Key Goals This Term

- How can we choose an appropriate model or algorithm to interpret biological data?
- What biological questions can we hope to answer with the data that we observe?
- How can we design better experiments?
- When can our conclusions be viewed as mechanistic?

Course Team

- Prof. Tommi Jaakkoa
- Prof. David Gifford
- Tim Danford

Lectures

- Sequence models (Lectures 1 – 4)
  - Sequence alignment
  - DNA sequence element discovery
- Single data source models (Lectures 5 – 12)
  - Microarray data binding models
  - Proteomic data models
  - Classification and clustering
- Integrated models (Lectures 13 – 22)
  - Network models
  - Functional modules
  - Dynamic models

Requirements

- Team Project
  - Topic chosen by you in consultation with us
  - Roughly four person teams
  - Intermediate (10 minute) and final presentation (20 minute) in-class presentation
- Problem Sets (4 – 5)
- One Final Quiz
- You must register even if you are a listener

Structure and Function of the Genome

Chromosomes

Human Genome

Comparative Genomics

Genes and their Products

Stem Cells

Diagram removed for copyright reasons.
Structure of chromosomes, genes, DNA double helix and base pairs.

DNA is Packaged into Nucleosomes

Nucleosomes consist of 140 bp DNA wrapped around 8 histone proteins

2 X (H2A, H2B, H3, H4)

Histones

Chromosomes are Arrays of Nucleosomes

Diagram removed for copyright reasons.

Alberts et al., Molecular Biology of the Cell

Structure of the Nucleus

Chromatin
-contains DNA and proteins formed into chromosomes

Nucleolus
-manufactures ribosomes

Diagram removed for copyright reasons.

Nuclear envelope
-allows the nucleus to control entry and exit of molecules

Alberts et al., Molecular Biology of the Cell

Structure and Function of the Genome

Chromosomes

Human Genome

Comparative Genomics

Genes and their Products

Stem Cells

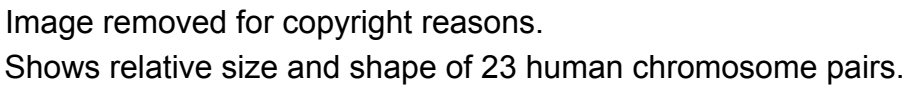


Image removed for copyright reasons.
Shows relative size and shape of 23 human chromosome pairs.

Human Gene Content: Surprisingly Few Genes

Only 1% of genome is genes

Protein-coding Gene Number: 30,000?

Human Genes:

- Tend to live in GC-rich regions
- Few new protein domains,
  many new domain architectures
- Big expansions of some families . . .
  Smell receptors
  Immunoglobulins
  Growth Factors

Human Genome Overview

Size of the genome	2.91 Gbp
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated-2001)	39,114
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. Y (5 genes/Mb)
Percent of base pairs spanned by genes	25.5
Percent of base pairs spanned by exons	1.1
Percent of base pairs spanned by introns	24.4
Percent of base pairs in intergenic DNA	74.5
Longest intergenic region	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

Venter et al., Science (2001)

National Center for Biotechnology Information (NCBI)
Human Genome Build 35, May 2004 Assembly (hg17)

Sequence name	Length (bp) including gaps
chr1	245,522,847
chr1_random	3,897,131
chr2	243,018,229
chr2_random	418,158
chr3	199,505,740
chr3_random	970,716
chr4	191,411,218
chr4_random	1,030,282
chr5	180,857,866
chr5_random	143,687
chr6	170,975,699
chr6_hla_hap1	139,182
chr6_hla_hap2	150,447
chr6_random	1,875,562
chr7	158,628,139
chr7_random	778,964
chr8	146,274,826
chr8_random	943,810
chr9	138,429,268
chr9_random	1,312,665
chr10	135,413,628
chr10_random	113,275
chr11	134,452,384
chr12	132,449,811
chr12_random	466,818
chr13	114,142,980
chr13_random	186,858
chr14	106,368,585
chr15	100,338,915
chr15_random	784,346
chr16	88,827,254
chr16_random	105,485
chr17	78,774,742
chr17_random	2,618,010
chr18	76,117,153
chr18_random	4,262
chr19	63,811,651
chr19_random	301,858
chr20	62,435,964
chr21	46,944,323
chr22	49,554,710
chr22_random	257,318
chrX	154,824,264
chrX_random	1,719,168
chrY	57,701,691
chrM	16,571
Total	3,095,016,460

Describing the Human Genome

chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
RH18061;RH80175	Displays region between STS markers RH18061;RH80175. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease

Oct4 Gene Locus (chr6)

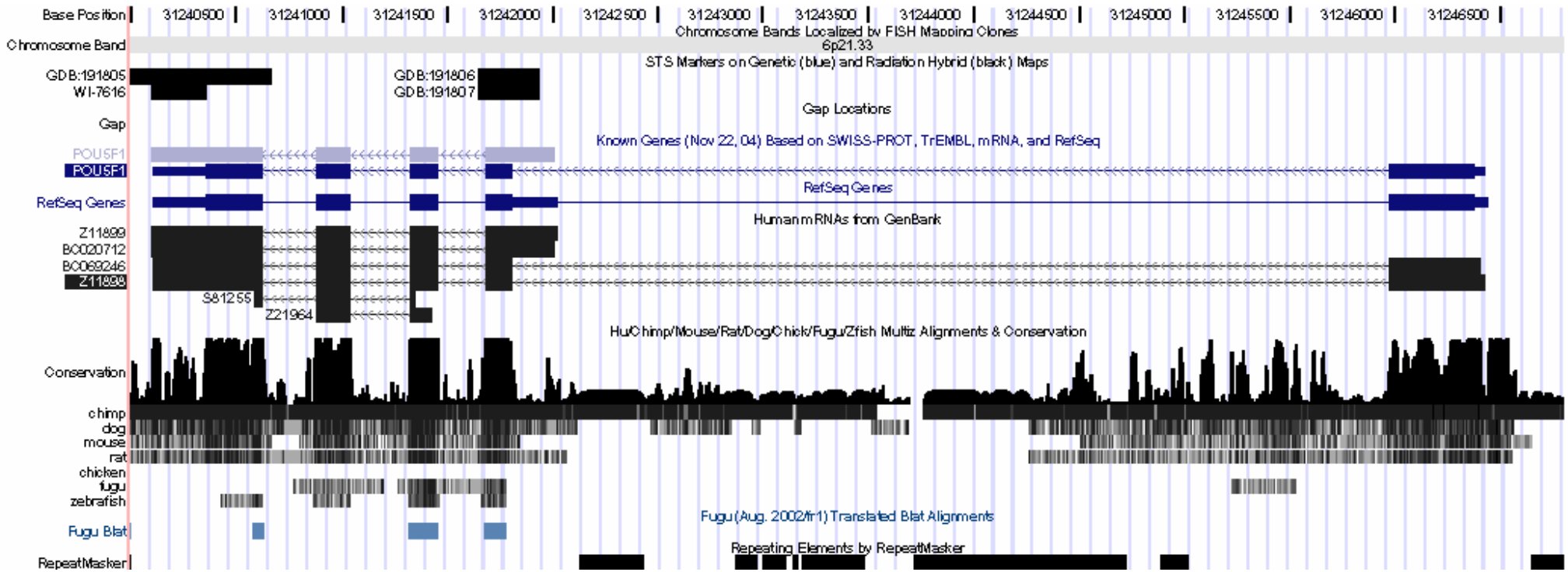
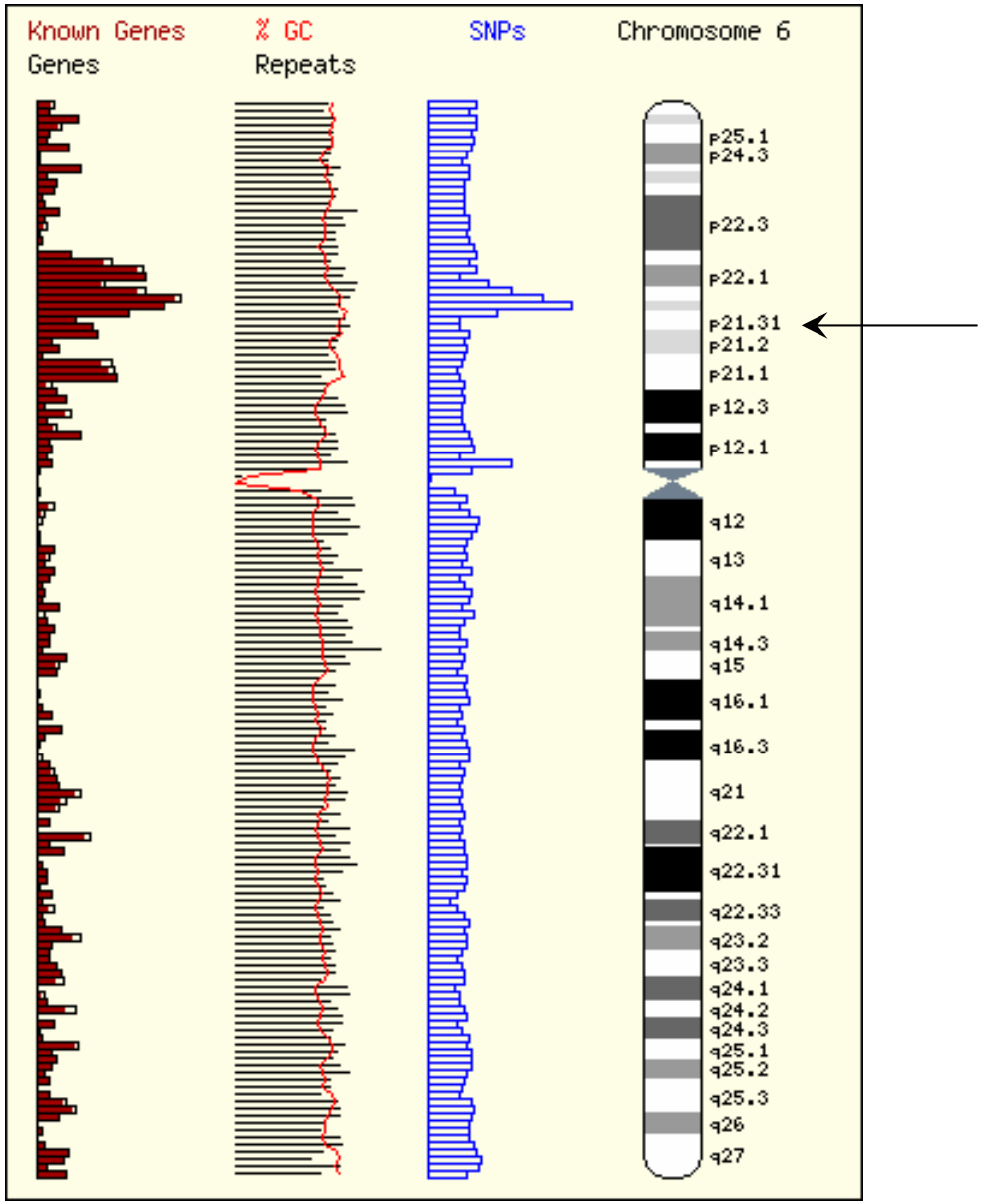


Image courtesy of the UCSC Genome Browser: http://genome.ucsc.edu/.


Known Genes
Genes
% GC
Repeats
SNPs
Chromosome 6
p25.1
p24.3
p22.3
p22.1
p21.31
p21.2
p21.1
p12.3
p12.1
q12
q13
q14.1
q14.3
q15
q16.1
q16.3
q21
q22.1
q22.31
q22.33
q23.2
q23.3
q24.1
q24.2
q24.3
q25.1
q25.2
q25.3
q26
q27

Oct4 Gene Locus (chr6)

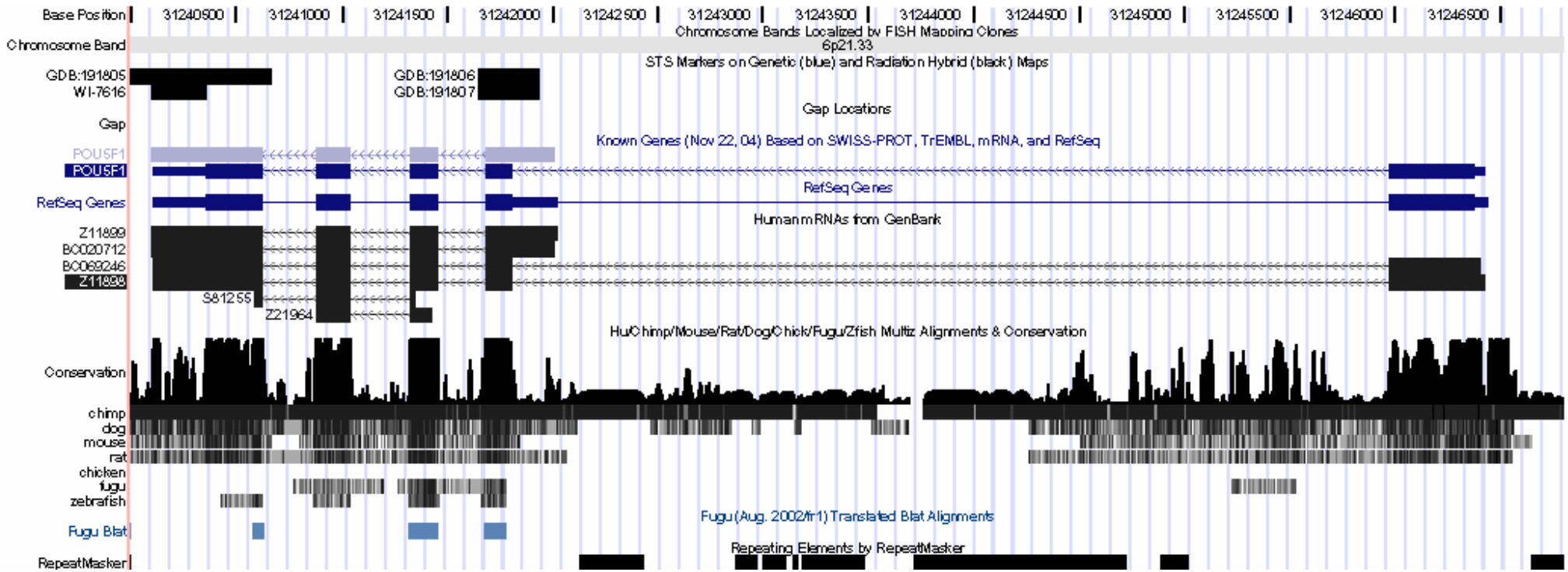


Image courtesy of the UCSC Genome Browser: http://genome.ucsc.edu/.

GDB:191805

Organism: Homo sapiens

Start: 31239867 **End:** 31240668

STS Marker GDB:191805

Chromosome: chr6 **Band:** 6p21.33

Left Primer: AGCTCATTGTCTAATGTCAT

Right Primer: CAGCTACATGGTGACTGAGT

Oct4 Gene Locus (chr6)

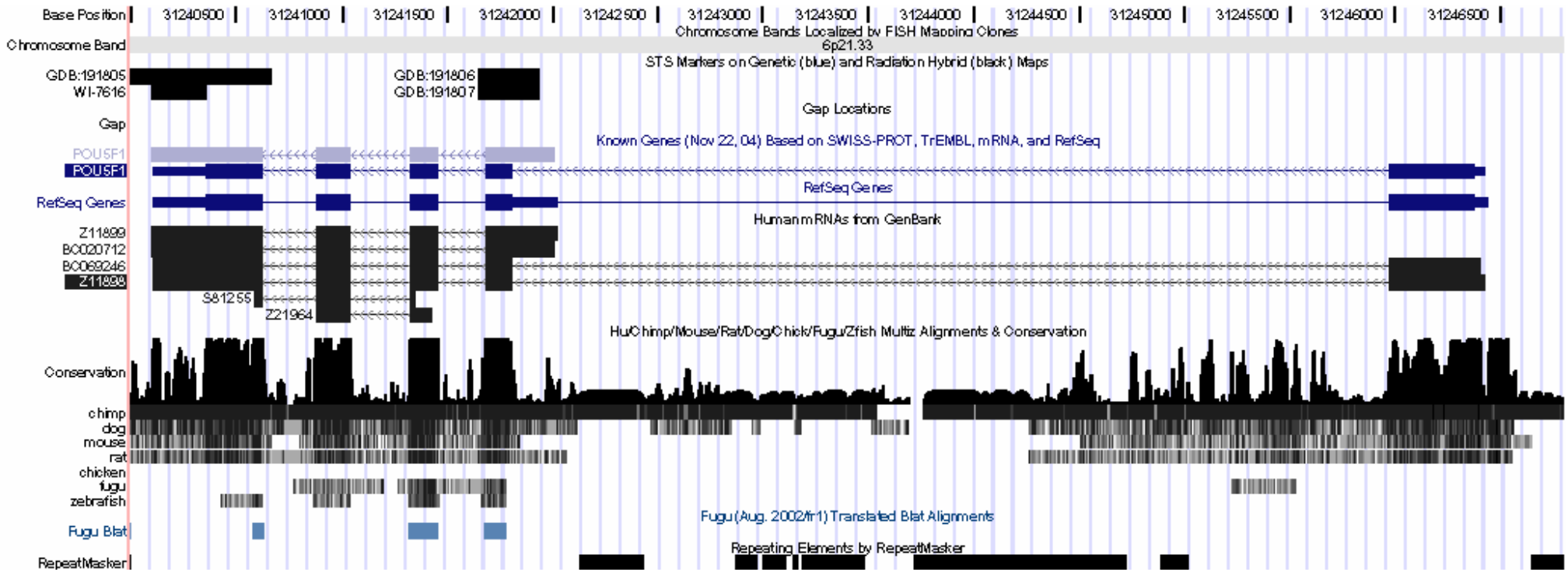


Image courtesy of the UCSC Genome Browser: http://genome.ucsc.edu/.

Repeat Masker

- Screens for interspersed repeats and low complexity DNA
- Interspersed repeats (Repeat database)
  - Primate database has 563 repeats comprising 664160 bp
  - Short Interspersed elements – SINEs (~80 bp Mariner, ~280 bp ALU)
  - Long interspersed elements – LINEs (6 – 8 Kb)
  - Transposable elements with long terminal repeats – LTRs (1.5 – 10 Kb)
- Low complexity DNA
  - 100 bp stretch >87% AT or >89% GC
  - 30 bp stretch >29 A/T (or GC) bases


SNPs and Their Utility


ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATAT
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACAT
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATC
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACAT
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTA
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTA
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTA
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGT
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTA
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGT
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCA
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGA
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG


ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGC
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATAT
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACAT
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGT
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATC
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTA
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACAT
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTA
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTAC
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCC
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTA
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTA
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGT
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTA
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGT
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGA
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTG

Human Variation:
Mutation Rate in Population

There are about 3 million known SNPs in human (about 1 every 1000 bases)

Average mutation rate 10^-6 (expect 1 mutation in any gene in 10^6 gametes)

Most mutation occurs in males: mutation rate is 2X higher in males than in females (Why?)

Rate of spontaneous abortion? Perhaps 50-70%

Humans are small population, that grew large fast

Two diagrams removed for copyright reasons.

Structure and Function of the Genome

Chromosomes

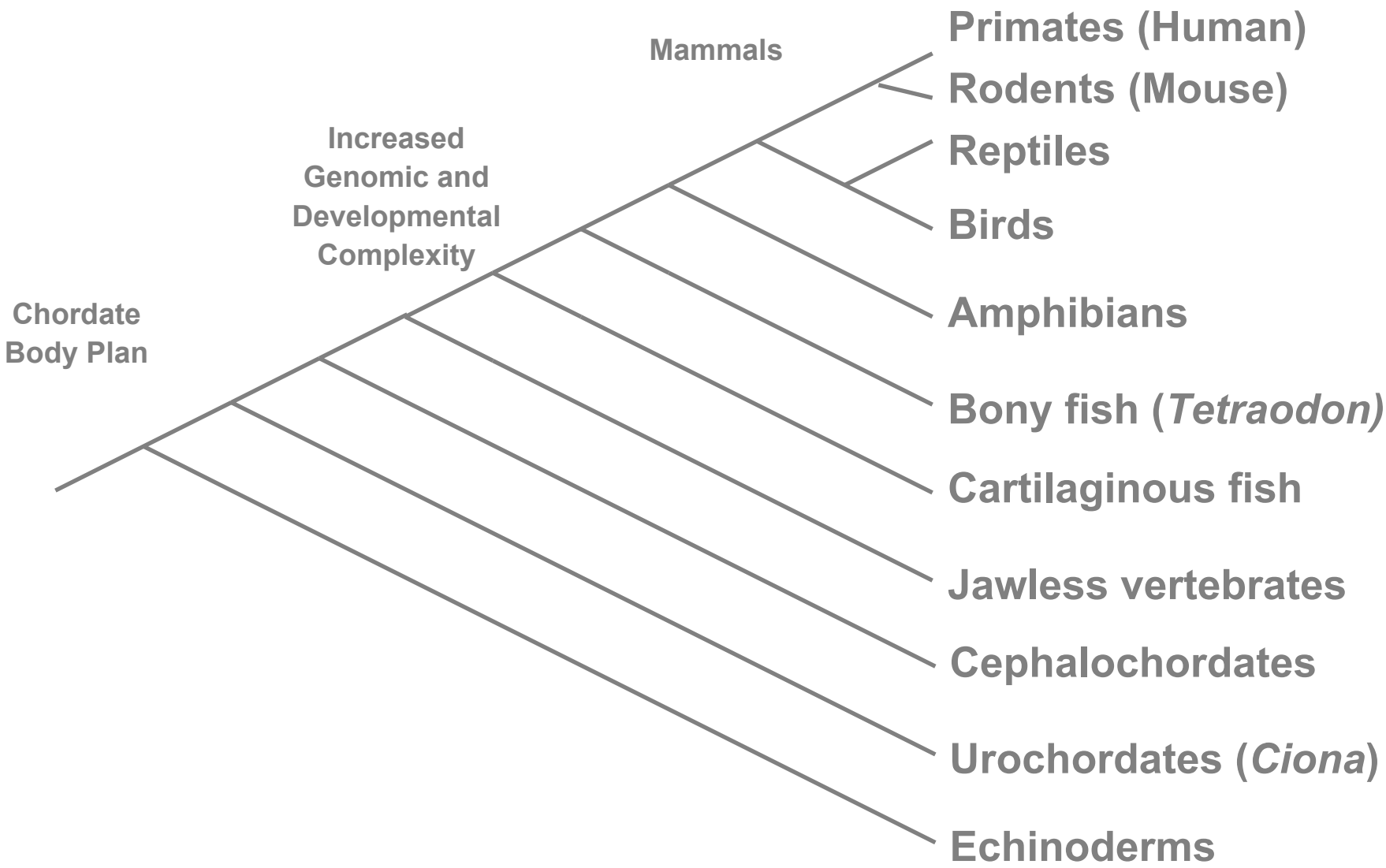
Human Genome

Comparative Genomics

Genes and their Products

Stem Cells

Insights into Evolution of Species



NCBI

Site Map
guide to NCBI resources

Cancer Chromosomes
chromosomal abnormalities

Clusters of Orthologous Groups
analysis of complete genomes

Gene
gene-related information

Genome
complete genome sequences

GEO
gene expression data

HomoloGene
orthologs between pairs of organisms

Map Viewer
map and genome displays

RefSeq
the reference

Genomic Biology

Genomic biology takes a holistic approach to molecular biology and evolution by studying the complete genome, its genes, and its protein expression patterns.

NCBI provides several genomic biology tools and resources, including organism-specific pages that include links to many web sites and databases relevant to that species. We invite you to explore the links provided on this page.

New Announcements

Entrez Genome Project

The Entrez Genome Project database is a searchable collection of large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms with organism-specific overviews functioning as portals for browsing and retrieval.

January 14, 2005
NCBI's annotation of the dog (*Canis familiaris*) genome assembly (build 1.1) is now available in the Map Viewer.

Genome Resources

- Aspergillus
- Bee
- Cat
- Chicken
- Chimp NEW
- Cow
- Dictyostelium
- Dog
- Frog
- Fruit Fly NEW
- Human
- Malaria
- Microbes
- Mosquito
- Mouse
- Nematode NEW
- Organelles
- Pig
- Plant Genomes
- Rat
- Retroviruses
- Sea Urchin
- Sheep
- Viral Genomes
- Zebrafish

Subscriptions

Comparative Genomics

	Genome size (MB)	Est #Genes
Human	3000	30,000
Mouse	3000	30,000
D. melanogaster	180	13,000
A. thaliana	100	25,000
C. elegans	97	19,000
S. cerevisiae	12	6,000
H. influenzae	1.8	1,700

Comparison of genome sequences can reveal sequence features that are conserved

Such sequences may be conserved due to constraints on function

-For example, protein coding sequences

Human-Mouse Comparative Analysis (I)

Photo of mouse removed for copyright reasons.

Mouse is 14% smaller, probably reflecting higher rate of deletion in mouse lineage

>90% of mouse and human genomes can be partitioned into corresponding regions of conserved synteny, reflecting segments in which the gene order in the most recent common ancestor has been conserved in both species

~30,000 protein-coding genes

Proportion of mouse genes without any homologue in human genome (and vice versa) is less than 1%

Human-Mouse Comparative Analysis (II)

Photo of mouse removed for copyright reasons.

Dozens of local gene family expansions have occurred in the mouse lineage

Most of these involve genes related to reproduction, immunity and olfaction, suggesting that these physiological systems have been the focus of extensive lineage-specific innovation in rodents

Human-Mouse Comparative Analysis (III)

Photo of mouse removed for copyright reasons.

~5% of genome sequences are conserved,
much more than can be explained by protein-coding sequences alone

These conserved sequences may be:
- binding sites for regulatory proteins
- genes for non-coding RNAs
- other?

Structure and Function of the Genome

Chromosomes

Human Genome

Comparative Genomics

Genes and their Products

Stem Cells

How can only ~30,000 genes specify a complex mammal?

Photos removed for copyright reasons.

Non-coding RNA Genes

Diagram removed for copyright reasons.
Schematic of replication, transcription and translation processes.

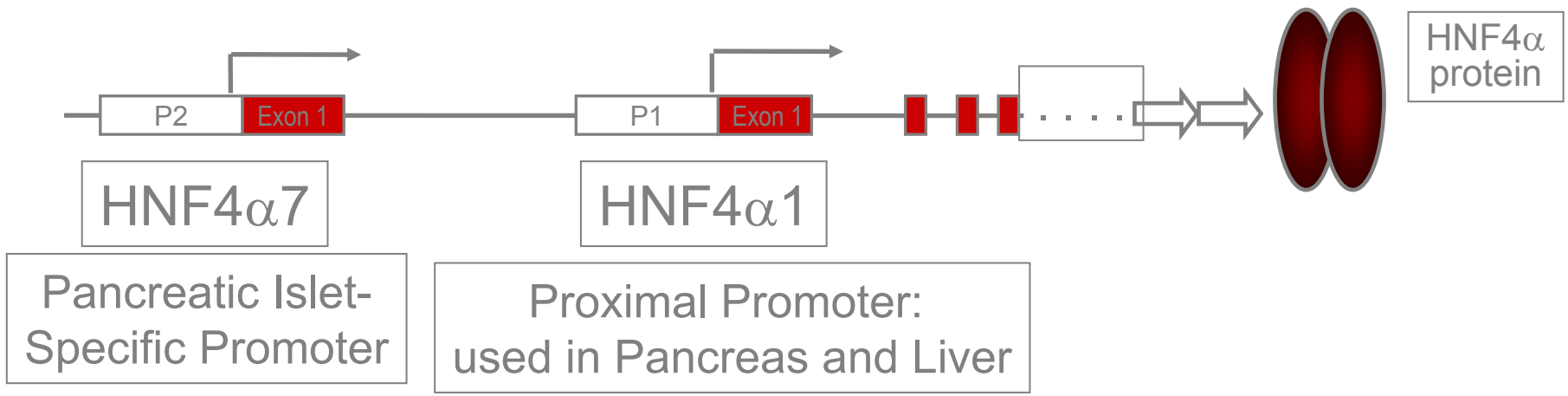
Genes and their Products

Gene Class	# Humans	Transcription Apparatus
Ribosomal RNA	~200	RNA Polymerase I
Protein-coding	~30,000	RNA Polymerase II
ncRNA	?	RNA polymerases II and III

Non-coding RNA Genes

Gene Class	Function
rRNA	Structural and functional component of ribosome
tRNA	Translational adapter
miRNA	Translational inhibition
snRNA	RNA splicing apparatus
snoRNA	rRNA processing
Large ncRNA	Gene regulation (XIST and X chromosome inactivation)

Tissue-specific Gene Transcription: HNF4a Gene Expression in Pancreas and Liver



Thomas et al. *Human Mol Gen* **10** (2001) 2089

Alternative Splicing Occurs Frequently with Human Genes

Diagram removed for copyright reasons.

Structure and Function of the Genome

Chromosomes

Human Genome

Comparative Genomics

Genes and their Products

Stem Cells

V. Hacker, 1895

Diagram removed for copyright reasons.

stammzelle

2 Kinds of Stem Cells

Embryonic

- most primitive
- can form all cell types
- immortal in culture
- plentiful

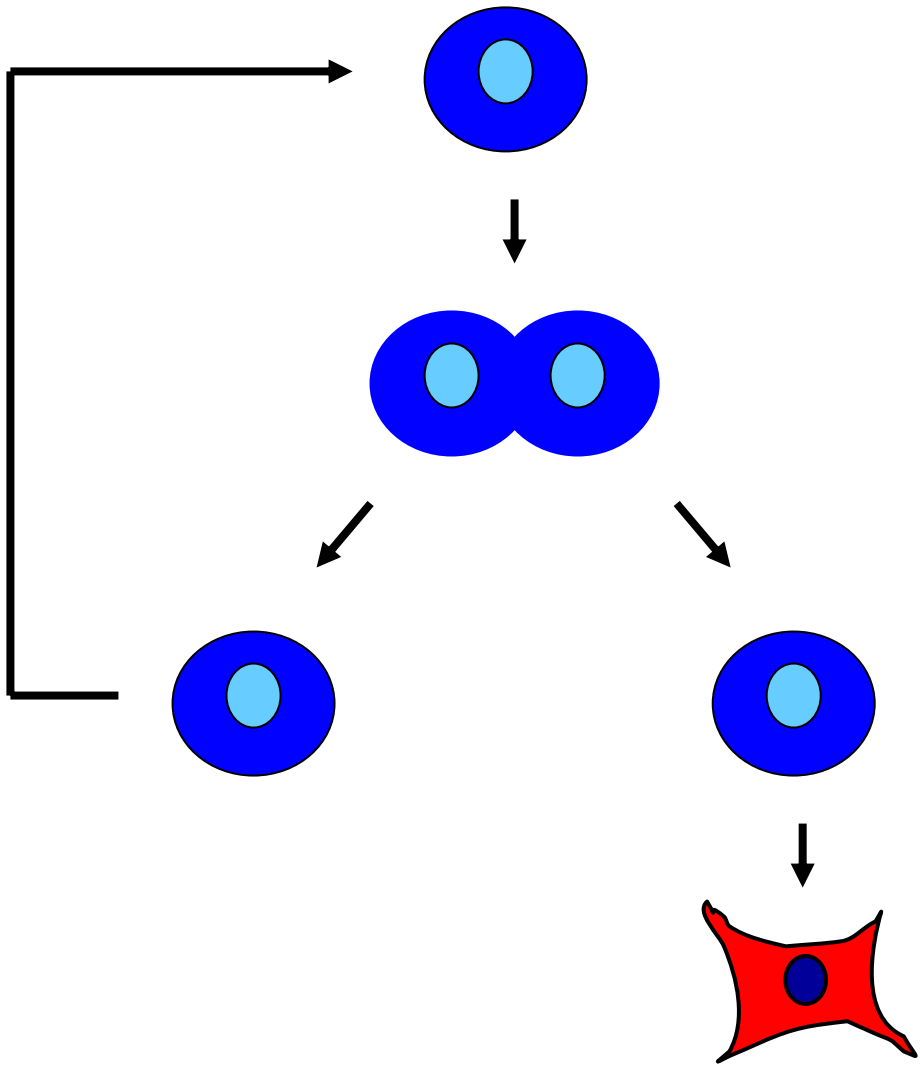
Adult

- organ specific
- can form few cell types
- limited lifespan
- hard to isolate

From blastocysts to human ES cells

Three pairs of photos removed for copyright reasons.

Stem Cells



self-renewal

differentiated cell

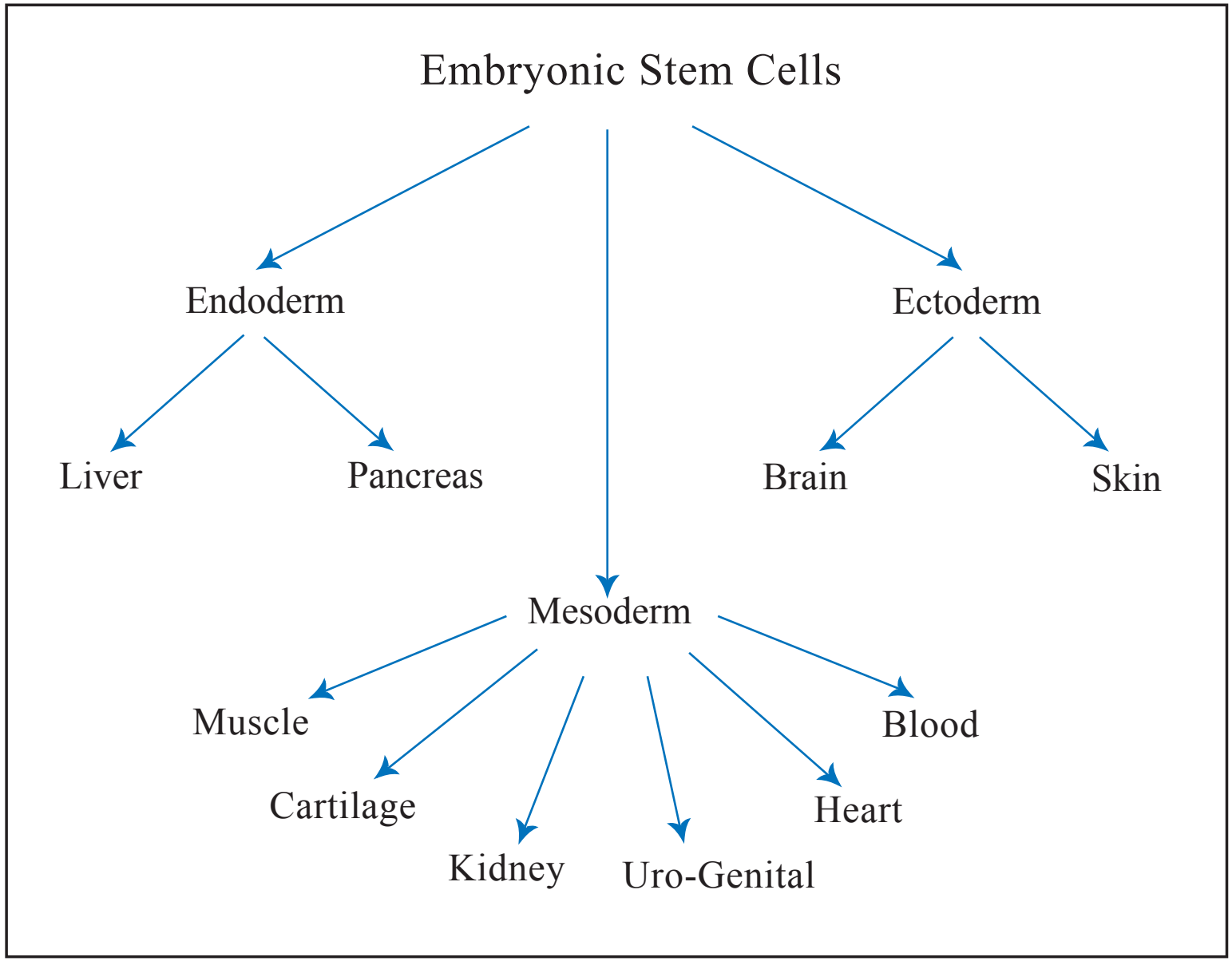
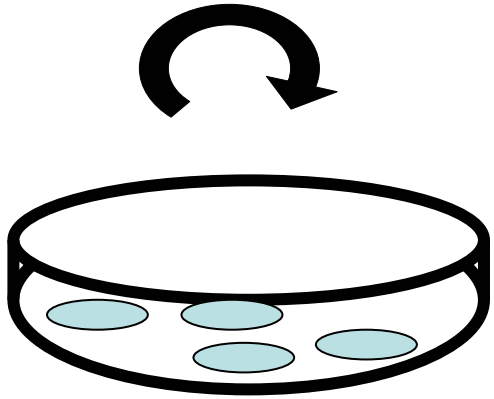
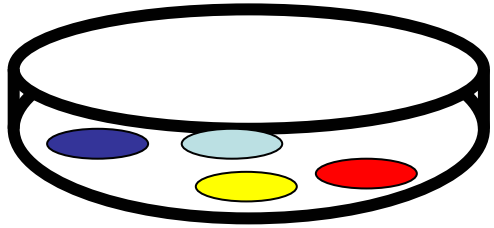


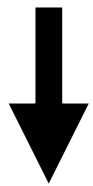
Figure by MIT OCW.





Human ES cells

Specialized cells



Human Heart cells

Photo removed for copyright reasons.

Terminology

stem cell

q arm

STS Marker

ribosome

mitochondria

synteny