

Genome Structure

Kinetics and Components

Genome

- The **genome** is **all the DNA** in a cell.
 - All the DNA on all the chromosomes
 - Includes genes, intergenic sequences, repeats
- Specifically, it is all the DNA in an organelle.
- Eukaryotes can have 2-3 genomes
 - Nuclear genome
 - Mitochondrial genome
 - Plastid genome
- If not specified, “genome” usually refers to the nuclear genome.

Components of the human Genome

- Human genome has 3.2 billion base pairs of DNA
- About 1.5-2% codes for proteins
- About 40-50% is repetitive, made by (retro)transposition
- What is the function of the remaining 50%?

Finding the function of genes

- Genes were originally defined in terms of phenotypes of mutants
- Now we have sequences of lots of DNA from a variety of organisms, so ...
- Which portions of DNA actually do something?
 - What do they do?
 - code for protein or some other product?
 - regulate expression?
 - used in replication, etc?

Much DNA in large genomes is non-coding

- Complex genomes have roughly 10x to 30x more DNA than is required to encode all the RNAs or proteins in the organism.
- Contributors to the non-coding DNA include:
 - Introns in genes
 - Regulatory elements of genes
 - Multiple copies of genes, including pseudogenes
 - Intergenic sequences
 - Interspersed repeats

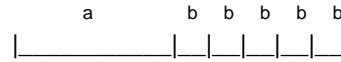
Distinct components in complex genomes

- Highly repeated DNA
 - R (repetition frequency) $\geq 100,000$
 - Almost no information, low complexity
- Moderately repeated DNA
 - $10 < R < 10,000$
 - Little information, moderate complexity
- “Single copy” DNA
 - $R=1$ or 2
 - Much information, high complexity

Reassociation kinetics measure sequence complexity

Sequence complexity is not the same as length

- **Complexity** is the number of base pairs of unique, i.e. nonrepeating, DNA.
- E.g. consider 1000 bp DNA.
 - 500 bp is sequence a, present in a single copy.
 - 500 bp is sequence b (100 bp) repeated 5X



$$L = \text{length} = 1000 \text{ bp} = a + 5b$$

$$N = \text{complexity} = 600 \text{ bp} = a + b$$

Less complex DNA renatures faster

Let a, b, ... z represent a string of base pairs in DNA that can hybridize. For simplicity in arithmetic, we will use 10 bp per letter.

DNA 1 = ab. This is very low sequence complexity, 2 letters or 20 bp.

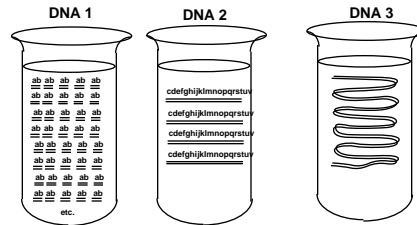
DNA 2 = cdefghijklmnopqrstuv. This is 10 times more complex (20 letters or 200 bp).

DNA 3 =

izyajczkblqfreighttrainrunninsofastelizabecthonqwtzxbifyoud
ontbelieveimleavingyoujustcountthedaysimgonerxcvwpowentdo
wntothecrossroadstriedtocatchariderobertjohnsonpzvmwcomeon
homeintomykitchentrad.

This is 100 times more complex (200 letters or 2000 bp).

Less complex DNA renatures faster, #2



For an equal mass/vol:

Molar concentration of each sequence:

150 microM

15 microM

1.5 microM

Relative rates of reassociation:

100

10

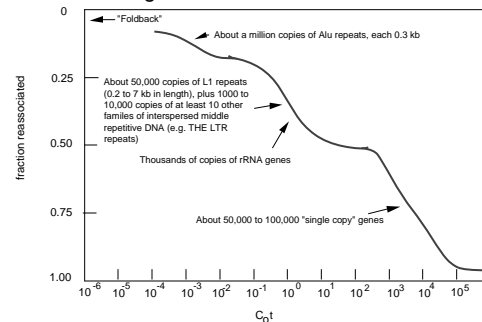
1

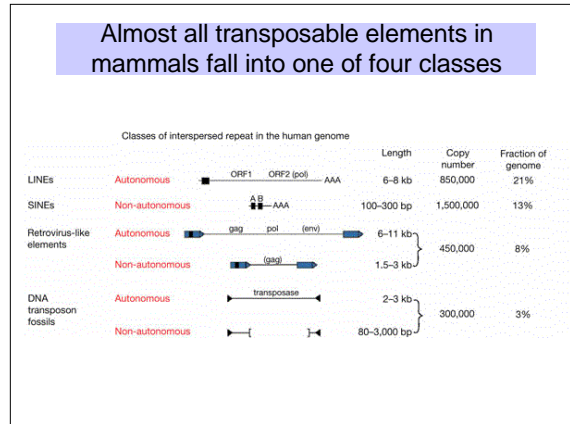
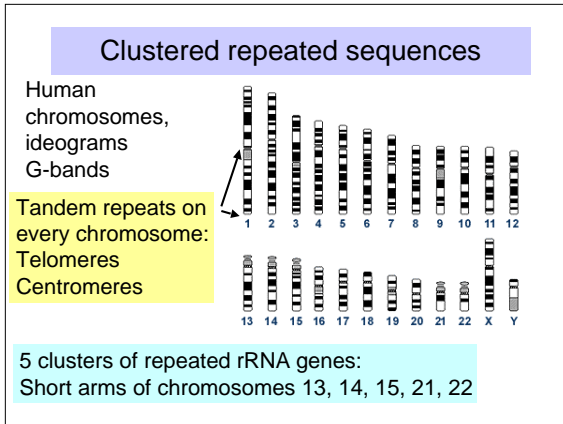
Insights from kinetics of renaturation

- Fast annealing DNA has low sequence complexity, but occupies about 40% of mammalian genomes.
- This is repeated DNA.
- Slow annealing DNA has high sequence complexity.
- This has genes (including a lot of intronic DNA) and a large amount of intergenic DNA.
- The repeats are interspersed throughout the genes and intergenic DNA.

Types of DNA in each kinetic component

Human genomic DNA





- ### Short interspersed repetitive elements: SINES
- Example: Alu repeats
 - Most abundant repeated DNA in primates
 - Short, about 300 bp
 - About 1 million copies
 - Likely derived from the gene for 7SL RNA
 - Cause new mutations in humans
 - They are **retrotransposons**
 - DNA segments that **move** via an **RNA intermediate**.
 - MIRs: Mammalian interspersed repeats
 - SINES found in all mammals
 - Analogous short retrotransposons found in genomes of all vertebrates.

- ### Long interspersed repetitive elements: LINEs
- Moderately abundant, long repeats
 - LINE1 family: most abundant
 - Up to 7000 bp long
 - About 50,000 copies
 - Retrotransposons
 - Encode reverse transcriptase and other enzymes required for transposition
 - No long terminal repeats (LTRs)
 - Cause new mutations in humans
 - Homologous repeats found in all mammals and many other animals

- ### Other common interspersed repeated sequences in humans
- LTR-containing retrotransposons
 - MaLR: mammalian, LTR retrotransposons
 - Endogenous retroviruses
 - MER4 (MEdium Reiterated repeat, family 4)
 - Repeats that resemble DNA transposons
 - MER1 and MER2
 - Mariner repeats
 - Were active early in mammalian evolution but are now inactive

- ### Finding repeats
- Compare a sequence to a database of known repeat sequences from the organism of interest
 - RepeatMasker
 - Arian Smit and P. Green, U. Wash.
 - <http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>
 - Try it on *INS* gene sequence

Strategy to determine DNA or RNA sequence

- Generate a *nested* set of fragments with one **common, labeled** end
- The other end terminates at one of the 4 nucleotides
- Electrophoretic resolution of the fragments allows the reading of the sequence:

Fragment of length 47 ends at G

48 A

49 T

Sequence is ...GAT....

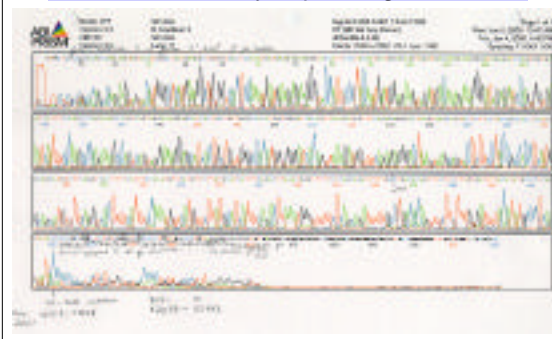
Common sequencing techniques

Technique	Common end	Label	Nt-specific end
DNA: Maxam & Gilbert	Restriction endonuclease	³² P	Base-specific chemical cleavage
DNA: Sanger	Primer for DNA polymerase	³² P or fluores-cence	Chain termination by dideoxy-nucleotides
RNA	Natural end of RNA	³² P	Nucleotide-specific enzymatic cleavage

Dideoxynucleotide sequencing

- The sequencing reactions generate a set of DNA molecules with one end in common.
- The other end terminates in an A, G, C or T, depending on which specific chain terminator was included in the reaction
 - E.g. ddATP to terminate at an A, which is complementary to a T in the template strand.
- Cycle Sequencing Movie:
 - <http://vector.cshl.org/resources/BiologyAnimationLibrary.htm>
- The Sanger dideoxynucleotide method is amenable to automation performed by robots.
- This approach is the one adapted for virtually all the whole-genome sequencing projects.

Example of output from automated dideoxysequencing

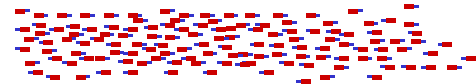


Shotgun sequencing of whole genomes

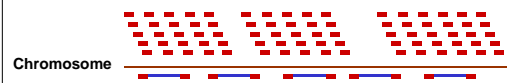
- Break total genomic DNA into small pieces (around 1000 bp in size) and clone into plasmids
- Sequence about 500 bp from each end.
- Use sequence alignments to assemble a final sequence.
- Requires that each bp be determined multiple times
 - about 3x coverage for small genomes (1-5 million bp)
 - about 10x coverage for large genomes (> 1 billion bp)

Shotgun sequencing and assembly

Sequence the ends of a huge number of small insert plasmids:



Align the sequences into contiguous assemblies (contigs):



The end sequences from mapped BAC contigs are used to assemble longer sequences from complex genomes. Gaps must be filled by directed sequencing.

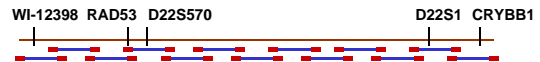
Genome sequencing after mapping

- Libraries of BACs have been screened and mapped to find overlapping arrays of contiguous clones (contigs)
 - E.g. find common restriction fragments in collections of clones
- Ends of the BACs are sequenced to provide markers through the genome
- Mapped contigs are then sequenced, using a combination of shotgun sequencing and directed sequencing

Directed sequencing of BAC contigs

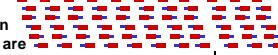
Chromosome 22 (part)

Anonymous markers and known genes mapped:



BAC contig, ends sequenced

Mapped BACs are broken into small pieces, which are shot-gun sequenced and assembled.



Gaps must be filled by alternate approaches, e.g. directed PCR.

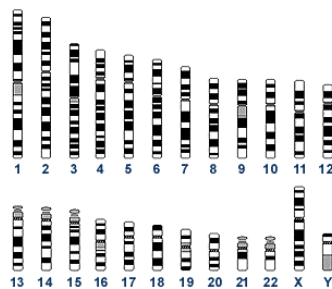
Chromosomes

Chromosomes organize and package genes inside cells

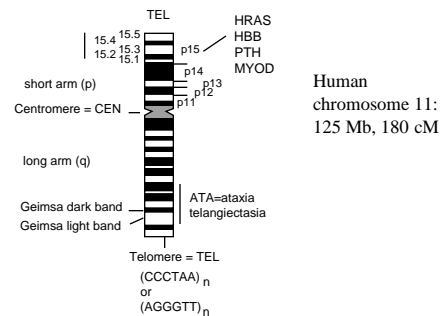
- Bind packaging proteins to DNA to make it more compact.
 - Histones +DNA = chromatin in eukaryotes
 - Fundamental subunit of chromatin = nucleosome
 - Virion proteins in viruses
 - HU (?) or other proteins in bacteria
- Loop chromatin and attach it to a matrix in nuclei

Human chromosomes, ideograms

Mitotic chromosomes are spread and stained with Geimsa. Those that stain are shown in black. G-bands (more A+T rich).

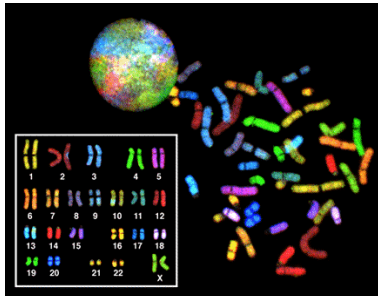


Bands and specialized regions of human chromosomes



Human chromosomes, spectral karyotype

Reagents specific to each chromosome. Chromosome painting.



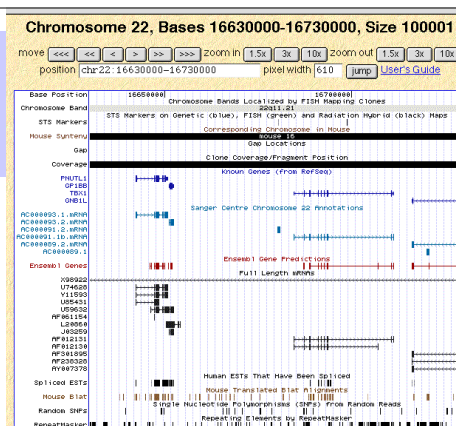
Distinctive and common features of chromosomes

- Distinctive proteins and DNA sequences have been used to develop chromosome painting reagents.
- Genomic DNA in vertebrates has long (megabase) stretches of G+C rich DNA, and other long stretches of A+T rich DNA
 - Called **isochores**
- Virtually all this DNA is organized into chromatin.
- Detailed views of sequence features: **Genome Browsers**
- Detailed information: **Genome databases**

Databases for genomic analysis

- Nucleic acid sequences
 - genomic and mRNA, including ESTs
- Protein sequences
- Protein structures
- Genetic and physical maps
- Organism-specific databases
- MedLine (PubMed)
- Online Mendelian Inheritance in Man (OMIM)

Human Genome Browser view



Identifying genes in genomic DNA sequences

- Identical to a known gene in the same species
- Highly significant match to a known gene in another species.
- Highly significant match to a spliced EST from the same or related species
- Parts of a gene may match portions of known genes at lower % identity
 - Assign potential functional domains by conserved motifs, e.g. protein kinase, ATPase, transmembrane domain
- Use **sequence alignment programs**

Common programs for sequence analysis

- RepeatMasker: find repeats in sequences
 - Also gives GC content
- BLAST: fast local alignment program
 - Search rapidly through sequence databases
 - Compare two sequences
- Gene finding by *ab initio* methods (genscan, GRAIL, etc.)
- PipMaker (to align 2 long genomic DNA sequences)
 - PipDispenser (<http://bio.cse.psu.edu>) for pre-computed human-mouse alignments
- Genome Browser: e.g. UCSC
 - sequence alignment portal: *Blat*

Results of BLAST search, *INS* vs. nr

L15440 (*INS* and flanking genes) vs. nr database

