



## **Human Genome**

### **Introduction**

In 2000, researchers from around the world published a draft sequence of the entire genome. 20 labs from 6 countries worked on the sequence.

The International Human Genome Project – international effort begun in 1995

All data from this project is available to all online – free!

### **Historical Context**

1865 - Mendel

1900 - 1925 – Heredity resides in chromosomes, DNA is the genetic material

1925 – 1950 - Structure of DNA, double helix

1950 – 1975 – How information is stored in DNA (DNA → RNA → protein)

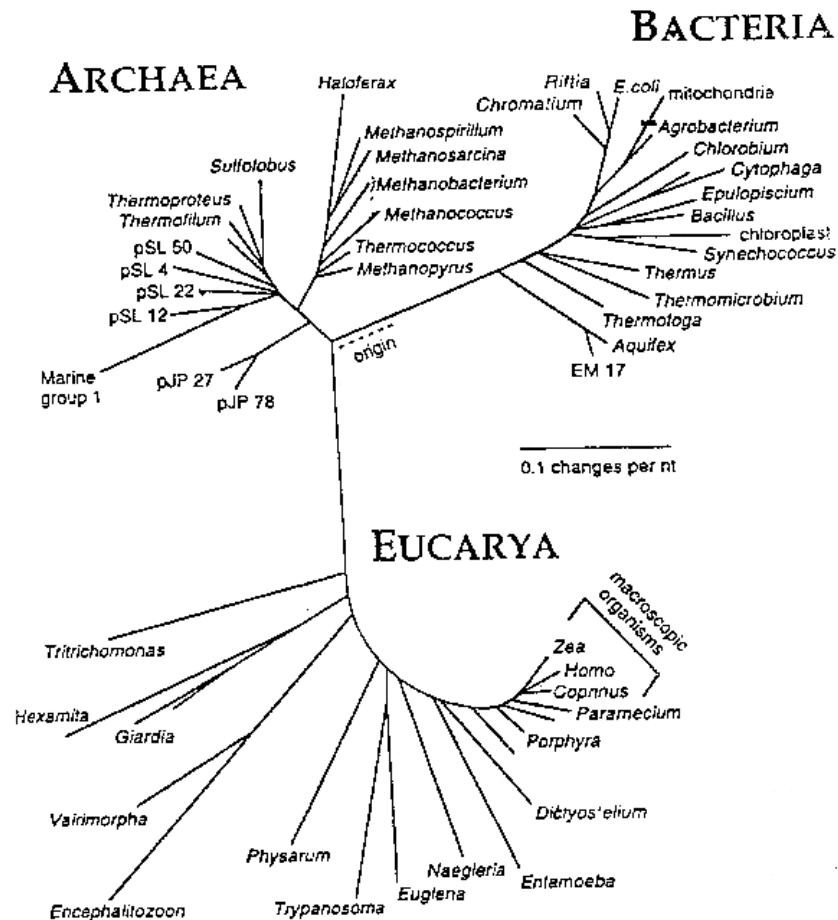
Lab techniques such as cloning and sequencing

1975 – 2000 – Sequencing of genes, small genomes, human genome

### **Why sequence the human and other genomes?**

- With the completion of the human genome we will no longer need to be “smart and lucky” to identify a gene (or genes) responsible for a disease, we can be “systematic.”
- Identification of genes and their functions will provide huge insight into molecular mechanisms and provide numerous targets for therapy. For example, finding that mutations leading to cystic fibrosis were located in a gene encoding a chloride ion channel led to a better understanding of the causes of ion imbalance in cystic fibrosis patients.
- Differences between individuals will again further our insight into molecular mechanisms and provide opportunities for individualized therapies. (Many different mutations in the chloride ion channel gene can lead to cystic fibrosis. A description of different treatments being tested for different CFTR mutations can be found in pages 7 – 10 of the lecture notes from our cystic fibrosis workshop. <http://www.colorado.edu/Outreach/BSI/k12activities/cysticfibrosis.html>)
- Comparison of the human genome to that of other organisms will further studies of the evolution of genes, gene families, and different life forms

- Sequence comparison has led to a complete revision of taxonomy and development of a completely new tree of life with three domains.



- Sequencing the human genome will also allow us to uncover things we could not imagine in advance. For example, studying the sequence of the human genome is beginning to lead to an understanding of how non-coding sequences may be regulating gene expression at several different levels.

### Sequencing Technology

1980s – 2 days and 1 person to get 600 bases of 10 samples

Today in the Molecular Cellular and Developmental Biology department at CU Boulder –

1 day and 1 person to get 600 bases of 192 samples

The human genome was sequenced on machines with even higher throughput

Cold Spring Harbor's DNA Interactive website has some online activities, information, and interviews on sequencing, the public vs private sector projects, cost etc. To access this information go to

[www.dnai.org](http://www.dnai.org)

Click on "genome."

Then click on "The Project."

### **Size**

The human genome is 3 billion bases long

New York Times spread out 6 pgs across would run from 4<sup>th</sup> St. to 142<sup>nd</sup> St.

The sequence would take up 69,900 double pages of the New York Times

If you printed the human genome in 12 font and stretched it out, it would run all the way from Penn Station, New York City to Union Station, Los Angeles.

It would take 142 large phone books to contain the entire human genome.

23 pairs of chromosomes ranging in size from

246,122,627 base pairs – Chromosome 1

44,626,493 base pairs – Chromosome 21

### **Differences among humans**

An average of 1 in 1200 bases differ between any two humans.

This is less than 0.1%.

The average 0.1% difference is responsible for inherited differences among humans (physical traits, genetically inherited diseases)

We refer to these differences as single nucleotide polymorphisms or SNPs

1.4 million SNPs have been identified and mapped.

SNPs allow studies of:

- Genome mapping – determining the location of genes linked to diseases or traits
- Evolution and migration of human populations
- Organization and evolution of the human genome
- Race – there is more variability within a given ethnic group than between ethnic groups. You cannot tell what "race" someone is by looking at their DNA sequence.

At Cold Spring Harbor's DNA Interactive site there is a page (genome fishing) that shows the size of the different chromosomes as well as the locations of SNPs. A second page (gene spots) shows a small sampling of important genes found on each chromosome.

[www.DNAi.org](http://www.DNAi.org)

Click on genome on the right hand side

Click on tour

Click on Genome Fishing or Gene Spots

## **Mutation and Crossing Over**

Most new mutations in humans arise during meiosis.

Mutations during meiosis occur two times more frequently in males compared to females.

Recombination rates are higher in distal regions of chromosomes and on shorter chromosome arms

Expect at least one crossover per chromosome arm in each meiosis

## **Coding sequences**

Surprisingly, the human genome has an estimated 30,000 – 40,000 genes.

This number is much lower than the previous estimate of 100,000 – 120,000 genes.

1 – 2% of the genome codes for protein

This is similar to the number of genes in mouse or mustard weed, and only twice as many genes as flies or nematodes.

## **Thus genome size and the number of genes do not account for vertebrate or human complexity.**

However, vertebrates have 5 times as many proteins as flies or worms.

Sequencing of the genomes of various organisms including human, mouse, fly and nematode has allowed us to observe that the complexity in vertebrates is largely due to

- alternative splicing (several proteins made from one gene)
- gene duplication and divergence resulting in large gene families
- evolution of new protein domains
- rearrangement of existing protein domains in unique combinations

## **Parasitic sequences**

46% of the genome is parasitic DNA sequences (transposable sequences)

These sequences are considered parasitic because they can copy themselves and move to a new place in the genome while leaving the original copy in place.

These parasitic sequences have selfish motives – they are concerned with their own reproduction and survival.

However they have been important in many of the innovations in our genome and have been important in shaping genomes.

- Some of our regulatory elements and genes originated from these sequences
- They have played an active role in shaping genomes by rearrangement and creating and shuffling genes.

These sequences may also be important in gene regulation.

To the researcher, these sequences are a rich paleontological record. They provide information that will further the study of evolution of genes, families, species.

Vertebrates vs other genomes (fly, nematode, plant)

- In vertebrates there is less removal of parasitic sequences. Vertebrates thus have older transposons than flies or nematodes.

- 60% of vertebrate transposable sequences two specific elements (LINE1 or Alu), other genomes have no predominant transposable sequence

#### Humans vs mouse

Transposition is much more common in mouse, the hominid genome is much more stable than any other genome studied to date. No one knows exactly why or what this means. 1 in 600 mutations in humans are due to transposition, compared to 1 in 10 for mouse.

At Cold Spring Harbor's DNA Interactive site there is a page that is a flyover visual of a piece of a chromosome. It shows how parasitic elements, introns (non-coding parts of a genes), exons (coding portions of genes), and gene families are arranged and interspersed in a representative piece of chromosome; the short arm of chromosome 11. This part of chromosome contains the  $\beta$ -globin and olfactory receptor genes as well as an intergenic region. Note: The movies have no audio so you will need to scroll down and read the text below the movie as the movie is playing.

[www.DNAi.org](http://www.DNAi.org)

Click on genome on the right hand side

Click on tour

Click on Flyover – choose one or more of the four pieces

#### GC content

GC rich regions are dark bands on chromosomes (associated with coding regions)

Average GC content – 41%

GC content of a small region varies from 60 – 70% for coding sequences to 30%.

#### CpG Islands

CG is greatly under-represented in the human genome (1/5 of the expected frequency).

Most CGs are methylated on the C,

When the C is methylated it leads to spontaneous deamination of C forming U

U is replaced with T by the repair system

Vertebrates methylate their DNA as a defense mechanism against bacterial infection.

We have specific molecules that recognize long stretches of unmodified DNA and direct an immune response against them.

CpG islands are areas in which CG are not methylated and occur at the expected frequency.

CpG Islands are of interest because they are associated with the 5' ends of genes (or pseudogenes).

There are 28,890 CpG islands not in the repeat sequences

Most have 60-70% GC content (expected for coding regions)

## **What's next**

Now we have the human genome sequenced – what's next?

The completion of the sequence does not mean that our understanding of the human genome is complete, rather it is just beginning.

The data analysis phase of the project will take longer than the sequencing project itself and will yield information we can not yet even imagine.

- Identifying genes - most of the 30,000 human genes have not yet been.
- Identifying gene products and their functions – the function of many of the identified genes is unknown.
- Identifying disease genes and designing treatments (some patient specific)
- Sequencing the genome of other organisms and animals
- Comparison of genomes to answer questions about evolution, both of specific organisms, and of genes and gene families.
- Identify regulatory sequences
  - “Traditional” regulatory sequences up and down stream of genes to which proteins bind to activate or repress expression.
  - Some transcribed sequences are not translated, instead the transcript itself acts in gene regulation
  - There is extensive DNA modification that is also thought to play a role in gene regulation
- Finish identifying SNPs and their association with different traits and populations. About 3 million SNPs have been identified to date.

**Activity**

Let's take a look at some random sequence. The following page contains a random set of bases.

There are 50 bases per line.

There are 46 lines for a total of 2300 bases.

How many of these pages would it take to print out the whole human genome (assume 3 billion base pairs as the size)?

Fill in the empty column of the table below. (Determine how many bases out of 2300 would be represent each characteristic)

<b>Characteristic</b>	<b>% of genome</b>	<b># of bases out of 2300</b>	<b>Color</b>
Coding DNA	1.5%		Green
Parasitic DNA	46%		Yellow
Differences between humans	0.1%		Pink

To generate a visual representation of the % of different types of sequences that make up the human genome you will shade in bases on the sample sequence. Using the colored highlighters supplied, color in correct number of bases on the following page. In reality these different sequences are interspersed throughout the genome. However, for this activity, it doesn't matter which particular bases you color in.

AGTCGTGCGTGGTACGAGACACACACAGGGTCTTAAACTTAGCTGAGCA  
GAGATGGACGTGATGTGCATGTCGTAGTCGTAGCTGTAGCTGATCGTGT  
GCCGCGTAGCTGCCCGTAGCTGTGTAATATACTGTATCGTAGCTGATC  
GTGACTGTACGTGATGCTGACGATCTGTGATGCTGAACACACAGCTATA  
GGTCGATGTGCAGCTAGCACGAGCTCGATGACGACGTAGCTGACACAC  
AGTGTAGACGTGTGACACACGTGCGGGCAAACGTTGACGCACGTGAC  
GTAGCTGGCAGCAGTCTAGCGAGACGTGCTGATGCGATGAAAAACGTG  
TACTGGTTTTATCGAGAGGGCGGCGGAGTACGATATTAGCTATTACTGA  
TGTCGATGTGATCGTAGTGAGCATGACTCGAACAGACGCAGGCAGTAA  
CAAAACGTGTAGCGCGCGGATATAACGTGATACGACAATATACGAG  
TACGTGTACGACCACACGTGACTATATCTACGTAGCTATATTATCGTATA  
TATATTTACGTGATTCAGTTACGATCGACCGTATATAGATGTCTGAACCAT  
CATCGACGATAGCAGTCAGTCTAACGTGATTAATTACTGCCTAGCATA  
TCGCAGTCTACGTAACCTATCGATACACGCTACACACACCAGTAACCATC  
AACACAAAAATTTTTAGGCATTACTTACAACAACTTTCTATAGAGGA  
GAGAGACTCCCCCCCCATCATATACGTACGTTGACGGATGTGGGACGT  
AGTGTACAGTGCTAGTGCTGTAGTGTGCTTATTTATCTTATGCTGTATTA  
CGTTTTTTGACACTGACTACGACACTACAGAAATAAGTAAGCCTATACG  
AACTCCCTAAGAAGAACTCACGACATACCGCCCGCGCTCTCCTCCTAG  
TCGATAAGCGATTATAGCGGGCGGATGTATGCGGGCGATGTAGGCGTGTAT  
ATATTCGGTATATTACGGGCGGACGCTCTCTAGCTCTATAGCGATCGAG  
CGTGTAGTACGAACGGGATATCTTCTCTATCTGATCTTATCTAGGACGG  
ATTCAGTGATCGGCGAACGTGTAGCACACGTACGGAGTATATATCGGAT  
CTTATCGTATTGCGGAGCTTAGCGTAGTGCGATCTTAGCGTGATCTGTAG  
TGCTGATGCTAATGCGCGTATGCGGAGCTGTAGCGATCTCGCGCGATCG  
AGGGGGGGCACGTATGTGACGGACACGCCAACGGTACAGGCTATTCT  
GAGGCAACCATAGTCGTGAACCATGCTTTGAGGCGTATTACGTGTAGTG  
CTGGCTGATGTGCGCGTAGTCGCGCGGAGGCGGAGTCGTTAACACAACC  
ACCAGTCGAGCGATTACGTGCGATTAACCCAACAAAAAAAAGTGTG  
TACGTGATGTGCGTGCCCGCGGCGATTCTCTTACCGCGGCCAACACGTA  
GTGCGATGTGCGTGAATATATGTCTCTCTTTTCTAAACCATGTTAGTCA  
TGGTAGCCCATCGTGCTGTGCCGCATATTATACAAAACGTGTCTCCCCCT  
AGGTCGTAATTCGTGATAGTCGTGTGATGTCGTGTGATGAACGTGTC  
GACTGTGTGACTGTAACACACCAACAGTGTGTACCCTACTTACGGCGG  
CGGGACGTACACGTGCCGCGATGCGCGATGCGCGATGTCGTAGGAGAG  
AGGATCGTGTACCACAGCTATCTACTGAGTCGTAGCTGTGAGCTGTGAT  
GGGAGGAACCCATAACCACTTCACTTATCGTACCCAGTCCAGTGTGCA  
TCGTGCTAGCCACACGTCTCTCCTCTGACCACGTCACCGTGACTGTGTAC  
GTGCTAGCTGATCGTGATCGTGACACCACGCATATGAACCAACATACCA  
TTATATATTATCCATCAGGTGGATCGTGTACGACGTGTACGTGTACGTGT  
ACCTAGCGTATCGATATTATATACGATCATATAGTACCCGCCGCCACAA  
AAAACGTATAGCTGCCCCCATGTGCTTAGACGAGCGTTAAAAAATG  
AGAAGGGGGACATTATCAGTCTATTGAAAGGATATTTACACCCCATTA  
GTGGCGAGGCCATTTATTATATTATATAGTGCTAGCTCCTCTTCTTATT  
CTTCTCTCTATTTACGGTAGCTGTACCCTACGGGGGGGATTCTTATGCGT  
AGGAAAAAAAACGTGTAGGCGGAGCTTTTATGCTAATCGGTATGCGA

## Supplemental Information

### Types of transposable sequences

long interspersed elements (LINEs – 21%)

have 2 open reading frames,

encode their own enzymes required to copy themselves and move  
(polymerase, recombinase)

Alu sequences are an example

short interspersed elements (SINEs – 13%)

do not encode their own proteins

use proteins encoded by LINEs to move

retrotransposons (8%)

encode reverse transcriptase and ability to move

other DNA transposons (2%)