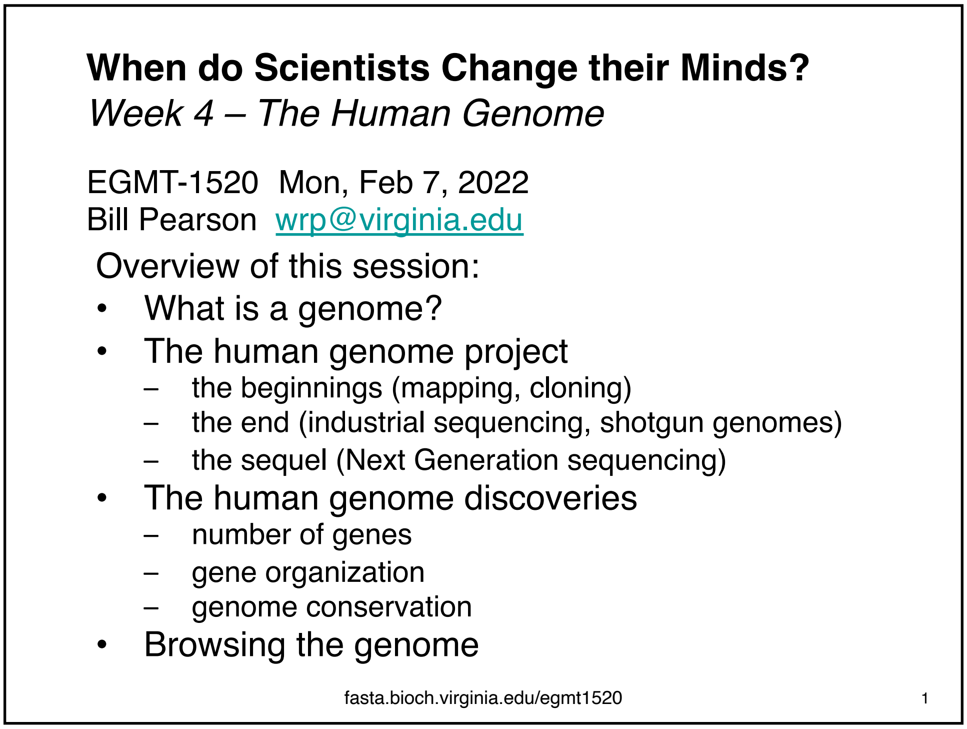


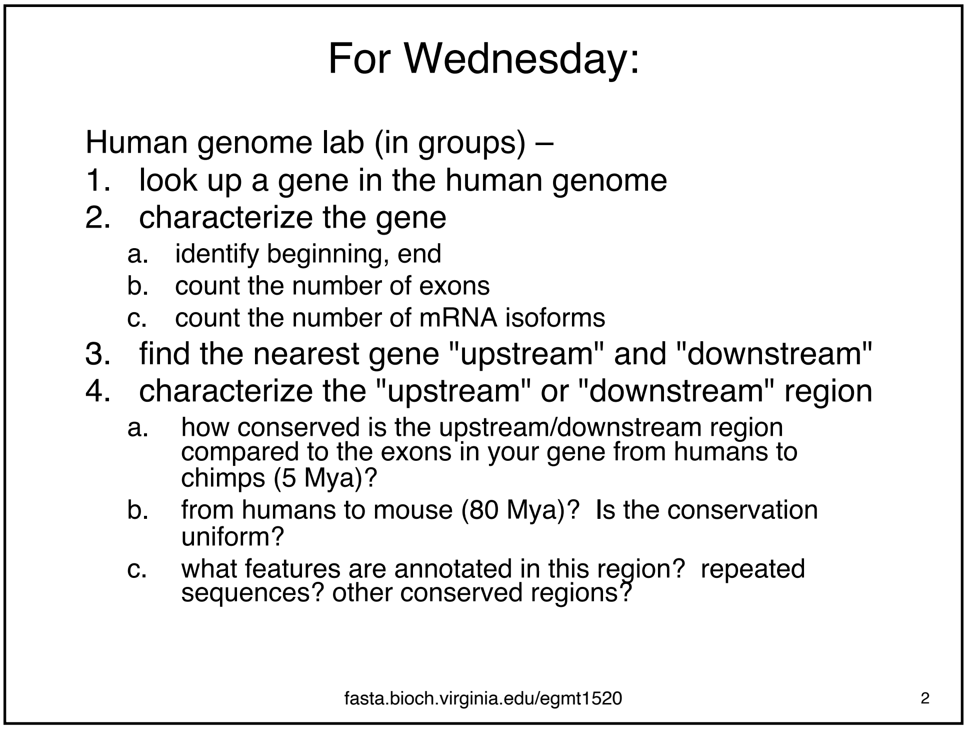
When do Scientists Change their Minds?

Week 4 – The Human Genome

EGMT-1520 Mon, Feb 7, 2022
Bill Pearson wrp@virginia.edu

Overview of this session:

- What is a genome?
- The human genome project
  - the beginnings (mapping, cloning)
  - the end (industrial sequencing, shotgun genomes)
  - the sequel (Next Generation sequencing)
- The human genome discoveries
  - number of genes
  - gene organization
  - genome conservation
- Browsing the genome



For Wednesday:

Human genome lab (in groups) –

1. look up a gene in the human genome
2. characterize the gene
   a. identify beginning, end
   b. count the number of exons
   c. count the number of mRNA isoforms
3. find the nearest gene "upstream" and "downstream"
4. characterize the "upstream" or "downstream" region
   a. how conserved is the upstream/downstream region compared to the exons in your gene from humans to chimps (5 Mya)?
   b. from humans to mouse (80 Mya)? Is the conservation uniform?
   c. what features are annotated in this region? repeated sequences? other conserved regions?

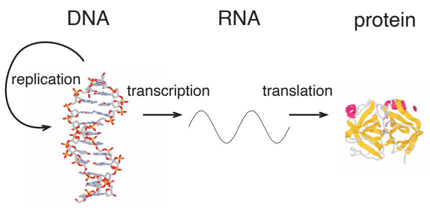
For Monday:

Repeat the human genome lab on a different gene.

1. report the name of the gene, and its chromosome location. Submit the URL of the UCSC genome browser page that shows the gene.
2. characterize the gene
   a. report the the length of the gene
   b. Is the gene on the forward or reverse strand?
   c. report the number of exons
3. report the name and coordinates of the nearest gene "upstream" and "downstream"
   a. Determine whether the gene is on the same strand, (forward/reverse) or on the opposite strand.

What does a gene look like?

- Central dogma:



DNA

RNA

protein

replication

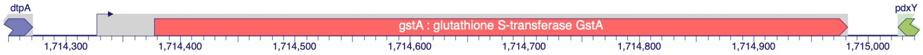
transcription

translation

- Parts of a gene:
  - start of transcription (beginning of the mRNA)
  - start of translation (beginning of the protein)
  - end of translation (end of the protein)
  - poly-A addition site (end of processed mRNA)
  - end of transcription

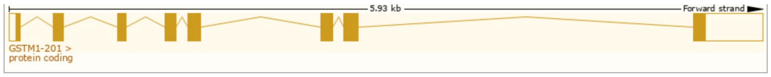
Prokaryotes (bacteria, archea) vs Eukaryotes

- Parts of a gene (prokaryotes, eukaryotes):
  - start of transcription (beginning of the mRNA)
  - start of translation (beginning of the protein)
  - end of translation (end of the protein)
  - poly-A addition site (end of processed mRNA)
  - end of transcription



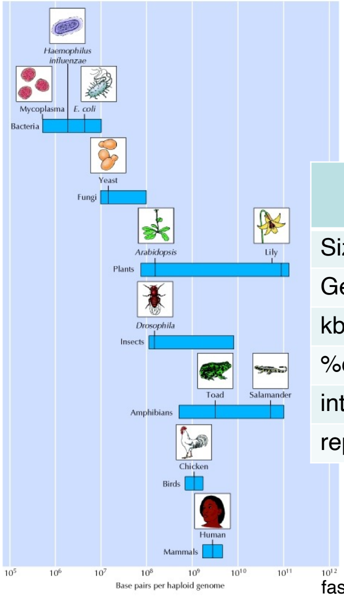
606 nt = 202 amino-acids

- Parts of a gene (eukaryotes):
  - introns and exons



gene: 5950 nt, mRNA: 1164 nt, protein: 218 aa

What is in a genome?

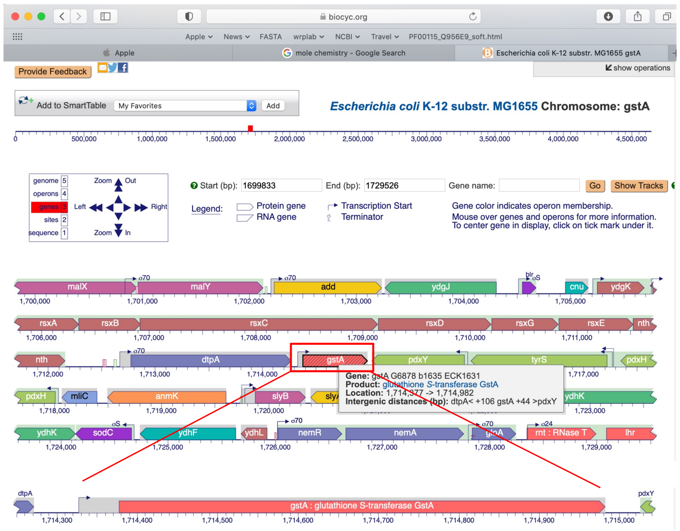


	E. col	Plas.	Yeast	Plant (ARATH)	Homo
Size(Mb)	4.64	22.8	12.5	115	3289
Genes	4288	5268	5770	25.5K	~25K
kb/Gene	0.95	4.34	2.09	4.53	27
%coding	87.8	52.6	70.5	28.8	1.3
introns	0	7406	272	107K	53K
repeat%	<1	<1	2.4	15	46

Pevsner, Table 16-1

Cooper, GM (2000) The Cell: A Molecular Approach. 2nd edition. Fig 4.1

Gene and genome complexity – E. coli
(4288 genes, 4.64 Mbase, 87.8% protein coding)



Provide Feedback

Add to SmartTable My Favorites Add

Escherichia coli K-12 substr. MG1655 Chromosome: gstA

Start (bp): 1699833 End (bp): 1729526 Gene name: Go Show Tracks

Legend: Protein gene, RNA gene, Transcription Start, Terminator

Gene color indicates operon membership.
Mouse over genes and operons for more information.
To center gene in display, click on tick mark under it.

Gene: gstA G6878 b1635 ECK1631
Product: glutathione *S*-transferase GstA
Location: 1,714,377 -> 1,714,982
Intergenic distances (bp): dtpA< +106 gstA +44 >pdxY

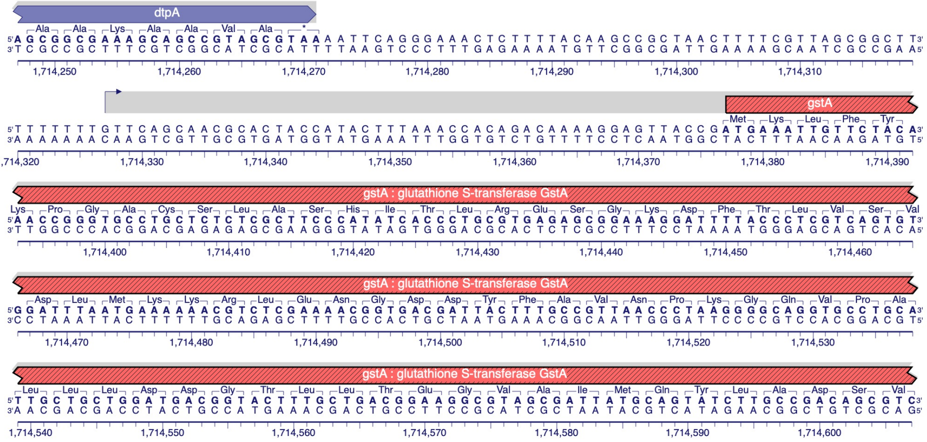
biocyc.org

fasta.bioch.virginia.edu/egmt1520

7

7

Gene and genome complexity – E. coli
(4288 genes, 4.64 Mbase, 87.8% protein coding)

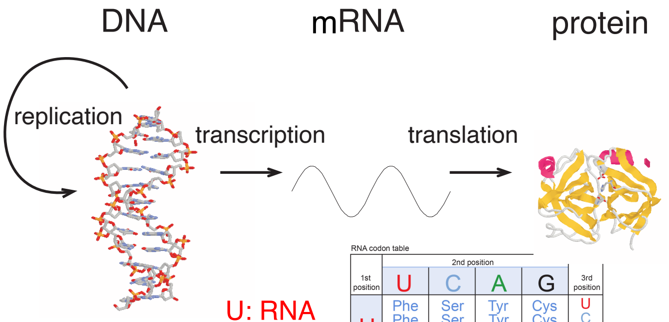

dtpA
Ala Ala Lys Ala Ala Val Ala *
5'A G C G G C G A A A G C A G C C G T A G C G T A A A A T T C A G G G A A A C T C T T T T A C A A G C C G C T A A C T T T T C G T T A G C G G C T T3'
3'T C G C C G C T T T C G T C G G C A T C G C A T T T T A A G T C C C T T T G A G A A A A T G T T C G G C G A T T G A A A A G C A A T C G C C G A A5'
1,714,250
1,714,260
1,714,270
1,714,280
1,714,290
1,714,300
1,714,310
gstA
Met Lys Leu Phe Tyr
5'T T T T T T T G T T C A G C A A C G C A C T A C C A T A C T T T A A A C C A C A G A C A A A A G G A G T T A C C G A T G A A A T T G T T C T A C A3'
3'A A A A A A A C A A G T C G T T G C G T G A T G G T A T G A A A T T T G G T G T C T G T T T T C C T C A A T G G C T A C T T T A A C A A G A T G T5'
1,714,320
1,714,330
1,714,340
1,714,350
1,714,360
1,714,370
1,714,380
1,714,390
gstA : glutathione S-transferase GstA
Lys Pro Gly Ala Cys Ser Leu Ala Ser His Ile Thr Leu Arg Glu Ser Gly Lys Asp Phe Thr Leu Val Ser Val
5'A A C C G G G T G C C T G C T C T C T C G C T T C C C A T A T C A C C C T G C G T G A G A G C G G A A A G G A T T T T A C C C T C G T C A G T G T3'
3'T T G G C C C A C G G A C G A G A G A G C G A A G G G T A T A G T G G G A C G C A C T C T C G C C T T T C C T A A A A T G G G A G C A G T C A C A5'
1,714,400
1,714,410
1,714,420
1,714,430
1,714,440
1,714,450
1,714,460
gstA : glutathione S-transferase GstA
Asp Leu Met Lys Lys Arg Leu Glu Asn Gly Asp Asp Tyr Phe Ala Val Asn Pro Lys Gly Gln Val Pro Ala
5'G G A T T T A A T G A A A A A A C G T C T C G A A A A C G G T G A C G A T T A C T T T G C C G T T A A C C C T A A G G G G C A G G T G C C T G C A3'
3'C C T A A A T T A C T T T T T T G C A G A G C T T T T G C C A C T G C T A A T G A A A C G G C A A T T G G G A T T C C C C G T C C A C G G A C G T5'
1,714,470
1,714,480
1,714,490
1,714,500
1,714,510
1,714,520
1,714,530
gstA : glutathione S-transferase GstA
Leu Leu Leu Asp Asp Gly Thr Leu Leu Thr Glu Gly Val Ala Ile Met Gln Tyr Leu Ala Asp Ser Val
5'T T G C T G C T G G A T G A C G G T A C T T T G C T G A C G G A A G G C G T A G C G A T T A T G C A G T A T C T T G C C G A C A G C G T C3'
3'A A C G A C G A C C T A C T G C C A T G A A A C G A C T G C C T T C C G C A T C G C T A A T A C G T C A T A G A A C G G C T G T C G C A G5'
1,714,540
1,714,550
1,714,560
1,714,570
1,714,580
1,714,590
1,714,600

fasta.bioch.virginia.edu/egmt1520

8

8

Central dogma



U: RNA
T: DNA

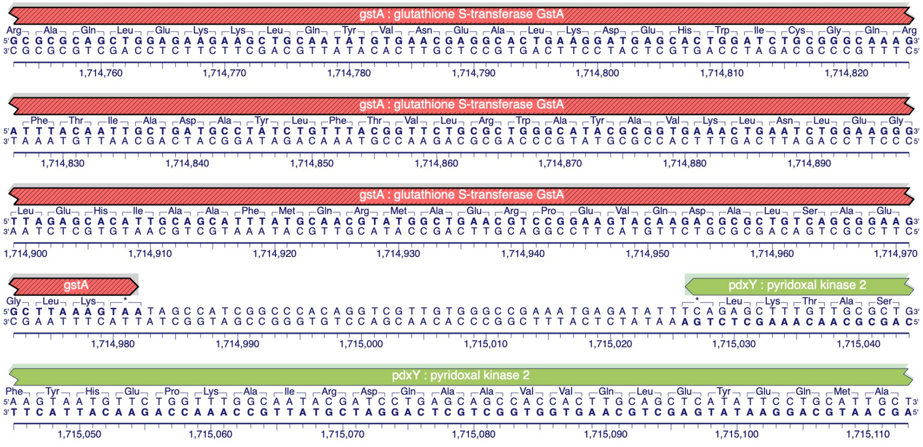
RNA codon table

1st position	2nd position U	C	A	G	3rd position
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr stop stop	Cys Cys stop Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G
	Amino Acids				

Ala: Alanine	Gln: Glutamine	Leu: Leucine	Ser: Serine
Arg: Arginine	Glu: Glutamic acid	Lys: Lysine	Thr: Threonine
Asn: Asparagine	Gly: Glycine	Met: Methionine	Trp: Tryptophane
Asp: Aspartic acid	His: Histidine	Phe: Phenylalanine	Tyr: Tyrosisne
Cys:Cysteine	Ile: Isoleucine	Pro: Proline	Val: Valine

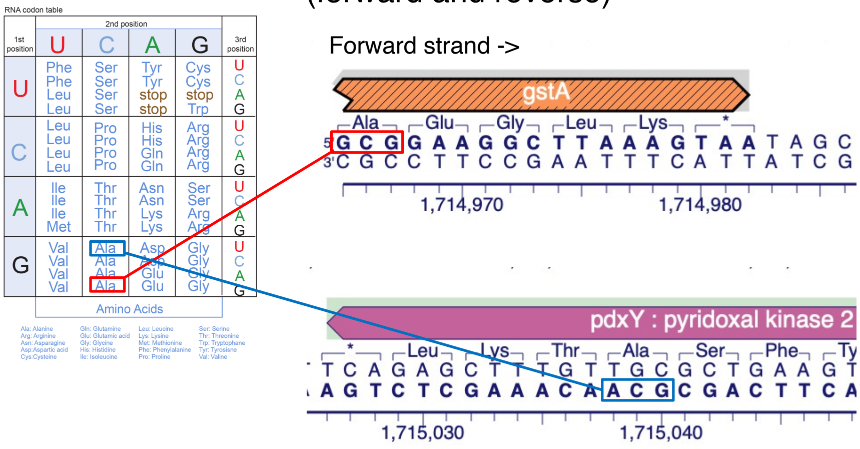
fasta.bioch.virginia.edu/egmt1520

Gene and genome complexity – E. coli (4288 genes, 4.64 Mbase, 87.8% protein coding)



fasta.bioch.virginia.edu/egmt1520

From gene to protein: translation and the genetic code (forward and reverse)



RNA codon table

1st position	U	C	A	G	3rd position
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr stop stop	Cys Cys stop Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

2nd position; Amino Acids

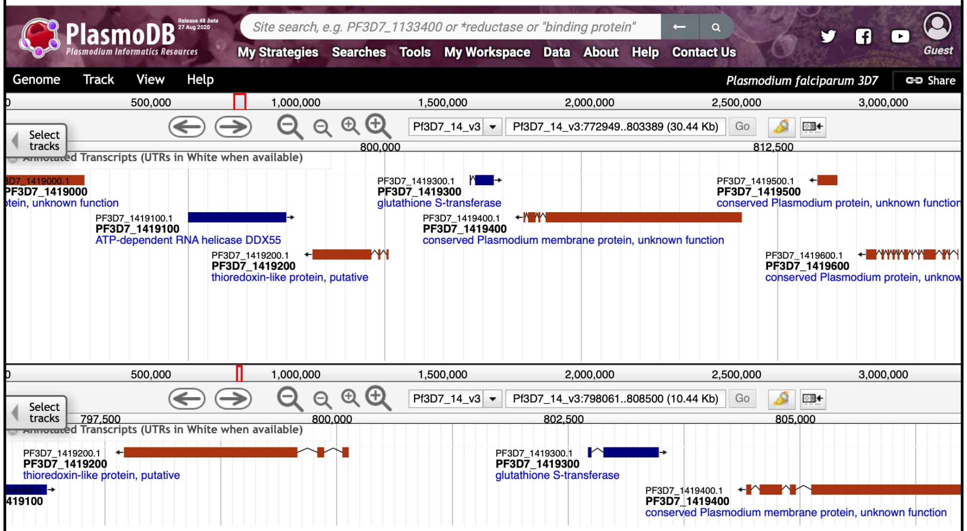
Ala: Alanine	Gln: Glutamine	Leu: Leucine	Ser: Serine
Arg: Arginine	Glu: Glutamic acid	Lys: Lysine	Thr: Threonine
Asn: Asparagine	Gly: Glycine	Met: Methionine	Trp: Tryptophane
Asp:Aspartic acid	His: Histidine	Phe: Phenylalanine	Tyr: Tyrosisne
Cys:Cysteine	Ile: Isoleucine	Pro: Proline	Val: Valine

Forward strand ->
gstA
Ala Glu Gly Leu Lys *
5'G C G G A A G G C T T A A A G T A A T A G C
3'C G C C T T C C G A A T T T C A T T A T C G
1,714,970
1,714,980
pdxY : pyridoxal kinase 2
* Leu Lys Thr Ala Ser Phe Ty
T C A G A G C T T T T G T T G C G C T G A A G T
A G T C T C G A A A A C A A C G C G A C T T C A
1,715,030
1,715,040

<- Reverse strand

fasta.bioch.virginia.edu/egmt1520

Gene and genome complexity – Plasmodium (5268 genes, 22.8 Mbase, 52.8% protein coding)



PlasmoDB
Plasmodium Informatics Resources
Release 48 beta 27 Aug 2020
Site search, e.g. PF3D7_1133400 or *reductase or "binding protein"
My Strategies
Searches
Tools
My Workspace
Data
About
Help
Contact Us
Guest
Genome
Track
View
Help
Plasmodium falciparum 3D7
Share
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
Select tracks
Pf3D7_14_v3
Pf3D7_14_v3:772949..803389 (30.44 Kb)
Go
800,000
812,500
Annotated Transcripts (UTRs in White when available)

3D7_1419000.1
PF3D7_1419000
tein, unknown function

PF3D7_1419100.1
PF3D7_1419100
ATP-dependent RNA helicase DDX55

PF3D7_1419200.1
PF3D7_1419200
thioredoxin-like protein, putative

PF3D7_1419300.1
PF3D7_1419300
glutathione S-transferase

PF3D7_1419400.1
PF3D7_1419400
conserved Plasmodium membrane protein, unknown function

PF3D7_1419500.1
PF3D7_1419500
conserved Plasmodium protein, unknown function

PF3D7_1419600.1
PF3D7_1419600
conserved Plasmodium protein, unknow

PF3D7_1419200.1
PF3D7_1419200
thioredoxin-like protein, putative

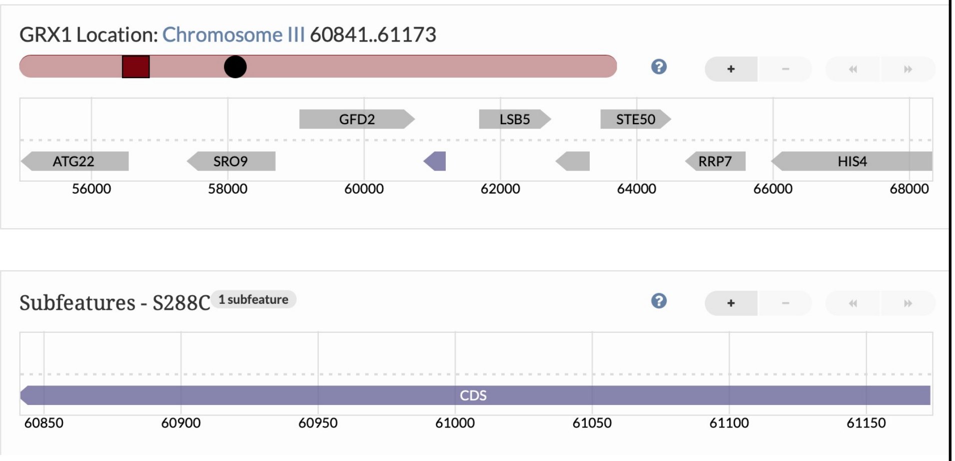
419100

PF3D7_1419300.1
PF3D7_1419300
glutathione S-transferase

PF3D7_1419400.1
PF3D7_1419400
conserved Plasmodium membrane protein, unknown function

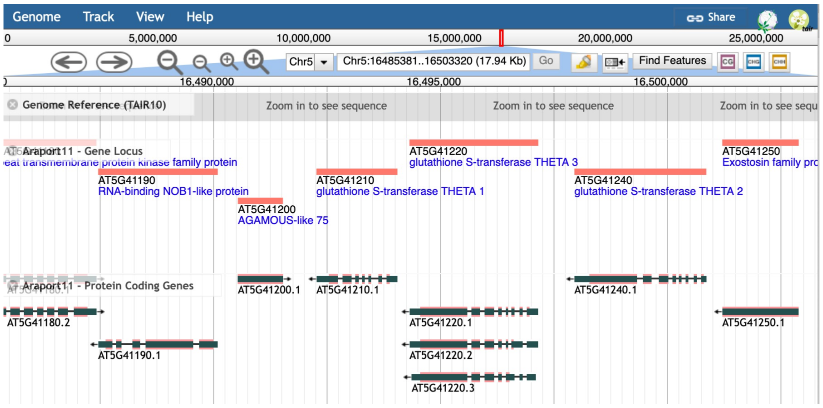
fasta.bioch.virginia.edu/egmt1520

Gene and genome complexity – Yeast (5770 genes, 12.5 Mbase, 70.5% protein coding)



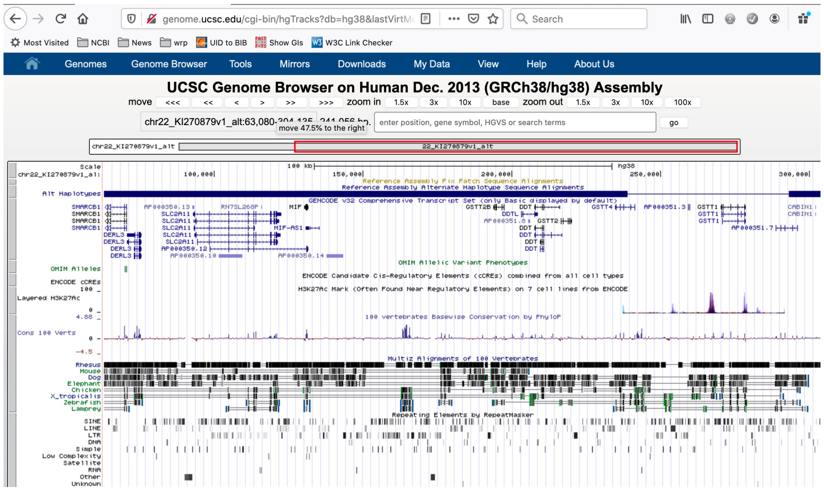
fasta.bioch.virginia.edu/egmt1520

Gene and genome complexity – Arabidopsis (mouse ear cress) (22,500 genes, 115 Mbase, 28 % protein coding)



fasta.bioch.virginia.edu/egmt1520

Gene and genome complexity – Human (~20,000 genes, 3,000 Mbase, 1.5% protein coding)

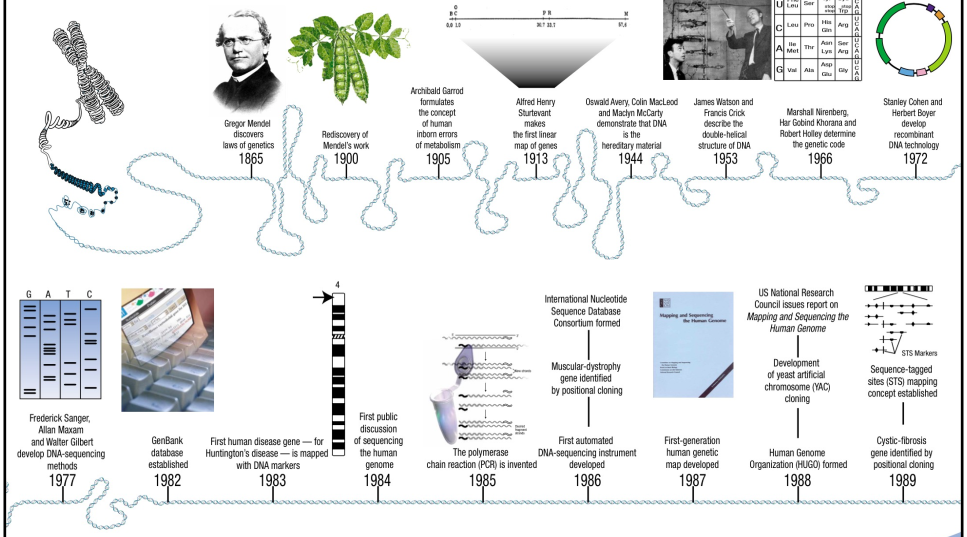


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr22_KI270879v1_alt:63,080-304,135 241,056 bp.

move 47.5% to the right

Landmarks in genetics and genomics



Gregor Mendel discovers laws of genetics 1865

Rediscovery of Mendel's work 1900

Archibald Garrod formulates the concept of human inborn errors of metabolism 1905

Alfred Henry Sturtevant makes the first linear map of genes 1913

Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material 1944

James Watson and Francis Crick describe the double-helical structure of DNA 1953

Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code 1966

Stanley Cohen and Herbert Boyer develop recombinant DNA technology 1972

Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods 1977

GenBank database established 1982

First human disease gene — for Huntington's disease — is mapped with DNA markers 1983

First public discussion of sequencing the human genome 1984

The polymerase chain reaction (PCR) is invented 1985

International Nucleotide Sequence Database Consortium formed

Muscular-dystrophy gene identified by positional cloning

First automated DNA-sequencing instrument developed 1986

First-generation human genetic map developed 1987

US National Research Council issues report on *Mapping and Sequencing the Human Genome*

Development of yeast artificial chromosome (YAC) cloning

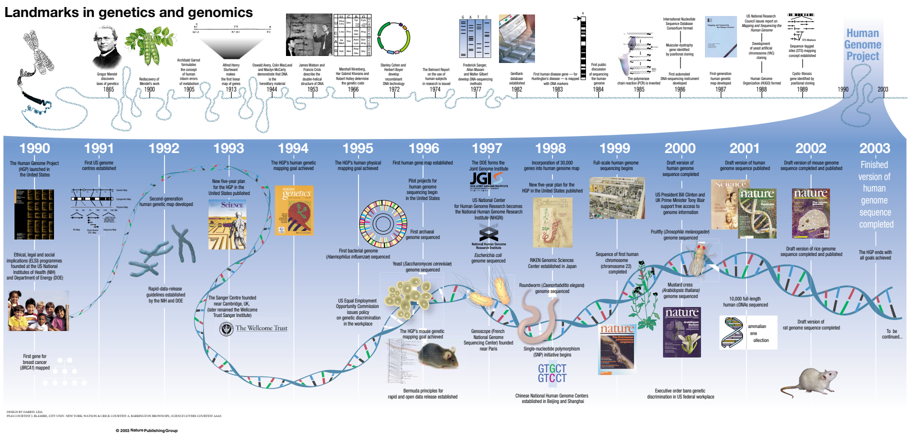
Human Genome Organization (HUGO) formed 1988

Sequence-tagged sites (STS) mapping concept established

Cystic-fibrosis gene identified by positional cloning 1989

Collins, Nature (2003) 422:835

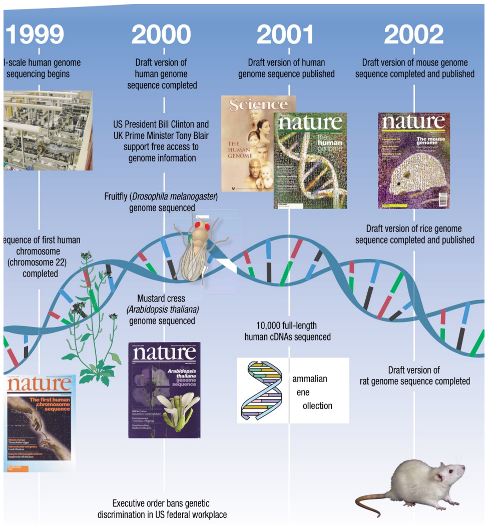
A history of genomes


Landmarks in genetics and genomics
Gregor Mendel discovers laws of genetics 1865
Rediscovery of Mendel's work 1900
Archibald Garrod formulates the concept of human inborn errors of metabolism 1905
Alfred Henry Sturtevant makes the first linear map of genes 1913
Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material 1944
James Watson and Francis Crick describe the double-helical structure of DNA 1953
Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code 1966
Stanley Cohen and Herbert Boyer develop recombinant DNA technology 1972
The Belmont Report on the use of human subjects in research is issued 1974
Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods 1977
GenBank database established 1982
First human disease gene — for Huntington's disease — is mapped with DNA markers 1983
First public discussion of sequencing the human genome 1984
The polymerase chain reaction (PCR) is invented 1985
International Nucleotide Sequence Database Consortium formed
Muscular-dystrophy gene identified by positional cloning
First automated DNA-sequencing instrument developed 1986
First-generation human genetic map developed 1987
US National Research Council issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed 1988
Sequence-tagged sites (STS) mapping concept established
Cystic-fibrosis gene identified by positional cloning 1989
Human Genome Project 1990 2003
1990 The Human Genome Project (HGP) launched in the United States
Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)
First gene for breast cancer (BRCA1) mapped
1991 First US genome centres established
1992 Second-generation human genetic map developed
Rapid-data-release guidelines established by the NIH and DOE
1993 New five-year plan for the HGP in the United States published
The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)
The Wellcome Trust
1994 The HGP's human genetic mapping goal achieved
1995 The HGP's human physical mapping goal achieved
First bacterial genome (Haemophilus influenzae) sequenced
US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996 First human gene map established
Pilot projects for human genome sequencing begin in the United States
First archaeal genome sequenced
Yeast (Saccharomyces cerevisiae) genome sequenced
The HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997 The DOE forms the Joint Genome Institute
US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)
Escherichia coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded near Paris
1998 Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the United States published
RIKEN Genomic Sciences Center established in Japan
Roundworm (Caenorhabditis elegans) genome sequenced
Single-nucleotide polymorphism (SNP) initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers established in Beijing and Shanghai
1999 Full-scale human genome sequencing begins
Sequence of first human chromosome (chromosome 22) completed
2000 Draft version of human genome sequence completed
US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information
Fruitfly (Drosophila melanogaster) genome sequenced
Mustard cress (Arabidopsis thaliana) genome sequenced
Executive order bans genetic discrimination in US federal workplace
2001 Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced
ammalian ene ollection
2002 Draft version of mouse genome sequence completed and published
Draft version of rice genome sequence completed and published
Draft version of rat genome sequence completed
2003 Finished version of human genome sequence completed
The HGP ends with all goals achieved
To be continued...
DESIGN BY DARRYL LEJA
PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; SCIENCE COVERS COURTESY AAAS
© 2003 Nature Publishing Group

Collins, Nature (2003) 422:835

fasta.bioch.virginia.edu/egmt1520

The Human Genome Project


1999
l-scale human genome sequencing begins
equence of first human chromosome (chromosome 22) completed
2000
Draft version of human genome sequence completed
US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information
Fruitfly (Drosophila melanogaster) genome sequenced
Mustard cress (Arabidopsis thaliana) genome sequenced
Executive order bans genetic discrimination in US federal workplace
2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced
ammalian
ene
ollection
2002
Draft version of mouse genome sequence completed and published
Draft version of rice genome sequence completed and published
Draft version of rat genome sequence completed

Collins, Nature (2003) 422:835

fasta.bioch.virginia.edu/egmt1520

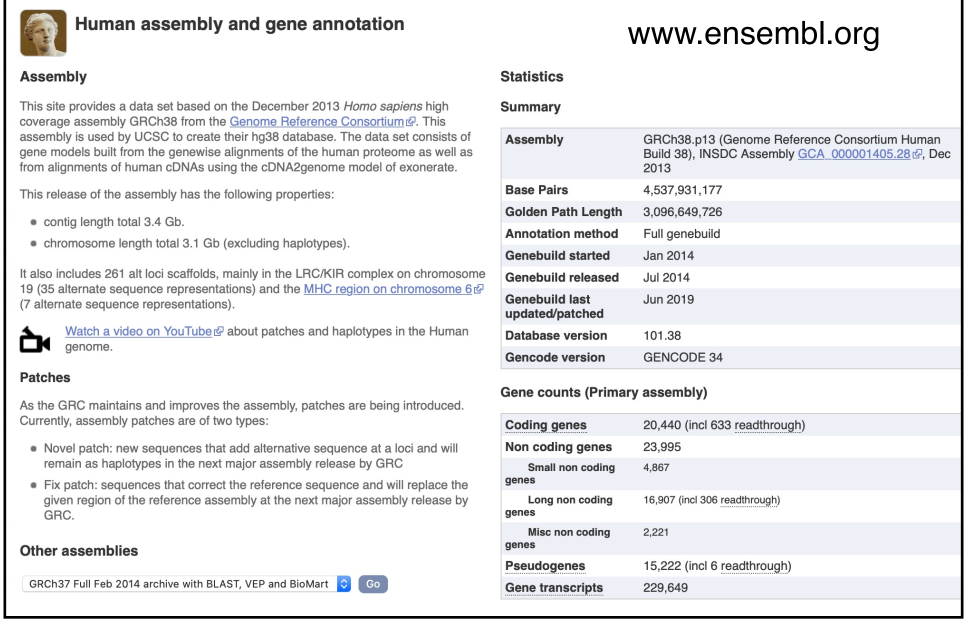
Sequencing capacity (2004)
40 million lanes/year ~ 8 billion bases – Whitehead Inst. (1 bacterial genome/day)
>40 billion bases/year, world-wide
(1000 bacterial genomes/year; 1 mammalian genome/year)

Sequencing capacity (2011) – Illumina sequencing 200 billion bases/week/machine, ~30,000 human genomes/year
Sequencing capacity (2015) – at least 300,000 human genomes/year

The human genome – initial insights

1. There were reported to be about 30,000 to 40,000 predicted protein-coding genes in the human genome. Currently, ENSEMBL reports 20,300 protein coding genes. Similar to Arabidopsis (plant, 26,000 genes) and pufferfish (21,000 genes), and marginally more genes than are found in many nematode and insect genomes (14,000).
2. More than 98% of the human genome does not code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs) (20%), short interspersed elements (SINEs) (13%), long terminal repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Thus half the human genome is derived from transposable elements.
3. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males.
4. More than 1.4 million single nucleotide polymorphisms (SNPs) were identified. SNPs are single nucleotide variations that occur once every 100 to 300 base pairs (bp). 36 million in Oct., 2014
5. Comparative sequencing of the mouse genome suggests that only about 5% of the human genome is under evolutionary selection.

The human genome
(GRCh38.p13 Dec. 2013)



www.ensembl.org

Human assembly and gene annotation

Assembly

This site provides a data set based on the December 2013 *Homo sapiens* high coverage assembly GRCh38 from the Genome Reference Consortium. This assembly is used by UCSC to create their hg38 database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb.
- chromosome length total 3.1 Gb (excluding haplotypes).

It also includes 261 alt loci scaffolds, mainly in the LRC/KIR complex on chromosome 19 (35 alternate sequence representations) and the MHC region on chromosome 6 (7 alternate sequence representations).

Watch a video on YouTube about patches and haplotypes in the Human genome.

Patches

As the GRC maintains and improves the assembly, patches are being introduced. Currently, assembly patches are of two types:

- Novel patch: new sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC
- Fix patch: sequences that correct the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC.

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

Statistics

Summary

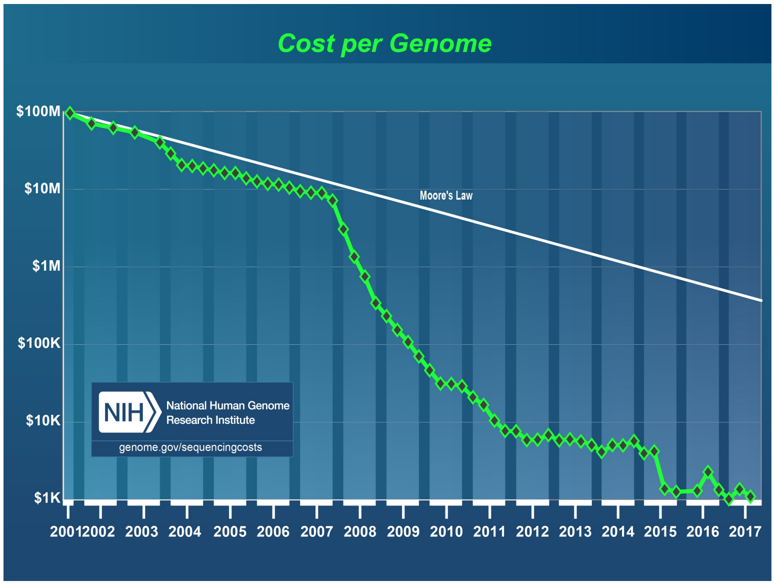
Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28, Dec 2013
Base Pairs	4,537,931,177
Golden Path Length	3,096,649,726
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2019
Database version	101.38
Gencode version	GENCODE 34

Gene counts (Primary assembly)

Coding genes	20,440 (incl 633 readthrough)
Non coding genes	23,995
Small non coding genes	4,867
Long non coding genes	16,907 (incl 306 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,222 (incl 6 readthrough)
Gene transcripts	229,649

21

Sequencing Costs

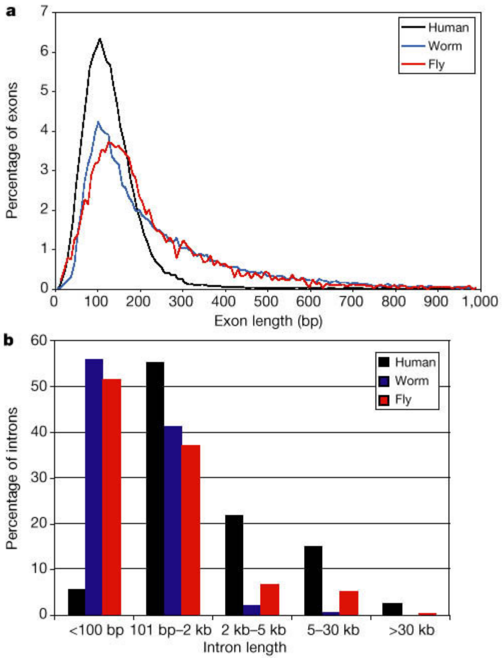


fasta.bioch.virginia.edu/egmt1520 22

22

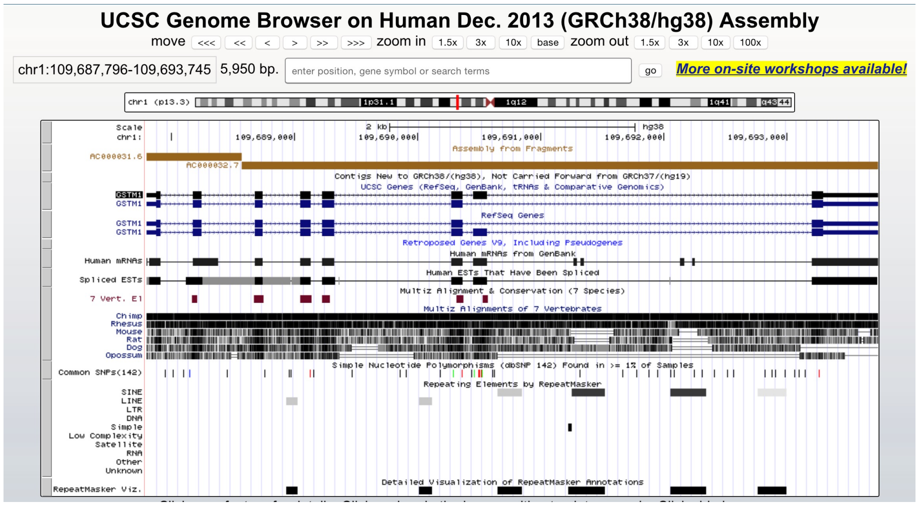
Exon Length

Intron Length



Lander *Nature* **409,** 860–921 (2001), Fig. 35

Retrieving the data: Genome Browsers (UCSC)

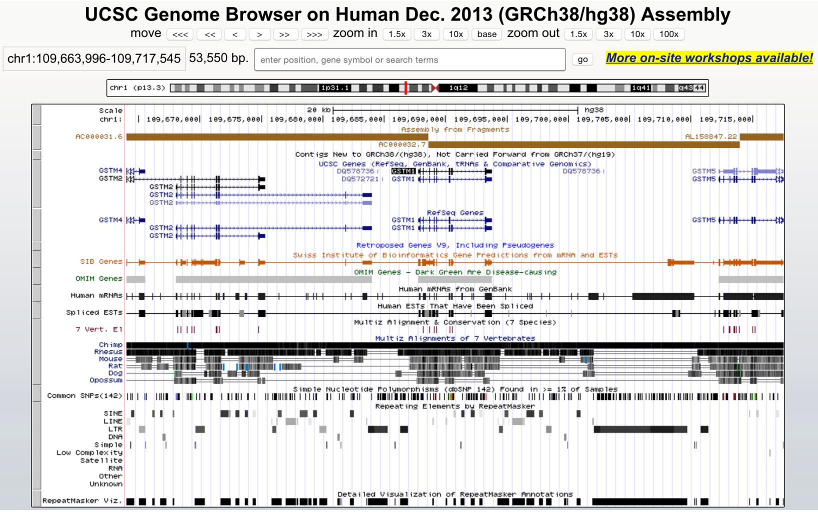


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:109,687,796-109,693,745 5,950 bp.

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
100x
enter position, gene symbol or search terms
go
More on-site workshops available!
chr1 (p13.3)
1p31.1
1q12
1q41
q43
44
Scale
chr1:
2 kb
hg38
109,689,000
109,690,000
109,691,000
109,692,000
109,693,000
Assembly from Fragments
AC000031.6
AC000032.7
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
GSTM1
GSTM1
RefSeq Genes
GSTM1
GSTM1
Retroposed Genes V9, Including Pseudogenes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.

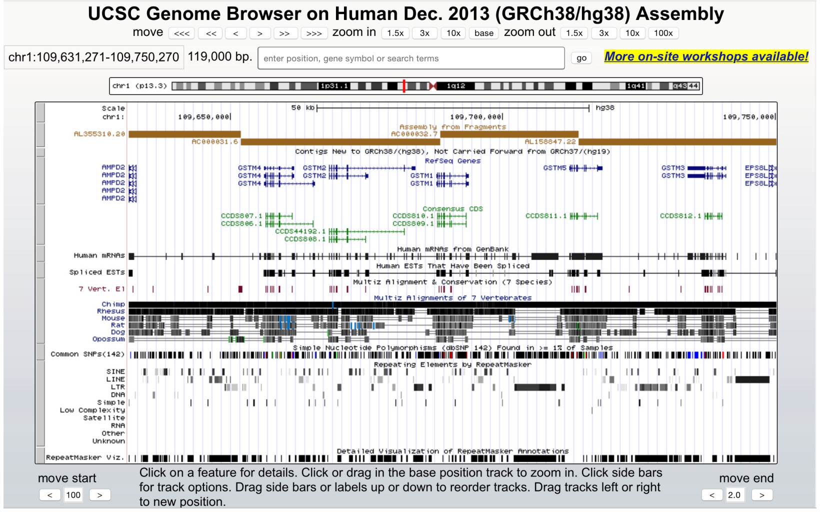
Genome Browsers (genome.ucsc.edu)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:109,663,996-109,717,545 53,550 bp.
enter position, gene symbol or search terms
go
More on-site workshops available!
chr1 (p13.3)
1p31.1
1q12
1q41
Scale
chr1:
20 kb
hg38
109,670,000
109,675,000
109,680,000
109,685,000
109,690,000
109,695,000
109,700,000
109,705,000
109,710,000
109,715,000
Assembly from Fragments
AC000031.6
AC000032.7
AL158847.22
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
GSTM4
GSTM2
GSTM1
GSTM5
DQ578736
DQ572721
RefSeq Genes
Retroposed Genes V9, Including Pseudogenes
Swiss Institute of Bioinformatics Gene Predictions from mRNA and ESTs
SIB Genes
OMIM Genes - Dark Green Are Disease-causing
OMIM Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.

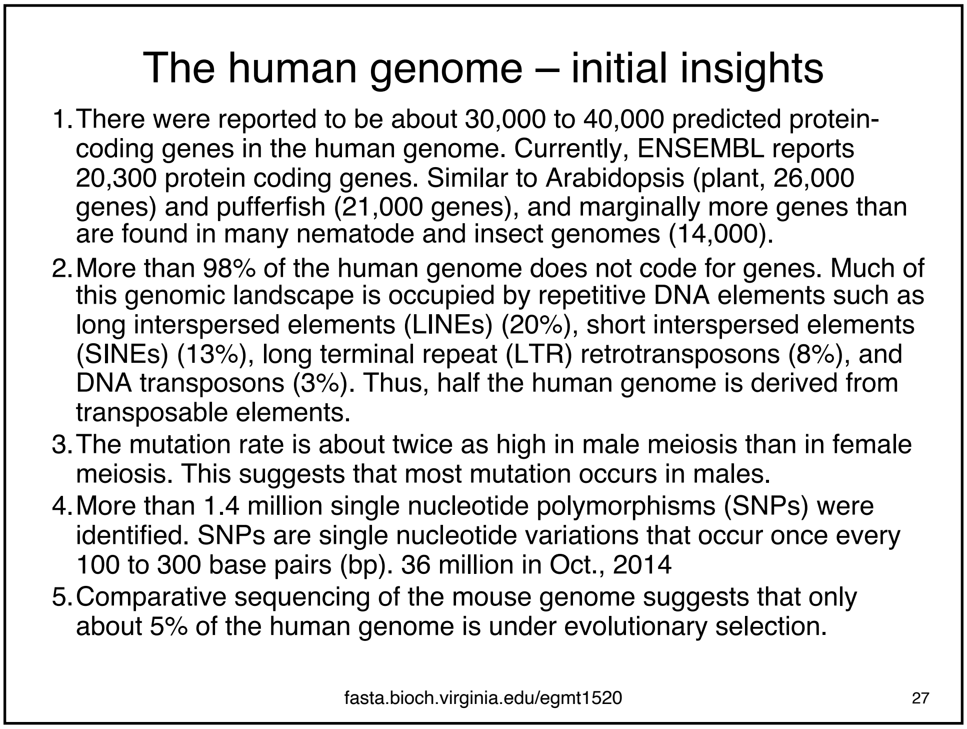
Genome Browsers (UCSC)



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

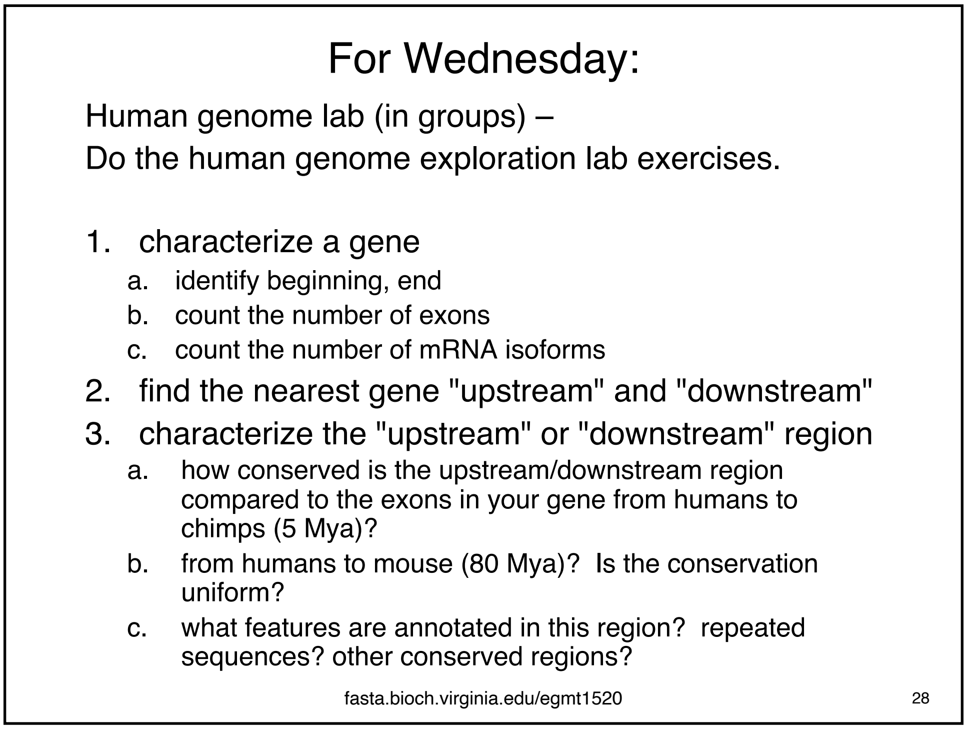
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr1:109,631,271-109,750,270 119,000 bp.
enter position, gene symbol or search terms
go
More on-site workshops available!
chr1 (p13.3)
1p31.1
1q12
1q41
Scale
chr1:
50 kb
hg38
109,650,000
109,700,000
109,750,000
Assembly from Fragments
AL355310.20
AC000032.7
AC000031.6
AL158847.22
Contigs New to GRCh38/(hg38), Not Carried Forward from GRCh37/(hg19)
RefSeq Genes
AMPD2
GSTM4
GSTM2
GSTM1
GSTM5
GSTM3
EPS8L3
Consensus CDS
CCDS807.1
CCDS806.1
CCDS44192.1
CCDS808.1
CCDS810.1
CCDS809.1
CCDS811.1
CCDS812.1
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Detailed Visualization of RepeatMasker Annotations
RepeatMasker Viz.
move start
< 100 >
move end
< 2.0 >



The human genome – initial insights

1. There were reported to be about 30,000 to 40,000 predicted protein-coding genes in the human genome. Currently, ENSEMBL reports 20,300 protein coding genes. Similar to Arabidopsis (plant, 26,000 genes) and pufferfish (21,000 genes), and marginally more genes than are found in many nematode and insect genomes (14,000).
2. More than 98% of the human genome does not code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs) (20%), short interspersed elements (SINEs) (13%), long terminal repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Thus, half the human genome is derived from transposable elements.
3. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males.
4. More than 1.4 million single nucleotide polymorphisms (SNPs) were identified. SNPs are single nucleotide variations that occur once every 100 to 300 base pairs (bp). 36 million in Oct., 2014
5. Comparative sequencing of the mouse genome suggests that only about 5% of the human genome is under evolutionary selection.

27

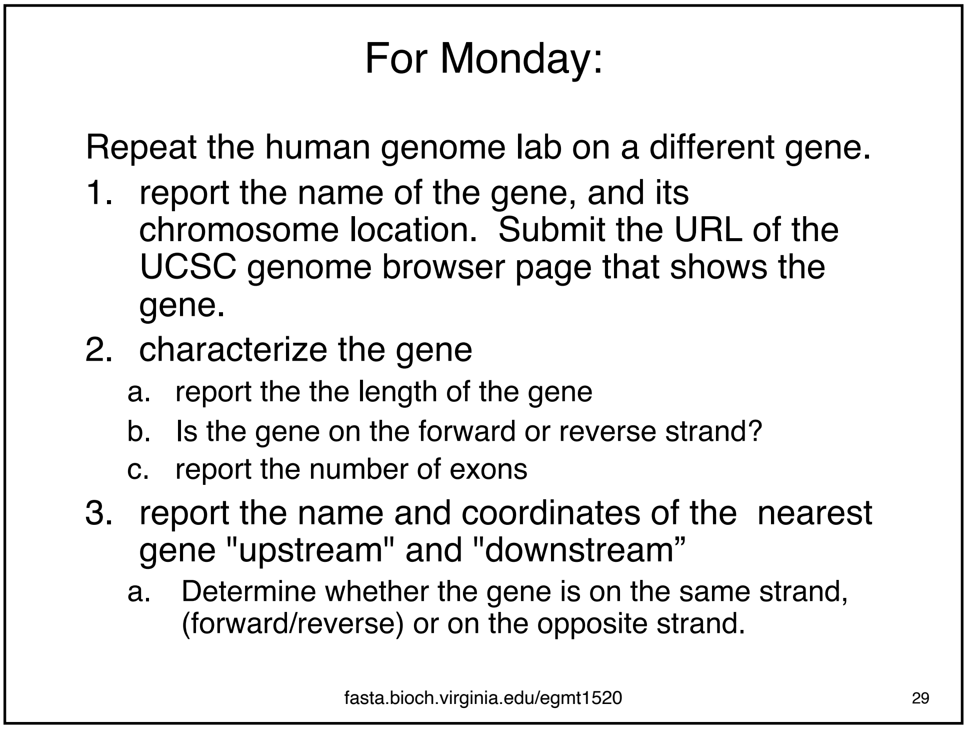


For Wednesday:

Human genome lab (in groups) –
Do the human genome exploration lab exercises.

1. characterize a gene
   a. identify beginning, end
   b. count the number of exons
   c. count the number of mRNA isoforms
2. find the nearest gene "upstream" and "downstream"
3. characterize the "upstream" or "downstream" region
   a. how conserved is the upstream/downstream region compared to the exons in your gene from humans to chimps (5 Mya)?
   b. from humans to mouse (80 Mya)? Is the conservation uniform?
   c. what features are annotated in this region? repeated sequences? other conserved regions?

28



For Monday:

Repeat the human genome lab on a different gene.

1. report the name of the gene, and its chromosome location. Submit the URL of the UCSC genome browser page that shows the gene.
2. characterize the gene
   a. report the the length of the gene
   b. Is the gene on the forward or reverse strand?
   c. report the number of exons
3. report the name and coordinates of the nearest gene "upstream" and "downstream”
   a. Determine whether the gene is on the same strand, (forward/reverse) or on the opposite strand.

29