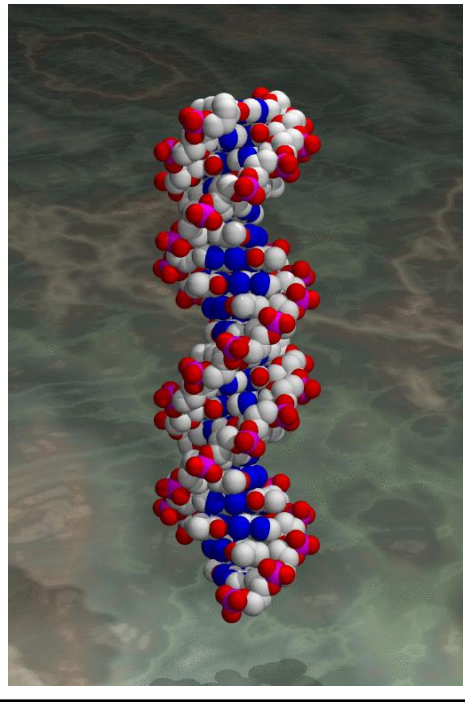
GENOMICS

Course: Molecular Biology (02022312)

Instructor: Dr. M A Srour

Textbook:
Watson J, Baker TA, Bell SP, Gann A, Levine M, Losick R (2008). Molecular Biology of the Gene, 6th ed. , Chap 20 pp. 703-31; chap 21 pp.739-82.
Malacinski GM (2003). Essentials of Molecular Biology, 4th ed. Chap 13, pp. 284-301

Lec # 13 Wed 11.04.2012



Genomics & Bioinformatics : definitions

- Genome: complete set of genetic information in a specific organism
- Genomics: The study of the structure and function of whole genomes> *structural and functional genomics*
- Structural genomics: the study of the sequences of genomes
- Functional genomics: the study of the pattern of genome-wide gene expression at various times or under various conditions
- Bioinformatics: study of the meaning (interpretation) of information contained in an organism's genome

Transcriptomics & Proteomics: definitions

- Transcriptome: the sum of all the different transcripts an organism can make in its lifetime
- Transcriptomics: the global study of an organism's transcripts

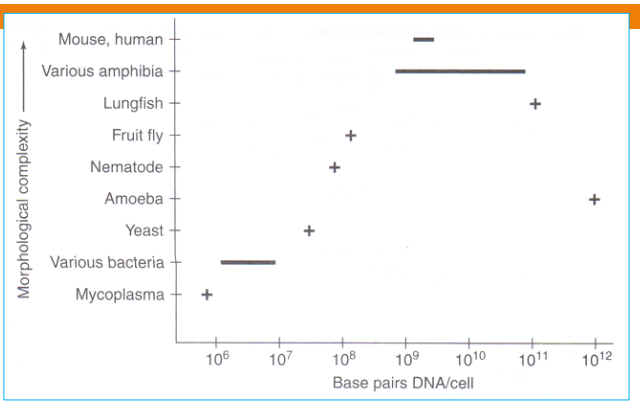
- Proteome: the complete set of the (functional) proteins in a cell/tissue.
- Proteomics: study that emphasizes the integration of structural, functional and quantitative relationship between proteins in a cell/tissue,

Genomics: the use of DNA as a starting point for discovery

- Pre-genomic era: protein (as starting point) >>>>> gene
- Genomic era: DNA (as starting point) >>>>> protein

- Genome size versus complexity:
- C-value paradox: lack of correlation between DNA content in an organism's genome and the size and level of complexity of the organism
- Genome size varies widely among eukaryotes and does not relate to morphological complexity (C-value paradox)
- Genome size and gene number: organisms with very different genome sizes may have similar numbers of genes

Genome size versus biological complexity



Biological complexity depends on gene number, regulation of gene expression, protein-interactions.

Which genomes should be sequenced?

- Priorities for sequencing whole genomes:
  - Ease of sequencing
  - Genome size
  - Breadth of interest of genome for scientific community
  - Unique feature of target organism for testing hypotheses
  - General relevance to understanding phylogenetic relationships of organisms

Sequencing genomes

- What information can be obtained from sequenced genomes?
  - Locate exactly coding regions of genes, defined by ORF
  - Spatial relationship among genes and distance between genes
  - Genome size and organism's complexity
- Genome of phage ΦX174 was the first to be sequenced and showed overlapping genes!

Milestones in genomic sequencing

Genome (Importance)	Size	Year
Phage ϕX174 (first genome)	5,375	1977
Hemophilus infiuenzae (bacterium, first organism)	1,830,000	1995
Mycoplasma genitalium (bacterium, smallest genome)	580,000	1995
Saccharomyces cerevisiae (yeast, first eukaryote)	12,068,000	1996
Methanococcus jannaschii (first archaeon)	1,660,000	1996
Escherichia coli (best studied bacterium)	4,639,221	1997
Borrelia burgdorferi (the spirochete that causes Lyme disease)	910,725	1997
Caenorhabditis elegans (first animal, roundworm)	97,000,000	1998
Arabidopsis thaliana (first plant, mustard family)	120,000,000	2000
Human chromosome 22 (first human chromosome)	53,000,000	1999
Drosophila melanogaster (a favorite genetic model)	180,000,000	2000
Human (working draft of the "holy grail" of genomics)	3,200,000,000	2001
Plasmodium falciparum (malaria parasite)	23,000,000	2002
Anopheles gambiae (major mosquito malaria carrier)	278,000,000	2002
Fugu rubripis (tiger pufferfish)	365,000,000	2002
Mus musculus (house mouse)	2,500,000,000	2002
Oryza sativa (rice, first cereal grain)	466,000,000	2002
Ciona intestinalis (sea squirt, primitive chordate)	117,000,000	2002
Human (finished sequence)	3,200,000,000	2003

The Human Genome Project

- Started in 1990 by a consortium led by Francis Collins, was expected to end in 2005,
  - Later the plan changed to a draft, end of 2000 and final draft in 2003
  - Method: map-then-sequence strategy
- In 1998, Celera (private company) led by Craig Venter, planned to complete a rough draft of HG by end 2000
  - Method: shot-gun sequencing
- A rough draft of HG was announced in June 26, 2000 by US president accompanied by Collins & Venter
- Today, 8 human genomes have been sequenced: 7 males and 1 female. This includes Dr Watson (Dec 2007) and Dr M Kriek (first female genome, the Netherland, May 2008)

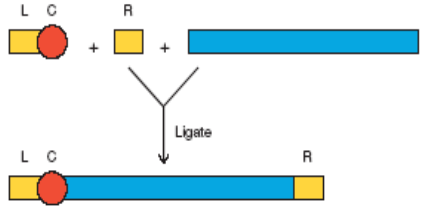
Genome sequence comparisons

- The direct comparisons of genome sequences between organisms is expected to be very informative, thus scientists are eager to sequence as many genomes as possible!!
- The divergence of genomic sequences between any two organisms reflect the evolutionary distance between those organisms

Vectors for large-scale genome projects

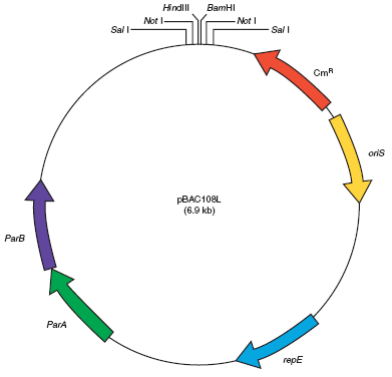
YAC

(Yeast Artificial Chromosome)



BAC

(Bacterial Artificial Chromosome)

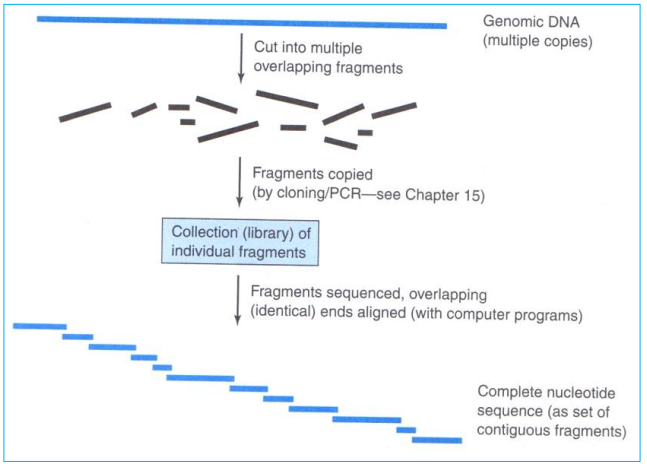


Clone-by-clone strategy: generation of a physical map

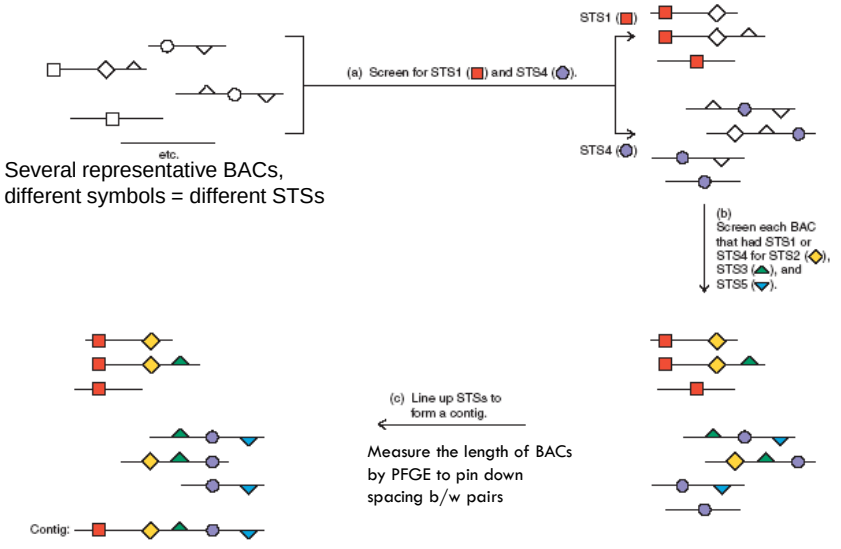
- It is systematic
- Method:
- Construct a genomic library using BACs
- Screen BACs using at least two STRs (STSs; seq tagged sites; 60-1000bp) spaced hundreds of kbs apart
- Mapping: screen BACs for several additional STRs, line BACs up in an overlapping fashion> Contig or contiguous DNAs containing long distances (actually overlapping)
- Go to finer mapping or sequencing of the contig

- Disadvantages of BACs: difficult to setup for large genomes

Clone-by-clone strategy: generation of a physical map



Clone-by-clone strategy: generation of a physical map



Shot-gun sequencing

Proposed in 1996 by Venter, Smith & Hood

No mapping is needed

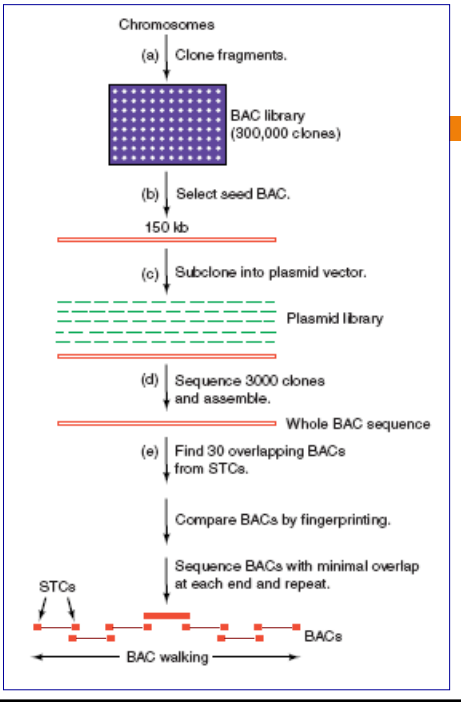
Starts with a set of BACs, each ~150 kb

Sequence both ends of each BAC, ~500bp

The 500bp serve as identity tag for each BAC, a sequence-tagged connector (STCs)

Fingerprint each BAC

Subdivide a SEED BAC into several pUC vectors (2kb/vector)



Shot-gun sequencing

- Systematic sequencing as shown in previous figure for shot-gun method would take a lot of time!!!
- *Modification:*
- sequencing of BACs randomly until there is 35 billion nts of sequence, equivalent to 10X of HG, giving high degree of coverage and accuracy> this should cover the whole HG
- Fed all data in a computer> find areas of overlap b/w clones, fit their sequence together> build the sequence of whole genome
- To overcome the gaps generated by sequencing small fragments (~2kb) in pUCs, they used the physical maps (the available STSs map)> thus this method is actually a hybrid of shot-gun and map-then-sequence

Working draft and finished version of human genome

- Working draft: ~ 90% complete, with error rate of 1 %
- Finished draft:
  - Contains ~ 99% of sequence that was possible to obtain
  - More accurate, with error rate of 0.001% and all sequences in proper order
  - Gene content 30,000-40,000 !!
  - ~half of genome composed of transposable elements

Applications of genomics

- Sequencing genomes: structural genomics
- Tow main applications:

1. Probing the pattern of gene expression in a given cell type at a given time (functional genomics)
   - DNA microarrays and Microchips
   - Serial analysis of gene expression
   - Deletion analysis
   - others
2. Finding genes involved in genetic traits, especially genetic diseases (positional cloning)

Applications of genomics:

Studying determinants of differences between individuals of a species

- SNPs (Single Nucleotide Polymorphism)
- Look for differences among individuals and prepare SNPs map for HG
- Potential uses of SNPs map
  - Study correlation b/w certain SNPs and genetic disease> can be used to screen subjects for tendency to develop a disease
  - Identify certain SNPs that correlate with good or poor response to drugs (pharmacogenomics)
  - Markers for measuring the frequency of recombination between genes