

## 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.

## Genomics

- **Genomics** is the study of genomes, including large chromosomal segments containing many genes.
- The *initial phase of genomics* aims to map and sequence an initial set of entire genomes.
- *Functional genomics* aims to deduce information about the function of DNA sequences.
  - Should continue long after the initial genome sequences have been completed.

## Genomics vs. Genetics

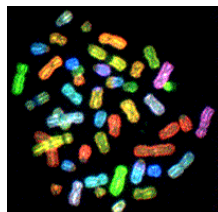
**Genetics:** study of **inherited phenotypes**

- Peter Goodfellow (1997, Nature Genetics 16:209-210):

"...I would define genetics as the study of inheritance and genomics as the study of genomes. The latter informs the former and includes the sequencing of genomes. The concept of functional genetics is a tautology (the whole point of genetics is to link genes with phenotypes). Functional genomics is the attachment of information about function to knowledge of DNA sequence' paradoxically, genetics is a major tool for functional genomics."

## Human genome

- 22 autosome pairs + 2 sex chromosomes
- 3 billion base pairs in the haploid genome
- Where and what are the 30,000 to 40,000 genes?
- Is there anything else interesting/important?



From NCBI web site, photo from T. Ried, Natl Human Genome Research Institute, NIH

## Components of the human Genome

- Human genome has 3.2 billion base pairs of DNA
- About 3% codes for proteins
- About 40-50% is repetitive, made by (retro)transposition
- What is the function of the remaining 50%?

### The Genomics Revolution

- Know (close to) all the genes in a genome, and the sequence of the proteins they encode.
- **BIOLOGY HAS BECOME A FINITE SCIENCE**
  - Hypotheses have to conform to what is present, not what you could imagine could happen.
- **No longer look at just individual genes**
  - Examine whole genomes or systems of genes

### Genomics, Genetics and Biochemistry

- Genetics: study of inherited phenotypes
- Genomics: study of genomes
- Biochemistry: study of the chemistry of living organisms and/or cells
- Revolution launched by full genome sequencing
  - Many biological problems now have finite (albeit complex) solutions.
  - New era will see an even greater interaction among these three disciplines

### 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?

### Genome Structure

- Distinct components of genomes
- Abundance and complexity of mRNA
- Normalized cDNA libraries and ESTs
- Genome sequences: gene numbers
- Comparative genomics

### 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

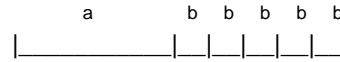
### Distinct components in complex genomes

- 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.

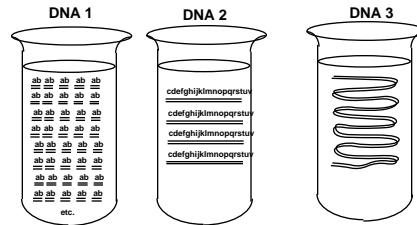
**DNA 2** = cdefghijklmnopqrstuv. This is 10 times more complex (20 letters or 200 bp).

**DNA 3** =

izyajczkblqfreighttrainrunnsofastelizabethcottonqwftzxbvifyou  
ontbelieveimleavingyoujustcountthedayssimgonerxcvwpowentdo  
wntothecrossroadstriedtocatcharideroberthjohnsonpzvmwcomeon  
homeintomykitchentrad.

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

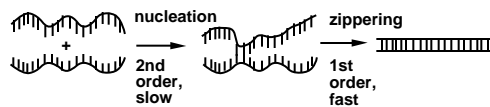
## Less complex DNA renatures faster, #2



For an equal mass/vol:

Molar concentration of each sequence:		
150 microM	15 microM	1.5 microM
Relative rates of reassociation:		
100	10	1

## Kinetics of renaturation are 2nd order



Denatured DNA  
(two single  
strands)

A short duplex  
forms at a region  
of complementarity.

Renatured  
DNA  
(two strands in  
duplex)

## Equations describing renaturation

Let  $C$  = concentration of single-stranded DNA at time  $t$   
(expressed as moles of nucleotides per liter).

The rate of loss of single-stranded (ss) DNA during renaturation is given by the following expression for a second-order rate process:

$$\frac{-dC}{dt} = kC^2 \quad \text{or} \quad \frac{dC}{C^2} = -kdt$$

Solving the differential equation yields:

$$\frac{C}{C_0} = \frac{1}{1 + kC_0t}$$

Time required for half-renaturation is inversely proportional to the rate constant

$$\frac{C}{C_0} = \frac{1}{1 + kC_0t}$$

At half renaturation,  $\frac{C}{C_0} = 0.5$ , and  $t = t_{1/2}$

$$C_0 t_{1/2} = \frac{1}{k}$$

$k$  in liters (mole nt)<sup>-1</sup> sec<sup>-1</sup>

Rate constant is inversely proportional to sequence complexity

$$k \propto \frac{\sqrt{L}}{N} \quad L = \text{length}; N = \text{complexity}$$

Empirically, the rate constant  $k$  has been measured as

$$k = 3 \times 10^5 \frac{\sqrt{L}}{N}$$

in 1.0 M Na<sup>+</sup> at T = T<sub>m</sub> - 25°C

Time required for half-renaturation is directly proportional to sequence complexity

$$C_0 t_{1/2} \propto \frac{N}{\sqrt{L}} \quad (4)$$

For a renaturation measurement, one usually shears DNA to a constant fragment length  $L$  (e.g. 400 bp). Then  $L$  is no longer a variable, and

$$C_0 t_{1/2} \propto N \quad (5)$$

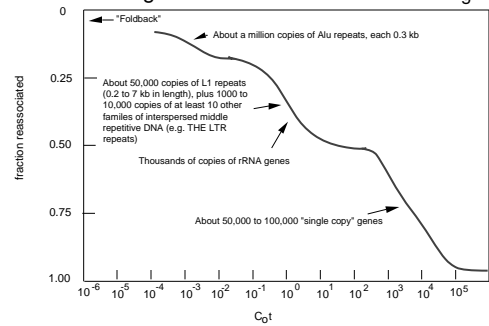
$$\frac{N^{\text{unknown}}}{N^{\text{standard}}} = \frac{C_0 t_{1/2}^{\text{unknown}}}{C_0 t_{1/2}^{\text{standard}}} \quad (6)$$

E.g. *E. coli*  $N = 4.639 \times 10^6$  bp

Types of DNA in each kinetic component

Human genomic DNA

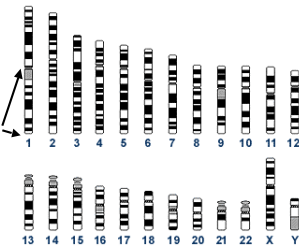
Fig. 1.7.5



Clustered repeated sequences

Human chromosomes, ideograms, G-bands

Tandem repeats on every chromosome: Telomeres, Centromeres



5 clusters of repeated rRNA genes: Short arms of chromosomes 13, 14, 15, 21, 22

Almost all transposable elements in mammals fall into one of four classes

Classes of interspersed repeat in the human genome		Length	Copy number	Fraction of genome
LINES	Autonomous: ORF1 ORF2 (pol) AAA	6-8 kb	850,000	21%
SINES	Non-autonomous: A B AAA	100-300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous: gag pol (env)	6-11 kb	450,000	8%
	Non-autonomous: (gag)	1.5-3 kb		
DNA transposon fossils	Autonomous: transposase	2-3 kb	300,000	3%
	Non-autonomous: [ ]	80-3,000 bp		

### 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