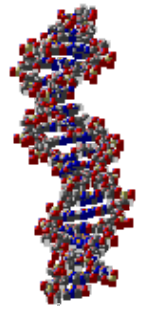
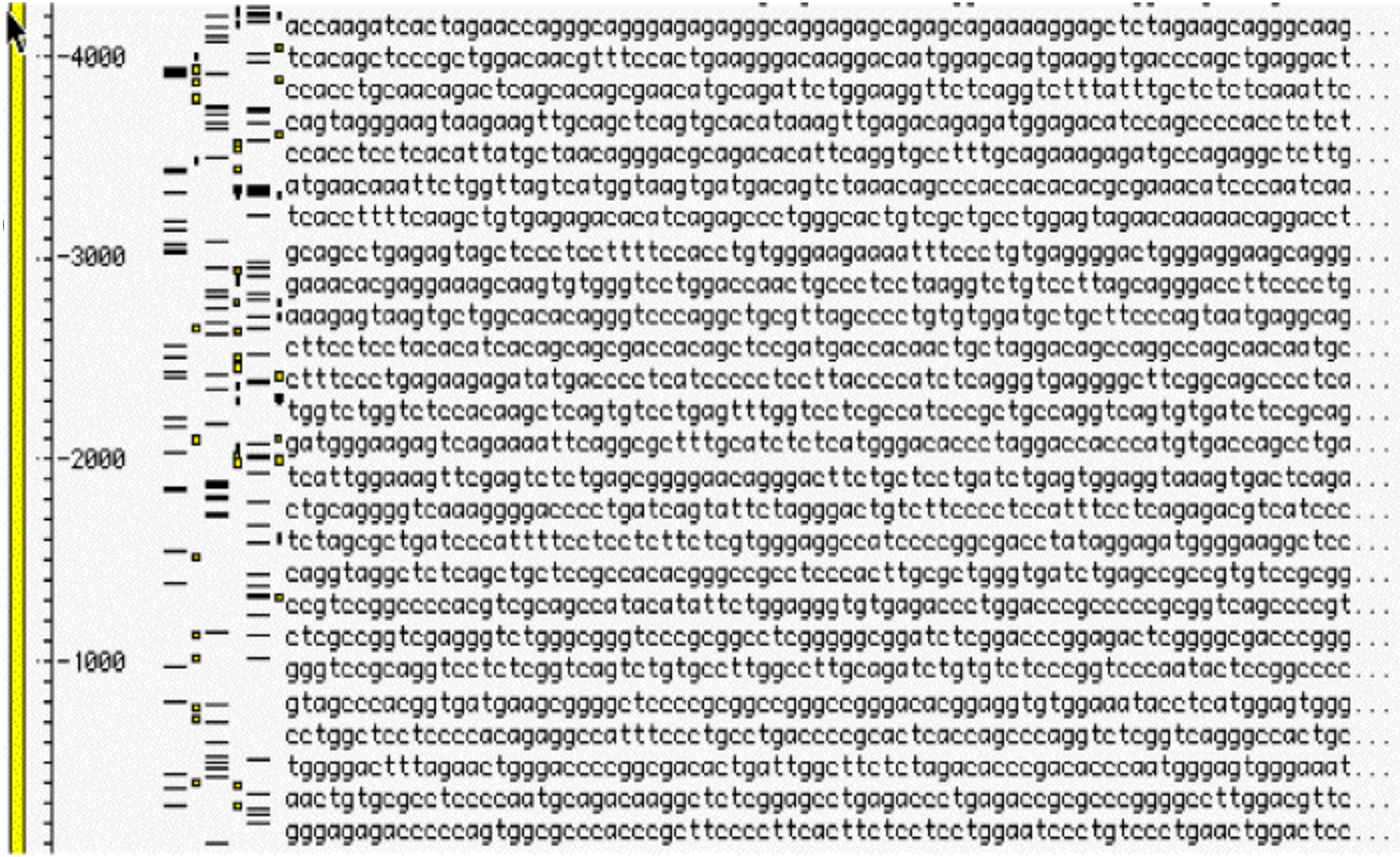
Human Genome Organization: An Update





Highlights of Human Genome Project Timetable

- Proposed in 1990 as 3 billion dollar joint venture between DOE and NIH with 15 year completion goal
- Private efforts by Celera Genomics in 1998 helped to accelerate project completion
- In 2000, working “draft” of human genome announced (95% complete). Draft sequence published in 2001.
- Work completed in April 2003 (only ~300 small gaps remaining)

Goals of the Human Genome Project

Create genetic and physical maps of the 22 autosomes and the X and Y chromosomes

Identify the entire set of genes in DNA

Determine the nucleotide sequence of 3 billion base pairs of DNA in the haploid genome

Analyze genetic variations among humans (identify single nucleotide polymorphisms called SNPs)

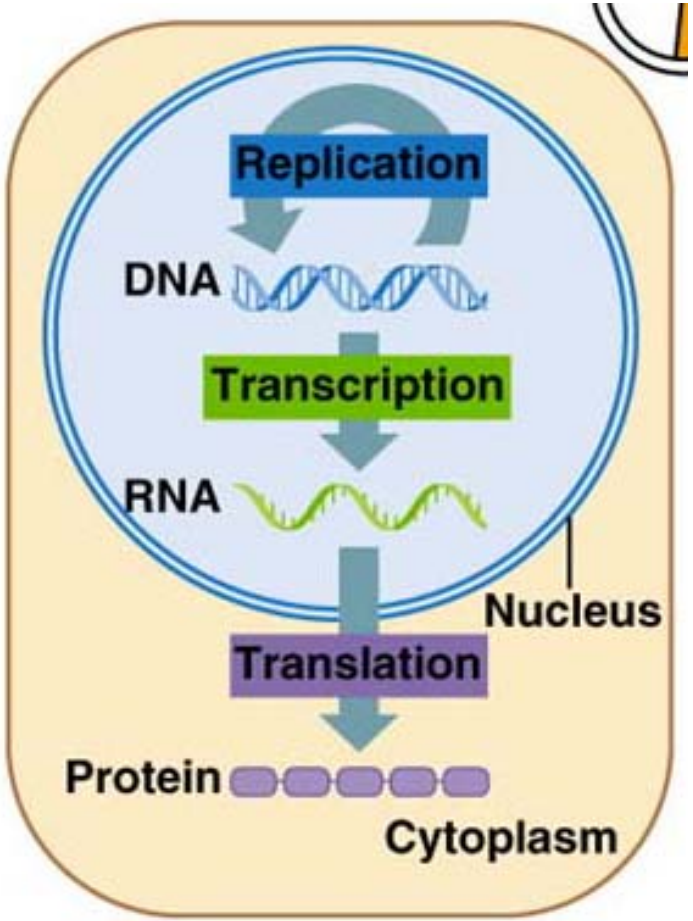
Map and sequence the genomes of model organisms (e.g., bacteria, yeast, nematodes, fruit flies, mice, etc)

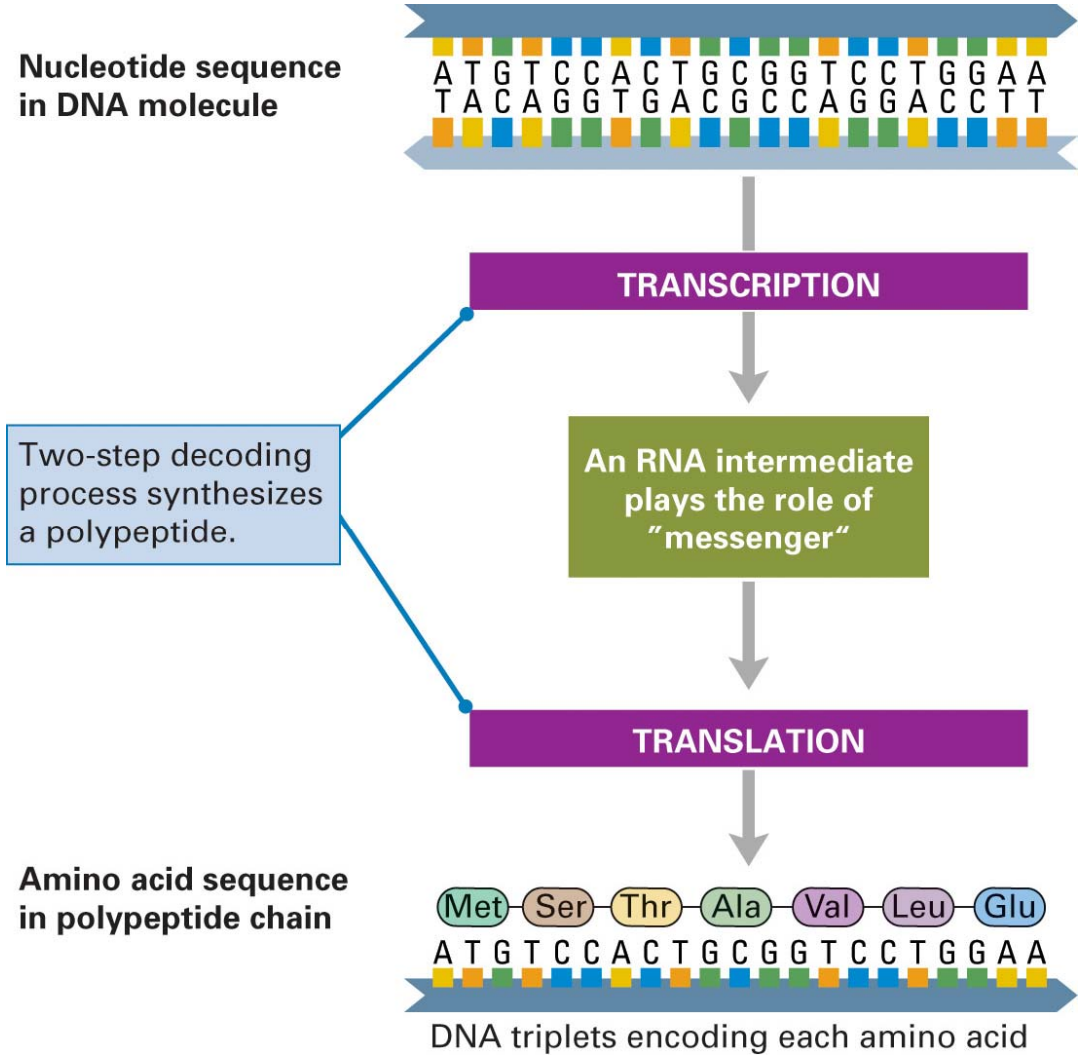
Develop the necessary laboratory and computational tools to assist in analyzing and understanding gene structure and function.

Disseminate genome information to scientists and the public

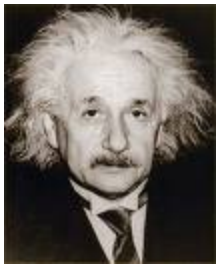
Examine ethical, social, and legal issues

Flow of Genetic Information in Eukaryotic Cells

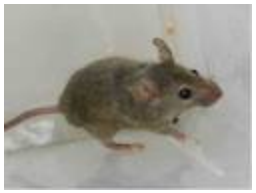


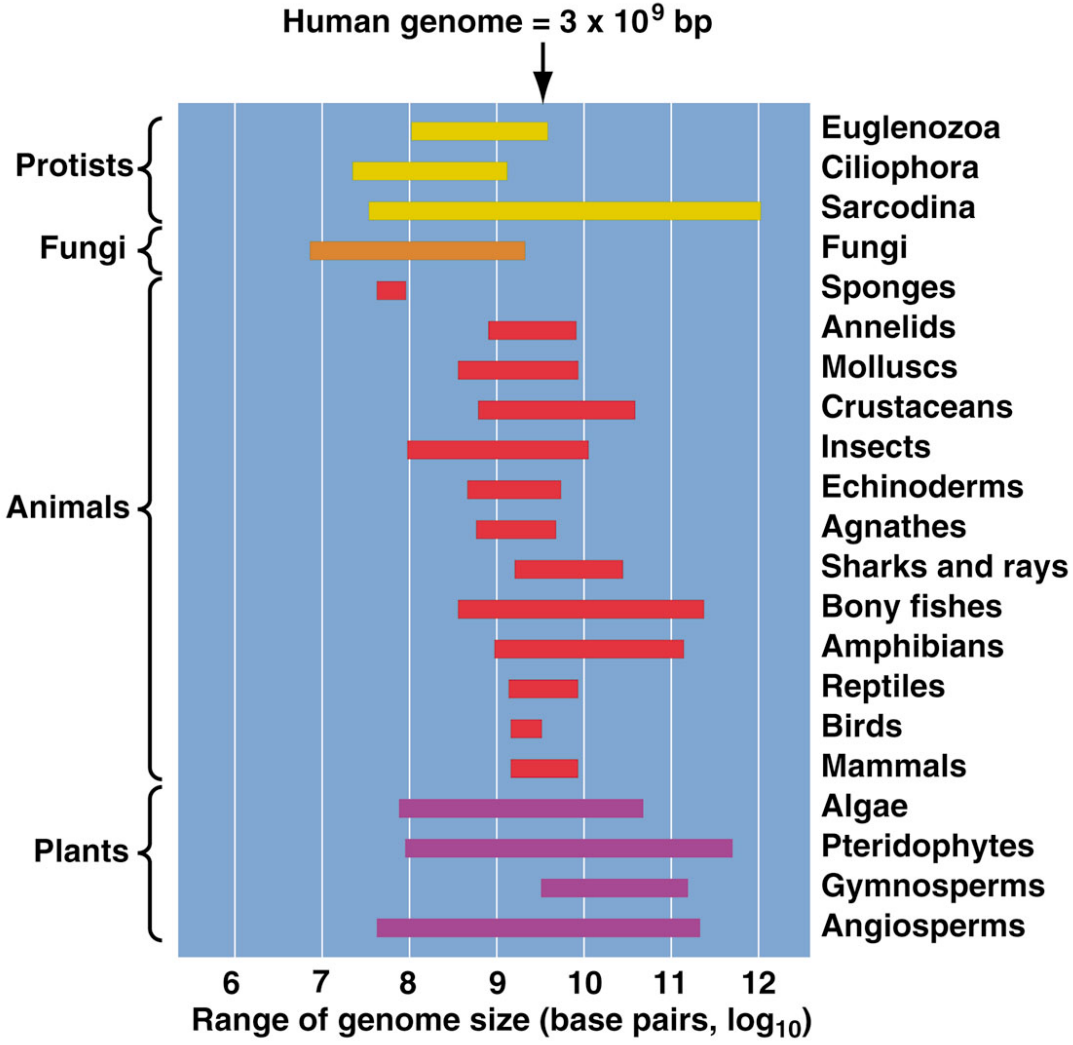


Complexity?



Which organism has the largest genome?





There is no correlation between complexity and genome size

Genome Comparisons: size of genome

Organism (scientific name)	Approximate size of genome (date completed)	Number of genes	Approximate percentage of genes shared with humans	Web access to genome databases
Bacterium (*Escherichia coli*)	4.1 million bp (1997)	4,403	Not determined	www.genome.wisc.edu/
Chicken (*Gallus gallus*)	1 billion bp (2004)	~20,000–23,000	60%	http://genomeold.wustl.edu/projects/chicken
Dog (*Canis familiaris*)	6.2 million bp (2003)	~18,400	75%	http://www.ncbi.gov/genome/guide/dog
Chimpanzee (*Pan troglodytes*)	~3 billion bp (initial draft, 2005)	~20,000–24,000	96%	http://www.nature.com/nature/focus/chimpgenome/index.html
Fruit fly (*Drosophila melanogaster*)	165 million bp (2000)	~13,600	50%	www.fruitfly.org
Humans (*Homo sapiens*)	~2.9 billion bp (2004)	~20,000–25,000	100%	www.doegenomes.org
Mouse (*Mus musculus*)	~2.5 billion bp (2002)	~30,000	~80%	www.informatics.jax.org
Plant (*Arabidopsis thaliana*)	119 million bp (2000)	~26,000	Not determined	www.arabidopsis.org
Rat (*Rattus norvegicus*)	~2.75 billion bp (2004)	~22,000	80%	www.hgsc.bcm.tmc.edu/projects/rat
Roundworm (*Caenorhabditis elegans*)	97 million bp (1998)	19,099	40%	genomeold.wustl.edu/projects/celegans
Yeast (*Saccharomyces cerevisiae*)	12 million bp (1996)	~5,700	30%	genomeold.wustl.edu/projects/yeast.index.php

Source: Nature Genome Gateway Web site www.nature.com/genomics/papers/.

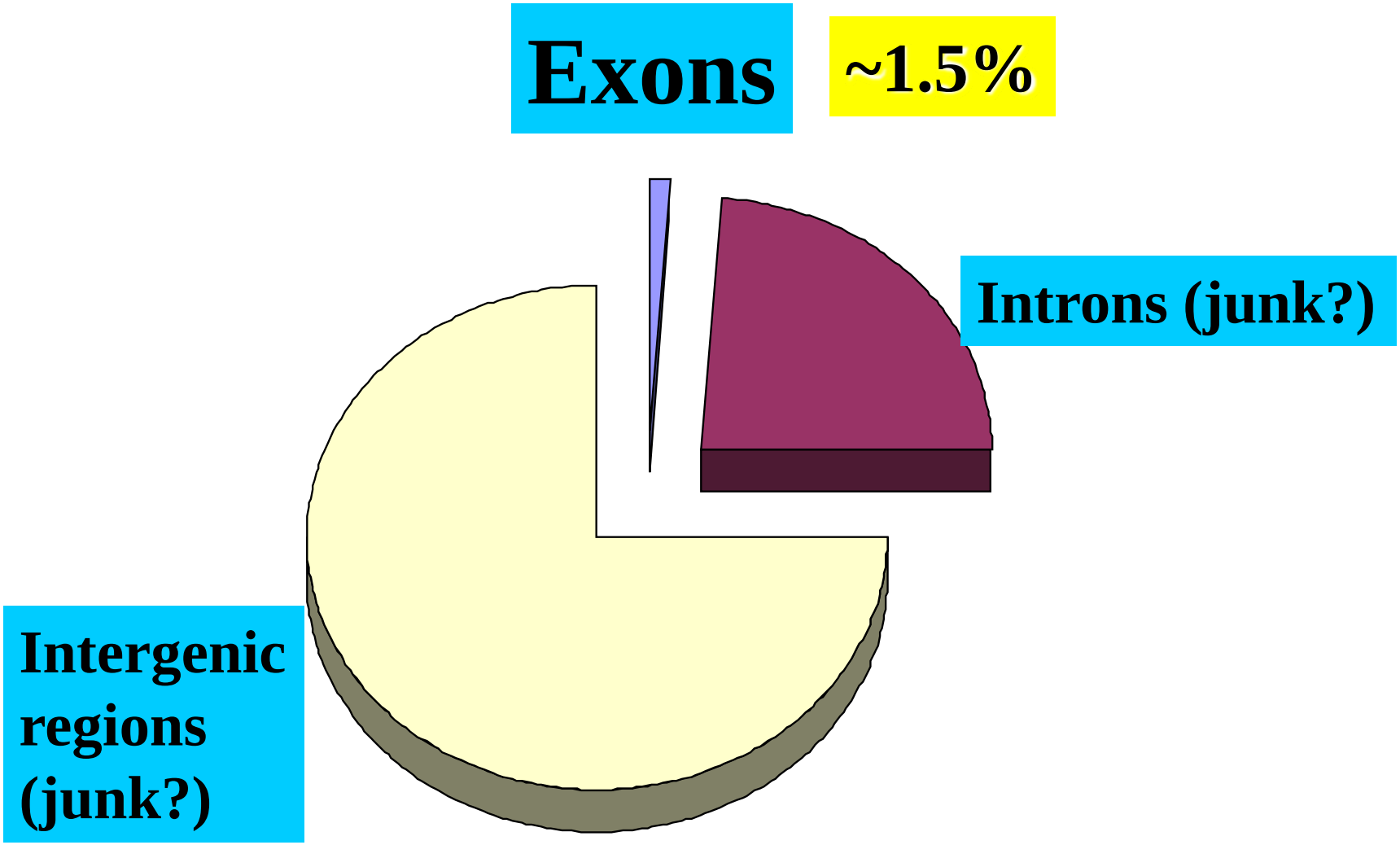
From: *Understanding the Human Genome Project* by Michael Palladino

Human gene insights:

- Average protein-coding gene size is ~30,000 base pairs with 8.8 exons separated by 7.8 introns.

  Note that largest gene is the dystrophin gene at ~2.4 million base pairs. Mutations in dystrophin cause muscular dystrophy.
- Chromosome 1 is the largest and has ~3000 genes. The smallest chromosome, the Y, has the fewest, ~250 genes.
- ~50% of human genes have no known function

>95% of our DNA consists of non-protein-coding DNA



What is the function of the “junk”?

- Regulatory roles necessary for controlling the expression of many genes. In some cases, encodes regulatory RNAs that influence gene expression
- Structural roles such as connecting adjacent genes, influencing the structure of the chromosome
- Other?

Number of genes in the human genome

- ~ 100,000 – 150,000 different proteins made by human cells
- **Complexity** of humans compared to other organisms AND the **large number of proteins**

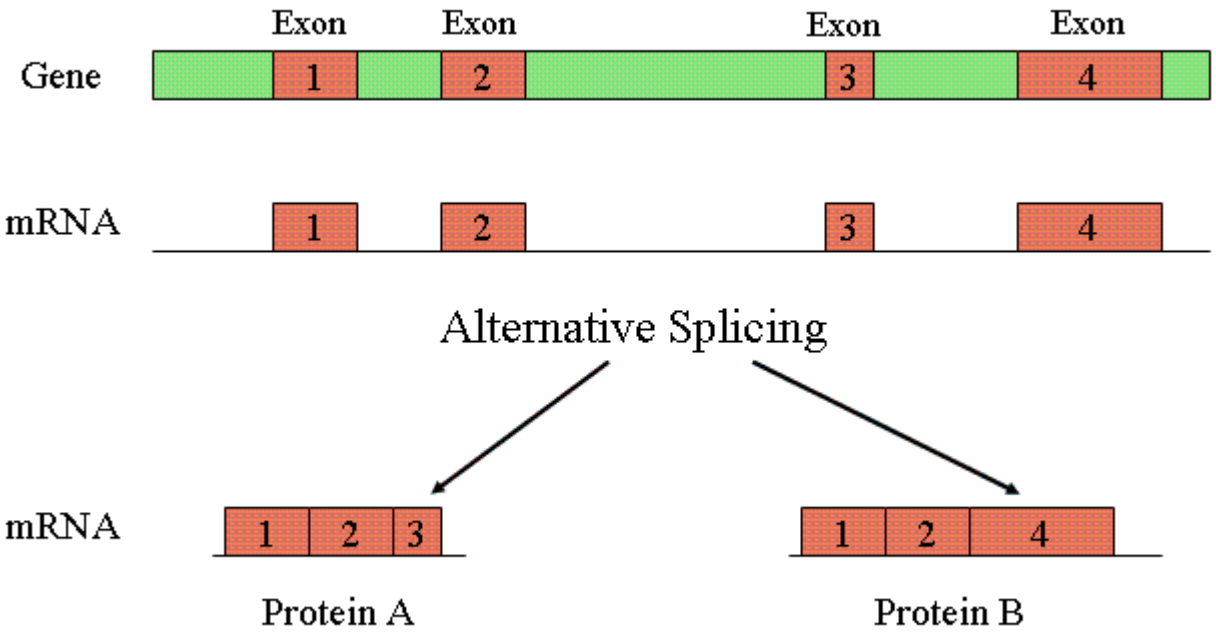
Number of genes at least 100,000

- HOWEVER, the number of protein-encoding genes is only ~20,000 to 25,000
- How can we explain this?

Genome Analysis shows:

- Large number of gene families with related functions

- Many genes code for multiple proteins through a complex process of mRNA processing called **alternative splicing**.



www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/

Genome Comparisons: number of genes

Organism (scientific name)	Approximate size of genome (date completed)	Number of genes	Approximate percentage of genes shared with humans	Web access to genome databases
Bacterium (*Escherichia coli*)	4.1 million bp (1997)	4,403	Not determined	www.genome.wisc.edu/
Chicken (*Gallus gallus*)	1 billion bp (2004)	~20,000–23,000	60%	http://genomeold.wustl.edu/projects/chicken
Dog (*Canis familiaris*)	6.2 million bp (2003)	~18,400	75%	http://www.ncbi.gov/genome/guide/dog
Chimpanzee (*Pan troglodytes*)	~3 billion bp (initial draft, 2005)	~20,000–24,000	96%	http://www.nature.com/nature/focus/chimpgenome/index.html
Fruit fly (*Drosophila melanogaster*)	165 million bp (2000)	~13,600	50%	www.fruitfly.org
Humans (*Homo sapiens*)	~2.9 billion bp (2004)	~20,000–25,000	100%	www.doegenomes.org
Mouse (*Mus musculus*)	~2.5 billion bp (2002)	~30,000	~80%	www.informatics.jax.org
Plant (*Arabidopsis thaliana*)	119 million bp (2000)	~26,000	Not determined	www.arabidopsis.org
Rat (*Rattus norvegicus*)	~2.75 billion bp (2004)	~22,000	80%	www.hgsc.bcm.tmc.edu/projects/rat
Roundworm (*Caenorhabditis elegans*)	97 million bp (1998)	19,099	40%	genomeold.wustl.edu/projects/celegans
Yeast (*Saccharomyces cerevisiae*)	12 million bp (1996)	~5,700	30%	genomeold.wustl.edu/projects/yeast.index.php

Source: Nature Genome Gateway Web site www.nature.com/genomics/papers/.

From: *Understanding the Human Genome Project* by Michael Palladino

What have we learned about ourselves from sequencing model organism genomes?

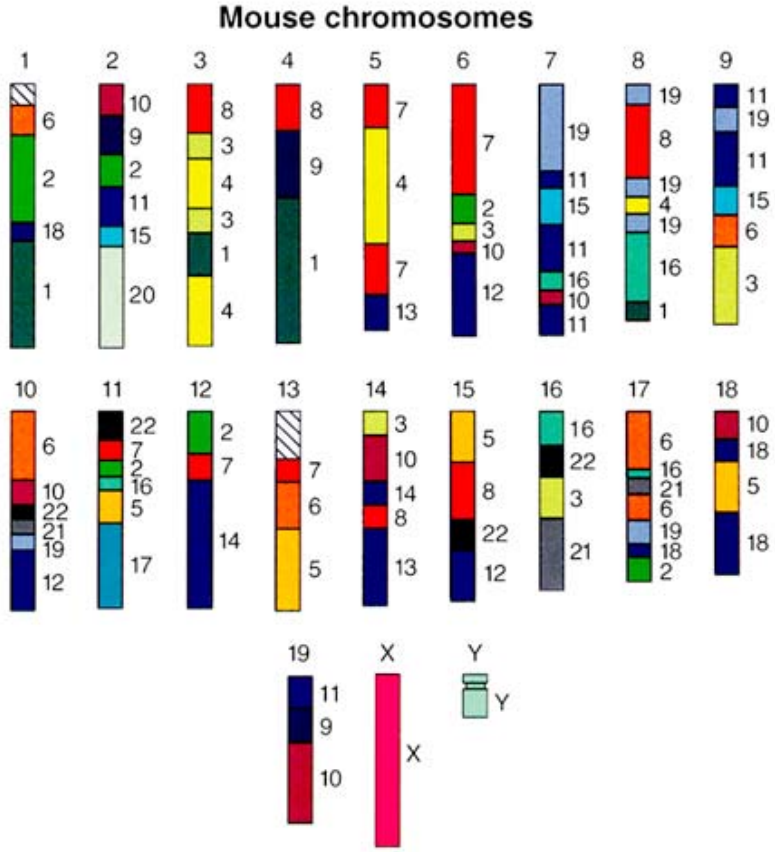
Just how unique are humans?

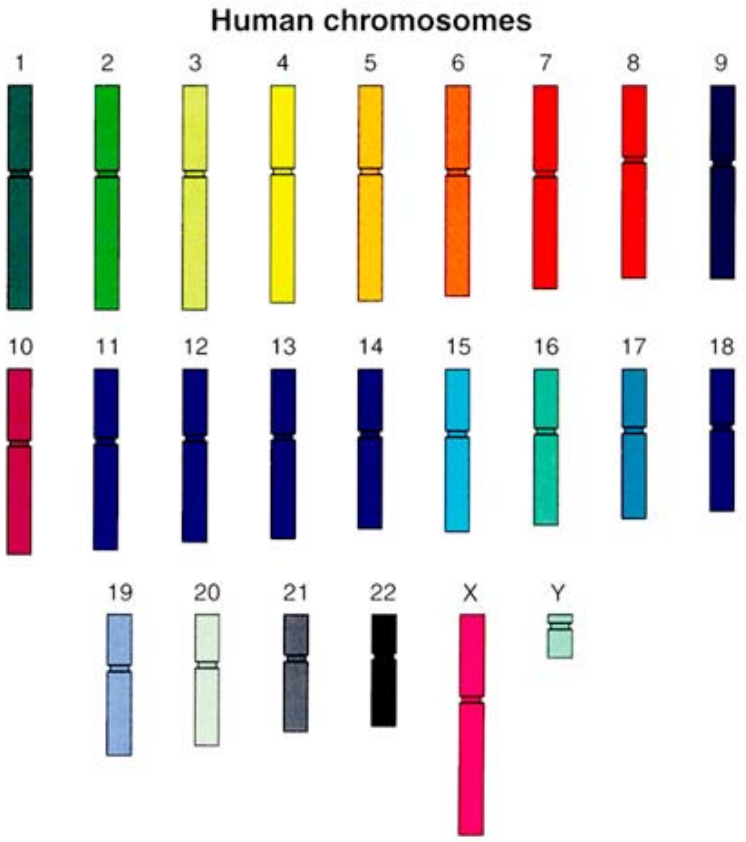
- Large numbers of genes are in common with other organisms
  - ~50% of our genes are also found in fruit flies
  - ~40% of our genes are also found in roundworms
  - ~30% of our genes are also found in yeast
  - ~80% of our genes are shared with the mouse and ~96% of our genes are shared with chimpanzees
  - ~100 of our genes are even shared with bacteria

Genomic comparisons between mice and men

- Both organisms have ~same number of genes
- Most of the common genes share the same intron and exon arrangement
- Nucleotide sequences within common gene exons are conserved to a high degree
- ~1/4 of alternatively spliced exons are specific either to human or mouse, such that species-specific proteins likely account for the differences between species.

Mouse and Human Genetic Similarities





Courtesy Lisa Stubbs
Oak Ridge National Laboratory

YGA 98-075R2

What else can we learn using model organisms?

Many genes determining body plan, organ development, and aging are nearly identical to genes in the fruit fly

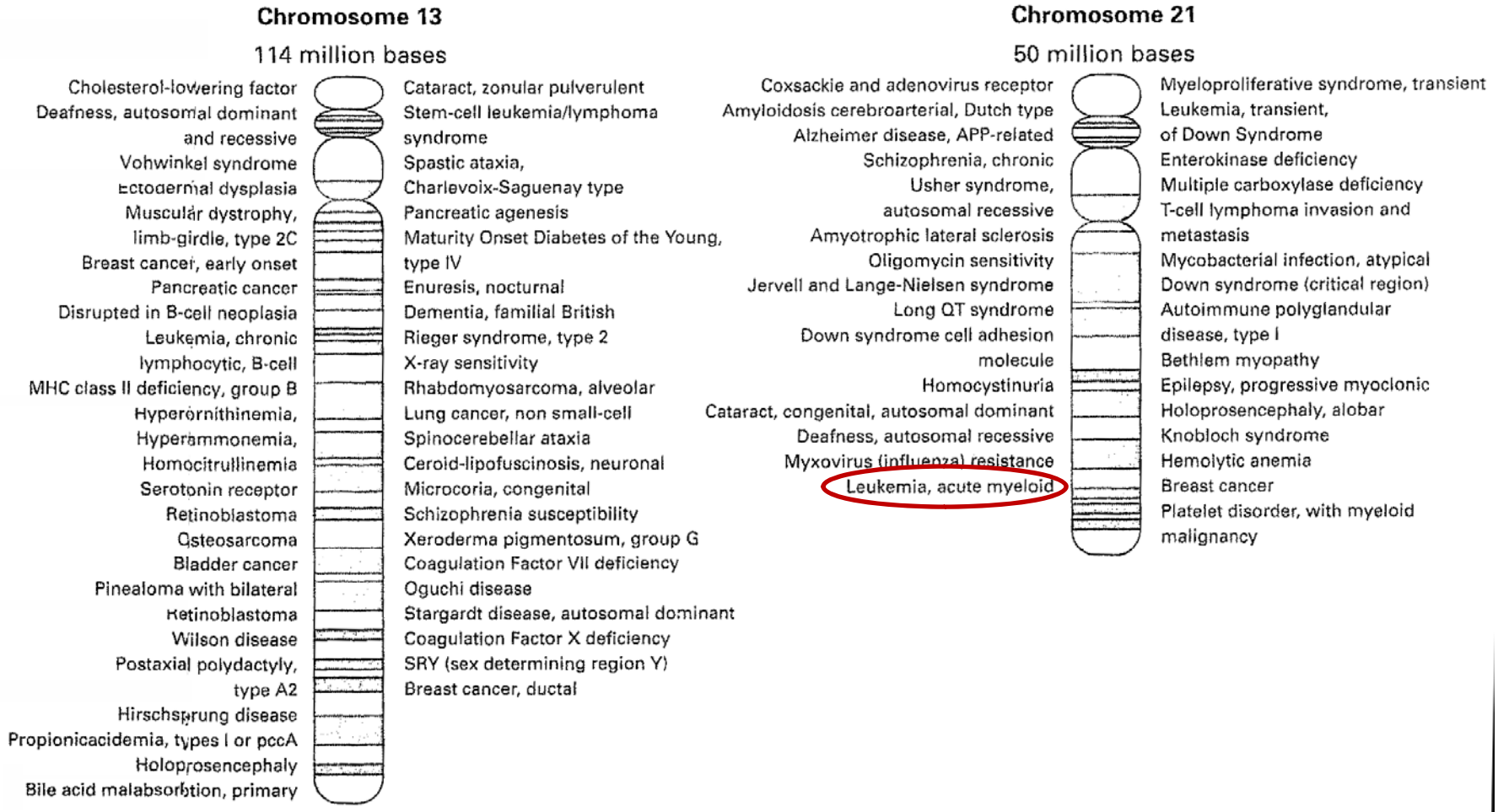
~61% of genes mutated in nearly 300 human disease conditions are found in the fruit fly. Genes include those involved in prostate cancer, pancreatic cancer, cardiac disease, cystic fibrosis, leukemia, and many other human genetic disorders.

Mapping Human Disease Genes

- Approximately 12 disease genes mapped by 1989

- Thousands of human disease genes have been identified and mapped as a result of the Human Genome Project

Disease genes on chromosomes 13 and 21



From *Understanding the Human Genome Project* by M. Palladino

Mapping the Cancer Genome

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley[1,2,3,4]*, Elaine R. Mardis[2,3]*, Li Ding[2,3], Bob Fulton[3], Michael D. McLellan[3], Ken Chen[3], David Dooling[3], Brian H. Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E. Larson[3], Dan C. Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W. Hillier[3,8], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R. Osborne[3], Patrick Minx[3], David Gordon[8], Asif Chinwalla[3], Yu Zhao[1], Rhonda E. Ries[1], Jacqueline E. Payton[5], Peter Westervelt[1,4], Michael H. Tomasson[1,4], Mark Watson[3,4,5], Jack Baty[6], Jennifer Ivanovich[4,7], Sharon Heath[1,4], William D. Shannon[1,4], Rakesh Nagarajan[4,5], Matthew J. Walter[1,4], Daniel C. Link[1,4], Timothy A. Graubert[1,4], John F. DiPersio[1,4] & Richard K. Wilson[2,3,4]

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults in the United States each year. The treatment of this disease has changed little in the past two decades, because most of the genetic events that initiate the disease remain undiscovered. Whole-genome sequencing is now possible at a reasonable cost and timeframe to use this approach for the unbiased discovery of tumour-specific somatic mutations that alter the protein-coding genes. Here we present the results obtained from sequencing a typical acute myeloid leukaemia genome, and its matched normal counterpart obtained from the same patient's skin. We discovered ten genes with acquired mutations; two were previously described mutations that are thought to contribute to tumour progression, and eight were new mutations present in virtually all tumour cells at presentation and relapse, the function of which is not yet known. Our study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to targeted therapies.

Nature **456** (November 6, 2008), 66-72.

Summary of genetic changes in protein-coding genes from AML patient

Table 2 | Non-synonymous somatic mutations detected in the AML sample

Gene	Consequence	Type	Solexa tumour reads WT:variant	Solexa skin reads WT:variant	Conservation score of mutant base	Mutations in other AML cases*
CDH24	Y590X	Nonsense	9:9	16:0	0.998	0/187
SLC15A1	W77X	Nonsense	15:12	19:0	1.000	0/187
KNDC1	L799F	Missense	7:8	20:0	NA	0/187
PTPRT	P1235L	Missense	9:13	16:0	1.000	0/187
GRINL1B	R176H	Missense	15:10	14:0	NA	0/187
GPR123	T38I	Missense	11:11	13:0	NA	0/187
EBI2	A338V	Missense	7:12	18:2	1.000	0/187
PCLKC	P1004L	Missense	19:9	15:1	0.98	0/187
FLT3	ITD	Indel	18:12	8:0	NA	51/185
NPM1	CATG ins	Indel	36:6	33:0	NA	43/180

Ins, insertion; WT, wild type.

*Patient cohort defined in ref. 23

Ley *et al*., 2008. *Nature* **456**: 66-72

Major findings

- Comparison of genome from normal skin compared to tumor cell from same patient
- 10 protein-coding gene differences: 2 already identified genetic alterations and 8 previously unknown mutations
- Alterations also noted in non-coding DNA
- Differences noted in genes from other AML patients suggests complex and diverse pathways to cancer onset and progression
- Prospect for personalized treatment strategies over time

Outgrowths from Human Genome Project and Future Prospects

- **Human Proteome Project** – to determine the structure and function of all human proteins (includes Protein Structure Initiative)
- **ENCODE** (ENCyclopedia of DNA Elements) to identify gene regulatory sequences
- **Microbial Genome Program** and **Genomes to Life Program** to sequence wide range of microbial genomes and to explore novel ways in which microbes can be used to develop new energy, remediate environmental waste, and other applications, respectively.
- **The Cancer Genome Atlas** (TCGA) to complete a catalogue of genomic changes involved in cancer

Summary

- Human genome consists of ~3 billion base pairs.
- Approximately 1.5% of genome codes for proteins. Other parts of genome vital for genome structural integrity and regulation.
- Fewer genes exist than originally expected (~20,000-25,000 genes instead of >100,000 or so, based on protein diversity). The functions of over 50% of proteins is unknown.
- Alternative splicing is the major mechanism to account for protein diversity (one gene codes for more than one protein).
- Comparative genomics using model organisms has increased our understanding of human gene structure and function since many genes are conserved between organisms. Many human disease genes have counterparts in some model organisms.
- The Human Genome Project provides a reference genome for projects that seek an understanding of genome changes in cancer and other diseases.