Next Generation Sequencing Hands-on Session

BMI7830

Gulcin Ozer, PhD

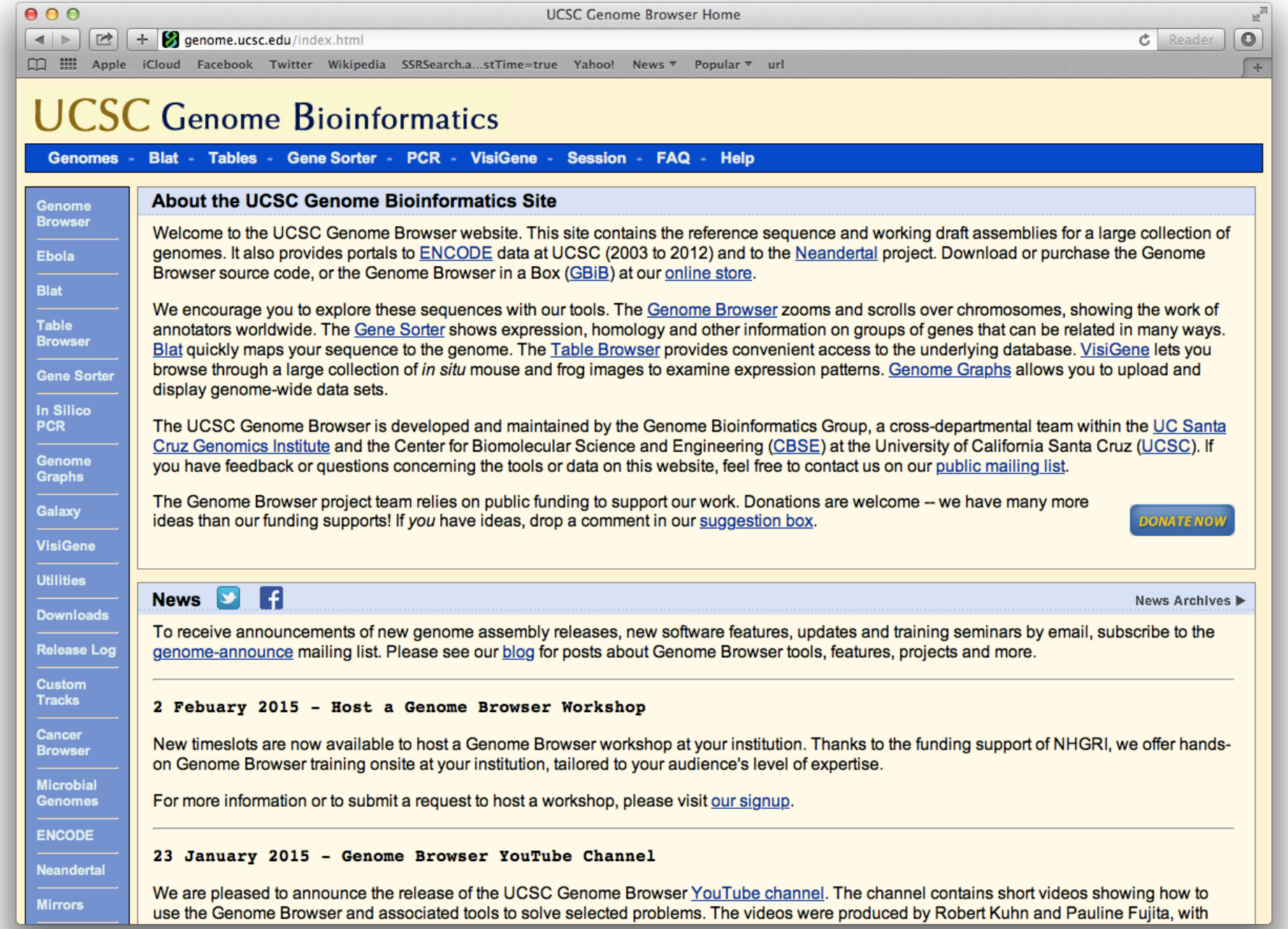
gulcin.ozer@osumc.edu

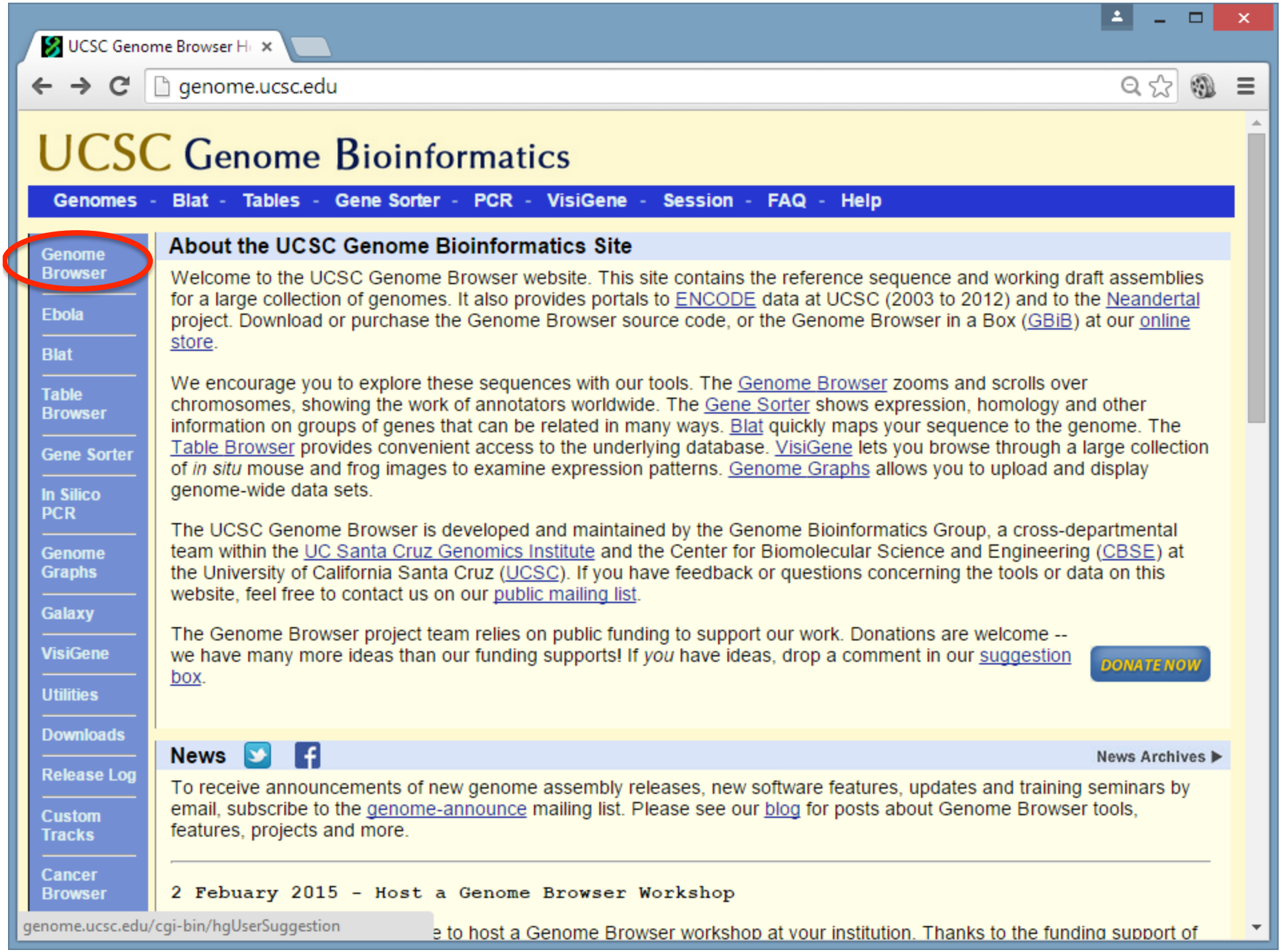
Department of Biomedical Informatics

The Ohio State University

Wexner Medical Center

November 3rd, 2015


UCSC Genome Browser Home
genome.ucsc.edu/index.html
Reader
Apple iCloud Facebook Twitter Wikipedia SSRSearch.a...stTime=true Yahoo! News Popular url
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
Ebola
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
Microbial Genomes
ENCODE
Neandertal
Mirrors
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute and the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.
The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives ▶
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
2 Febuary 2015 - Host a Genome Browser Workshop
New timeslots are now available to host a Genome Browser workshop at your institution. Thanks to the funding support of NHGRI, we offer hands-on Genome Browser training onsite at your institution, tailored to your audience's level of expertise.
For more information or to submit a request to host a workshop, please visit our signup.
23 January 2015 - Genome Browser YouTube Channel
We are pleased to announce the release of the UCSC Genome Browser YouTube channel. The channel contains short videos showing how to use the Genome Browser and associated tools to solve selected problems. The videos were produced by Robert Kuhn and Pauline Fujita, with


UCSC Genome Browser H
genome.ucsc.edu
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
Ebola
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute and the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.
The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
2 Febuary 2015 - Host a Genome Browser Workshop
genome.ucsc.edu/cgi-bin/hgUserSuggestion

Human (Homo sapiens) G

genome.ucsc.edu/cgi-bin/hgGateway

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	enter position, gene symbol or search terms

submit

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display

Human Genome Browser – hg19 assembly (sequences)

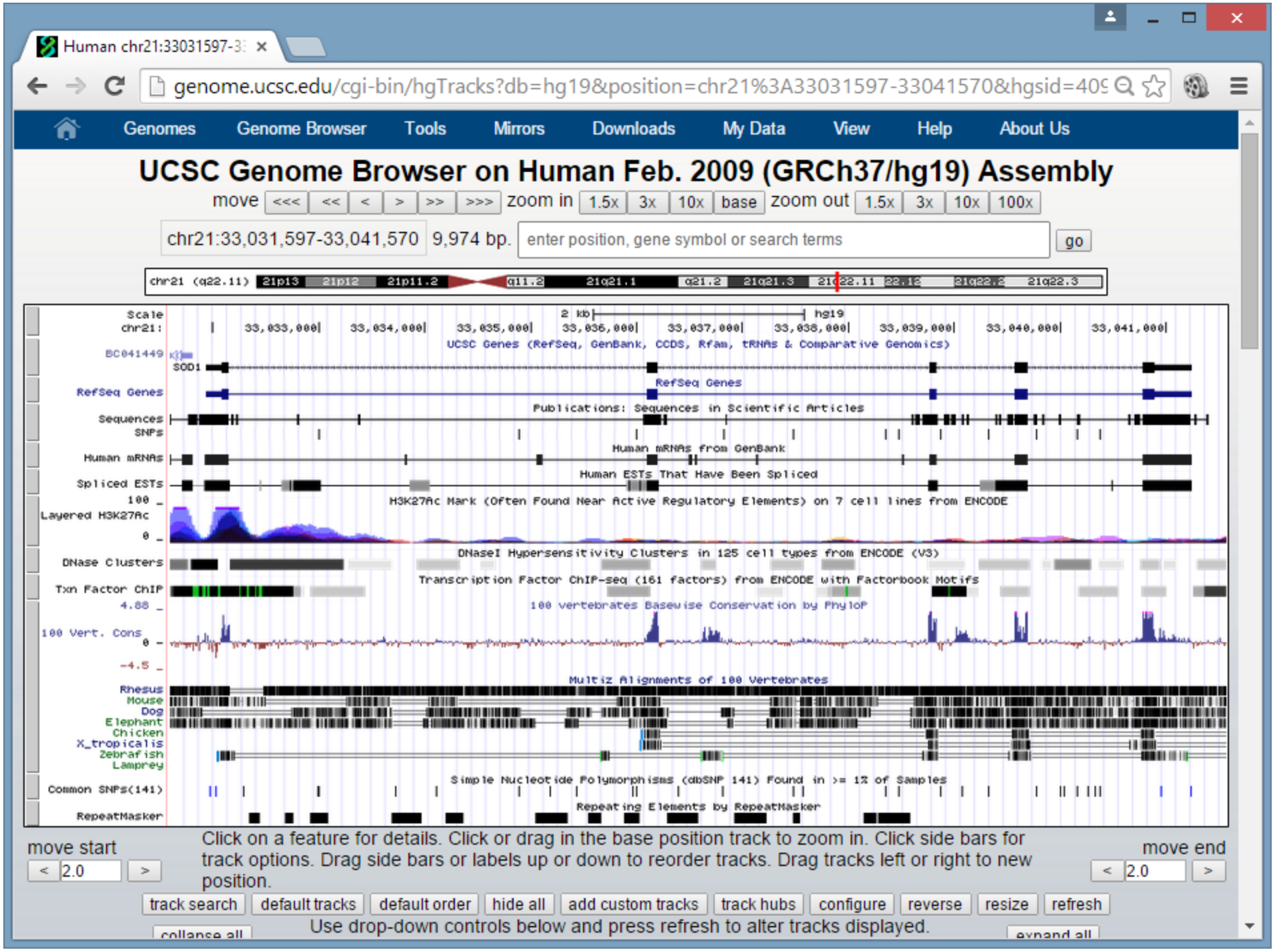
The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

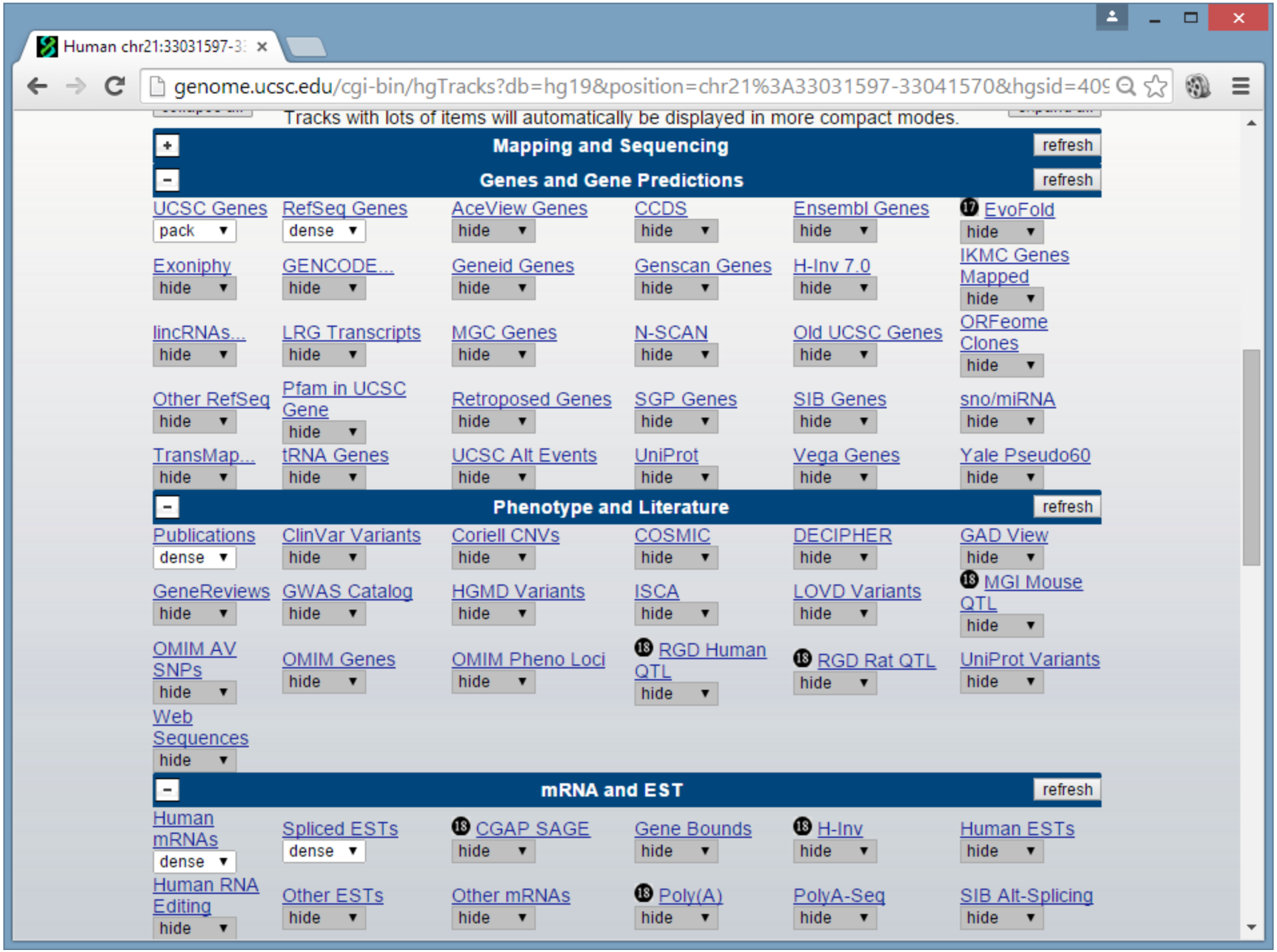
Homo sapiens
(Graphic courtesy of CBSE)

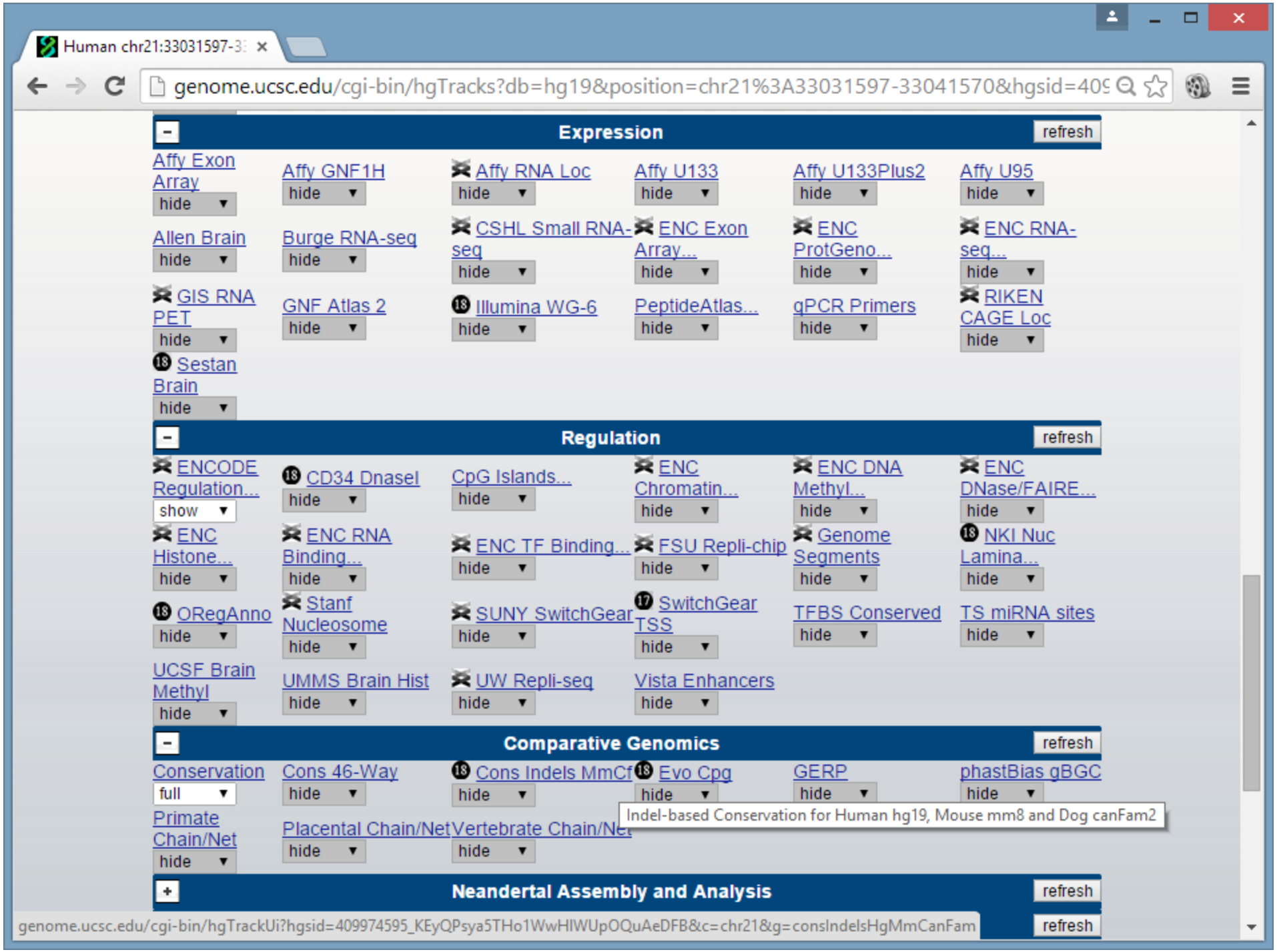
Sample position queries

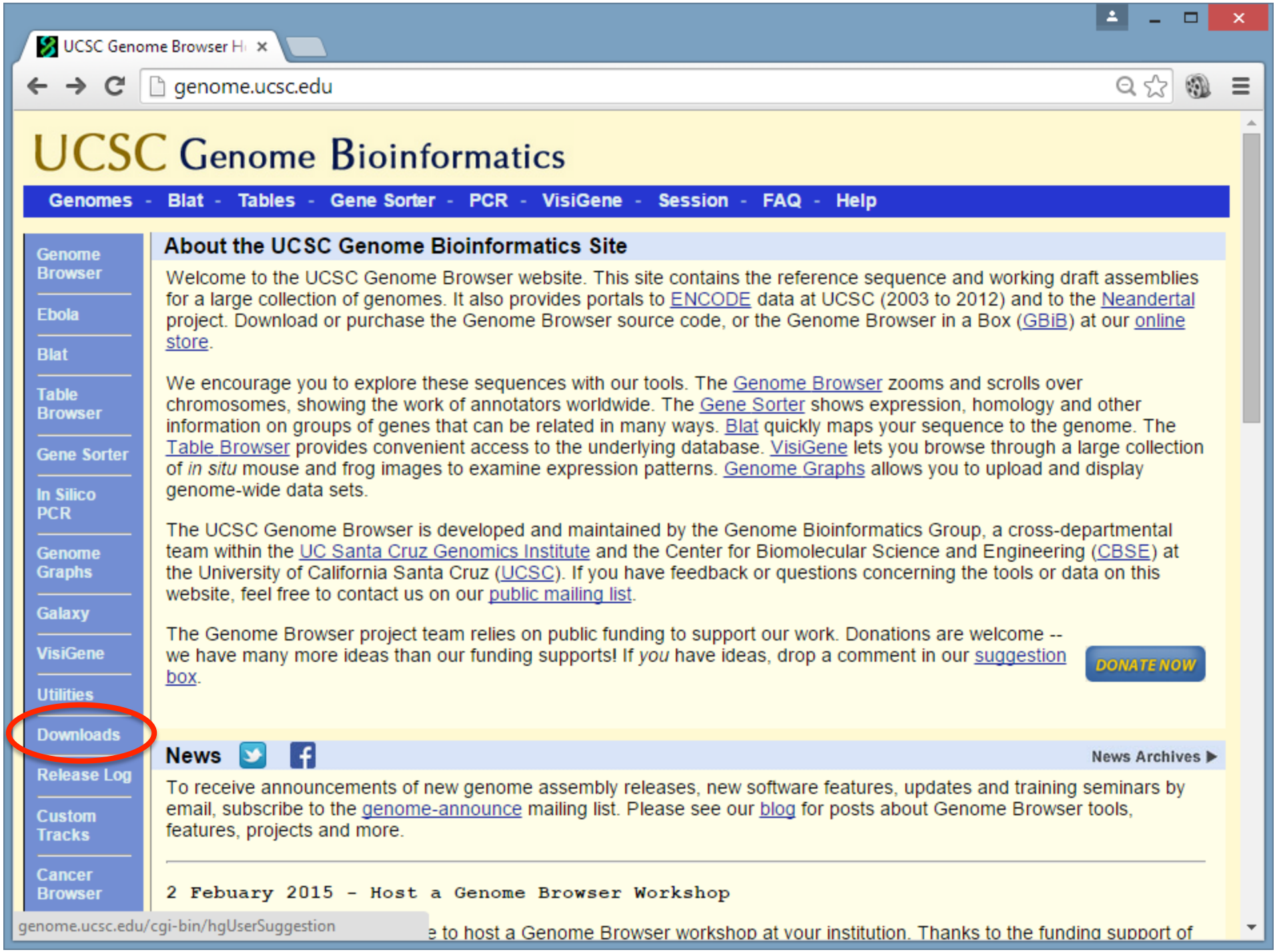
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

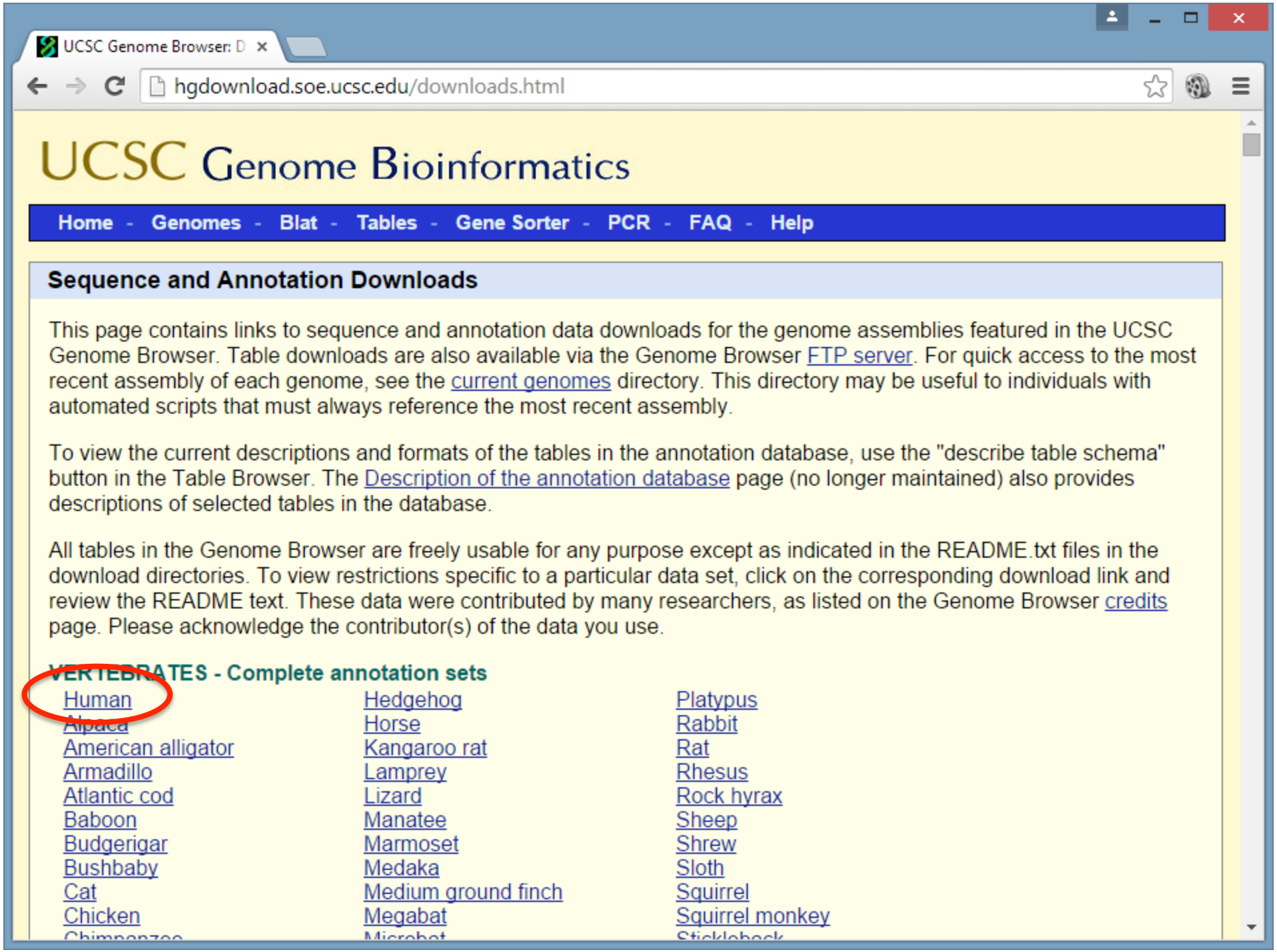
Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061:RH80175	Displays region between genome landmarks, such as the STS markers

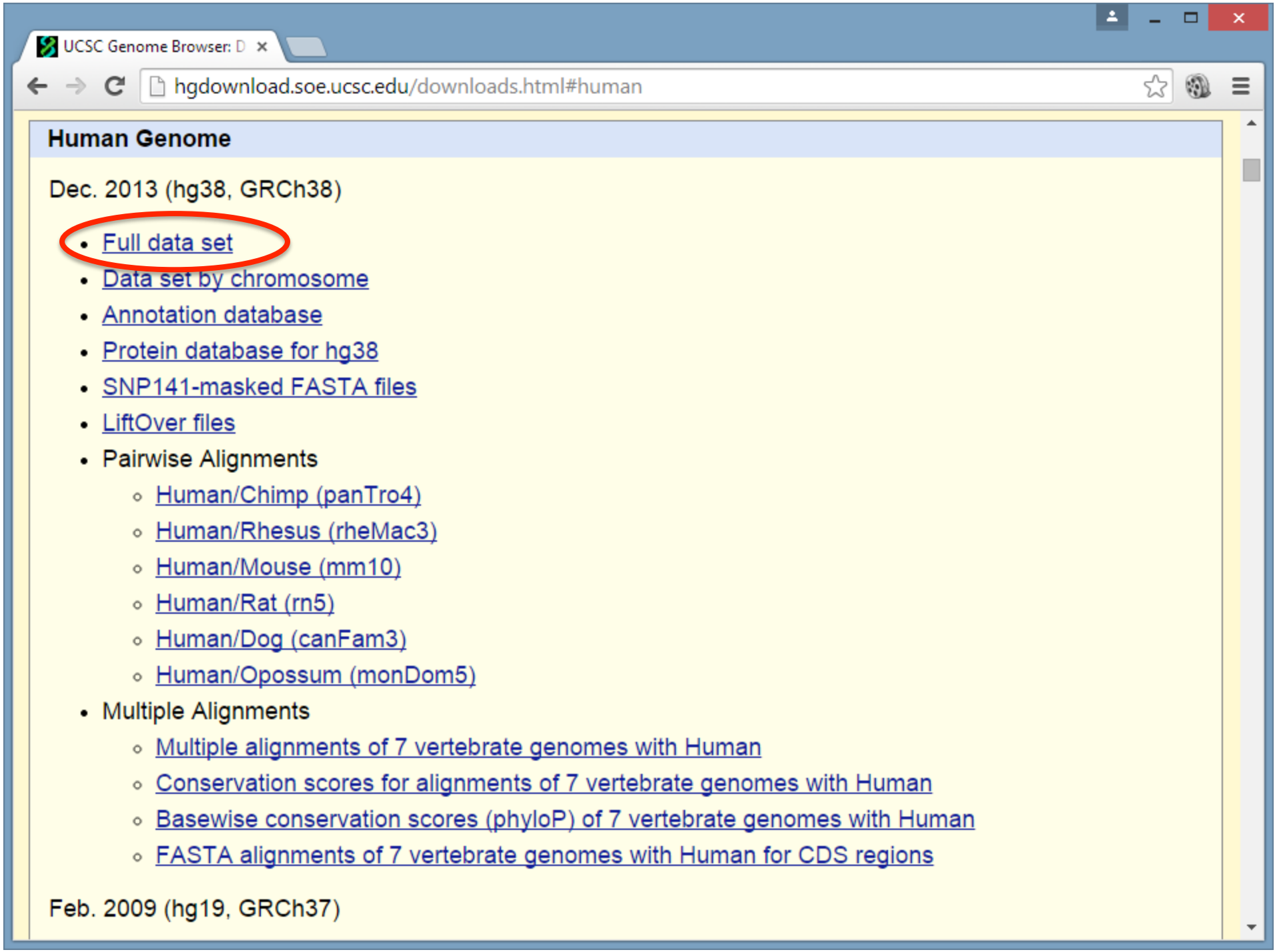

Human chr21:33031597-3
genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr21%3A33031597-33041570&hgsid=40
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr21:33,031,597-33,041,570 9,974 bp.
enter position, gene symbol or search terms
go
chr21 (q22.11)
Scale
chr21:
2 kb
hg19
33,033,000
33,034,000
33,035,000
33,036,000
33,037,000
33,038,000
33,039,000
33,040,000
33,041,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BC041449
SOD1
RefSeq Genes
Sequences
Publications: Sequences in Scientific Articles
SNPs
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(141)
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
move start
move end
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
Use drop-down controls below and press refresh to alter tracks displayed.
collapse all
expand all

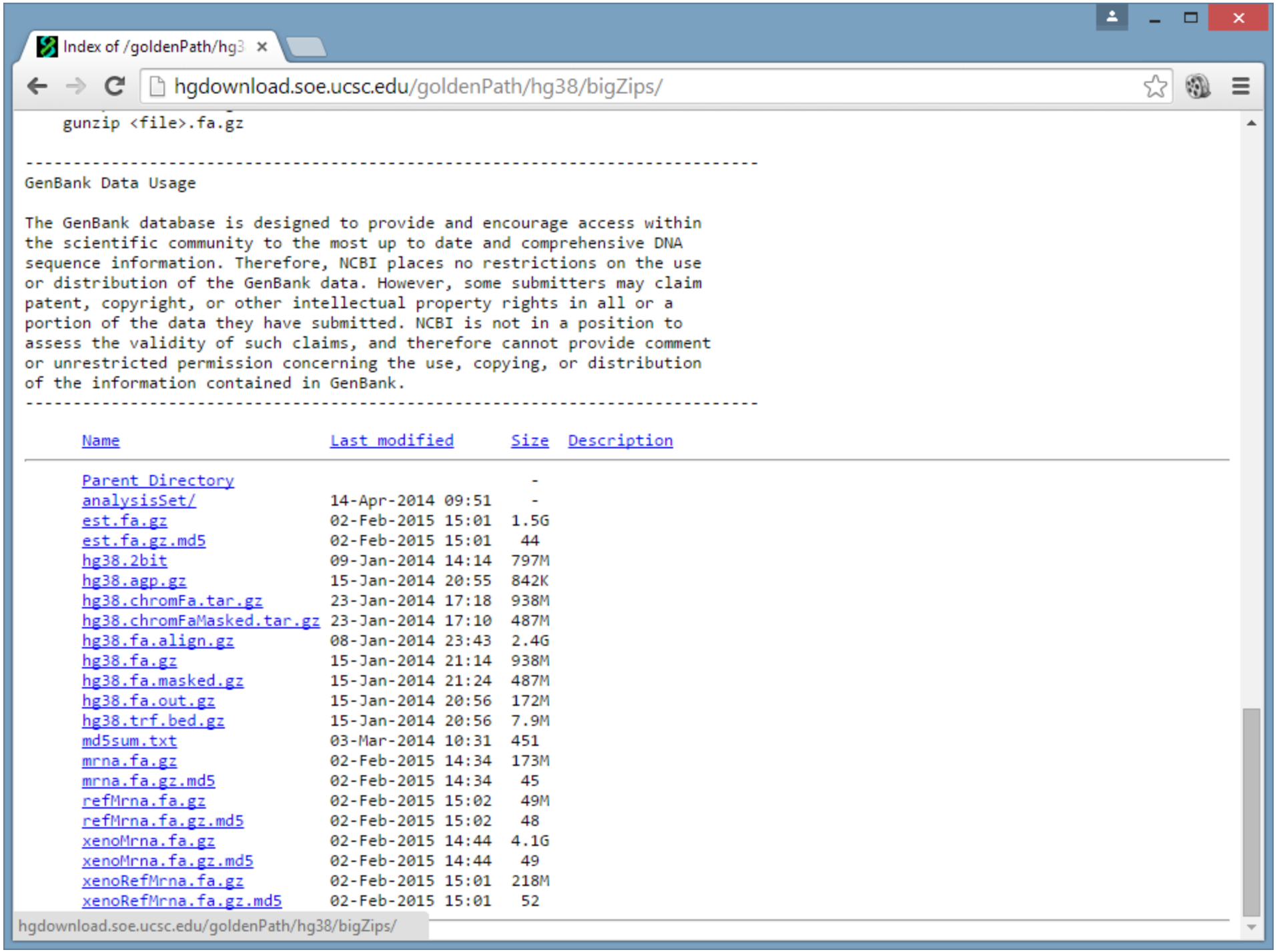

Human chr21:33031597-3
genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr21%3A33031597-33041570&hgsid=40
Tracks with lots of items will automatically be displayed in more compact modes.
Mapping and Sequencing
refresh
Genes and Gene Predictions
refresh
UCSC Genes pack
RefSeq Genes dense
AceView Genes hide
CCDS hide
Ensembl Genes hide
EvoFold hide
Exoniphy hide
GENCODE... hide
Geneid Genes hide
Genscan Genes hide
H-Inv 7.0 hide
IKMC Genes Mapped hide
lincRNAs... hide
LRG Transcripts hide
MGC Genes hide
N-SCAN hide
Old UCSC Genes hide
ORFeome Clones hide
Other RefSeq hide
Pfam in UCSC Gene hide
Retroposed Genes hide
SGP Genes hide
SIB Genes hide
sno/miRNA hide
TransMap... hide
tRNA Genes hide
UCSC Alt Events hide
UniProt hide
Vega Genes hide
Yale Pseudo60 hide
Phenotype and Literature
refresh
Publications dense
ClinVar Variants hide
Coriell CNVs hide
COSMIC hide
DECIPHER hide
GAD View hide
GeneReviews hide
GWAS Catalog hide
HGMD Variants hide
ISCA hide
LOVD Variants hide
MGI Mouse QTL hide
OMIM AV SNPs hide
OMIM Genes hide
OMIM Pheno Loci hide
RGD Human QTL hide
RGD Rat QTL hide
UniProt Variants hide
Web Sequences hide
mRNA and EST
refresh
Human mRNAs dense
Spliced ESTs dense
CGAP SAGE hide
Gene Bounds hide
H-Inv hide
Human ESTs hide
Human RNA Editing hide
Other ESTs hide
Other mRNAs hide
Poly(A) hide
PolyA-Seq hide
SIB Alt-Splicing hide


Human chr21:33031597-3
genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr21%3A33031597-33041570&hgsid=40
Expression
refresh
Affy Exon Array
hide
Affy GNF1H
hide
Affy RNA Loc
hide
Affy U133
hide
Affy U133Plus2
hide
Affy U95
hide
Allen Brain
hide
Burge RNA-seq
hide
CSHL Small RNA-seq
hide
ENC Exon Array...
hide
ENC ProtGeno...
hide
ENC RNA-seq...
hide
GIS RNA PET
hide
GNF Atlas 2
hide
Illumina WG-6
hide
PeptideAtlas...
hide
qPCR Primers
hide
RIKEN CAGE Loc
hide
Sestan Brain
hide
Regulation
refresh
ENCODE Regulation...
show
CD34 DnaseI
hide
CpG Islands...
hide
ENC Chromatin...
hide
ENC DNA Methyl...
hide
ENC DNase/FAIRE...
hide
ENC Histone...
hide
ENC RNA Binding...
hide
ENC TF Binding...
hide
FSU Repli-chip
hide
Genome Segments
hide
NKI Nuc Lamina...
hide
ORegAnno
hide
Stanf Nucleosome
hide
SUNY SwitchGear
hide
SwitchGear TSS
hide
TFBS Conserved
hide
TS miRNA sites
hide
UCSF Brain Methyl
hide
UMMS Brain Hist
hide
UW Repli-seq
hide
Vista Enhancers
hide
Comparative Genomics
refresh
Conservation
full
Cons 46-Way
hide
Cons Indels MmCf
hide
Evo Cpg
hide
GERP
hide
phastBias gBGC
hide
Primate Chain/Net
hide
Placental Chain/Net
hide
Vertebrate Chain/Net
hide
Indel-based Conservation for Human hg19, Mouse mm8 and Dog canFam2
Neandertal Assembly and Analysis
refresh
genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=409974595_KEyQPsya5THo1WwHlWUpOQuAeDFB&c=chr21&g=consIndelsHgMmCanFam
refresh


UCSC Genome Browser H
genome.ucsc.edu
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
Ebola
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute and the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.
The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
2 Febuary 2015 - Host a Genome Browser Workshop
genome.ucsc.edu/cgi-bin/hgUserSuggestion

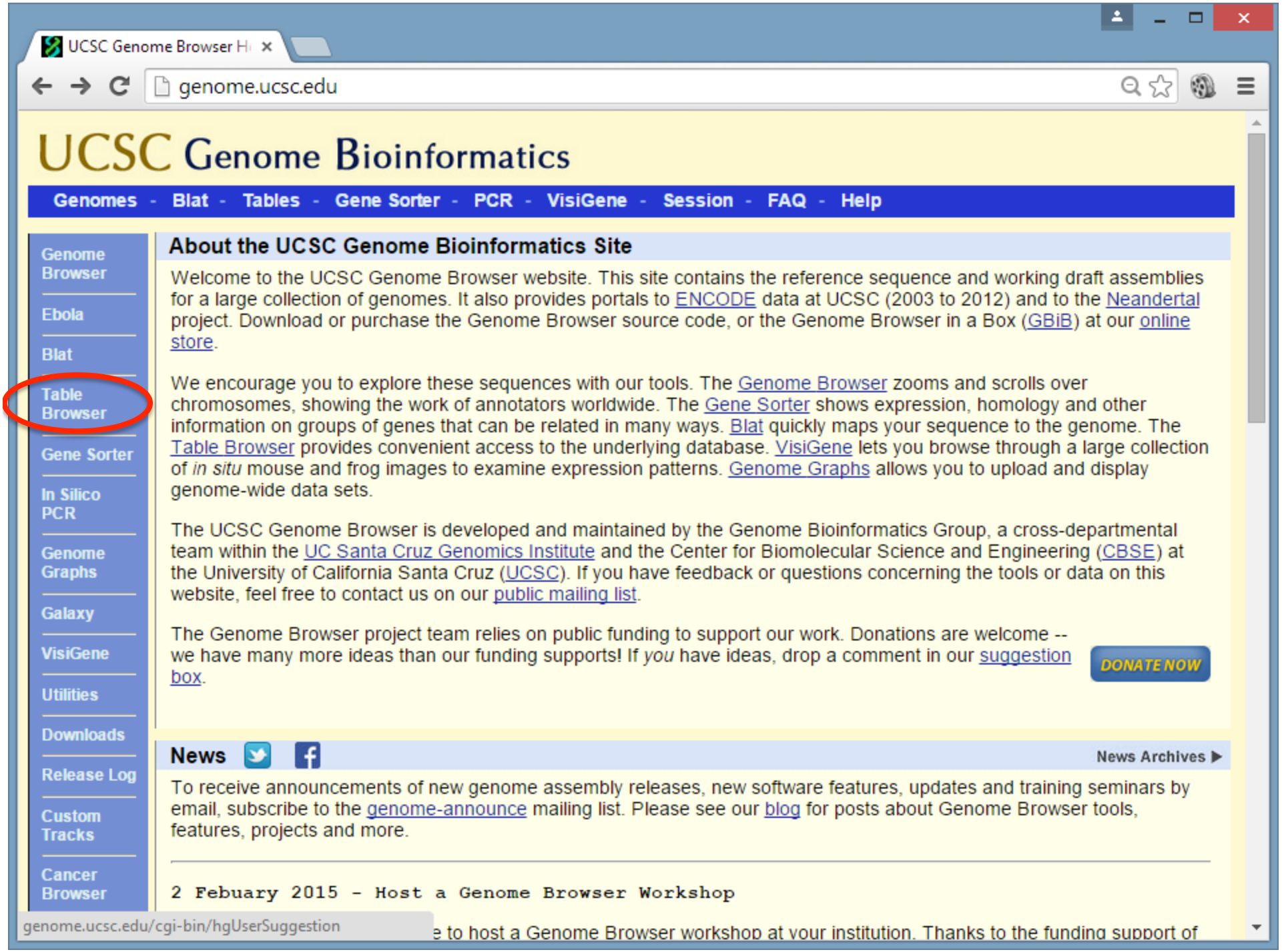

UCSC Genome Browser: D
hgdownload.soe.ucsc.edu/downloads.html
UCSC Genome Bioinformatics
Home - Genomes - Blat - Tables - Gene Sorter - PCR - FAQ - Help
Sequence and Annotation Downloads
This page contains links to sequence and annotation data downloads for the genome assemblies featured in the UCSC Genome Browser. Table downloads are also available via the Genome Browser FTP server. For quick access to the most recent assembly of each genome, see the current genomes directory. This directory may be useful to individuals with automated scripts that must always reference the most recent assembly.
To view the current descriptions and formats of the tables in the annotation database, use the "describe table schema" button in the Table Browser. The Description of the annotation database page (no longer maintained) also provides descriptions of selected tables in the database.
All tables in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. To view restrictions specific to a particular data set, click on the corresponding download link and review the README text. These data were contributed by many researchers, as listed on the Genome Browser credits page. Please acknowledge the contributor(s) of the data you use.
VERTEBRATES - Complete annotation sets
Human
Alpaca
American alligator
Armadillo
Atlantic cod
Baboon
Budgerigar
Bushbaby
Cat
Chicken
Chimpanzee
Hedgehog
Horse
Kangaroo rat
Lamprey
Lizard
Manatee
Marmoset
Medaka
Medium ground finch
Megabat
Microbat
Platypus
Rabbit
Rat
Rhesus
Rock hyrax
Sheep
Shrew
Sloth
Squirrel
Squirrel monkey
Stickleback


UCSC Genome Browser: D
hgdownload.soe.ucsc.edu/downloads.html#human
Human Genome
Dec. 2013 (hg38, GRCh38)
Full data set
Data set by chromosome
Annotation database
Protein database for hg38
SNP141-masked FASTA files
LiftOver files
Pairwise Alignments
Human/Chimp (panTro4)
Human/Rhesus (rheMac3)
Human/Mouse (mm10)
Human/Rat (rn5)
Human/Dog (canFam3)
Human/Opossum (monDom5)
Multiple Alignments
Multiple alignments of 7 vertebrate genomes with Human
Conservation scores for alignments of 7 vertebrate genomes with Human
Basewise conservation scores (phyloP) of 7 vertebrate genomes with Human
FASTA alignments of 7 vertebrate genomes with Human for CDS regions
Feb. 2009 (hg19, GRCh37)




```
    gunzip <file>.fa.gz

--------------------------------------------------------------------------------
GenBank Data Usage

The GenBank database is designed to provide and encourage access within
the scientific community to the most up to date and comprehensive DNA
sequence information. Therefore, NCBI places no restrictions on the use
or distribution of the GenBank data. However, some submitters may claim
patent, copyright, or other intellectual property rights in all or a
portion of the data they have submitted. NCBI is not in a position to
assess the validity of such claims, and therefore cannot provide comment
or unrestricted permission concerning the use, copying, or distribution
of the information contained in GenBank.
--------------------------------------------------------------------------------
```

Name	Last modified	Size	Description
Parent Directory		-	
analysisSet/	14-Apr-2014 09:51	-	
est.fa.gz	02-Feb-2015 15:01	1.5G	
est.fa.gz.md5	02-Feb-2015 15:01	44	
hg38.2bit	09-Jan-2014 14:14	797M	
hg38.agp.gz	15-Jan-2014 20:55	842K	
hg38.chromFa.tar.gz	23-Jan-2014 17:18	938M	
hg38.chromFaMasked.tar.gz	23-Jan-2014 17:10	487M	
hg38.fa.align.gz	08-Jan-2014 23:43	2.4G	
hg38.fa.gz	15-Jan-2014 21:14	938M	
hg38.fa.masked.gz	15-Jan-2014 21:24	487M	
hg38.fa.out.gz	15-Jan-2014 20:56	172M	
hg38.trf.bed.gz	15-Jan-2014 20:56	7.9M	
md5sum.txt	03-Mar-2014 10:31	451	
mrna.fa.gz	02-Feb-2015 14:34	173M	
mrna.fa.gz.md5	02-Feb-2015 14:34	45	
refMrna.fa.gz	02-Feb-2015 15:02	49M	
refMrna.fa.gz.md5	02-Feb-2015 15:02	48	
xenoMrna.fa.gz	02-Feb-2015 14:44	4.1G	
xenoMrna.fa.gz.md5	02-Feb-2015 14:44	49	
xenoRefMrna.fa.gz	02-Feb-2015 15:01	218M	
xenoRefMrna.fa.gz.md5	02-Feb-2015 15:01	52	


UCSC Genome Browser H
genome.ucsc.edu
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
Ebola
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Utilities
Downloads
Release Log
Custom Tracks
Cancer Browser
About the UCSC Genome Bioinformatics Site
Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.
The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute and the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.
The Genome Browser project team relies on public funding to support our work. Donations are welcome -- we have many more ideas than our funding supports! If you have ideas, drop a comment in our suggestion box.
DONATE NOW
News
News Archives
To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list. Please see our blog for posts about Genome Browser tools, features, projects and more.
2 Febuary 2015 - Host a Genome Browser Workshop
genome.ucsc.edu/cgi-bin/hgUserSuggestion

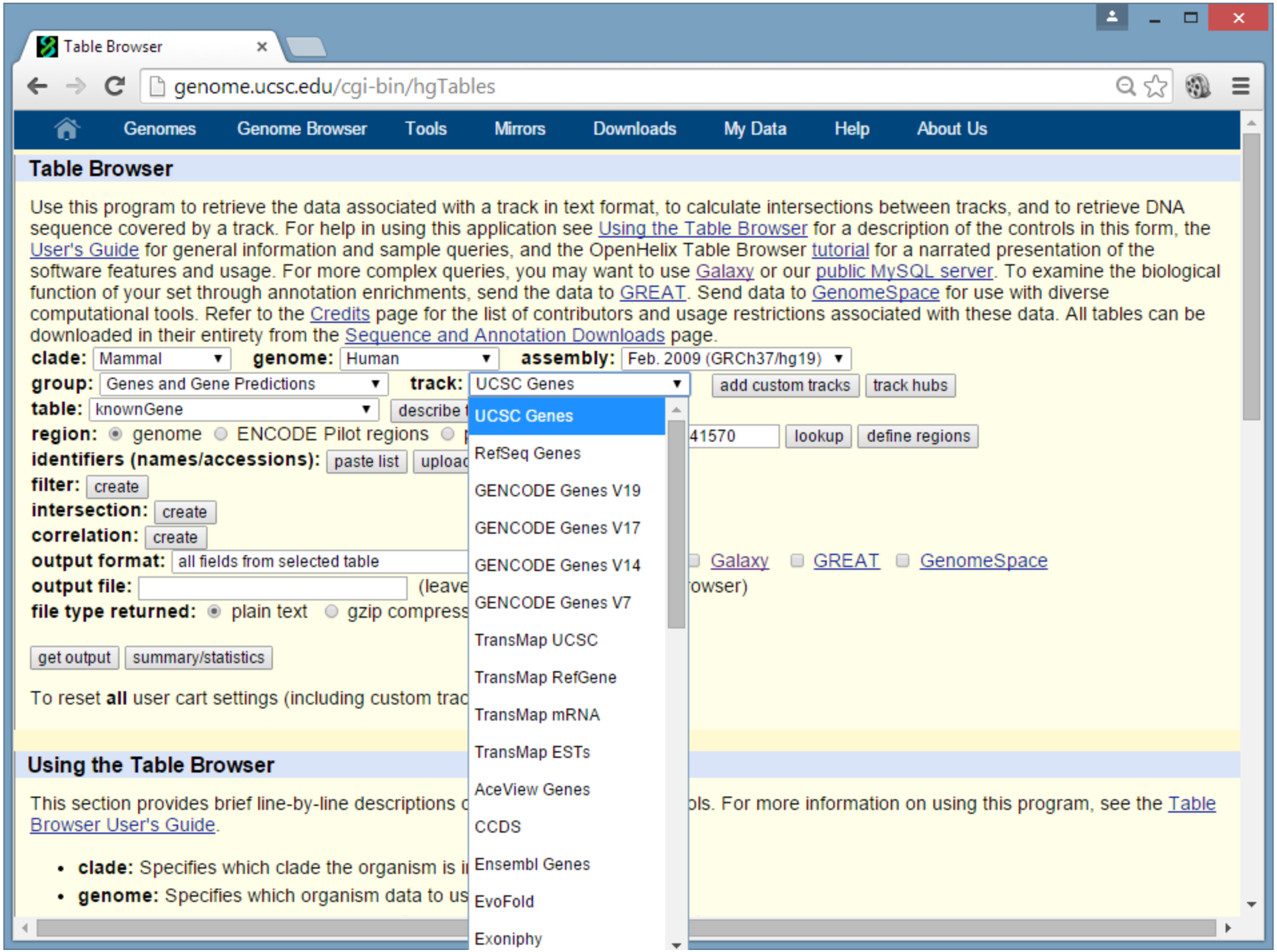


Table Browser
genome.ucsc.edu/cgi-bin/hgTables
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help
About Us
Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.
clade: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions
track: UCSC Genes
add custom tracks
track hubs
table: knownGene
describe
region: genome
ENCODE Pilot regions
41570
lookup
define regions
identifiers (names/accessions):
paste list
filter: create
intersection: create
correlation: create
output format: all fields from selected table
Galaxy
GREAT
GenomeSpace
output file:
file type returned: plain text
gzip compress
get output
summary/statistics
To reset all user cart settings (including custom trac
UCSC Genes
RefSeq Genes
GENCODE Genes V19
GENCODE Genes V17
GENCODE Genes V14
GENCODE Genes V7
TransMap UCSC
TransMap RefGene
TransMap mRNA
TransMap ESTs
AceView Genes
CCDS
Ensembl Genes
EvoFold
Exoniphy
Using the Table Browser
This section provides brief line-by-line descriptions
ols. For more information on using this program, see the Table Browser User's Guide.
clade: Specifies which clade the organism is i
genome: Specifies which organism data to us

Common Data Formats

- Fastq (reads with quality information)
- SAM/BAM (detailed alignment information)
- BED
- WIG
- GTF
- VCF

Sequence Reads with Quality Information

FASTQ is the most commonly used format

Wexner Medical Center

Sequence Alignment/Map (SAM) Format

- Standard format for short read alignment
  - SAM is plain text
  - BAM is binary and indexed
  - Details of the format

  http://samtools.github.io/hts-specs/SAMv1.pdf
- Two sections
  - Header
  - Alignment

The S

- Each
type
- File f
name
used

record

ole

ams

Tag	Description
@HD	The header line. The first line if present.
VN*	Format version. *Accepted format*: /^[0-9]+\.[0-9]+$/.
SO	Sorting order of alignments. *Valid values*: **unknown** (default), **unsorted**, **queryname** and **coordinate**. For coordinate sort, the major sort key is the RNAME field, with order defined by the order of @SQ lines in the header. The minor sort key is the POS field. For alignments with equal RNAME and POS, order is arbitrary. All alignments with '*' in RNAME field follow alignments with some other value but otherwise are in arbitrary order.
@SQ	Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order.
SN*	Reference sequence name. Each @SQ line must have a unique SN tag. The value of this field is used in the alignment records in RNAME and PNEXT fields. Regular expression: [!-)+-<>-~][!-~]*
LN*	Reference sequence length. *Range*: $[1,2^{31}-1]$
AS	Genome assembly identifier.
M5	MD5 checksum of the sequence in the uppercase, excluding spaces but including pads (as '*'s).
SP	Species.
UR	URI of the sequence. This value may start with one of the standard protocols, e.g http: or ftp:. If it does not start with one of these protocols, it is assumed to be a file-system path.
@RG	Read group. Unordered multiple @RG lines are allowed.
ID*	Read group identifier. Each @RG line must have a unique ID. The value of ID is used in the RG tags of alignment records. Must be unique among all read groups in header section. Read group IDs may be modified when merging SAM files in order to handle collisions.
CN	Name of sequencing center producing the read.
DS	Description.
DT	Date the run was produced (ISO8601 date or date/time).
FO	Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Multi-base flows are encoded in IUPAC format, and non-nucleotide flows by various other characters. *Format*: /*|[ACMGRSVTWYHKDBN]+/
KS	The array of nucleotide bases that correspond to the key sequence of each read.
LB	Library.
PG	Programs used for processing the read group.
PI	Predicted median insert size.
PL	Platform/technology used to produce the reads. *Valid values*: **CAPILLARY**, **LS454**, **ILLUMINA**, **SOLID**, **HELICOS**, **IONTORRENT** and **PACBIO**.
PU	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier.
SM	Sample. Use pool name where a pool is being sequenced.
@PG	Program.
ID*	Program record identifier. Each @PG line must have a unique ID. The value of ID is used in the alignment PG tag and PP tags of other @PG lines. PG IDs may be modified when merging SAM files in order to handle collisions.
PN	Program name
CL	Command line
PP	Previous @PG-ID. Must match another @PG header's ID tag. @PG records may be chained using PP tag, with the last record in the chain having no PP tag. This chain defines the order of programs that have been applied to the alignment. PP values may be modified when merging SAM files in order to handle collisions of PG IDs. The first PG record in a chain (i.e. the one referred to

The SAM/BAM Format - Alignment

Each line has 11 mandatory fields

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

The SAM/BAM Format - Alignment

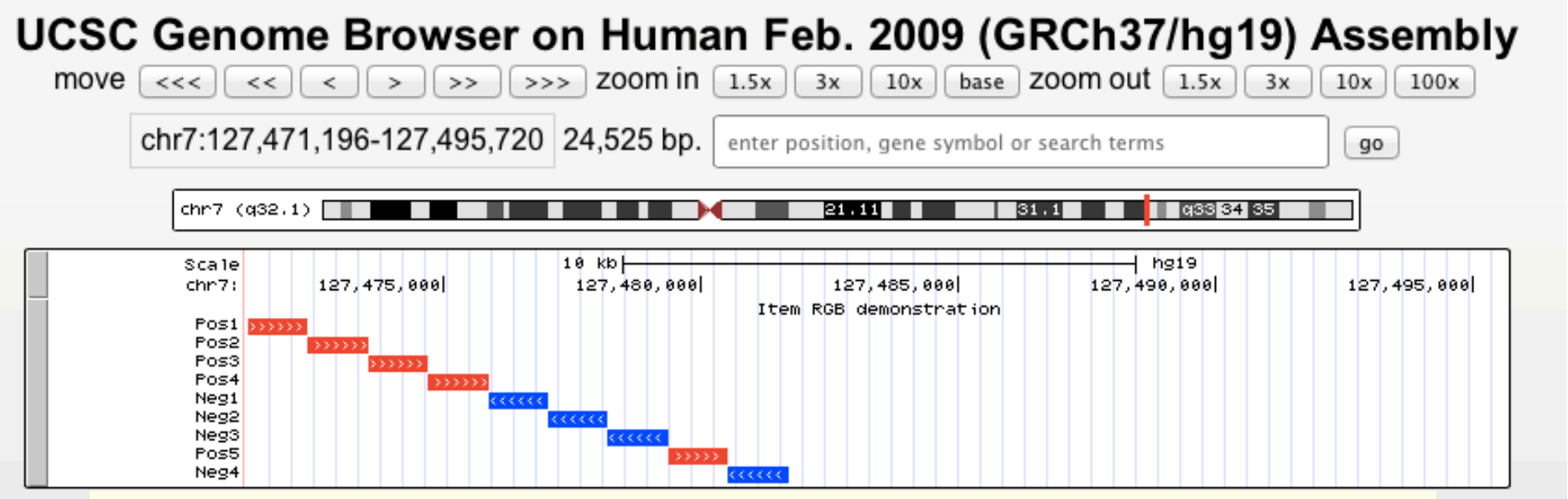
```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:chrM LN:16299
@SQ     SN:chr1 LN:197195432
@SQ     SN:chr2 LN:181748087
@SQ     SN:chr3 LN:159599783
@SQ     SN:chr4 LN:155630120
@SQ     SN:chr5 LN:152537259
@SQ     SN:chr6 LN:149517037
@SQ     SN:chr7 LN:152524553
@SQ     SN:chr8 LN:131738871
@SQ     SN:chr9 LN:124076172
@SQ     SN:chr10        LN:129993255
@SQ     SN:chr11        LN:121843856
@SQ     SN:chr12        LN:121257530
@SQ     SN:chr13        LN:120284312
@SQ     SN:chr14        LN:125194864
@SQ     SN:chr15        LN:103494974
@SQ     SN:chr16        LN:98319150
@SQ     SN:chr17        LN:95272651
@SQ     SN:chr18        LN:90772031
@SQ     SN:chr19        LN:61342430
@SQ     SN:chrX LN:166650296
@SQ     SN:chrY LN:15902555
@RG     ID:110826_I631_FCB01VGACXX_L6_MOUytnRAQDIAAPEI-2        PL:illumina     PU:110826_I631_FCB01VGACXX_L6   LB:MOUytnRAQDIAAPEI-2   SM:PF216
@RG     ID:110901_I875_FCC00L1ACXX_L6_MOUytnRAQDIABPEI-2        PL:illumina     PU:110901_I875_FCC00L1ACXX_L6   LB:MOUytnRAQDIABPEI-2   SM:PF216
@RG     ID:110910_I658_FCD059YACXX_L2_MOUytnRAQDIAAPEI-2        PL:illumina     PU:110910_I658_FCD059YACXX_L2   LB:MOUytnRAQDIAAPEI-2   SM:PF216
@RG     ID:110910_I658_FCD059YACXX_L5_MOUytnRAQDIABPEI-2        PL:illumina     PU:110910_I658_FCD059YACXX_L5   LB:MOUytnRAQDIABPEI-2   SM:PF216
@RG     ID:110910_I658_FCD059YACXX_L8_MOUytnRAQDIBAPEI-2        PL:illumina     PU:110910_I658_FCD059YACXX_L8   LB:MOUytnRAQDIBAPEI-2   SM:PF216
@PG     ID:bwa  PN:bwa  VN:0.5.9-r16
FCD059YACXX:2:1105:1812:30275   99      chrM    1       60      90M     =       353     441     GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAACACAAAGGTTTGGTCCTGGCCTTA
@@@DDDFDDHGHHJIIIJIJJJIJIJIGHIJGGGIEDHIIADHIIJFIGIJJJJIIJGGIGIJIFGHGIJJJBE@CEHHFHFFFF@BECE      X0:i:1  X1:i:0  MD:Z:90 RG:Z:110910_I658_FCD059YACXX_L2_MOUytnRAQDIAAPEI-2
XG:i:0  AM:i:37 NM:i:0  SM:i:37 XM:i:0  XO:i:0  MQ:i:60 XT:A:U

FCC00L1ACXX:6:1105:10157:68134  163     chrM    1       60      90M     =       372     460     GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAACACAAAGGTTTGGTCCTGGCCTTA
@BCFFFFFGHHHHJJJJJJJJJJIJJJJIJJJIJGIJJIJJJJJJJJIJGHGHIIJJJJIJJIIJJIJIJIJ@HGCEHGGHGHFFFFFEEE      X0:i:1  X1:i:0  MD:Z:90 RG:Z:110901_I875_FCC00L1ACXX_L6_MOUytnRAQDIABPEI-2
XG:i:0  AM:i:37 NM:i:0  SM:i:37 XM:i:0  XO:i:0  MQ:i:60 XT:A:U

FCC00L1ACXX:6:2307:15656:134100 163     chrM    1       60      90M     =       355     443     GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAACACAAAGGTTTGGTCCTGGCCTTA
B@@FFFF>DHHHHJIJJJIJIJJIIJIIJJJJIJJIJEGJJJJGHGHJJHHHIJGHEIIIJJJIIJIJJIIJJ3CEHHHGHFFFFDEEC      X0:i:1  X1:i:0  MD:Z:90 RG:Z:110901_I875_FCC00L1ACXX_L6_MOUytnRAQDIABPEI-2
XG:i:0  AM:i:37 NM:i:0  SM:i:37 XM:i:0  XO:i:0  MQ:i:60 XT:A:U

FCB01VGACXX:6:1208:1975:93153   99      chrM    3       60      88M2S   =       359     445     TAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAACACAAAGGTTTAGTCCTGGCCTTATA
@;?DDB8?=CFDFGIIIIG>CEG@FFGHAAGHEGGF9CCF*?GI>FH>B@FDH9DH:48BFFG:3CFHDAC).=@)).77??C(67?###      X0:i:1  X1:i:0  XC:i:88 MD:Z:75G12       RG:Z:110826_I631_FCB01VGACXX_L6_MOUytnRAQDIAAPEI-2
XG:i:0  AM:i:37 NM:i:1  SM:i:37 XM:i:1  XO:i:0  MQ:i:60 XT:A:U
```

CIGAR string - 9M32N8M

```
REF:   AGCTAGCATCGTGTAACCGGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
READ:            GTGTAACCC.............................TCAGAATA
```

Wexner Medical Center

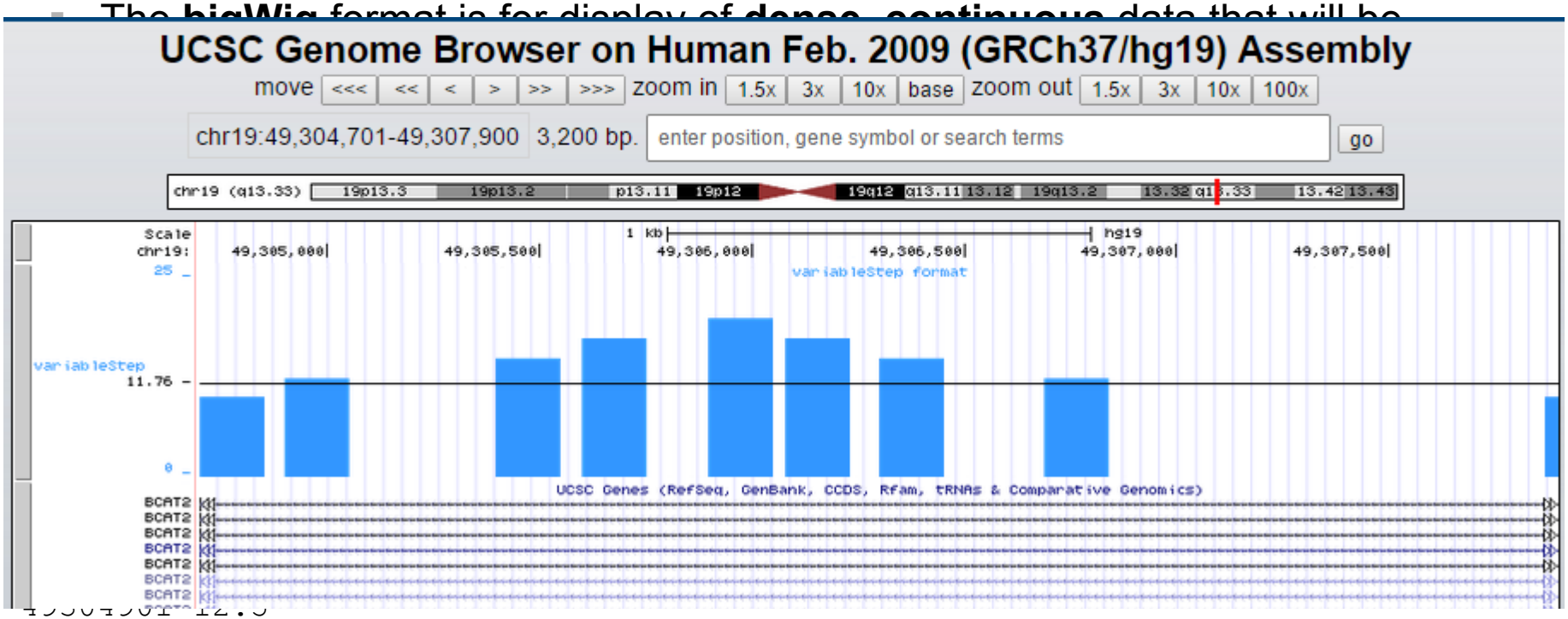
BED file format




```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

WIG File Format

- Wiggle format (WIG) allows the display of continuous-valued data in a track format.
- The **bigWig** format is for display of **dense, continuous** data that will be




```
49305401 15.0
49305601 17.5
49305901 20.0
49306081 17.5
49306301 15.0
49306691 12.5
49307871 10.0
```

Wexner Medical Center

VCF File Format

- Variant Call Format (VCF) is a flexible and extendable format for variation data

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

https://wiki.nci.nih.gov/display/TCGA/TCGA+Variant+Call+Format+(VCF)+1.1+Specification

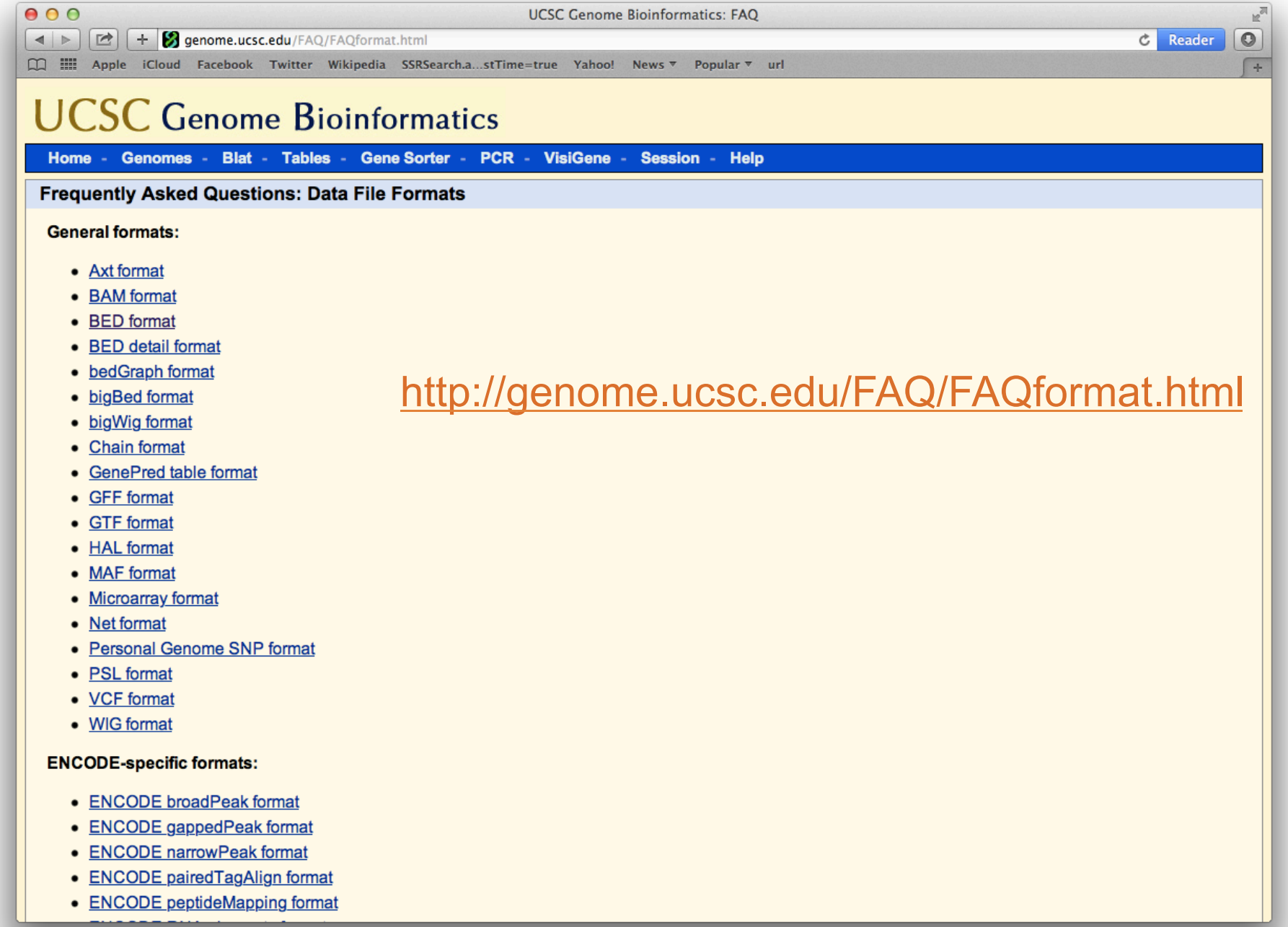
GTF File Format

- Gene Transfer Format
- Structure

```
<seqname> <source> <feature> <start> <end> <score>
<strand> <frame> [attributes] [comments]
```

- Simple example with 3 translated exons. Order of rows is not important.

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```


UCSC Genome Bioinformatics: FAQ
genome.ucsc.edu/FAQ/FAQformat.html
Reader
Apple iCloud Facebook Twitter Wikipedia SSRSearch.a...stTime=true Yahoo! News Popular url
UCSC Genome Bioinformatics
Home - Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - Help
Frequently Asked Questions: Data File Formats
General formats:
Axt format
BAM format
BED format
BED detail format
bedGraph format
bigBed format
bigWig format
Chain format
GenePred table format
GFF format
GTF format
HAL format
MAF format
Microarray format
Net format
Personal Genome SNP format
PSL format
VCF format
WIG format
ENCODE-specific formats:
ENCODE broadPeak format
ENCODE gappedPeak format
ENCODE narrowPeak format
ENCODE pairedTagAlign format
ENCODE peptideMapping format
http://genome.ucsc.edu/FAQ/FAQformat.html

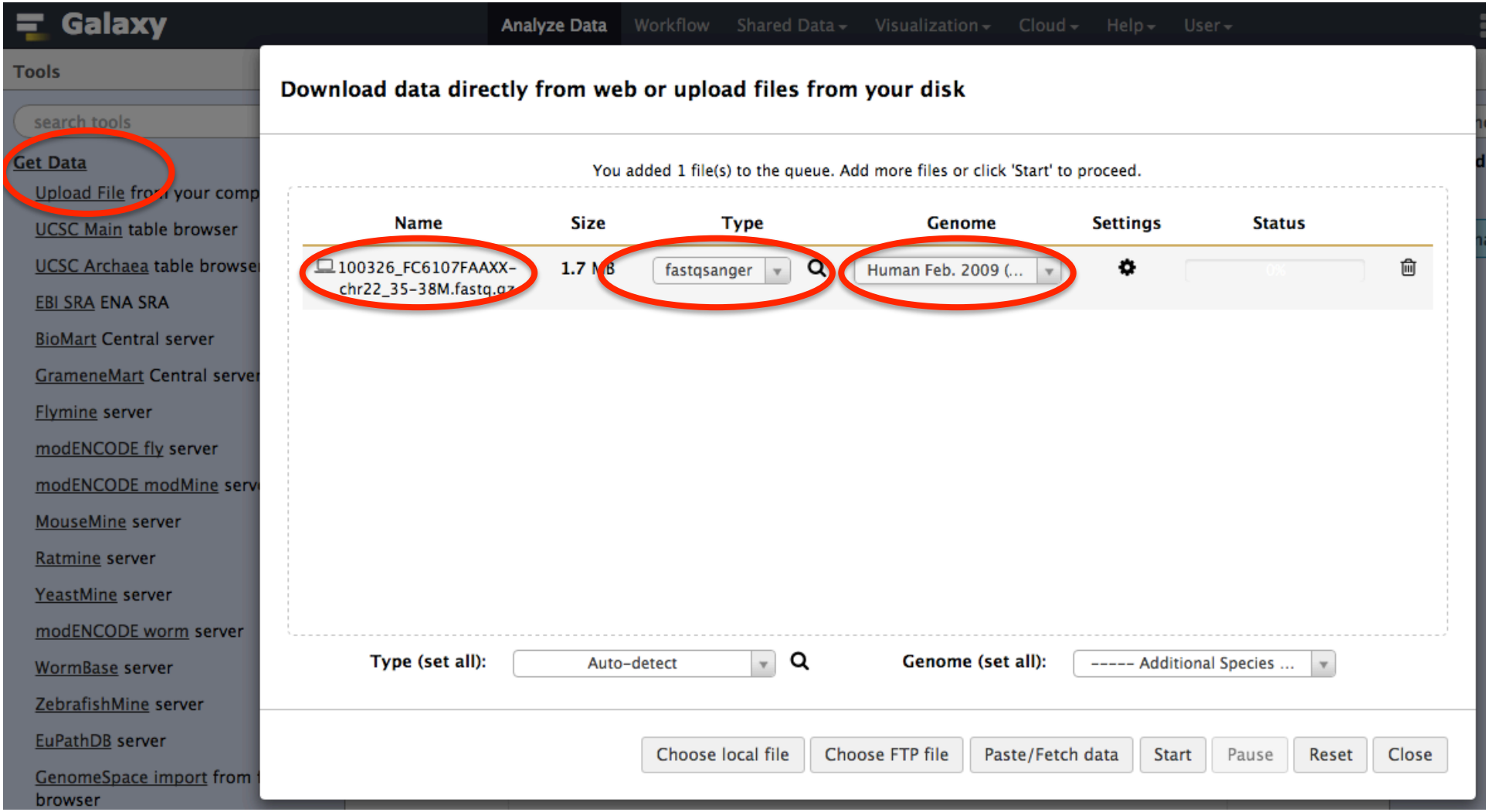
https://osu.box.com/exome-seq-example

Create a Galaxy Account
https://usegalaxy.org/

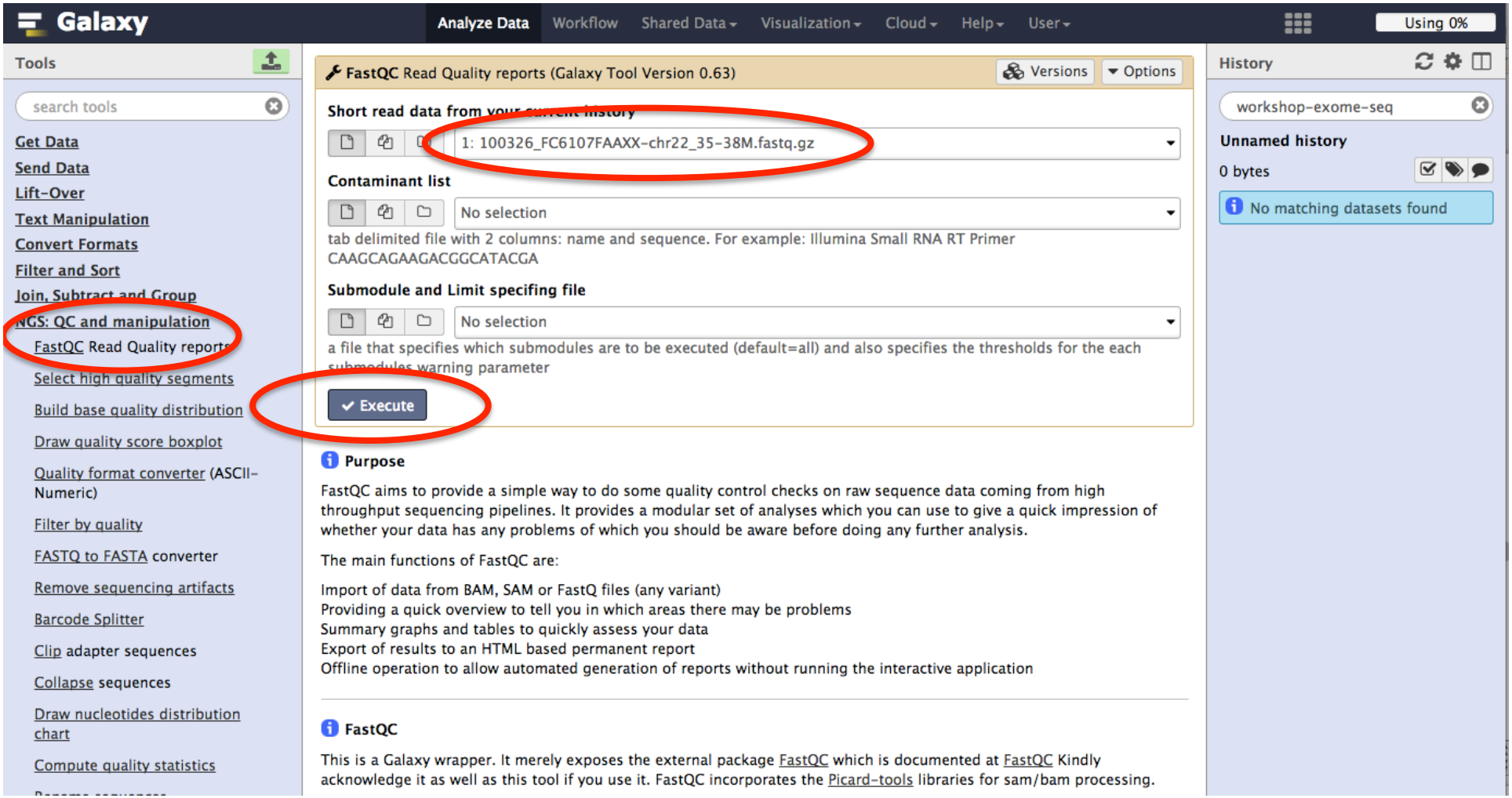
History window

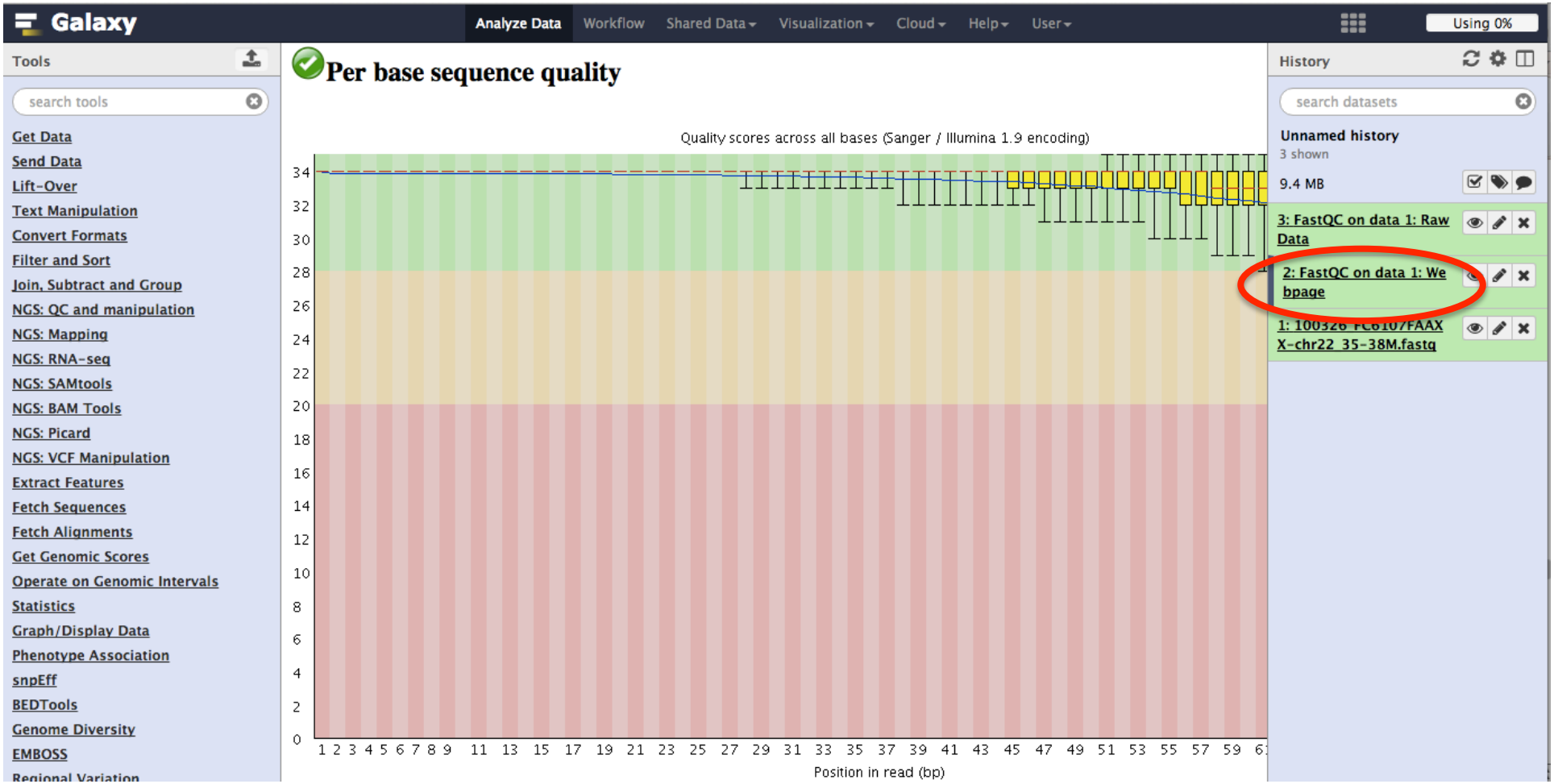
- ✓ All analysis steps are saved
- ✓ Data is not overwritten
- ✓ Output can be visualized and downloaded
- ✓ Can create a workflow to repeat an analysis

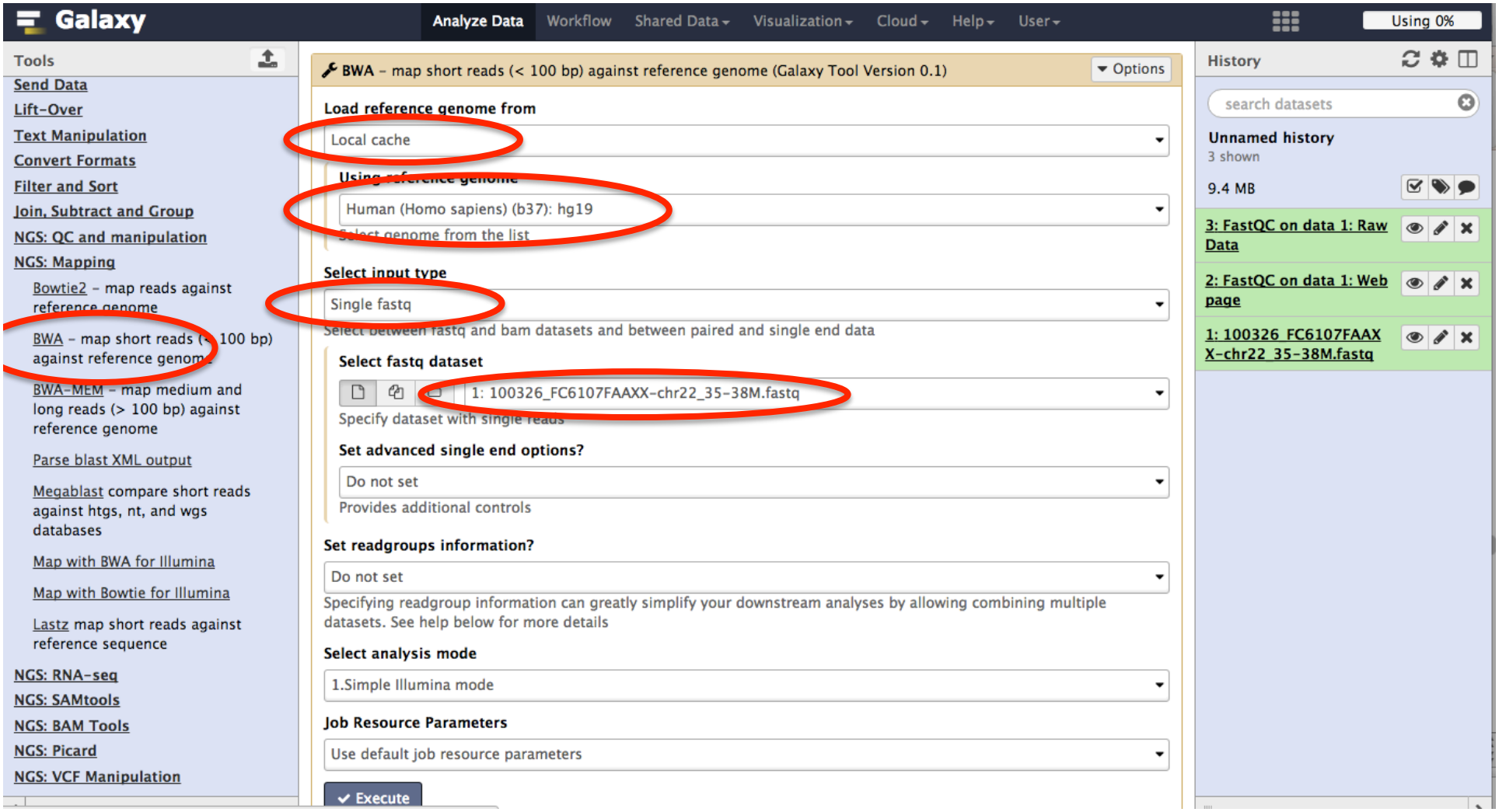
- ✓ Public instance
- ✓ No need of programming/cluster computing experience
- ✓ Integrates many bioinformatics tools within one interface
- ✓ Keeps track of all the steps performed in an analysis

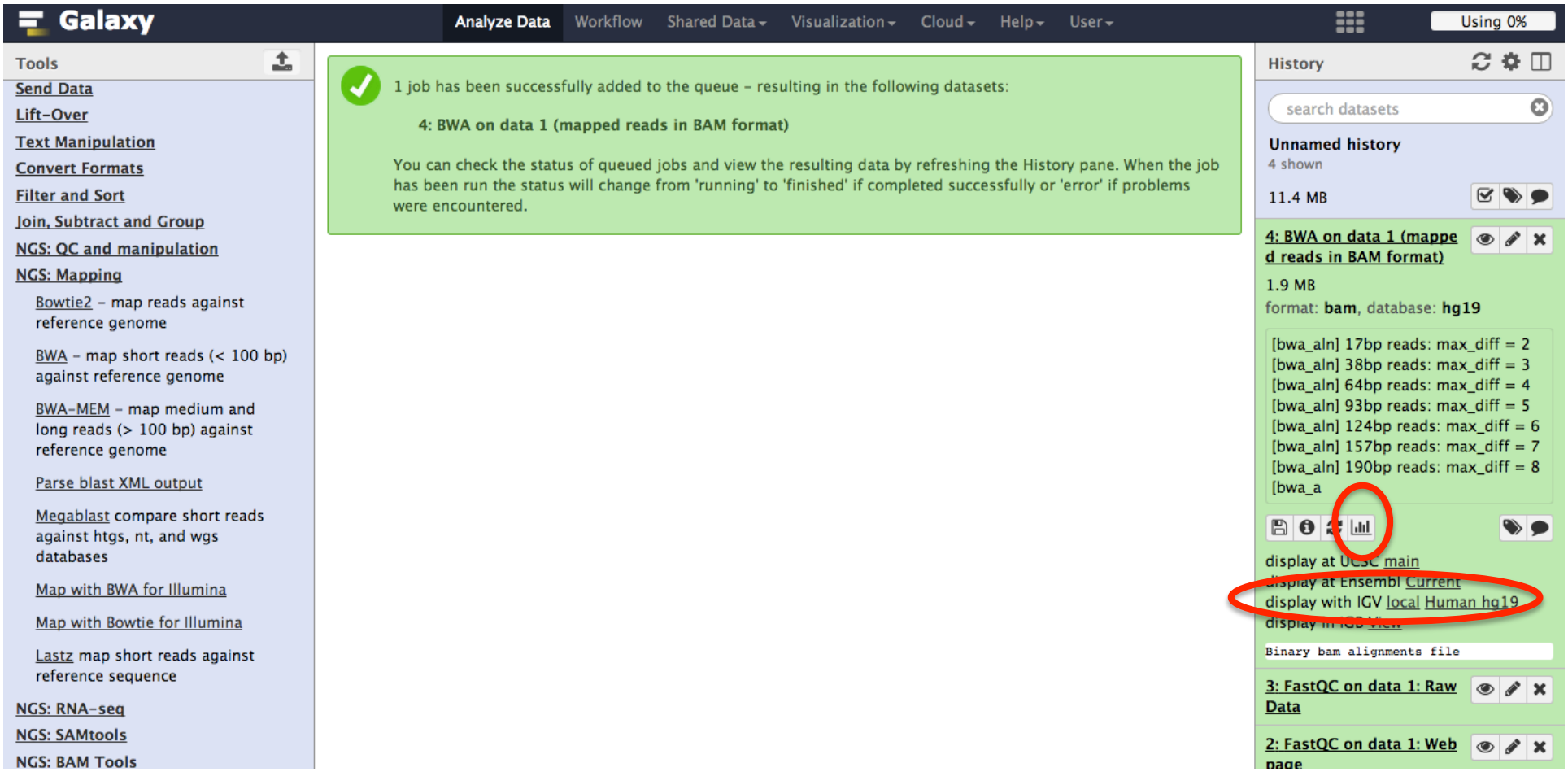

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Tools
search tools
Get Data
Upload File from your comp
UCSC Main table browser
UCSC Archaea table browse
EBI SRA ENA SRA
BioMart Central server
GrameneMart Central serve
Flymine server
modENCODE fly server
modENCODE modMine serv
MouseMine server
Ratmine server
YeastMine server
modENCODE worm server
WormBase server
ZebrafishMine server
EuPathDB server
GenomeSpace import from
browser
Download data directly from web or upload files from your disk
You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.
Name
Size
Type
Genome
Settings
Status
100326_FC6107FAAXX-chr22_35-38M.fastq.gz
1.7 MB
fastqsanger
Human Feb. 2009 (...
0%
Type (set all):
Auto-detect
Genome (set all):
----- Additional Species ...
Choose local file
Choose FTP file
Paste/Fetch data
Start
Pause
Reset
Close

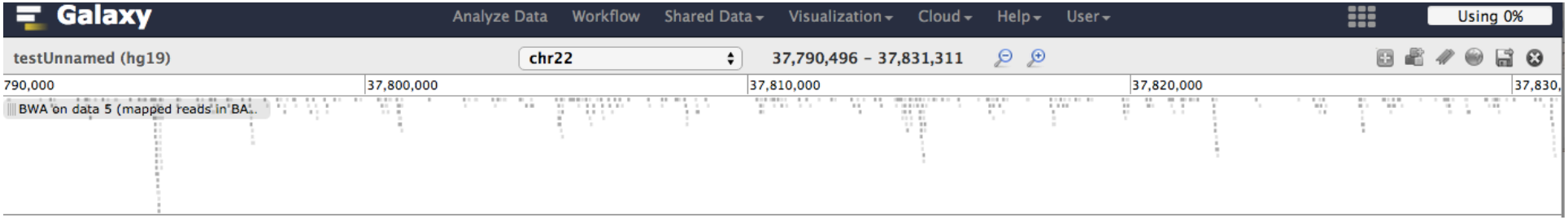
Example dataset: https://osu.app.box.com/exome-seq-example

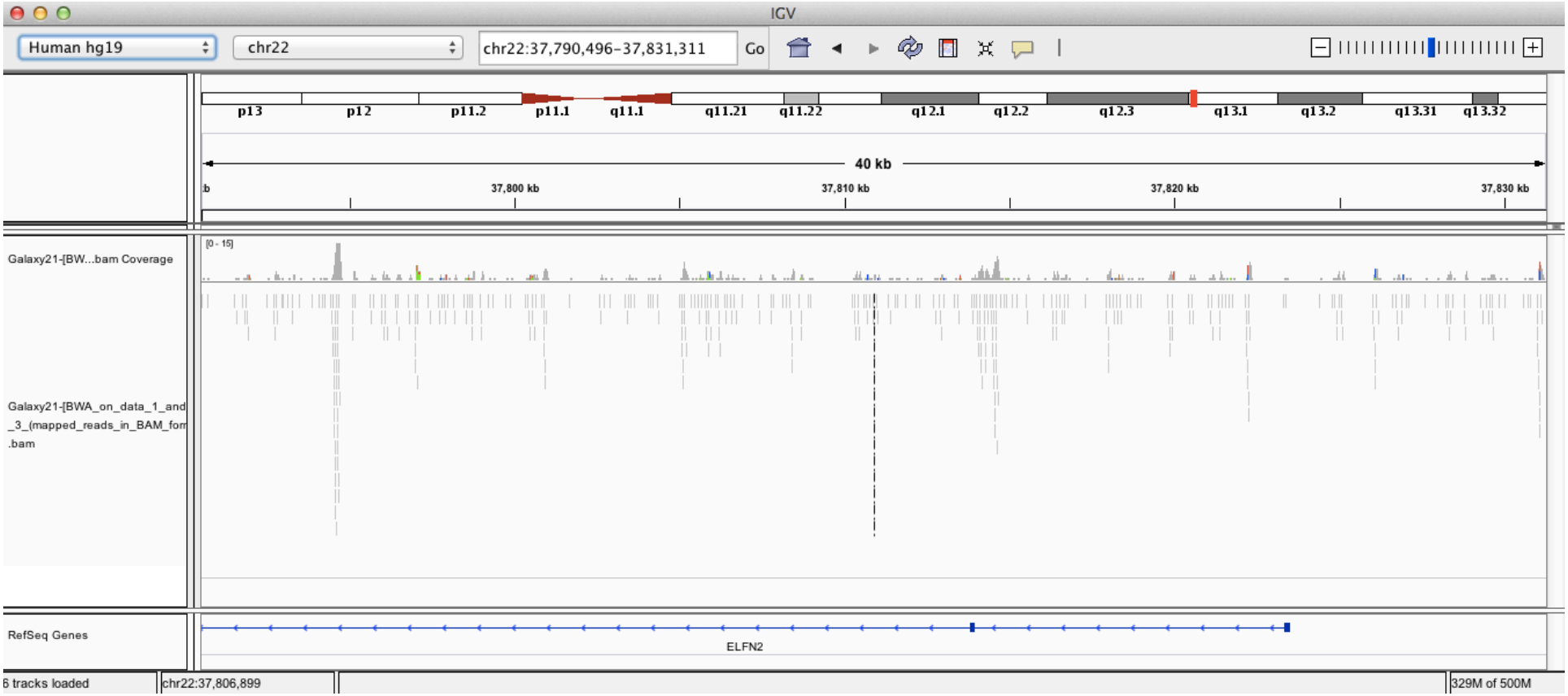


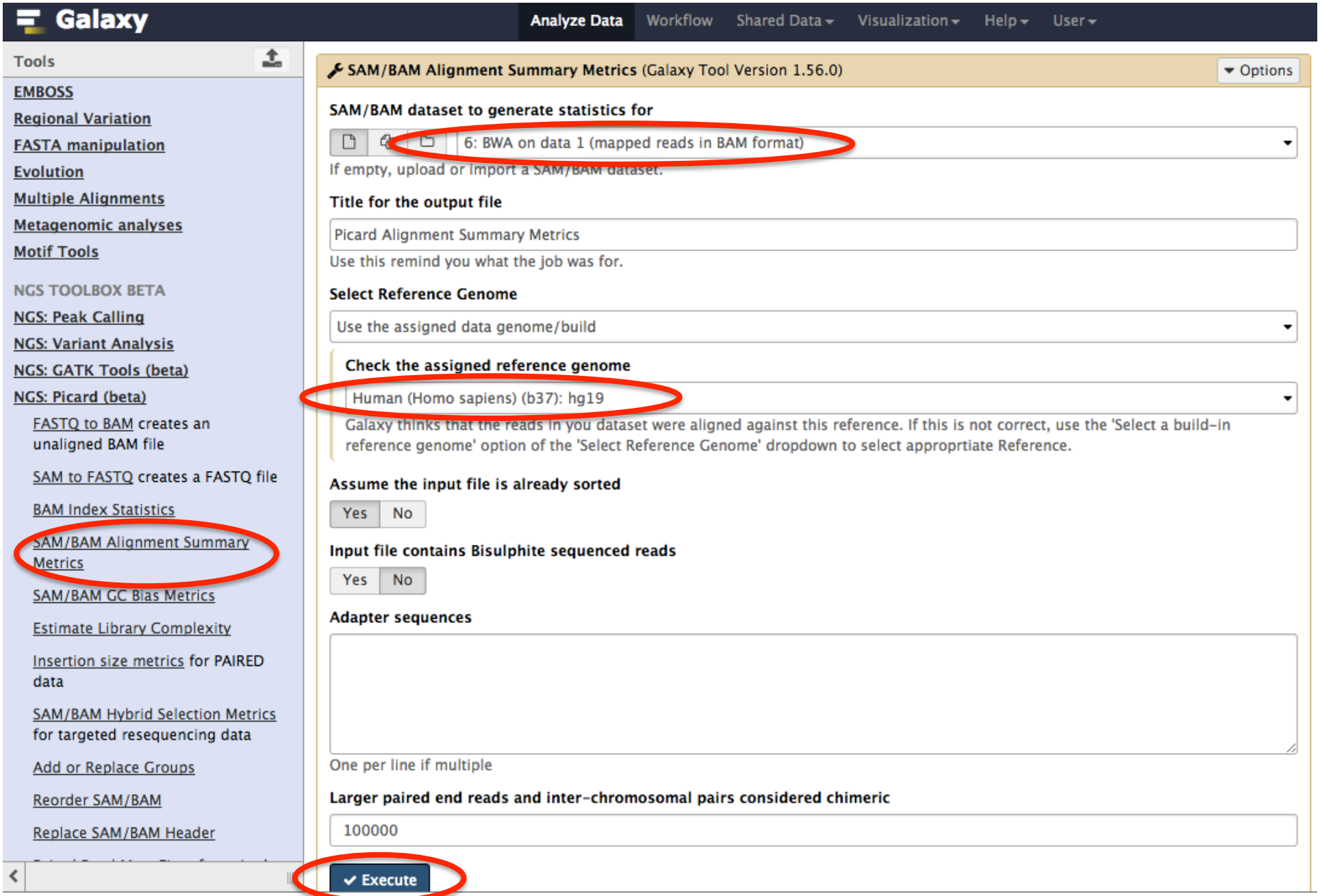

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
NGS: QC and manipulation
NGS: Mapping
NGS: RNA-seq
NGS: SAMtools
NGS: BAM Tools
NGS: Picard
NGS: VCF Manipulation
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Phenotype Association
snpEff
BEDTools
Genome Diversity
EMBOSS
Regional Variation
Per base sequence quality
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
History
search datasets
Unnamed history
3 shown
9.4 MB
3: FastQC on data 1: Raw Data
2: FastQC on data 1: Webpage
1: 100326_FC6107FAAXX-chr22_35-38M.fastq


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
Send Data
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
NGS: QC and manipulation
NGS: Mapping
Bowtie2 - map reads against reference genome
BWA - map short reads (< 100 bp) against reference genome
BWA-MEM - map medium and long reads (> 100 bp) against reference genome
Parse blast XML output
Megablast compare short reads against htgs, nt, and wgs databases
Map with BWA for Illumina
Map with Bowtie for Illumina
Lastz map short reads against reference sequence
NGS: RNA-seq
NGS: SAMtools
NGS: BAM Tools
NGS: Picard
NGS: VCF Manipulation
BWA - map short reads (< 100 bp) against reference genome (Galaxy Tool Version 0.1)
Options
Load reference genome from
Local cache
Using reference genome
Human (Homo sapiens) (b37): hg19
Select genome from the list
Select input type
Single fastq
Select between fastq and bam datasets and between paired and single end data
Select fastq dataset
1: 100326_FC6107FAAXX-chr22_35-38M.fastq
Specify dataset with single reads
Set advanced single end options?
Do not set
Provides additional controls
Set readgroups information?
Do not set
Specifying readgroup information can greatly simplify your downstream analyses by allowing combining multiple datasets. See help below for more details
Select analysis mode
1.Simple Illumina mode
Job Resource Parameters
Use default job resource parameters
Execute
History
search datasets
Unnamed history
3 shown
9.4 MB
3: FastQC on data 1: Raw Data
2: FastQC on data 1: Web page
1: 100326_FC6107FAAXX-chr22_35-38M.fastq


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
Send Data
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
NGS: QC and manipulation
NGS: Mapping
Bowtie2 – map reads against reference genome
BWA – map short reads (< 100 bp) against reference genome
BWA-MEM – map medium and long reads (> 100 bp) against reference genome
Parse blast XML output
Megablast compare short reads against htgs, nt, and wgs databases
Map with BWA for Illumina
Map with Bowtie for Illumina
Lastz map short reads against reference sequence
NGS: RNA-seq
NGS: SAMtools
NGS: BAM Tools
1 job has been successfully added to the queue – resulting in the following datasets:
4: BWA on data 1 (mapped reads in BAM format)
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
search datasets
Unnamed history
4 shown
11.4 MB
4: BWA on data 1 (mapped reads in BAM format)
1.9 MB
format: bam, database: hg19
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_a
display at UCSC main
display at Ensembl Current
display with IGV local Human hg19
Binary bam alignments file
3: FastQC on data 1: Raw Data
2: FastQC on data 1: Web page


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
testUnnamed (hg19)
chr22
37,790,496 - 37,831,311
790,000
37,800,000
37,810,000
37,820,000
37,830,
BWA on data 5 (mapped reads in BA..


IGV
Human hg19
chr22
chr22:37,790,496-37,831,311
Go
p13
p12
p11.2
p11.1
q11.1
q11.21
q11.22
q12.1
q12.2
q12.3
q13.1
q13.2
q13.31
q13.32
40 kb
37,800 kb
37,810 kb
37,820 kb
37,830 kb
Galaxy21-[BW...bam Coverage
[0 - 15]
Galaxy21-[BWA_on_data_1_and
_3_(mapped_reads_in_BAM_for
.bam
RefSeq Genes
ELFN2
6 tracks loaded
chr22:37,806,899
329M of 500M


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
EMBOSS
Regional Variation
FASTA manipulation
Evolution
Multiple Alignments
Metagenomic analyses
Motif Tools
NGS TOOLBOX BETA
NGS: Peak Calling
NGS: Variant Analysis
NGS: GATK Tools (beta)
NGS: Picard (beta)
FASTQ to BAM creates an unaligned BAM file
SAM to FASTQ creates a FASTQ file
BAM Index Statistics
SAM/BAM Alignment Summary Metrics
SAM/BAM GC Bias Metrics
Estimate Library Complexity
Insertion size metrics for PAIRED data
SAM/BAM Hybrid Selection Metrics for targeted resequencing data
Add or Replace Groups
Reorder SAM/BAM
Replace SAM/BAM Header
SAM/BAM Alignment Summary Metrics (Galaxy Tool Version 1.56.0)
Options
SAM/BAM dataset to generate statistics for
6: BWA on data 1 (mapped reads in BAM format)
If empty, upload or import a SAM/BAM dataset.
Title for the output file
Picard Alignment Summary Metrics
Use this remind you what the job was for.
Select Reference Genome
Use the assigned data genome/build
Check the assigned reference genome
Human (Homo sapiens) (b37): hg19
Galaxy thinks that the reads in you dataset were aligned against this reference. If this is not correct, use the 'Select a build-in reference genome' option of the 'Select Reference Genome' dropdown to select approprtiate Reference.
Assume the input file is already sorted
Yes
No
Input file contains Bisulphite sequenced reads
Yes
No
Adapter sequences
One per line if multiple
Larger paired end reads and inter-chromosomal pairs considered chimeric
100000
Execute

Tools

EMBOSS
Regional Variation
FASTA manipulation
Evolution
Multiple Alignments
Metagenomic analyses
Motif Tools

NGS TOOLBOX BETA
NGS: Peak Calling
NGS: Variant Analysis
NGS: GATK Tools (beta)
NGS: Picard (beta)

FASTQ to BAM creates an unaligned BAM file
SAM to FASTQ creates a FASTQ file
BAM Index Statistics
SAM/BAM Alignment Summary Metrics
SAM/BAM GC Bias Metrics
Estimate Library Complexity
Insertion size metrics for PAIRED data
SAM/BAM Hybrid Selection Metrics for targeted resequencing data
Add or Replace Groups
Reorder SAM/BAM
Replace SAM/BAM Header

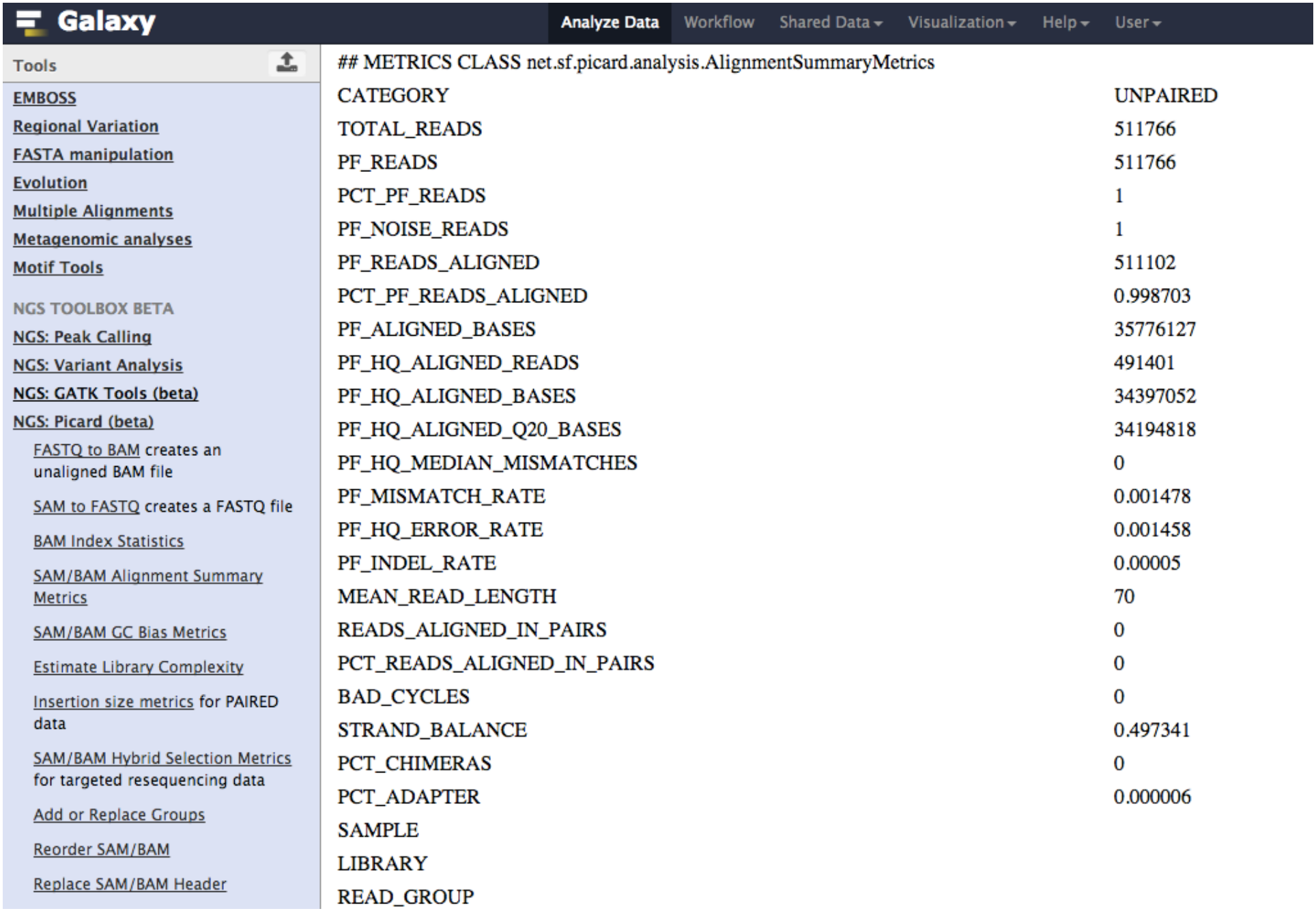
Option	Description
INPUT=File	SAM or BAM file Required.
OUTPUT=File	File to write insert size metrics to Required.
REFERENCE_SEQUENCE=File	Reference sequence file Required.
ASSUME_SORTED=Boolean	If true (default), unsorted SAM/BAM files will be considerd coordinate sorted
MAX_INSERT_SIZE=Integer	Paired end reads above this insert size will be considered chimeric along with inter-chromosomal pairs. Default value: 100000.
ADAPTER_SEQUENCE=String	This option may be specified 0 or more times.
IS_BISULFITE_SEQUENCED=Boolean	Whether the SAM or BAM file consists of bisulfite sequenced reads. Default value: false.
CREATE_MD5_FILE=Boolean	Whether to create an MD5 digest for any BAM files created.

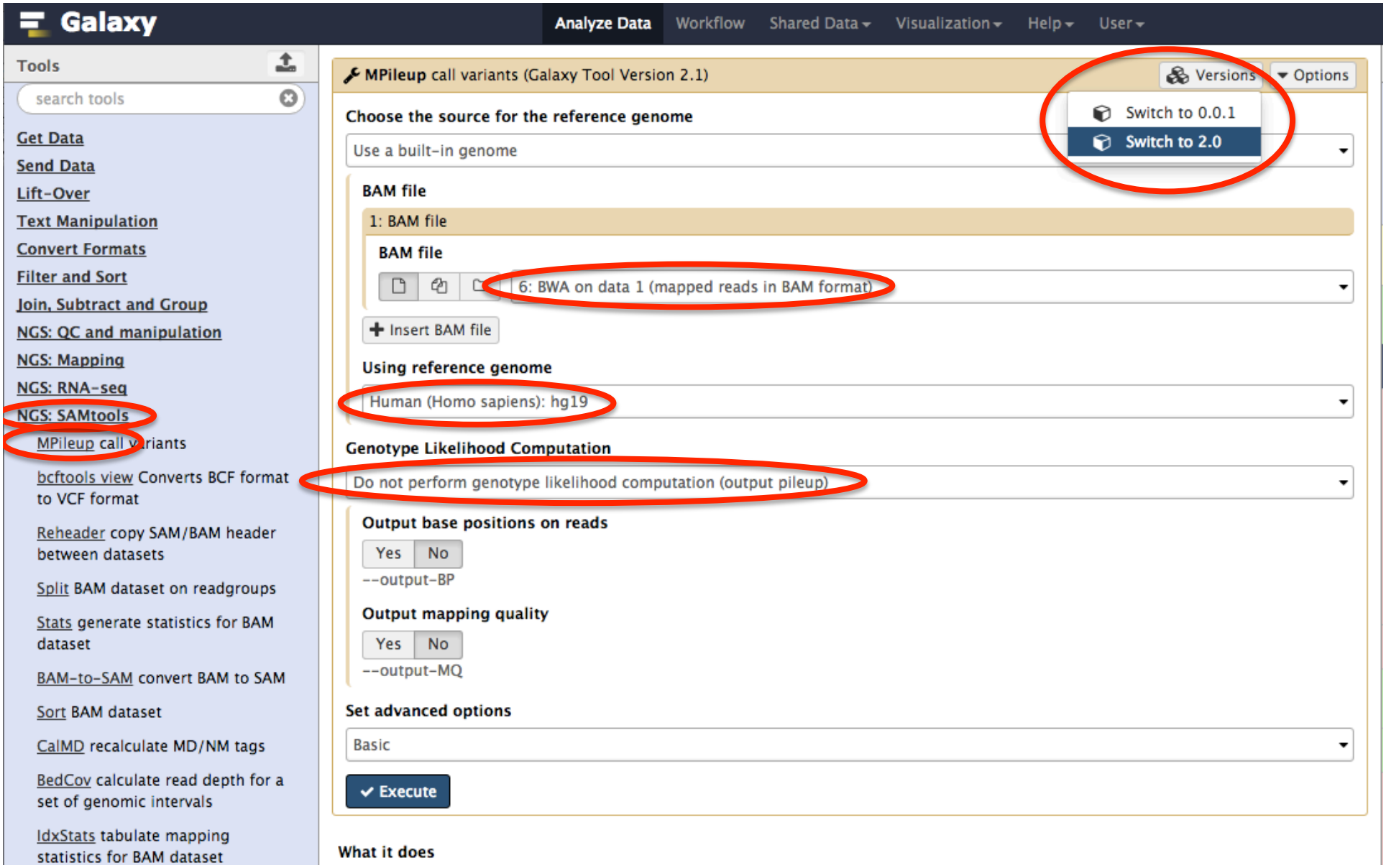
The output produced by the tool has the following columns:

1. CATEGORY: One of either UNPAIRED (for a fragment run), FIRST_OF_PAIR when metrics are for only the first read
2. TOTAL_READS: The total number of reads including all PF and non-PF reads. When CATEGORY equals PAIR this valu
3. PF_READS: The number of PF reads where PF is defined as passing Illumina's filter.
4. PCT_PF_READS: The percentage of reads that are PF (PF_READS / TOTAL_READS)
5. PF_NOISE_READS: The number of PF reads that are marked as noise reads. A noise read is one which is composed
6. PF_READS_ALIGNED: The number of PF reads that were aligned to the reference sequence. This includes reads tha
7. PCT_PF_READS_ALIGNED: The percentage of PF reads that aligned to the reference sequence. PF_READS_ALIGNED / P
8. PF_HQ_ALIGNED_READS: The number of PF reads that were aligned to the reference sequence with a mapping qualit
9. PF_HQ_ALIGNED_BASES: The number of bases aligned to the reference sequence in reads that were mapped at high
10. PF_HQ_ALIGNED_Q20_BASES: The subest of PF_HQ_ALIGNED_BASES where the base call quality was Q20 or higher.
11. PF_HQ_MEDIAN_MISMATCHES: The median number of mismatches versus the reference sequence in reads that were ali
12. PF_HQ_ERROR_RATE: The percentage of bases that mismatch the reference in PF HQ aligned reads.
13. MEAN_READ_LENGTH: The mean read length of the set of reads examined. When looking at the data for a single la
14. READS_ALIGNED_IN_PAIRS: The number of aligned reads who's mate pair was also aligned to the reference.
15. PCT_READS_ALIGNED_IN_PAIRS: The percentage of reads who's mate pair was also aligned to the reference. READS_
16. BAD_CYCLES: The number of instrument cycles in which 80% or more of base calls were no-calls.
17. STRAND_BALANCE: The number of PF reads aligned to the positive strand of the genome divided by the number of
18. PCT_CHIMERAS: The percentage of reads that map outside of a maximum insert size (usually 100kb) or that have
19. PCT_ADAPTER: The percentage of PF reads that are unaligned and match to a known adapter sequence right from t

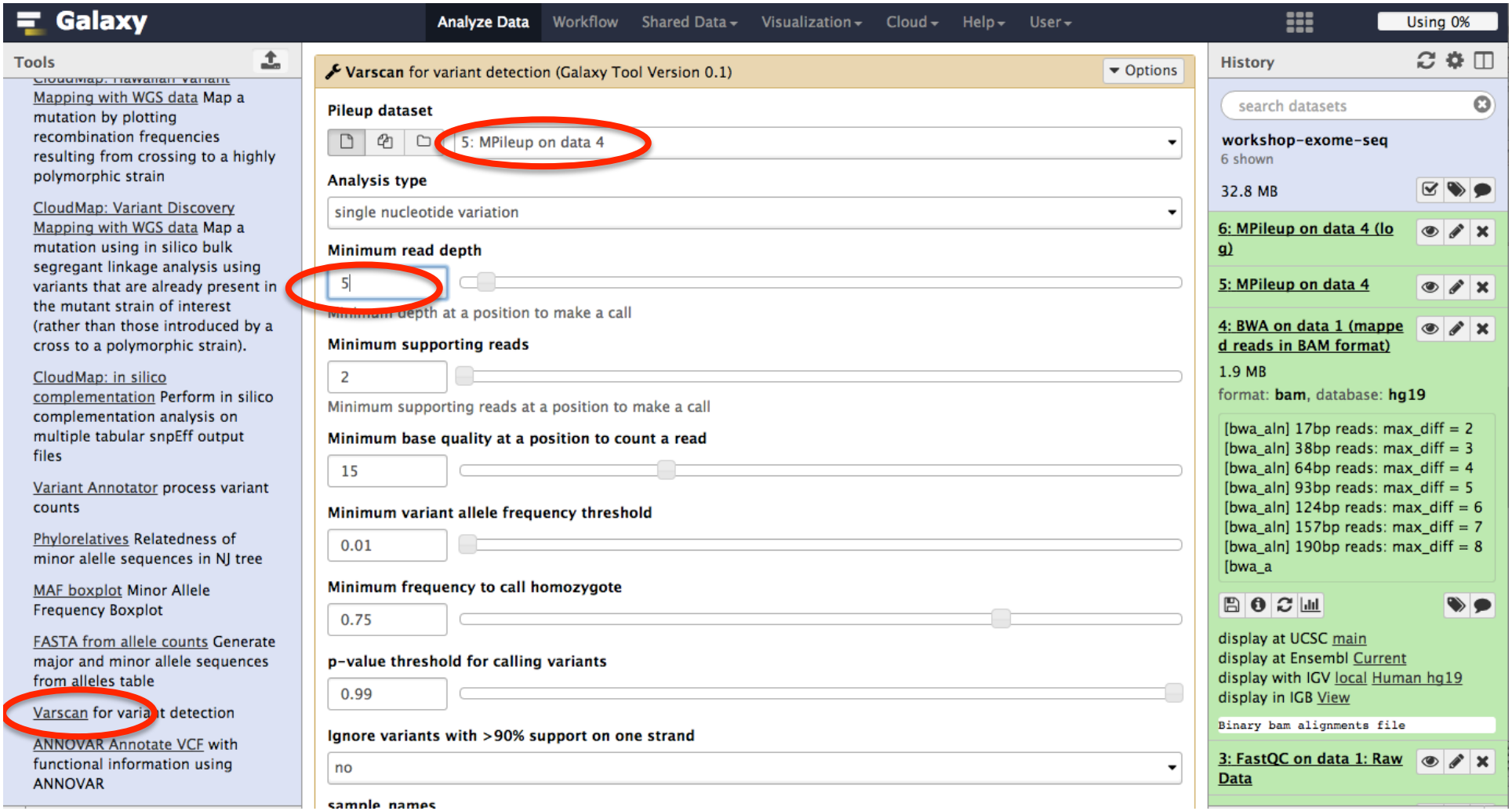
Warning on SAM/BAM quality

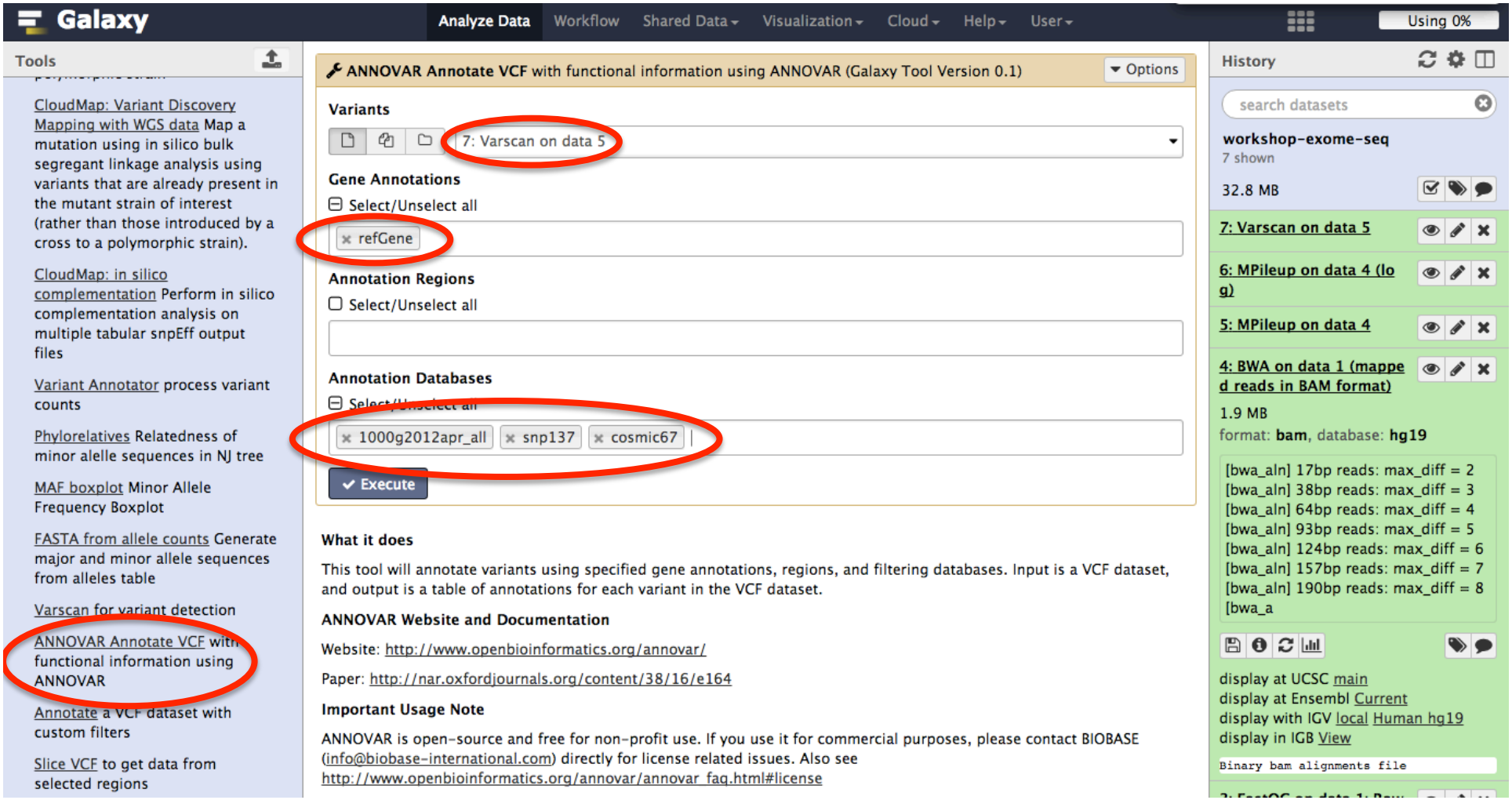
Many SAM/BAM files produced externally and uploaded to Galaxy do not fully conform to SAM/BAM specifications. Galaxy deals with this by using the **LENIENT** flag when it runs Picard, which allows reads to be discarded if they're empty or don't map. This appears to be the only way to deal with SAM/BAM that cannot be parsed.

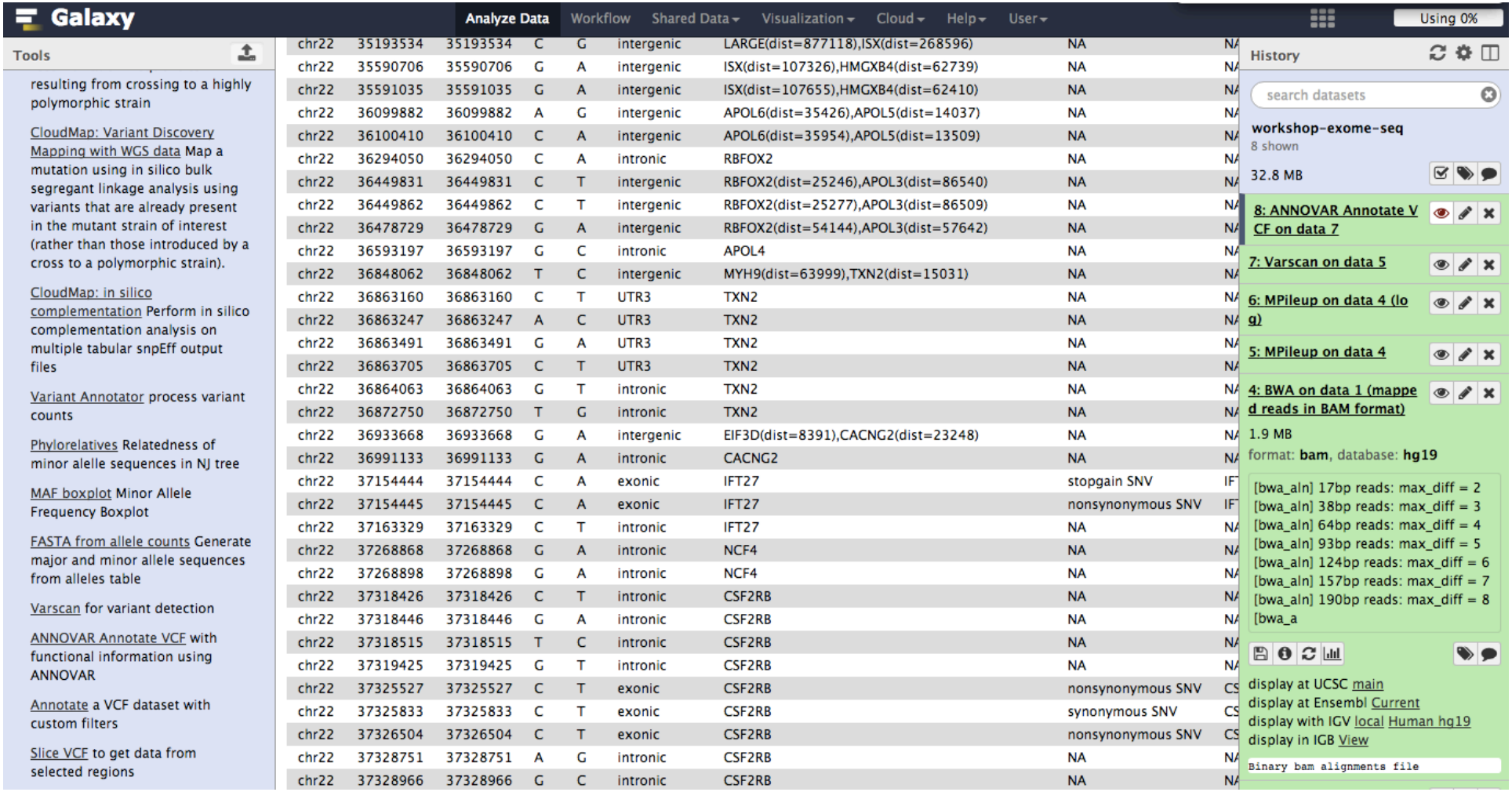

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
EMBOSS
Regional Variation
FASTA manipulation
Evolution
Multiple Alignments
Metagenomic analyses
Motif Tools
NGS TOOLBOX BETA
NGS: Peak Calling
NGS: Variant Analysis
NGS: GATK Tools (beta)
NGS: Picard (beta)
FASTQ to BAM creates an unaligned BAM file
SAM to FASTQ creates a FASTQ file
BAM Index Statistics
SAM/BAM Alignment Summary Metrics
SAM/BAM GC Bias Metrics
Estimate Library Complexity
Insertion size metrics for PAIRED data
SAM/BAM Hybrid Selection Metrics for targeted resequencing data
Add or Replace Groups
Reorder SAM/BAM
Replace SAM/BAM Header
METRICS CLASS net.sf.picard.analysis.AlignmentSummaryMetrics
CATEGORY UNPAIRED
TOTAL_READS 511766
PF_READS 511766
PCT_PF_READS 1
PF_NOISE_READS 1
PF_READS_ALIGNED 511102
PCT_PF_READS_ALIGNED 0.998703
PF_ALIGNED_BASES 35776127
PF_HQ_ALIGNED_READS 491401
PF_HQ_ALIGNED_BASES 34397052
PF_HQ_ALIGNED_Q20_BASES 34194818
PF_HQ_MEDIAN_MISMATCHES 0
PF_MISMATCH_RATE 0.001478
PF_HQ_ERROR_RATE 0.001458
PF_INDEL_RATE 0.00005
MEAN_READ_LENGTH 70
READS_ALIGNED_IN_PAIRS 0
PCT_READS_ALIGNED_IN_PAIRS 0
BAD_CYCLES 0
STRAND_BALANCE 0.497341
PCT_CHIMERAS 0
PCT_ADAPTER 0.000006
SAMPLE
LIBRARY
READ_GROUP

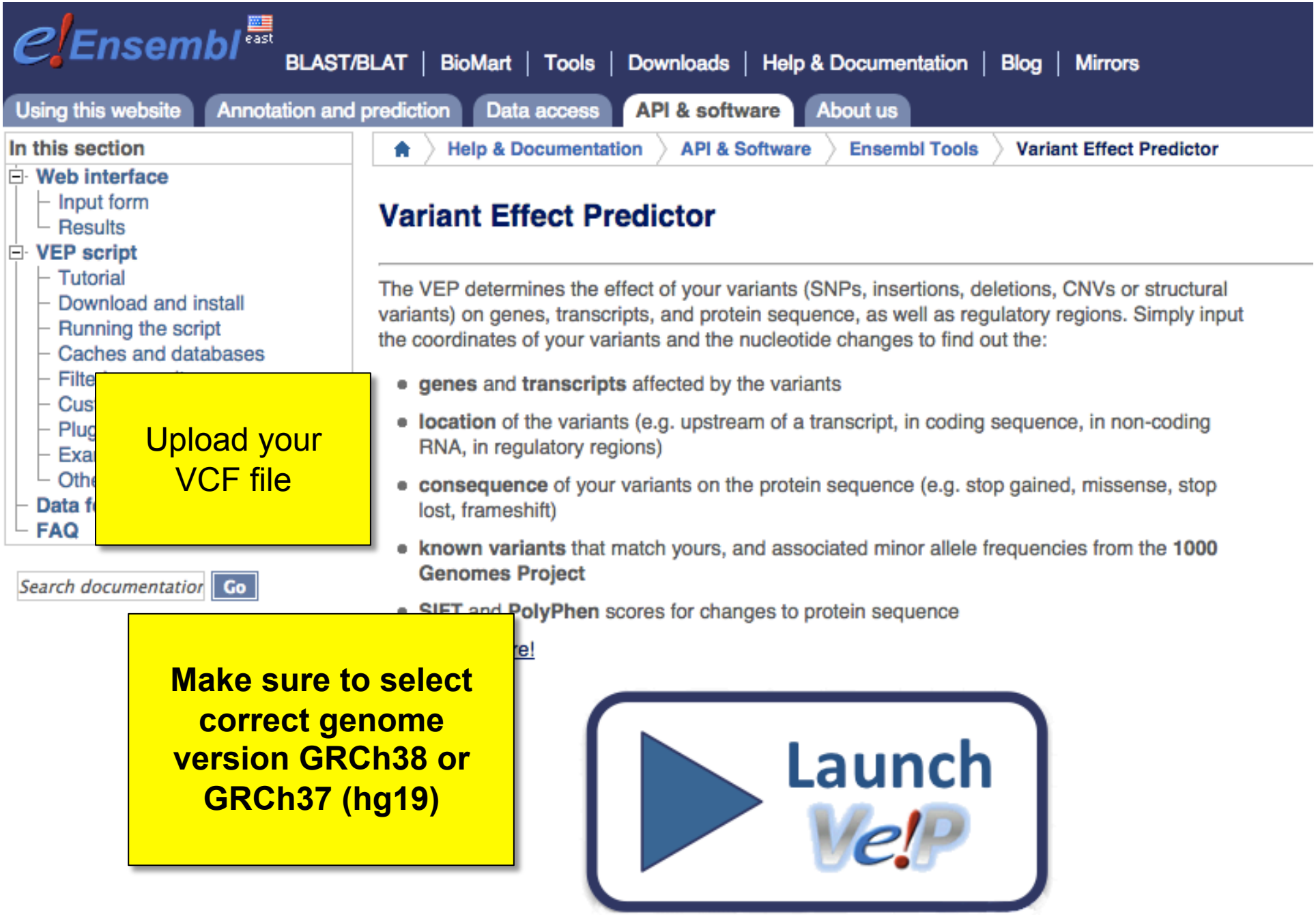

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
NGS: QC and manipulation
NGS: Mapping
NGS: RNA-seq
NGS: SAMtools
MPileup call variants
bcftools view Converts BCF format to VCF format
Reheader copy SAM/BAM header between datasets
Split BAM dataset on readgroups
Stats generate statistics for BAM dataset
BAM-to-SAM convert BAM to SAM
Sort BAM dataset
CalMD recalculate MD/NM tags
BedCov calculate read depth for a set of genomic intervals
IdxStats tabulate mapping statistics for BAM dataset
MPileup call variants (Galaxy Tool Version 2.1)
Versions
Options
Switch to 0.0.1
Switch to 2.0
Choose the source for the reference genome
Use a built-in genome
BAM file
1: BAM file
BAM file
6: BWA on data 1 (mapped reads in BAM format)
Insert BAM file
Using reference genome
Human (Homo sapiens): hg19
Genotype Likelihood Computation
Do not perform genotype likelihood computation (output pileup)
Output base positions on reads
Yes
No
--output-BP
Output mapping quality
Yes
No
--output-MQ
Set advanced options
Basic
Execute
What it does

NGS: Variant Analysis




Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
CloudMap: Variant Discovery Mapping with WGS data Map a mutation using in silico bulk segregant linkage analysis using variants that are already present in the mutant strain of interest (rather than those introduced by a cross to a polymorphic strain).
CloudMap: in silico complementation Perform in silico complementation analysis on multiple tabular snpEff output files
Variant Annotator process variant counts
Phylorelatives Relatedness of minor alelle sequences in NJ tree
MAF boxplot Minor Allele Frequency Boxplot
FASTA from allele counts Generate major and minor allele sequences from alleles table
Varscan for variant detection
ANNOVAR Annotate VCF with functional information using ANNOVAR
Annotate a VCF dataset with custom filters
Slice VCF to get data from selected regions
ANNOVAR Annotate VCF with functional information using ANNOVAR (Galaxy Tool Version 0.1)
Options
Variants
7: Varscan on data 5
Gene Annotations
Select/Unselect all
refGene
Annotation Regions
Select/Unselect all
Annotation Databases
Select/Unselect all
1000g2012apr_all
snp137
cosmic67
Execute
What it does
This tool will annotate variants using specified gene annotations, regions, and filtering databases. Input is a VCF dataset, and output is a table of annotations for each variant in the VCF dataset.
ANNOVAR Website and Documentation
Website: http://www.openbioinformatics.org/annovar/
Paper: http://nar.oxfordjournals.org/content/38/16/e164
Important Usage Note
ANNOVAR is open-source and free for non-profit use. If you use it for commercial purposes, please contact BIOBASE (info@biobase-international.com) directly for license related issues. Also see http://www.openbioinformatics.org/annovar/annovar_faq.html#license
History
search datasets
workshop-exome-seq
7 shown
32.8 MB
7: Varscan on data 5
6: MPileup on data 4 (log)
5: MPileup on data 4
4: BWA on data 1 (mapped reads in BAM format)
1.9 MB
format: bam, database: hg19
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_a
display at UCSC main
display at Ensembl Current
display with IGV local Human hg19
display in IGB View
Binary bam alignments file


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
resulting from crossing to a highly polymorphic strain
CloudMap: Variant Discovery Mapping with WGS data Map a mutation using in silico bulk segregant linkage analysis using variants that are already present in the mutant strain of interest (rather than those introduced by a cross to a polymorphic strain).
CloudMap: in silico complementation Perform in silico complementation analysis on multiple tabular snpEff output files
Variant Annotator process variant counts
Phylorelatives Relatedness of minor alelle sequences in NJ tree
MAF boxplot Minor Allele Frequency Boxplot
FASTA from allele counts Generate major and minor allele sequences from alleles table
Varscan for variant detection
ANNOVAR Annotate VCF with functional information using ANNOVAR
Annotate a VCF dataset with custom filters
Slice VCF to get data from selected regions
chr22 35193534 35193534 C G intergenic LARGE(dist=877118),ISX(dist=268596) NA
chr22 35590706 35590706 G A intergenic ISX(dist=107326),HMGXB4(dist=62739) NA
chr22 35591035 35591035 G A intergenic ISX(dist=107655),HMGXB4(dist=62410) NA
chr22 36099882 36099882 A G intergenic APOL6(dist=35426),APOL5(dist=14037) NA
chr22 36100410 36100410 C A intergenic APOL6(dist=35954),APOL5(dist=13509) NA
chr22 36294050 36294050 C A intronic RBFOX2 NA
chr22 36449831 36449831 C T intergenic RBFOX2(dist=25246),APOL3(dist=86540) NA
chr22 36449862 36449862 C T intergenic RBFOX2(dist=25277),APOL3(dist=86509) NA
chr22 36478729 36478729 G A intergenic RBFOX2(dist=54144),APOL3(dist=57642) NA
chr22 36593197 36593197 G C intronic APOL4 NA
chr22 36848062 36848062 T C intergenic MYH9(dist=63999),TXN2(dist=15031) NA
chr22 36863160 36863160 C T UTR3 TXN2 NA
chr22 36863247 36863247 A C UTR3 TXN2 NA
chr22 36863491 36863491 G A UTR3 TXN2 NA
chr22 36863705 36863705 C T UTR3 TXN2 NA
chr22 36864063 36864063 G T intronic TXN2 NA
chr22 36872750 36872750 T G intronic TXN2 NA
chr22 36933668 36933668 G A intergenic EIF3D(dist=8391),CACNG2(dist=23248) NA
chr22 36991133 36991133 G A intronic CACNG2 NA
chr22 37154444 37154444 C A exonic IFT27 stopgain SNV
chr22 37154445 37154445 C A exonic IFT27 nonsynonymous SNV
chr22 37163329 37163329 C T intronic IFT27 NA
chr22 37268868 37268868 G A intronic NCF4 NA
chr22 37268898 37268898 G A intronic NCF4 NA
chr22 37318426 37318426 C T intronic CSF2RB NA
chr22 37318446 37318446 G A intronic CSF2RB NA
chr22 37318515 37318515 T C intronic CSF2RB NA
chr22 37319425 37319425 G T intronic CSF2RB NA
chr22 37325527 37325527 C T exonic CSF2RB nonsynonymous SNV
chr22 37325833 37325833 C T exonic CSF2RB synonymous SNV
chr22 37326504 37326504 C T exonic CSF2RB nonsynonymous SNV
chr22 37328751 37328751 A G intronic CSF2RB NA
chr22 37328966 37328966 G C intronic CSF2RB NA
History
search datasets
workshop-exome-seq
8 shown
32.8 MB
8: ANNOVAR Annotate VCF on data 7
7: Varscan on data 5
6: MPileup on data 4 (log)
5: MPileup on data 4
4: BWA on data 1 (mapped reads in BAM format)
1.9 MB
format: bam, database: hg19
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_a
display at UCSC main
display at Ensembl Current
display with IGV local Human hg19
display in IGB View
Binary bam alignments file



http://www.ensembl.org/info/docs/tools/vep/index.html

IGV
Integrative Genomics Viewer

Wexner Medical Center

java.com: Java + You

http://www.java.com/en/

java web start launcher

Apple Yahoo! Google Maps YouTube Wikipedia News (13) Popular

Galaxy | Downloads | Integrat... | java.com: Java + You

Java™

Search

Download Help

JAVA + YOU, DOWNLOAD TODAY!

Free Java Download

» What is Java? » Do I have Java? » Need Help?

Feedback

About Java

JavaOne Java Developer

Alice Java + Alice

Greenfoot Java + Greenfoot

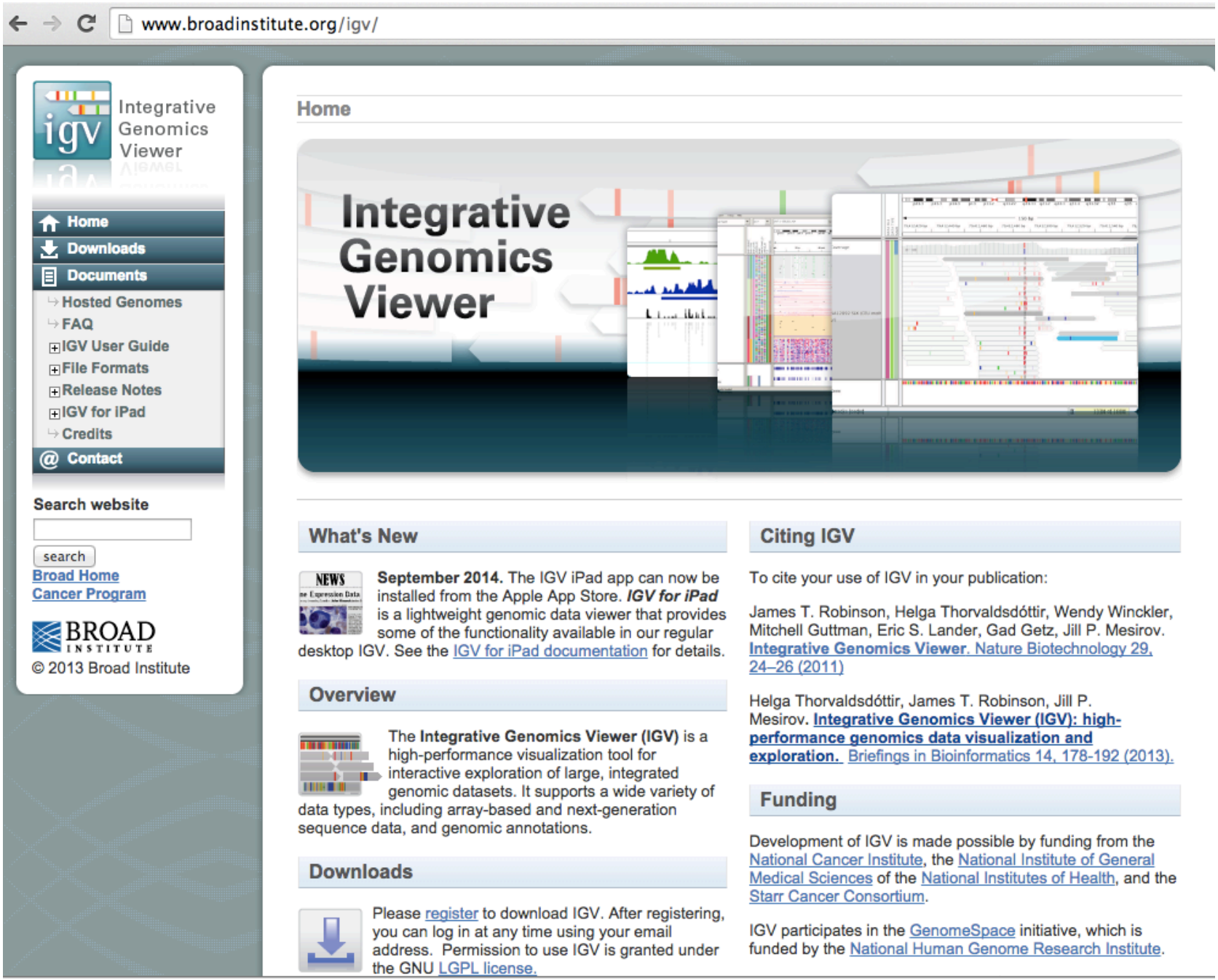
Java + BlueJ **Java + BlueJ**

ORACLE ACADEMY Oracle Academy

Java magazine Get it now for FREE! Java Magazine


Java Downloads for All Operating Systems
http://www.java.com/en/download/manual.jsp
java web start launcher
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (13)
Popular
Galaxy
Downloads | Integrat...
Java Downloads for ...
Search
Java
Download
Help
Available Operating Systems
» Windows
» Mac
» Linux
» Solaris
Help Resources
» Troubleshoot Java
Java 7
» Where can I get Java 7?
JDK
» Looking for the JDK?
Java Downloads for All Operating Systems
Recommended Version 8 Update 31
Select the file according to your operating system from the list below to get the latest Java for your computer.
> Remove Older Versions
> What is Java?
By downloading Java you acknowledge that you have read and accepted the terms of the end user license agreement
Windows
Which should I choose?
Windows Online (32-bit)
filesize: 624 KB
Instructions
Windows Offline (32-bit)
filesize: 29.0 MB
Instructions
Windows Offline (64-bit)
filesize: 89.1 MB
Instructions
After installing Java, you may need to restart your browser in order to enable Java in your browser.
If you use 32-bit and 64-bit browsers interchangeably, you will need to install both 32-bit and 64-bit Java in order to have the Java plug-in for both browsers. » FAQ about 64-bit Java for Windows
Feedback

http://www.broadinstitute.org/igv/



Integrative Genomics Viewer (IGV) (Version 2.3)

Install IGV

Options for installing and running IGV:

1. (Mac only) Download and run the Mac applcation; or
2. (All systems) Use the Java Web Start buttons (Mac users: see below for limitations); or
3. (All systems) Download the binary distribution and run IGV from the command line.

1. Mac Application

Download and unzip the Mac App archive, then double-click the IGV application to run it. The application can be moved to the "Applications" folder, or anywhere else. ***Note: This requires Java 7. Mac users with Java 6 (JRE 1.6) should use the binary distribution archive or the Java Web Start buttons below***.	Download Mac App

2. Java Web Start

The buttons below use Java Web Start (JWS) to install and launch IGV directly from our web site.

***Mac Users**: The Java Web Start option does not work for some users due to security settings. The recommended solution is to use the bundled Mac App from the link above. Alternatively you can try to work around this try right-clicking on the buttons and saving the "jnlp" file, then right-click on the "jnlp" file and select "Open With Java Web Start".

Chrome: Chrome does not automatically launch the Java Webstart files by default. Instead, the launch buttons below will download a "jnlp" file. This should appear in the lower left corner of the browser. Double-click the downloaded file to run, or if on a Mac right-click and select "Open With Java Web Start"

Windows users: To run with more than 1.2 GB of memory you must install 64-bit Java. ***Most Windows installs do not include 64-bit Java by default, even if the operating system is 64-bit***. Attempting to use the 2GB or greater launch options with 32-bit Java will result in the error *"could not create virtual machine"*.

Launch	Launch	Launch	Launch
Launch with 750 MB	Launch with 1.2 GB Maximum usable memory for Windows OS with 32-bit Java.	Launch with 2 GB Maximum usable memory for 32-bit MacOS.	Launch with 10 GB For large memory machines with 64-bit Java.

Acknowledgement

Selen Yilmaz, MS

Wexner Medical Center

Gulcin Ozer, PhD
340C Lincoln Tower
1800 Cannon Dr.

Gulcin.Ozer@osumc.edu

Wexner Medical Center