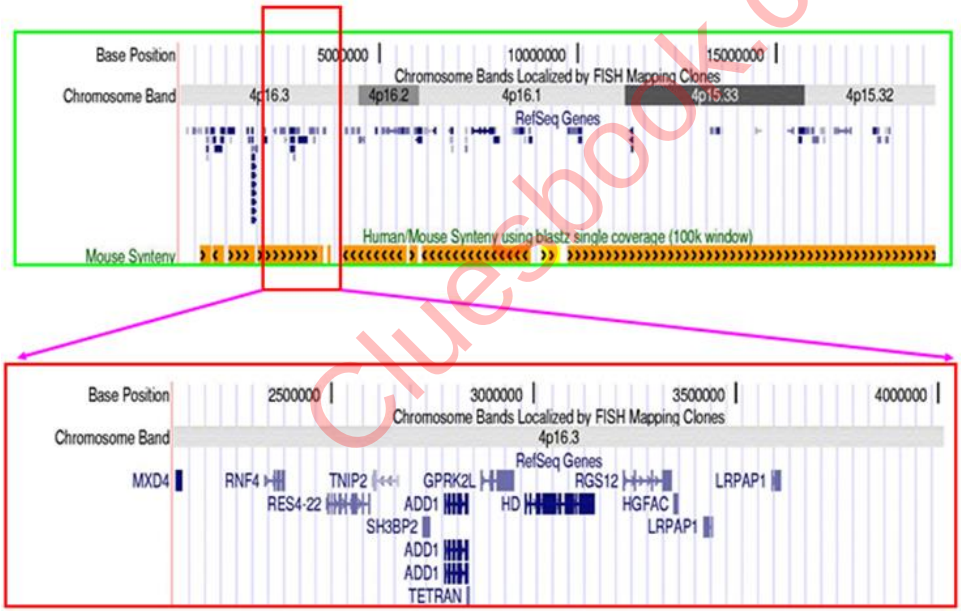
The browser takes you from early maps of the genome . . .



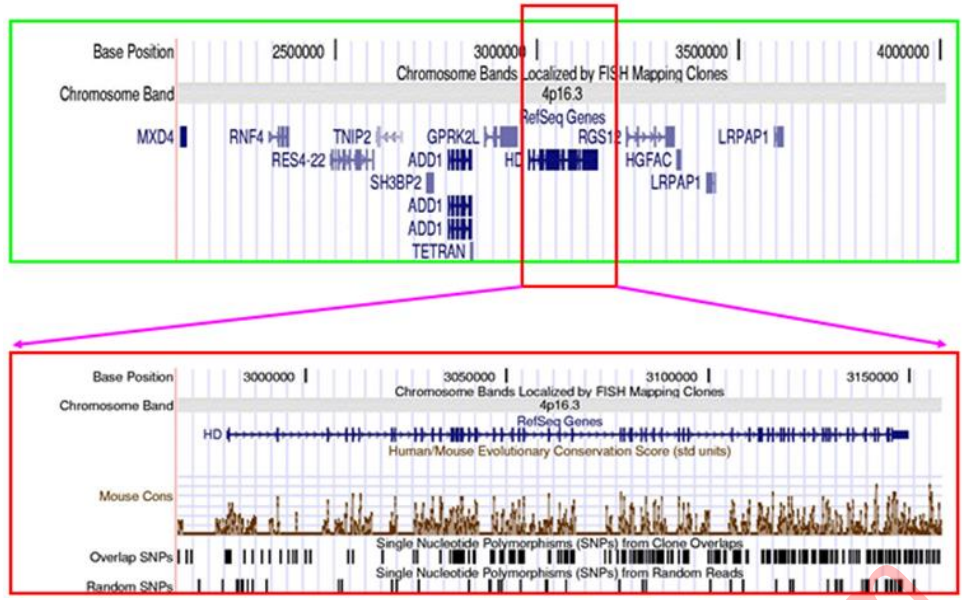
Multi Resolution View



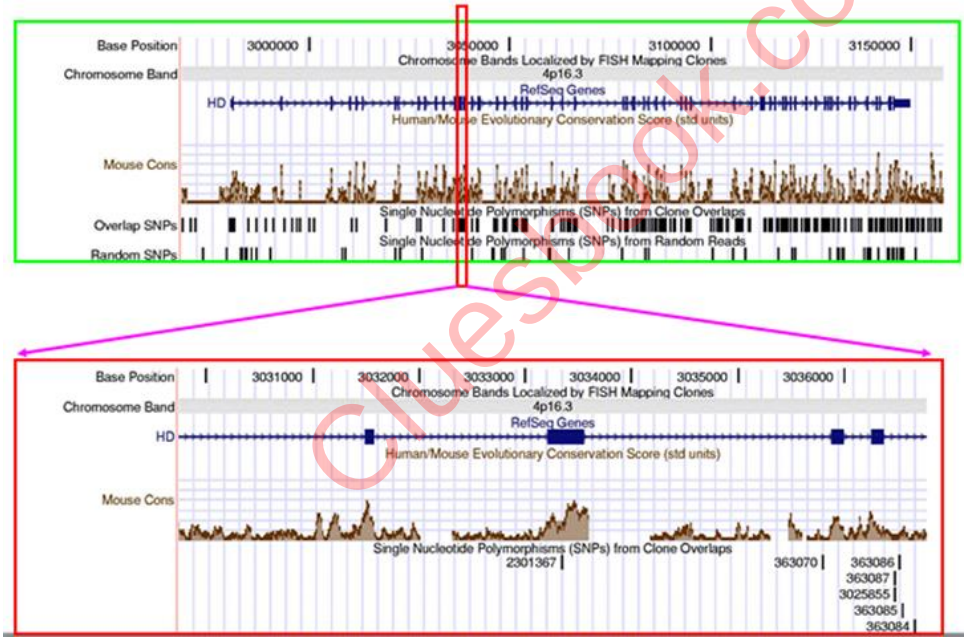
Gene Cluster View



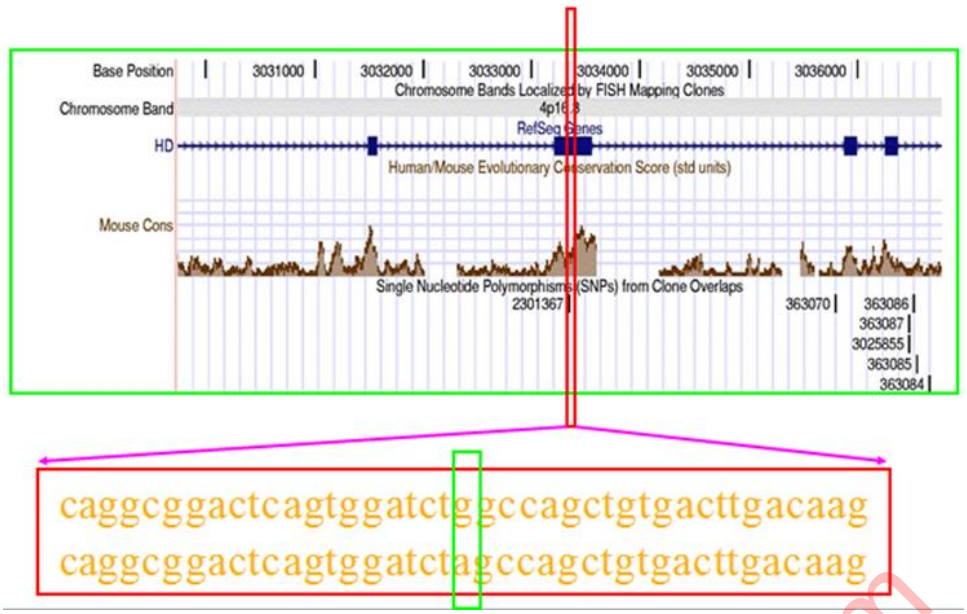
Single Gene View



Single Exon View



Single Base/Nucleotide View



Genome Browser - UCSC

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

Topic 52

Genome Browser - Ensembl

Ensembl Genome Browser

Ensembl is a joint project between 3 organizations

EMBL

EBI

WTSI

Ensembl Genome Browser

Ensembl does not gather any genome project directly

Works in relation with the sequencing centers that generate the genome assembly

4 Main Databases

- Ensembl Core Database
- Ensembl EST Database
- Ensembl Compara Database
- Ensembl Variation Database

Ensembl

Ensembl core databases store genome sequence and annotation information

Gene, transcript, and protein models

Databases stores information about cDNA and protein alignments

Ensembl Genome Browser

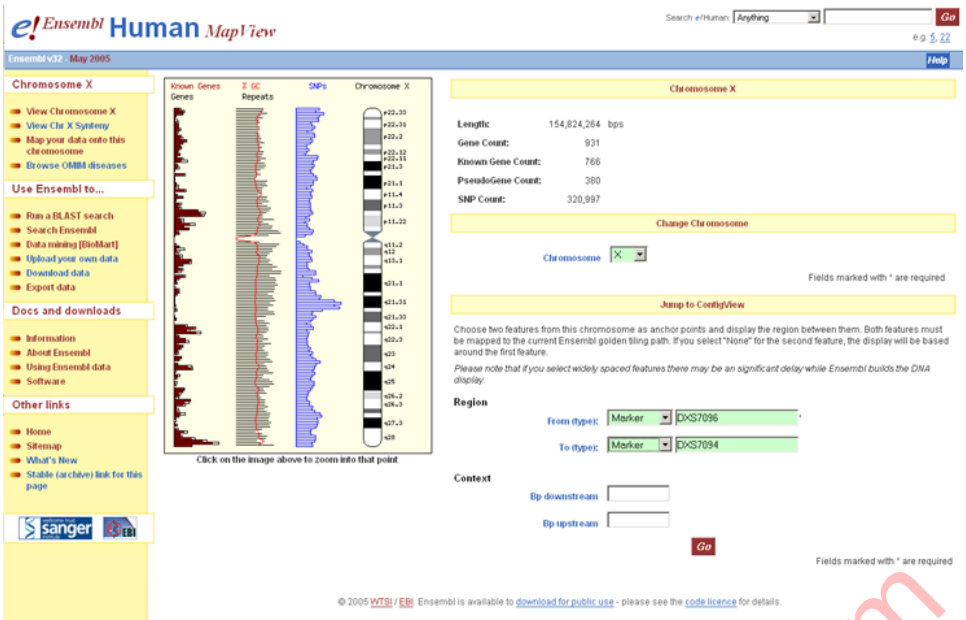
Within one genome: regulatory elements, gene order etc

Comparative studies: Evolution, conserved regions rearrangements

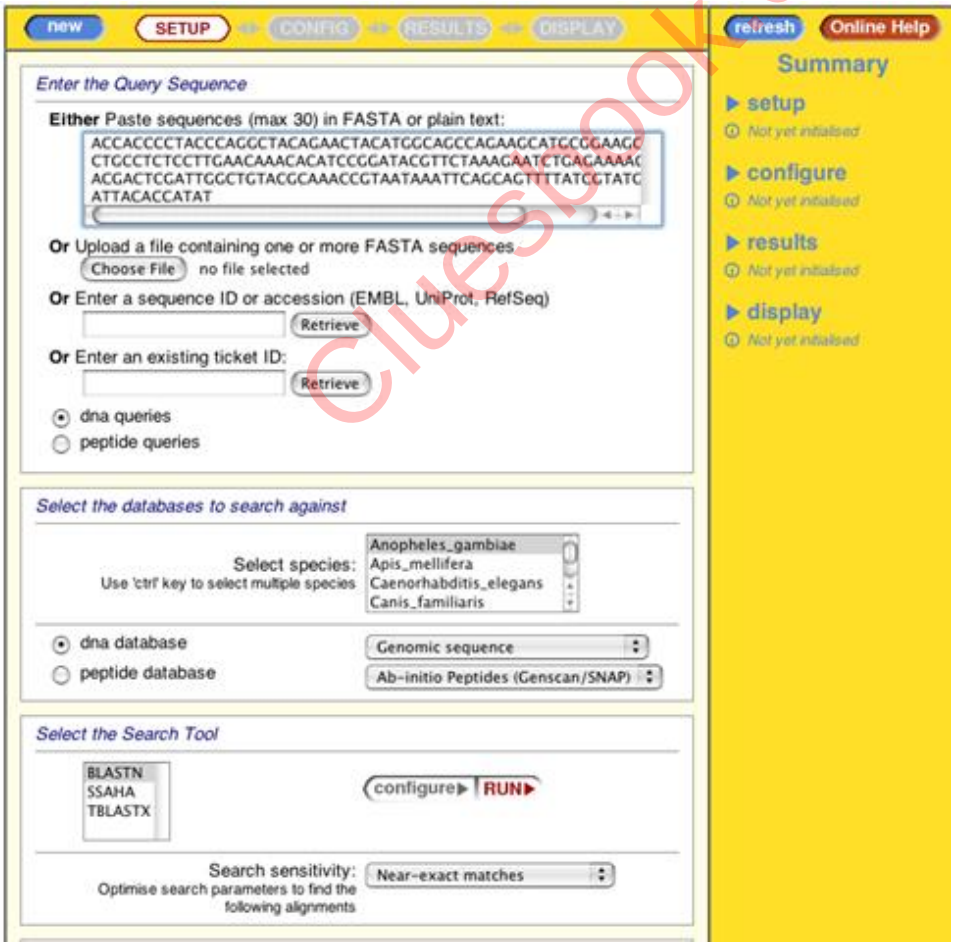
Gene quality and prediction

Ensembl Genome Browser – Home page

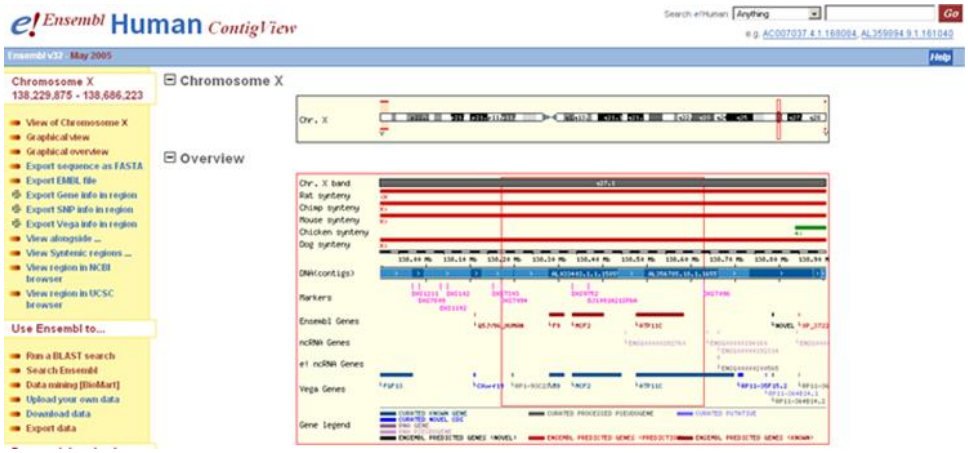
Ensembl Genome Browser – Map View



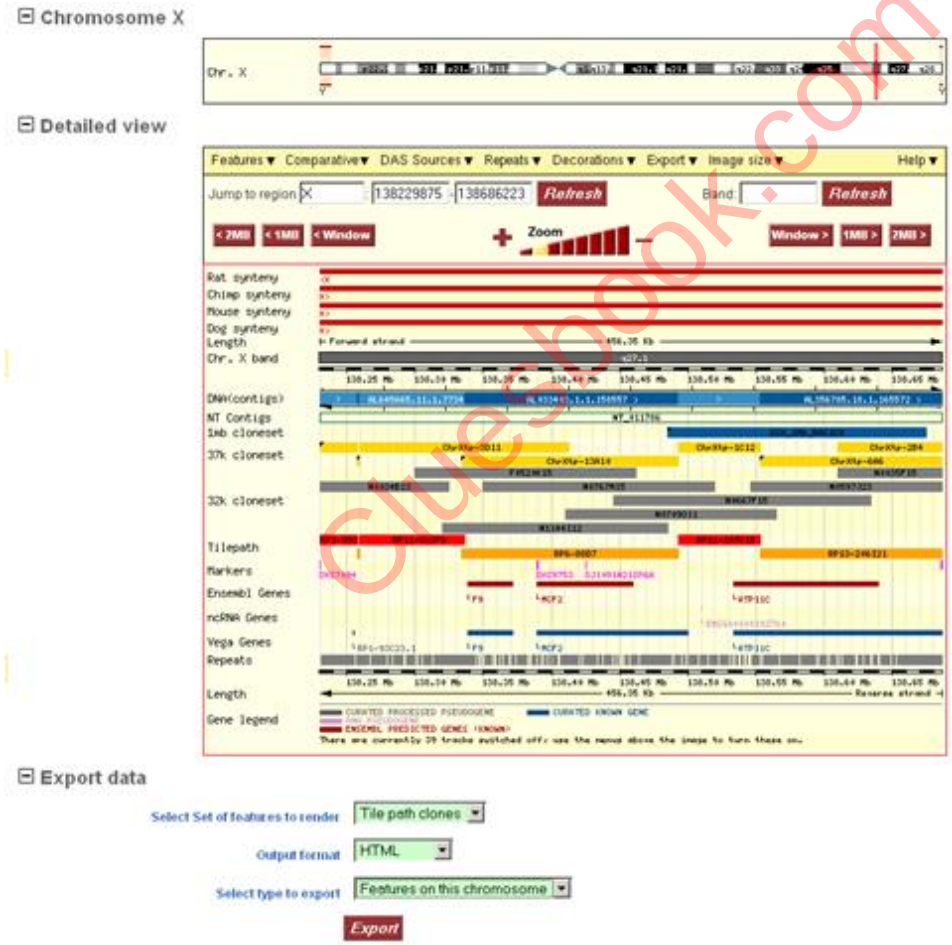
Ensembl Genome Browser – Blast



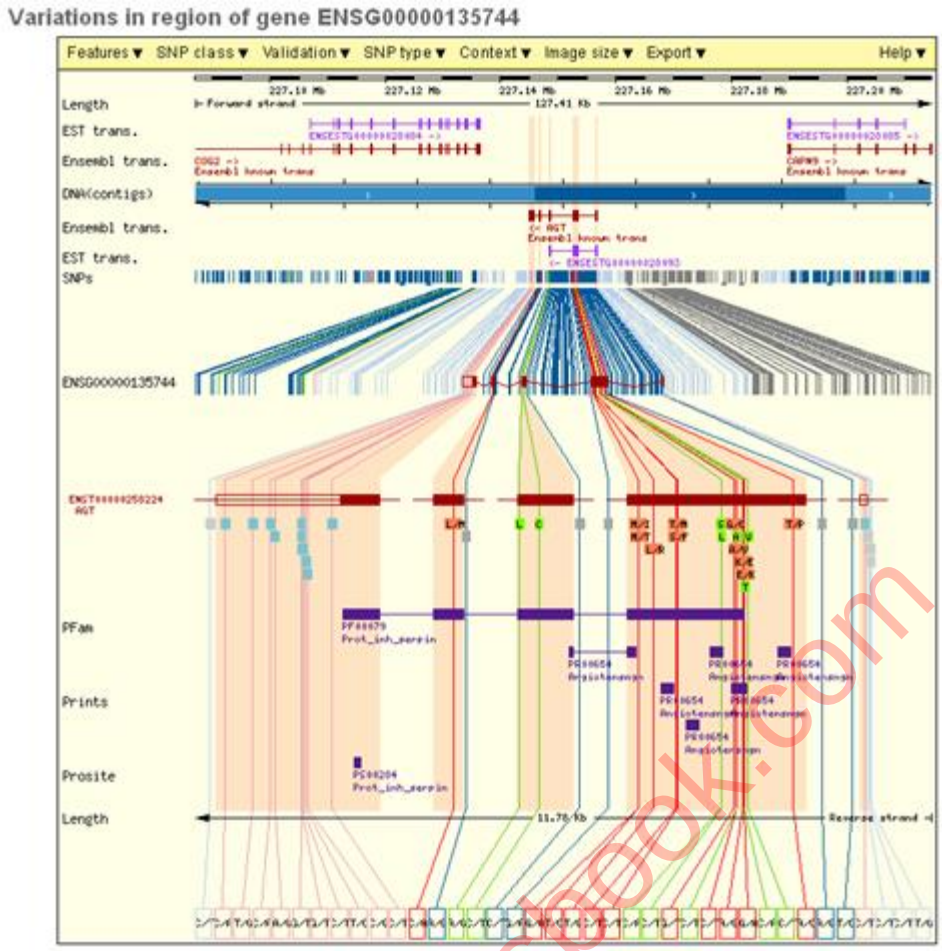
Ensembl Genome Browser – Contig View



Ensembl Genome Browser – Cyto View



Ensembl Genome Browser – SNP View



Ensembl Genome Browser – Marker View

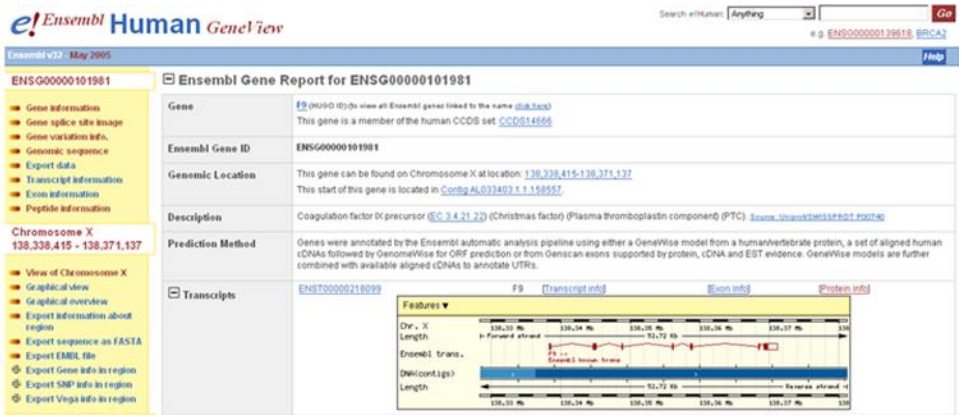
Chromosome Map Marker DXS9752

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 Left Primer: 259 TTTTCAGGTTAATGGACACGC Right Primer: CCATTTTGCAGCCGTAATTT

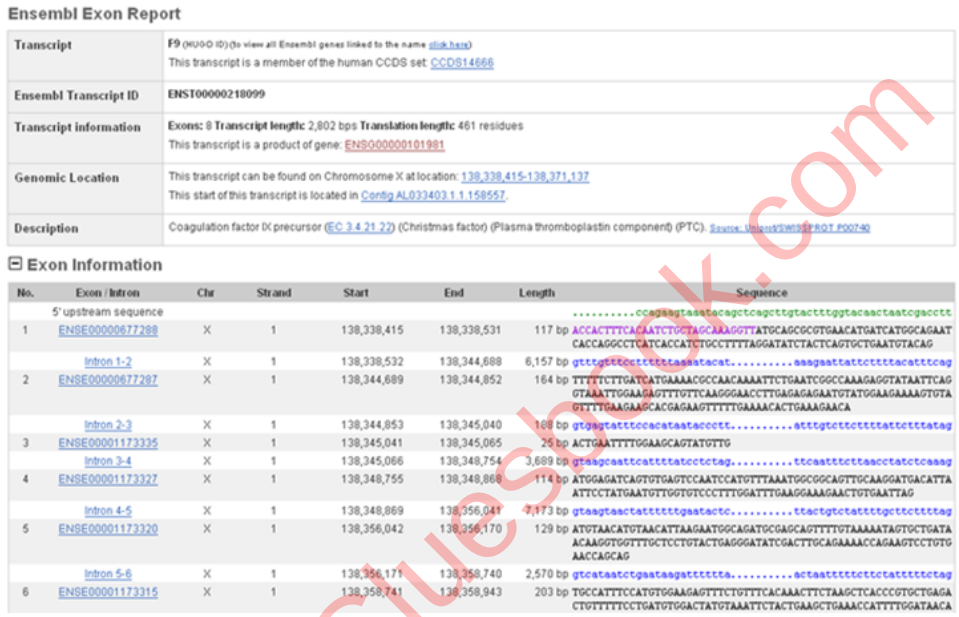
Marker DXS9752 map locations

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

Ensembl Genome Browsers – Gene View



Ensembl Genome Browsers – Exons View



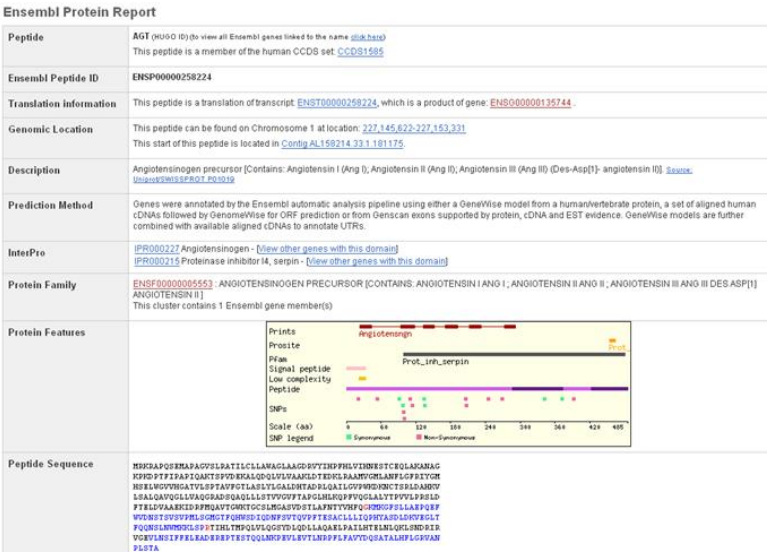
Ensembl Exon Report

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: UniprotKSWISSPROT P00740

Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Length	Sequence
	5' upstream sequence						.....ccagtaagtaaatacagctcagcttgtactttggtacaactaatcgacctt
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	gtttgttccttttttaaaatacat.....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.....atttgtcttctttattctttatag
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	gtaagtaactatttttttgaatactc.....ttactgtctattttgcttcttttag
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.....actaatttttcttctattttctag
6	ENSE00001173315	X	1	138,358,741	138,358,943	203 bp	TGCCATTTCCATGTGGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGA CTGTTTTTCCTGATGTGGACTATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACA

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: UniprotKSWISSPROT 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	Prints, Prosite, Pfam, Signal peptide, Low complexity, Peptide, SNPs, Scale (aa), SNP legend (Synonymous, Non-Synonymous)
Peptide Sequence	MRKRAPQSRMAPGVSLRATILCLLAWAGLAACGDPVTIKPFHLVIHNNESTCEQLAKANAG KFPDPTFIPAPIQARTSPVDEKALQDQLVLVAAKLDTEEKLRAAMVGMLANFLGFPITGM HSELWGVVHGATVLSPTAVFGTLASLYLGALDHTADRLQAILGVPWKDKNCTSRLDAHGV LSALQAVQGLLVAQGRADSQAQLLLSTVVGVFTAPGLHLKQPFVQGLALYTPVVLPRSLD FTELDVAAEKIDRFMQAVTGWKTGCSLMGASVDSTLAFNTYVNFQGKMKGFSLLAEPQEF WVDNSTSVSVPMLSGMGTFQHWSDIQDNFSVTQVFFTESACLLLIQPHYASDLDKVEGLT FQQNSLNWMGKLSPRTIHLTMPQLVLQGSYDLQDLLAQAELPAILHTELNLQKLSNDPIR VGEVLNSIFPELEADERRPTESTQQLNKPEVLEVTLNRPFLFAVVDQSATALHFLGRVAN PLSTA

Ensembl Genome Browser

Homo Sapiens - Human

Pan troglodytes- Chimpanzee

Macaca mulatta - Rhesus monkey

Mus musculus - Mouse

Rattas norvegicus - Rat

Canis familiaris - Dog

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

Topic 53

Genetics and Genomics

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

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

Privacy and confidentiality of genetic information

Fairness in the use of genetic information by insurers, employers, courts, schools, adoption agencies

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

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

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

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

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

Fairness in access to advanced genomic technologies.

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

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

Topic 54

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.
Genes transcribed, which means that we can identify them

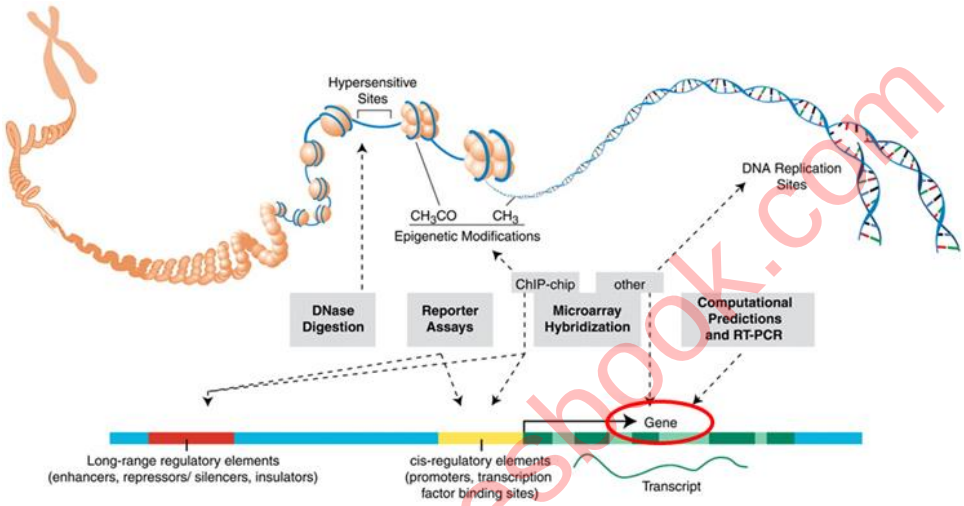
Traditionally cDNA /EST sequencing

More recently by microarray
Structural Annotation – Finding genes and other biologically relevant sites

Annotation/function can be mapped to different levels:

Organism, cellular, molecular level

Genome Annotations



Genome Annotations-Protein Coding Genes

Genes can be identified *in vitro* using computational methods

Protein-coding genes have recognizable features

Open reading frames

Codon bias

Known transcription and translational start and stop motifs

Splice consensus sequences at intron-exon boundaries
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

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

Harder to find

No poly-A tailed—

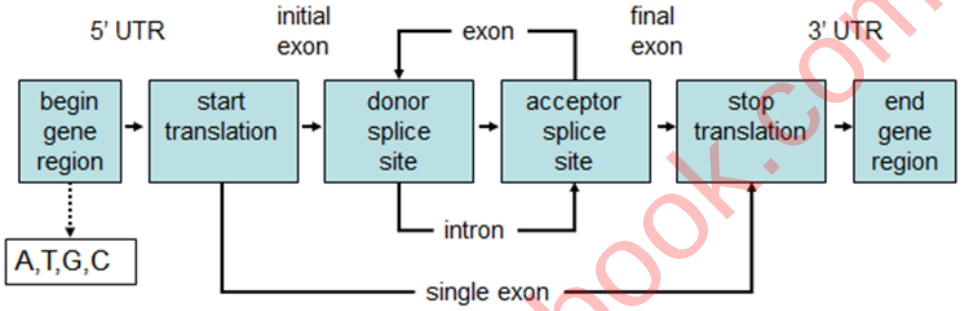
No ORF

Rely on sequence divergence at nucleotide not protein level, so homology is harder to detect

Most gene-discovery programs makes use of machine learning algorithm.

Two approaches;

Artificial Neural Networks and Hidden Markov Models.



Gene Discovery Programs- Diagram for an HMM for Gene Discovery

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

Topic 55

Genetics and Genomics

Genome Annotations – Transcription Factors

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

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)

Transcription factors bind to specific DNA sequences

Most transcription factors can bind to a range of similar sequences.

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

Genome Annotations

7 characterized binding sites for certain transcription factor

TCCGGAAGC
TCCGGATGC
TCCGGATCT
CATGGATGC
CCAGGAAGT
GGTGGATGC
ACCGGATGC

Consensus Sequence

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

Frequency matrix and its graphical depiction, a sequence logo

A	111007200
T	302000502
G	110770060
C	254000015

Genome Annotations-Transcription Factors

Genome Annotations - Conclusion

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

Topic 56

Genetics and Genomics

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

Arabidopsis Thaliana – A model Plant

Arabidopsis has 5 chromosomes (2n=10)

Contains about 135 megabases of sequence

Encodes approximately 27,000 genes and 35,000 proteins.
Has 35% unique genes

Has 37% genes that exist as members of large gene families (families of 5 or more members)
An estimated 58-60% of the Arabidopsis genome exists as large segmental duplications
Arabidopsis genome contains genes encoding RNA polymerase subunits not seen in other eukaryotic organisms
Arabidopsis has genes unique to plants – approximately 150 unique protein families were found, including 16 unique families of transcription factors
Arabidopsis has 5 chromosomes (2n=10)

Contains about 135 Mb of sequence

Encodes ~ 27,000 genes and 35,000 proteins.

Topic 57

Genetics and Genomics

Mouse Genome – A Model Organism in Animal Biology and Genetics

Number of chromosome 2n=40

Human and mouse genomes have conserved blocks of genetic material

Humans and mice suffer from similar diseases

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: Chromosomes and their Sizes

Chromosome	cM	Mbp	Chromosome	cM	Mbp
1	98.5	195	11	88.0	122
2	103.9	182	12	63.9	120
3	82.7	160	13	67.3	120
4	88.6	157	14	66.4	125
5	90.2	152	15	59.0	104
6	79.0	150	16	57.8	91
7	89.1	145	17	61.3	95
8	76.2	129	18	59.4	91
9	75.1	125	19	56.9	61
10	77.9	131			

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)

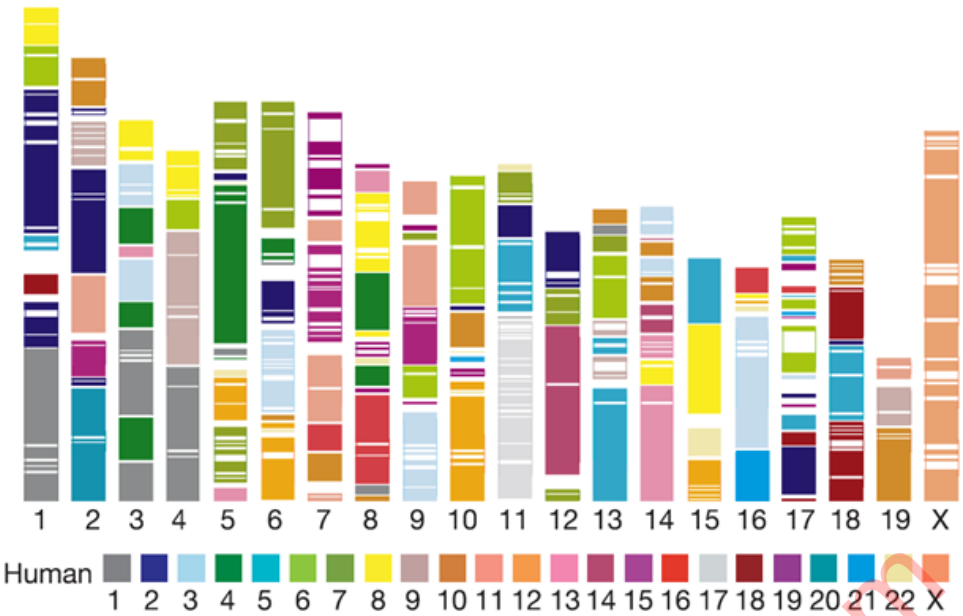
The (G+C) content for each of the mouse chromosome is relatively similar

Mouse Genome

Mouse Genome

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

Human and Mouse: Chromosomes Orthologs



Mouse Genome

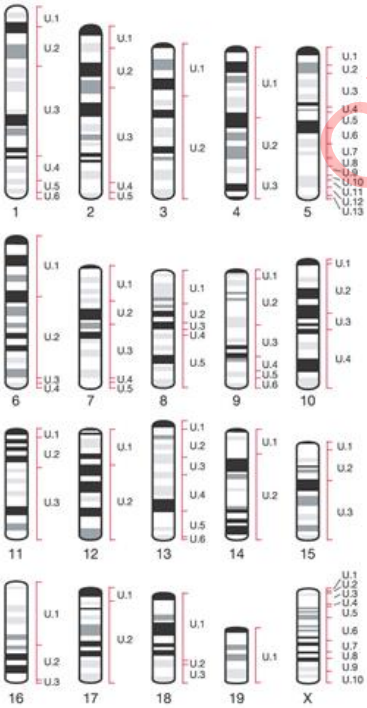
Mouse Genome Contigs

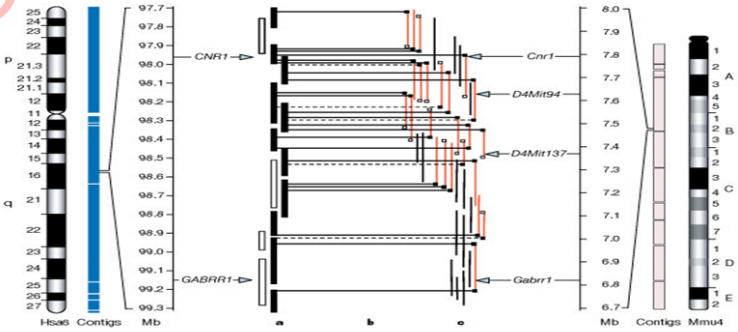
Most mouse BAC contigs contained multiple mouse markers.

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

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

Cytogenetic View of Chromosomes





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

Human–Mouse Homology Clone Map

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

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

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

Topic 58

Genome Comparison -Human and Mouse

Genome Comparison-Human and Mouse

Comparison

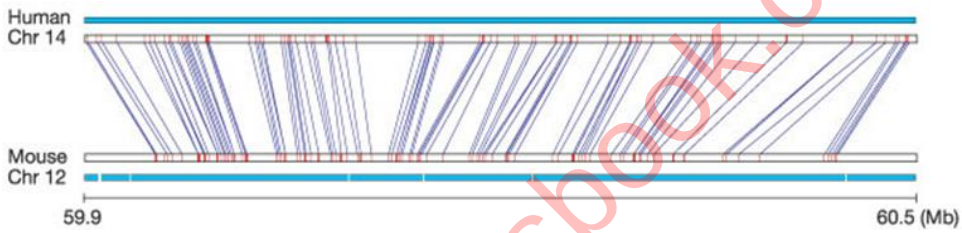
The mouse genome is about 14% smaller than the human genome

Over 90% of mouse and human genomes can be partitioned into corresponding regions of conserved synteny
At the nucleotide level, 40% of the human genome can be aligned to the mouse genome
Mouse and human genomes each seem to contain ~ 23,000 protein-coding genes.

The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) is < 1%.
Despite marked differences in activity of transposable elements, similar types of repeat sequences have accumulated in the corresponding genomic regions

Comparison - Human and Mouse Genome

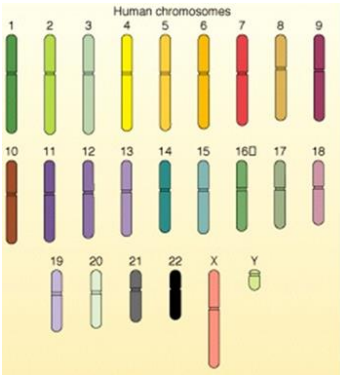
A typical 510-kb segment of mouse chromosome 12 that shares common ancestry with a 600-kb section of human chromosome 14.

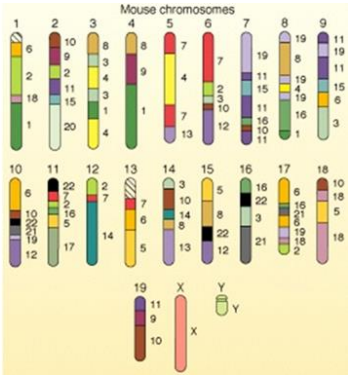


Blue lines connect the reciprocal unique matches in the two genomes.
Mouse genome contains fewer CpG islands than human.

(G+C) content and density of CpG islands shows more variability in human than mouse
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.
The density of CpG islands is relatively homogenous for all mouse chromosomes and more variable in human.
Approximately 99% of mouse genes have a homologue in the human genome

Comparison – Human and Mouse Genome





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.

Topic 59

Mapping of Disease Genes

Disease gene mapping is one of the main objectives of genotyping

Different approaches for identification 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.

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

Mostly identifying and isolating the protein product of the gene.

Such genes usually produce large amounts of well-known and well studied proteins.

Two common methods used

- Gene-specific oligonucleotides
- Antibody Methods

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

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.

Different approaches are used for identification of disease genes

Topic 60

DNA Microarrays

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

There are several synonyms of DNA microarrays such as

- DNA chips
- Gene chips
- DNA arrays
- Gene arrays
- Biochips

Microarray - Principle

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

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

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

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

Just sequencing genomes is not sufficient

Thousands of genes remain without an assigned function

Each spot of DNA, (probe), represents a single gene

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

Genome-wide studies and genotyping

Evaluating microRNA levels

Determine which genes have biological significance in a system

Discovery of new genes, pathways, and cellular trafficking

Topic 61

Types of Microarrays

DNA microarrays

Exon arrays

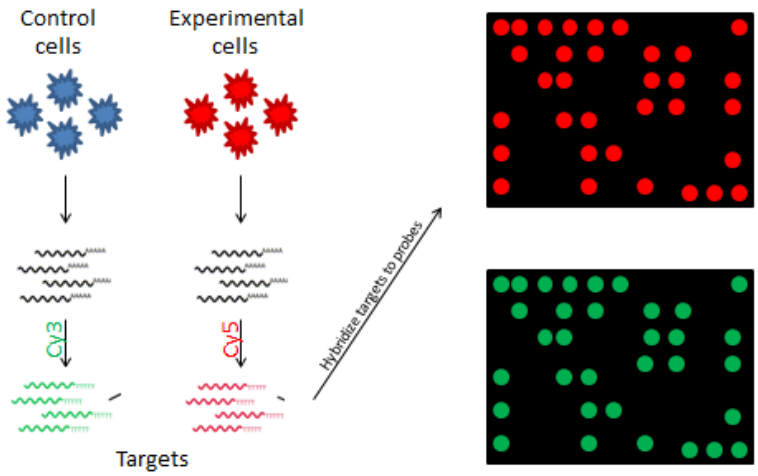
Comparative Genomic Hybridization array

DNA microarrays- such as cDNA microarrays and oligonucleotide microarrays, SNPs

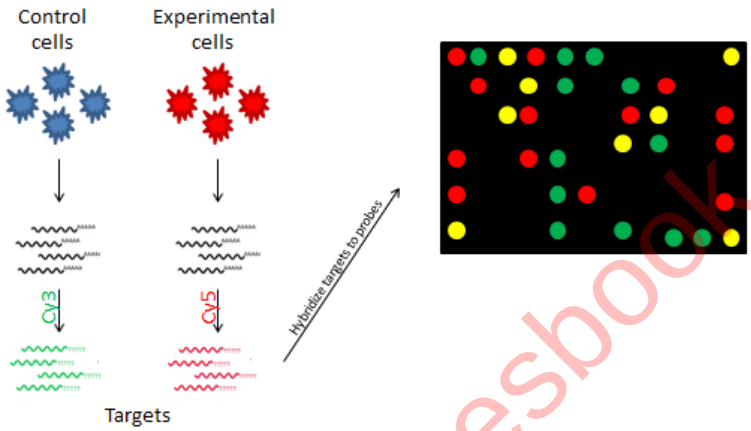
Exon arrays-Alternative splice variant detection

CGH arrays- Comparative genomic hybridization

Oligonucleotide Microarrays



cDNA Arrays



Protein microarrays (protein-protein interactions)

Tissue microarrays

Cellular microarrays (also called transfection microarrays

Expression Arrays

Chemical compound microarrays

Antibody microarrays (proteomics)

Carbohydrate arrays (glycoarrays)

Re-sequencing arrays

Microarray Require

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

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

Microarray and SNPs

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

Methylation Microarrays

Methylation arrays determine methylated DNA

Determining which regions of DNA are methylated ultimately to study epigenetics

Topic 62

Microarrays Formats

Main Formats

Cartridge-based

- Spotted
- Electronic

Spotted Glass Slide

Tissue Section Slide

Cartridge-Based Chips

High density arrays of DNA oligos within a plastic housing

One sample = One chip (Affymetrix, Agilent, Applied Biosystems etc)

Generally used with expression and DNA arrays

Cartridge-Based Expression Microarrays

Involves Fluorescently tagged biotinylated cRNA

One chip per sample

Uses single fluorescent dye

Spotted Glass Arrays

Uses cDNA, Oligonucleotide, protein, antibody

Robotically spotted cDNAs or Oligonucleotides

Printed on Nylon, Plastic, or Glass microscope slide

Spotted cDNA and Oligo Glass Arrays

Involves two dyes on the same slide

Red dye-Cy5

Green dye-Cy3

Control and experimental cDNA

Main Formats

Cartridge-based

- Spotted
- Electronic
- Spotted Glass Slide
- Tissue Section Slide

Topic 63

Microarray Procedure

Collect Samples

Isolate mRNA

Create Labelled DNA

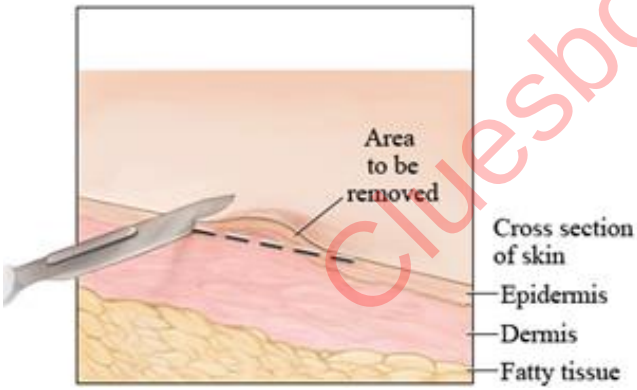
Hybridization

Microarray Scanner

Analyze Data

Microarray Procedure

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



Extract the RNA from the samples.

After isolating the RNA, isolate the mRNA from the rRNA and tRNA.

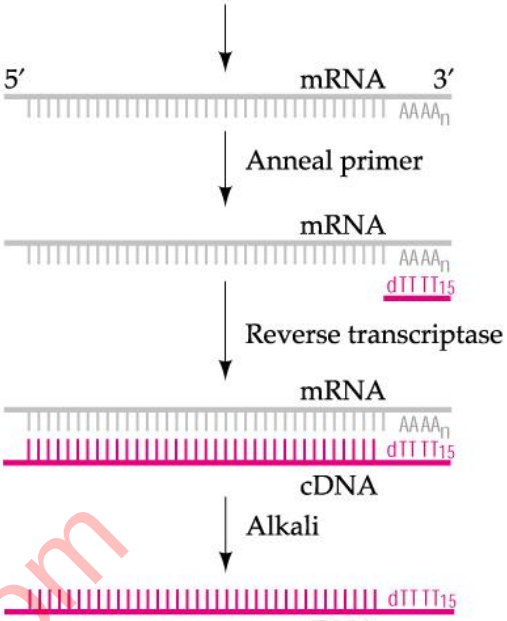
mRNA has a poly-A tail, use a column containing beads with poly-T tails to bind the mRNA

Rinse with buffer to release the mRNA from the beads. The buffer disrupts the pH, disrupting the hybrid bonds.

Microarray Procedure - 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 – 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 – Lasers

Laser scans array and produces images

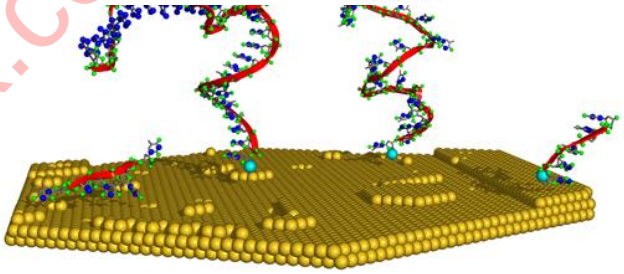
One laser for each color, e.g. one for green, one for red

Microarray Procedure – Microarray Scanner

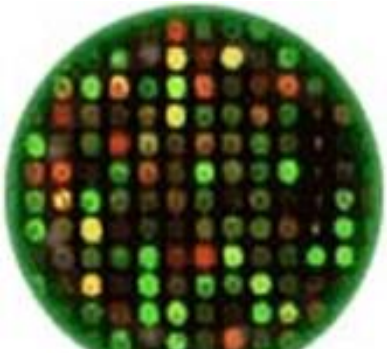
The scanner has a laser, a computer, and a camera.

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.



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

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

YELLOW - both samples hybridized equally to the target DNA.

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

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

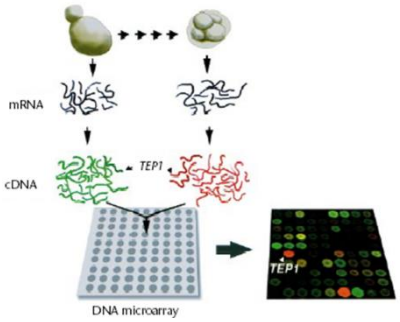
Microarray Procedure

Collect Samples

Isolate mRNA

Create Labelled DNA

Hybridization



Microarray Scanner

Analyze Data

Topic 64

Microarray Chips

There are two types of DNA Chips/Microarrays:

cDNA based microarray

Oligonucleotide based microarray

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

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.

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

DNA Based Microarray Chips

Almost similar type procedure

Gene expression profiling

Discovery of drugs

Diagnostics and genetic engineering

Alternative splicing detection

Functional genomics

DNA sequencing

Toxicological research (Toxicogenomics)

Microarray Chips

There are 2 types of DNA Chips/Microarrays:

- cDNA based microarray
- Oligonucleotide based microarray

Topic 65

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

Different parts of DNA can be used to study gene expression

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.

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

Microarray -Limitations

Cross-hybridization of sequences with high identity

Chip to chip variation

The real limitation is Bioinformatics

Expensive – repeat experiments

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

Pharmacogenomics

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

Toxicogenomics

Microarray technology allows us to research the impact of toxins on cells.

Some toxins can change the genetic profiles of cells.

Biological Applications

Many biological discovery

New and better molecular diagnostics

New molecular targets for therapy

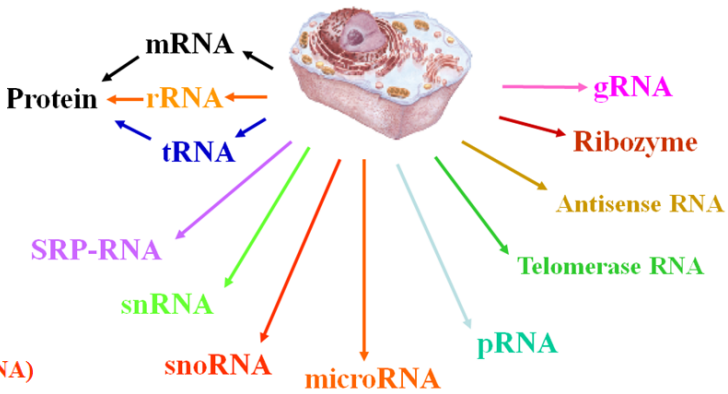
Finding and refining biological pathways

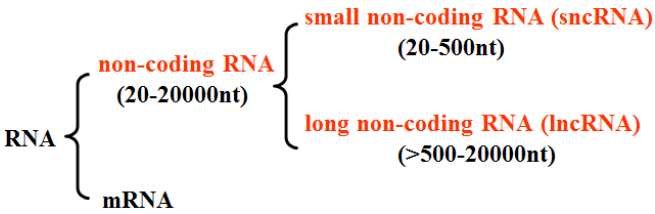
Topic 66

Types of RNA and RNomics

Types of RNA

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





Types of RNA – Housekeeping RNAs

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

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 – 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.

Topic 67

Non-coding DNA

Types of Non-coding DNA

Non- coding functional RNA

Cis- and Trans- regulatory elements

Introns

Pseudogenes

Repeat sequences, transposons and viral elements

Telomeres

Non-coding Functional RNA

The RNA molecules which are not translated into proteins.

Examples: Ribosomal RNA, transfer RNA & micro RNA

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.

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.

Repeat Sequences, Transposons & Viral Elements

Transposons and Retrotransposons are mobile genetic element

Retrotransposons: LINEs, SINEs

Telomeres

Telomeres are regions of repetitive DNA

Located at the end of a chromosome

They provide protection from chromosomal deterioration during DNA replication

Functions

Essential for chromosome structure

Genome Protection

Enhancers, silencers, promotors, insulator

Genetic switches

Regulation of gene expression.

Topic 68

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

Types of Non-coding RNA

Ribosomal RNA

Transfer RNA

Small nuclear RNA

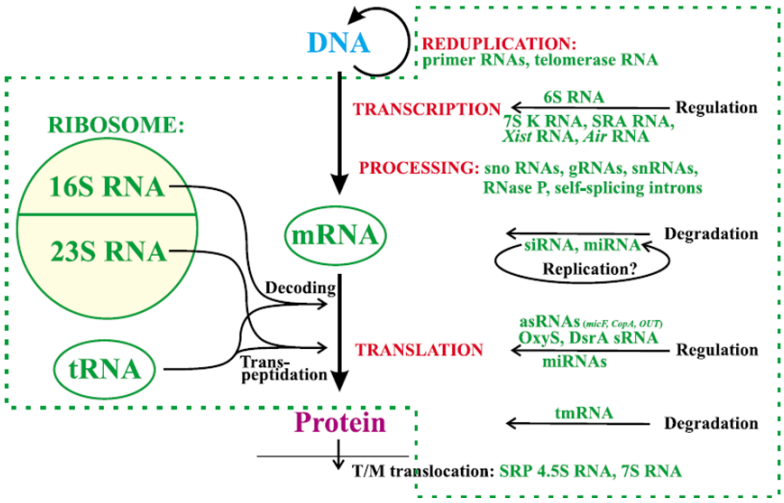
Small nucleolar RNA

Short interfering

Micro RNA

Long non-coding RNA

Functional Diversity of Non-coding RNA



Non-coding RNAs in Human Genome

tRNA	600	SRP RNA	1
18S rRNA	200	RNase P RNA	1
5.8S rRNA	200	Telomerase RNA	1
28S rRNA	200	RNase MRP	1
5S rRNA	200	Y RNA	5
snoRNA	300	Vault	4
miRNA	250	7SK RNA	1
U1	40	Xist	1
U2	30	H19	1
U4	30	BIC	1
U5	30		
U6	20		
U4atac	5	Antisense RNAs	1000s?
U6atac	5	Cis reg regions	100s?
U11	5	Others	?
U12	5		

Non-coding RNA

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

Topic 69

MicroRNA (miRNA)

Small non-coding double stranded RNAs

Approximately 19-22 nucleotides long

Repress activity of complementary mRNAs

Regulate 30% of mammalian genes

Described in invertebrates and vertebrates: worms, fungi, plants, and mammals

Many are conserved between vertebrates and invertebrates

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

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

Hairpin precursor ~70 nt (pre-miRNA)

Mature miRNA ~22 nt (miRNA)

Drosha and Pasha are part of the “Microprocessor” protein complex (~600 - 650kDa)

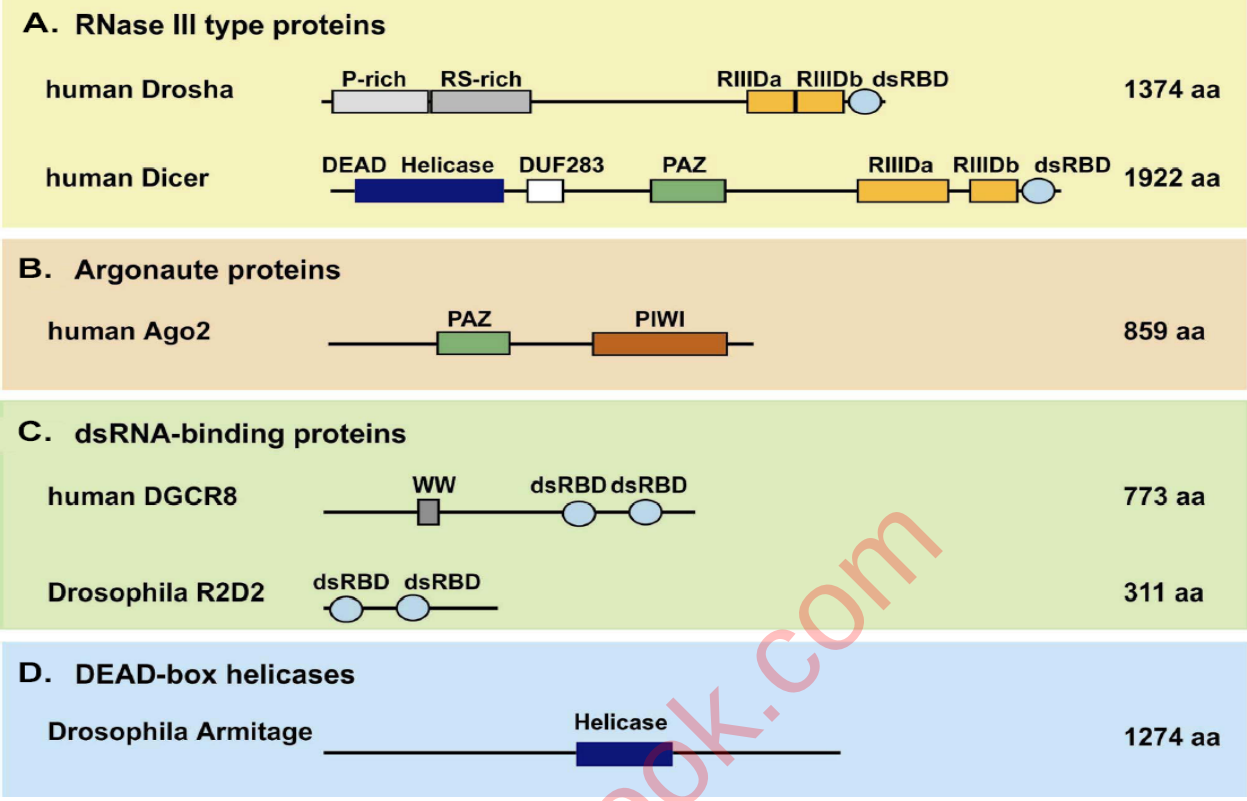
Drosha and Dicer are RNase III enzymes

Pasha is a dsRNA binding protein

Exportin 5 is a member of the karyopherin nucleocytoplasmic transport factors

Argonautes are RNase H enzymes

MicroRNA (miRNA)



~1900 discovered in Homo sapiens

Mouse and human have highly conserved miRNA genes

MicroRNA (miRNA) Conclusion

Small non-coding double stranded RNAs

Approximately 19-22 nucleotides long

30% of mammalian mRNAs are regulated by miRNAs.

Topic 70

- **Genetics and Genomics**

Biogenesis of MicroRNA (miRNA)

- **Biogenesis of MicroRNA (miRNA)**

MicroRNA (miRNA)

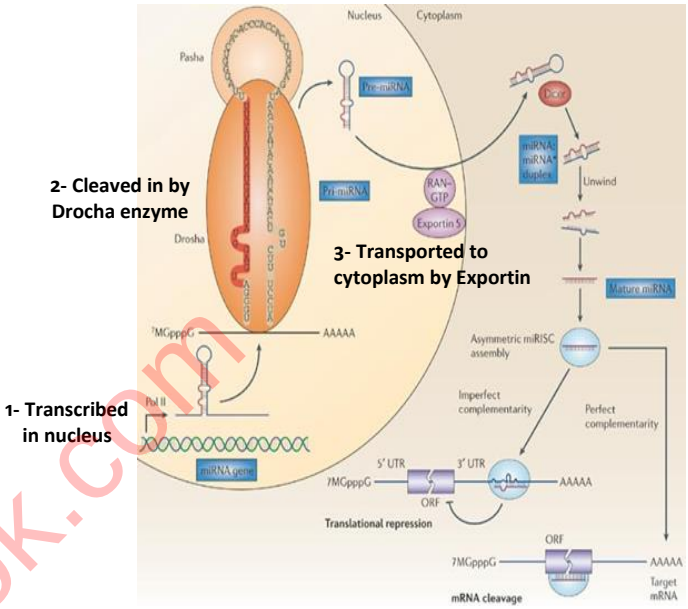
- Small non-coding double stranded RNAs
- Approximately 19-22 nucleotides long
- miRNA originates with ssRNA that forms a hairpin secondary structure
- **Biogenesis of MicroRNA (miRNA)**

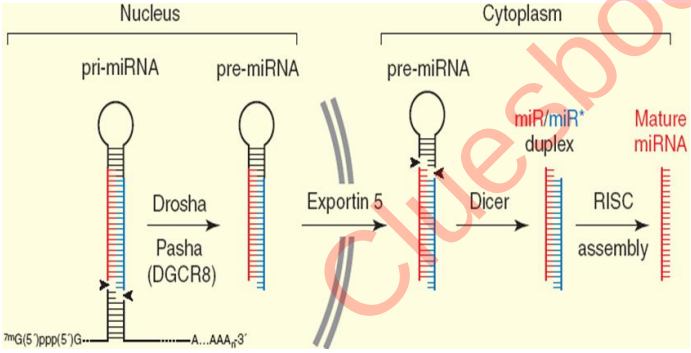
MicroRNA (miRNA)

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

Biogenesis of 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)

MicroRNA (miRNA)

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

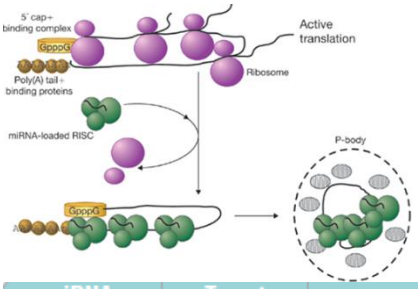
Topic 71

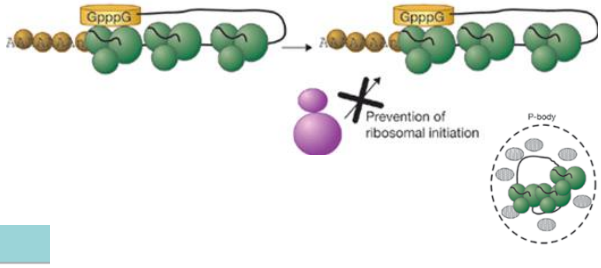
Functions of miRNA

- 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





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

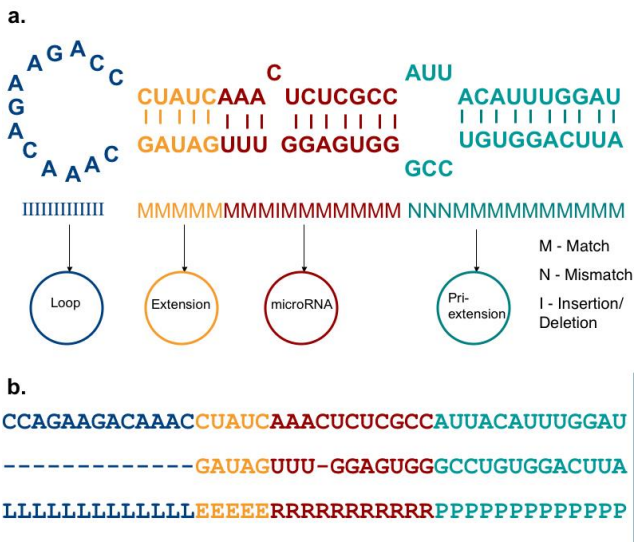
- miRNAs can function as tumour suppressors and oncogenes.
- Gene therapies that use miRNAs might be an effective approach to blocking tumour progression.
- miRNAs such as let-7, which has been shown to negatively regulate the Ras oncogenes, and miR-15 and miR-16, which negatively regulate BCL2, are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

Topic 72

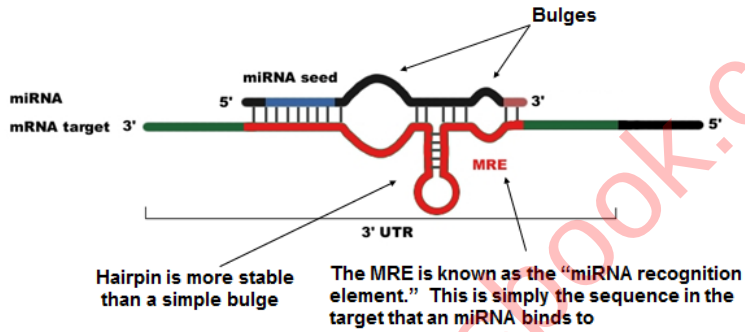
- **miRNA - Mode of Action**

Mode of Action of miRNA

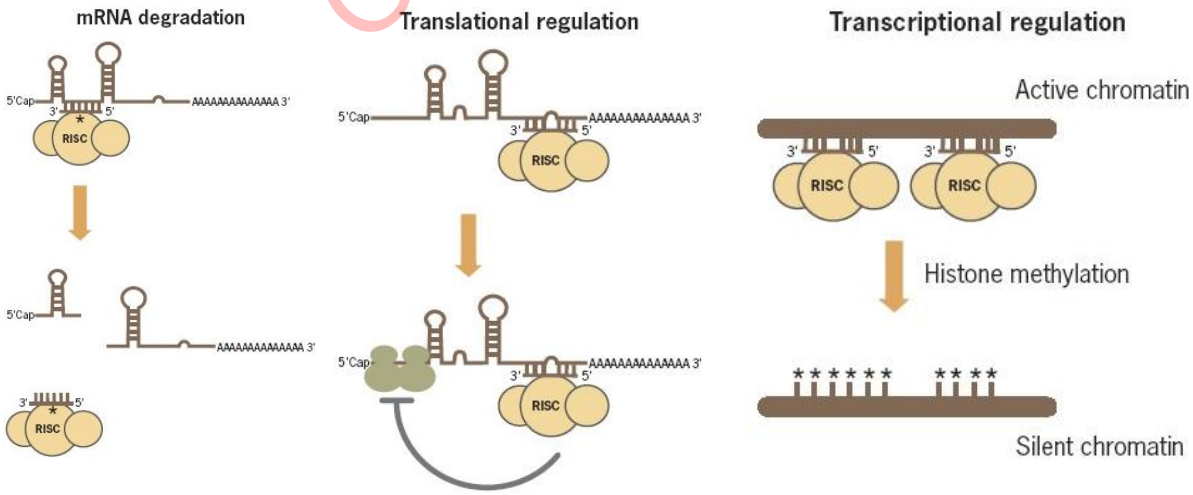
- Three different mode of action of miRNA
- mRNA degradation
- Transcriptional regulation
- Translational regulation
- **Structure of miRNA at Nucleotide level**



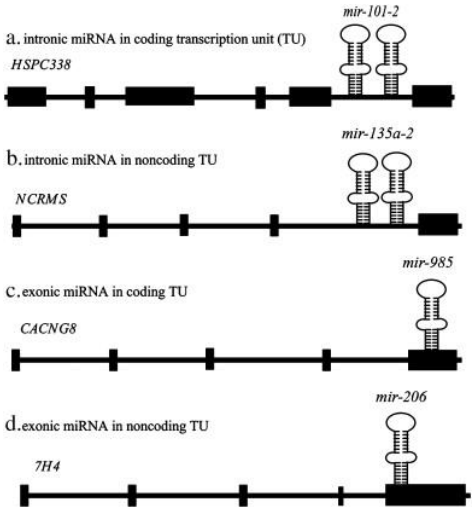
miRNA Binding to Target (mRNA)



Mode of Action of miRNA - mRNA Degradation



Schematic Illustration of the Genomic Organization and Structure of miRNA Genes



miRNA - Mode of Action

Mode of action of miRNA

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

Lecture 73

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
- Xist appears to recruit a specific histone isoform which maintains the chromosome in inactive state
- Additionally, Xist containing complexes recruit histone deacetylases and methylases
- 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

shRNA and Genetic Imprinting

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

Lecture 74

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

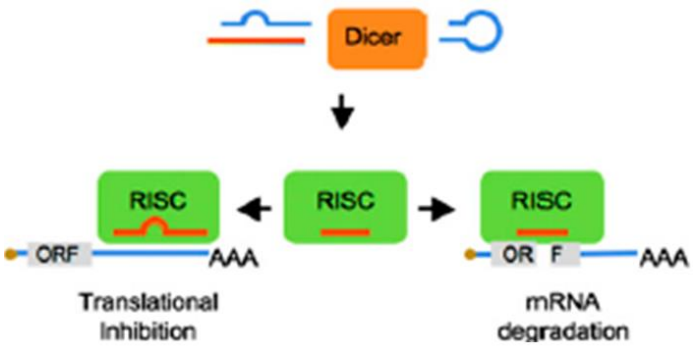
- Unwinding of double-stranded siRNA by ATP independent helicase.
- Active components of RISC are endonucleases called argonaute proteins which cleave the target mRNA strand.
- RNA Induced Silencing Complex

RNA Induced Silencing Complex - Structure

RNA Induced Silencing Complex

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).



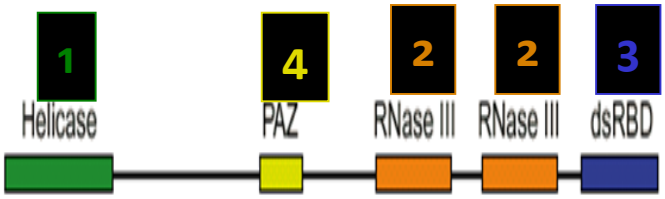
- RISC bound to partially complementary mRNA induces translational repression: results in destabilization of target mRNA
- The more RISC bound; greater the inhibitory effect
- RISC bound to perfectly complementary mRNA (to miRNA) are cleaved and degraded by RISC in a process similar to RNAi
- RISC is a large (~500-kDa) RNA-multiprotein complex, which triggers mRNA degradation in response to siRNA

Lecture 75

- **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 cleaves double-stranded RNA (dsRNA) and pre-microRNA into short double-stranded RNA fragments called small interfering RNA and microRNA
- 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 homologs exist in many organisms including C. elegans, Drosphila, yeast and humans
- Amino-terminal helicase domain
- Dual RNase III motifs in the carboxy terminal segment
- dsRNA binding domain
- PAZ domain - 110-130 amino acids
-

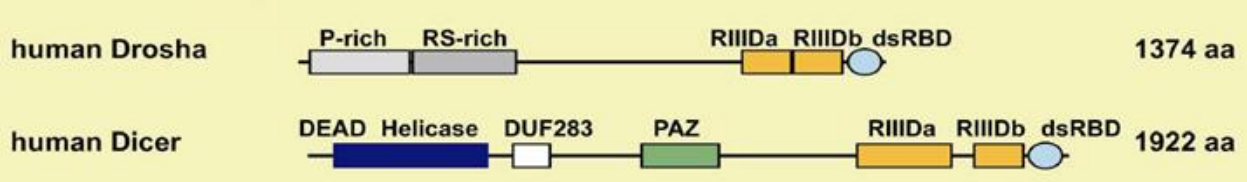


Drosha

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

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

Dicer and Drosha - Humans



Dicer and Drosha

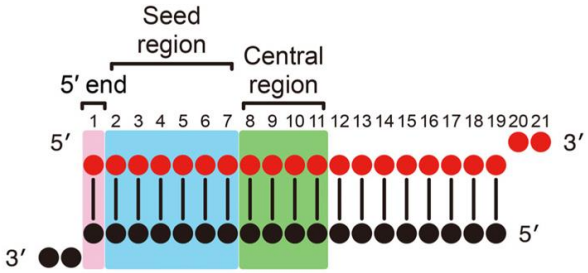
- Dicer is an enzyme that is part of the RNase III family
- Drosha belongs to class 2 ribonuclease III enzyme family

Lecture 76

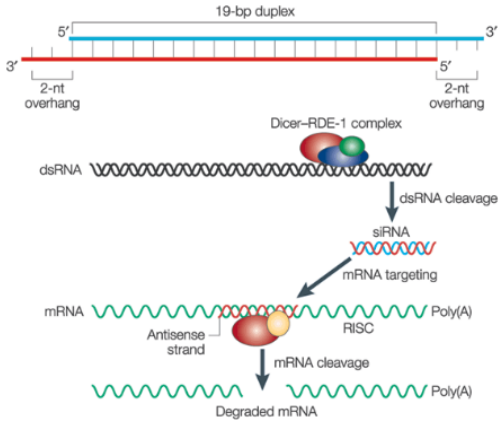
- **Small Interfering RNA (siRNA)**

siRNA

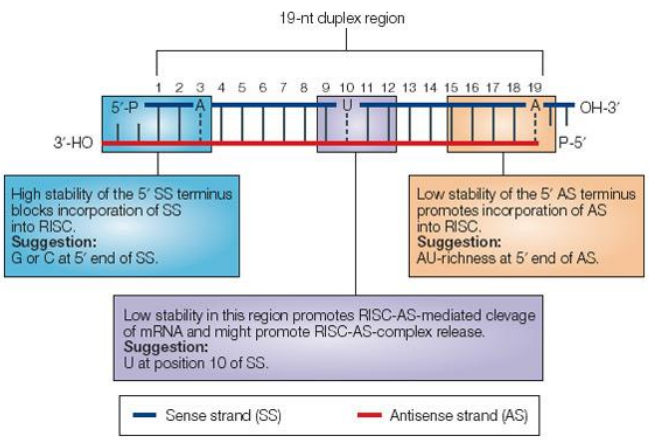
- Small interfering RNA (siRNA), sometimes known as short interfering RNA, are a class 20-25 nucleotide-long RNA molecules that interfere with the expression of genes.
- They are naturally produced as part of the RNA interference (RNAi) pathway by the enzyme Dicer.
- Exogenously introduced by investigators to bring about knockdown of a particular gene.
- 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.



- short (usually 21-nt) double-strand of RNA (dsRNA) with 2-nt overhangs on either end.
- Short double-strand of RNA (dsRNA) with 2-nt overhangs on either end.
- Small interfering RNAs that have an integral role in the phenomenon of RNA interference(RNAi),a form of post-transcriptional gene silencing
- A single base pair difference between the siRNA template and the target mRNA is enough to block the process.

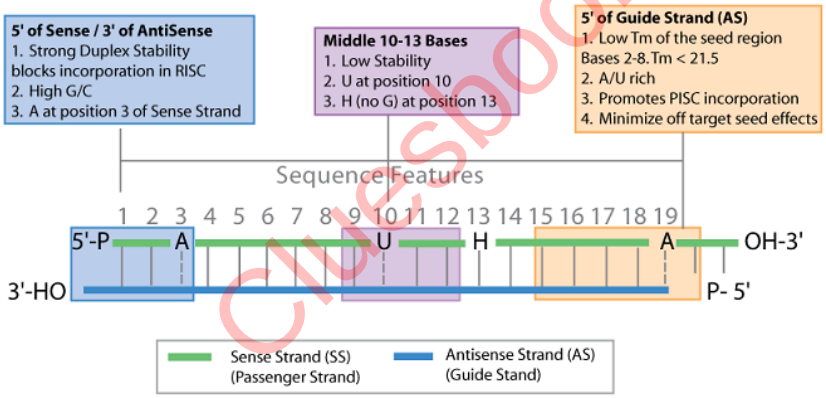


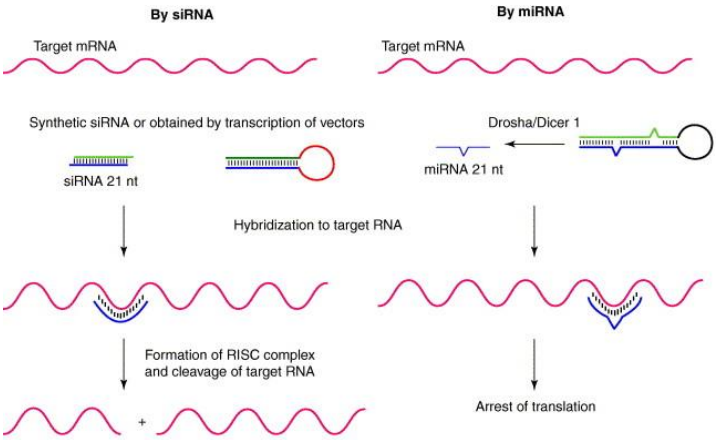
- **Small interfering RNA**
- Several different methods of expression
- Several different methods of delivery
- Dicer (type III RNAse III) cleaves long dsRNA into siRNA 21-25nt dsRNA from exogenous sources
- Small interfering RNA (siRNA), sometimes known as short interfering RNA, are a class 20-25 nucleotide-long RNA molecules that interfere with the expression of genes.



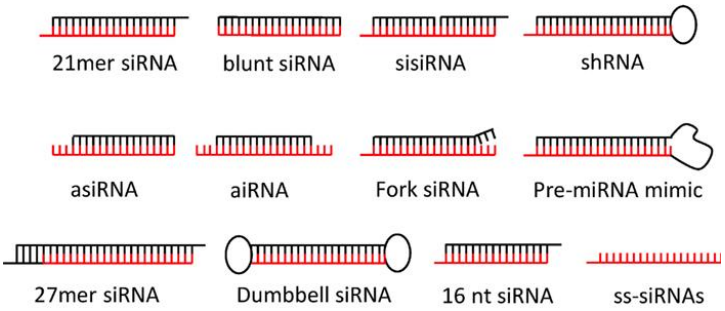
Lecture 77

- **Design of siRNA**
- 21-23 nt dsRNA,
- GC% slightly < 50%
- Perfect complimentary to target mRNA
- Targeting 3'UTR works better than 5' UTR

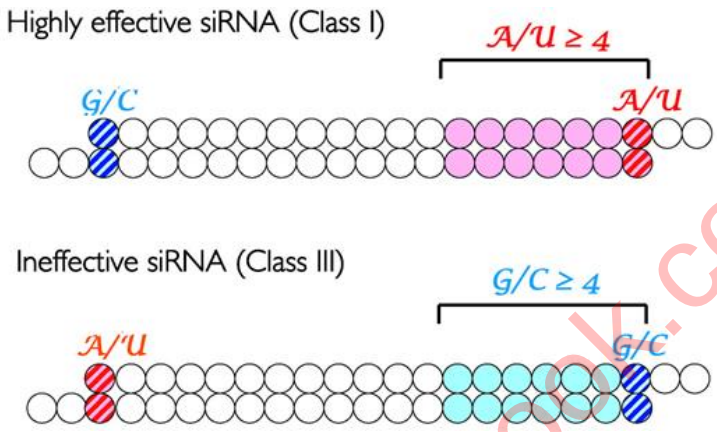


- 

- **siRNA/miRNA: Difference in mode of action**



Different 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

Lecture 78

Applications of siRNA

Applications of siRNA

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

Applications of siRNA

    - Applications for cancer prevention, infections and developmental defects

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

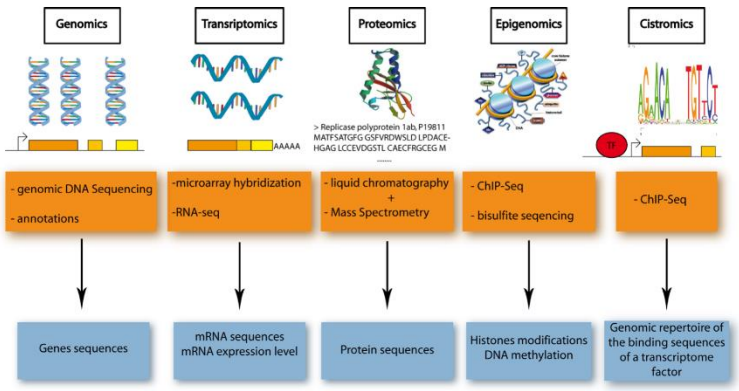
- Determining protein function
    - Easier than a knockout
    - Used for partial knockdowns

Lecture 79

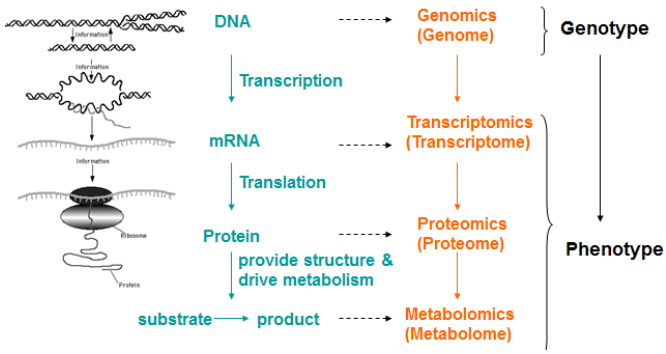
- **Transcriptomics**

Transcriptome

- The population of mRNAs expressed by a genome at any given time
- The complete collection of transcribed elements of the genome
- The study of characteristics and regulation of the functional RNA transcript population of a cell/s or organism at a specific time



Transcriptomics



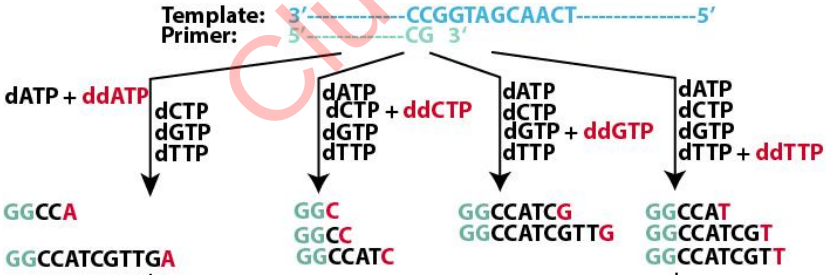
- Percentage of the genetic code that is transcribed into RNA molecules (depends on development, environment, time of the day, tissue)
- Collection of all gene transcripts present in a given cell/tissue at a given time
- Genes and pathways involved in biological processes
- Genes with similar expression may be functionally related and under the same genetic control mechanism
- Help elucidating the function of unknown genes based on their spatial and temporal expression
- Identifies marker genes for diagnosis of diseases

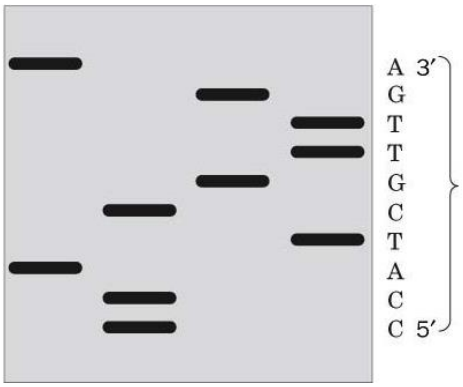
Lecture 80

Sequencing Techniques

Sanger's Method

- Frederick Sanger (MRC Centre Cambridge, UK)
- DNA sequencing using **chain-terminating inhibitors** (1977)
- **Fred Sanger**
- Nobel prize in Chemistry 1958 and 1980 ,Sanger Institute, University of Cambridge



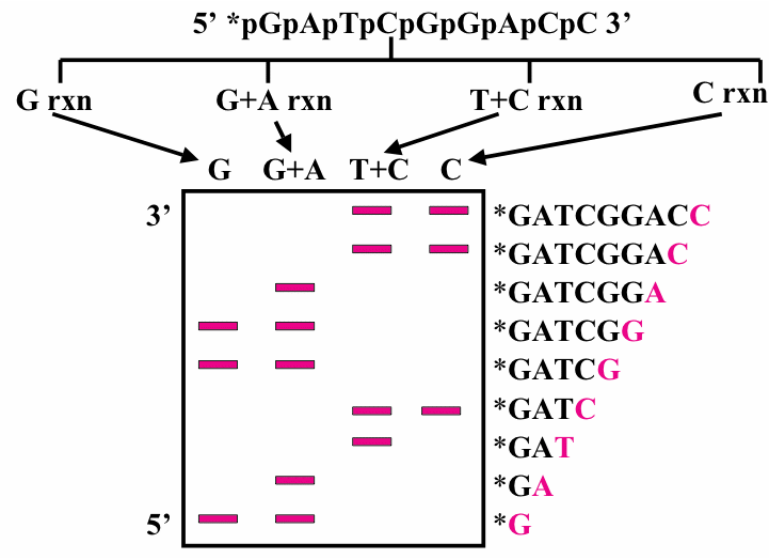
- 

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

Base Modifications

- Purines (A+G) are depurinated by formic acid
- Gs methylated by dimethyl sulfate
- Pyrimidines (C+T) are hydrolized using hydrazine
- NaCl retards hydrazine reaction with **T** in **C-only** reaction
- **Maxam-Gilbert's Method**
- One base modification per DNA molecule
- Cleavage is done where the base is modified



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

Lecture 81

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

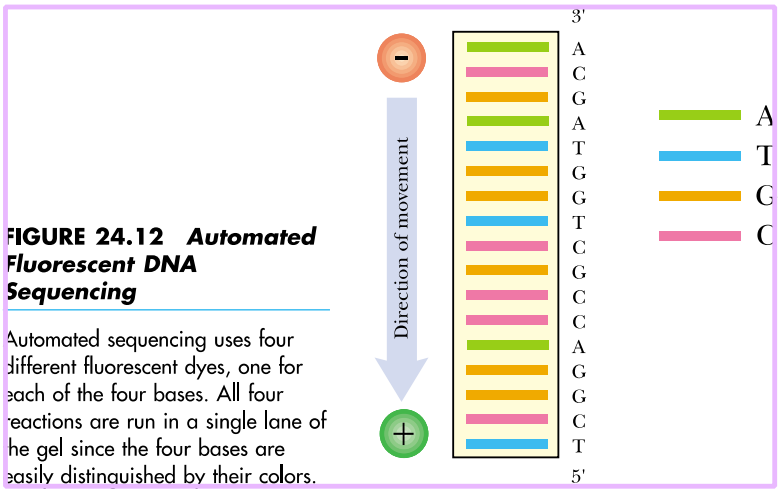


FIGURE 24.12 ***Automated Fluorescent DNA Sequencing***

Automated sequencing uses four different fluorescent dyes, one for each of the four bases. All four reactions are run in a single lane of the gel since the four bases are easily distinguished by their colors.

The Emergence of DNA Chip Technology

DNA arrays are used for a variety of purposes, including sequencing. Large numbers of probes are bound to the chip and hybridization with target DNA occurs on the chip surface.

Earlier DNA technology was largely based on gel electrophoresis, an approach that is both difficult to automate and labor intensive. DNA chips were developed to allow automated side-by-side analysis of multiple DNA sequences. In practice the simultaneous analysis of thousands of DNA sequences is possible. The first chip was introduced by a company called Affymetrix in California in the early 1990's. Since then DNA chips have been used for a variety of purposes including sequencing, detection of mutations and gene expression. DNA chips all rely on hybridization between single-stranded DNA permanently attached to the chip and DNA (or RNA) in solution. Many different DNA molecules are attached to a single chip forming an array of spots on a solid support (the chip). The DNA or RNA to be analyzed must be labeled, usually with fluorescent dyes. Hybridization at each spot is scanned and the signals are analyzed by appropriate software to generate colorful data arrays. Two major variants of the DNA chip exist. Earlier chips mostly used short oligonucleotides. However, it is also possible to attach full length cDNA molecules. Prefabricated cDNA or oligonucleotides may be attached to the chip. Alternatively, oligonucleotides may be synthesized directly onto the surface of the chip by a modification of the phosphoramidite method described in Chapter 21. Modern arrays may have 100,000 or more oligonucleotides mounted on a single chip.

The Oligonucleotide Array Detector

DNA arrays can detect the presence of multiple small fragments of DNA sequence. A computer then compiles the overall sequence.

The **oligonucleotide array detector** simultaneously detects and identifies lots of short DNA fragments (i.e., oligonucleotides). It can be used both for diagnostic purposes and for large scale DNA sequencing. The key principle involved is DNA-DNA hybridization (see Ch. 21).

Consider a piece of DNA of unknown sequence. This is denatured to give single strands and one of these is tested for hybridization to a known probe sequence of say, eight bases (an octonucleotide; e.g., CGCGCCCG). If the unknown DNA binds to the probe, then the probe sequence occurs somewhere in the complementary strand of the unknown DNA. The unknown DNA is then tested for hybridization to all other possible stretches of eight bases, one at a time, to see which are found.

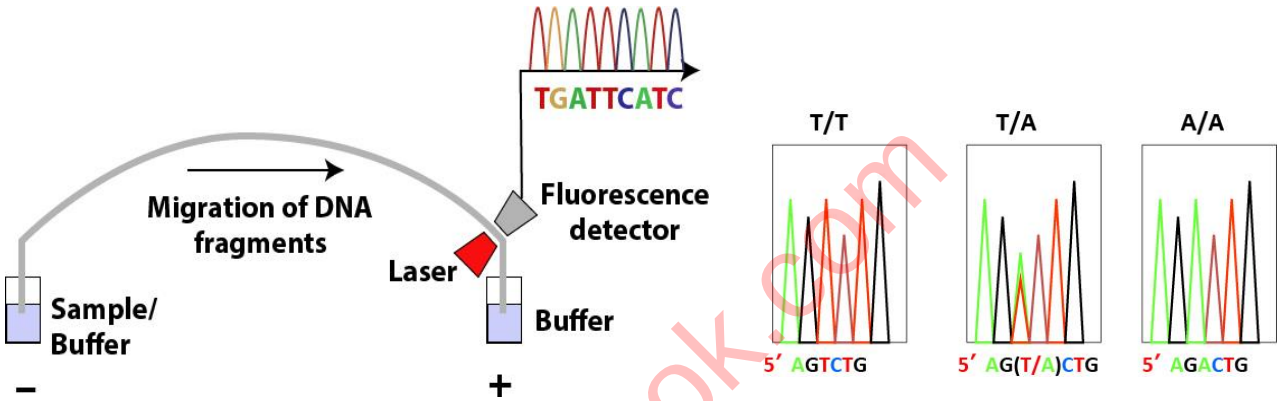
oligonucleotide array detector Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or DNA chip

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

- 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



Conclusions

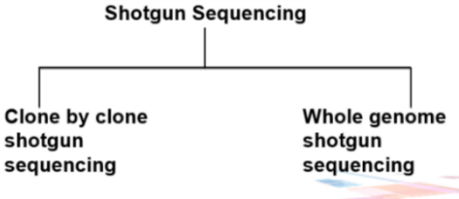
- Use of flourescent dyes and an automated system improved the effeciency of sequencing

Lecture 82

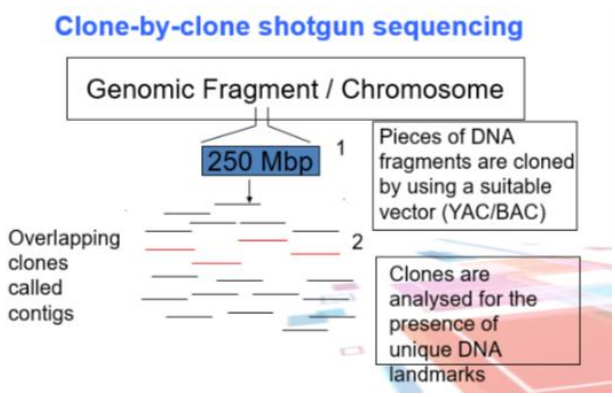
Shotgun Sequencing

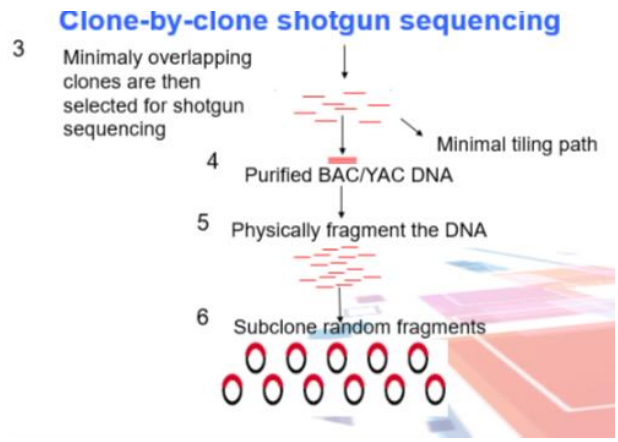
Shotgun sequencing

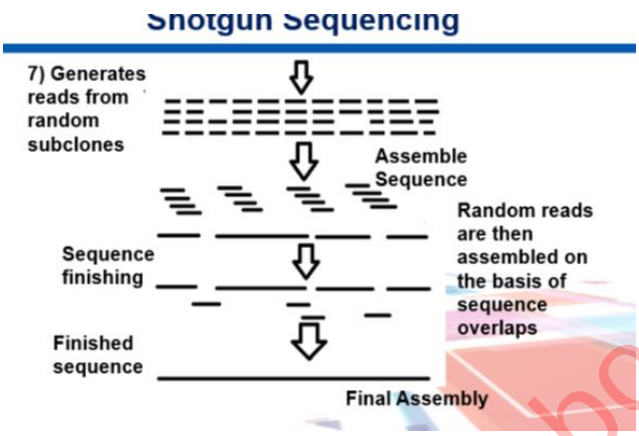
- The DNA is broken up randomly into numerous small segments, which are sequenced
- Named by analogy with the rapidly expanding, quasi-random firing pattern of a shotgun



- **Clone by clone shotgun sequencing**
- Also known as hierarchical shotgun sequencing or map based 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

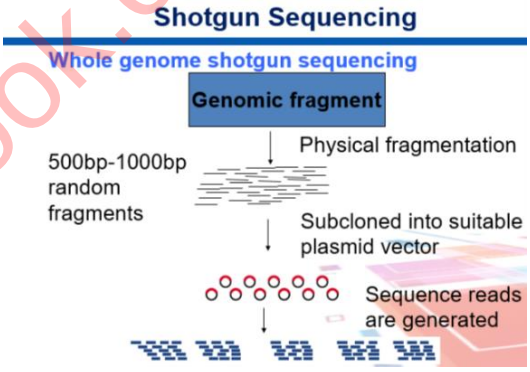


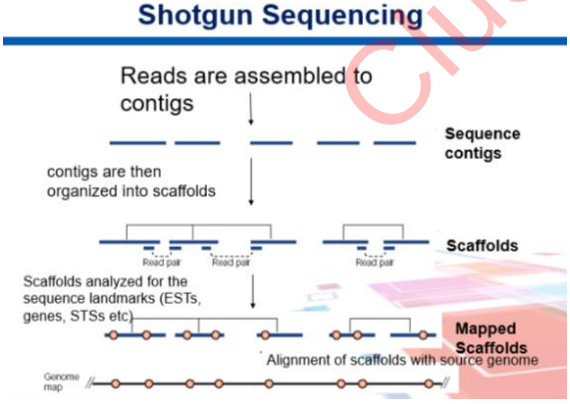




Whole Genome Shotgun Sequencing (WGS)

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





Shotgun sequencing

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

Lecture 83

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

Historical Perspective

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

Historical Perspective

- It led to the development of Next Generation Sequencing Techniques
- 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
- 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
- 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 **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

NGS Technologies

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

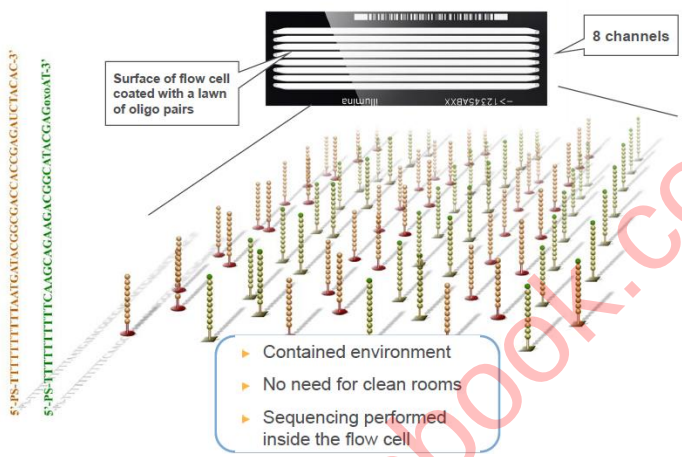
NGS

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

NGS

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

The flow cell design

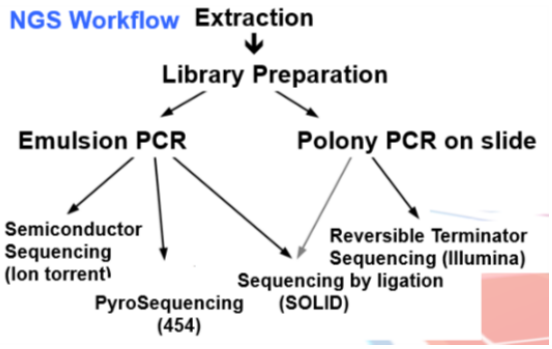


Generations of sequencing

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

Generations of sequencing

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



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

Conclusions

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

Lecture 84

NGS Platforms

NGS Technologies

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

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/

	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**

Illumina

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

Applied Biosystems Inc (ABI)

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

Ion Torrent

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

Lecture 85

454 Pyrosequencing

Sequencing in 454 machines is performed by pyrosequencing

Principle

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

DNA_n + dNTP à

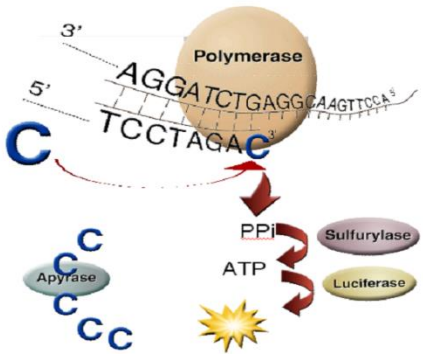
DNA_{n+1} + PP_i

Principle

PPi is used to generate ATP from adenosine phosphosulfate (APS)

- APS + PP_I à ATP

ATP and luciferase generate light by conversion of luciferin to oxyluciferin



Principle

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

loading

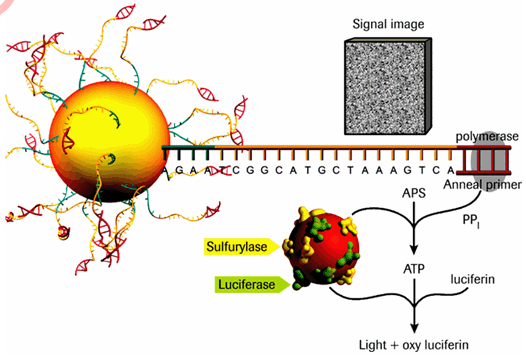
- After amplification, the emulsion shell is broken and the clonally amplified beads are ready for loading onto the fibre-optic PicoTiterDevice for sequencing

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

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



Sequencing

- CCD camera records light flashes from each well

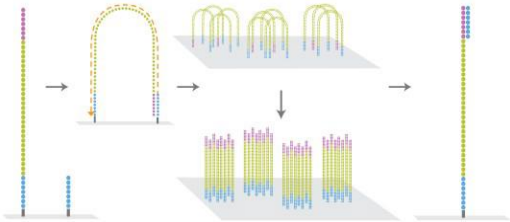
Conclusions

- Sequencing in 454 machines is performed by pyrosequencing

Lecture 86

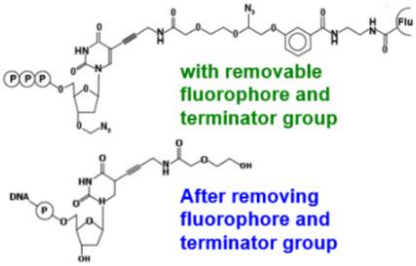
Illumina Sequencing

- 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

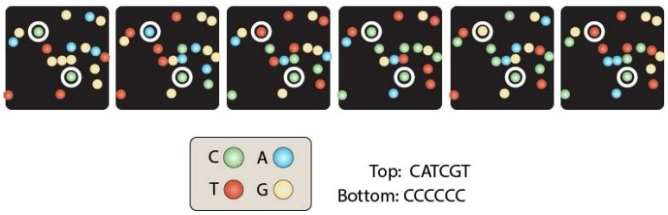


Sequencing

- Illumina uses
- reversible terminators with flourophores



Sequencing

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

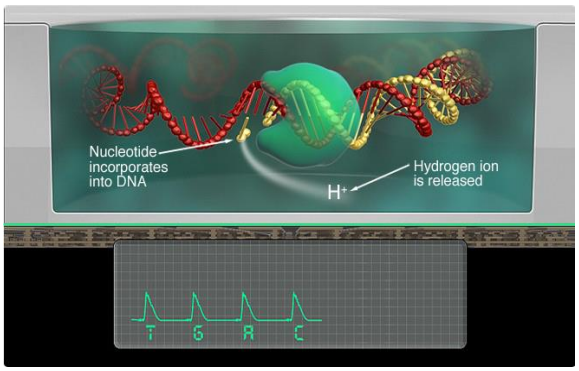
- **Conclusions**
- Illumina performs sequencing by synthesis (SBS) using reversible terminators with flourophores

Lecutre 87

Ion Semiconductor Sequencing

NGS Issues

- Alternating phases of nucleotide incorporation, signal detection and dephasing
- PCR biases
- Duplicates, PCR errors
- **Between 2nd and 3rd generation**
- **Still use 'wash-and-scan'**
- Ion Torrent
- Non-optical sequencing
- Emulsion PCR
- **Between 2nd and 3rd generation**
- **Still use 'wash-and-scan'**
- Helicose Genetic Analysis System
- Single molecule sequencing
- **Semiconductor Chips**

- Three types of semiconductor chips:
- **314 – 20Mb**
- **316 - 200Mb**
- **318 – 1Gb**
- 

Lecture 88

3rd Generation sequencing

Introduction

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

3GS Plateforms

- Helicos
- Pacific Biosciences
- Oxford nanopore

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

- **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

Pacific Biosciences

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

- **Pacific Biosciences Pacbio RS**

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

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

- **Oxford Nanopore**
- MinION
- GridION System
-

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

Advantages of single molecule sequencing

- Less sample preparation (no PCR)
- No amplification
- no PCR errors
- fewer contamination issues
- no GC-bias
- analyze every sample (unPCRable / unclonable)
- analyze low quality DNA (museum, archeological, forensics samples)
- Absolute quantification
- Sequence RNA directly

Lecture 89

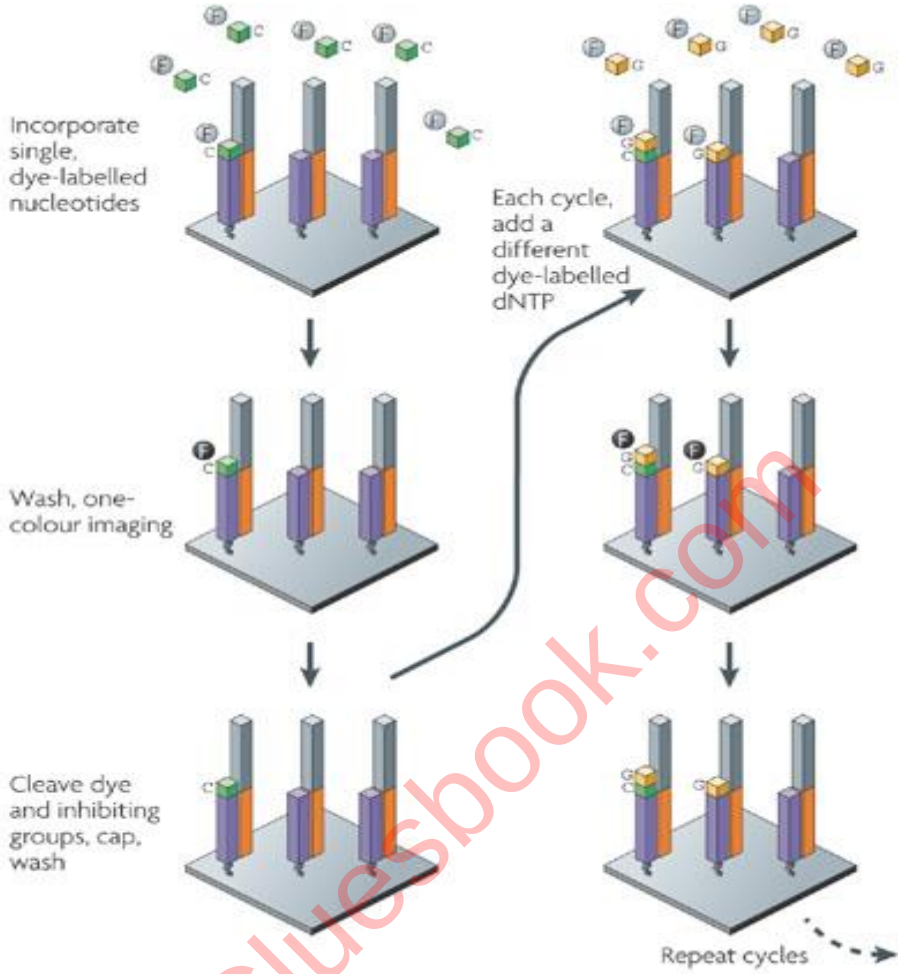
Helicose sequencing

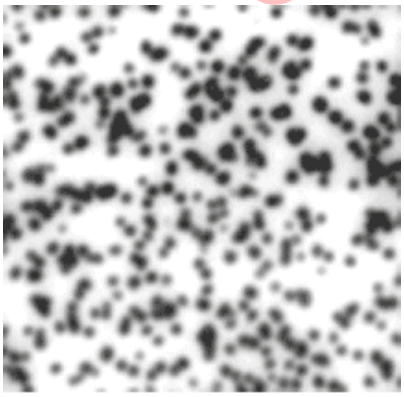
Helicos

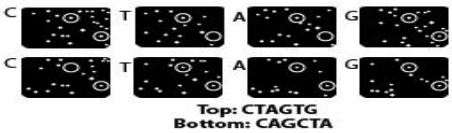
- Images the extension of individual DNA molecules using a defined primer and individual fluorescently labeled nucleotides, which contain a "Virtual Terminator" preventing incorporation of multiple nucleotides per cycle

Helicos

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



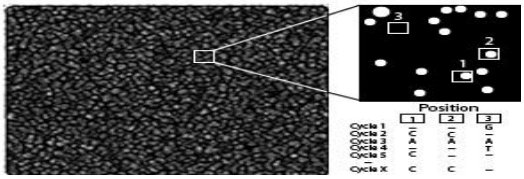




Top: CTAGTG
Bottom: CAGCTA

Nucleotides flown sequentially

(Dark nucleotides: incorporation not detected)



Conclusions

- Helicos was an effort towards single molecule sequencing using illumina like approach

- It was an important step towards third generation sequecing

Lecture 90

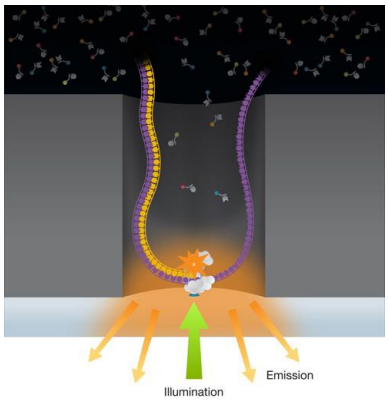
Pacbio sequencing

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

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
- 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



SMRT sequencing

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

Lecture 91

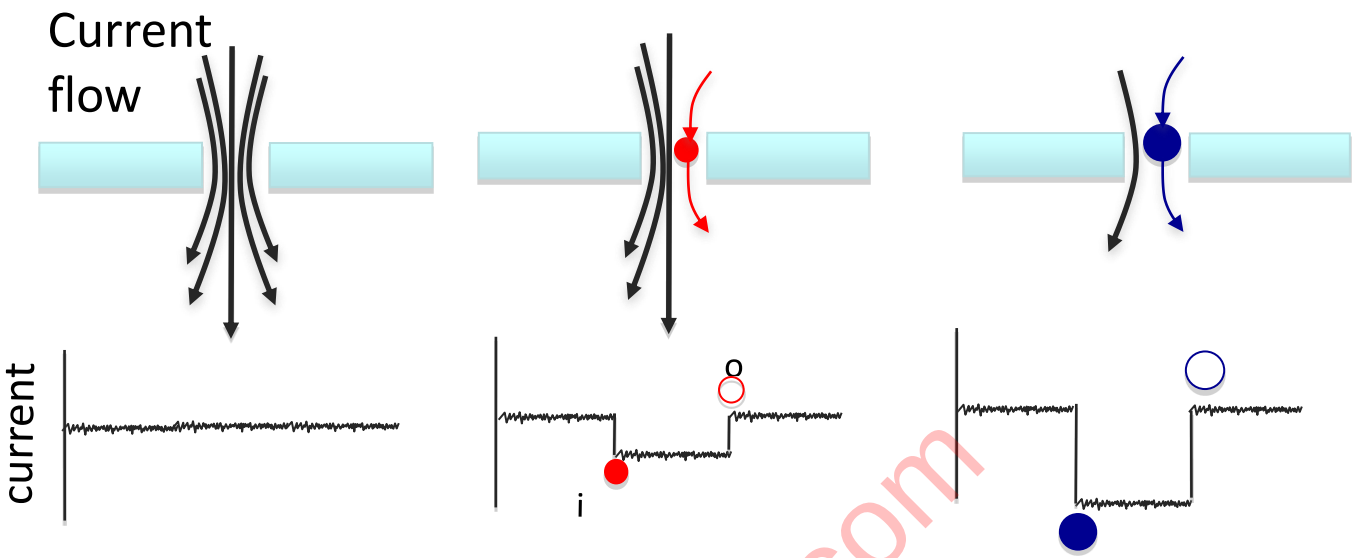
Oxford Nanopore sequencing

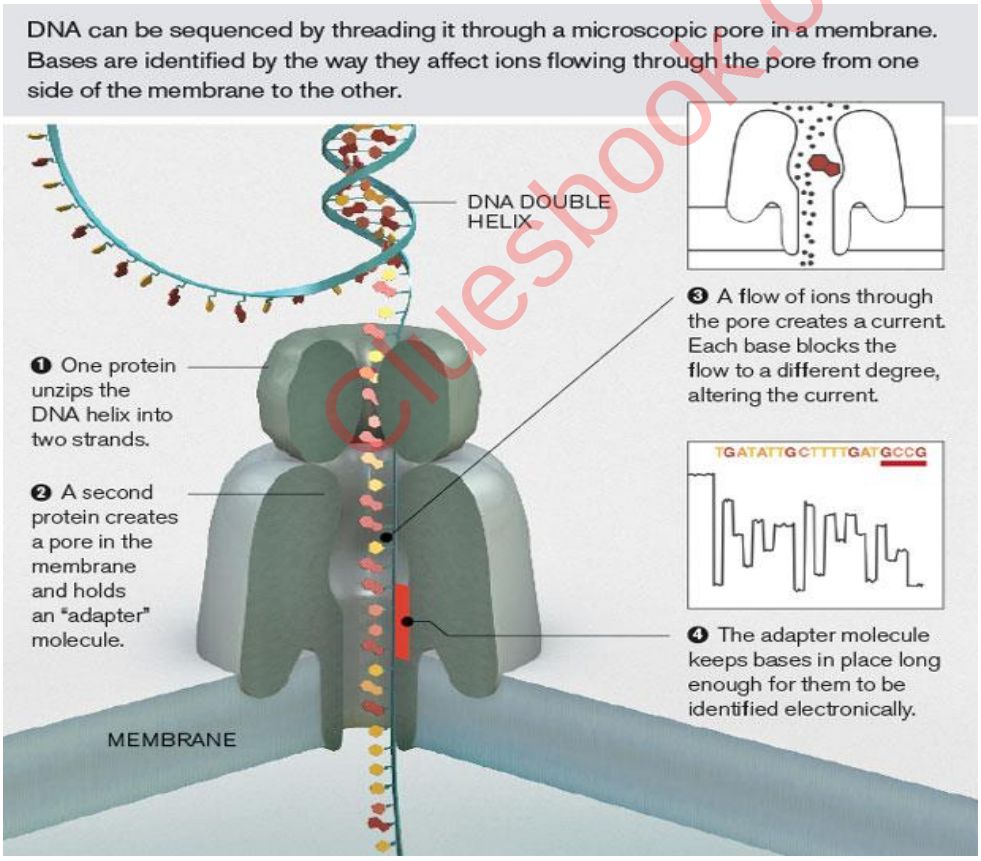
Nanopore sequencing

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

Nanopore sequencing

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





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

Lecture 92

Comparison of NGS Methods

Which Plate-form to use

Depends on

- Biological question
- Budget
- Analysis team's expertise

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

	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
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

	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

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

Lecture 93

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?

Issues

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

Issues

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

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)

Historical viewpoint

A tangled knot;

One gene – one function

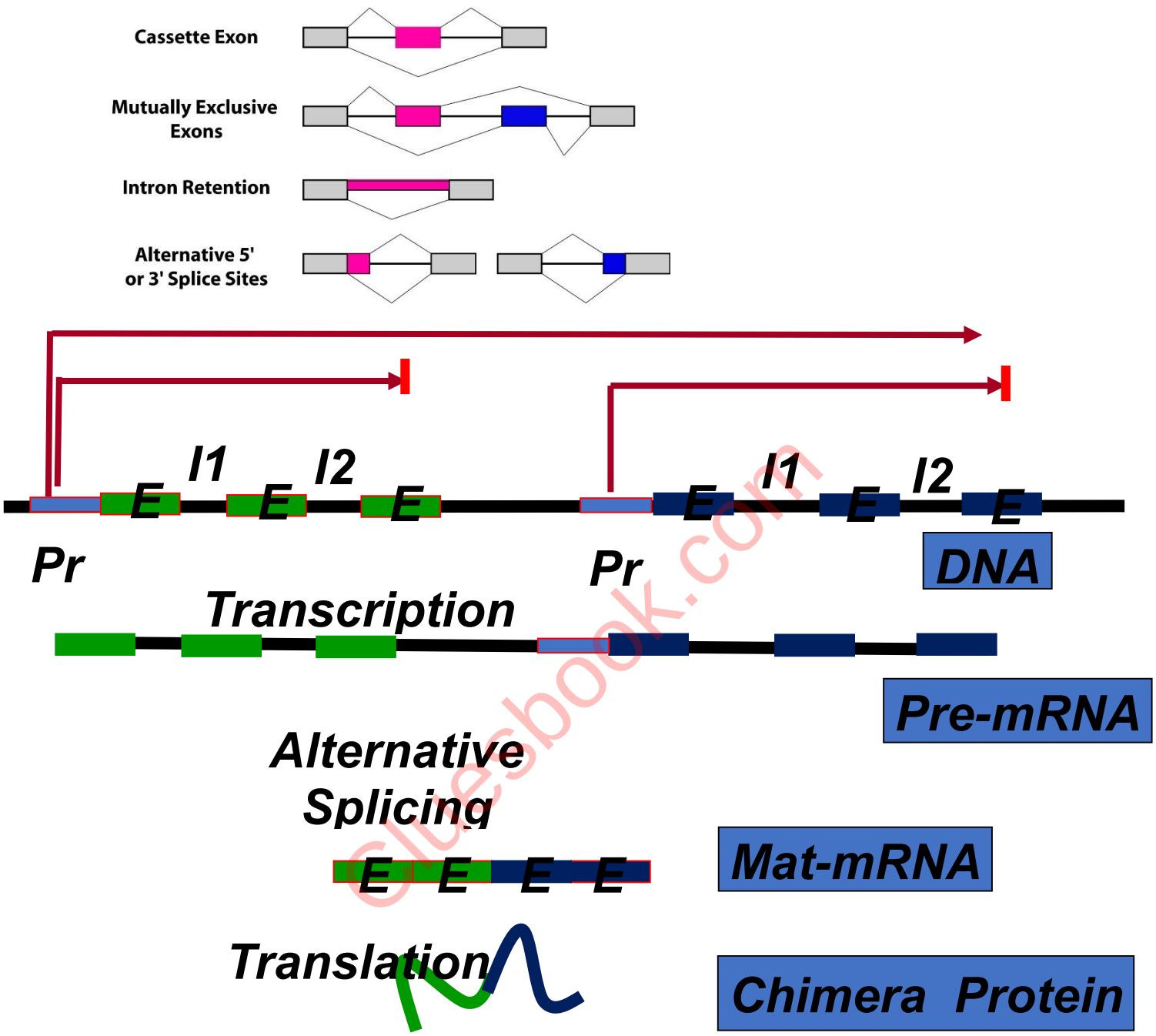
One gene – one enzyme

One gene – one polypeptide

One gene – A number of polypeptides

Number of genes – one polypeptide

Alternative splicing



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

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

Lecture 94

Approaches

Predictions are derived from different computational methods

Two famous approaches;

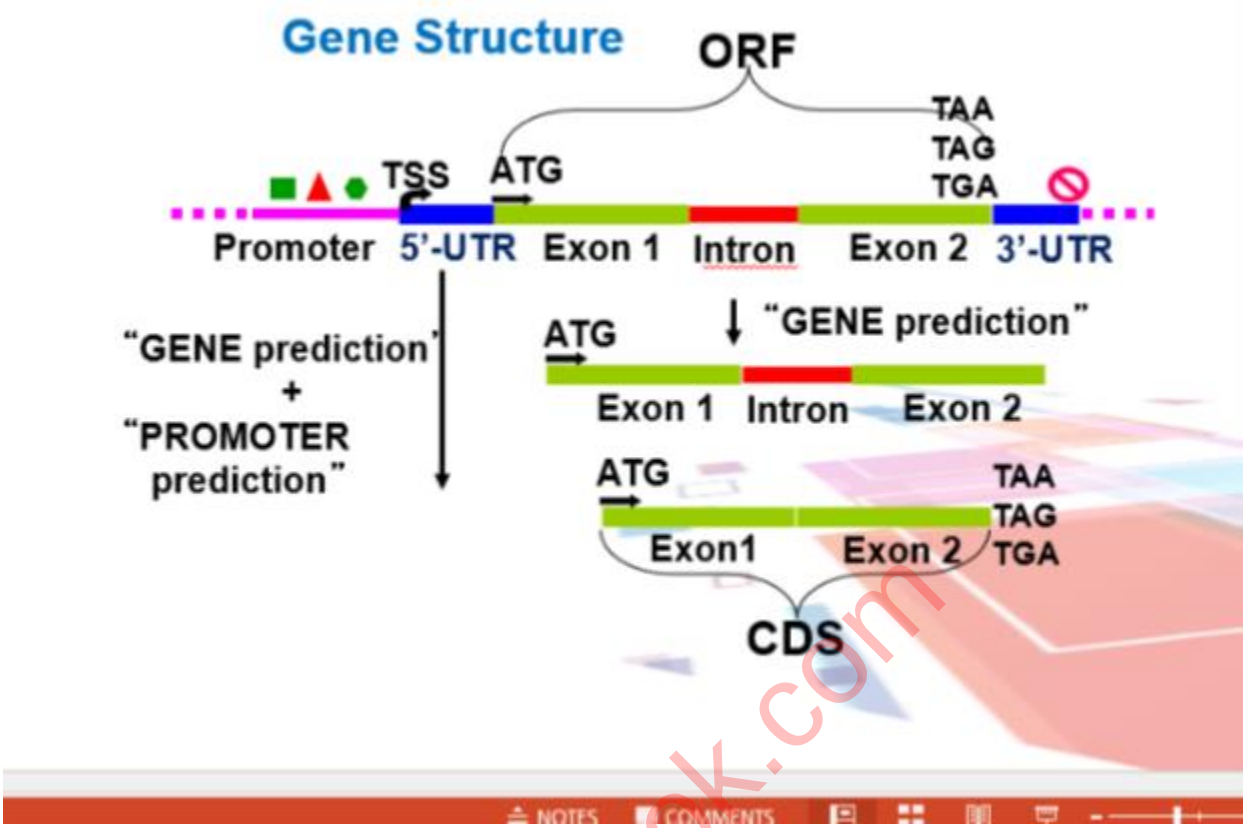
***"Ab initio"* gene finding**

Comparative Approach

***"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
- ***Comparative* gene finding**
- Look for genes by comparing segments of sequence with those of known genes
- ***Comparative* gene finding**
- Tends to underestimate since limited to recognize only genes similar to what have been seen before



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)

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

Conclusions

There are two famous approaches for gene predictions

***"Ab initio"* gene finding**

Comparative Approach

Finding eukaryotic genes is complicated

Lecture 95

Gene Prediction

Gene prediction Involves;

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

Coding Sequence or CDS(coding DNA sequence)

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

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

Coding Sequence or CDS (coding DNAsequence)

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

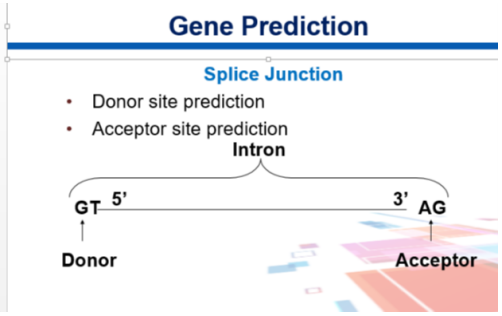
ORF (Open Reading Frame)

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

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



1st letter	Second Letter: U		C		A		G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Intron-Splicing Site-Exon

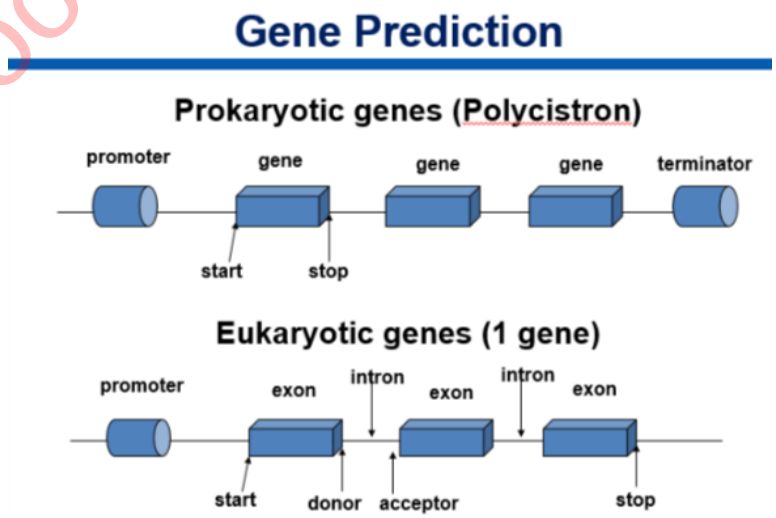
GTCTTTTTGAGATAGAAAACAAAAAAGAACTTATAGCTCCAAATATCTTTCTTTGGATAACAACTCAAATAATATTT
GCTTTCTTGTGATTGTAGATGGTTTGGGAACCGAAGTCAATTTGGACGGGATTGCTTACTACAATAATGTTATTAAT
GCTCTTCTAGAGAAAGGTATTGAAGCATTCCTAATAAATTTGCTCCTAGTATAATTGTGGTTTATGCATAAGAACTT
CTCCTGCCAGGTATTCAGCCTTATGTAACTTTGTACCACTGGGACCTTCCTTTGCATCTCCAGGAATCAATGAATGG
ATGGTTAAATAGGAAGATTGTGTAAGTATTAAAATGCCAGGTGTTTAAATATGAAGTTACTGCATAT

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)



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**

Signal

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

Content

exons, introns, UTRs, promoter regions

Conclusions

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

Conclusions

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

Lecture 96

- **Genetics and Genomics**

Genes Predictions - Prokaryotes

- Locating genes within genomic sequence.
- Defining initiation termination sites of genes.
- Extracting the coding region of each gene.
- Identifying function for coding region.
- A region of the genome that codes for a functional component such as an RNA or protein.

Identify coding regions computationally

from raw genomic sequence data

Translation utilizes a trinucleotide coding system: codons

- Translation begins at a start codon
- Translation ends at a stop codon
- 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 - Gene

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

Genes Predictions - Gene

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

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 – 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 – 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

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

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

Lecture 97

Genes Predictions - Eukaryotes

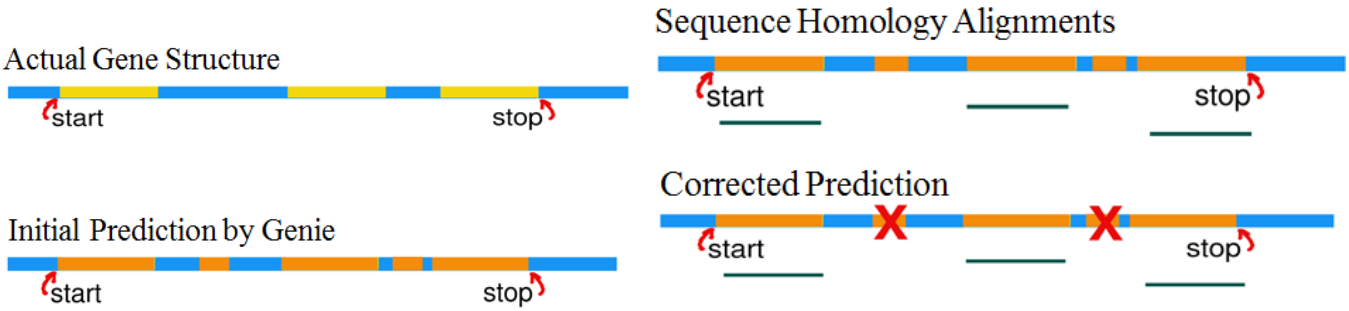
- Tools for finding genes in eukaryotes
- Genie
- Fgenes
- Genscan

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 Eukaryotes - Genie

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



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.

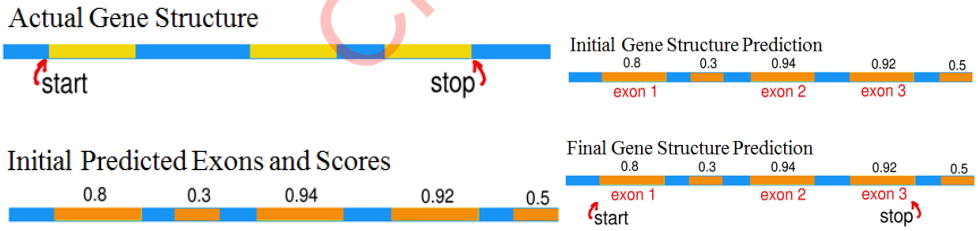
Fgenes is about 70% accurate in most mammalian genomes.

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.



Advantages	Disadvantages
• Alternative gene structures are reported. • Also attempts to identify putative promoter and poly-A sites.	• User cannot train models. • Only human model-based version is available for unrestricted public use.

Genes Predictions Eukaryotes - Genscan

Models for different states (GHMMs)

State 1 and 2: Exons and Introns

Length

Composition

State 3: Splice junctions

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 Eukaryotes – Genscan Output

Gene structure

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

Genes Predictions Eukaryotes – Genscan Output

- Poly-adenylation site
- Genscan is 80% accurate on human sequences.
- Eukaryotic gene structures can be quite complex.
- The best approaches to gene finding in eukaryotes combine different methods.

Lecture 98

- **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

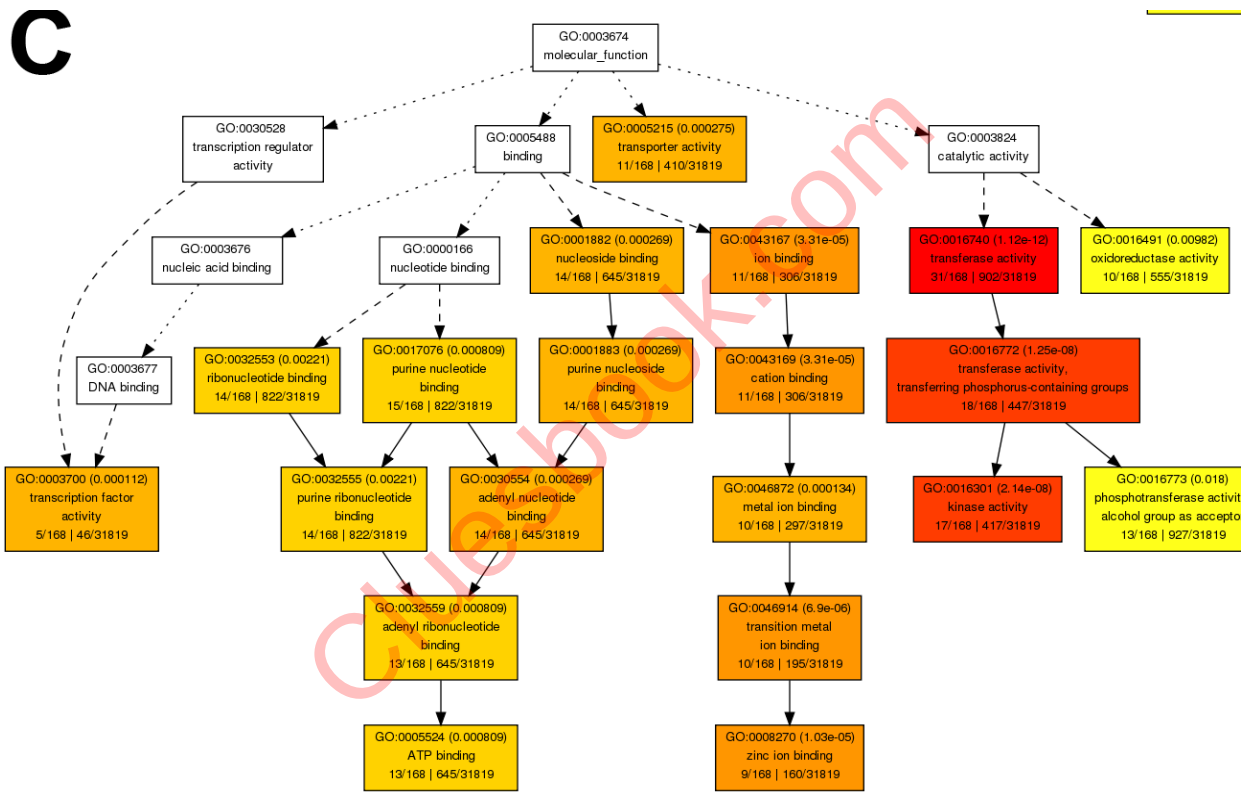
GO Terms

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

Three main types

Biological Process (BP) Cellular Components (CC)

Molecule Function (MF)



DAVID Tool

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

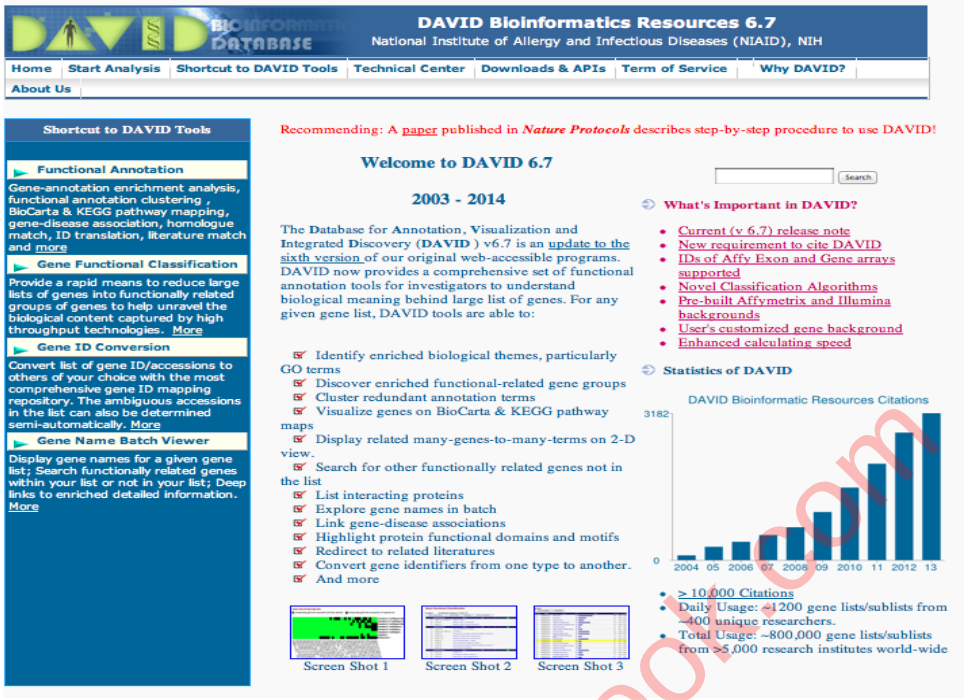
http://david.abcc.ncifcrf.gov/

- **GO Functional Analysis**

DAVID Tool

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

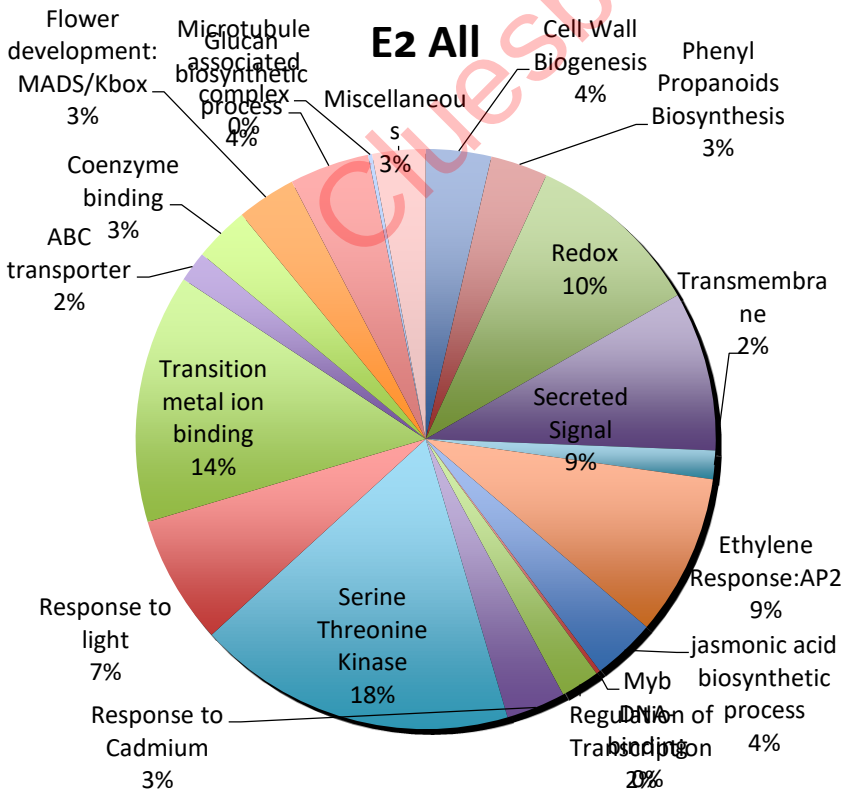
http://david.abcc.ncifcrf.gov/



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



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

Lecture 99

Gene Network

Introduction

A network is a topology that connects its components based upon some relationship with each other.

- Networks are implemented as graphs.

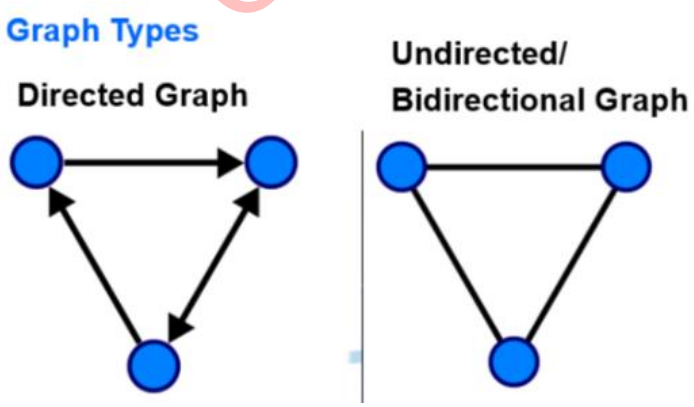
Graphs

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

Graph Types

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**

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

Lecture 100

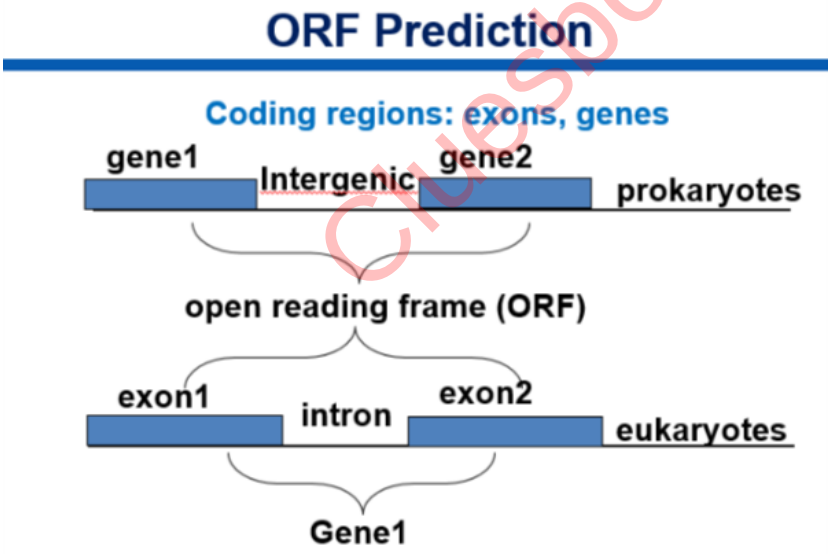
ORF Prediction

ORF (Open Reading Frame)

Gene finding, specially in prokaryotes starts form searching for an open reading frames (ORF)

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 and gene finding

ORF provide an important evidence in gene finding

Generally longer ORFs are preferred

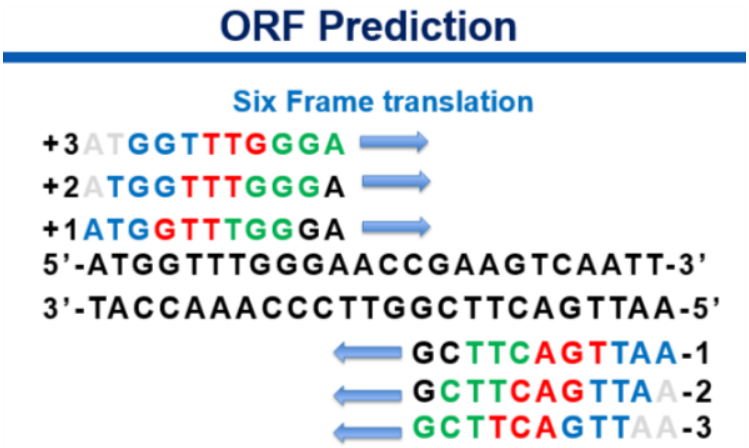
However presence of ORF not necessarily means the region is translated to a functional product

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

Reading Frames

Three on forward strand and three on complementary strand



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)

Lecture 101

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

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

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

Codon usage and likelihood ratio

An improvement may be;

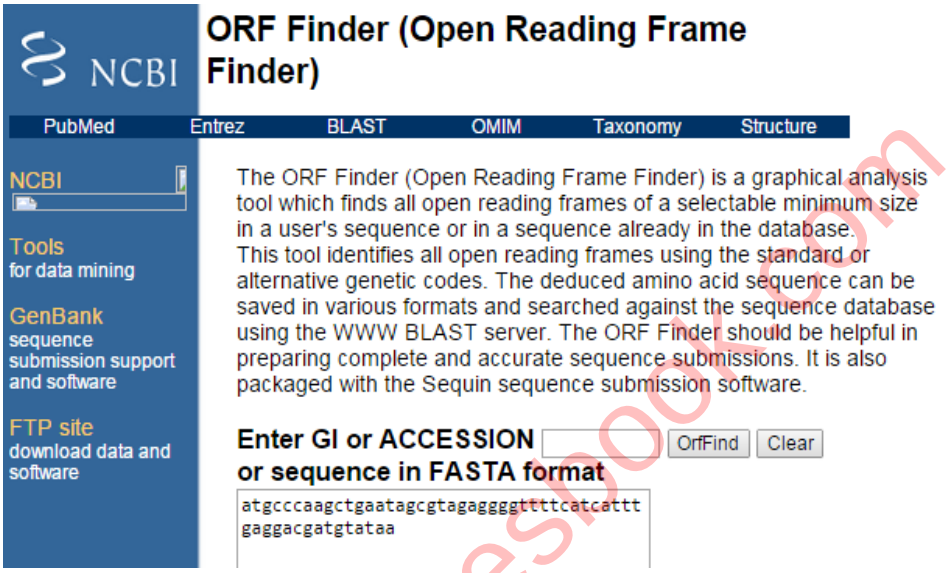
In-Frame hexamer count

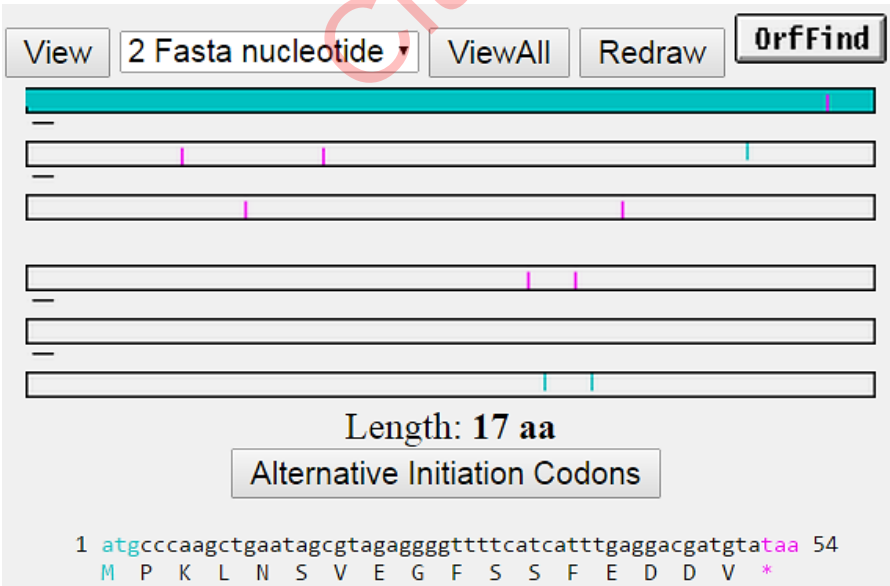
i.e frequencies of pairs of consecutive codons

ORF Finders

Tools are mainly based on pattern finding algorithms

- NCBI's ORF Finder
- ORF Investigator
- OrfPredictor





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

Lecture 102

TSS Prediction

Translation Start Site (TSS)

Translation starts with ATG that codes for methionine in a polypeptide

Assumption

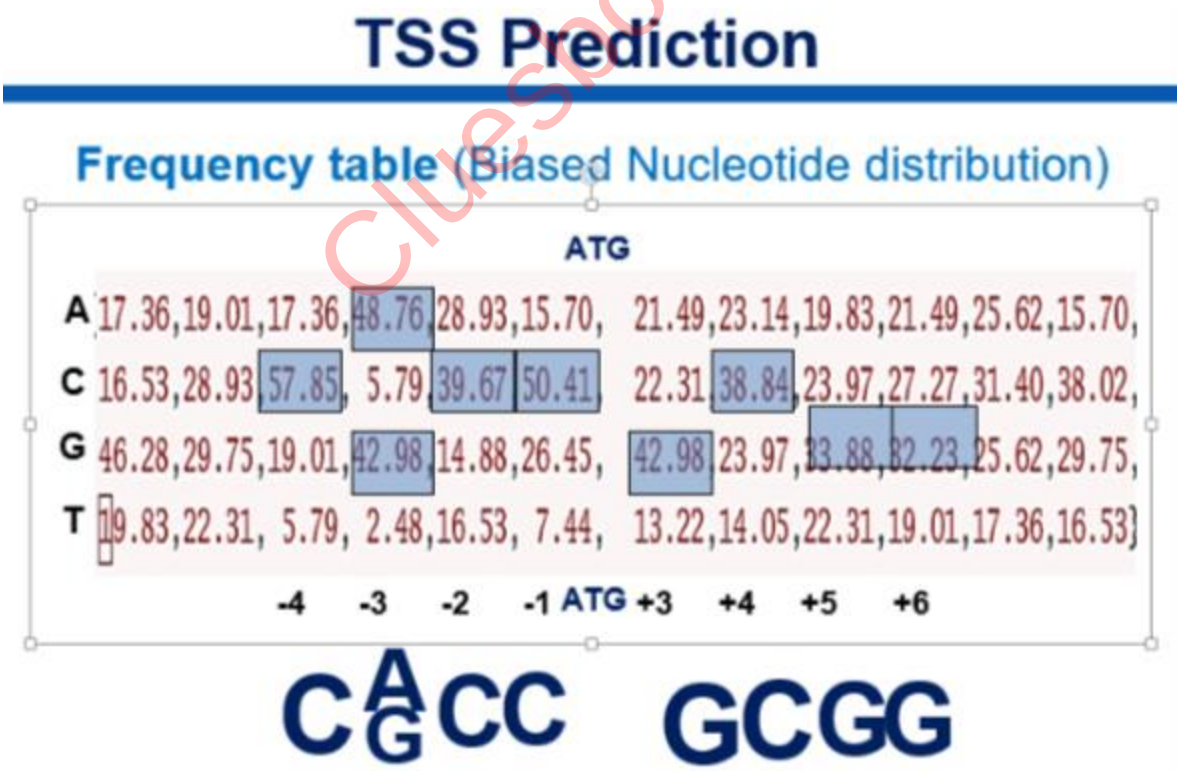
Certain nucleotides prefer to be around TSS than others

The "biased" nucleotide distribution is information is a basis for translation start 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

$F_i(X)=\Sigma \log(C_i(X)/N_i(X))$



Example

Which one is more probable to be a Translation Start?

CACC ATA GC

TCGA ATG TT

Solution

We can use frequency table and the scoring function as under;

$S_i = \Sigma \log (F_i (X)/0.25)$

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

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

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

Conclusions

- TSS prediction can be an important step in gene prediction
- TSS can be predicted while using the frequency of neighboring nucleotides

Lecture 103

Prediction of splice junctions

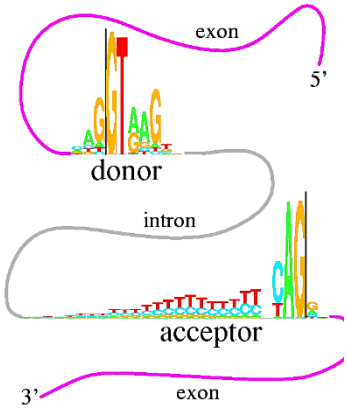
Splice Junctions

Donor site

- Coding region | GT

Acceptor

- YAG | coding region
- **Canonical form**
    - GT-AG: 99.24%



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

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.

Sequence LOGOS

- Overall height at position is proportional to the information content

Sequence LOGOS

- Proportions of each nucleotide/amino acid are in relation to their observed frequency, with most frequent on top, next most frequent below
- **Non Canonical Splice Junctions**
- In addition to canonical **GT-AG (99.24%)**;

- **GC-AG: 0.69%**
- **AT-AC: 0.05%**
- **Others: 0.02%**
- **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
- **Information Content (IC)**
- **$S_i = \Sigma \log (F_i (X)/0.25)$**
- Use "information content as a criterion for determining the length of flanks
- **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

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

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

$\Sigma \log (F_i (X)/0.25)$

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

Conclusions

Like TSS, the flanks of splice junctions show "biased" distributions of nucleotides in certain positions

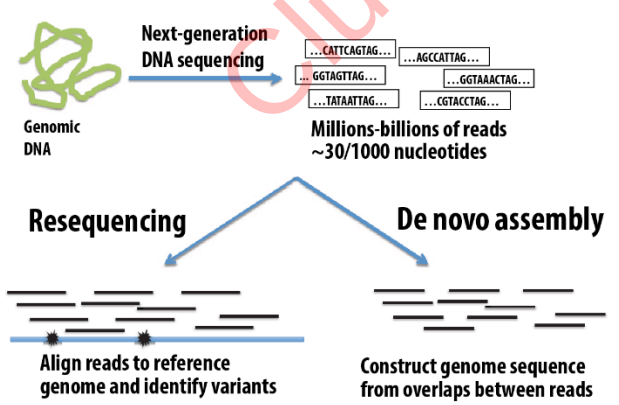
- These biased distributions of nucleotides can be used for prediction of splice junctions

Lecture 104

- **Genome Assembly**

Outline

- Introduction
- Vocabulary terms
- Overview
- Conclusions
- **Genome Assembly**

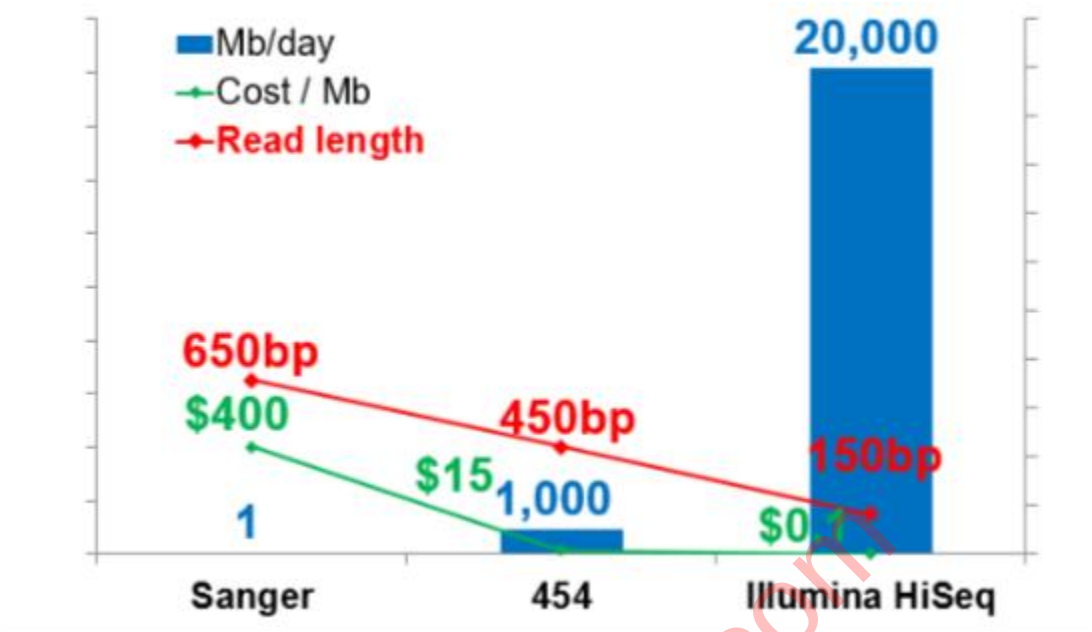


Genome Assembly

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** **---**
ATGCTANNNNNNNNNNNNNNNNNNNN**AGCTA---**

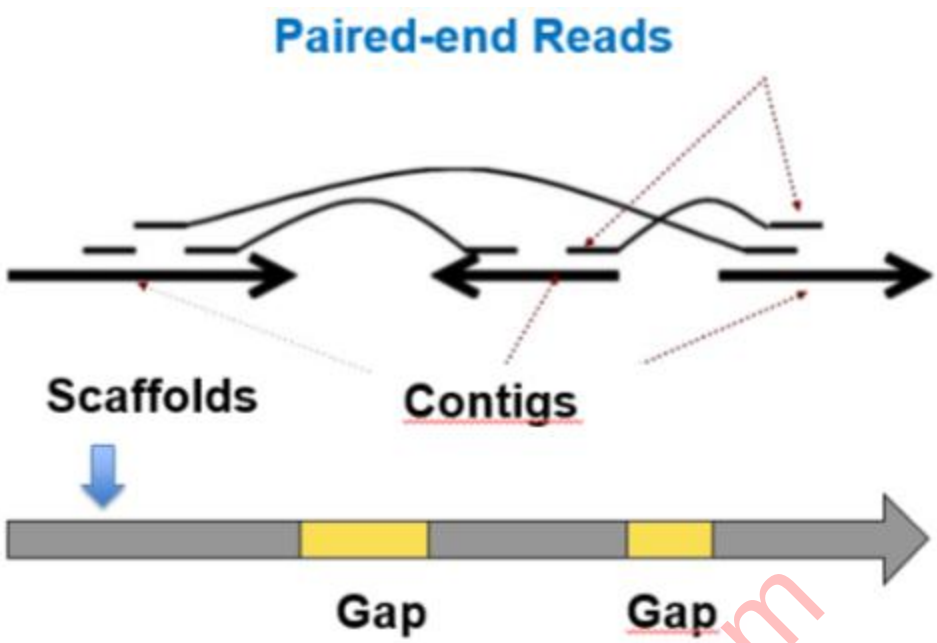
- **Genome Assembly**

Vocabulary

Captured Gap:

A gap within a scaffold

- The order and orientation of the contigs spanning the gap is known



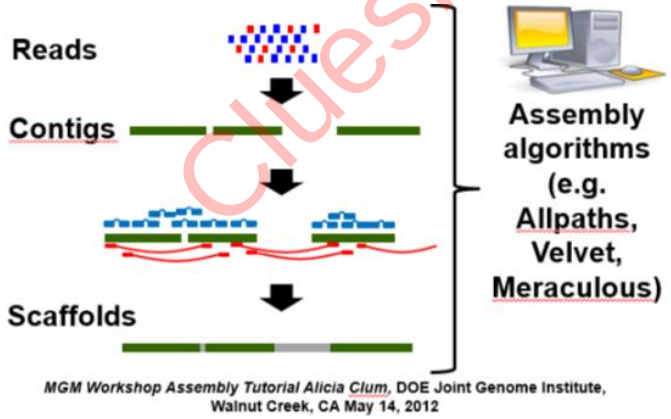
•

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



•

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

Lecture 106

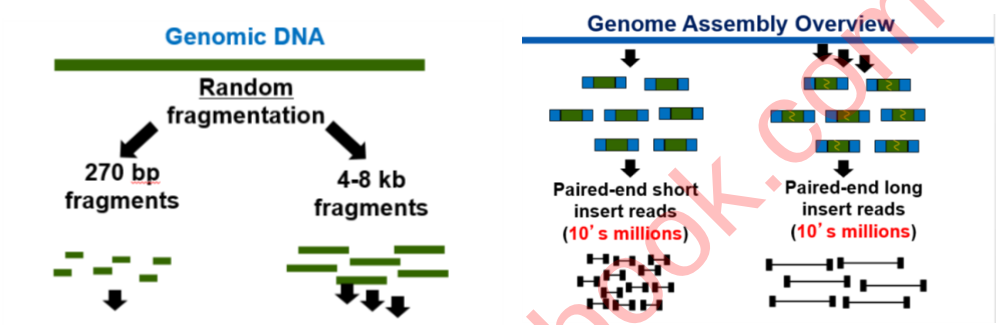
Genome Assembly Overview

Outline

- Sample preparation
- Sequencing
- Assembly
- Finishing
- Conclusions

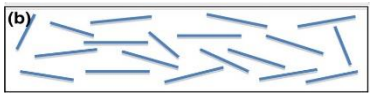
Sample preparation

DNA is collected from the biological sample followed by library preparations



Sequencing

The output from the sequencer consists of many billions short, unordered DNA fragments (strings of ATGCs) from random positions in genome

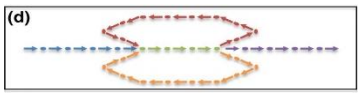


Assembly

The short fragments are compared with each other to discover how they overlap

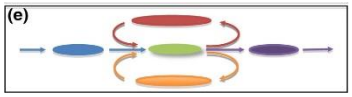
Assembly

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

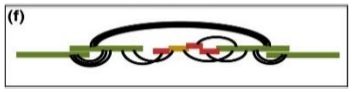


Refinement and simplification

The assembly graph is refined to correct errors and simplify into the initial set of contigs, shown as large ovals connected by edges



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

Conclusions

Genome assembly is performed as follows;

- Sample preparation
- Sequencing
- Assembly
- Refinement

Lecture 107

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

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
- This will affect:
  - How many reads will be required to attain sufficient coverage (typically **10x to 100x**)

  This will affect:
  - How many reads will be required to attain sufficient coverage (typically **10x to 100x**)

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

C = NL/ G

N: Total number of reads

L: Length of a Read

G: Genome size

Represented as **X**

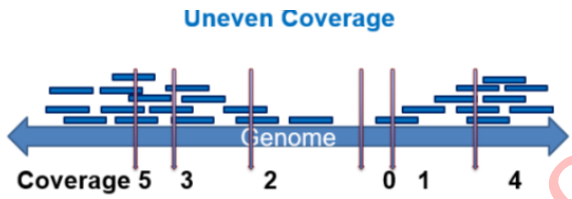
Example

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**

Uneven Coverage



Effect of genome size

- This will help decide which sequencing technology to use and what computational resources will be needed

Conclusions

- Genome size is an important factor in Genome Sequencing.
- This will help decide which sequencing technology and computational resources will be needed

Lecture 108

- **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

Repeat content

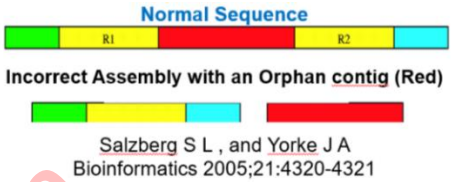
A major proportion of eukaryotic genome is made up of repeats.

Repeats are the most common source of Mis-assemblies

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**

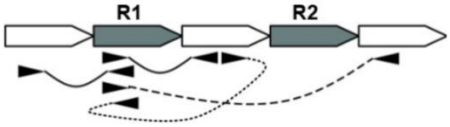


Double-barreled Shotgun Assembly

- DNA is randomly sheared into fragments (inserts)
- The ends of each insert are sequenced, resulting in two reads per fragment (double-barreled sequencing)
- The original DNA sequence is reconstructed from the reads (assembly step)

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 →→**



Conclusions

- Most of the eukaryotic genomes possess repeats
- Repeats pose special challenges to assembly
- Using mate pair libraries can help fix these issues

Lecture 109

- **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 genomes helps in;
- Can themselves be improved in some cases

Caution

- Can cause errors while architectures are different

Factors

Following factors may be important in planning an Assembly project;

- Size of Genome
- Repeat content
- Related species Assembled
- **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

- **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

Lecture 110

- **Greedy Graph Algorithm**

Outline

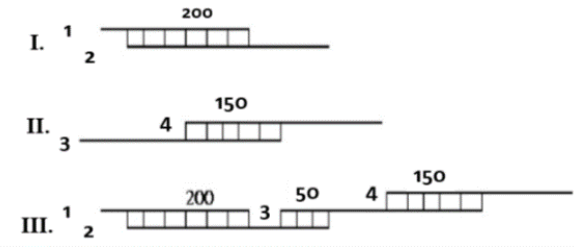
- Introduction
- Steps
- Advantages
- Limitations
- Greedy Graph based Assemblers
- Conclusion

Introduction

- Greedy Graph algorithms represent the simplest, most intuitive, solution to the assembly problem.

Steps

Greedy Graph algorithms works as follows:

- Compare all reads or contigs in a pairwise fashion to identify overlapping sequences
- Greedy Graph algorithms works as follows:
- Merge the sequences that overlap each other the best
- 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

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

Limitations

- Greedy assemblers can detect false overlaps and high-scoring ones that are resulted from repetitive sequences.
- Graph traversal using greedy approach may cause algorithm to become stuck in local maxima, which produces a suboptimal solution for the assembly problem
- The local maxima will increase the gaps between contigs in the assembly finishing process

Greedy Graph based Assemblers:

- SSAKE
- VCAKE
- SHARCGS

Conclusion:

- Greedy graph approach is simple, easy and fast

Lecture 111

- **ALLPATHS-LG**

Genome Assemblers

Variants of simple sequence alignment programs to piece together vast quantities of fragments generated by the sequencers

Introduction

It's a short read genome assembler from the Computational Research and Development group at the Broad Institute

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**

Difference between ALLPATHS-LG and Traditional Assemblers

- **ALLPATHS-LG assemblies are not necessarily linear, but instead are presented in the form of a graph**

Linearized Graph Assemblies

- **An assembly consisting of contigs and scaffolds with embedded ambiguity codes**
- **ALLPATHS-LG**

• 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

Computational Requirements

- **64 bit linux**
- **Memory requirements**
  - **About 160 bytes per genome base implying**

Advantages

- **Relatively fast runtime**
- **Can use long reads only for small genomes**

Conclusion

- **It's a short read genome assembler from the Computational Research and Development group at the Broad Institute**

References

- **http://www.broadinstitute.org/software/allpaths-lg/blog/**
- **Assembly Tutorial by Michael Schatz**
- **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**

Lecture 112

- **Lander Waterman Curve**

Outline

- Introduction
- Assumptions
- Lander Waterman Model
- Variables used in calculation
- Calculations
- Results
- Conclusion

Introduction

- Lander/Waterman suggested in their 1988 paper that the number of times a base is sequenced follows a Poisson distribution

Assumptions

There are two key assumptions they made;

- Reads will be randomly distributed in the genome

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 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

Variables used in Calculations

Variables used in calculation are;

G = haploid genome length in bp (base pairs)

L = sequenced read length in bp

Variables used in calculation are;

N = Number of reads sequenced

T = amount of overlap needed for detection in bp

Calculations

Coverage can be defined as 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

$$\mathbf{C} = \frac{\boldsymbol{LN}}{\boldsymbol{G}}$$

Gap Length

Let $\mathbf{e^{-C}}$ 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.

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
- 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

Lecture 113

- **Assembly Pipeline**

Strategy

Overall process can be divided into following major phases;

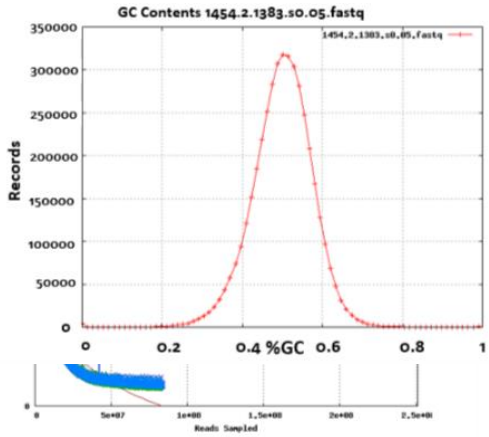
- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- Repeat resolution
- **Assembly Pipeline**

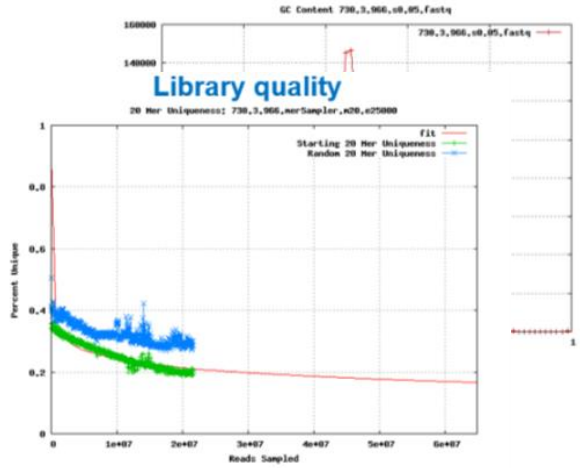
Quality screening

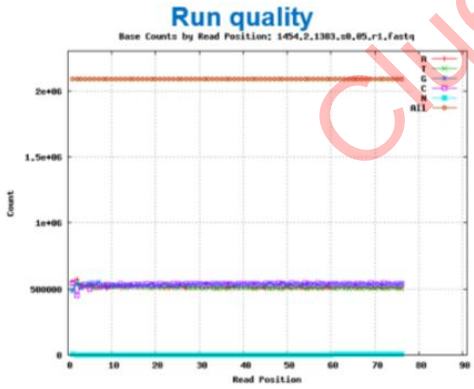
Screening is done while satisfying the following matrices;

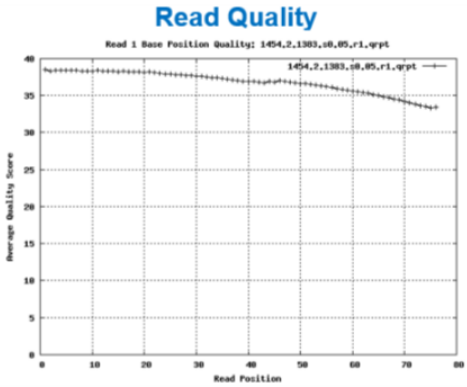
- Genome properties (GC)
- Library quality
- Sequencing Run Quality

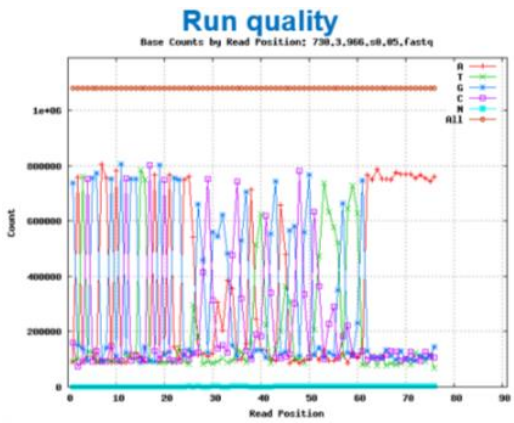
- Read Quality











Lecture 114

- **Assembly Pipeline**

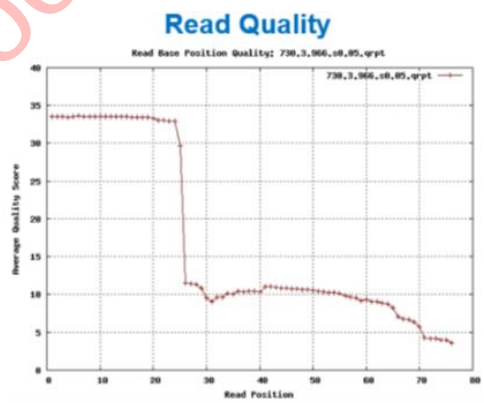
Strategy

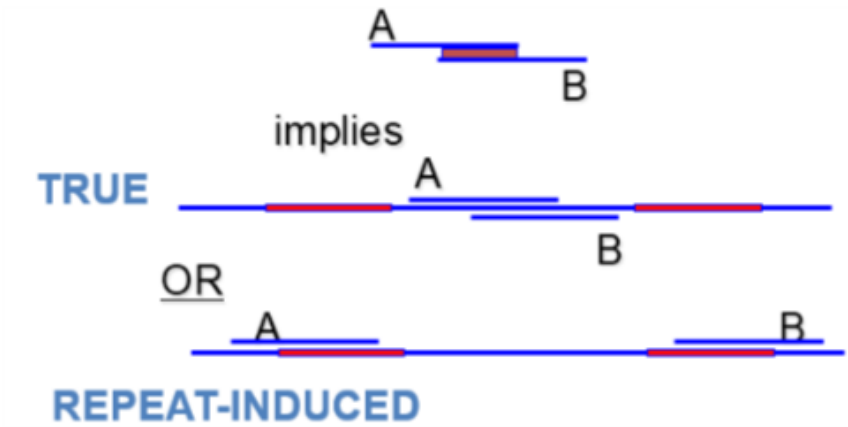
Overall process can be divided into following major phases;

- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- Repeat resolution

Overlapping

Find all overlaps ≥ 40bp allowing 6% mismatch.

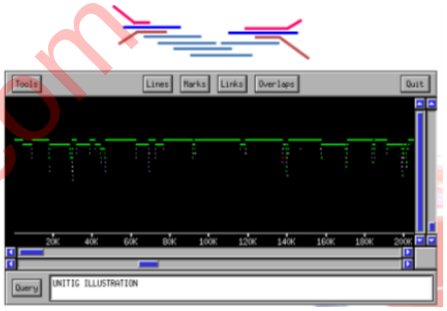




Strategy

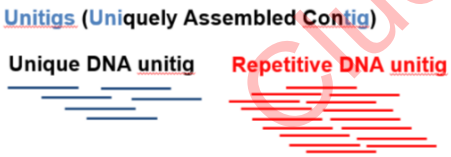
Overall process can be divided into following major phases;

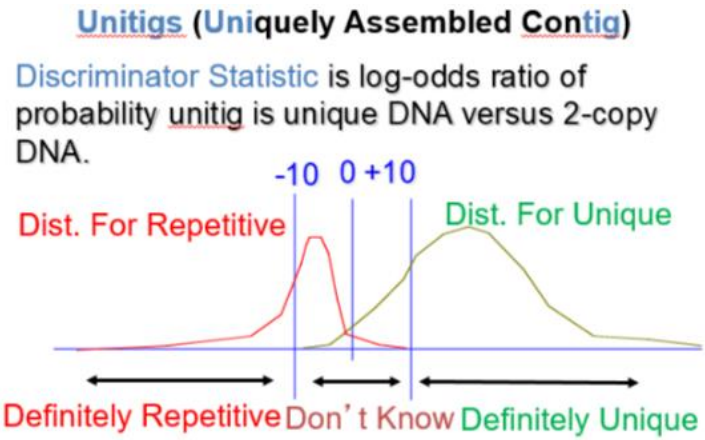
- Quality Screening
- Overlapping
- Unitiging
- Scaffolding
- Repeat resolution



Compute all overlap consistent sub-assemblies

Unitigs (Uniquely Assembled Contig)





Repeat Resolution

pre-assembly

find fragments that belong to repeats

- statistically (most existing assemblers)
- repeat database (*RepeatMasker*)

during assembly

- detect "tangles" indicative of repeats (Pevzner, Tang, Waterman 2001)
- **Repeat Resolution**
- **post-assembly:** find repetitive regions and potential mis-assemblies.
- *Reputer, RepeatMasker*
- "unhappy" mate-pairs (too close, too far, mis-oriented)

Lecture 115

- **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**

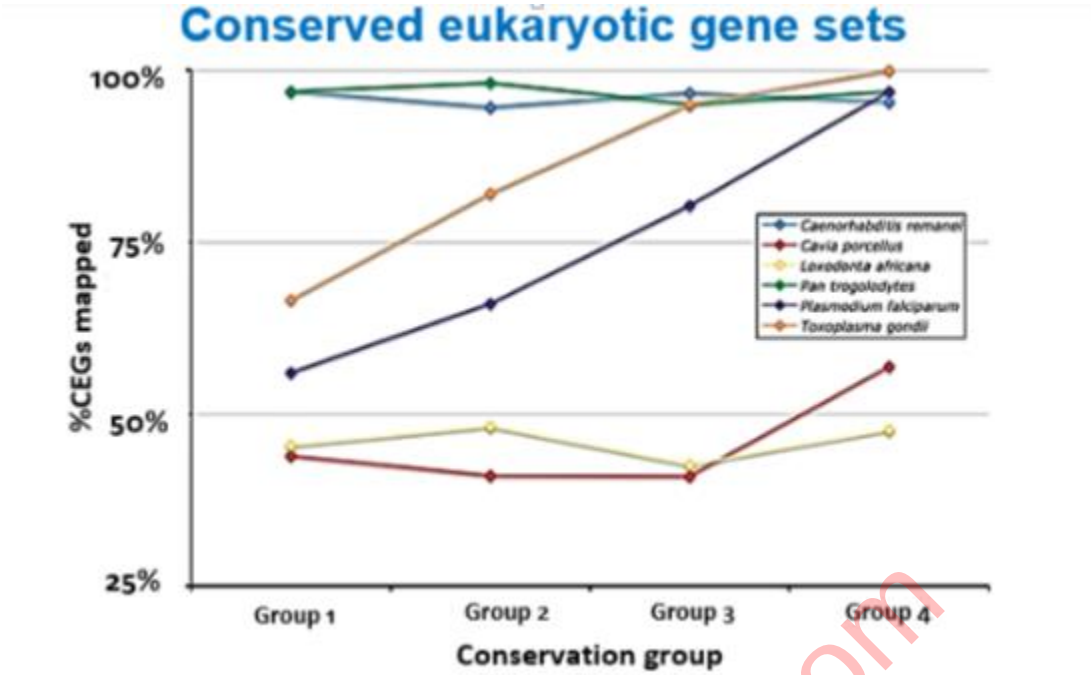
N50

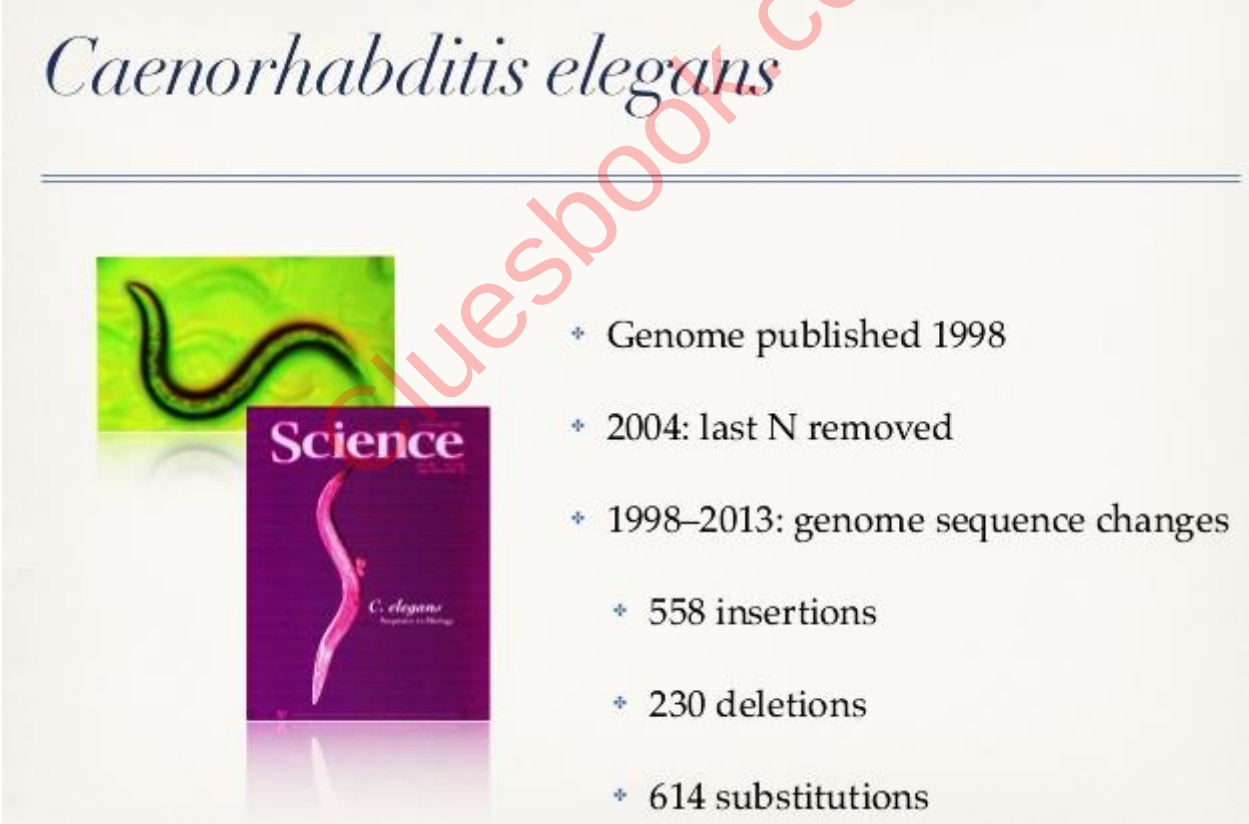
Most common measure of assembly quality;

- N50 = length of the shortest contig in a set making up 50% of the assembly length

- **Quality of Assembled Genome**

Conserved eukaryotic gene sets





- **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

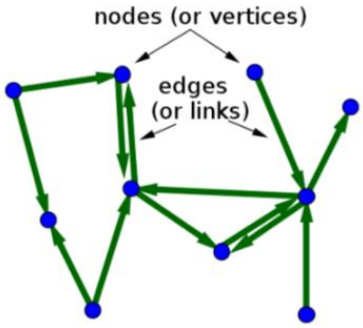
- Mayo/UIUC Summer Course in Computational Biology
- M. Schatz, S. Salzberg, K. Bradnam, K. Krampis, D. Zerbino, J. J. Cook, M. Pop, G. Sutton
- *Alicia Clum,* DOE Joint Genome Institute, Walnut Creek, CA
- Thank you!

Lecture 116

- **Graphs**

Outline

- Introduction
- Types of graphs
- Graph Algorithms for Assembly
- Conclusion

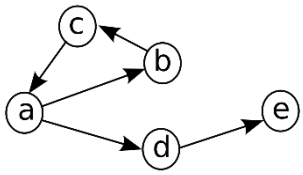


Introduction

- 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**

Types of Graphs

- Directed Graphs
- Undirected Graphs

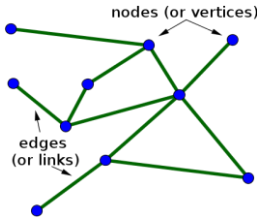


Directed Graphs

- If the edges may only be traversed in one direction, the graph is known as directed graph

Undirected Graphs

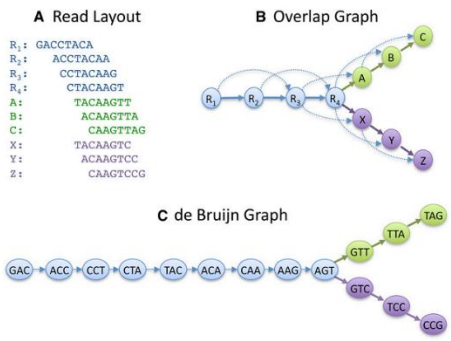
- Undirected graph is set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional



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**

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

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.

Eulerian Circuit

- An Eulerian path which begins and ends on the same vertex.
- It starts and ends on the same vertex.

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

Lecture 117

- **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 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 Graph

- Each node corresponds to reads
- An edge connects two nodes if an overlap was identified between the corresponding reads

Layout stage

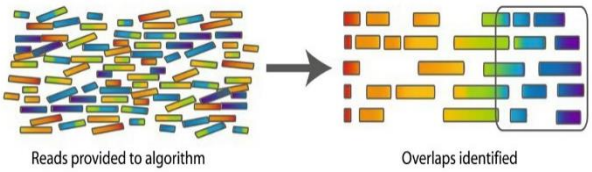
- During this stage, overlap graph is analyzed in order to identify paths through the graph that correspond to the genome being assembled

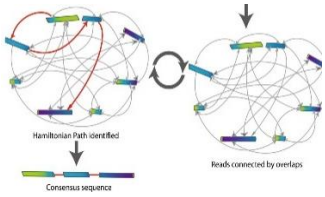
Layout stage

- The goal is to find single shortest Hamiltonian path that visits each node in graph exactly once

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





Advantages

- Suitable for large size genomes
- Works for both short and long reads
- OLC approach is capable of handling NGS data
- Generate correct order of contigs
- OLC recognizes repeats but may collapse then into one
  - ABBBCBBD => ABCBD

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

Limitations

- No efficient algorithm to find the Hamiltonian path

OLC Based Assemblers

- Arachne
- Celera Assembler (CABOG)
- Newbler
- Minimus
- Edena
- CAP
- PCAP

Conclusion

- The OLC strategy is arguably the most successful assembly strategy in practical setting.

Lecture 118

De Bruijn Graph

Outline

Introduction

Steps

Problems in Graph Construction

Advantages

Limitations

DBG based Assemblers

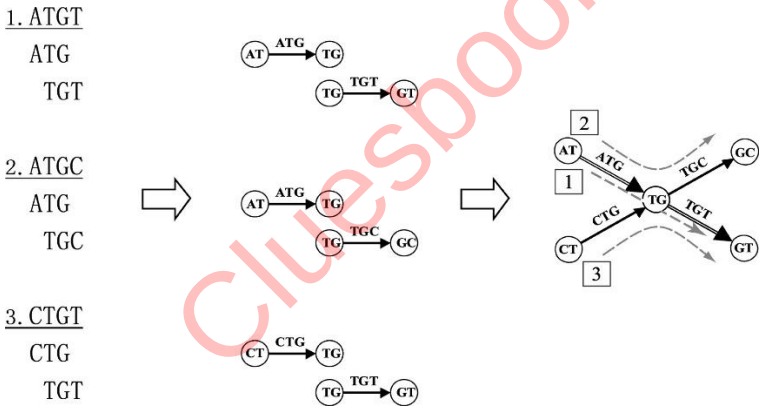
Conclusion

Introduction

This approach for genome assembly is most widely applied to short reads from the Solexa and SOLiD Platforms

Steps

- Generate overlapping substrings of length k from the reads
- These substrings are called **Kmers**
- Generate De Bruijn graph
- Nodes represent all possible fixed length strings or **Kmers**
- Edges represent overlap of k-1 nucleotides
- Two nodes are linked with an edge if they share a k-1 mer
- Connect one Kmer to another if the two k-mers completely overlap except for one nucleotide at each end
- Third, look for a Euler cycle, representing a candidate genome because it visits every edge of graph exactly once



DBG Based Assemblers

- EULER(-SR)
- Velvet
- ALLPATH(S2)
- ABySS
- SOAPdenovo

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

Conclusion

- This approach for genome assembly is most widely applied to short reads from the Solexa and SOLiD Platforms.

Lecture 119

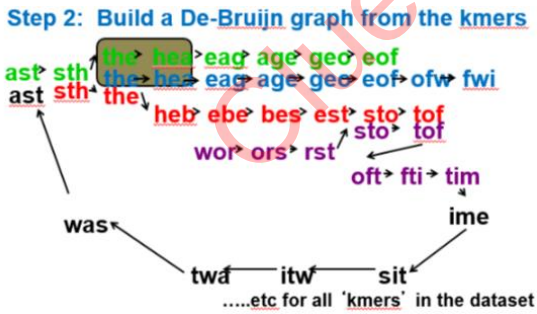
- **De Bruijn Graph Example**

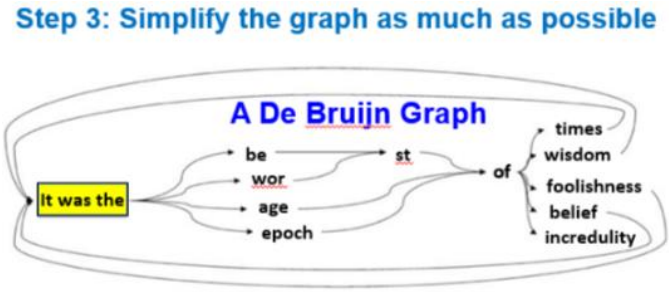
Deconstruction

- The first step of the De Bruijn assembler is to deconstruct the sequencing reads into its constitutive "kmers".
- The first step of the De Bruijn assembler is to deconstruct the sequencing reads into its constitutive "kmers".
- "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,.... "

Traditional Method

- All-vs-all assemblers fail due to immense computational
- A **million (10^6)** reads requires a **trillion (10^{12})** pairwise alignments



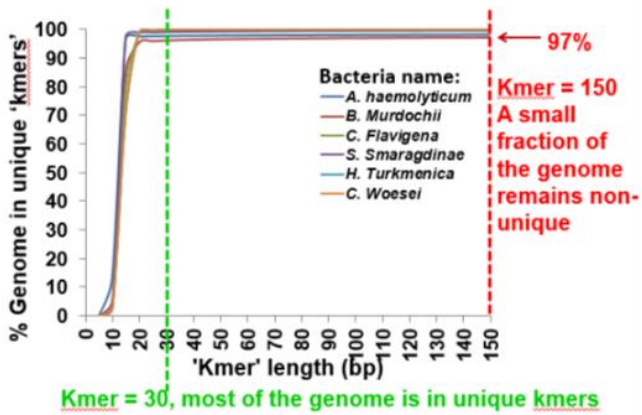


The final assembly (k=3)

wor times itwasthe foolishness st wisdom incredulity age epoch be of belief

A better assembly (k=20)

Itwasthebestoftimesitwastheworstoftimesitwasthe ageofwisdomitwastheageoffoolis…



Conclusions

- Using an apropriate size Kmer and storing information stored in millions of reads is an important aspact of de Bruijn graph Algorithm

Lecture 120

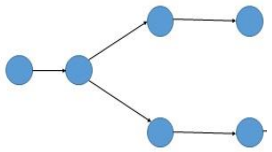
- **Assembly Issues**

Problems in Graph Construction

- Spurs
- Bubbles
- Converging and diverging paths
- Cycles

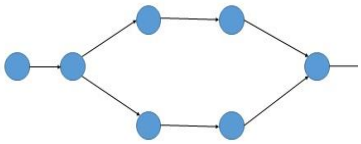
Spurs

- Short dead-end branches (divergences) of the main path
- Possible causes include sequencing errors toward one end of a read, and low coverage
-



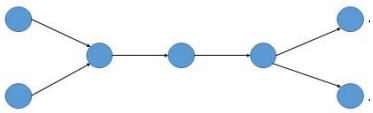
Bubbles

- 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



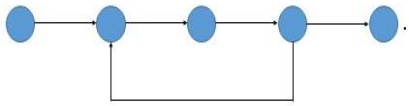
Converging and Diverging Paths

- 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



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
- Nodes that are unambiguously connected by an edge are merged together

Limitations

- Loss of information due to splitting into k mers
- Exponential number of Eulerian paths because of number of repeats in genome
- Assembler has to find most probable path
- Increase in memory consumption and runtime
- **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

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.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

- Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, *10*(4), 354-366.
- http://gcat.davidson.edu/phast/debruijn.html

Lecture 121

Proteomics Introduction

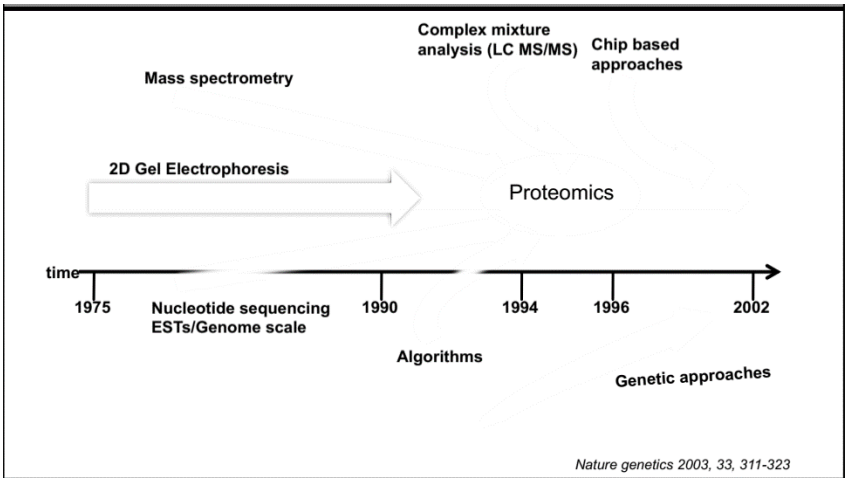
Proteomics is the study of all the proteins found in an organism.

- Introduction to Proteomic
- The terms “proteomics” and “proteome” were coined by Marc Wilkins and colleagues in 1994 and mirror the terms “genomics” and “genome.

- 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”.

Introduction of Protein

- Protein Extraction
- Protein Separation
- Protein Identification
- Protein Characterization
- Proteomics
- Abundance
    - Identity
    - Quantity
- Function
    - Expression
    - Asssay



- Protein chemistry to proteomics
- Diversification of proteomics technologies
- Advent of novel technology platforms

Lecture 122

- **Comparative Proteomics**

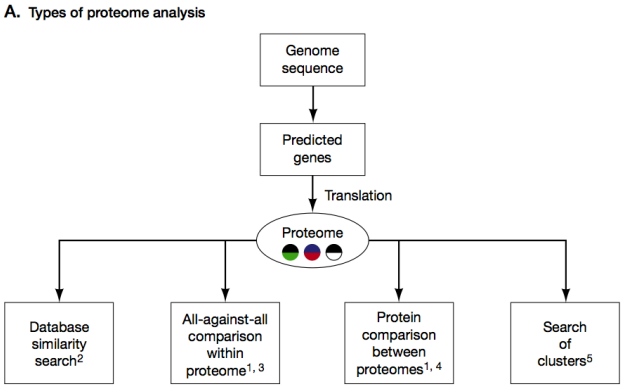
Introduction

- For a new Genome, predicted genes are translated
- The collection of protein sequences encoded by the genome makes up the proteome of that individual
- **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

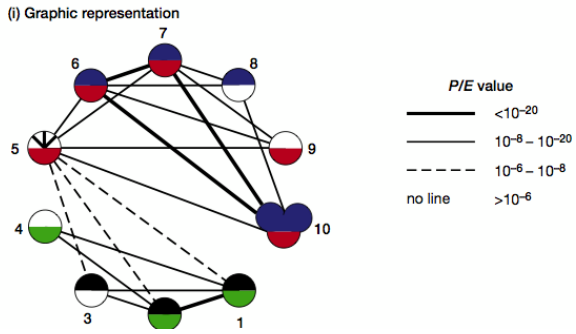
Cluster Analysis

- 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



Clustering by subgraph

- Each sequence is a vertex
- Significant alignment score is an edge
- Trimming by removing weak edges (High P/E)



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

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

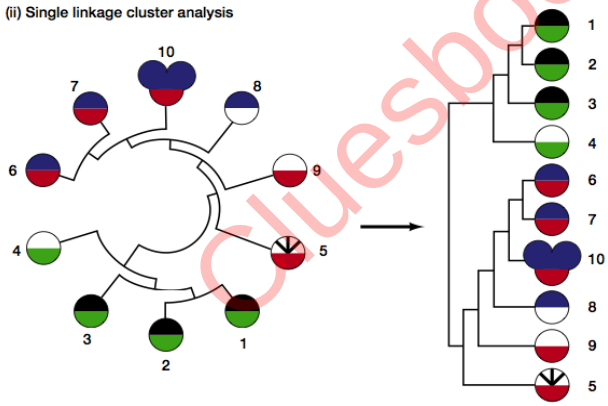


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

- **Comparative Proteomics**

Core Proteome

- All-against-all comparison provides an indication of Gene/Protein families
- Unique set of proteins is core proteome

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.

Conclusions

- Genome is translated into proteome
- Self comparison of proteome yields gene families and duplications
- Unique set of proteins is core proteome

Lecture 123

- **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**

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**
- **Between-Proteome comparisons**

Conclusions

- Proteome comparison helps finding orthologs, gene families and protein domains
- Domain comparison reveals their biological roles

Lecture 124

- **Genetics and Genomics**

Proteomics, Types and Techniques

- **Proteomics, Types and Techniques**

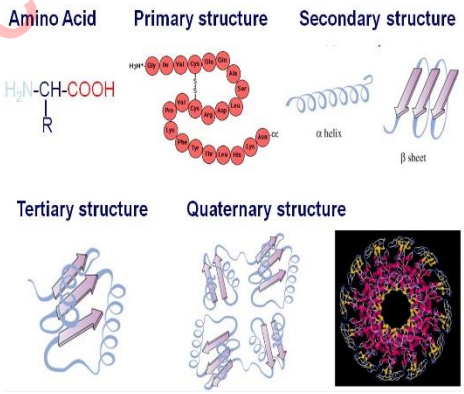
Study of Proteomics

- Proteomics is the study of all the proteins found in an organism
- Different Types

GENOME	PROTEOME
▪ 4 nucleotides	▪ 20 amino acids
▪ Double helix	▪ Each protein has unique 3D shape
▪ Same in all cells	▪ Differs with cell type

Structural Proteomics

- Structural proteomics deals with mapping out the 3-D structure and nature of protein and/or proteins complexes

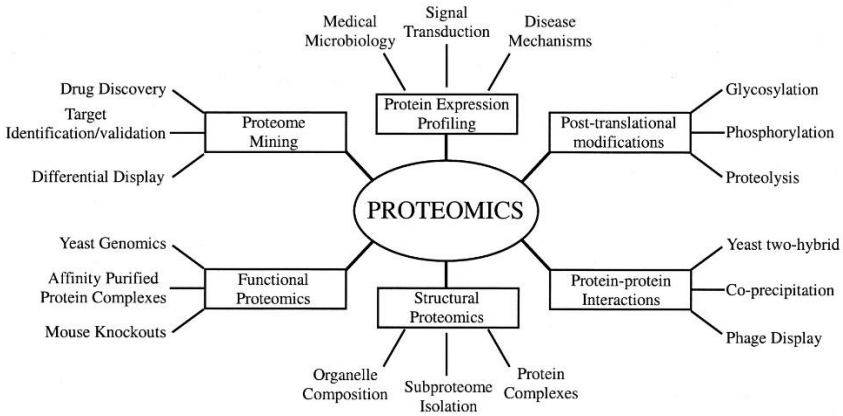


Functional Proteomics

- Posttranslational modifications
  - Protein-protein, protein-ligand interactions
  - Sequence-structure-function relationships

Expression Proteomics

- Expression proteomics refers to the quantitative study of protein expression between samples differing by some variable.
  - Analysis of differential protein expression



Study of Proteomics

- Protein profiling
- Predicting protein structure
- Protein networks

Protein Profiling

- Determination of the proteins that make up a given proteome
- Proteomes vary by cell type
- Proteomes vary by stage of cell development
- Some proteins abundant, others very rare

Protein Profiling Techniques

- Two-dimensional gel electrophoresis
- Chemical protein sequencing
- Protein sequencing by mass spectrometry

Proteins Structure Determination Techniques

- X-ray crystallography reliable but slow, not all protein crystallize
- Computer structure-prediction programs not reliable for all proteins

Proteomics - Conclusion

- Proteomics is the study of all the proteins found in an organism
- Different Types

Lecture 125

Human Proteome - Characteristics

Proteomics

- Proteomics is the study of proteins that are generated from the genetic code of an organism.
- 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 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

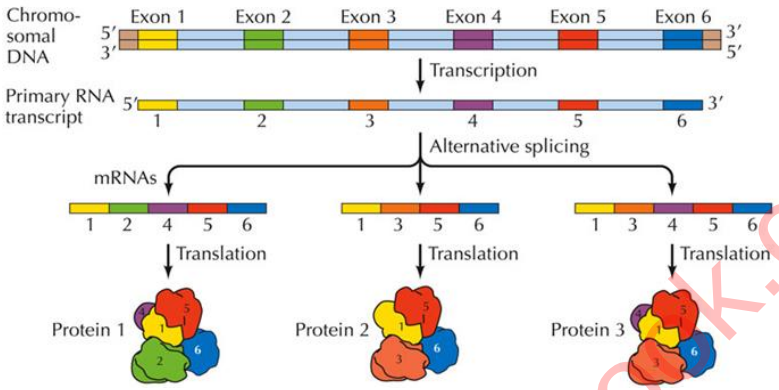
Human Proteome

- Less than 2 % of human diseases result from a single gene defect
- 98 % complex diseases like cancer are reflected in a modified protein network

Multiple Functions

- One protein or peptide may have multiple functions.
- Complex regulation of protein function.
- Proteins modifications

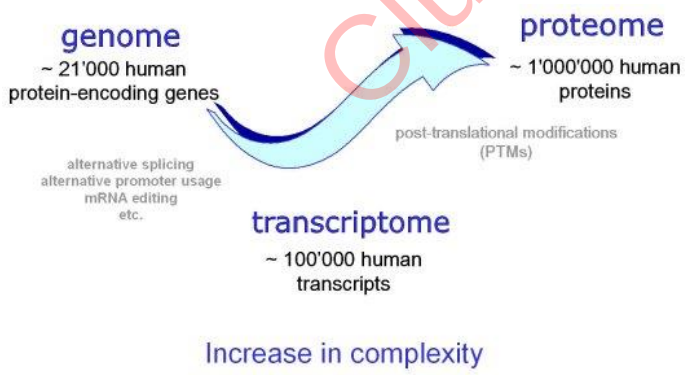
Proteomics – More Number of Proteins than Genes



More Complex Proteome than Invertebrates

- Proteome more complex than those of invertebrates.

Human Genome and Proteome



Proteome Complexity increases from Yeast to Humans

- Complexity of proteome increase from yeast to humans
    - More genes
    - Genes shuffling increases

Proteome – More Chemical Modifications

- Alternative RNA Splicing –
- Humans exhibit significantly more chemical modification of proteins

Human Genome and Human Proteome

- Human Genes ~ 21,000 - 23,000
- Human Transcripts ~ 100,000
- Human Proteins ~ 250,000- 1,000,000

Lecture 126

why Proteomics

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

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

- 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

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.

<u>Advantages of study of proteomics</u>

- 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.

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
- 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.....

Lecture 127

Biological Importance of proteins

- Cellular Function
- Structural importance

- Enzymes
- Harmones
- Transport protein
- Messenger protein
- Antibodies
- Plasma protein
- Protein in diet

Pharmaceutical Importance of proteins

- Proteins as pharmaceuticals
- Proteins applications
- Whey proteins health effects
- Iron chelate Protein
- Zinc chelate Protein
- Tumor markers

Cellular Function

- Structural support
- Biochemical reactions of cell
- Role of proteins in plasma membrane

Structural Support

- Microfilaments:
- Distribution of Organelles
- Formation of mitotic Spindle
- Actin Filaments : Cell movements
- Intermediate Filaments
- Architectural support inside cells

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
  R

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

Enzymes

There are three types of enzymes

- Metabolic enzymes
- Digestive enzymes
- Food enzymes

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

Antibodies

- These are protein components of immune
  system

- Have main function to bind antigens.

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.

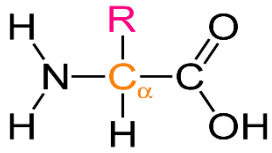
Proteins in Diet

Sources of Proteins

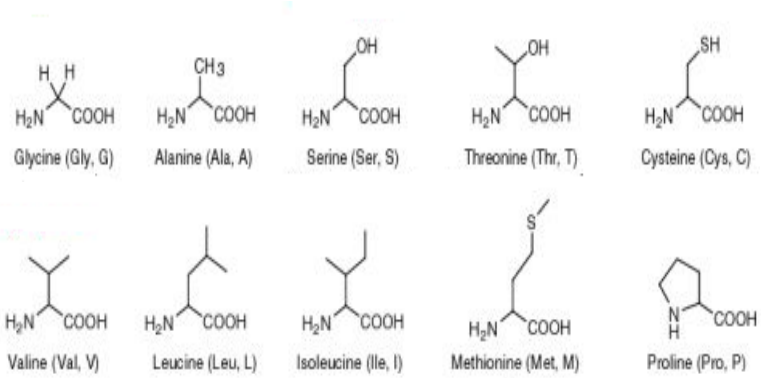
- Meat
- Beef
- Poultry
- Fish
- Egg
- Beans
- Milk

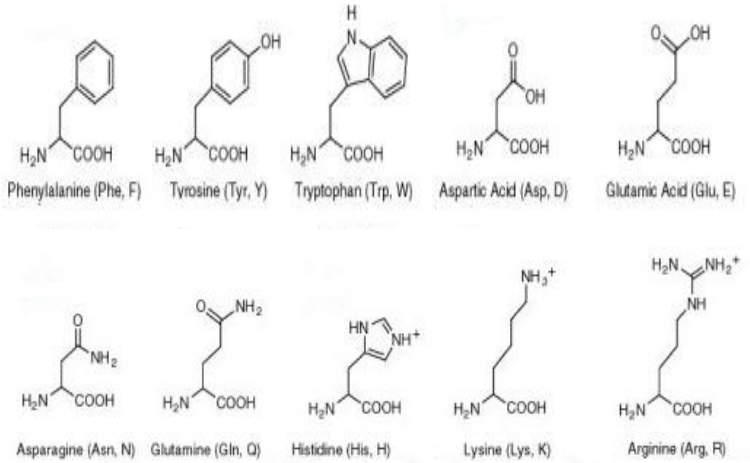
Lecture 128

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.

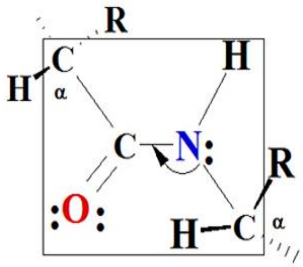


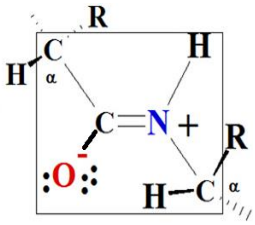


- In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond.
- Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides.

Lecture 129

Primary structure of proteins



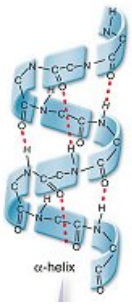


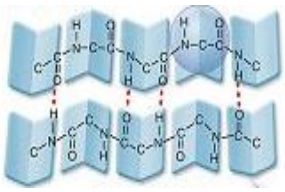
- Primary structure or covalent structure of protein refers to the amino acid sequence of its polypeptide chain.
- Each type of protein has a unique amino acid sequence.
- Linus Pauling and Robert Corey carefully analyzed the peptide bond.
- Their findings laid the foundation for our present understanding of protein structure.
- They demonstrated that the peptide C - N bond is somewhat shorter than the C - N bond in a simple amine.
- 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.
- 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.
- 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.

Lecture 130

Secondary structure of proteins

- Secondary structure of proteins refers to the local conformation of some part of a polypeptide.
- A few types of secondary structures are particularly stable and occur widely in proteins.
- The most prominent are:-
    - α-helix
    - β- conformations.

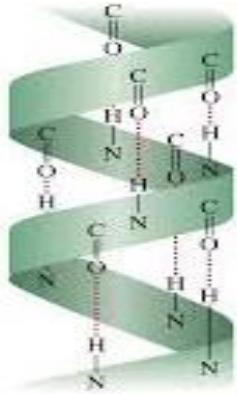




Lecture 131

α- Helix

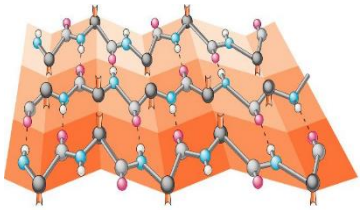
- The simplest arrangement which a polypeptide chain could assume with its rigid peptide bonds is a helical structure, which Pauling and Corey called the **α-helix**.
- The helical twist of the α-helix found in all proteins is right-handed.
- The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis.
- 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.
- About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure.



Lecture 132

β- Pleated Sheets

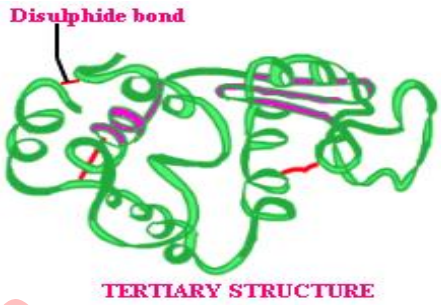
- Pauling and Corey predicted a second type of secondary structure which they called **β-sheets.**
- This is a more extended conformation of polypeptide chains.
- The backbone of the polypeptide chain is extended into a zigzag structure.
- The zigzag polypeptide chains are arranged side by side to form a structure resembling a series of pleats.
- The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions.
- Hydrogen bonds are formed between adjacent segments of polypeptide chain.
- The adjacent polypeptide chains in a sheet can be either parallel or antiparallel.



Lecture 133

Tertiary Structure of Proteins

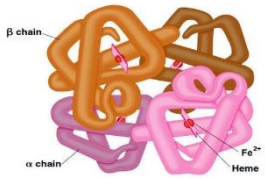
- The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's **tertiary structure.**



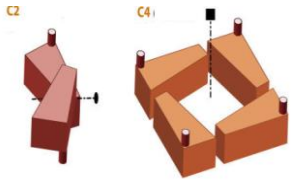
- It includes longer-range aspects of amino acid sequence.
- Amino acids that are far apart in the polypeptide chain may interact within the completely folded structure of a protein.
- Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak interactions (and sometimes by covalent bonds) between the segments.
- Large polypeptide chains usually fold into two or more globular clusters known as **domains,** which often give these proteins a bi- or multilobal appearance.

Lecture 134

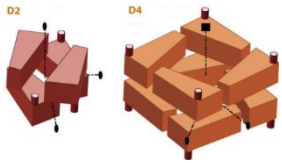
Quaternary Structure of Proteins



- 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.
- A multi-subunit protein is also referred to as a multimer.

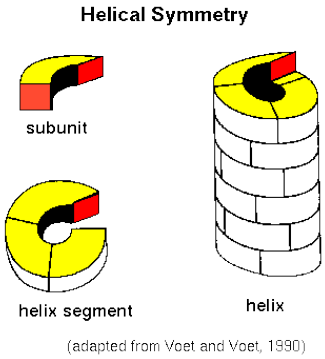


- A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer.
- Identical subunits of multimeric proteins are generally arranged in a symmetric patterns.



- Oligomers can have either rotational symmetry or helical symmetry.
- There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis.

- A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present.
- More complex rotational symmetries include icosahedral symmetry.
- An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces.
- The other major type of symmetry found in oligomers is helical symmetry.



Lecture 135

Life and Death of a Protein

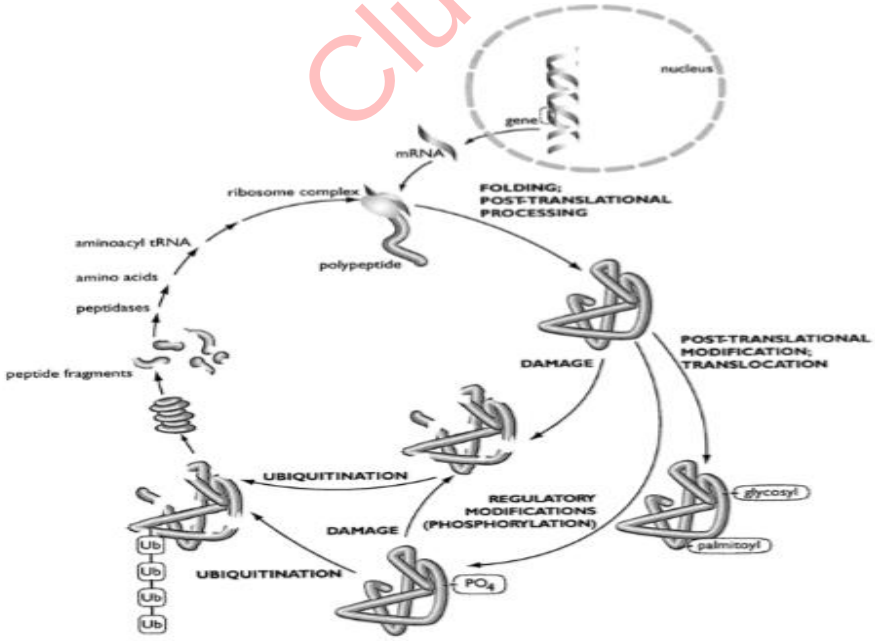
Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes.

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.

Life Cycle of the Cell



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.

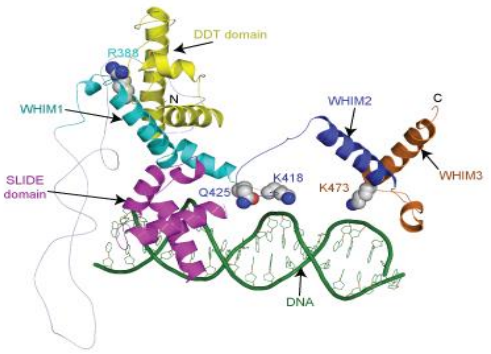
- The activities of many proteins are then controlled by posttranslational modifications.
- The most prominent and best-understood of these is phosphorylation of serine, threonine, or tyrosine residues.
- Phosphorylation may activate or inactivate enzymes, alter proteinprotein interactions and associations, 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.
- 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.

Lecture 136

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.



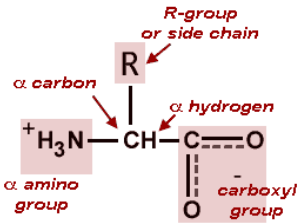
- In other cases, some key amino acids occur in a reproducible relationship to each other in a sequence, even though various substitutions in other amino acids occur.
- Longer amino acid sequences often form domains, which confer specific properties or functions on a protein.
- Some domain structures 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.

Lecture 136

- **Genetics and Genomics**

Type of Proteins/ Families

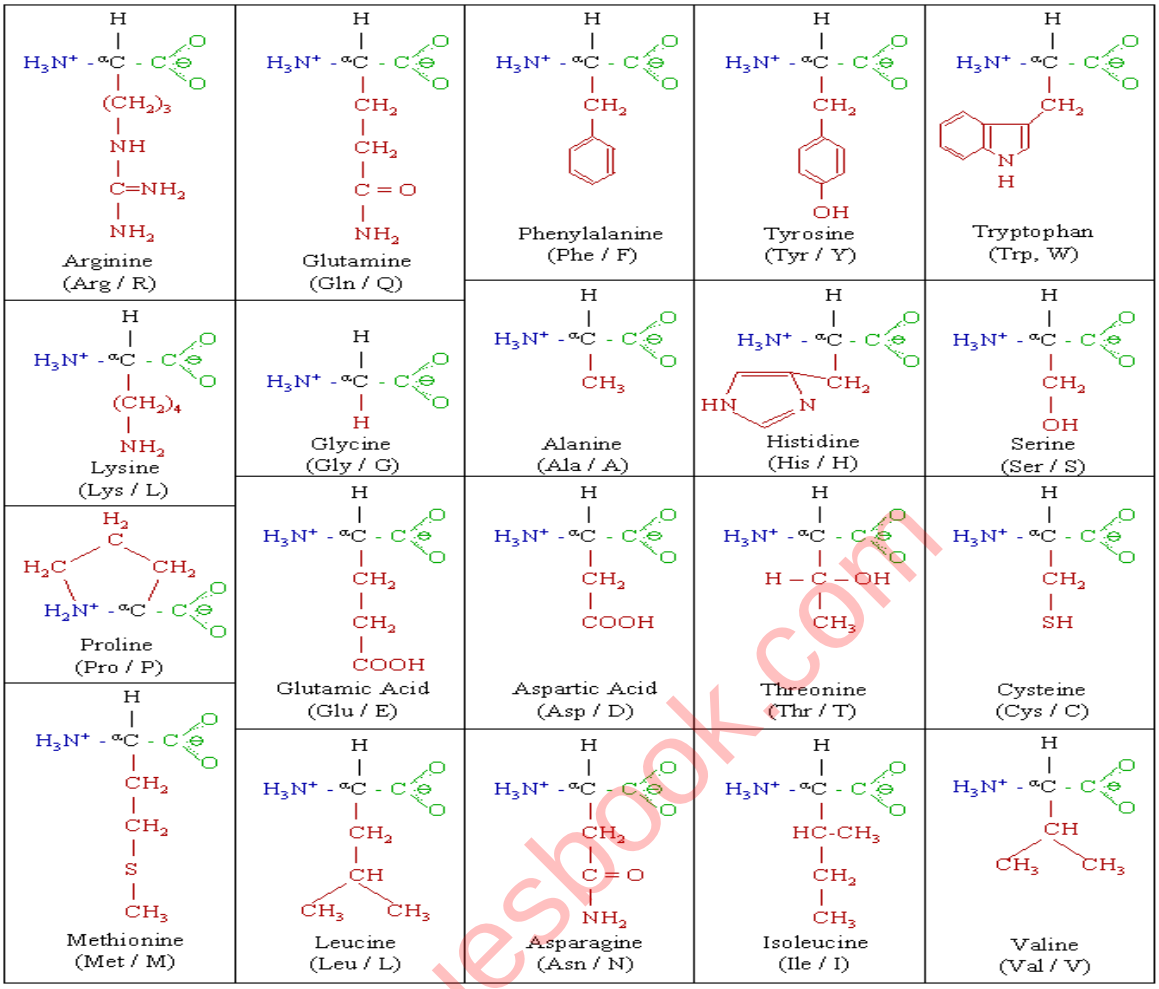
- **Type of Proteins/Families**



Protein Structure

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

- **Amino Acids: Building Blocks of Proteins**



Proteins

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

Fibrous Proteins – Cytoskeletal Proteins

- Actin
- Coronin
- Dystrophin
- Keratin

- Tubulin
- Collagen
- Elastin
- Fibronectin

Globular Proteins - Major Types

- Plasma proteins
- Hemoproteins
- Cell adhesion
- Transmembrane transport proteins
- Hormones and growth factors
- Receptors
- DNA-binding protein
- Immune system proteins
- Nutrient storage/transport
- Chaperone proteins
- Enzymes

Complexes with multiple components including proteins

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

Types of Proteins - Conclusion

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

Lecture 138

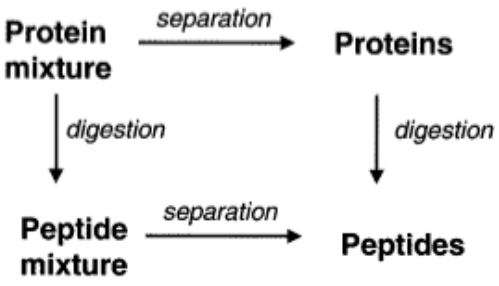
Analytical Protein and Peptide Separations

- 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 components of the mixture.



Lecture 139

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

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

Reductants

- Reductants e.g. e.g., dithiothreitol [DTT], mercaptoethanol, thiourea are used.
- These reduce disulfide bonds or prevent protein oxidation

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

Enzymes

- Enzymes like DNAse and RNAse are used.
- Enzymes digest contaminating nucleic acids, carbohydrates, and lipids present in the soup.

Lecture 140

Analytical Protein and Peptide Separations

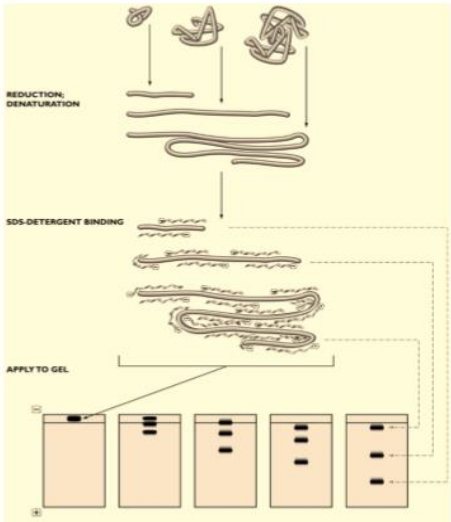
- ***Protein Separations Before Digestion***
- The three principal separation approaches used with intact proteins are;
  1. 1D-SDS-PAGE
  2. 2D-SDS-PAGE
  3. preparative isoelectric focusing (IEF).
- The protein mixture may be separated into a relatively small number of fractions as in 1D-SDS-PAGE and 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.

Lecture 141

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.

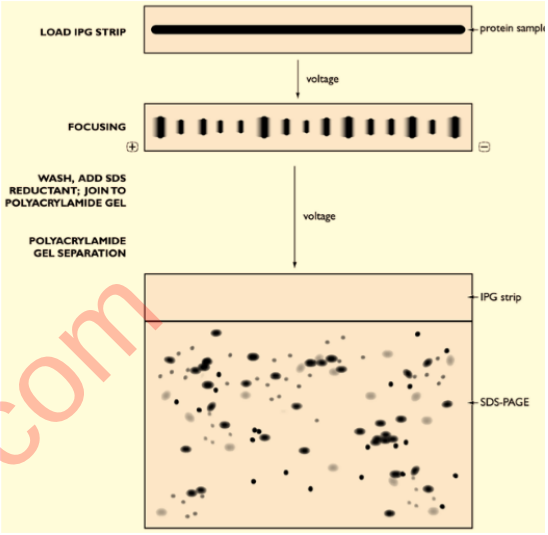


Extent of cross-linking in 1D-SDS PAGE

- 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

Lecture 142.

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

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.

Lecture 143

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

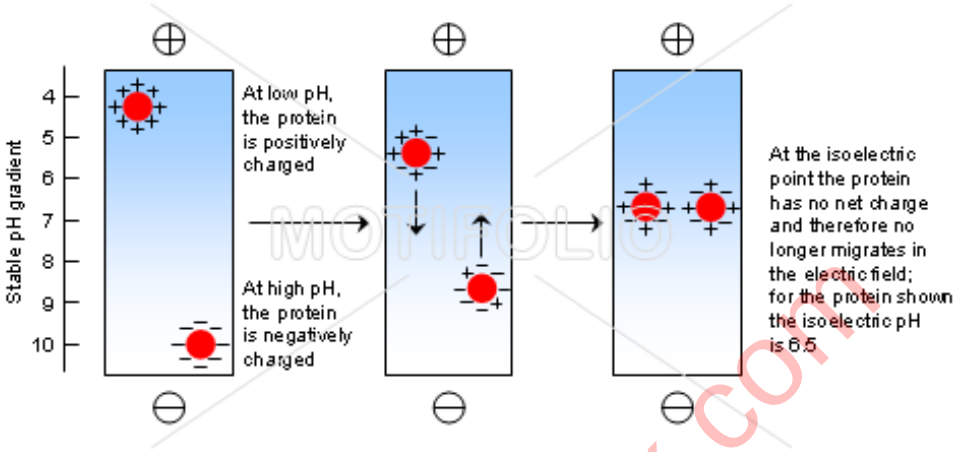
- A third problem with 2D-SDS-PAGE is the relatively small dynamic range of protein staining as a detection technique. Spot densities reflect about a 100-fold range of protein concentrations, at best. This means that staining of 2D-gels allows the visualization of abundant proteins, whereas less abundant proteins frequently cannot be detected.

Lecture 144

- **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 agentscan be used to maintain solubility of hydrophobic proteins.

Lecture 145

High-Performance Liquid Chromatography

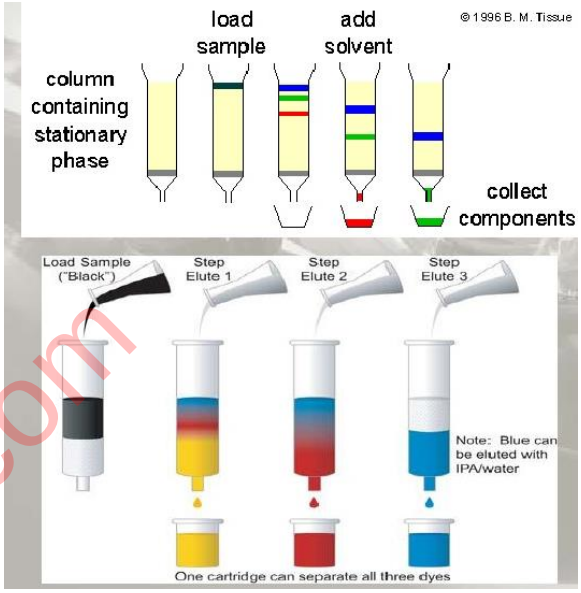
Introduction

- 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



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.

Lecture 146

Protein Separations After Digestion

Introduction

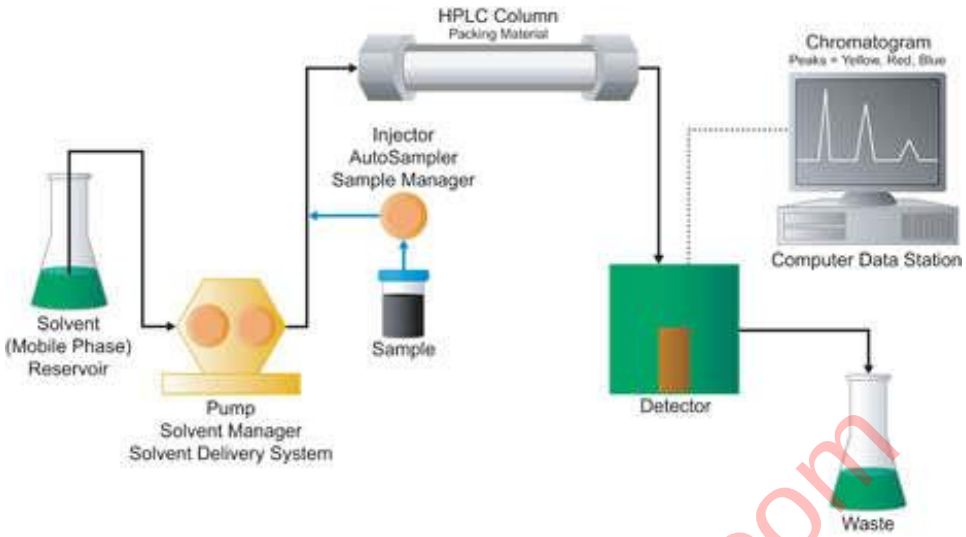
- In this approach, the proteins in the sample are first digested into a mixture of peptides, then the peptides are separated prior to analysis.
- The extreme application of this approach would be to digest a complete cell or tissue extract to peptides and then perform MS analysis on the mixture.
- Indeed, 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.

Lecture 147

Tandem LC Approaches for Peptide Analysis

Introduction

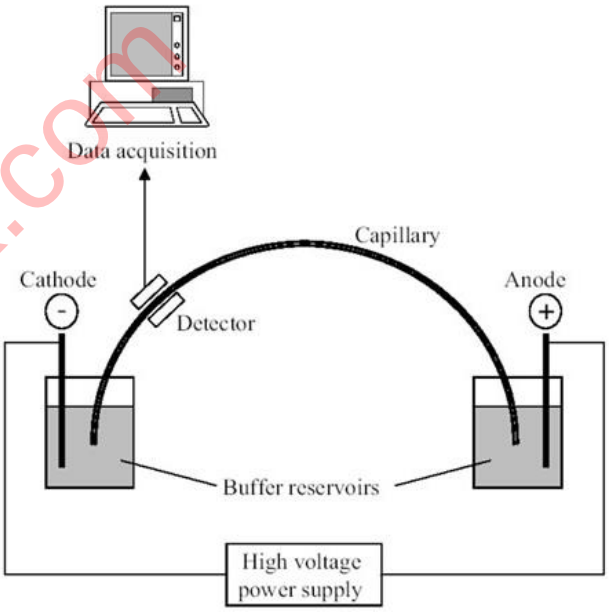
- Certainly the most widely used approach to analysis of peptide mixtures is HPLC.
- The diversity of stationary phases and separation modes gives HPLC considerable resolving power.
- The use of combined separation modes in series is referred to as “tandem HPLC.”
- The idea behind tandem LC is that the 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

Lecture 148

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.