Genome Structure

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.

Human genome

• 22 autosome pairs + 2 sex chromosomes

• 3 billion base pairs in the haploid genome

• Where and what the 30,000 to

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

40,000 genes?

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

Finding the function of genes

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) >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

a b b b b b |\_\_\_\_\_\_\_\_\_\_\_|\_\_|\_\_|\_\_|\_\_|\_\_|

L = length = 1000 bp = a + 5b

N = complexity = 600 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** = izyajczkblqfreighttrainrunninsofastelizabethcottonqwftzxvbifyoud 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:**

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

DNA Transposons

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

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 **retrotranposons**

– 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