Genomic biology

Genome
Genome Project

GenBank

The NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis. GenBank consists of several divisions, most of which can be accessed through the Nucleotide database. The exceptions are the EST and GSS divisions, which are accessed through the Nucleotide EST and Nucleotide GSS databases, respectively.

Genome

Contains sequence and map data from the whole genomes of over 1000 organisms. The genomes represent

NCBI

Genome

Contains sequence and map data from the whole genomes of over 1000 organisms. The records regard both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.

Sequence Read Archive (SRA)

The Short Read Archive (SRA) stores sequencing data from the next generation of sequencing platforms

Genome

Genome Search

Limits Advanced

Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Using Genome

- Help
- Browse by Organism
- Download / FTP
- Sumbit a genome

Custom resources

- Human Genome
- Microbes
- Organelles
- Plants
- Viruses

Other Resources

- BioProject
- BioSample
- Assembly
- Protein Clusters
- Map Viewer

Genome Tools

- BLAST the Human Genome
- Genomic groups BLAST
- NCBI remap
- Genome Decoration Page

Genome Annotation and Analysis

- Eukaryotic Genome Annotation
- Prokaryotic Genome Annotation
- PASC (Pairwise Sequence Comparison)
- TaxPlot (3-way Genome Comparison)

External Resources

- GOLD - Genome On Line Database
- Enseble Genome Browser
- Bacteria Genomes at Sanger
- Large-Scale Genome Sequencing (NHGRI)

Overview | Eukaryotes | Prokaryotes | Viruses

First Previous **Shown: 1 - 100 out of 6249 items** Next Last

Organism/Name	Kingdom (All)	Group (All)	SubGroup (All)					
Abalone shriveling syndrome-associated virus	Viruses	dsDNA viruses, no RNA stage	unclassified	0.035	1	-	-	1
Abelson murine leukemia virus	Viruses	Retro-transcribing viruses	Retroviridae	0.006	1	-	-	1
Abiotrophia defectiva	Bacteria	Firmicutes	Bacilli	3.48	-	-	-	1
Abutilon Brazil virus	Viruses	ssDNA viruses	Geminiviridae	0.005	2	-	-	1
Abutilon mosaic Bolivia virus	Viruses	ssDNA viruses	Geminiviridae	0.005	2	-	-	1
Abutilon mosaic virus	Viruses	ssDNA viruses	Geminiviridae	0.005	2	-	-	1
Acacia mangium	Eukaryotes	Plants	Land Plants	0	13	-	-	1
Acanthamoeba castellanii	Eukaryotes	Protists	Other Protists	46.43	-	1	-	1
Acanthamoeba polyphaga mimivirus	Viruses	dsDNA viruses, no RNA stage	Mimiviridae	1.18	1	-	-	1
Acanthascus dawsoni	Eukaryotes	Animals	Other Animals	0	-	-	-	1
Acanthocheilonema viteae	Eukaryotes	Animals	Roundworms	0	-	-	-	1
Acanthocystis turfacea Chlorella virus 1	Viruses	dsDNA viruses, no RNA stage	Phycodnaviridae	0.29	1	-	-	1
Acaryochloris marina	Bacteria	Cyanobacteria	Chroococcales	8.36	1	-	9	1
Acaryochloris phage A-HIS1	Viruses	dsDNA viruses, no RNA stage	Siphoviridae	0	-	-	-	1
Acaryochloris sp. CCMEE 5410	Bacteria	Cyanobacteria	Chroococcales	0	-	-	-	1
Acetivibrio cellulolyticus	Bacteria	Firmicutes	Clostridia	6.14	-	-	-	1
Acetobacter aceti	Bacteria	Proteobacteria	Alphaproteobacteria	3.58	-	-	1	2
Acetobacter pasteurianus	Bacteria	Proteobacteria	Alphaproteobacteria	3.34	1	-	6	8
Acetobacter pomorum	Bacteria	Proteobacteria	Alphaproteobacteria	2.88	-	-	-	1
Acetobacter tropicalis	Bacteria	Proteobacteria	Alphaproteobacteria	3.72	-	-	-	2

Organism/Name	BioProject	Group: All	SubGroup: Fishes	Size (Mb)	GC%	Assembly	Chrs	Organelles
Danio rerio	PRJNA167	Animals	Fishes	0	-		-	-
Danio rerio	PRJNA13922	Animals	Fishes	1400.99	36.90	Zv9	25	1
Danio rerio	PRJNA38201	Animals	Fishes	0	-		-	-
Takifugu rubripes	PRJNA12054	Animals	Fishes	0.016	44.20		-	1	-		-	13
Gasterosteus aculeatus	PRJNA11773	Animals	Fishes	0.016	44.70		-	1	-		-	13
Gasterosteus aculeatus	PRJNA11774	Animals	Fishes	0	-		-	-	-		-	-
Gasterosteus aculeatus	PRJNA12389	Animals	Fishes	0	-		-	-	-		-	-
Gasterosteus aculeatus	PRJNA13579	Animals	Fishes	446.61	44.60	GasAcu_Jan2006	-	-	-	AANH01	-	-
Tetraodon nigroviridis	PRJNA12350	Animals	Fishes	342.4	46.60	TetNig_Feb2004	-	-	-	CAAE01	-	-
Tetraodon nigroviridis	PRJNA15573	Animals	Fishes	0.016	46.90		-	1	-		-	13
Tetraodon nigroviridis	PRJNA20435	Animals	Fishes	0	-		-	-	-		-	-
Tetraodon nigroviridis	PRJNA33819	Animals	Fishes	0	-		-	-	-		-	-
Oncorhynchus mykiss	PRJNA11824	Animals	Fishes	0.017	46.00		-	1	-		-	13
Oncorhynchus	PRJNA12371	Animals	Fishes	0	-		-	-	-		-	-

Genome information for zebrafish (Danio rerio)

Lineage: Eukaryota[844]; Metazoa[299]; Chordata[113]; Craniata[108]; Vertebrata[107]; Euteleostomi[104]; Actinopterygii[25]; Neopterygii[25]; Teleostei[24]; Ostariophysi[3]; Cypriniformes[2]; Cyprinidae[2]; Danio[1]

The reference sequence (RefSeq) genome assembly is provided by NCBI using assembly instructions provided by the Wellcome Trust Sanger Institute. The assembled genome is distributed internationally by FTP and can be viewed in browsers provided by NCBI, Ensembl, and the University of Santa Cruz (UCSC).

Chromosomes

Click on chromosome name to open Map Viewer

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

17 18 19 20 21 22 23 24 25

Assembly and Annotation

Default assembly

Assembly Name	Zv9
Last sequence update	
Highest level of assembly	some chromosomes assembled
Size (total bases)	1,412,448,247
Number of genes	28,733
Number of proteins	27,391

Mitochondrial Genome

Last record update	01-Feb-2010
Last sequence update	02-Aug-2001

The Human Genome Project (HGP)

- Was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health.
- The Wellcome Trust (U.K.) became a major partner.
- Additional contributions came from Japan, France, Germany, China, and others.

Project goals were to

- *identify* all the approximately 20,000-25,000 genes in human DNA,
- *determine* the sequences of the 3 billion chemical base pairs that make up human DNA,
- *store* this information in databases,
- *improve* tools for data analysis,
- *transfer* related technologies to the private sector, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.

CELERA

16 February 2001
Vol 291, Issue 5507, Pages 1145-1434

Overview | Eukaryotes | Prokaryotes | Viruses

First Previous **Shown: 1 - 100 out of 304 items** Next Last

Organism/Name	BioProject	Group (All)	SubGroup (Mammals)	Size (Mb)	GC%	Assembly	Chrs	Organelles	Plasmids	WGS	Scaffolds	Genes	Prote
Homo sapiens	PRJNA168	Animals	Mammals	3095.69	41.58	GRCh37.p6	24	1	-		383	32143	360
Homo sapiens	PRJNA1431	Animals	Mammals	2695.72	40.80	Hs_Celera_WGSA	24	-	-	AADD01	8832	-	-
Homo sapiens	PRJNA16133	Animals	Mammals	158.33	40.90	CRA_TCAGchr7v2	1	-	-		6	1526	19
Homo sapiens	PRJNA20837	Animals	Mammals	2809.55	40.90	Homo sapiens HuRef	24	-	-	ABBA01	188408	31708	343
Homo sapiens	PRJNA28335	Animals	Mammals	41.67	40.80	Watson-partial	-	-	-	ABKV01	-	-	-
Homo sapiens	PRJNA28911	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA28919	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA28957	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA29429	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA30559	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens neanderthalensis	PRJNA30941	Animals	Mammals	0.017	44.40		-	1	-		-	13	3
Homo sapiens	PRJNA30977	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33237	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33783	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33831	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33835	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33847	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33851	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33859	Animals	Mammals	0	-		-	-	-		-	-	-
Homo sapiens	PRJNA33865	Animals	Mammals	0	-		-	-	-		-	-	-

Genome information for human (Homo sapiens)

Lineage: Eukaryota[844]; Metazoa[299]; Chordata[113]; Craniata[108]; Vertebrata[107]; Euteleostomi[104]; Mammalia[66]; Eutheria[62]; Euarchontoglires[32]; Primates[15]; Haplorrhini[13]; Catarrhini[11]; Hominidae[4]; Homo[1]

The reference sequence (RefSeq) genome assembly is provided by NCBI using assembly instructions provided by the International Human Genome Project. The assembled genome is distributed internationally by FTP and the same assembly can be viewed in browsers provided by NCBI, Ensembl, and the University of Santa Cruz (UCSC).The reference genome is annotated More...

Chromosomes

Click on chromosome name to open Map Viewer

Assembly and Annotation

Default assembly

Assembly Name	GRCh37.p5
Last sequence update	06-Mar-2009
Highest level of assembly	some chromosomes assembled
Size (total bases)	3,101,788,170
Number of genes	36,036
Number of proteins	32,130

Mitochondrial Genome

Last record update	30-Apr-2010
Last sequence update	08-Jul-2009

J. Craig Venter INSTITUTE

HuRef Genome Browser

JCVI has published the first diploid genome of an individual—Dr. Venter, in PLoS Biology.

This analysis and assembly of Dr. Venter's DNA is the first look at both sets of an individual's chromosomes.

All the data for the first human diploid genome has been deposited at NCBI, and JCVI researchers have also developed a NEW GENOME BROWSER that highlights the newly discovered variation.

PLoS Biol. 2007 Sep 4;5(10):e254.

The diploid genome sequence of an individual human.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC.

J. Craig Venter Institute, Rockville, Maryland, USA. slevy@jcvi.org

Comment in:

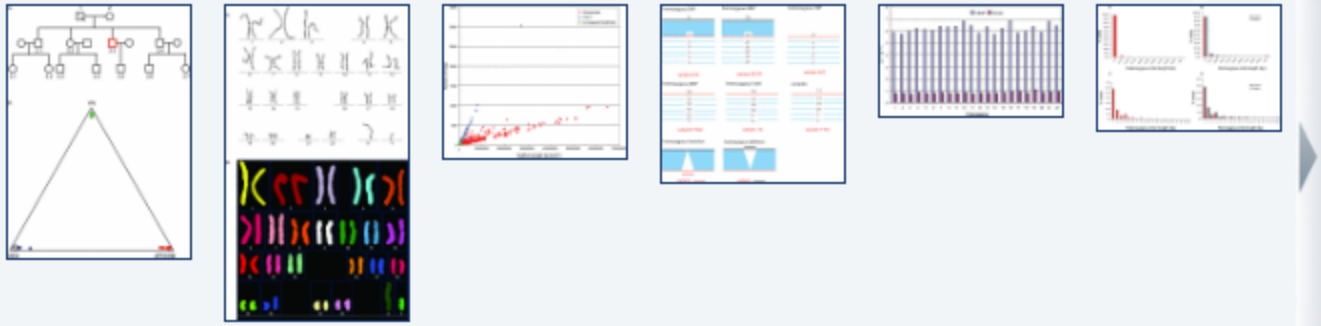
PLoS Biol. 2007 Oct;5(10):e266.

Abstract

Presented here is a genome sequence of an individual human. It was produced from approximately 32 million random DNA fragments, sequenced by Sanger dideoxy technology and assembled into 4,528 scaffolds, comprising 2,810 million bases (Mb) of contiguous sequence with approximately 7.5-fold coverage for any given region. We developed a modified version of the Celera assembler to facilitate the identification and comparison of alternate alleles within this individual diploid genome. Comparison of this genome and the National Center for Biotechnology Information human reference assembly revealed more than 4.1 million DNA variants, encompassing 12.3 Mb. These variants (of which 1,288,319 were novel) included 3,213,401 single nucleotide polymorphisms (SNPs), 53,823 block substitutions (2-206 bp), 292,102 heterozygous insertion/deletion events (indels) (1-571 bp), 559,473 homozygous indels (1-82,711 bp), 90 inversions, as well as numerous segmental duplications and copy number variation regions. Non-SNP DNA variation accounts for 22% of all events identified in the donor, however they involve 74% of all variant bases. This suggests an important role for non-SNP genetic alterations in defining the diploid genome structure. Moreover, 44% of genes were heterozygous for one or more variants. Using a novel haplotype assembly strategy, we were able to span 1.5 Gb of genome sequence in segments >200 kb, providing further precision to the diploid nature of the genome. These data depict a definitive molecular portrait of a diploid human genome that provides a starting point for future genome comparisons and enables an era of individualized genomic information.

PMID: 17803354 [PubMed - indexed for MEDLINE] PMCID: PMC1964779 Free PMC Article

Images from this publication. See all images (15) Free text



Publication Types, MeSH Terms

Related citations

The sequence of the human genome. [Scienc

Towards a comprehensive structural var map of an individual human g [Genome B

The complete genome of an individual by massively parallel DNA sequencing [Natu

Review [Analysis and application of SNP haplotype in the human [Yi Chuan Xue Ba

Review The Human Genome Project--an overview. [Med Res Re

See re

Cited by over 100 PubMed Cent articles

Improved detection of rare genetic varian diseases. [PLoS On

Assessment of genetic variation for the L retrotransposon from r [BMC Bioinformatic

Frequent and efficient use of the sister c for DNA double-strand break rep [PLoS B

All links from this record

Related Citations

Gene

Nucleotide

Nucleotide (RefSeq)

Nucleotide (Weighted)

Protein (RefSeq)

Protein (Weighted)

References for this PMC Article

Related Project

The HuRef assembly represents a composite haploid version of the diploid genome sequence from a single individual.

The highest scoring allele is represented in the consensus sequence.

Overview | Genomes

Genome information for human (Homo sapiens)

Lineage: Eukaryota[844]; Metazoa[299]; Chordata[113]; Craniata[108]; Vertebrata[107]; Euteleostomi[104]; Mammalia[66]; Eutheria[62]; Euarchontoglires[32]; Primates[15]; Haplorrhini[13]; Catarrhini[11]; Hominidae[4]; Homo[1]

This reference sequence (RefSeq) genome assembly is based on the GenBank submission of the J. Craig Venter genome assembly. Annotation displayed on the RefSeq genome records and in the Map Viewer is calculated by the NCBI genome annotation pipeline.

Chromosomes

Click on chromosome name to open Map Viewer

Assembly and Annotation

No assembly data available for this organism genome

Related BioProjects

Type	Count
RefSeq Genome	1

comments on this story

Stories by subject
- Biotechnology
- Business
- Cell and molecular biology
- Genetics
- Health and medicine

Stories by keywords
- Craig Venter
- Human genome sequencing
- Next-generation sequencing technology
- James Watson
- 454 LifeSciences

This article elsewhere

Blogs linking to this article

Add to Connotea
Add to Digg
Add to Facebook
Add to Newsvine
Add to Del.icio.us

Published online 16 April 2008 | **452**, 788 (2008) | doi:10.1038/452788b

News

James Watson's genome sequenced at high speed

New-generation technology takes just four months and costs a fraction of old method.

Meredith Wadman

The first full genome to be sequenced using next-generation rapid-sequencing technology is published today (see page 872)[1], marking another milestone in the extraordinarily fastmoving field of human genome sequencing.

It took just four months, a handful of scientists and less than US$1.5 million to sequence the 6 billion base pairs of DNA pioneer James Watson. The achievement is first proof of principle that these rapid-sequencing machines can decipher large, complex genomes (see page 819)[2]. Made in this case by Connecticut-based 454 Life Sciences — a division of Roche Diagnostics — they allow many more sequencing reactions to proceed at the same time, on the same surface, than the previous generation of machines that produced the inaugural human genomes[3,4]. That change has had big pay-offs in speed, efficiency and, ultimately, cost (see Table 1).

James Watson's is not the first full genome to be published; that distinction goes to genomics entrepreneur J. Craig Venter, whose genome was sequenced using previous-generation machines[5] at a cost of $100 million. "Venter's genome was at the end of the last generation," says 454 founder Jonathan Rothberg of the Rothberg Institute for Childhood Diseases in Guilford, Connecticut, who is the Watson paper's lead author. "We did this work in January 2007, with last January's technology," he points out; the raw data were released in May 2007. "It just keeps on getting better and cheaper." Rothberg calls Watson "the first of the rest of us", but the low cost the team managed is still a far cry from the '$1,000 genome'

most recent | commented

- Coral bleaching goes from bad to worse
  19 November 2010
- US clamp-down on alcoholic energy drinks
  18 November 2010
- Nobel prize committee under fire
  18 November 2010
- Funding woes afflict African herbal therapy institute
  18 November 2010
- Disease-in-a-dish approach gives clues to Rett syndrome
  17 November 2010

Related stories
- Ready or not
  09 April 2008
- Watson's folly
  24 October 2007
- Common sense for our genomes
  17 October 2007
- Genomics: The personal side of genom
  03 October 2007
- All about Craig: the first 'full' genome sequence
  05 September 2007

The Ensembl Project

The Ensembl Project

was started in 1999, some years before the draft human genome was completed. The goal was to automatically annotate the genome, integrate this annotation with other available biological data and make all this publicly available via the web.

Ensembl is a joint project between European Bioinformatics Institute (EBI), the Wellcome Trust Sanger Institute (WTSI).

Both institutes are located in the Wellcome Trust Genome Campus in Hinxton, south of the city of Cambridge, United Kingdom.

Since the website's launch in July 2000, many more genomes have been added to Ensembl and the range of available data has also expanded to include comparative genomics, variation and regulatory data.

Find a Species

The main Ensembl site focuses on vertebrate genomes - scroll down for links to our sister sites covering invertebrates, plants, bacteria, etc.

Species tree

Static image (PDF) · Interactive image (requires Java)

Ensembl Species


Alpaca
Vicugna pacos
vicPac1


Guinea Pig
Cavia porcellus
cavPor3


Platypus
Ornithorhynchus anatinus
OANA5


Anole Lizard
Anolis carolinensis
AnoCar2.0


Hedgehog
Erinaceus europaeus
HEDGEHOG


Rabbit
Oryctolagus cuniculus
oryCun2


Armadillo
Dasypus novemcinctus
dasNov2


Horse
Equus caballus
EquCab2


Rat
Rattus norvegicus
RGSC3.4


Baboon (preview - assembly only)
Papio hamadryas


Human
Homo sapiens
GRCh37


Saccharomyces cerevisiae
Saccharomyces cerevisiae
EF3


Bushbaby
Otolemur garnettii
BUSHBABY1


Hyrax
Procavia capensis
proCap1


Sheep (preview - assembly only)
Ovis aries


Caenorhabditis elegans
WS220


Kangaroo rat
Dipodomys ordii
dipOrd1


Shrew
Sorex araneus
COMMON_SHREW1


Ciona intestinalis
JGI2


Lamprey (preview new assembly)
Petromyzon marinus


Sloth
Choloepus hoffmanni
choHof1


Ciona savignyi
CSAV2.0


Lesser hedgehog tenrec
Echinops telfairi
TENREC


Squirrel
Spermophilus tridecemlineatus
SQUIRREL


Cat
Felis catus
CAT


Macaque
Macaca mulatta
MMUL_1


Stickleback
Gasterosteus aculeatus
BROADS1


Chicken
Gallus gallus
WASHUC2


Marmoset
Callithrix jacchus
C_jacchus3.2.1


Tarsier
Tarsius syrichta
tarSyr1


Chimpanzee
Pan troglodytes
CHIMP2.1


Medaka
Oryzias latipes
MEDAKA1


Tasmanian devil
Sarcophilus harrisii
DEVIL7.0


Cow
Bos taurus
UMD3.1


Megabat
Pteropus vampyrus
pteVam1


Tetraodon
Tetraodon nigroviridis
TETRAODON8

Search Ensembl Human

Search for: Go

e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

Description

Human (*Homo sapiens*)

Assembly

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly GRCh37 (GCA_000001405.6) from the Genome Reference Consortium. This assembly is used by UCSC to create their hg19 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:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Human (GRCh37)
About this species
Description
Genome Statistics
Assembly and Genebuild
Top 40 InterPro hits
Top 500 InterPro hits
What's New
Sample entry points
Karyotype
Location (6:133017695-1331
Gene (BRCA2)
Transcript (FOXP2-203)
Variation (rs1333049)
Regulation (ENSR00001348
Configure this page
Manage your data
Export data
Bookmark this page
Search Ensembl Hu
Search for:
e.g. BRCA2 or 6:133017695-13316115
Description
Human (*Homo sapiens*)
Assembly
This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly GRCh37 (GCA_000001405.6) from th
database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments
This release of the assembly has the following properties:
- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine haplotypic regions, mainly in the MHC region of chromosome 6.

As the GRC maintains and improves the assembly, patches are being introduced. Patch release five (GRCh37.p5) was included in Ens

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

Search for: Go

e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

Assembly and Genebuild

Summary

Assembly:	GRCh37.p5, Feb
Database version:	64.37
Base Pairs:	3,283,984,159
Golden Path Length:	3,101,804,739
Genebuild by:	Ensembl
Genebuild method:	Full genebuild
Genebuild started:	Jul 2010
Genebuild released:	Apr 2011
Genebuild last updated/patched:	Sep 2011

Gene counts

Known protein-coding genes:	20,469
Novel protein-coding genes:	431
Pseudogenes:	14,266
RNA genes:	12,499
Immunoglobulin/T-cell receptor gene segments:	562
Gene exons:	640,185
Gene transcripts:	178,191

Other

Genscan gene predictions:	47,019
Short Variants (SNPs, indels, somatic mutations):	30,099,223
Structural variants:	1,772,315

Permanent link - View in archive site

Base Pairs (whole assembly)

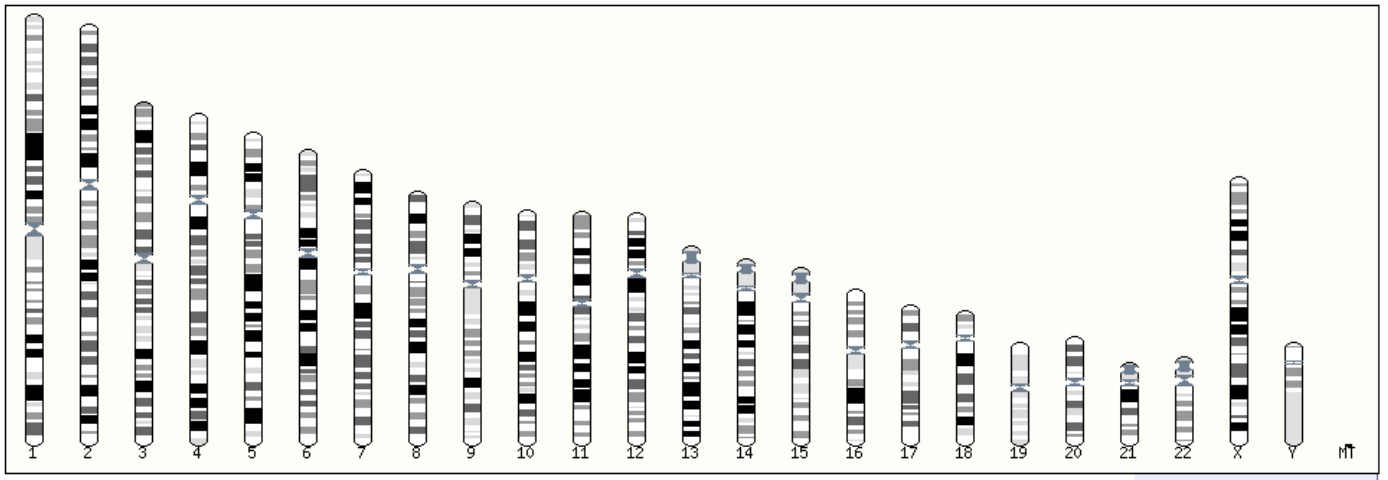
The total number of base pairs; the sum of all sequences in the database. This includes redundant regions such as haplotypic sequences and the pseudo-autosomal region (PAR) of the Y chromosome in human, and gaps in *Drosophila melanogaster*.

Golden Path

The "golden path" is the length of the reference assembly. It consists of the sum of all top-level sequences, omitting any redundant regions such as haplotypes and PARs.

Karyotype

Whole genome help



Export Image

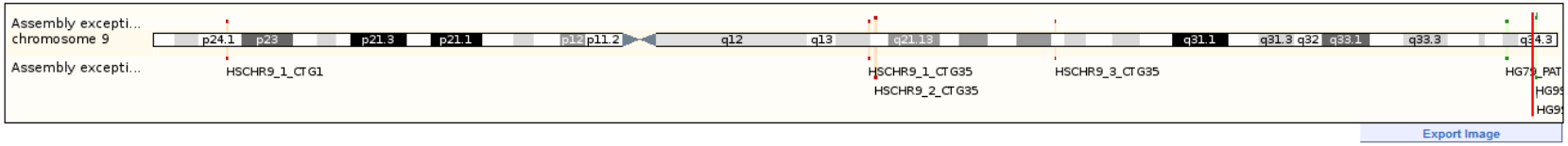
Click on the image above to jump to a chromosome, or click and drag to select a region

Summary

Assembly:	GRCh37.p5, Feb 2009
Database version:	64.37
Base Pairs:	3,283,984,159
Golden Path Length:	3,101,804,739
Genebuild by:	Ensembl
Genebuild method:	Full genebuild
Genebuild started:	Jul 2010
Genebuild released:	Apr 2011
Genebuild last updated/patched:	Sep 2011

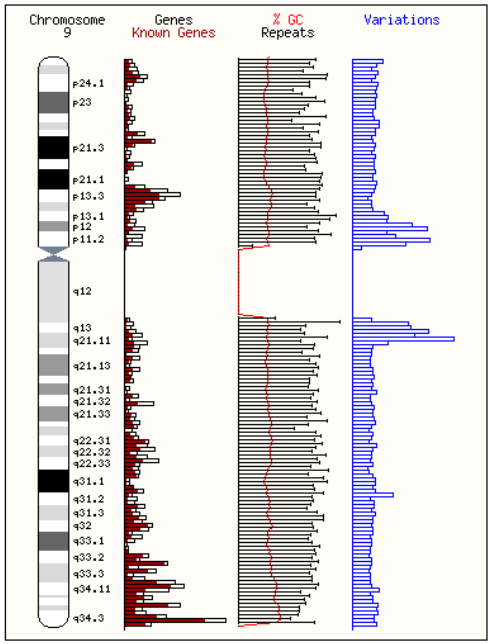
Gene counts

Known protein-coding genes:	20,469
Novel protein-coding genes:	431
Pseudogenes:	14,266
RNA genes:	12,499
Immunoglobulin/T-cell receptor gene segments:	562
Gene exons:	640,185
Gene transcripts:	178,191



Single chromosome information

Chromosome summary help



Click on the image above to zoom into that point

Jump to Chromosome: 9 Go

Chromosome Statistics

Length (bps):	141,213,431
Known Protein-coding Genes:	788
Novel Protein-coding Genes:	12
Pseudogene Genes:	693
miRNA Genes:	69
rRNA Genes:	19
snRNA Genes:	66
snoRNA Genes:	51
Misc RNA Genes:	55
SNPs:	1,674,619

ays

sked Questions
ls

sk

Region in detail

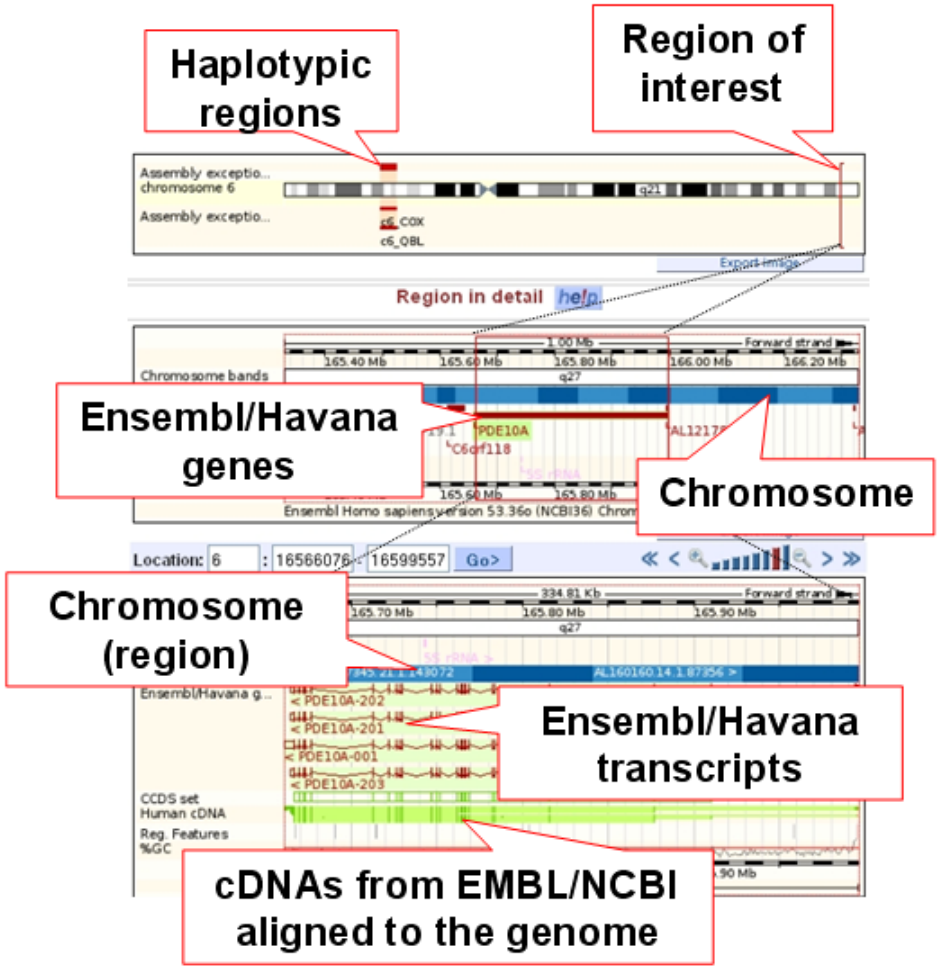
Example page

Region in detail allows you to browse genes, variations, sequence conservation, and other annotation along the genome. There are three main panels (or images), shown haplotypes or patches in red or green, respectively.

The next panel is called the Top panel. It shows an overview of Ensembl and Havana genes along the chromosome (depicted as a blue bar). The individual contigs that m genes are shown as gold or red bars, pseudogenes are grey, and non-coding genes are blue.

The main panel is further explored in Figure 2.

FIGURE 1



Region in detail

The Main panel is shown in **figure 2**. Data tracks (genes, cDNA alignments, etc.) are shown above and below the blue bar. Tracks above the blue bar are on the forward strand. Reorder the tracks by using the vertical blue bar at the left, or in the Track order menu in the configuration panel. Add or change data tracks using the **configure**

To zoom in or change the display, use our zoom slide, or enter in basepairs manually. Alternatively, click any gene or transcript, or click and drag the mouse and form a

FIGURE 2

Though the HGP completed in 2003, analyses of the data will continue.

The Genome Reference Consortium (GRC) aims to improve the representation of the reference human genome.

Genome Reference Consortium

The goal of the Genome Reference Consortium (GRC) is to ensure that the human and mouse reference assemblies are biologically relevant by closing gaps, fixing errors and representing complex variation.

The Genome Reference Consortium (GRC)

Some genomic regions (gaps) were recalcitrant to closure with existing technology.

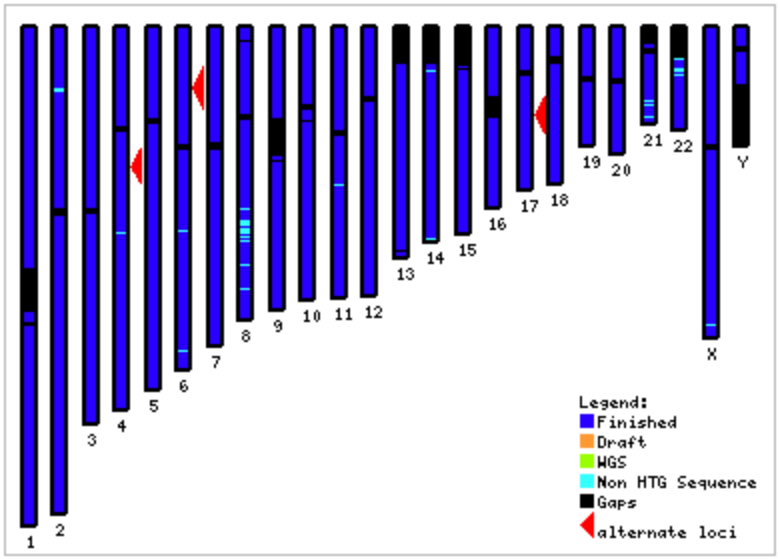
These regions are so variable that they are best represented by multiple sequences in order to capture all of the sequences potentially available at these loci.

GRC works to correct misrepresented loci, to close gaps, to provide alternate assemblies for complex or structurally variant genomic loci.

Genome Reference Consortium

Human Genome Overview

Information concerning continuing improvement of the human genome.



GRCh37: A graphical representation of the latest human assembly. The genome is colored with respect to the genomic component used to build the genome assembly at that location. The red triangles mark regions where alternate loci have been provided.

The most recent assembly for human is GRCh37 (download the assembly). This is the first assembly produced by the GRC and is considered the next version of NCBI Build 36 (also known as hg18). Improvements in this assembly include:

- **Closure of 25 unspanned gaps found in Build 36**
- **Resolution of over 150 issues reported as problems in Build 36**
- **Addition of alternate loci for three complex regions, including the MHC region.**
- **Standardization of AGPs, including the addition of biological gap information.**

GRCh37 is a haploid assembly, constructed from multiple individuals and can be divided into a ' primary assembly ' and a set of ' alternate loci '. The primary assembly represents the assembled chromosomes, plus any unlocalized or unplaced sequence that represent the non-redundant, haploid assembly .The alternate loci represent regions for which there is large scale variation and an alternate tilng path is available for this region. An example of such a region can be found at chromosome 17q21.31, often known as the MAPT locus. This region was described as carrying an inversion polymorphism (PMID: 15654335) and has been associated with various phenotypes (PMID: 16718704 ; PMID: 18628315). The version of this region in Build 36 was actually a mosaic of both haplotypes (as tracked in HG-77) and has been resolved in GRCh37 thanks to data described in Zody et al., 2008 (PMID: 19165922).

Information on alternate loci

Chromosome region with alternate loci	Length of region	Number of alternate contigs in region	View Region
UGT2B17 region (chr4:69,170,077-69,877,175)	707,099 bp	1 contig	view
MHC region (chr6: 28,477,797-33,448,354)	4,970,558 bp	7 contigs	view
MAPT region (chr17: 43,384,864-44,913,631)	1,528,768 bp	1 contig	view

The most recent assembly for human is GRCh37

The GRCh37 genome assembly

- is a haploid assembly, constructed from multiple individuals and can be divided into a primary assembly and a set of alternate loci.
- The primary assembly represents the assembled chromosomes, plus any unlocalized or unplaced sequence that represent the non-redundant, haploid assembly.
- The alternate loci represent regions for which there is large scale variation and an alternate path is available for this region.

An example of alternate loci

An example of such a region can be found at chromosome 17q21.31, often known as the MAPT locus.

This region was described as carrying an inversion polymorphism and has been associated with two phenotypes. The version of this region in Build 36 was actually a mosaic of both haplotypes and has been resolved in GRCh37.

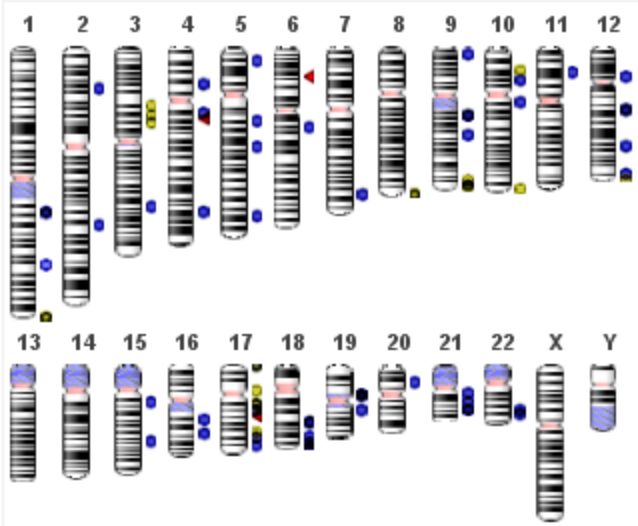
- In addition GRC provides patch releases in order to provide updated information for a particular region without changing the chromosome coordinates.

- Patches are small bits of sequences which have an alignment to the Primary Assembly so that the relationship between the patch sequence and the chromosome sequence is clear.

Genome Reference Consortium

Human Genome Overview

Information concerning the continuing improvement of the human genome.



Regions containing alternate-loci
Regions containing fix patches
Regions containing novel patches

An ideogram representation of the latest human assembly (not showing unplaced or unlocalized sequences).

The GRC is working hard to provide the best possible reference assembly for human. We do this by both generating multiple representations (alternate loci) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as patches . This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

Getting Data

GRCh37 (Latest Major Release): FTP
GRCh37 patch release 2 (Latest Minor Release): FTP
Information on regions under review: FTP

Next assembly update
The next assembly update will be a minor update (only patches) and will happen in December 2010

GRCh37 Patch Release 2 | GRCh37 Patch Release 1 | GRCh37

GRCh37 Patch Release 2 (GRCh37.p2)

Release date: August 16, 2010
Release type: minor
Release notes: In this release, 70 patches were added, 52 were of type Novel and 18 were of type Fix. There were 20 issues resolved in this release. Additionally, the MT sequence was added to this assembly.
Assembly accession: GCA_000001405.3

GRC News and Updates

13 Oct 2010

Zebrafish genome joins GRC

15 Sep 2010

GRC now has a blog!

Welcome to the GRC blog. Check back here for information on assembly upda tips on using the asse...

see all

Recently Resolved Human Iss

Human (HG-115) Nov19, 2

AC091390.1 was sequenced and contains the missing data from AC005088.3 We will update the TPF wit the inclusion of AC091390.1, this will be reflected in Build38.

Human (HG-121) Nov19, 2

The gap between AC060780.18 and AC109326.11 has be removed.

see all

References

Whole Genome Papers

The HGP Reference Assembly
The Venter Genome Assembly

Human Chromosome Papers

Chr1 Chr2 Chr3 Chr4 Chr5 Chr6
Chr7 Chr8 Chr9 Chr10 Chr11
Chr12 Chr13 Chr14 Chr15 Chr16

Patches provide updated information for a particular region without changing the chromosome coordinates.

'Fix' patches represent improved regions. 'Novel' patches represent new alternate loci that we have added to the assembly.

The gap between AC060780.18 and AC109326.11 has be removed.

see all

References

Whole Genome Papers

The HGP Reference Assembly

The Venter Genome Assembly

Human Chromosome Papers

Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22 ChrX ChrY

GRCh37 Patch Release 2 | GRCh37 Patch Release 1 | GRCh37

GRCh37 Patch Release 2 (GRCh37.p2)

Release date: August 16, 2010

Release type: minor

Release notes: In this release, 70 patches were added, 52 were of type Novel and 18 were of type Fix. There were 20 issues resolved in this release. Additionally, the MT sequence was added to this assembly.

Assembly accession: GCA_000001405.3

Human Region Information for GRCh37.p2

Region Name	Region Type	Alt Locus ID	Chr	Start	Stop	Patch Type
MHC	Alternate locus	GL000250.1	6	28477797	33448354	na
MHC	Alternate locus	GL000251.1	6	28477797	33448354	na
MHC	Alternate locus	GL000252.1	6	28477797	33448354	na
MHC	Alternate locus	GL000253.1	6	28477797	33448354	na
MHC	Alternate locus	GL000254.1	6	28477797	33448354	na
MHC	Alternate locus	GL000255.1	6	28477797	33448354	na
MHC	Alternate locus	GL000256.1	6	28477797	33448354	na
UGT2B17	Alternate locus	GL000257.1	4	69170077	69878175	na
MAPT	Alternate locus	GL000258.1	17	43384864	44913631	na
ABO	Patch	GL339450.1	9	136049442	136369192	fix
SMA	Patch	GL339449.1	5	68505250	70910270	novel
EPPK1_SPATC1	Patch	GL383535.1	8	144743526	145146062	fix
PECAM1	Patch	GL383558.1	17	62273514	62649312	fix
VPRBP	Patch	GL383523.1	3	51416109	51584055	fix
SCXB	Patch	GL383536.1	8	145285645	145659901	fix
DNAH12	Patch	GL383524.1	3	57369478	57399969	fix
FAM23A_MRC1	Patch	GL383543.1	10	17613209	18252930	fix
SOCS7	Patch	GL383559.1	17	36372617	36711255	fix
MYO19	Patch	GL383560.1	17	34442621	35005379	fix
REGION27	Patch	GL383561.1	17	21250948	21566608	fix
FAM101B	Patch	GL383562.1	17	252429	296626	fix
SLC25A26	Patch	GL383525.1	3	66270271	66308065	fix
REGION17	Patch	GL383544.1	10	133258319	133381404	fix
GALNT9	Patch	GL383548.1	12	132806993	132967794	fix
REGION12	Patch	GL383537.1	9	139136890	139252828	fix
REGION12	Patch	GL383538.1	9	139136890	139252828	fix