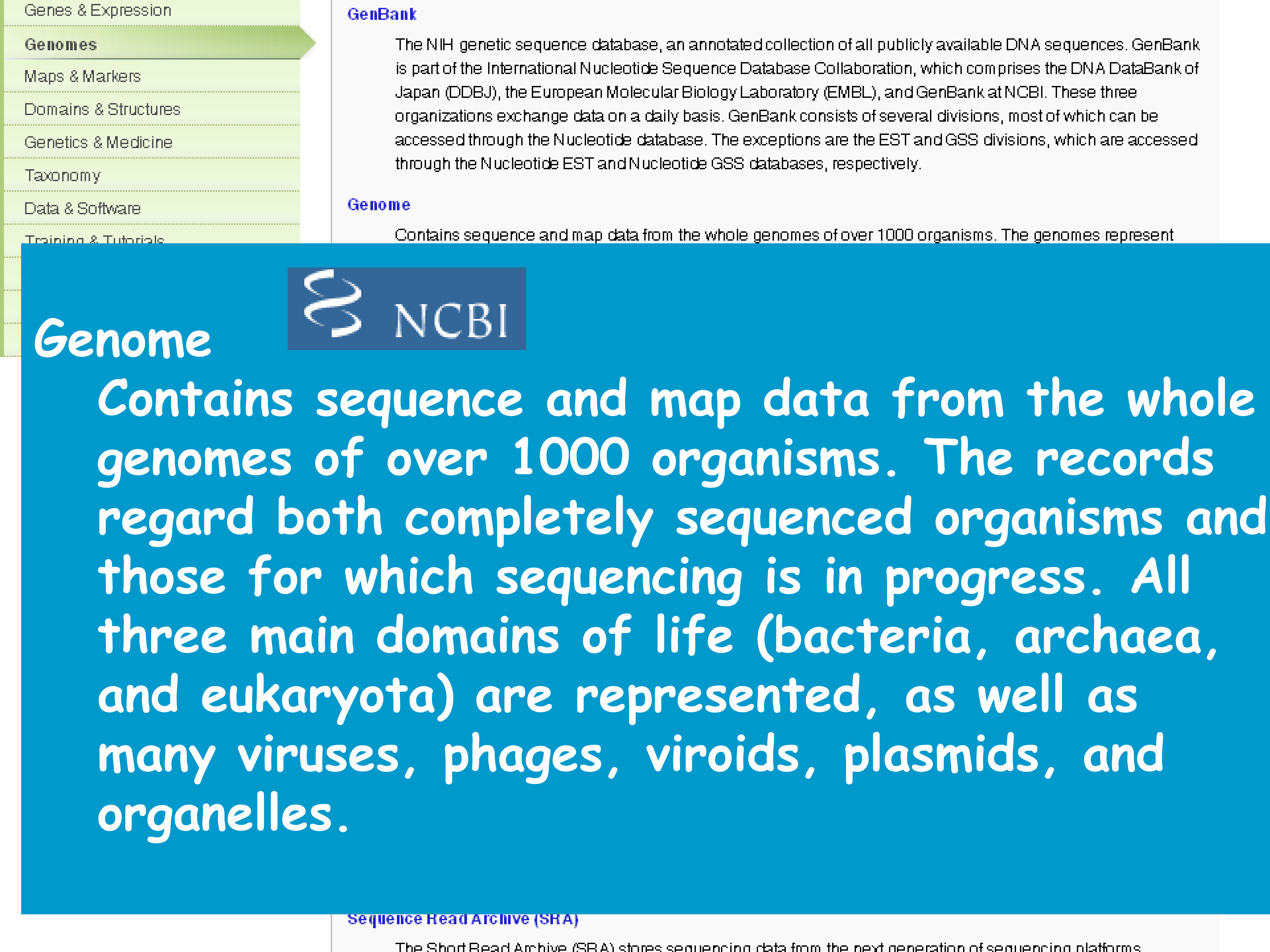
Genomic biology

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

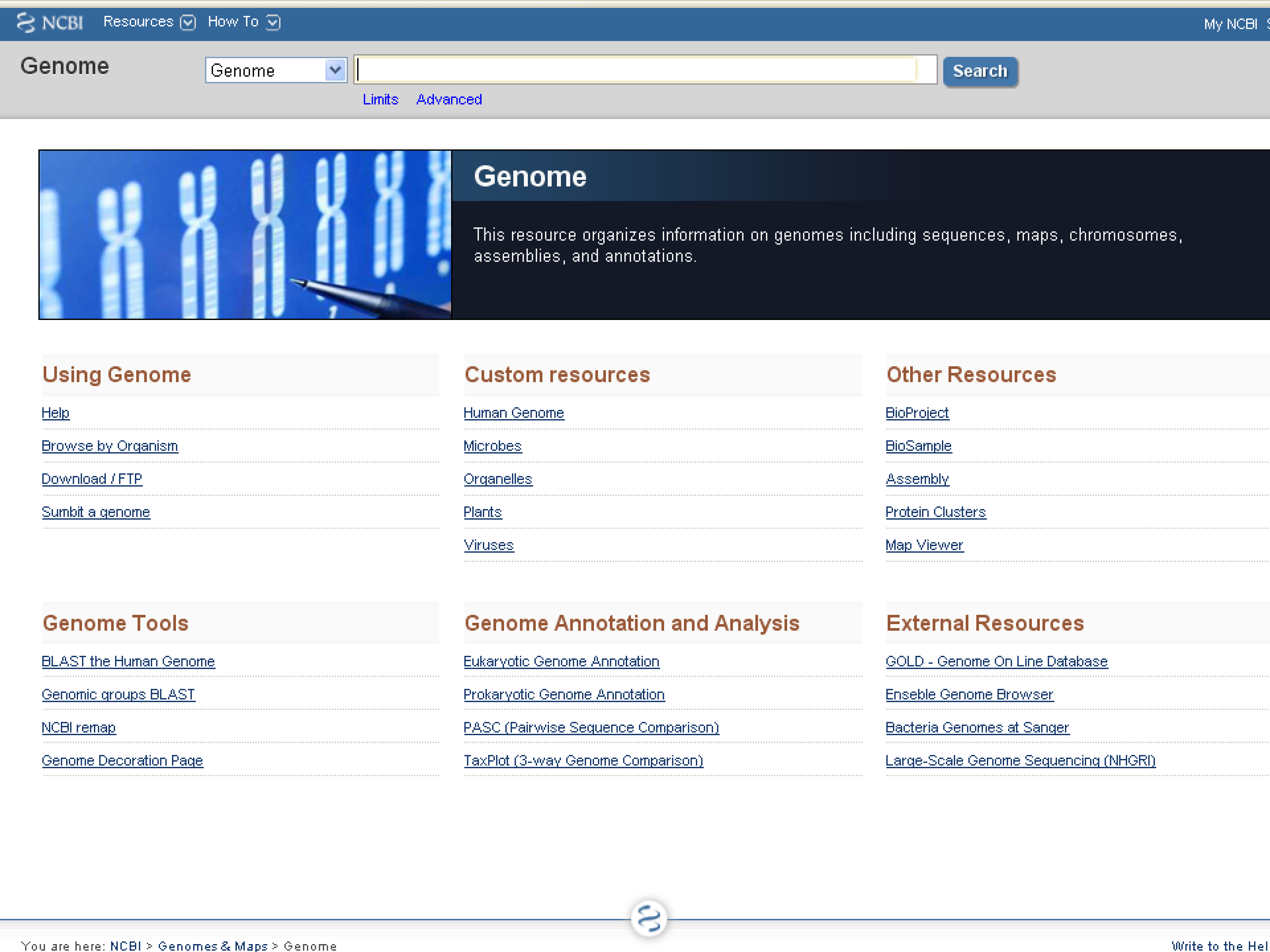
Sequence Read Archive (SRA)

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

Genome

NCBI

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.



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)

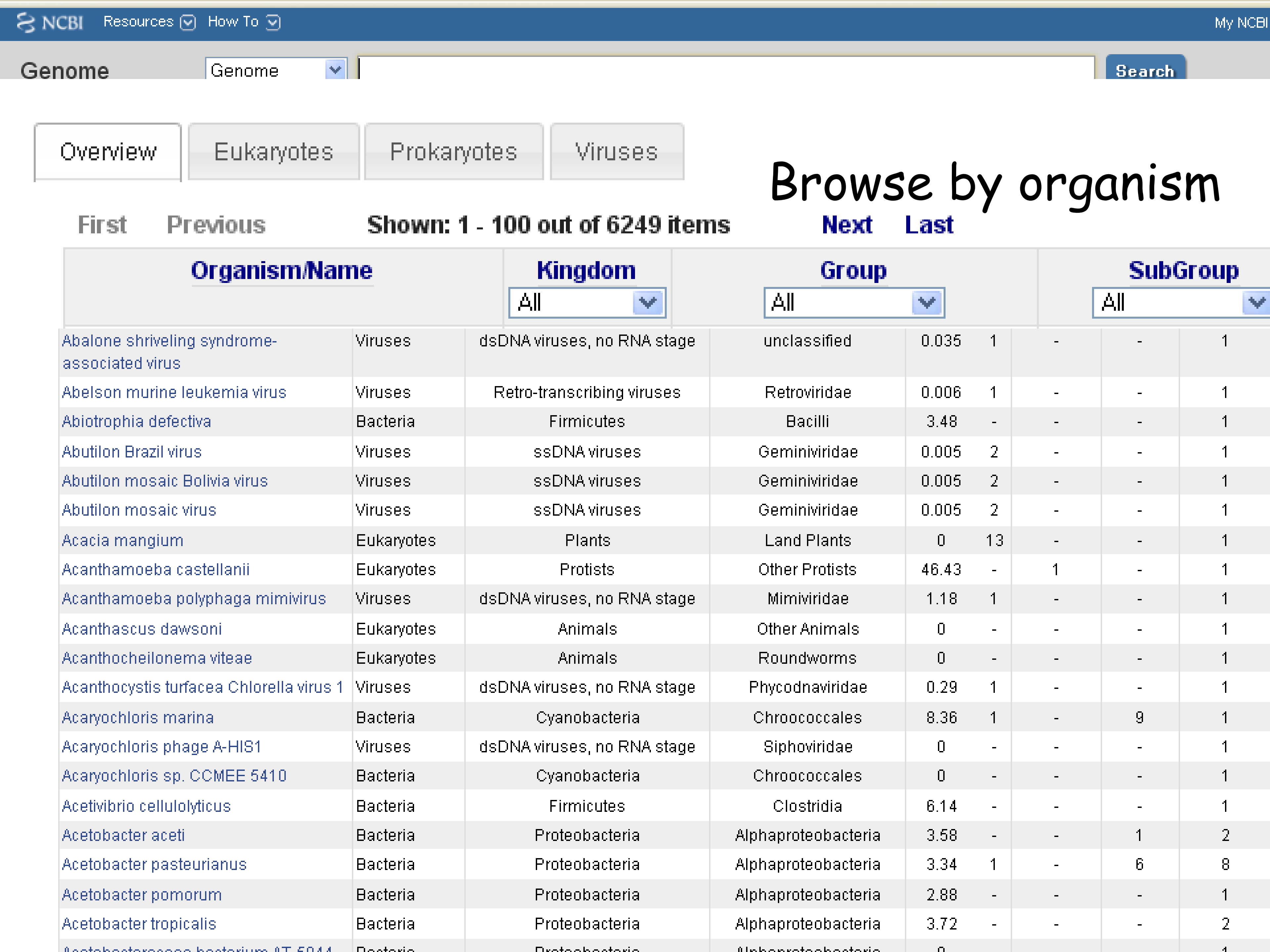
External Resources

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

You are here: NCBI > Genomes & Maps > Genome

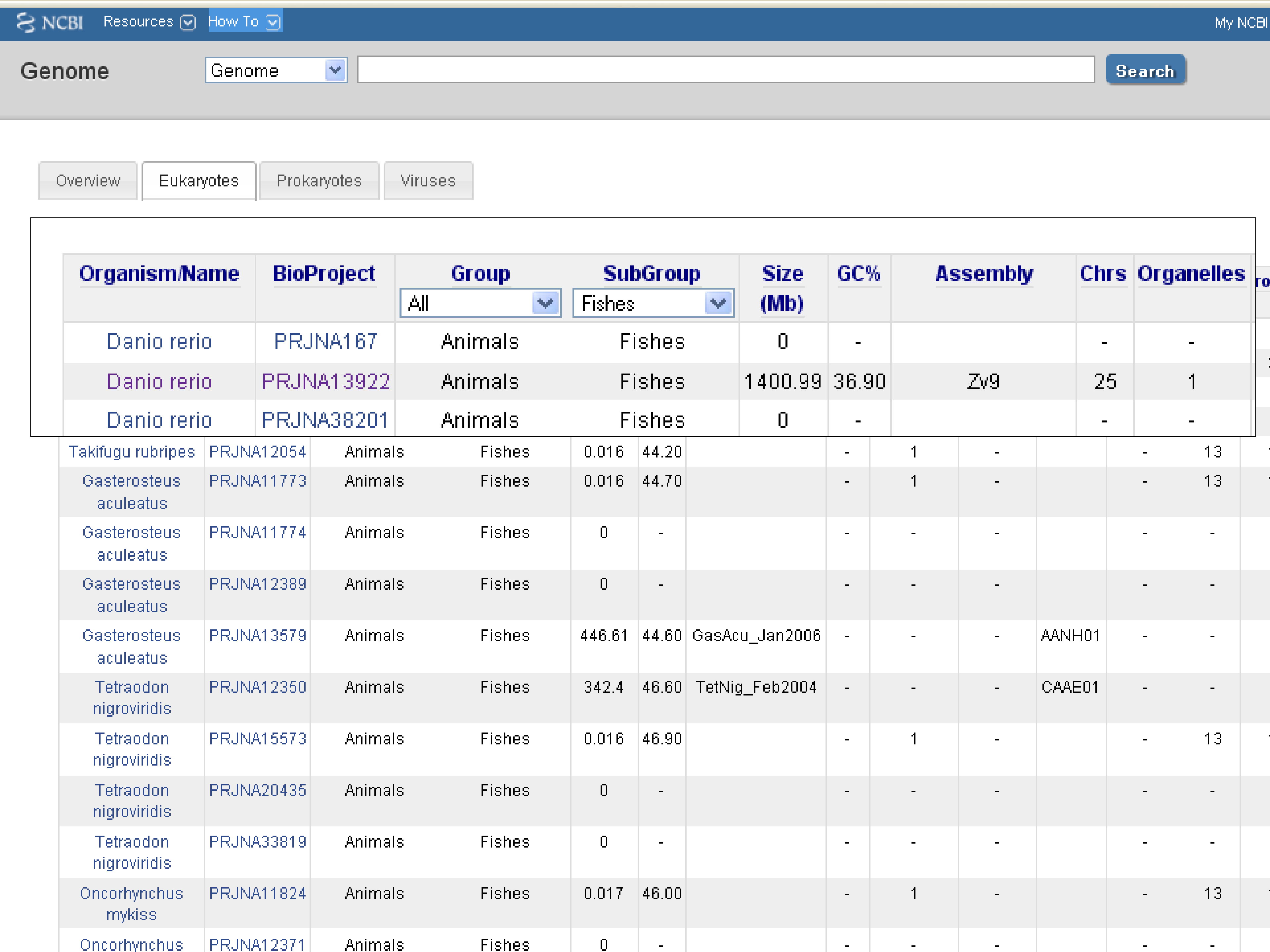
Write to the Hel

Browse by organism



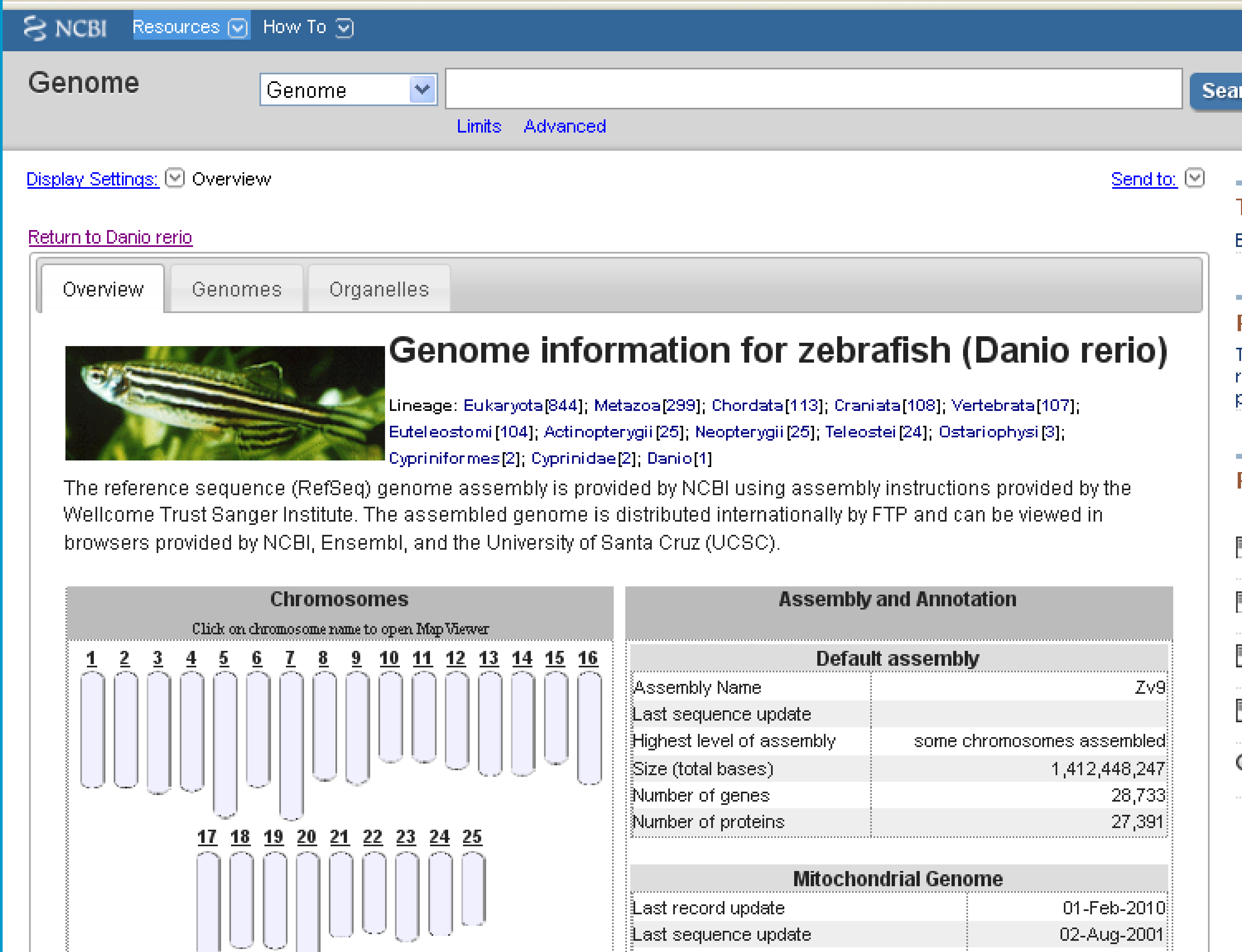
First Previous Shown: 1 - 100 out of 6249 items Next Last

Organism/Name	Kingdom	Group	SubGroup					
	All	All	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
Acetobacteraceae bacterium AT-5844	Bacteria	Proteobacteria	Alphaproteobacteria	0				1



Organism/Name	BioProject	Group All	SubGroup Fishes	Size (Mb)	GC%	Assembly	Chrs	Organelles	ro
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).

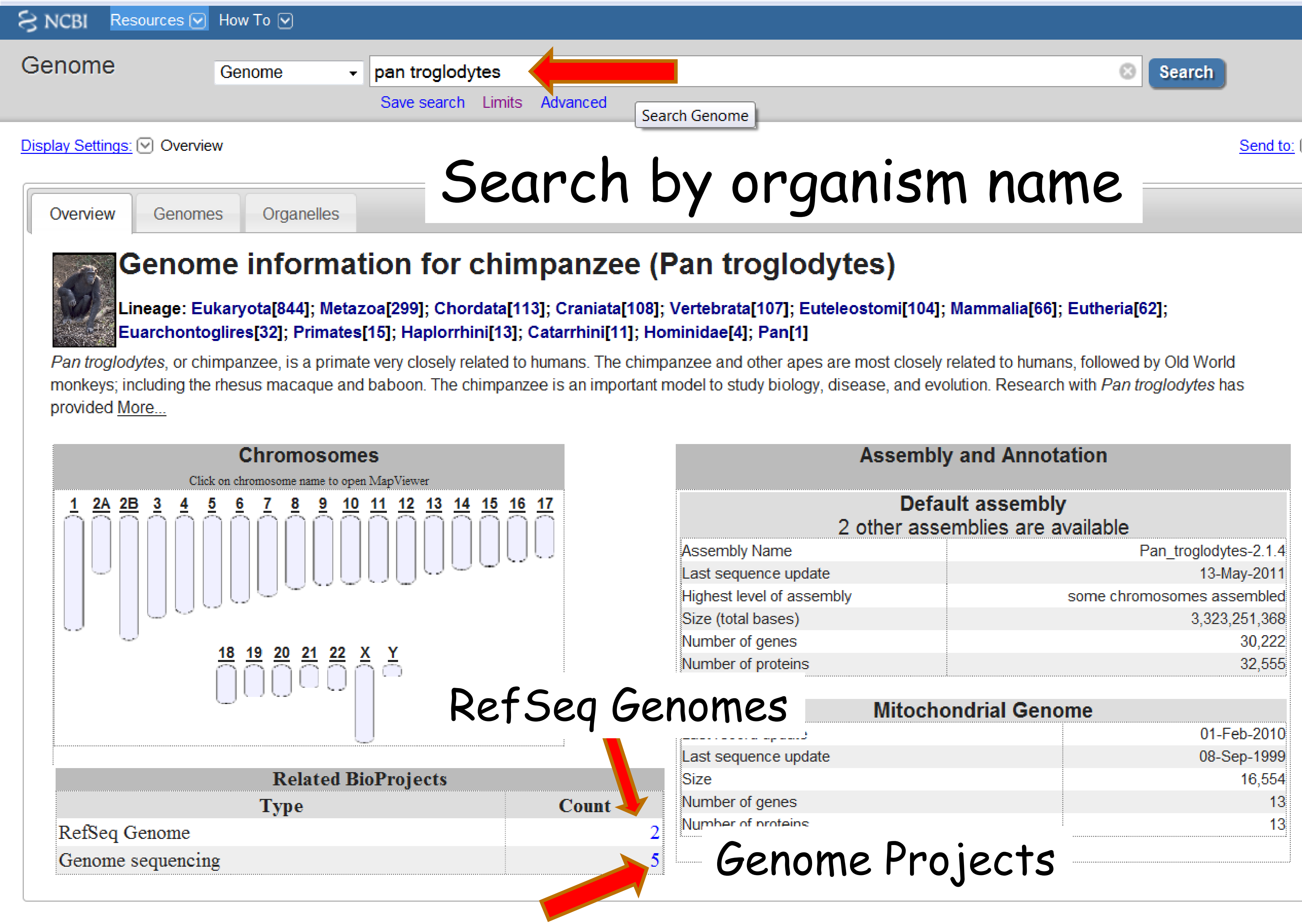
Chromosomes

Click on chromosome name to open Map Viewer

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



pan troglodytes

Search by organism name

Genome information for chimpanzee (Pan troglodytes)

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]; Pan[1]

Pan troglodytes, or chimpanzee, is a primate very closely related to humans. The chimpanzee and other apes are most closely related to humans, followed by Old World monkeys; including the rhesus macaque and baboon. The chimpanzee is an important model to study biology, disease, and evolution. Research with *Pan troglodytes* has provided More...

Chromosomes

Click on chromosome name to open MapViewer

Assembly and Annotation

Default assembly
2 other assemblies are available

Assembly Name	Pan_troglodytes-2.1.4
Last sequence update	13-May-2011
Highest level of assembly	some chromosomes assembled
Size (total bases)	3,323,251,368
Number of genes	30,222
Number of proteins	32,555

RefSeq Genomes

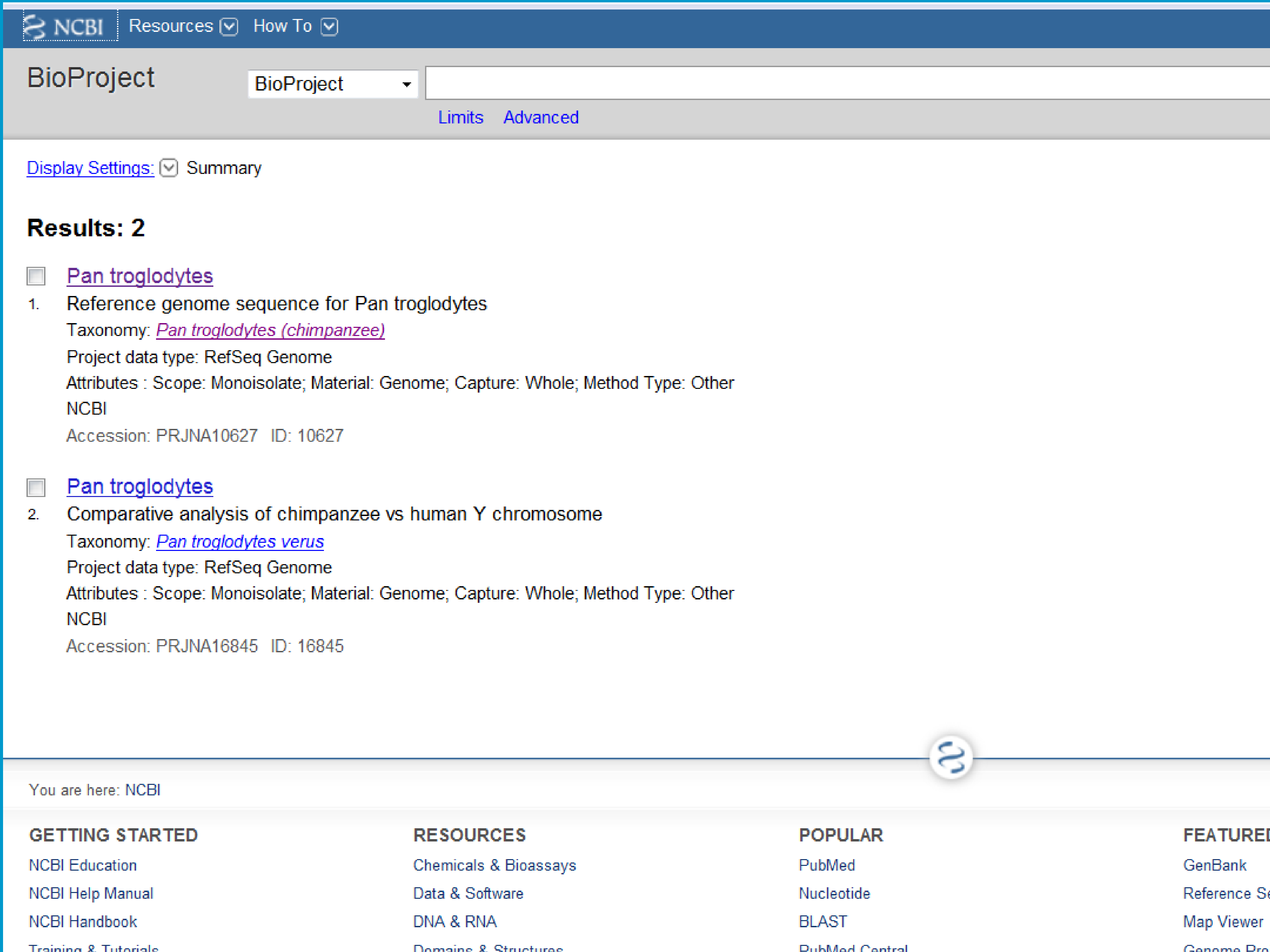
Mitochondrial Genome

Last record update	01-Feb-2010
Last sequence update	08-Sep-1999
Size	16,554
Number of genes	13
Number of proteins	13

Related BioProjects

Type	Count
RefSeq Genome	2
Genome sequencing	5

Genome Projects



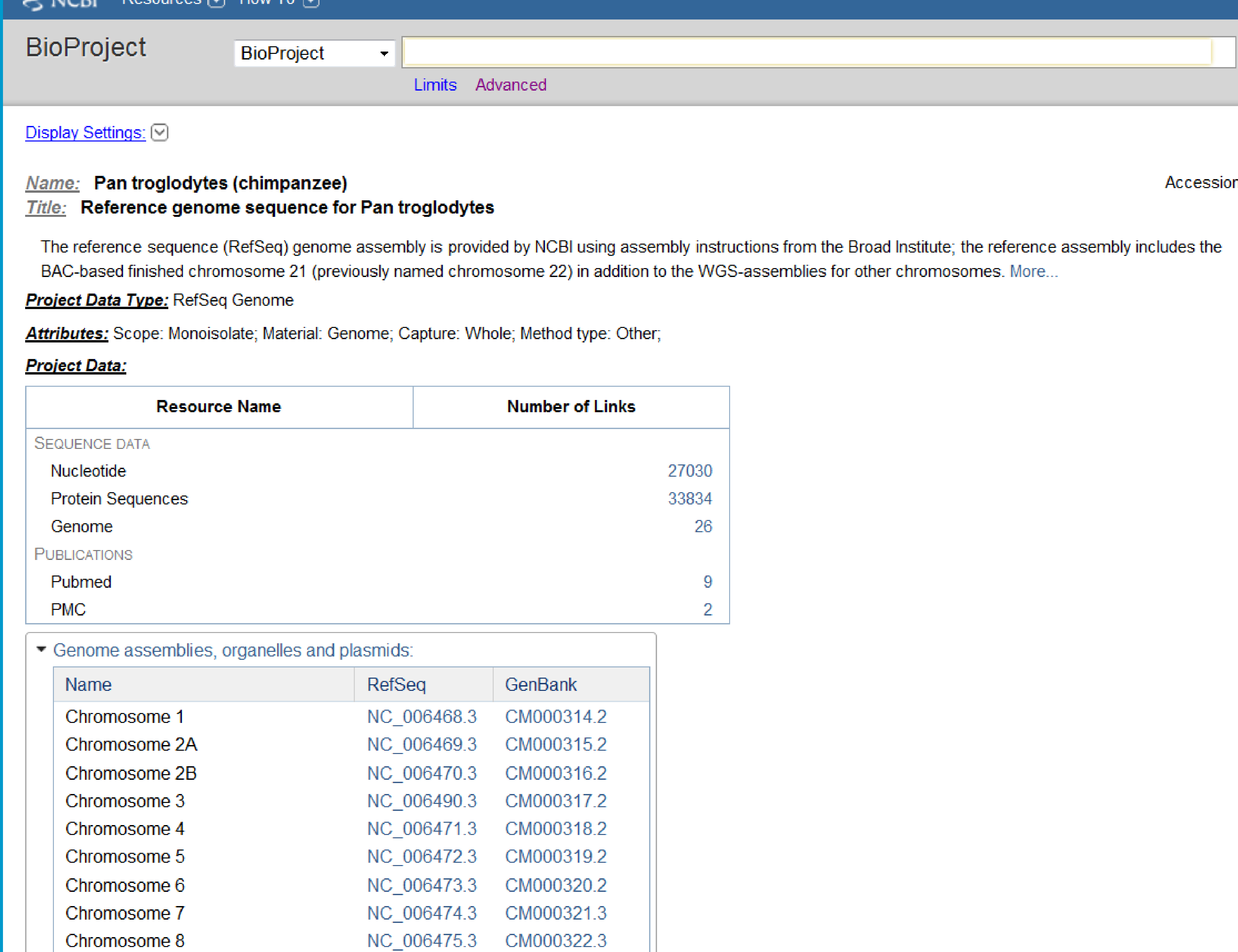
Results: 2

☐ Pan troglodytes

1. Reference genome sequence for Pan troglodytes
   Taxonomy: *Pan troglodytes (chimpanzee)*
   Project data type: RefSeq Genome
   Attributes : Scope: Monoisolate; Material: Genome; Capture: Whole; Method Type: Other
   NCBI
   Accession: PRJNA10627 ID: 10627

☐ Pan troglodytes

2. Comparative analysis of chimpanzee vs human Y chromosome
   Taxonomy: *Pan troglodytes verus*
   Project data type: RefSeq Genome
   Attributes : Scope: Monoisolate; Material: Genome; Capture: Whole; Method Type: Other
   NCBI
   Accession: PRJNA16845 ID: 16845



Name: **Pan troglodytes (chimpanzee)**

Title: **Reference genome sequence for Pan troglodytes**

The reference sequence (RefSeq) genome assembly is provided by NCBI using assembly instructions from the Broad Institute; the reference assembly includes the BAC-based finished chromosome 21 (previously named chromosome 22) in addition to the WGS-assemblies for other chromosomes. More...

Project Data Type: RefSeq Genome

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Other;

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide	27030
Protein Sequences	33834
Genome	26
PUBLICATIONS	
Pubmed	9
PMC	2

Genome assemblies, organelles and plasmids:

Name	RefSeq	GenBank
Chromosome 1	NC_006468.3	CM000314.2
Chromosome 2A	NC_006469.3	CM000315.2
Chromosome 2B	NC_006470.3	CM000316.2
Chromosome 3	NC_006490.3	CM000317.2
Chromosome 4	NC_006471.3	CM000318.2
Chromosome 5	NC_006472.3	CM000319.2
Chromosome 6	NC_006473.3	CM000320.2
Chromosome 7	NC_006474.3	CM000321.3
Chromosome 8	NC_006475.3	CM000322.3

The Human Genome Project (HGP)

Was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health.

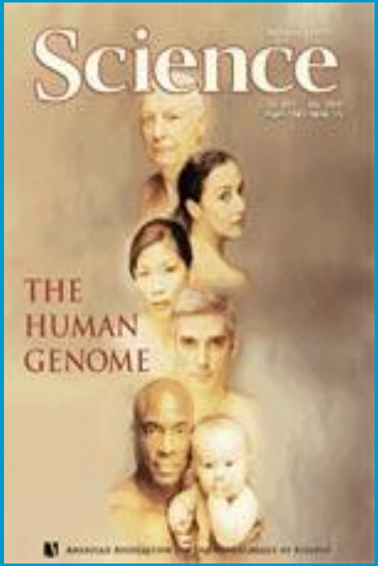
The Wellcome Trust (U.K.) became a major partner.

Additional contributions came from Japan, France, Germany, China, and others.

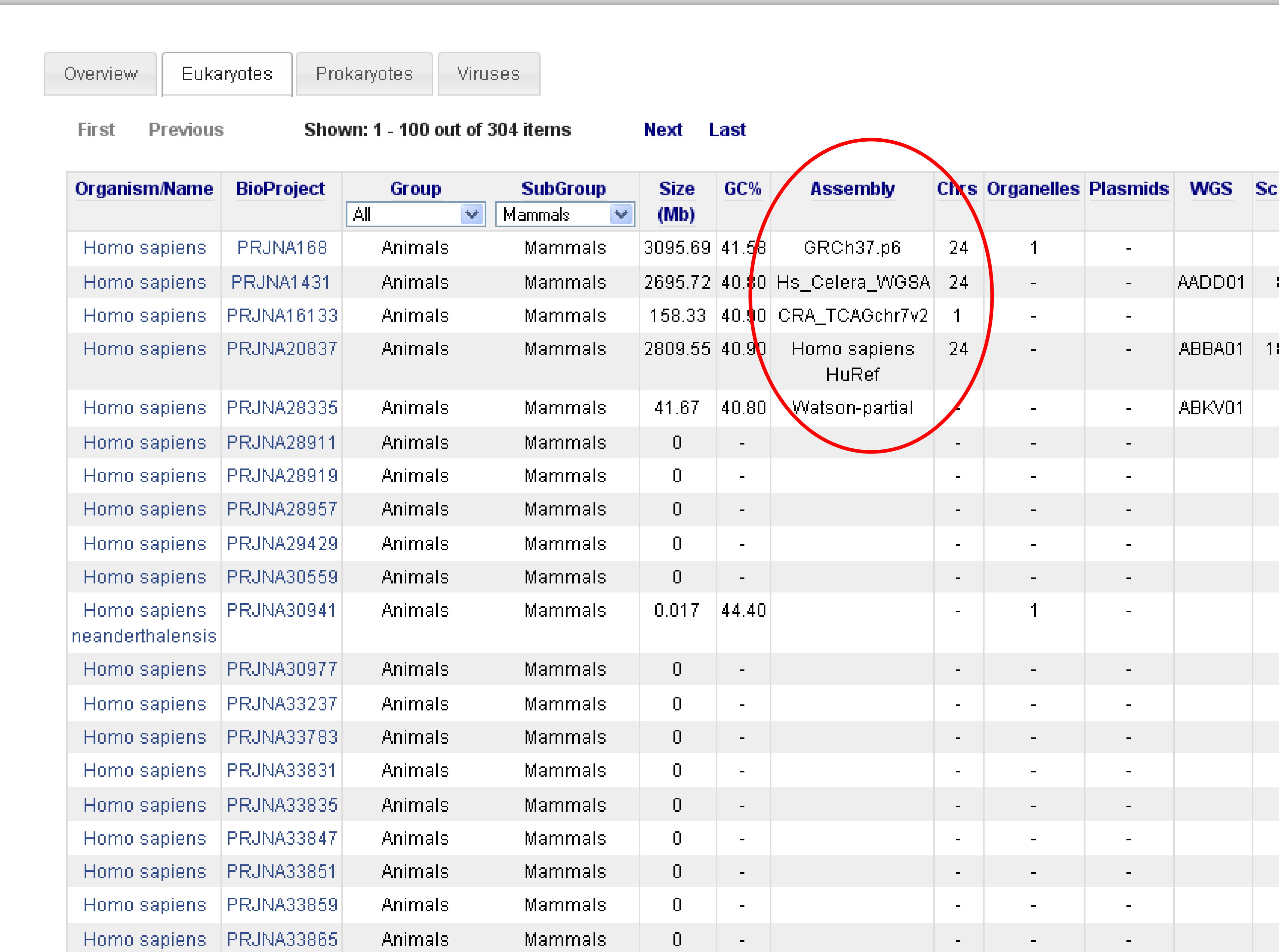
Project goals were to

- *identify* all the human genes (20,000-25,000),
- *determine* the sequences of the 3 billion base pairs,
- *store* this information in databases,
- *improve* tools for data analysis,

CELERA

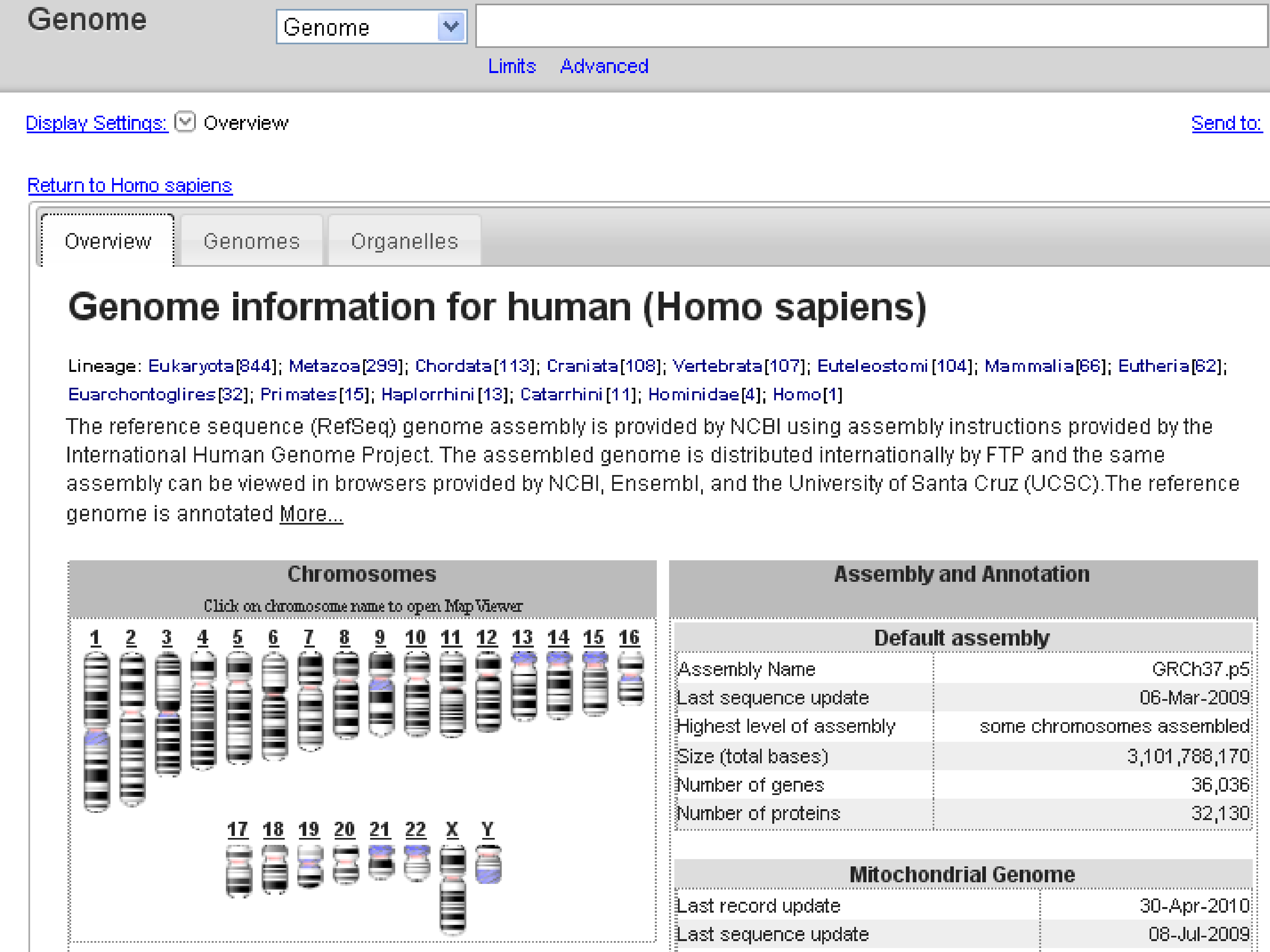


16 February 2001
Vol 291, Issue 5507, Pages 1145-1434



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

Organism/Name	BioProject	Group	SubGroup	Size (Mb)	GC%	Assembly	Chrs	Organelles	Plasmids	WGS	Sc
		All	Mammals								
Homo sapiens	PRJNA168	Animals	Mammals	3095.69	41.58	GRCh37.p6	24	1	-		
Homo sapiens	PRJNA1431	Animals	Mammals	2695.72	40.80	Hs_Celera_WGSA	24	-	-	AADD01	
Homo sapiens	PRJNA16133	Animals	Mammals	158.33	40.90	CRA_TCAGchr7v2	1	-	-		
Homo sapiens	PRJNA20837	Animals	Mammals	2809.55	40.90	Homo sapiens HuRef	24	-	-	ABBA01	1
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	-		
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

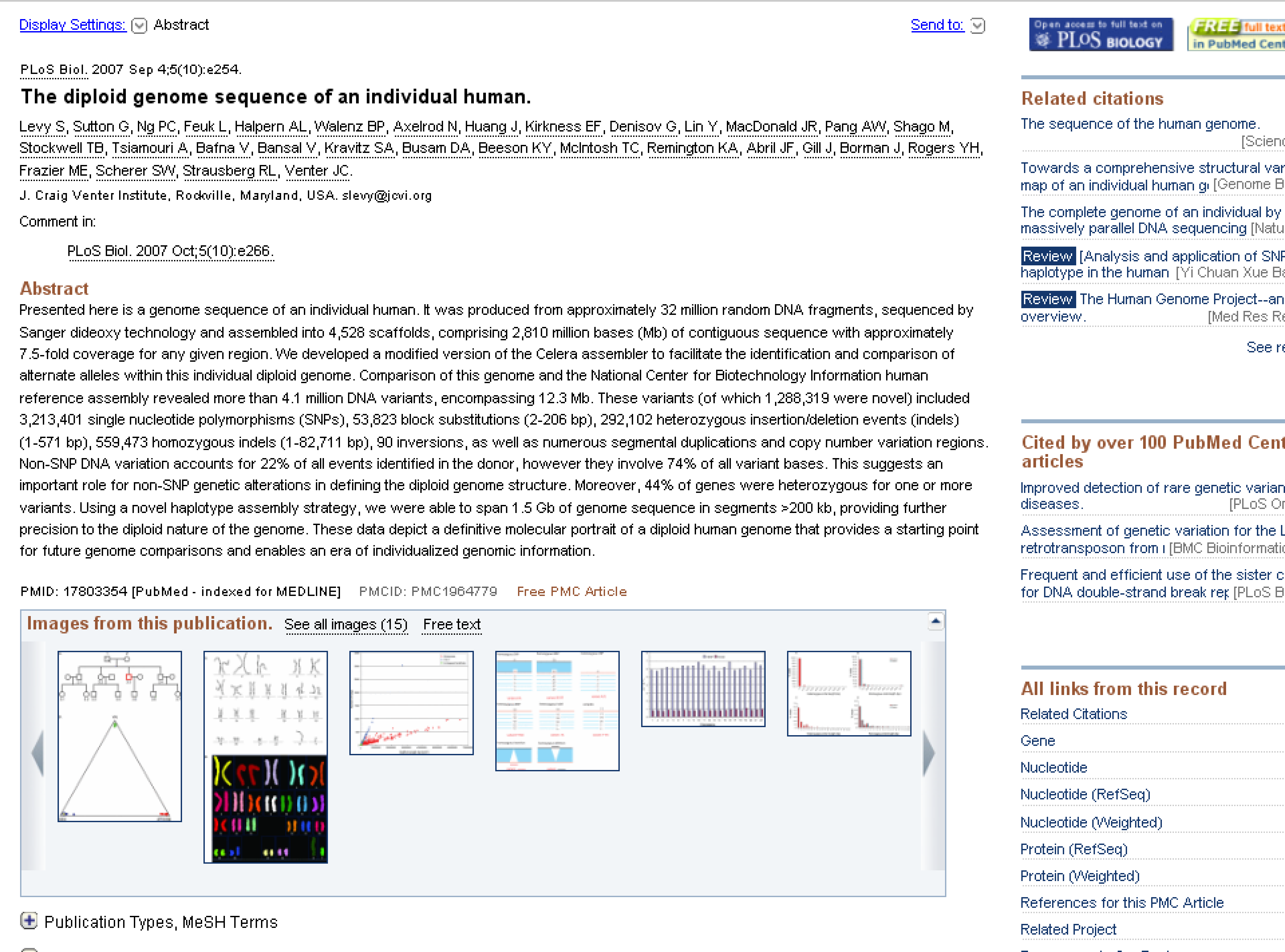
Mitochondrial Genome

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

J. Craig Venter™
INSTITUTE

HuRef Genome

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



Display Settings: Abstract

Send to:

Open access to full text on PLoS BIOLOGY
FREE full text in PubMed Cent

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 [BMC Bioinformati

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)

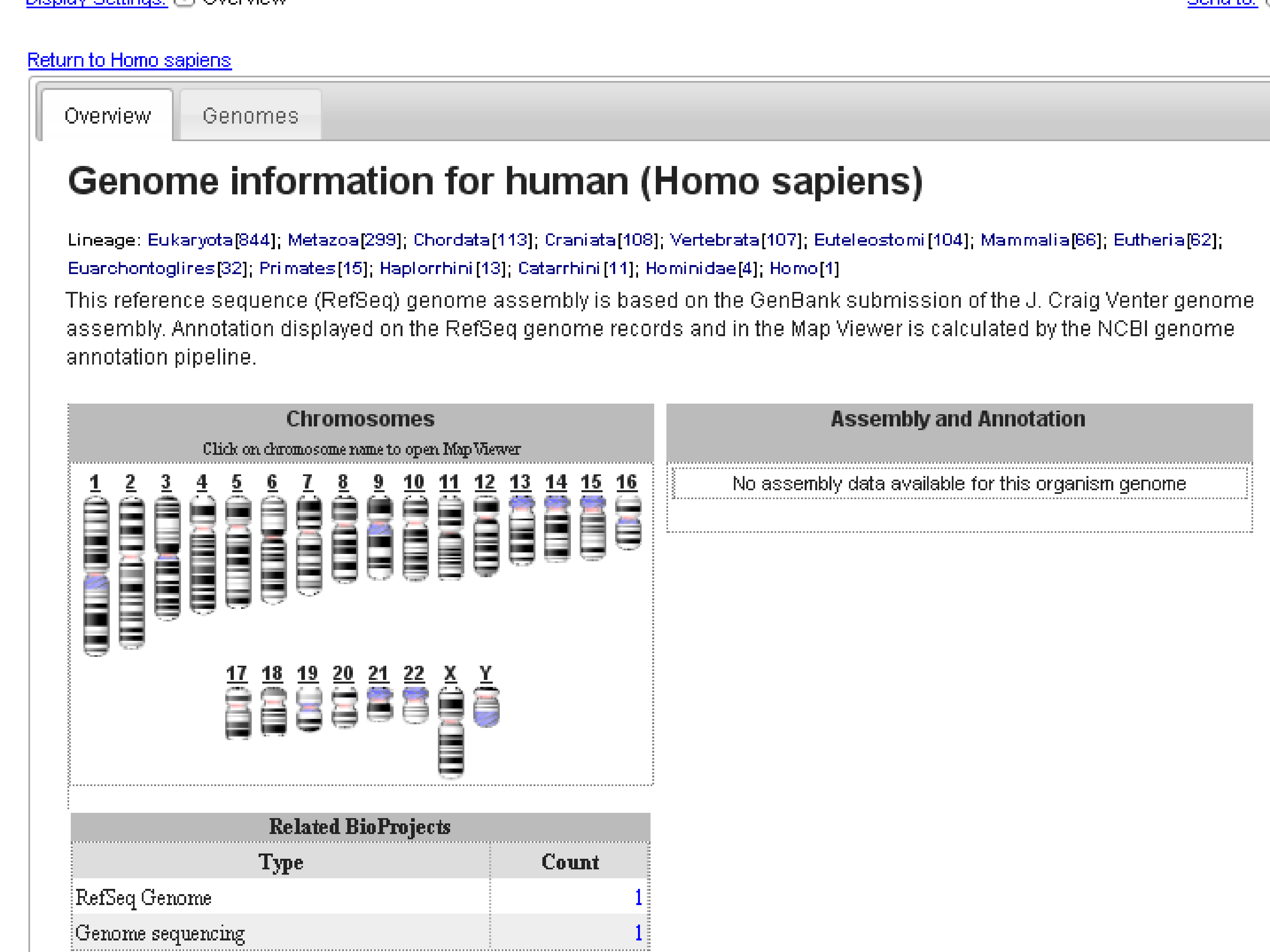
References for this PMC Article

Related Project

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

All the data for the first human diploid genome has been deposited at NCBI.

The highest scoring allele is represented in the consensus sequence.



Display Settings: ☐ Overview

Send to:

Return to Homo sapiens

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

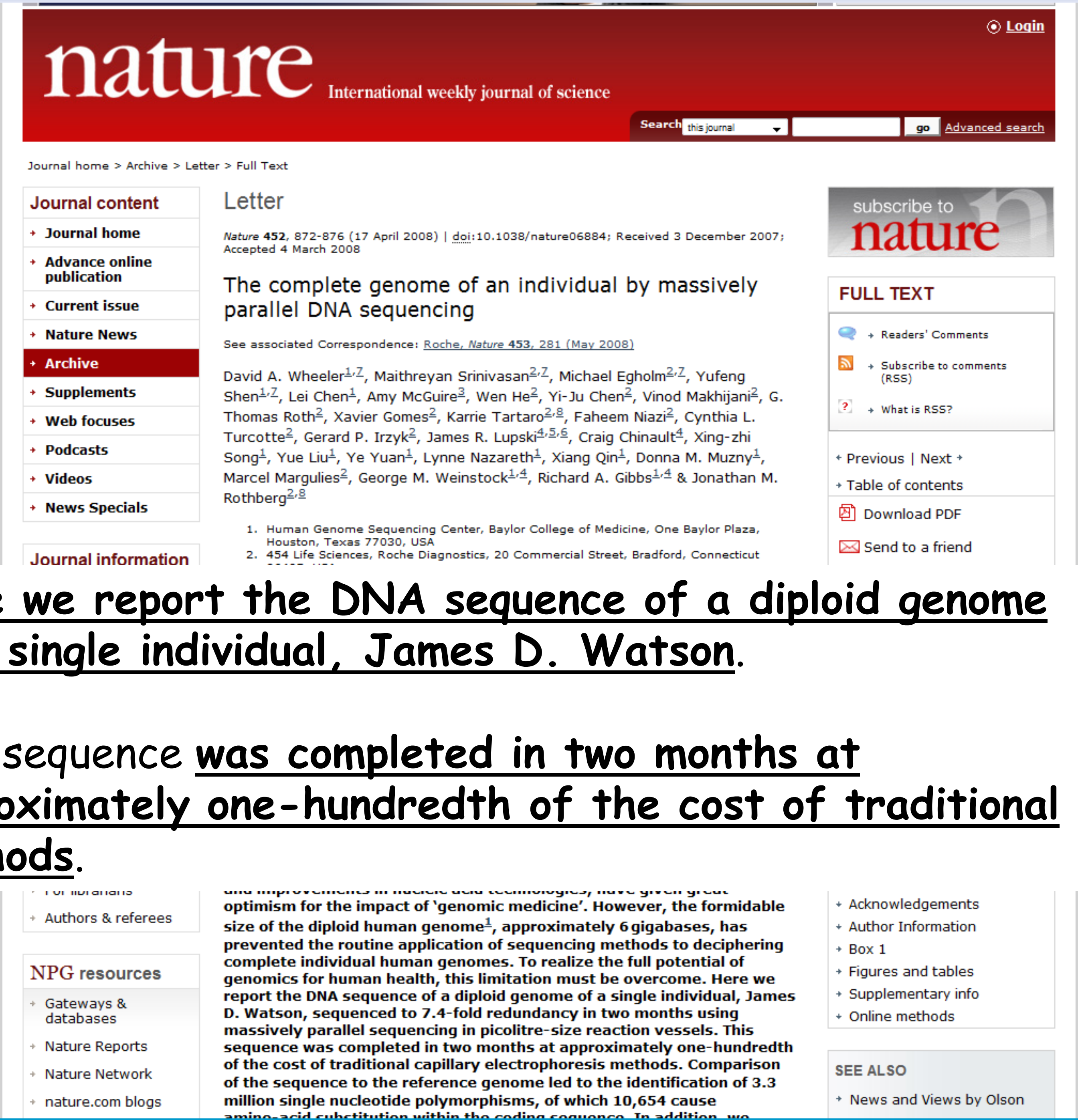
Click on chromosome name to open Map Viewer

Assembly and Annotation

No assembly data available for this organism genome

Related BioProjects

Type	Count
RefSeq Genome	1
Genome sequencing	1



Journal home > Archive > Letter > Full Text

Letter

Nature **452**, 872-876 (17 April 2008) | doi:10.1038/nature06884; Received 3 December 2007; Accepted 4 March 2008

The complete genome of an individual by massively parallel DNA sequencing

See associated Correspondence: Roche, *Nature* **453**, 281 (May 2008)

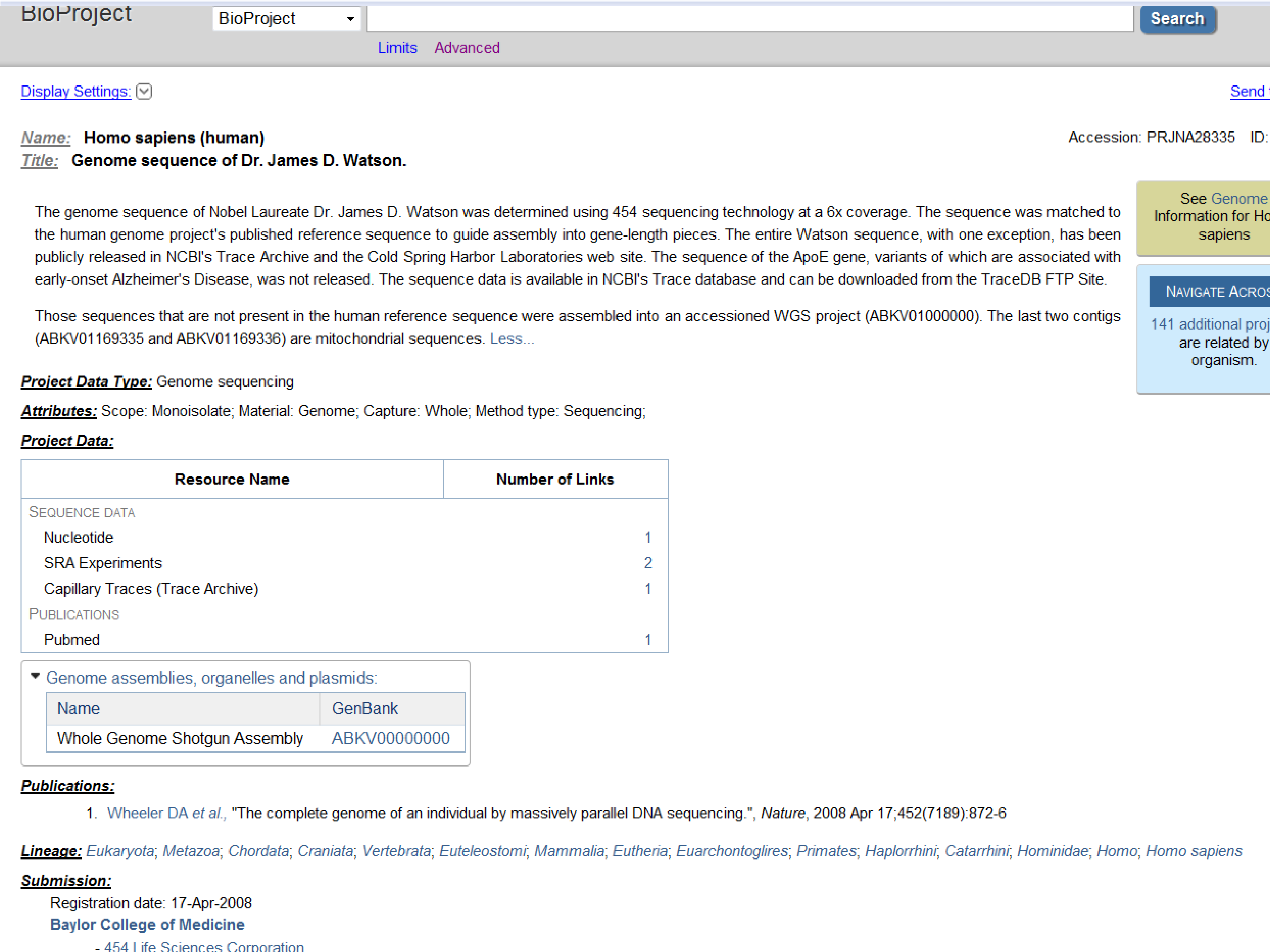
David A. Wheeler[1,7], Maithreyan Srinivasan[2,7], Michael Egholm[2,7], Yufeng Shen[1,7], Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2,8], Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2,8]

1. Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA
2. 454 Life Sciences, Roche Diagnostics, 20 Commercial Street, Bradford, Connecticut

Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson.

This sequence **was completed in two months at approximately one-hundredth of the cost of traditional methods**.

and improvements in nucleic acid technologies, have given great optimism for the impact of 'genomic medicine'. However, the formidable size of the diploid human genome[1], approximately 6 gigabases, has prevented the routine application of sequencing methods to deciphering complete individual human genomes. To realize the full potential of genomics for human health, this limitation must be overcome. Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in two months using massively parallel sequencing in picolitre-size reaction vessels. This sequence was completed in two months at approximately one-hundredth of the cost of traditional capillary electrophoresis methods. Comparison of the sequence to the reference genome led to the identification of 3.3 million single nucleotide polymorphisms, of which 10,654 cause amino-acid substitution within the coding sequence. In addition, we



Name: **Homo sapiens (human)**

Accession: PRJNA28335 ID:

Title: **Genome sequence of Dr. James D. Watson.**

The genome sequence of Nobel Laureate Dr. James D. Watson was determined using 454 sequencing technology at a 6x coverage. The sequence was matched to the human genome project's published reference sequence to guide assembly into gene-length pieces. The entire Watson sequence, with one exception, has been publicly released in NCBI's Trace Archive and the Cold Spring Harbor Laboratories web site. The sequence of the ApoE gene, variants of which are associated with early-onset Alzheimer's Disease, was not released. The sequence data is available in NCBI's Trace database and can be downloaded from the TraceDB FTP Site.

Those sequences that are not present in the human reference sequence were assembled into an accessioned WGS project (ABKV01000000). The last two contigs (ABKV01169335 and ABKV01169336) are mitochondrial sequences. Less...

See Genome Information for Ho sapiens

NAVIGATE ACROS

141 additional proj are related by organism.

Project Data Type: Genome sequencing

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing;

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide	1
SRA Experiments	2
Capillary Traces (Trace Archive)	1
PUBLICATIONS	
Pubmed	1

▼ Genome assemblies, organelles and plasmids:

Name	GenBank
Whole Genome Shotgun Assembly	ABKV00000000

Publications:

1. Wheeler DA *et al.*, "The complete genome of an individual by massively parallel DNA sequencing.", *Nature*, 2008 Apr 17;452(7189):872-6

Lineage: *Eukaryota*; *Metazoa*; *Chordata*; *Craniata*; *Vertebrata*; *Euteleostomi*; *Mammalia*; *Eutheria*; *Euarchontoglires*; *Primates*; *Haplorrhini*; *Catarrhini*; *Hominidae*; *Homo*; *Homo sapiens*

Submission:

Registration date: 17-Apr-2008

Baylor College of Medicine

- 454 Life Sciences Corporation

The Ensembl Project

- European Bioinformatics Institute (EBI),
- 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

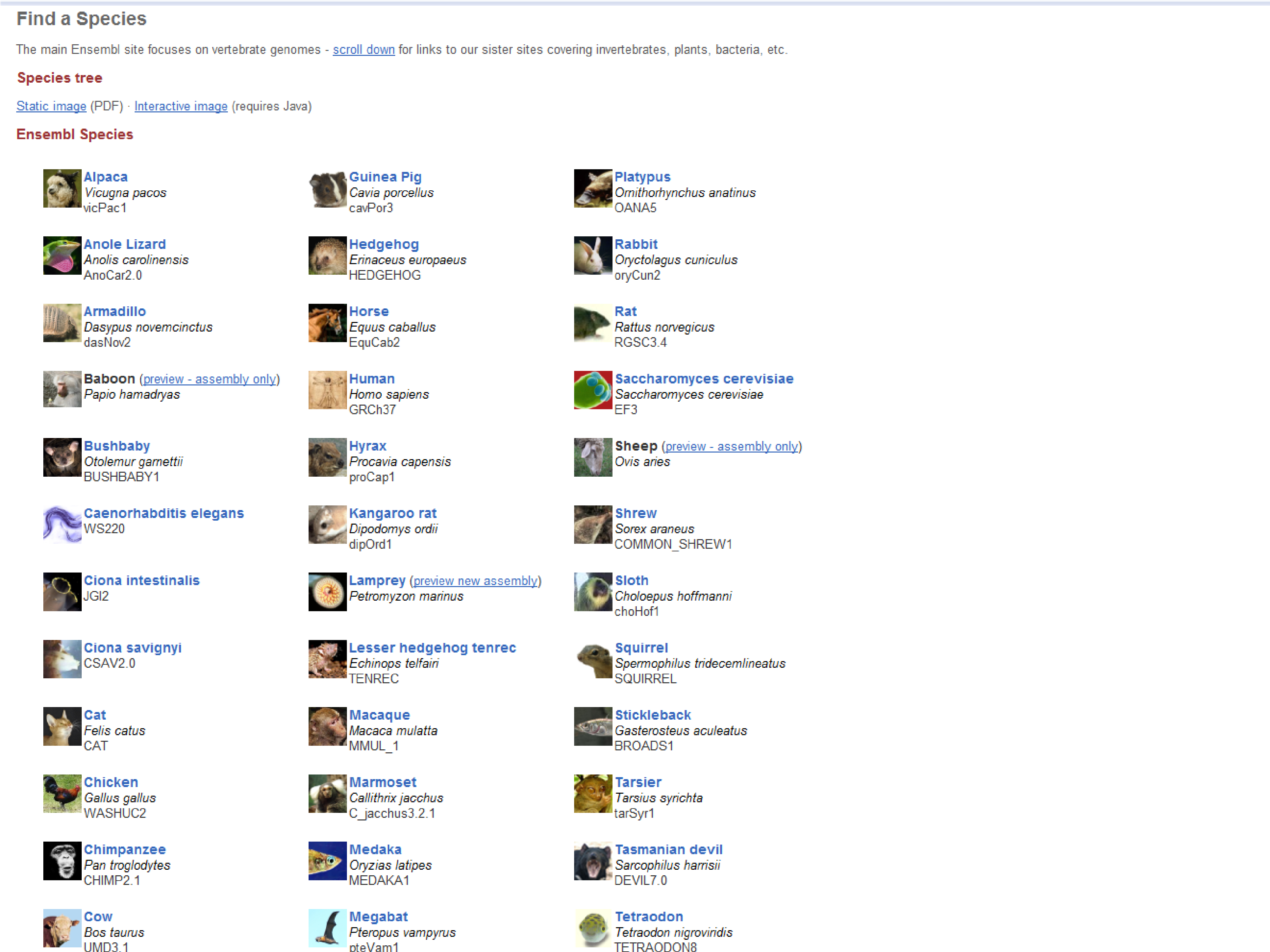
The Ensembl Project

was started in 1999, some years before the draft human genome was completed.

The goal was to automatically annotate the genome and integrate this annotation with other available biological data.

- Since the website's launch in 2000, many more genomes have been added;

- the available information 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

Anole Lizard
Anolis carolinensis
AnoCar2.0

Armadillo
Dasypus novemcinctus
dasNov2

Baboon (preview - assembly only)
Papio hamadryas

Bushbaby
Otolemur garnettii
BUSHBABY1

Caenorhabditis elegans
WS220

Ciona intestinalis
JGI2

Ciona savignyi
CSAV2.0

Cat
Felis catus
CAT

Chicken
Gallus gallus
WASHUC2

Chimpanzee
Pan troglodytes
CHIMP2.1

Cow
Bos taurus
UMD3.1

Guinea Pig
Cavia porcellus
cavPor3

Hedgehog
Erinaceus europaeus
HEDGEHOG

Horse
Equus caballus
EquCab2

Human
Homo sapiens
GRCh37

Hyrax
Procavia capensis
proCap1

Kangaroo rat
Dipodomys ordii
dipOrd1

Lamprey (preview new assembly)
Petromyzon marinus

Lesser hedgehog tenrec
Echinops telfairi
TENREC

Macaque
Macaca mulatta
MMUL_1

Marmoset
Callithrix jacchus
C_jacchus3.2.1

Medaka
Oryzias latipes
MEDAKA1

Megabat
Pteropus vampyrus
pteVam1

Platypus
Ornithorhynchus anatinus
OANA5

Rabbit
Oryctolagus cuniculus
oryCun2

Rat
Rattus norvegicus
RGSC3.4

Saccharomyces cerevisiae
Saccharomyces cerevisiae
EF3

Sheep (preview - assembly only)
Ovis aries

Shrew
Sorex araneus
COMMON_SHREW1

Sloth
Choloepus hoffmanni
choHof1

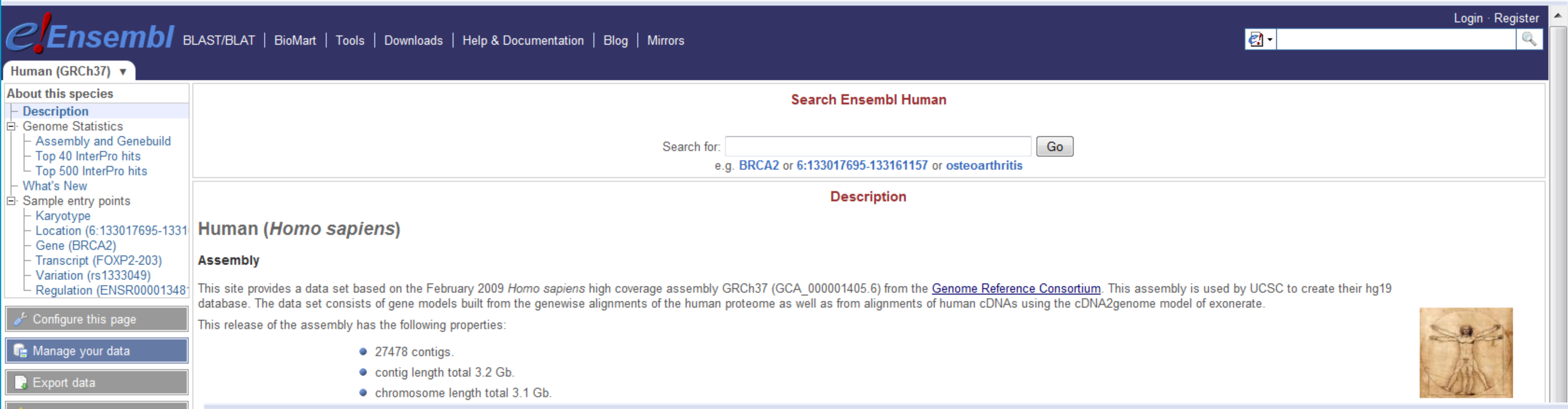
Squirrel
Spermophilus tridecemlineatus
SQUIRREL

Stickleback
Gasterosteus aculeatus
BROADS1

Tarsier
Tarsius syrichta
tarSyr1

Tasmanian devil
Sarcophilus harrisii
DEVIL7.0

Tetraodon
Tetraodon nigroviridis
TETRAODON8



Search Ensembl Human

Search for: Go

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

Description

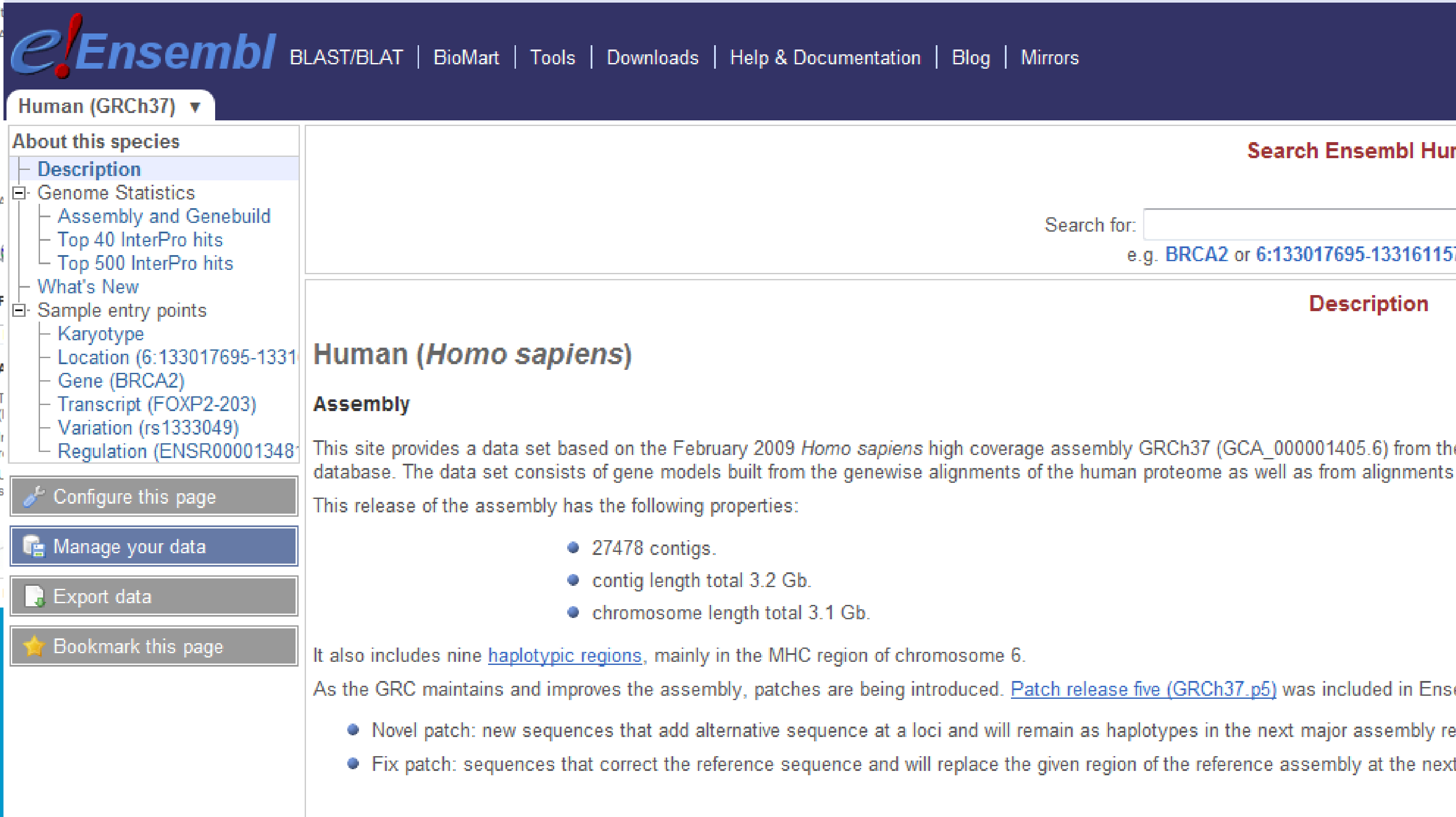
Human (*Homo sapiens*)

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.



Search Ensembl Hur

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 the database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments

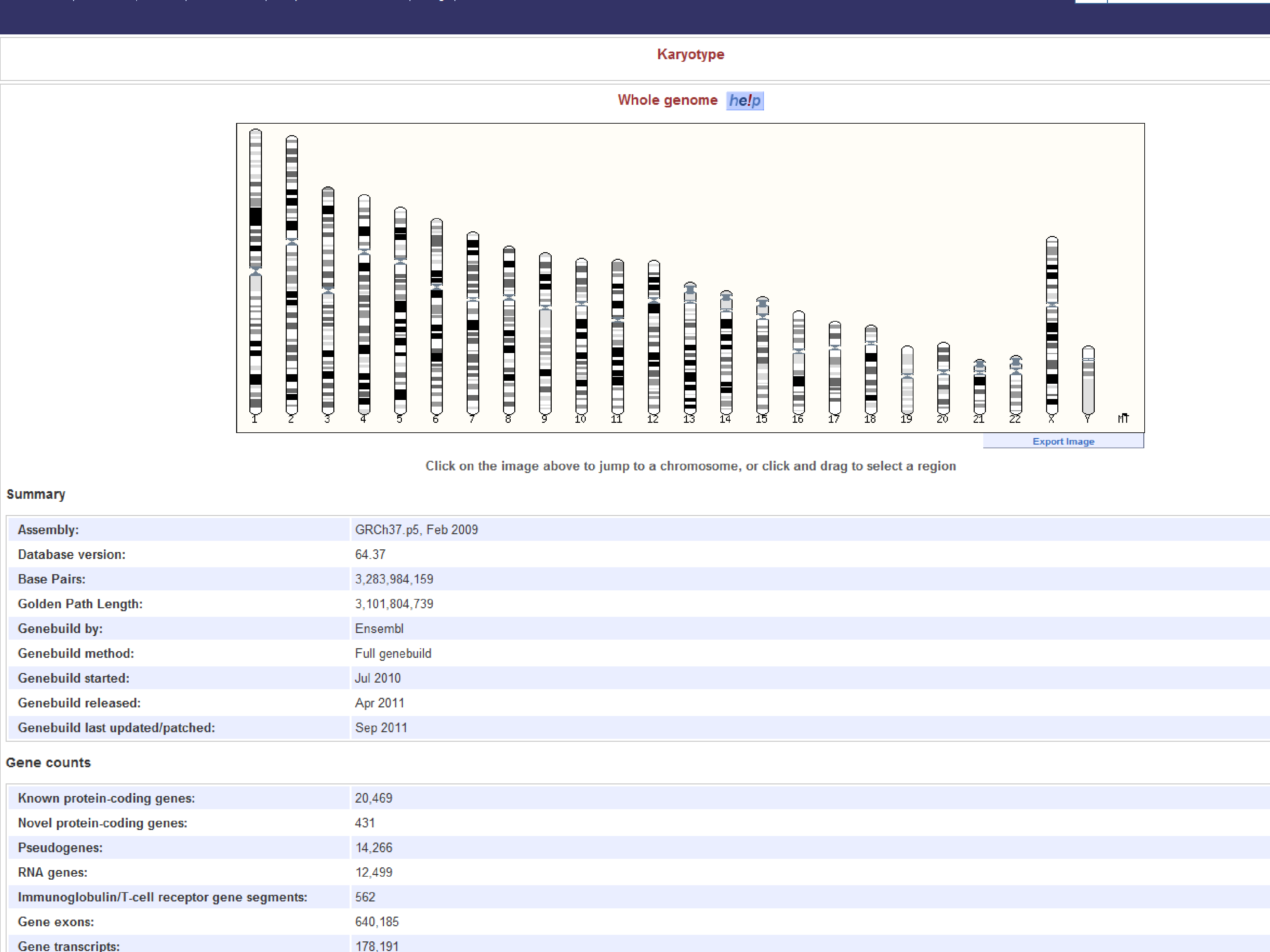
This release of the assembly has the following properties:

- 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 Ense

- 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



Karyotype

Whole genome

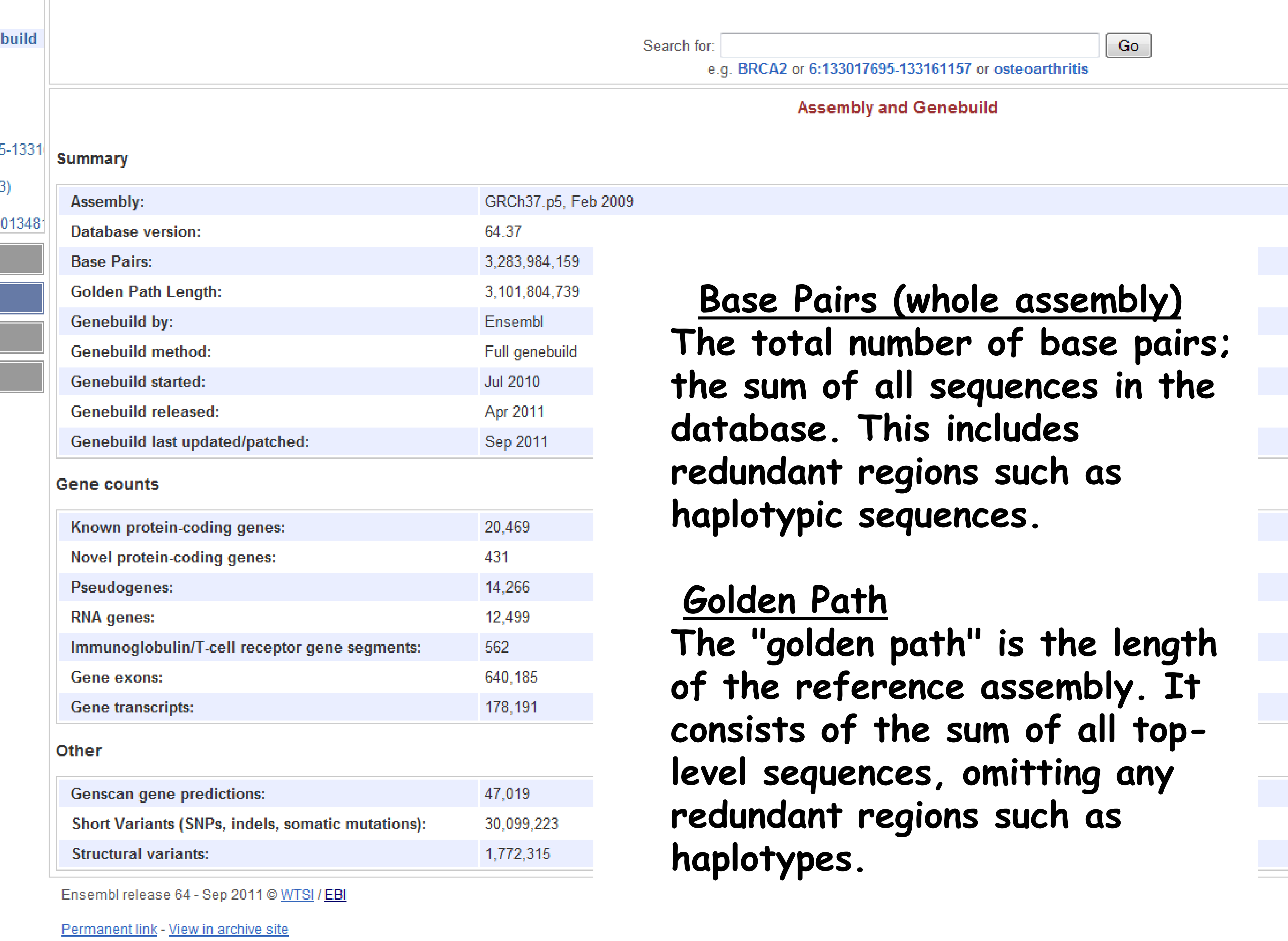
Click on the image above to jump to a chromosome, or click and drag to select a region

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



Assembly and Genebuild

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

Other

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

Ensembl release 64 - Sep 2011 © WTSI / EBI

Permanent link - View in archive site

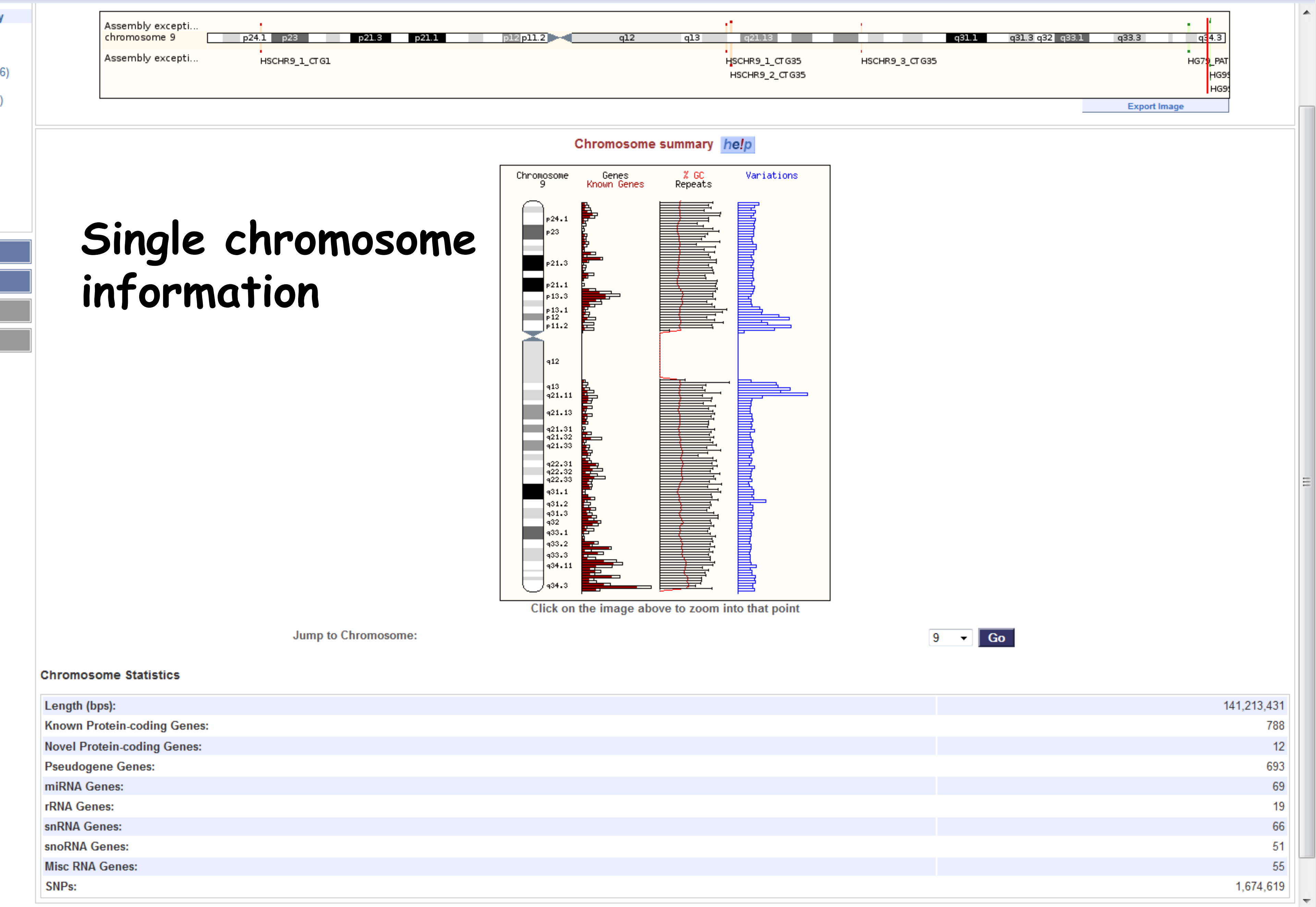
Base Pairs (whole assembly)

The total number of base pairs; the sum of all sequences in the database. This includes redundant regions such as haplotypic sequences.

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.

Single chromosome information



Chromosome summary he!p

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

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

wellcome trust sanger institute

The Wellcome Trust Sanger Institute

THE Genome CENTER AT WASHINGTON UNIVERSITY

The Genome Center at Washington University

EMBL-EBI

The European Bioinformatics Institute

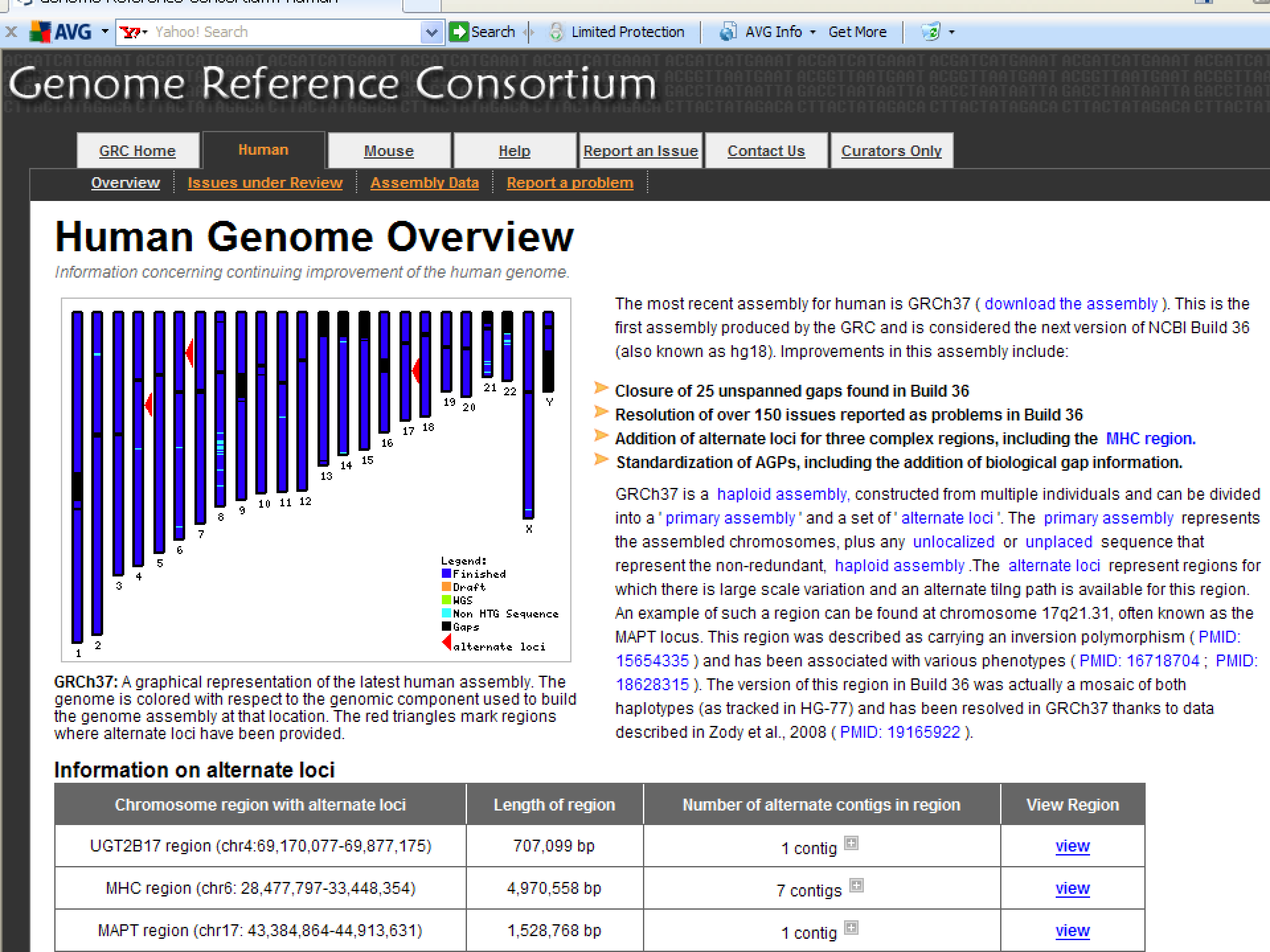
NCBI

The National Center for Biotechnology Information

Focused on the human and mouse reference assemblies to close gaps, fix errors and represent complex variation.

The Genome Reference Consortium (GRC)

The gap regions are so variable that they are best represented by multiple sequences



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

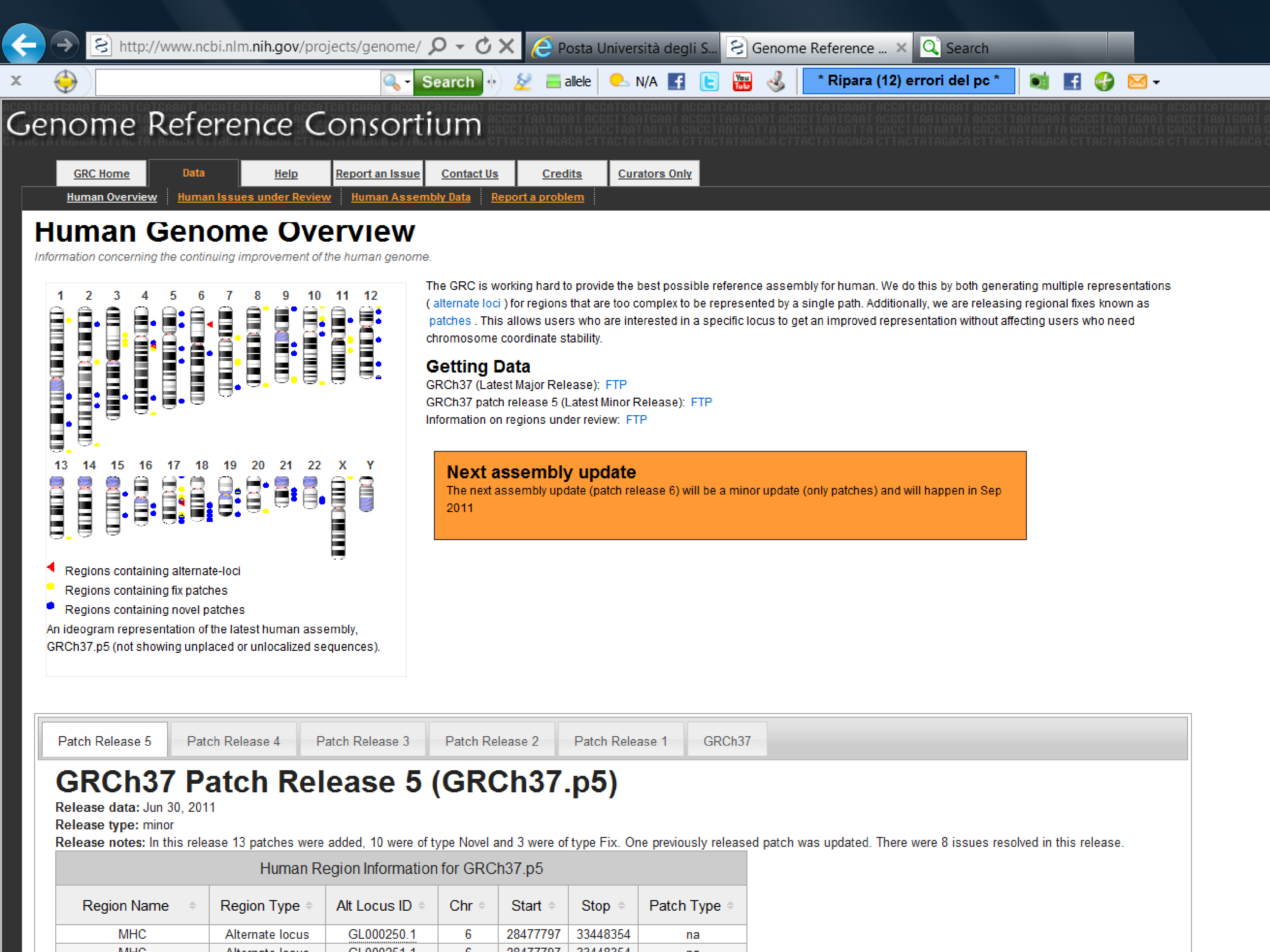
- is a haploid assembly, constructed from multiple individuals and can be divided into a primary assembly and a set of alternate loci.
- The primary assembly represents the assembled chromosomes, plus any unlocalized or unplaced sequence that represent the non-redundant, haploid assembly.
- The alternate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

An example of alternate loci

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.


http://www.ncbi.nlm.nih.gov/projects/genome/
Posta Università degli S...
Genome Reference ...
Search
Search
allele
N/A
* Ripara (12) errori del pc *

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, GRCh37.p5 (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 5 (Latest Minor Release): FTP

Information on regions under review: FTP

Next assembly update

The next assembly update (patch release 6) will be a minor update (only patches) and will happen in Sep 2011

GRCh37 Patch Release 5 (GRCh37.p5)

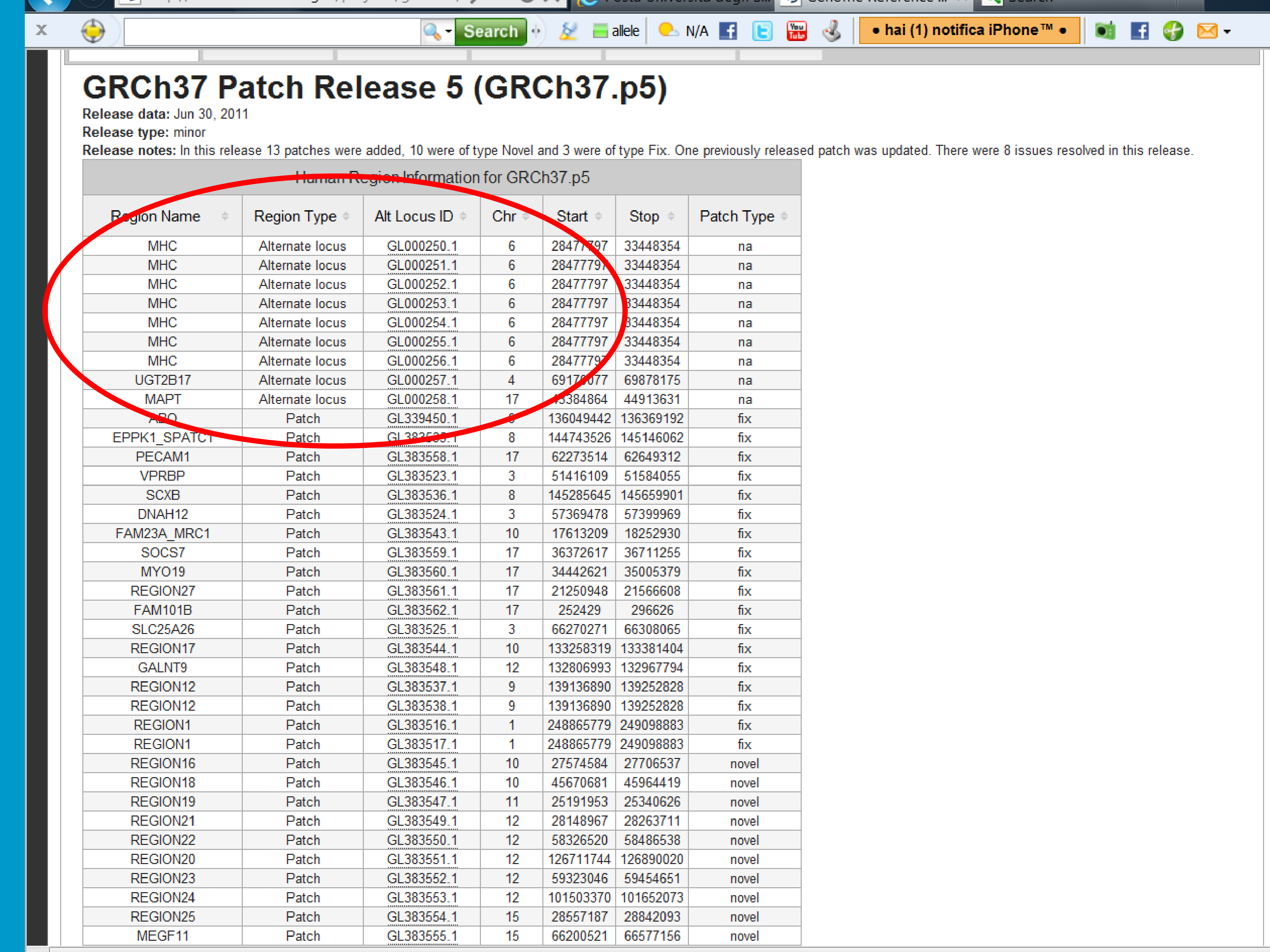
Release data: Jun 30, 2011

Release type: minor

Release notes: In this release 13 patches were added, 10 were of type Novel and 3 were of type Fix. One previously released patch was updated. There were 8 issues resolved in this release.

Human Region Information for GRCh37.p5

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



GRCh37 Patch Release 5 (GRCh37.p5)

Release data: Jun 30, 2011

Release type: minor

Release notes: In this release 13 patches were added, 10 were of type Novel and 3 were of type Fix. One previously released patch was updated. There were 8 issues resolved in this release.

Human Region Information for GRCh37.p5

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
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
REGION1	Patch	GL383516.1	1	248865779	249098883	fix
REGION1	Patch	GL383517.1	1	248865779	249098883	fix
REGION16	Patch	GL383545.1	10	27574584	27706537	novel
REGION18	Patch	GL383546.1	10	45670681	45964419	novel
REGION19	Patch	GL383547.1	11	25191953	25340626	novel
REGION21	Patch	GL383549.1	12	28148967	28263711	novel
REGION22	Patch	GL383550.1	12	58326520	58486538	novel
REGION20	Patch	GL383551.1	12	126711744	126890020	novel
REGION23	Patch	GL383552.1	12	59323046	59454651	novel
REGION24	Patch	GL383553.1	12	101503370	101652073	novel
REGION25	Patch	GL383554.1	15	28557187	28842093	novel
MEGF11	Patch	GL383555.1	15	66200521	66577156	novel

patch releases

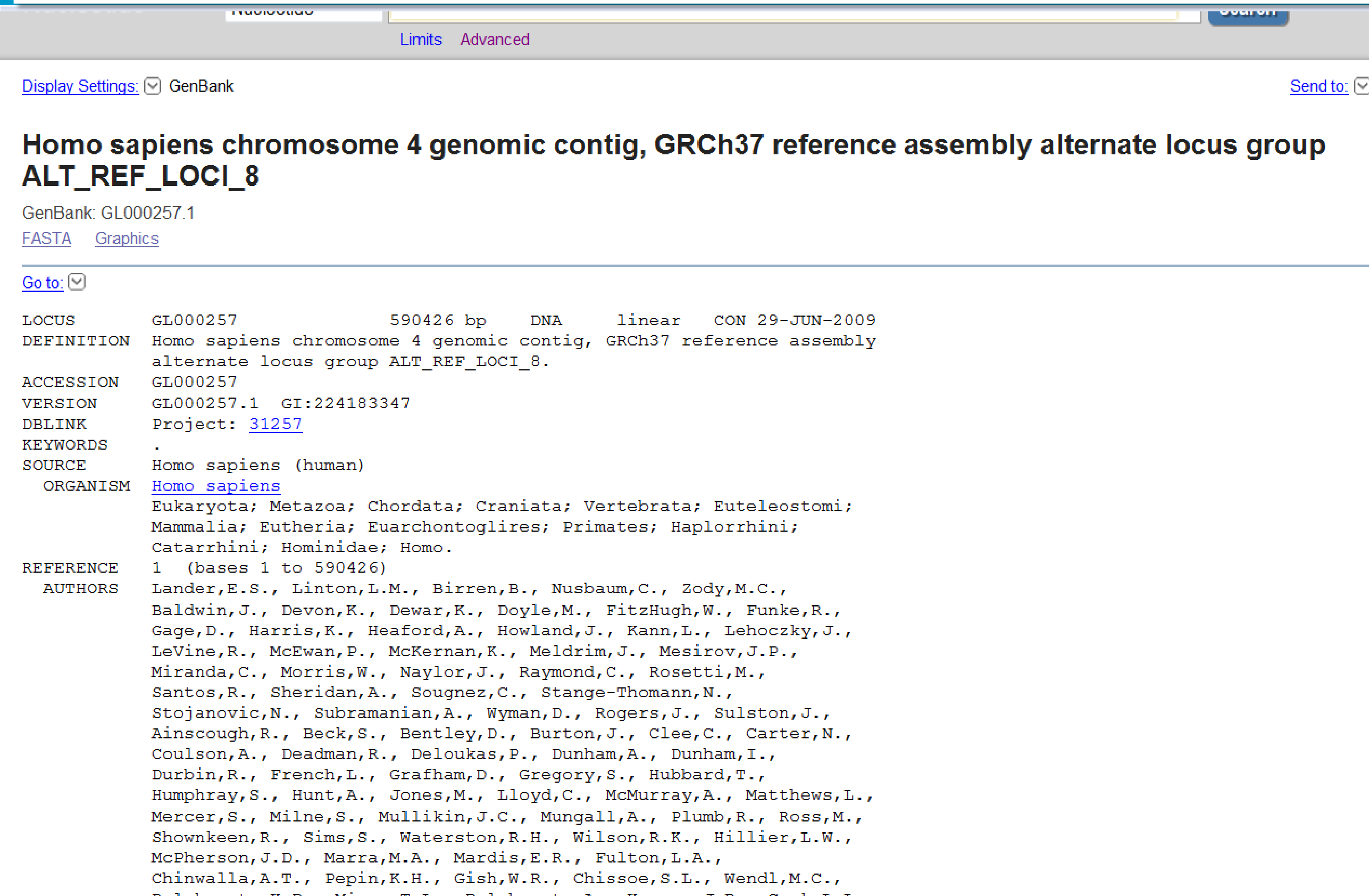
provide updated information for a particular region without changing the chromosome coordinates.

Patches are small bits of sequences which can be aligned to the Primary Assembly.

'Fix' patches represent improved regions. Are released to correct an error in the assembly and will be removed when the new full assembly is released.

‘Novel’ patches represent new alternate loci not in the last full assembly release and will be retained in the next full assembly release.

An Alternate Loci Nucleotide Record



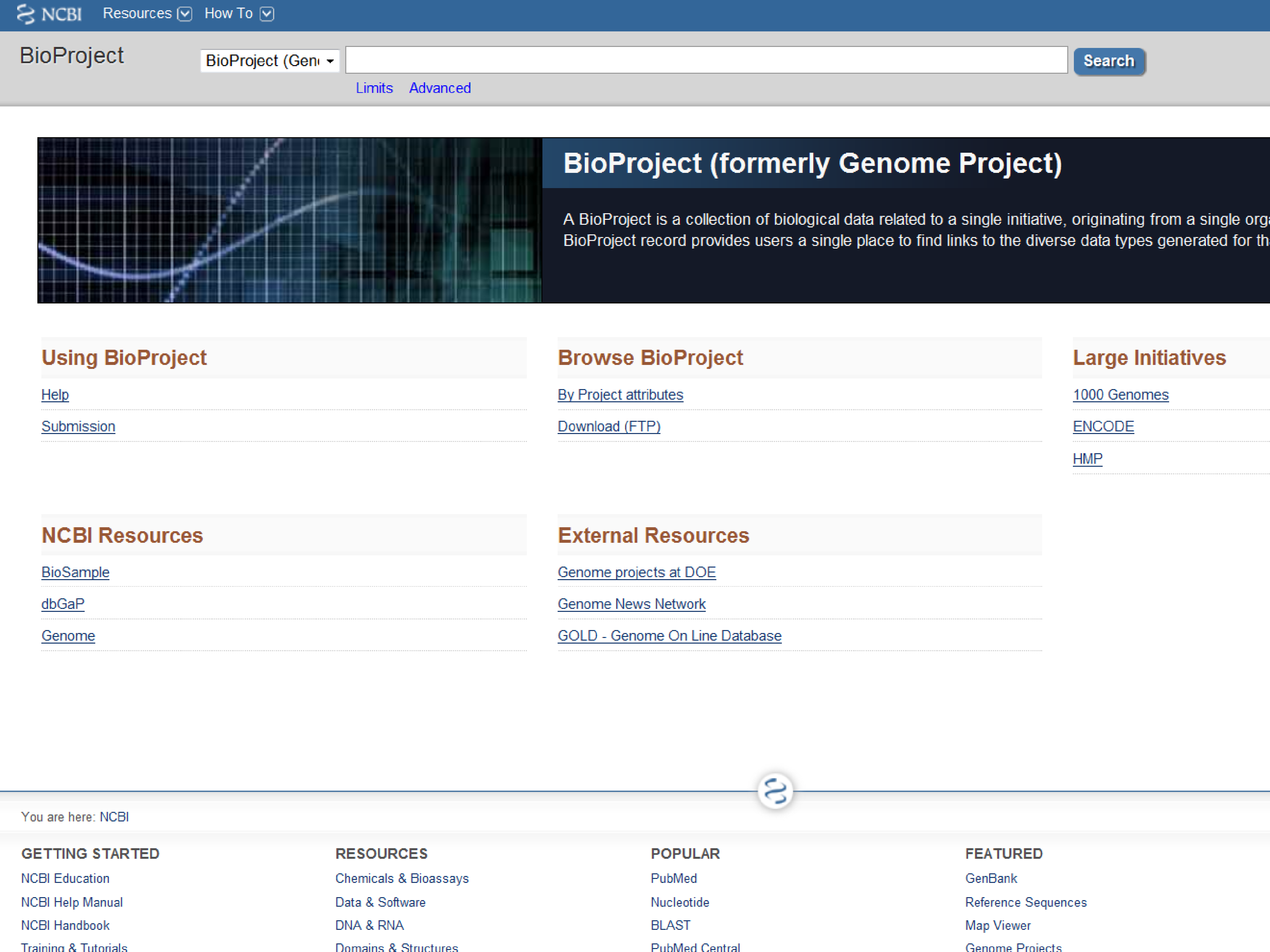
Display Settings: GenBank

Homo sapiens chromosome 4 genomic contig, GRCh37 reference assembly alternate locus group ALT_REF_LOCI_8

GenBank: GL000257.1

```
LOCUS       GL000257              590426 bp    DNA     linear   CON 29-JUN-2009
DEFINITION  Homo sapiens chromosome 4 genomic contig, GRCh37 reference assembly
            alternate locus group ALT_REF_LOCI_8.
ACCESSION   GL000257
VERSION     GL000257.1  GI:224183347
DBLINK      Project: 31257
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 590426)
  AUTHORS   Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C.,
            Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R.,
            Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J.,
            LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P.,
            Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M.,
            Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N.,
            Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J.,
            Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N.,
            Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I.,
            Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T.,
            Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L.,
            Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M.,
            Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W.,
            McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A.,
            Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C.,
```

BioProject



BioProject (formerly Genome Project)

A BioProject is a collection of biological data related to a single initiative, originating from a single org
BioProject record provides users a single place to find links to the diverse data types generated for th

Using BioProject

- Help
- Submission

Browse BioProject

- By Project attributes
- Download (FTP)

Large Initiatives

- 1000 Genomes
- ENCODE
- HMP

NCBI Resources

- BioSample
- dbGaP
- Genome

External Resources

- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

BioProject (formerly Genome Project)

Is a collection of genomics, functional genomics, and genetics projects and links to their resulting datasets.

It provides a reliable mechanism to access specific datasets that can be difficult to find.

Database content is exchanged with other members of the International Nucleotide Sequence Database Collaboration (INSDC).

Bioprojects support a variety of projects

from a focused genome sequencing project

↓

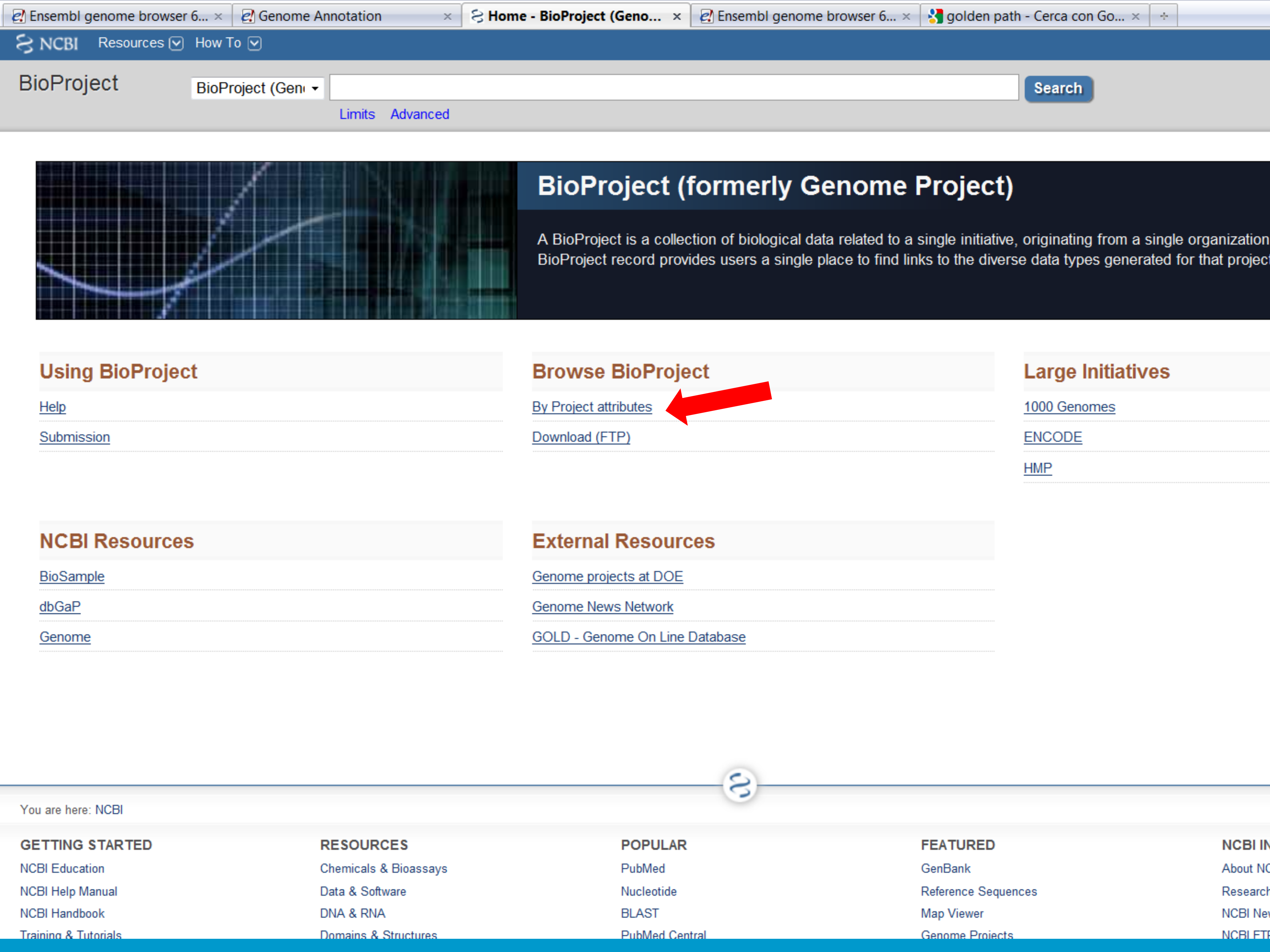
to a large international collaboration with multiple sub-projects

Project records can be established for:

- Genome sequencing and assembly
- Transcriptome sequencing and expression
- Targeted locus sequencing
- Genetic or RH Maps
- Epigenetics
- Phenotype or Genotype
- Variation detection

Access to BioProject records by

- query,
- browsing,
- following a link from another NCBI database.
    - Links may be found in several databases (Gene, Nucleotide..).


Ensembl genome browser 6...
Genome Annotation
Home - BioProject (Geno...
Ensembl genome browser 6...
golden path - Cerca con Go...
NCBI
Resources
How To
BioProject
BioProject (Gen
Search
Limits
Advanced

BioProject (formerly Genome Project)

A BioProject is a collection of biological data related to a single initiative, originating from a single organization
BioProject record provides users a single place to find links to the diverse data types generated for that projec

Using BioProject

- Help
- Submission

Browse BioProject

- By Project attributes
- Download (FTP)

Large Initiatives

- 1000 Genomes
- ENCODE
- HMP

NCBI Resources

- BioSample
- dbGaP
- Genome

External Resources

- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

You are here: NCBI

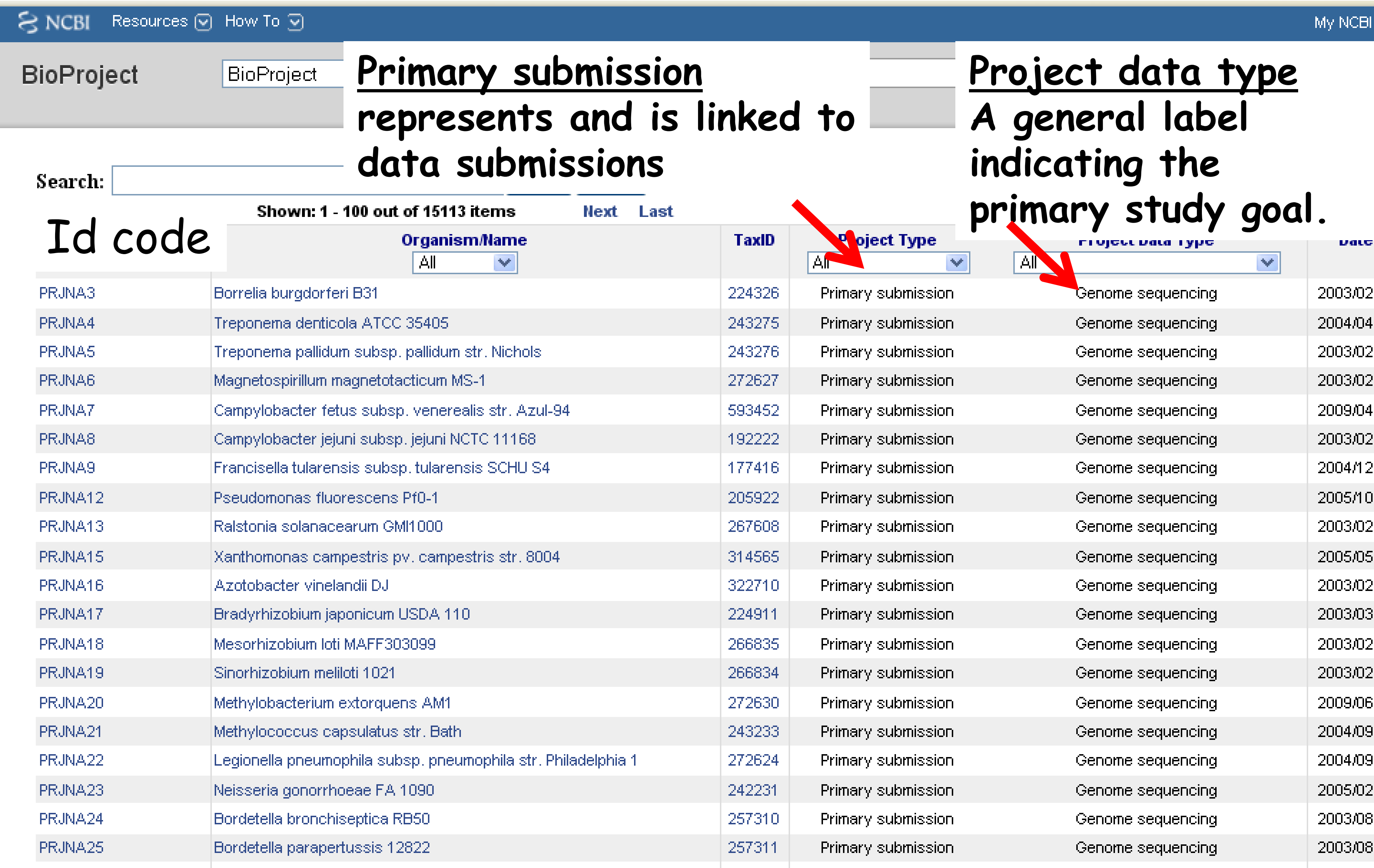
GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI IN
NCBI Education	Chemicals & Bioassays	PubMed	GenBank	About NC
NCBI Help Manual	Data & Software	Nucleotide	Reference Sequences	Research
NCBI Handbook	DNA & RNA	BLAST	Map Viewer	NCBI Ne
Training & Tutorials	Domains & Structures	PubMed Central	Genome Projects	NCBI FTP

Browsing "by project attribute"

Primary submission
represents and is linked to data submissions

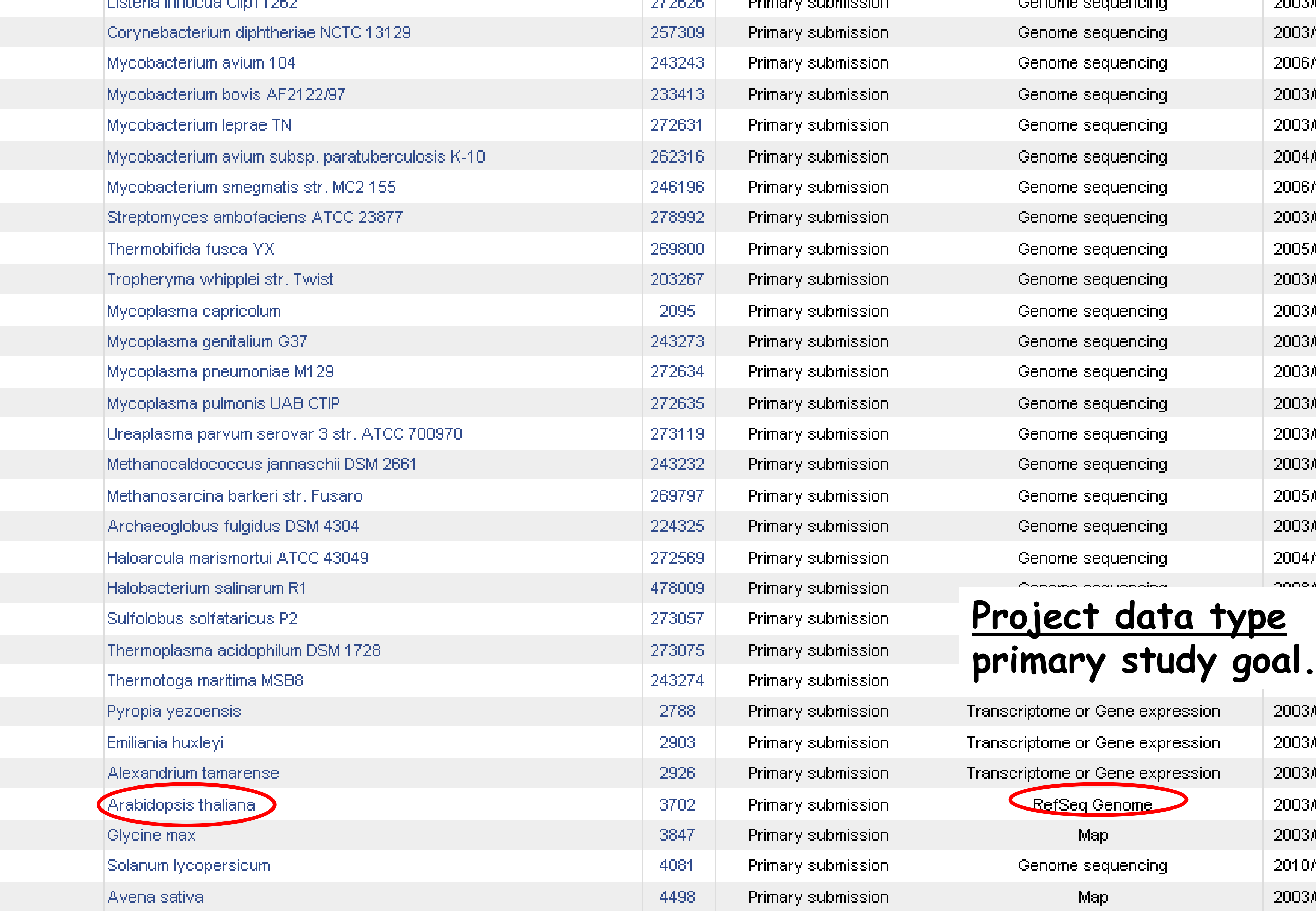
Project data type
A general label indicating the primary study goal.

Id code



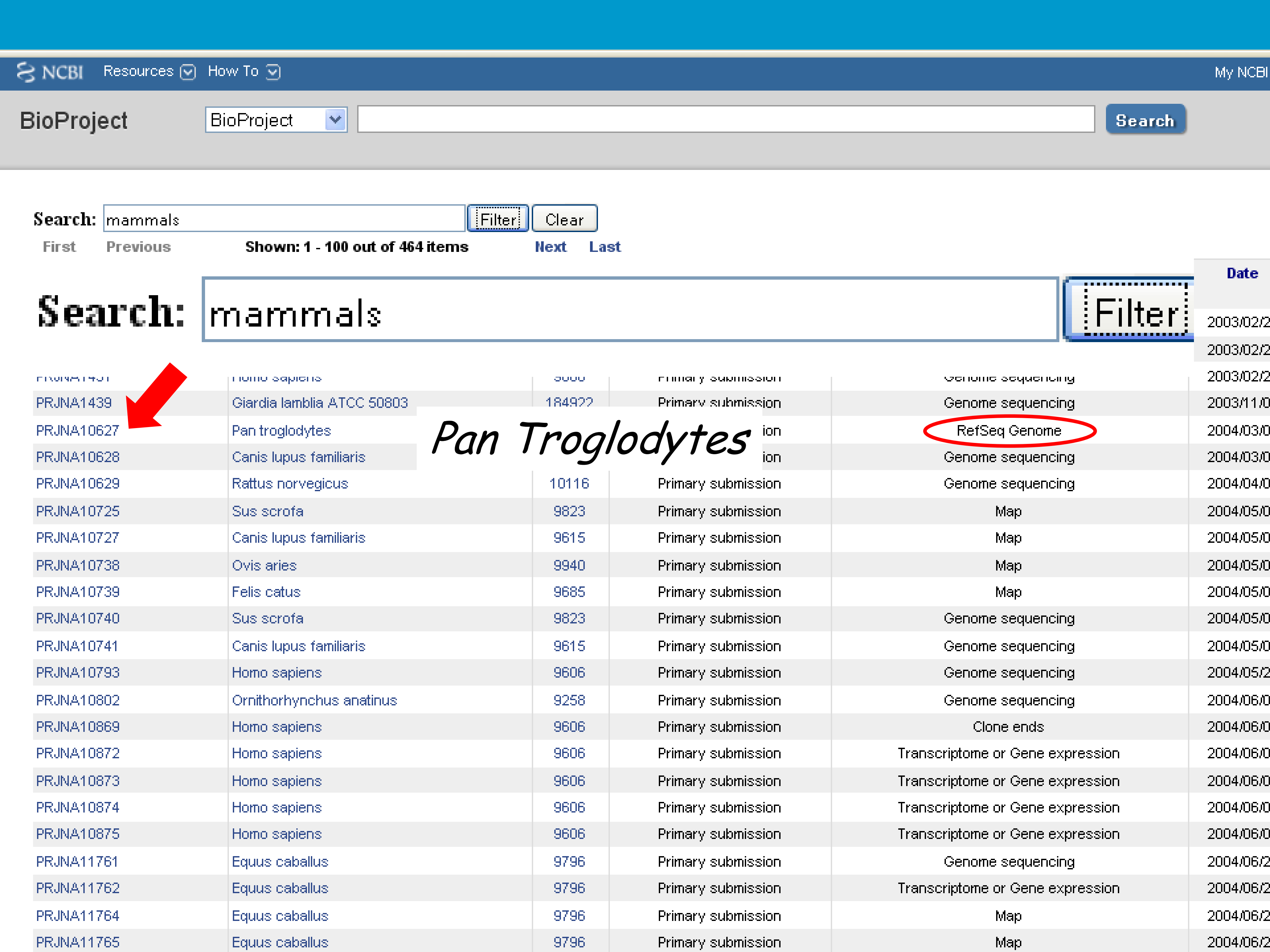
Shown: 1 - 100 out of 15113 items Next Last

	Organism/Name All	TaxID	P oject Type All	Project Data Type All	Date
PRJNA3	Borrelia burgdorferi B31	224326	Primary submission	Genome sequencing	2003/02
PRJNA4	Treponema denticola ATCC 35405	243275	Primary submission	Genome sequencing	2004/04
PRJNA5	Treponema pallidum subsp. pallidum str. Nichols	243276	Primary submission	Genome sequencing	2003/02
PRJNA6	Magnetospirillum magnetotacticum MS-1	272627	Primary submission	Genome sequencing	2003/02
PRJNA7	Campylobacter fetus subsp. venerealis str. Azul-94	593452	Primary submission	Genome sequencing	2009/04
PRJNA8	Campylobacter jejuni subsp. jejuni NCTC 11168	192222	Primary submission	Genome sequencing	2003/02
PRJNA9	Francisella tularensis subsp. tularensis SCHU S4	177416	Primary submission	Genome sequencing	2004/12
PRJNA12	Pseudomonas fluorescens Pf0-1	205922	Primary submission	Genome sequencing	2005/10
PRJNA13	Ralstonia solanacearum GMI1000	267608	Primary submission	Genome sequencing	2003/02
PRJNA15	Xanthomonas campestris pv. campestris str. 8004	314565	Primary submission	Genome sequencing	2005/05
PRJNA16	Azotobacter vinelandii DJ	322710	Primary submission	Genome sequencing	2003/02
PRJNA17	Bradyrhizobium japonicum USDA 110	224911	Primary submission	Genome sequencing	2003/03
PRJNA18	Mesorhizobium loti MAFF303099	266835	Primary submission	Genome sequencing	2003/02
PRJNA19	Sinorhizobium meliloti 1021	266834	Primary submission	Genome sequencing	2003/02
PRJNA20	Methylobacterium extorquens AM1	272630	Primary submission	Genome sequencing	2009/06
PRJNA21	Methylococcus capsulatus str. Bath	243233	Primary submission	Genome sequencing	2004/09
PRJNA22	Legionella pneumophila subsp. pneumophila str. Philadelphia 1	272624	Primary submission	Genome sequencing	2004/09
PRJNA23	Neisseria gonorrhoeae FA 1090	242231	Primary submission	Genome sequencing	2005/02
PRJNA24	Bordetella bronchiseptica RB50	257310	Primary submission	Genome sequencing	2003/08
PRJNA25	Bordetella parapertussis 12822	257311	Primary submission	Genome sequencing	2003/08



Listeria innocua Clip11262	272626	Primary submission	Genome sequencing	2003/
Corynebacterium diphtheriae NCTC 13129	257309	Primary submission	Genome sequencing	2003/
Mycobacterium avium 104	243243	Primary submission	Genome sequencing	2006/
Mycobacterium bovis AF2122/97	233413	Primary submission	Genome sequencing	2003/
Mycobacterium leprae TN	272631	Primary submission	Genome sequencing	2003/
Mycobacterium avium subsp. paratuberculosis K-10	262316	Primary submission	Genome sequencing	2004/
Mycobacterium smegmatis str. MC2 155	246196	Primary submission	Genome sequencing	2006/
Streptomyces ambofaciens ATCC 23877	278992	Primary submission	Genome sequencing	2003/
Thermobifida fusca YX	269800	Primary submission	Genome sequencing	2005/
Tropheryma whipplei str. Twist	203267	Primary submission	Genome sequencing	2003/
Mycoplasma capricolum	2095	Primary submission	Genome sequencing	2003/
Mycoplasma genitalium G37	243273	Primary submission	Genome sequencing	2003/
Mycoplasma pneumoniae M129	272634	Primary submission	Genome sequencing	2003/
Mycoplasma pulmonis UAB CTIP	272635	Primary submission	Genome sequencing	2003/
Ureaplasma parvum serovar 3 str. ATCC 700970	273119	Primary submission	Genome sequencing	2003/
Methanocaldococcus jannaschii DSM 2661	243232	Primary submission	Genome sequencing	2003/
Methanosarcina barkeri str. Fusaro	269797	Primary submission	Genome sequencing	2005/
Archaeoglobus fulgidus DSM 4304	224325	Primary submission	Genome sequencing	2003/
Haloarcula marismortui ATCC 43049	272569	Primary submission	Genome sequencing	2004/
Halobacterium salinarum R1	478009	Primary submission	Genome sequencing	2008/
Sulfolobus solfataricus P2	273057	Primary submission		
Thermoplasma acidophilum DSM 1728	273075	Primary submission		
Thermotoga maritima MSB8	243274	Primary submission		
Pyropia yezoensis	2788	Primary submission	Transcriptome or Gene expression	2003/
Emiliania huxleyi	2903	Primary submission	Transcriptome or Gene expression	2003/
Alexandrium tamarense	2926	Primary submission	Transcriptome or Gene expression	2003/
Arabidopsis thaliana	3702	Primary submission	RefSeq Genome	2003/
Glycine max	3847	Primary submission	Map	2003/
Solanum lycopersicum	4081	Primary submission	Genome sequencing	2010/
Avena sativa	4498	Primary submission	Map	2003/

Project data type
primary study goal.



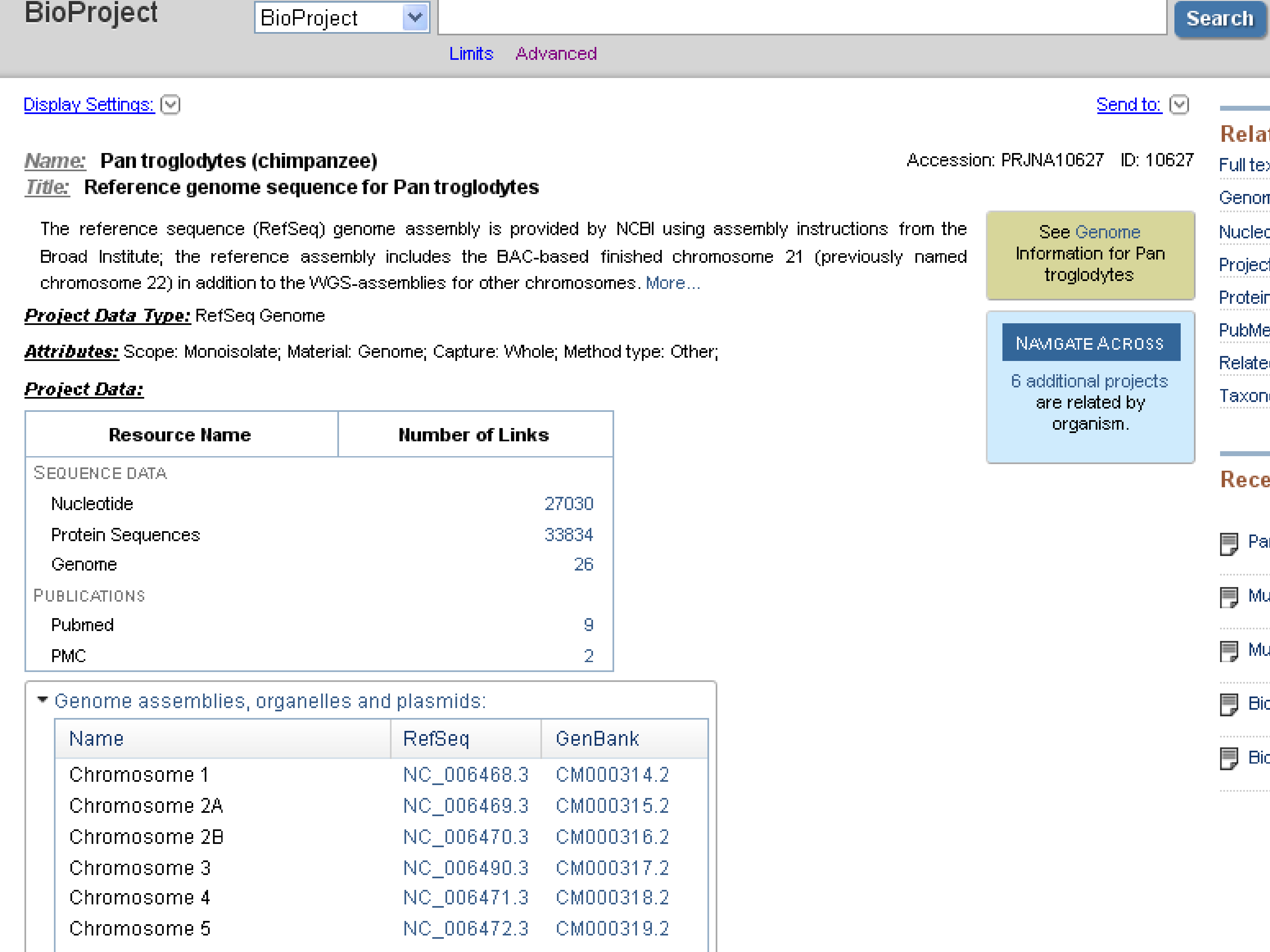
Search: mammals [Filter] [Clear]

First Previous **Shown: 1 - 100 out of 464 items** Next Last

Search: mammals [Filter]

Pan Troglodytes

					Date
					2003/02/2
					2003/02/2
PRJNA1431	Homo sapiens	9606	Primary submission	Genome sequencing	2003/02/2
PRJNA1439	Giardia lamblia ATCC 50803	184922	Primary submission	Genome sequencing	2003/11/0
PRJNA10627	Pan troglodytes		ion	RefSeq Genome	2004/03/0
PRJNA10628	Canis lupus familiaris		ion	Genome sequencing	2004/03/0
PRJNA10629	Rattus norvegicus	10116	Primary submission	Genome sequencing	2004/04/0
PRJNA10725	Sus scrofa	9823	Primary submission	Map	2004/05/0
PRJNA10727	Canis lupus familiaris	9615	Primary submission	Map	2004/05/0
PRJNA10738	Ovis aries	9940	Primary submission	Map	2004/05/0
PRJNA10739	Felis catus	9685	Primary submission	Map	2004/05/0
PRJNA10740	Sus scrofa	9823	Primary submission	Genome sequencing	2004/05/0
PRJNA10741	Canis lupus familiaris	9615	Primary submission	Genome sequencing	2004/05/0
PRJNA10793	Homo sapiens	9606	Primary submission	Genome sequencing	2004/05/2
PRJNA10802	Ornithorhynchus anatinus	9258	Primary submission	Genome sequencing	2004/06/0
PRJNA10869	Homo sapiens	9606	Primary submission	Clone ends	2004/06/0
PRJNA10872	Homo sapiens	9606	Primary submission	Transcriptome or Gene expression	2004/06/0
PRJNA10873	Homo sapiens	9606	Primary submission	Transcriptome or Gene expression	2004/06/0
PRJNA10874	Homo sapiens	9606	Primary submission	Transcriptome or Gene expression	2004/06/0
PRJNA10875	Homo sapiens	9606	Primary submission	Transcriptome or Gene expression	2004/06/0
PRJNA11761	Equus caballus	9796	Primary submission	Genome sequencing	2004/06/2
PRJNA11762	Equus caballus	9796	Primary submission	Transcriptome or Gene expression	2004/06/2
PRJNA11764	Equus caballus	9796	Primary submission	Map	2004/06/2
PRJNA11765	Equus caballus	9796	Primary submission	Map	2004/06/2



Name: **Pan troglodytes (chimpanzee)**

Accession: PRJNA10627 ID: 10627

Title: **Reference genome sequence for Pan troglodytes**

The reference sequence (RefSeq) genome assembly is provided by NCBI using assembly instructions from the Broad Institute; the reference assembly includes the BAC-based finished chromosome 21 (previously named chromosome 22) in addition to the WGS-assemblies for other chromosomes. More...

See Genome Information for Pan troglodytes

NAVIGATE ACROSS

6 additional projects are related by organism.

Project Data Type: RefSeq Genome

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Other;

Project Data:

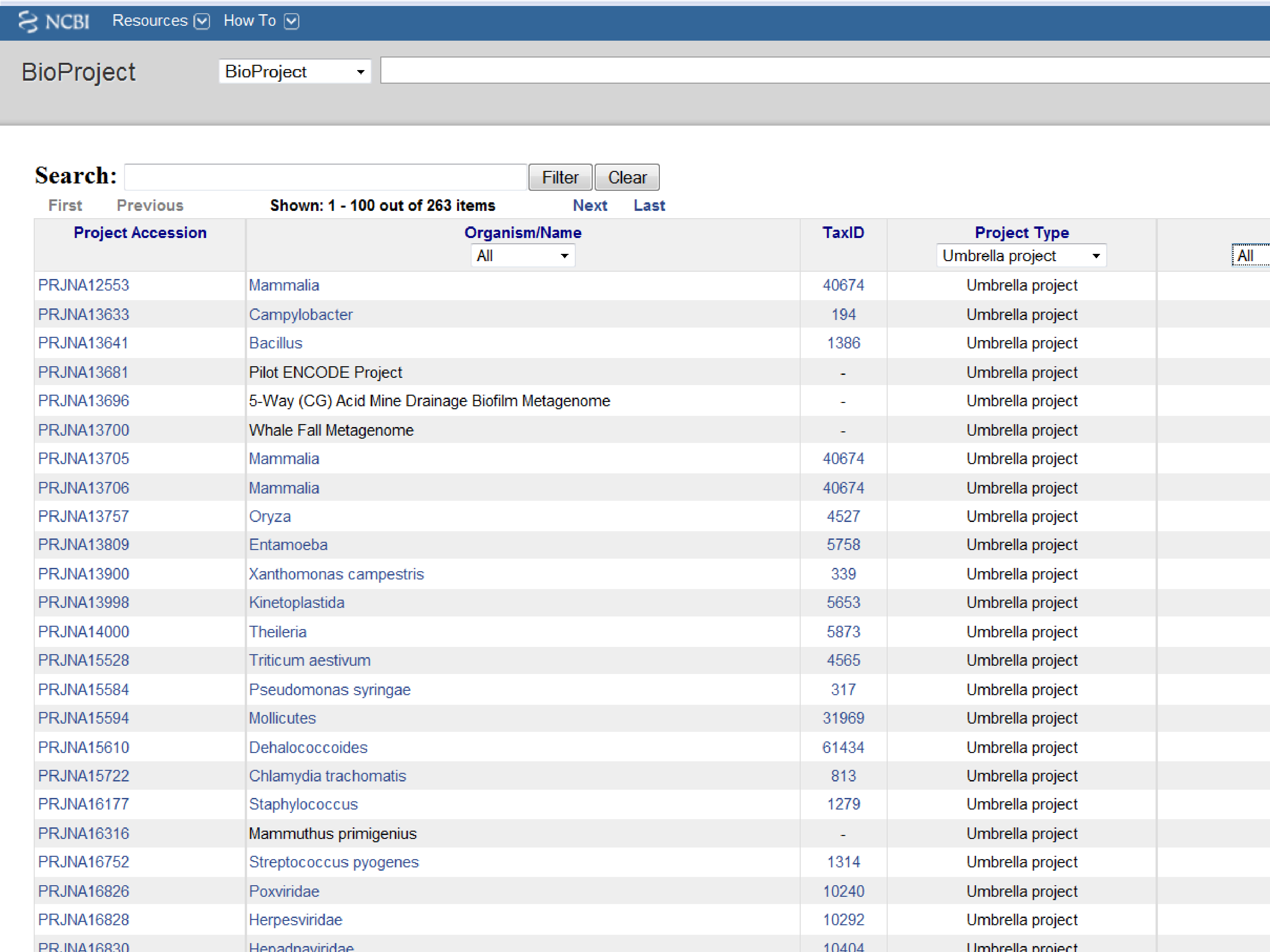
Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide	27030
Protein Sequences	33834
Genome	26
PUBLICATIONS	
Pubmed	9
PMC	2

Genome assemblies, organelles and plasmids:

Name	RefSeq	GenBank
Chromosome 1	NC_006468.3	CM000314.2
Chromosome 2A	NC_006469.3	CM000315.2
Chromosome 2B	NC_006470.3	CM000316.2
Chromosome 3	NC_006490.3	CM000317.2
Chromosome 4	NC_006471.3	CM000318.2
Chromosome 5	NC_006472.3	CM000319.2

Umbrella project

- Administrative project that is created to group multiple projects that are related by a single effort from a single submitter or group of submitters, but represent distinct studies that differ in methodology, sample material, or resulting data type.

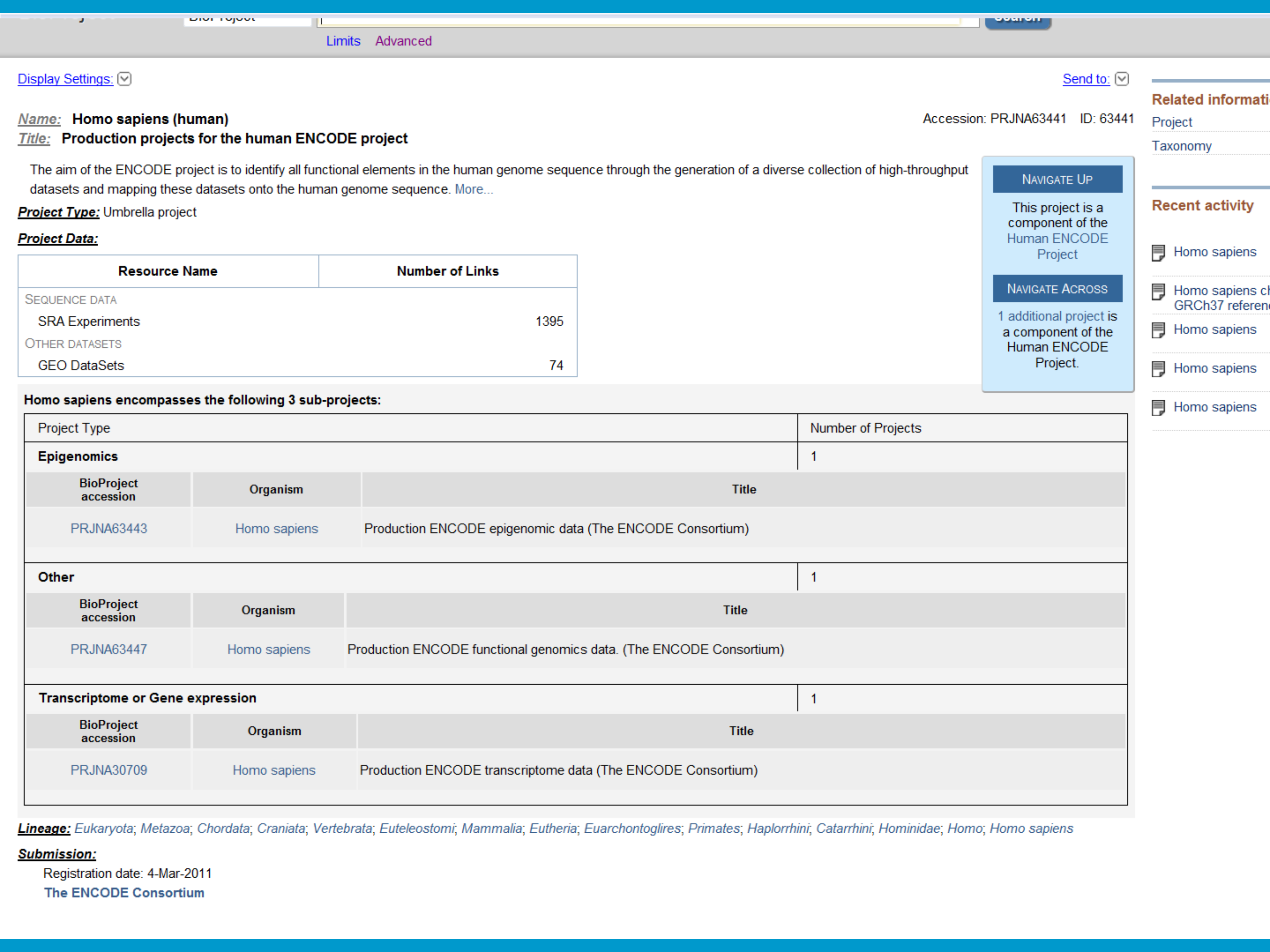


BioProject

Search:

First Previous **Shown: 1 - 100 out of 263 items** Next Last

Project Accession	Organism/Name All	TaxID	Project Type Umbrella project	All
PRJNA12553	Mammalia	40674	Umbrella project	
PRJNA13633	Campylobacter	194	Umbrella project	
PRJNA13641	Bacillus	1386	Umbrella project	
PRJNA13681	Pilot ENCODE Project	-	Umbrella project	
PRJNA13696	5-Way (CG) Acid Mine Drainage Biofilm Metagenome	-	Umbrella project	
PRJNA13700	Whale Fall Metagenome	-	Umbrella project	
PRJNA13705	Mammalia	40674	Umbrella project	
PRJNA13706	Mammalia	40674	Umbrella project	
PRJNA13757	Oryza	4527	Umbrella project	
PRJNA13809	Entamoeba	5758	Umbrella project	
PRJNA13900	Xanthomonas campestris	339	Umbrella project	
PRJNA13998	Kinetoplastida	5653	Umbrella project	
PRJNA14000	Theileria	5873	Umbrella project	
PRJNA15528	Triticum aestivum	4565	Umbrella project	
PRJNA15584	Pseudomonas syringae	317	Umbrella project	
PRJNA15594	Mollicutes	31969	Umbrella project	
PRJNA15610	Dehalococcoides	61434	Umbrella project	
PRJNA15722	Chlamydia trachomatis	813	Umbrella project	
PRJNA16177	Staphylococcus	1279	Umbrella project	
PRJNA16316	Mammuthus primigenius	-	Umbrella project	
PRJNA16752	Streptococcus pyogenes	1314	Umbrella project	
PRJNA16826	Poxviridae	10240	Umbrella project	
PRJNA16828	Herpesviridae	10292	Umbrella project	
PRJNA16830	Hepadnaviridae	10404	Umbrella project	



Name: Homo sapiens (human)   Accession: PRJNA63441   ID: 63441

Title: Production projects for the human ENCODE project

The aim of the ENCODE project is to identify all functional elements in the human genome sequence through the generation of a diverse collection of high-throughput datasets and mapping these datasets onto the human genome sequence. More...

Project Type: Umbrella project

Project Data:

Resource Name	Number of Links
Sequence data	
SRA Experiments	1395
Other datasets	
GEO DataSets	74

Navigate Up

This project is a component of the Human ENCODE Project

Navigate Across

1 additional project is a component of the Human ENCODE Project.

Homo sapiens encompasses the following 3 sub-projects:

Project Type	Number of Projects
Epigenomics	1

BioProject accession	Organism	Title
PRJNA63443	Homo sapiens	Production ENCODE epigenomic data (The ENCODE Consortium)

Project Type	Number of Projects
Other	1

BioProject accession	Organism	Title
PRJNA63447	Homo sapiens	Production ENCODE functional genomics data. (The ENCODE Consortium)

Project Type	Number of Projects
Transcriptome or Gene expression	1

BioProject accession	Organism	Title
PRJNA30709	Homo sapiens	Production ENCODE transcriptome data (The ENCODE Consortium)

Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens*

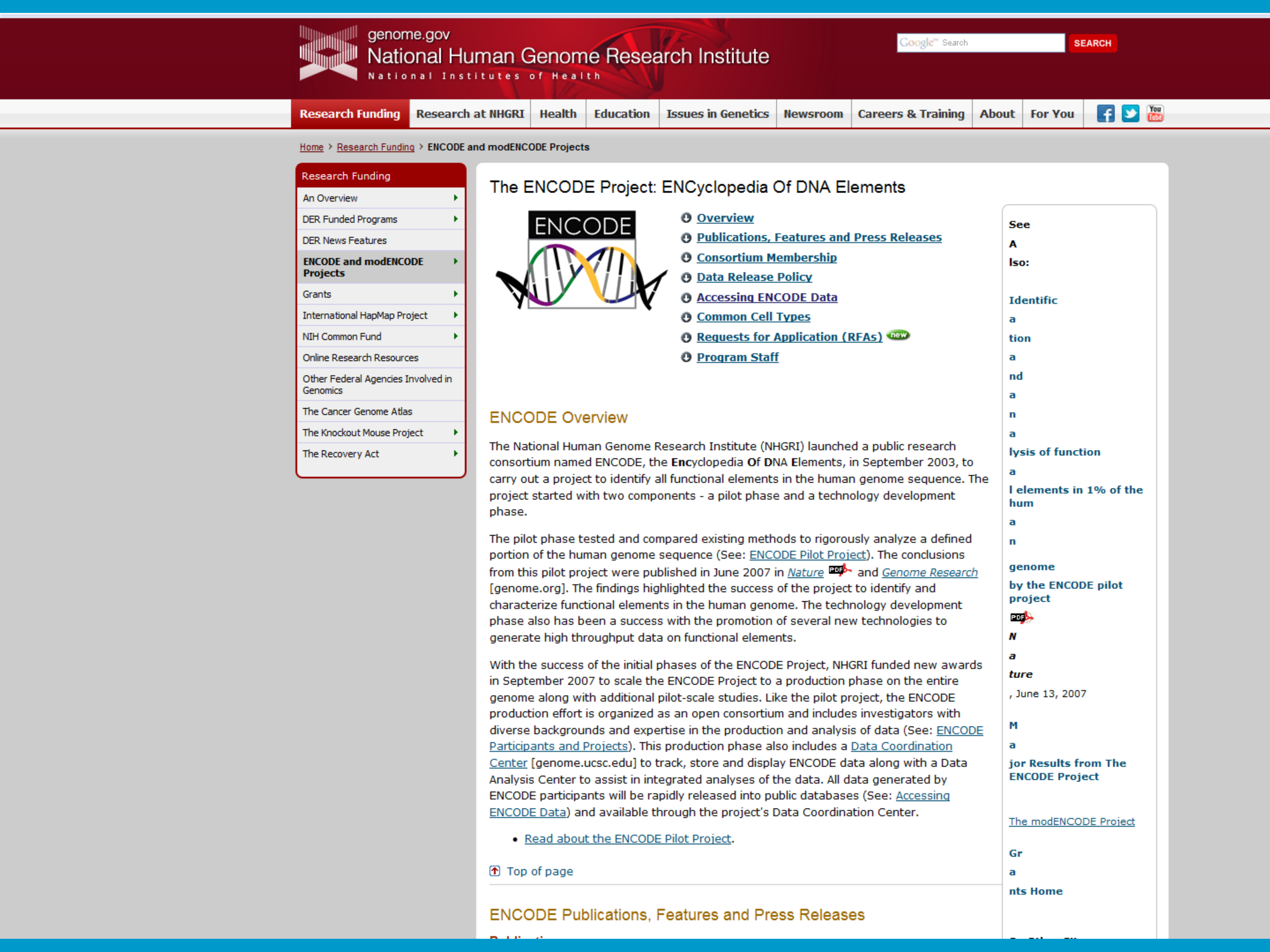
Submission:

Registration date: 4-Mar-2011

The ENCODE Consortium

ENCODE PROJECT

ENCyclopedia
Of DNA Elements



Research Funding
- An Overview
- DER Funded Programs
- DER News Features
- **ENCODE and modENCODE Projects**
- Grants
- International HapMap Project
- NIH Common Fund
- Online Research Resources
- Other Federal Agencies Involved in Genomics
- The Cancer Genome Atlas
- The Knockout Mouse Project
- The Recovery Act

The ENCODE Project: ENCyclopedia Of DNA Elements

ENCODE Overview

The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the **Enc**yclopedia **O**f **D**NA **E**lements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. The project started with two components - a pilot phase and a technology development phase.

The pilot phase tested and compared existing methods to rigorously analyze a defined portion of the human genome sequence (See: ENCODE Pilot Project). The conclusions from this pilot project were published in June 2007 in *Nature* PDF and *Genome Research* [genome.org]. The findings highlighted the success of the project to identify and characterize functional elements in the human genome. The technology development phase also has been a success with the promotion of several new technologies to generate high throughput data on functional elements.

With the success of the initial phases of the ENCODE Project, NHGRI funded new awards in September 2007 to scale the ENCODE Project to a production phase on the entire genome along with additional pilot-scale studies. Like the pilot project, the ENCODE production effort is organized as an open consortium and includes investigators with diverse backgrounds and expertise in the production and analysis of data (See: ENCODE Participants and Projects). This production phase also includes a Data Coordination Center [genome.ucsc.edu] to track, store and display ENCODE data along with a Data Analysis Center to assist in integrated analyses of the data. All data generated by ENCODE participants will be rapidly released into public databases (See: Accessing ENCODE Data) and available through the project's Data Coordination Center.

- Read about the ENCODE Pilot Project.

Top of page

ENCODE Publications, Features and Press Releases

See Also:

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project PDF ***Nature***, June 13, 2007

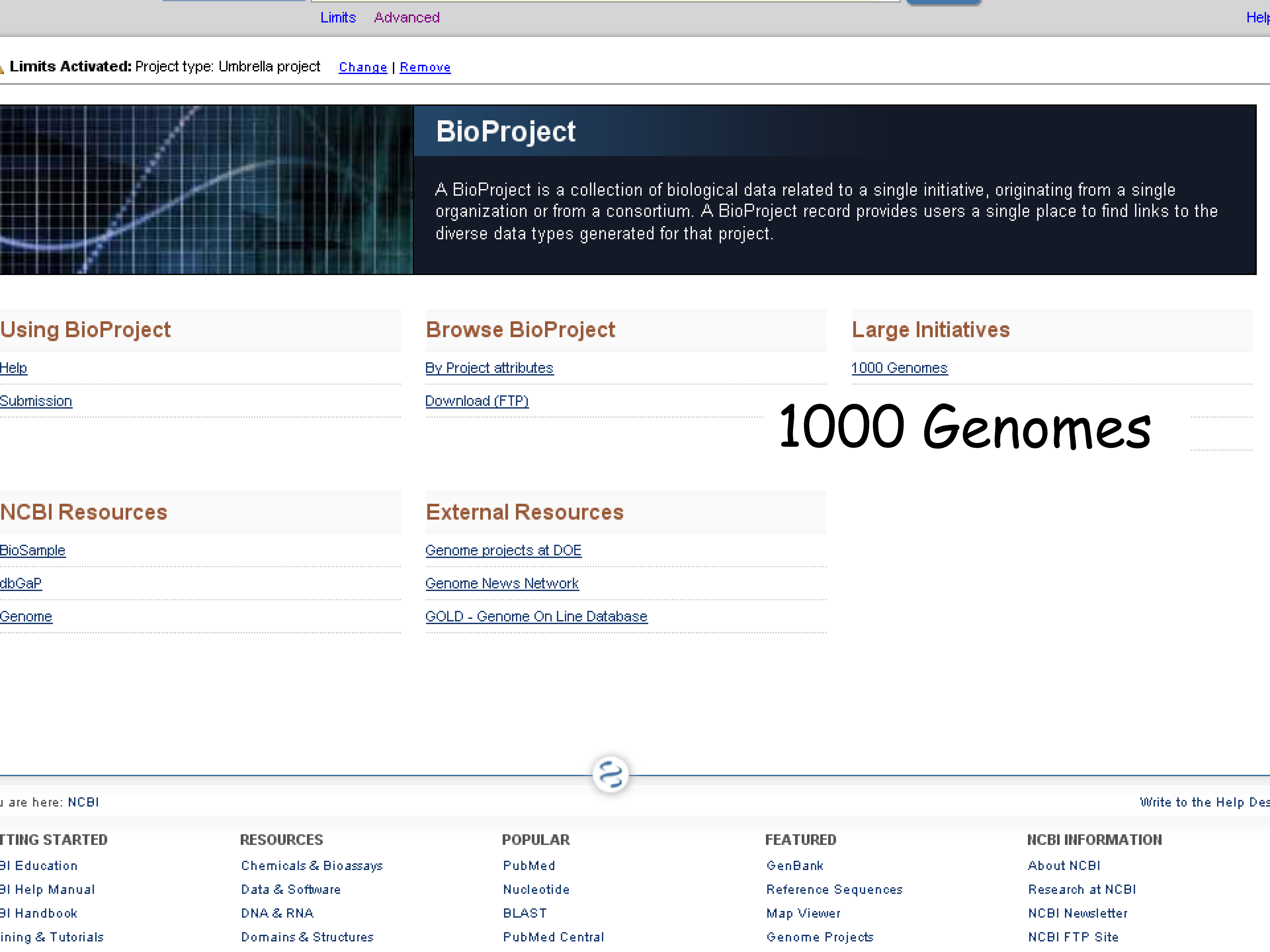
Major Results from The ENCODE Project

The modENCODE Project

Grants Home

- The aim of the ENCODE project is to identify all functional elements in the human genome, including coding and regulatory regions.

- The basic approach has been comparative genomics



Limits Advanced Help

Limits Activated: Project type: Umbrella project Change | Remove

BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Using BioProject

- Help
- Submission

Browse BioProject

- By Project attributes
- Download (FTP)

Large Initiatives

- 1000 Genomes

1000 Genomes

NCBI Resources

- BioSample
- dbGaP
- Genome

External Resources

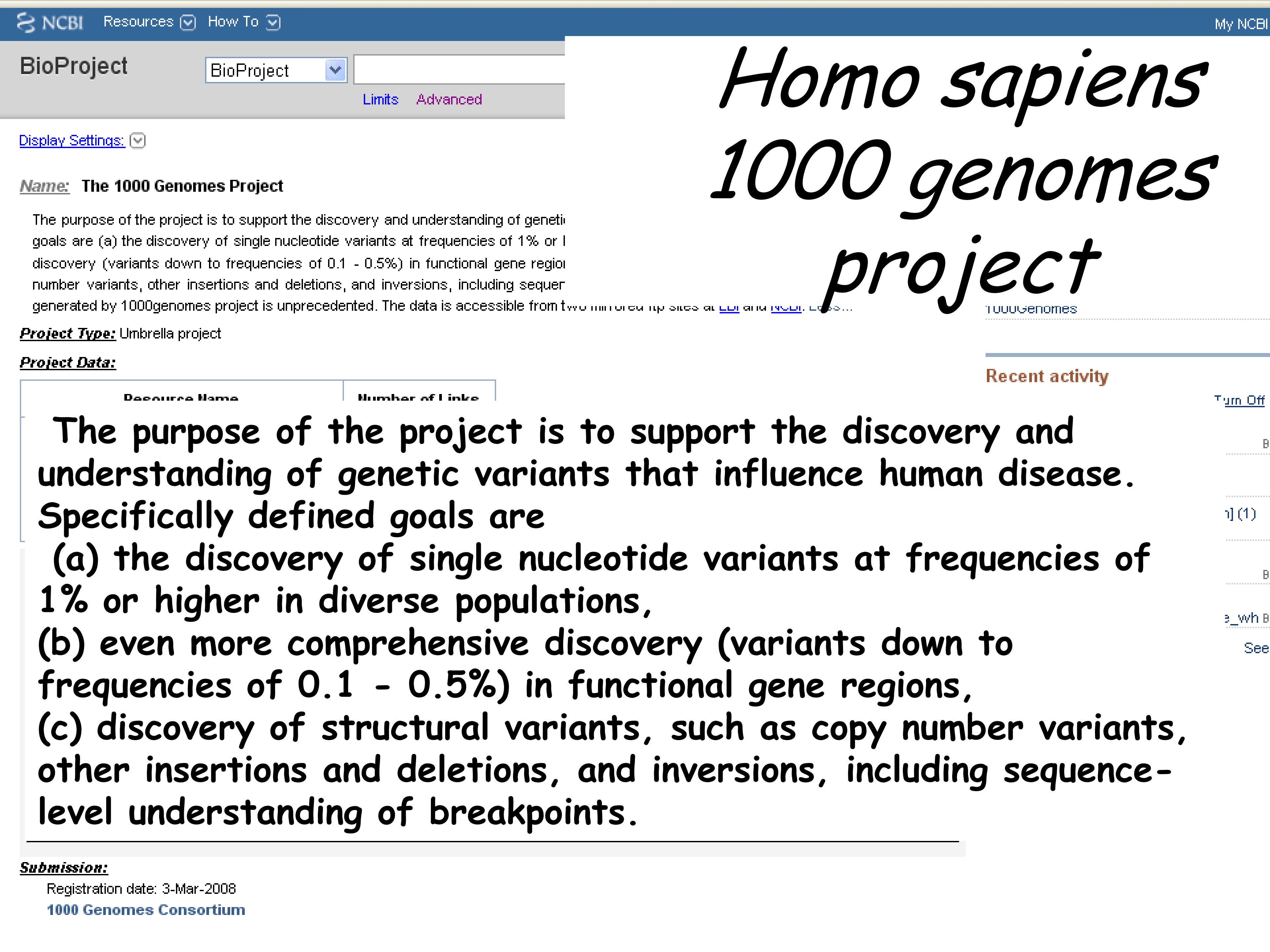
- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

u are here: NCBI

Write to the Help Des

TTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
3I Education	Chemicals & Bioassays	PubMed	GenBank	About NCBI
3I Help Manual	Data & Software	Nucleotide	Reference Sequences	Research at NCBI
3I Handbook	DNA & RNA	BLAST	Map Viewer	NCBI Newsletter
ining & Tutorials	Domains & Structures	PubMed Central	Genome Projects	NCBI FTP Site

Homo sapiens 1000 genomes project



BioProject

Name: The 1000 Genomes Project

The purpose of the project is to support the discovery and understanding of geneti
goals are (a) the discovery of single nucleotide variants at frequencies of 1% or l
discovery (variants down to frequencies of 0.1 - 0.5%) in functional gene regio
number variants, other insertions and deletions, and inversions, including sequer
generated by 1000genomes project is unprecedented. The data is accessible from two mirrored ftp sites at EBI and NCBI. Less...

1000Genomes

Project Type: Umbrella project

Project Data:

Resource Name	Number of Links

Submission:

Registration date: 3-Mar-2008

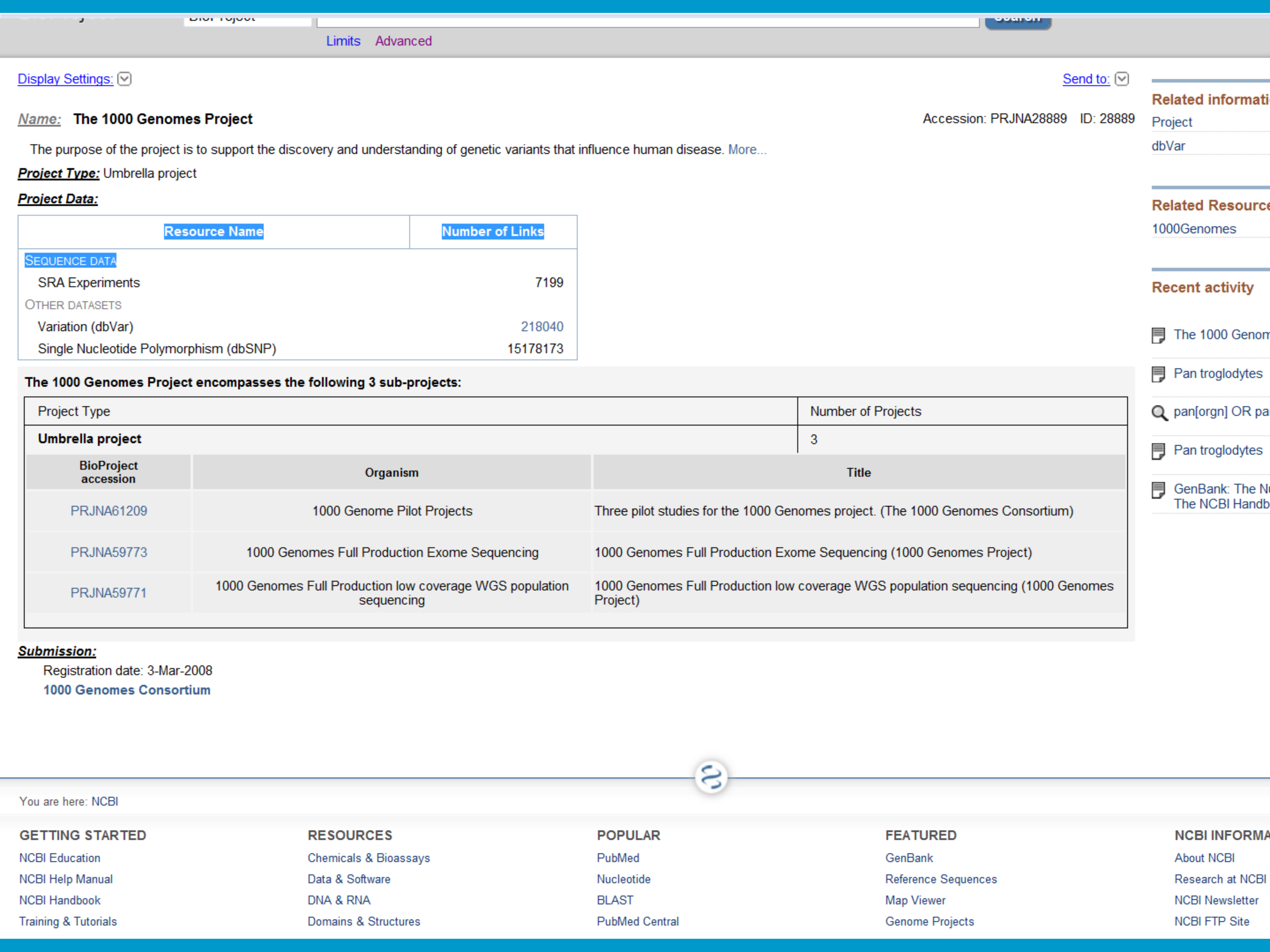
1000 Genomes Consortium

The purpose of the project is to support the discovery and understanding of genetic variants that influence human disease. Specifically defined goals are

(a) the discovery of single nucleotide variants at frequencies of 1% or higher in diverse populations,

(b) even more comprehensive discovery (variants down to frequencies of 0.1 - 0.5%) in functional gene regions,

(c) discovery of structural variants, such as copy number variants, other insertions and deletions, and inversions, including sequence-level understanding of breakpoints.



Name: **The 1000 Genomes Project**

Accession: PRJNA28889 ID: 28889

The purpose of the project is to support the discovery and understanding of genetic variants that influence human disease. More...

Project Type: Umbrella project

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	7199
OTHER DATASETS	
Variation (dbVar)	218040
Single Nucleotide Polymorphism (dbSNP)	15178173

The 1000 Genomes Project encompasses the following 3 sub-projects:

Project Type	Number of Projects
Umbrella project	3

BioProject accession	Organism	Title
PRJNA61209	1000 Genome Pilot Projects	Three pilot studies for the 1000 Genomes project. (The 1000 Genomes Consortium)
PRJNA59773	1000 Genomes Full Production Exome Sequencing	1000 Genomes Full Production Exome Sequencing (1000 Genomes Project)
PRJNA59771	1000 Genomes Full Production low coverage WGS population sequencing	1000 Genomes Full Production low coverage WGS population sequencing (1000 Genomes Project)

Submission:

Registration date: 3-Mar-2008

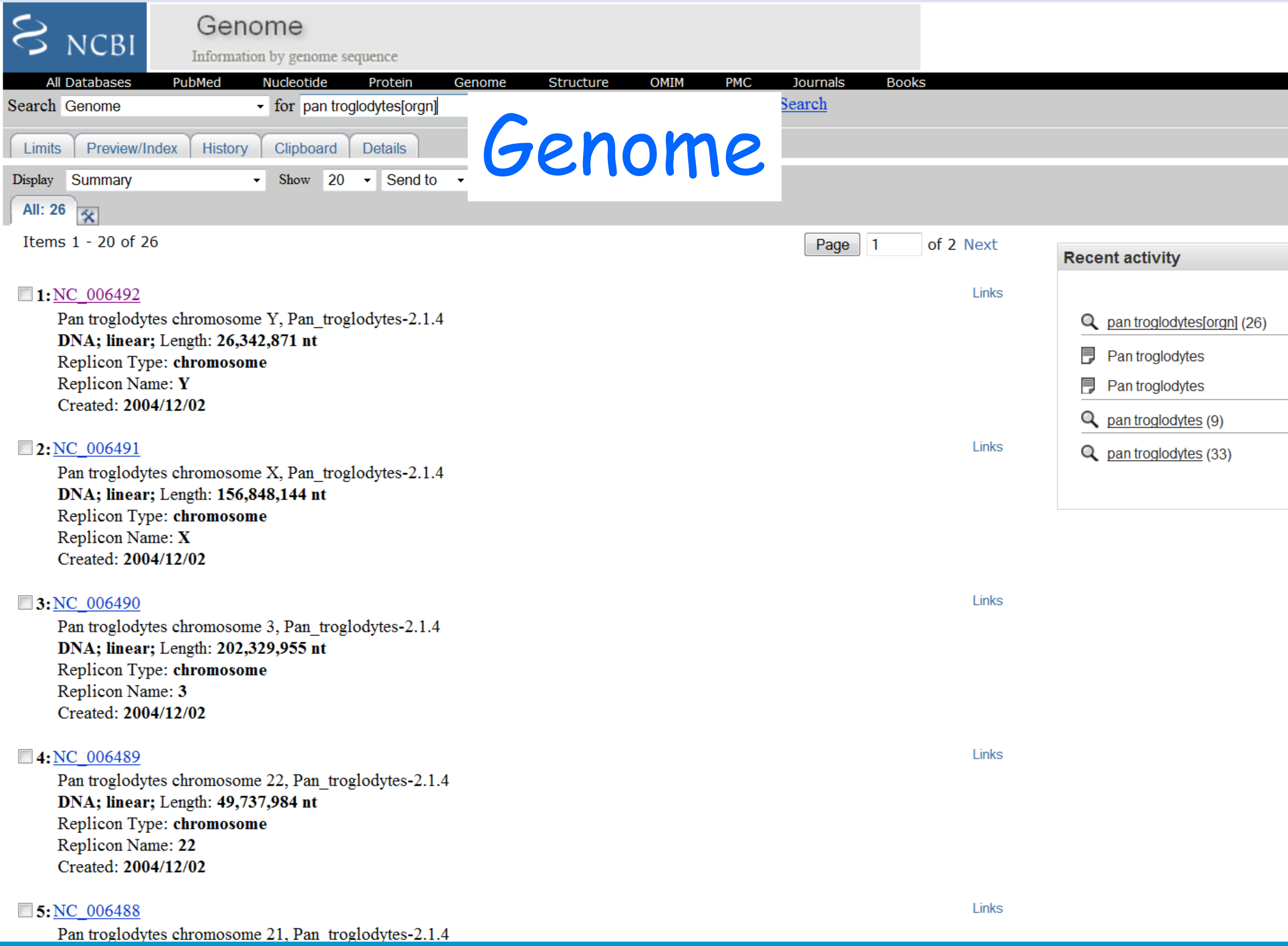
1000 Genomes Consortium

Searching for information on the chimpanzee (*Pan Troglodytes*) genome

Pan troglodytes, or chimpanzee, is a primate very closely related to humans. The chimpanzee is an important model to study biology, disease, and evolution.

Link rete

Genome



Search Genome for pan troglodytes[orgn]

Display Summary Show 20

All: 26

Items 1 - 20 of 26

Page 1 of 2

Recent activity

- pan troglodytes[orgn] (26)
- Pan troglodytes
- Pan troglodytes
- pan troglodytes (9)
- pan troglodytes (33)

□ **1:** NC_006492
Pan troglodytes chromosome Y, Pan_troglodytes-2.1.4
DNA; linear; Length: **26,342,871 nt**
Replicon Type: **chromosome**
Replicon Name: **Y**
Created: **2004/12/02**

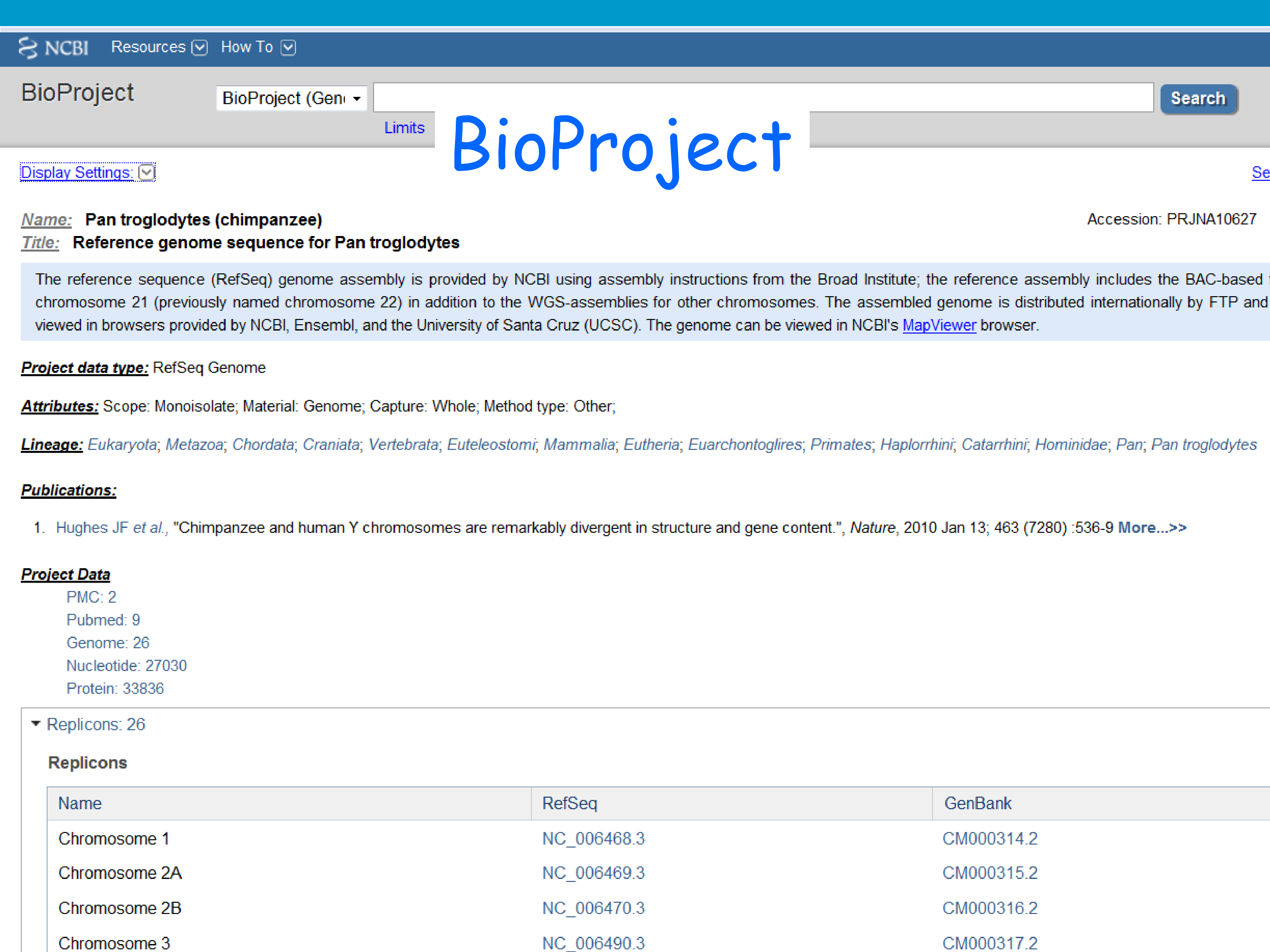
□ **2:** NC_006491
Pan troglodytes chromosome X, Pan_troglodytes-2.1.4
DNA; linear; Length: **156,848,144 nt**
Replicon Type: **chromosome**
Replicon Name: **X**
Created: **2004/12/02**

□ **3:** NC_006490
Pan troglodytes chromosome 3, Pan_troglodytes-2.1.4
DNA; linear; Length: **202,329,955 nt**
Replicon Type: **chromosome**
Replicon Name: **3**
Created: **2004/12/02**

□ **4:** NC_006489
Pan troglodytes chromosome 22, Pan_troglodytes-2.1.4
DNA; linear; Length: **49,737,984 nt**
Replicon Type: **chromosome**
Replicon Name: **22**
Created: **2004/12/02**

□ **5:** NC_006488
Pan troglodytes chromosome 21, Pan_troglodytes-2.1.4

BioProject



Name: **Pan troglodytes (chimpanzee)**

Accession: PRJNA10627

Title: **Reference genome sequence for Pan troglodytes**

The reference sequence (RefSeq) genome assembly is provided by NCBI using assembly instructions from the Broad Institute; the reference assembly includes the BAC-based chromosome 21 (previously named chromosome 22) in addition to the WGS-assemblies for other chromosomes. The assembled genome is distributed internationally by FTP and viewed in browsers provided by NCBI, Ensembl, and the University of Santa Cruz (UCSC). The genome can be viewed in NCBI's MapViewer browser.

Project data type: RefSeq Genome

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Other;

Lineage: *Eukaryota*; *Metazoa*; *Chordata*; *Craniata*; *Vertebrata*; *Euteleostomi*; *Mammalia*; *Eutheria*; *Euarchontoglires*; *Primates*; *Haplorrhini*; *Catarrhini*; *Hominidae*; *Pan*; *Pan troglodytes*

Publications:

1. Hughes JF *et al.*, "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.", *Nature*, 2010 Jan 13; 463 (7280) :536-9 **More...>>**

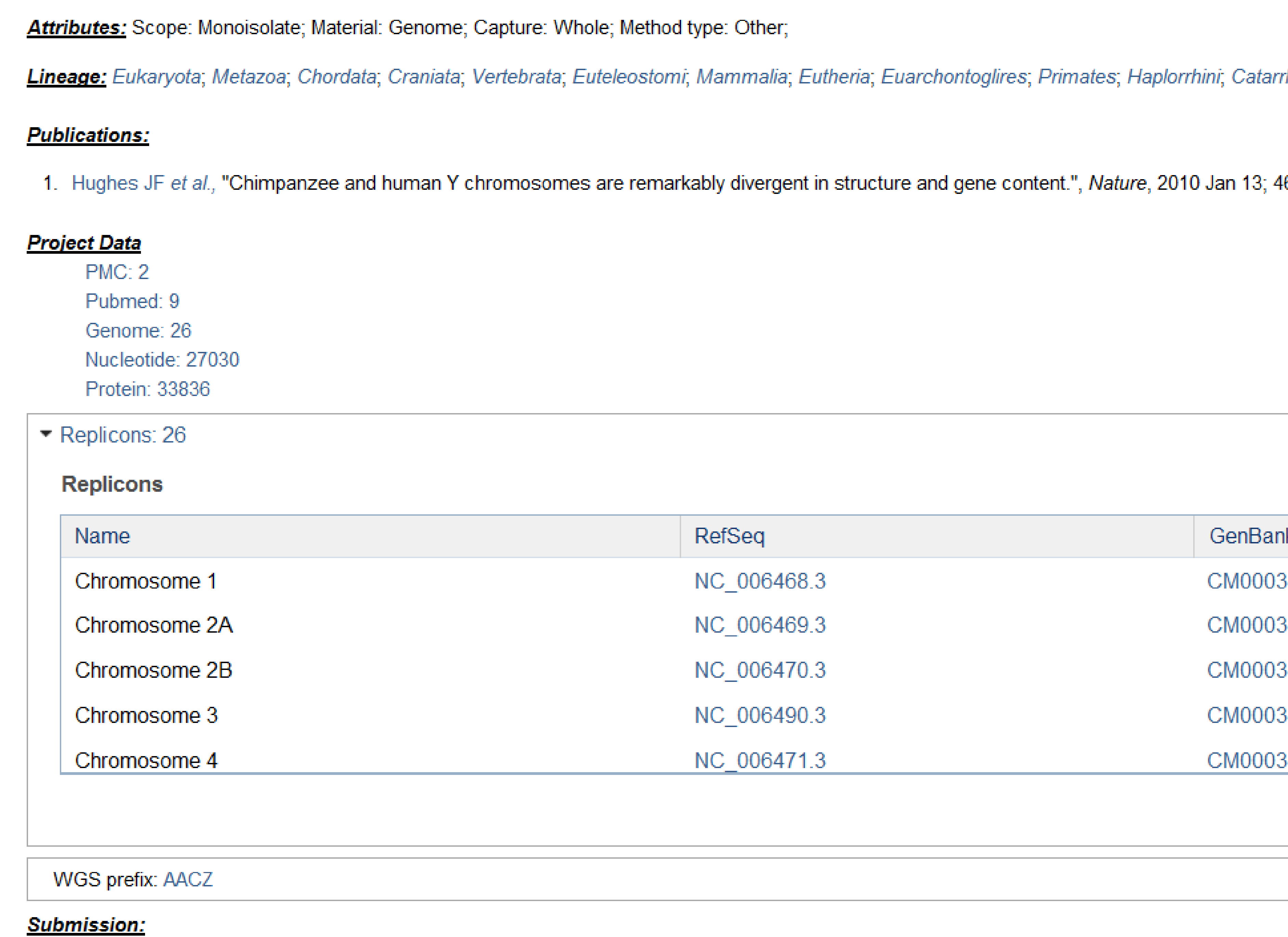
Project Data

PMC: 2
Pubmed: 9
Genome: 26
Nucleotide: 27030
Protein: 33836

▾ Replicons: 26

Replicons

Name	RefSeq	GenBank
Chromosome 1	NC_006468.3	CM000314.2
Chromosome 2A	NC_006469.3	CM000315.2
Chromosome 2B	NC_006470.3	CM000316.2
Chromosome 3	NC_006490.3	CM000317.2



Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Other;

Lineage: *Eukaryota*; *Metazoa*; *Chordata*; *Craniata*; *Vertebrata*; *Euteleostomi*; *Mammalia*; *Eutheria*; *Euarchontoglires*; *Primates*; *Haplorrhini*; *Catarr*

Publications:

1. Hughes JF *et al.*, "Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.", *Nature*, 2010 Jan 13; 46

Project Data

PMC: 2
Pubmed: 9
Genome: 26
Nucleotide: 27030
Protein: 33836

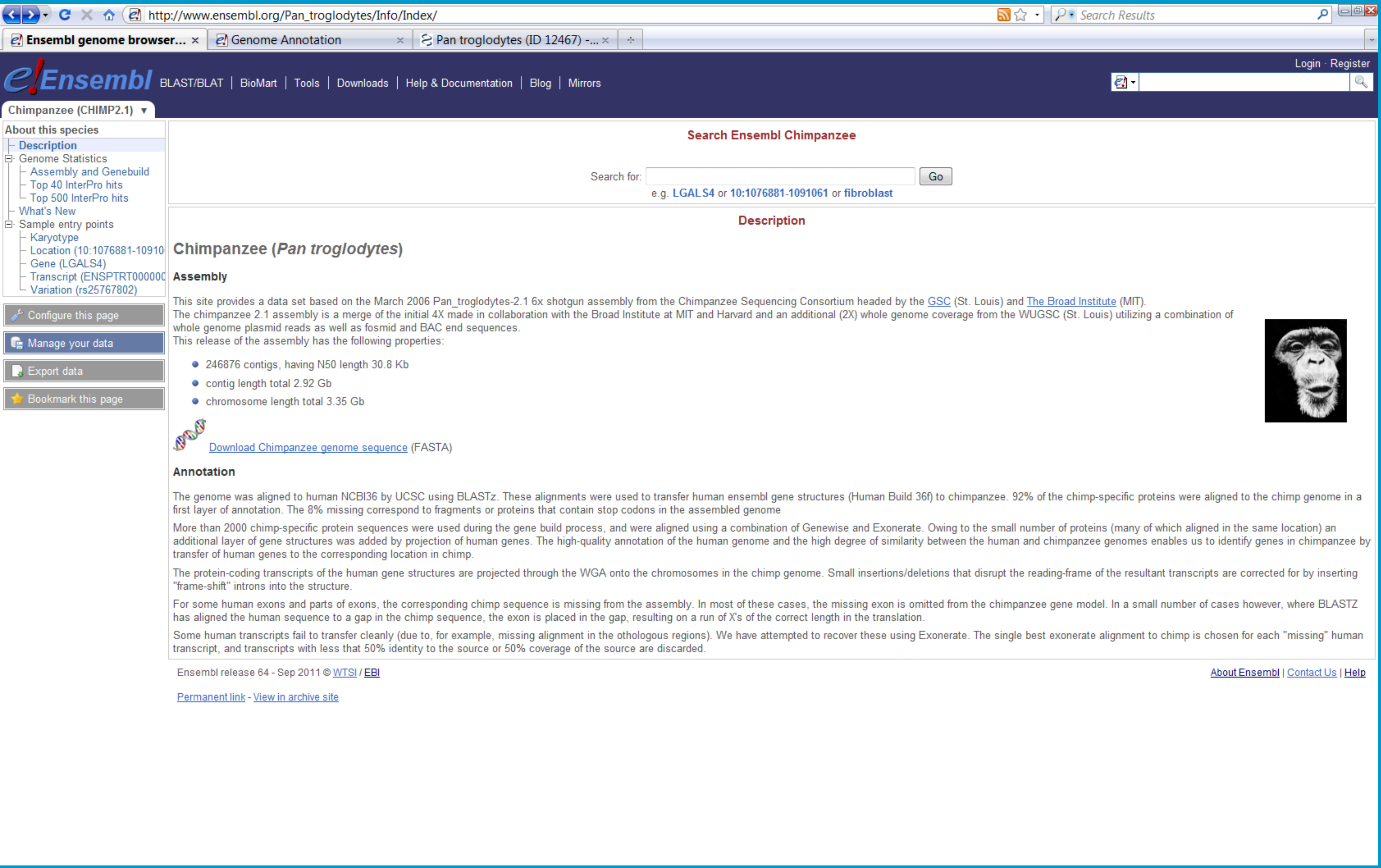
▾ Replicons: 26

Replicons

Name	RefSeq	GenBan
Chromosome 1	NC_006468.3	CM0003
Chromosome 2A	NC_006469.3	CM0003
Chromosome 2B	NC_006470.3	CM0003
Chromosome 3	NC_006490.3	CM0003
Chromosome 4	NC_006471.3	CM0003

WGS prefix: AACZ

Submission:



Search Ensembl Chimpanzee

Search for: Go

e.g. LGALS4 or 10:1076881-1091061 or fibroblast

Description

Chimpanzee (*Pan troglodytes*)

Assembly

This site provides a data set based on the March 2006 Pan_troglodytes-2.1 6x shotgun assembly from the Chimpanzee Sequencing Consortium headed by the GSC (St. Louis) and The Broad Institute (MIT).
The chimpanzee 2.1 assembly is a merge of the initial 4X made in collaboration with the Broad Institute at MIT and Harvard and an additional (2X) whole genome coverage from the WUGSC (St. Louis) utilizing a combination of whole genome plasmid reads as well as fosmid and BAC end sequences.
This release of the assembly has the following properties:

- 246876 contigs, having N50 length 30.8 Kb
- contig length total 2.92 Gb
- chromosome length total 3.35 Gb

Download Chimpanzee genome sequence (FASTA)

Annotation

The genome was aligned to human NCBI36 by UCSC using BLASTz. These alignments were used to transfer human ensembl gene structures (Human Build 36f) to chimpanzee. 92% of the chimp-specific proteins were aligned to the chimp genome in a first layer of annotation. The 8% missing correspond to fragments or proteins that contain stop codons in the assembled genome

More than 2000 chimp-specific protein sequences were used during the gene build process, and were aligned using a combination of Genewise and Exonerate. Owing to the small number of proteins (many of which aligned in the same location) an additional layer of gene structures was added by projection of human genes. The high-quality annotation of the human genome and the high degree of similarity between the human and chimpanzee genomes enables us to identify genes in chimpanzee by transfer of human genes to the corresponding location in chimp.

The protein-coding transcripts of the human gene structures are projected through the WGA onto the chromosomes in the chimp genome. Small insertions/deletions that disrupt the reading-frame of the resultant transcripts are corrected for by inserting "frame-shift" introns into the structure.

For some human exons and parts of exons, the corresponding chimp sequence is missing from the assembly. In most of these cases, the missing exon is omitted from the chimpanzee gene model. In a small number of cases however, where BLASTZ has aligned the human sequence to a gap in the chimp sequence, the exon is placed in the gap, resulting on a run of X's of the correct length in the translation.

Some human transcripts fail to transfer cleanly (due to, for example, missing alignment in the othologous regions). We have attempted to recover these using Exonerate. The single best exonerate alignment to chimp is chosen for each "missing" human transcript, and transcripts with less that 50% identity to the source or 50% coverage of the source are discarded.

Ensembl release 64 - Sep 2011 © WTSI / EBI

About Ensembl | Contact Us | Help

Permanent link - View in archive site