D756–D761 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 6 November 2019*
doi: 10.1093/nar/gkz1012

UCSC Genome Browser enters 20th year

Christopher M. Lee[1,*], Galt P. Barber[1], Jonathan Casper[1], Hiram Clawson[1], Mark Diekhans[1], Jairo Navarro Gonzalez[1], Angie S. Hinrichs[1], Brian T. Lee[1], Luis R. Nassar[1], Conner C. Powell[1], Brian J. Raney[1], Kate R. Rosenbloom[1], Daniel Schmelter[1], Matthew L. Speir[1], Ann S. Zweig[1], David Haussler[1,2], Maximilian Haeussler[1], Robert M. Kuhn[1] and W. James Kent[1]

[1]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA and [2]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Received September 11, 2019; Revised October 16, 2019; Editorial Decision October 17, 2019; Accepted October 25, 2019

ABSTRACT

The University of California Santa Cruz Genome Browser website (https://genome.ucsc.edu) enters its 20th year of providing high-quality genomics data visualization and genome annotations to the research community. In the past year, we have added a new option to our web BLAT tool that allows search against all genomes, a single-cell expression viewer (https://cells.ucsc.edu), a 'lollipop' plot display mode for high-density variation data, a RESTful API for data extraction and a custom-track backup feature. New datasets include Tabula Muris single-cell expression data, GeneHancer regulatory annotations, The Cancer Genome Atlas Pan-Cancer variants, Genome Reference Consortium Patch sequences, new ENCODE transcription factor binding site peaks and clusters, the Database of Genomic Variants Gold Standard Variants, Genomenon Mastermind variants and three new multi-species alignment tracks.

INTRODUCTION

The University of California Santa Cruz (UCSC) Genome Browser (1) is a web-based viewer for genome sequence data and annotations. The UCSC Genome Browser team has steadily added data and software features to the website since first coming online in July 2000, and currently hosts 206 assemblies from 105 species. We continue to add new species and assemblies, but, as most of our user base focuses on human and mouse, the vast majority of annotation data are focused on those two species. In addition to displaying and interacting with data hosted on UCSC servers, users can also upload and display their own data in the form of custom tracks (a single set of annotations) or track hubs (externally hosted collections of tracks). We encourage track hub creators to contact us about having their hubs added to our list of public track hubs.

A large part of our recent effort has been in making track hubs easier to use and configure, as well as encouraging researchers to submit their track hubs to the list of 'public hubs' (accessible from the blue top navigation bar as 'My Data → Track Hubs'), so that other researchers can more easily discover and access their data. A track hub consists of data files and configuration files that organize the data into annotation tracks, adding labels and descriptions and optionally grouping related datasets into composite tracks. A guide to the various options available for creating a track hub is available in the Track Hub Database Definition page (https://genome.ucsc.edu/goldenPath/help/trackDb/trackDbHub.html).

Along with displaying and hosting data, our site also provides a number of tools for interacting with, comparing and downloading those data. Some, such as the Table Browser (2) and BLAT (3), are available for direct use on the web. Other tools are available for download—we provide a directory of over 200 utilities ('Downloads → Utilities') that can be used to download and examine data files, or convert between formats. We also maintain a public MySQL server ('Downloads → MySQL Access') for accessing our annotation data, and with this update have added a RESTful API ('Downloads → REST API'), both of which can be queried to download and access the data we host. The API is described in more detail below.

We continue to support local 'mirroring' of the Genome Browser software and data ('Mirrors → Mirroring Instructions'). The Genome Browser in a Box (GBiB) (4) is a fully pre-configured virtual machine that is easy to set up and run. For enterprise solutions, the Genome Browser in the Cloud (GBiC) is a tool that automates the setup of a mirror instance on a bare-metal or cloud server. Both provide dynamic access to the data on UCSC's servers as well as the option of downloading data locally for offline use. This last

[*]To whom correspondence should be addressed. Email: chmalee@ucsc.edu

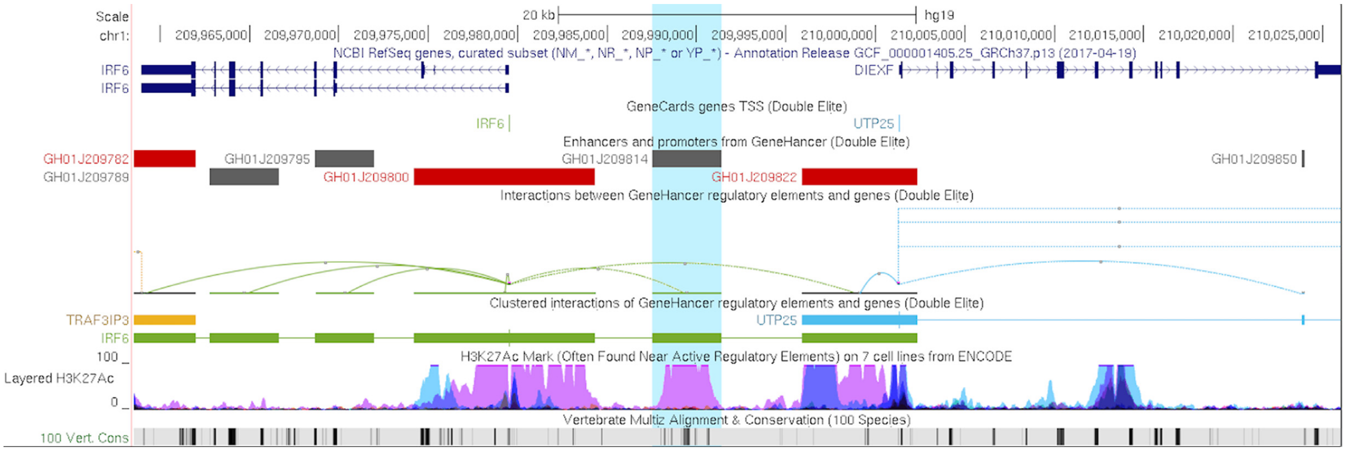

Scale
chr1:
20 kb
hg19
209,965,000
209,970,000
209,975,000
209,980,000
209,985,000
209,990,000
209,995,000
210,000,000
210,005,000
210,010,000
210,015,000
210,020,000
210,025,000
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001405.25_GRCh37.p13 (2017-04-19)
IRF6
IRF6
DIEXF
GeneCards genes TSS (Double Elite)
IRF6
UTP25
Enhancers and promoters from GeneHancer (Double Elite)
GH01J209782
GH01J209795
GH01J209814
GH01J209850
GH01J209789
GH01J209800
GH01J209822
Interactions between GeneHancer regulatory elements and genes (Double Elite)
Clustered interactions of GeneHancer regulatory elements and genes (Double Elite)
TRAF3IP3
UTP25
IRF6
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
Vertebrate Multiz Alignment & Conservation (100 Species)
100 Vert. Cons

Figure 1. The GeneHancer track set and the interaction track format with hills instead of the default valleys. The GeneHancer track relates enhancer and promoters to their interactions with nearby genes. In this display, a highlighted GH01J209814 enhancer is associated with the gene IRF6 (interferon regulatory factor 6) located about 10 kb upstream.

option may be useful for navigating clinical or proprietary data or using the Browser behind a restrictive firewall.

ANNOTATIONS AND VISUALIZATIONS

Over the past year, we have continued to update and add tracks and assemblies. A full list of the new data is available in Supplementary Table S1 (assemblies) and Supplementary Table S2 (tracks), but a few items deserve special mention due to their accompanying visualization features: the TCGA Pan-Cancer dataset and the accompanying new 'lollipop' display mode, the GeneHancer database of regulatory elements and the accompanying improvements to the recently introduced interact format, and the Tabula Muris single-cell RNA expression dataset and accompanying single-cell browser. We have also expanded our support of human genome assembly patches released by the Genome Reference Consortium, and have added annotations from many tracks onto these new sequences to better support researchers studying human variation.

GeneHancer and interact track format improvements

Last year (5) we introduced a new 'interact' track format that draws connecting arcs between pairs of genomic regions. This year brings the release of the first native track utilizing this format, the GeneHancer track from the GeneCards group (6). GeneHancer is a database of human regulatory elements (enhancers and promoters) and their inferred target genes. These elements were synthesized from over 1 million regulatory elements obtained from seven genome-wide databases: ENCODE (7), Ensembl (8), FANTOM5 (9), VISTA (10), dbSUPER (11), EPDNew (12) and UCNEbase (13). The highly filtered 'double elite' dataset is displayed by default, and the full dataset can be enabled from the track configuration page. This track is available on the GRCh37/hg19 and GRCh38/hg38 assemblies. In support of this track, a number of enhancements were made to the interact track format, including an inverted view (hills instead of valleys, Figure 1), 'pack' and 'squish' visibility modes, and a cluster mode, which combines directional interactions with the same target or source into a single item.

TCGA Pan-Cancer data and the new 'lollipop' plot display

The new TCGA Pan-Cancer mutations track for the hg38 assembly presents somatic variants called from 33 whole-genome sequencing projects, as part of The Cancer Genome Atlas. Variant calls were downloaded from the Genomic Data Commons TCGA portal (14) and transformed into one track per cancer type as well as one merged track. As shown in Figure 2, this track features a new 'lollipop' plot display type, where the position of a lollipop circle corresponds to the genomic coordinate of the variant, and the height of the lollipop stem reflects the number of samples in which the variant was found. Lollipop circles are shaded according to their score relative to the other variants in the current viewing window. This mode allows variants common to many samples (or high-scoring in the general case) to stand out in a sea of potentially spurious cancer mutations. The TCGA Pan-Cancer track also supports filtering of items by sample count, gender and TCGA Project ID.

Tabula Muris and the single-cell browser

The GRCm38/mm10 assembly features a new native track, Tabula Muris (15). The Tabula Muris transcriptome compendium contains single-cell sequencing data from roughly 100 000 cells and 20 different organs/tissues. The data have been organized into three separate tracks: an overview bar graph track of median expression per cell type, an individual wiggle graph of the number of reads at a particular base pair, and a track showing an analysis of splice junctions.

Along with this dataset, we have created a new tool for visualization of single-cell data, termed the 'single-cell browser' (Figure 3). The single-cell browser is available from the landing page through the menu ('Projects → UCSC Cell Browser'), or directly via https://cells.ucsc.edu. The single-cell browser presents an interactive display from the data produced by a standard Cellranger, Seurat (16) or SCANPY (17) run (Figure 3, left panel). This tool supports clicking or

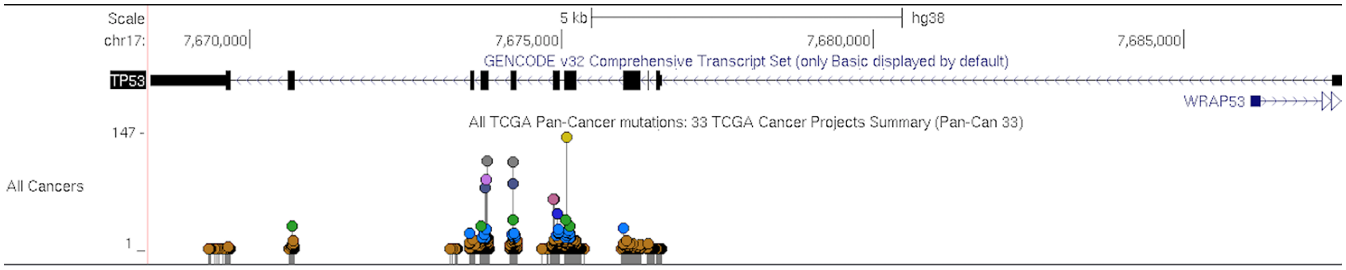
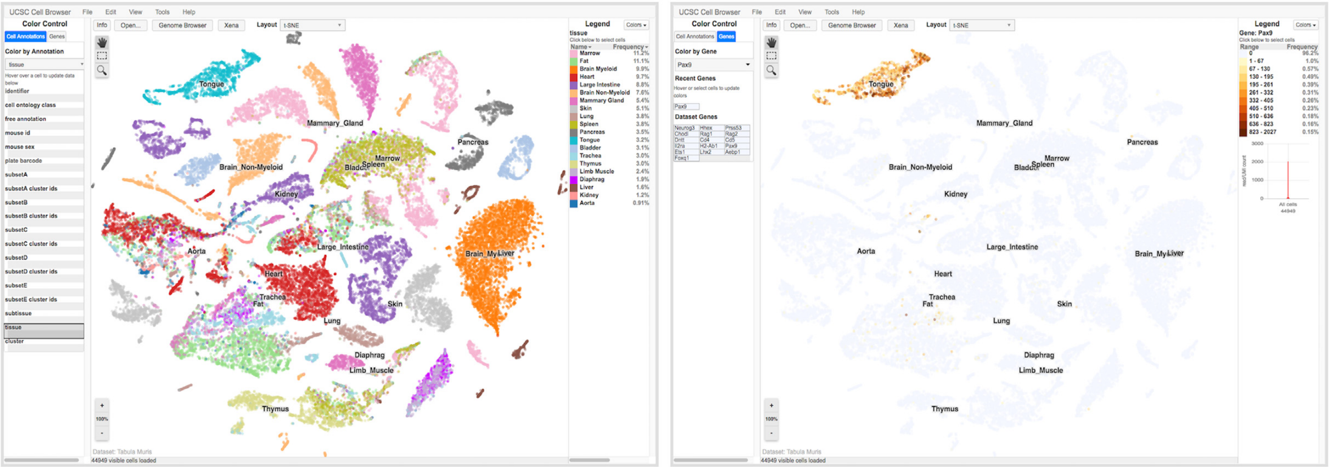

Scale
5 kb
hg38
chr17:
7,670,000
7,675,000
7,680,000
7,685,000
GENCODE v32 Comprehensive Transcript Set (only Basic displayed by default)
TP53
WRAP53
All TCGA Pan-Cancer mutations: 33 TCGA Cancer Projects Summary (Pan-Can 33)
147 -
All Cancers
1 _

Figure 2. The TCGA Pan-Cancer track at the TP53 locus. The lollipop display showcases SNPs found in cancers that were not present in normal tissue. Lollipops are shaded according to their score relative to the other SNPs in the current view. The tall yellow and gray lollipops stand out to reveal hotspots indicated across ∼120 people and multiple cancers.



Color by Annotation

tissue

Hover over a cell to update data below

identifier
cell ontology class
free annotation
mouse id
mouse sex
plate barcode
subsetA
subsetA cluster ids
subsetB
subsetB cluster ids
subsetC
subsetC cluster ids
subsetD
subsetD cluster ids
subsetE
subsetE cluster ids
subtissue
tissue
cluster

tissue

Click below to select cells

Name	Frequency
Marrow	11.2%
Fat	11.1%
Brain Myeloid	9.9%
Heart	9.7%
Large Intestine	8.8%
Brain Non-Myeloid	7.6%
Mammary Gland	5.4%
Skin	5.1%
Lung	3.8%
Spleen	3.8%
Pancreas	3.5%
Tongue	3.2%
Bladder	3.1%
Trachea	3.0%
Thymus	3.0%
Limb Muscle	2.4%
Diaphrag	1.9%
Liver	1.6%
Kidney	1.2%
Aorta	0.91%

Dataset: Tabula Muris
44949 visible cells loaded

Color by Gene

Pax9

Recent Genes

Hover or select cells to update colors

Pax9

Dataset Genes

Neurog3	Hhex	Prss53
Chodl	Rag1	Rag2
Dntt	Cd4	Cd5
Il2ra	H2-Ab1	Pax9
Ets1	Lhx2	Aebp1
Foxq1		

Gene: Pax9

Click below to select cells

Range	Frequency
0	96.2%
1 - 67	1.0%
67 - 130	0.57%
130 - 195	0.49%
195 - 261	0.39%
261 - 332	0.31%
332 - 405	0.26%
405 - 510	0.23%
510 - 636	0.18%
636 - 823	0.16%
823 - 2027	0.15%

Dataset: Tabula Muris
44949 visible cells loaded

Figure 3. The single-cell browser. The left-hand panel shows the default view upon loading the Tabula Muris dataset in the cell browser. Cells are shaded according to which tissue cluster they belong to, and each tissue cluster is labeled accordingly. Hovering the mouse pointer over an individual cell populates the fields on the left-hand panel with the metadata associated with each cell. The right-hand panel shows the result of filtering all of the cells for Pax9 gene expression.

hovering on cells to obtain metadata about each cell, coloring cells by gene (Figure 3, right panel), and clicking to show cluster-specific marker genes. In addition to the Tabula Muris data, there are 16 other single-cell experiments available to explore on the single-cell browser website.

Human patch sequences

With the Genome Reference Consortium (GRC) deciding to indefinitely postpone the next coordinate-changing human assembly GRCh39 (https://www.ncbi.nlm.nih.gov/grc/human), we have committed to modifying how we handle the patch sequences released by the GRC. Previously, when a new patch sequence release was announced, tracks were created against the target assembly (e.g. hg38) that allowed these sequences to be viewed in their genomic context. Now, for the hg38 assembly only, we are incorporating the patch sequences directly into our assembly database with remapped or recomputed annotations. This has numerous consequences: BLAT results now include matches on _alt and _fix sequences, position searches may lead to new sequences in addition to initial assembly release sequences, and some tracks include annotations on new sequences. In the case of certain third-party datasets (like GENCODE or RefSeq), we now directly incorporate their existing patch sequence annotations. Annotations computed by us, such as RepeatMasker, CpG Islands, and Augustus, have been expanded to include the patch sequences. We have also remapped annotations from main chromosomes to patch sequences for several popular tracks such as GTEx Genes and ENCODE Regulation. Tracks that include annotations on patch sequences in GRCh38.p12 are marked with a 'P12' icon next to the track label in the track controls section. To avoid breaking pre-existing pipelines, the sequence files available in the hg38 bigZips download directory (e.g. https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz) will remain unchanged. Instead, new subdirectories have been added in bigZips: 'p12' for files that include patch sequences from the GRCh38.p12 release, 'initial' to explicitly provide unchanged files, and 'latest' to provide stable links to files from the most recent patch release to support pipelines that do include patches. In addition to the extended data, the multiregion feature ('View → Multi-Region') now supports autocomplete search for fix and patch sequences to facilitate the viewing of alternate sequences in their genomic context. More information about our incorporation of sequence patches can be found at http://genome.ucsc.edu/blog/patches/.

ENCODE transcription factor binding site peaks and clusters

The ENCODE Transcription Factor Binding Site Peaks and Clusters track sets show regions of transcription factor binding derived from a large collection of ChIP-seq experiments performed by the ENCODE project between February 2011 and November 2018 (ENCODE phases 2 and 3). The tracks are organized into two types: the Peaks set contains the underlying ChIP-seq peaks and can be optionally filtered by cell type or transcription factor, while the Clusters track provides a summary display of occupancy regions for each cluster. This track is available for hg38 and hg19.

DGV Gold Standard Variants

The Database of Genomic Variants (18) is a curated set of large structural variants that appear in healthy individuals. The Gold Standard set is created by starting with all of the variants in DGV that are found in at least two different studies and at least two different samples, filtering out low-quality variants and then merging what is left according to a 50% minimum overlap. The variants are then combined into a 'boxplot'-like record, where the highest quality variant in the cluster forms the 'inner box', and the whiskers define the maximal coordinates of the copy number variants. This track is available for hg19.

Genomenon Mastermind

The Mastermind search engine from Genomenon (https://www.genomenon.com/mastermind) mines full-text publications for disease–gene–variant associations. We have created a track on hg19 that shows variants that have been indexed by the search engine, with links to Genomenon for more details such as references in which the variant was reported and associated with disease.

Gene set updates

This year we have continued to add Ensembl gene sets and GENCODE track sets as they became available, most recently adding Ensembl v95 for most assemblies, GENCODE v31 for hg19 and hg38, and M22 for mm10. Along with updating the GENCODE track sets, we have also updated the default gene set for hg38 to GENCODE v29, and switched to using the ENST* transcripts as the default transcript names instead of the previous uc* identifiers. Finally, there are two new tracks in the NCBI RefSeq track set for the hg38 and hg19 assemblies: a 'Diffs' track that shows differences between the transcript sequences provided by RefSeq and the reference genome, and a track for transcripts that are selected as the most clinically relevant by the Human Gene Mutation Database (HGMD) curators (19). For more information about the different gene sets available, see the following page: https://genome.ucsc.edu/FAQ/FAQgenes.html.

Multiple alignments for chicken, *Drosophila* and *Caenorhabditis elegans*

This year we have created three new multiple alignment tracks: one for chicken (GRCg6a/galGal6) that features 76 other vertebrates, one for *Drosophila melanogaster* (BGDP Release 6/dm6) that features 123 other insects and one for *C. elegans* (WBcel235/ce11) that features 134 other worms.

New public track hubs

A number of groups have created new hubs to share with the general UCSC Genome Browser user community, notably UniBind (20), DASHR v2.0 (21), human p53 Binding and Expression Resource (22), Primate x4 NeuroDiff and Human CRISPRa (23) and the Vertebrate Genomes Project assemblies (24).

SOFTWARE AND TOOL IMPROVEMENTS

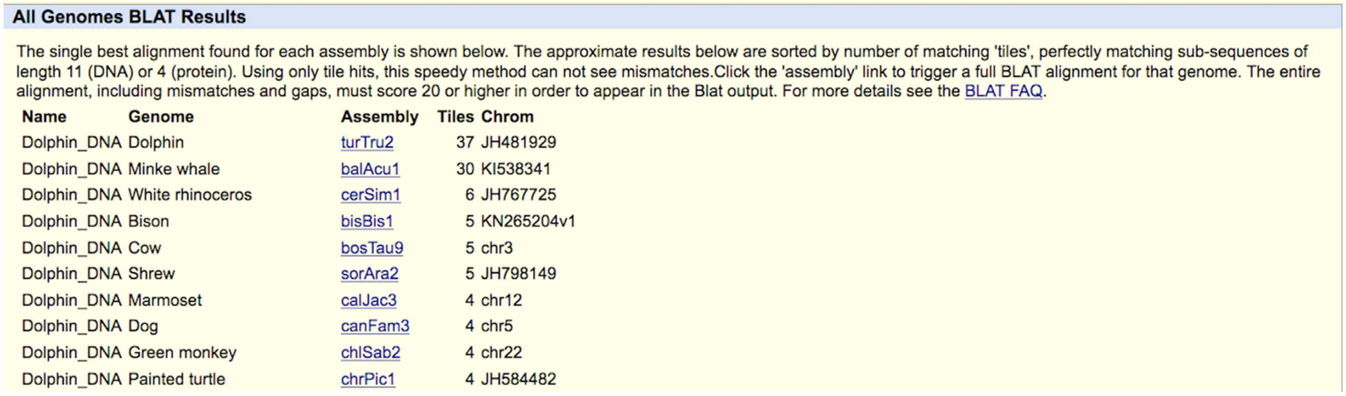
BLAT against all genomes

The BLAST-like Alignment Tool (BLAT) was one of the first features of the Browser, and remains one of the most popular tools on the site. Normally used to discover the exon–intron structure of a piece of mRNA or to find sequence similarity across a particular genome sequence, BLAT ('Tools → Blat') can now, with a single click, search across all of the default organisms available in the UCSC database as well as any attached assembly hubs. Each of over 100 genomes is searched in parallel, resulting in a quick return of matched 'tiles', sorted by the organism with the most matches (Figure 4), as a heuristic indicator of which species are more likely to have true alignments. Clicking into each organism causes a full BLAT search to execute for that organism and leads to the standard BLAT results section.

Instead of carrying out a full BLAT search against each genome, the BLAT All feature only checks for matching 11-mers (or 4-mers in the case of protein queries), which are called 'tiles'. Mismatches and gaps are not accounted for, which substantially speeds up the search while still allowing users to get an overall idea of sequence homology. Results can then be explored further with the standard BLAT interface. For more information on the BLAT All feature, we have updated the BLAT documentation available here: https://genome.ucsc.edu/FAQ/FAQblat.html#blat9.

REST API

Previously, users who wanted to download track or sequence data would need to grab the dataset in its entirety from our downloads server, or use the Table Browser's graphical interface. Unfortunately, scripting against these resources was never practical for dynamic queries like getting track data for >1000 regions or accessing UCSC data via R/Python/Perl or another programming environment. To better support those use cases, we have created a RESTful API that returns JSON-formatted data for a variety of data retrieval queries. The API is accessed via the URL https://api.genome.ucsc.edu and has two primary functions: listing available datasets and downloading data from said datasets. Notable endpoints include /list/tracks/ for listing all of the tracks for a given database or hub, and /getData/sequence/ and /getData/track/ for obtaining genome sequence and annotation data, respectively. For a user guide including the full list of endpoints, please see the documentation at https://genome.ucsc.edu/goldenPath/help/api.html.



All Genomes BLAT Results

The single best alignment found for each assembly is shown below. The approximate results below are sorted by number of matching 'tiles', perfectly matching sub-sequences of length 11 (DNA) or 4 (protein). Using only tile hits, this speedy method can not see mismatches.Click the 'assembly' link to trigger a full BLAT alignment for that genome. The entire alignment, including mismatches and gaps, must score 20 or higher in order to appear in the Blat output. For more details see the BLAT FAQ.

Name	Genome	Assembly	Tiles	Chrom
Dolphin_DNA	Dolphin	turTru2	37	JH481929
Dolphin_DNA	Minke whale	balAcu1	30	KI538341
Dolphin_DNA	White rhinoceros	cerSim1	6	JH767725
Dolphin_DNA	Bison	bisBis1	5	KN265204v1
Dolphin_DNA	Cow	bosTau9	5	chr3
Dolphin_DNA	Shrew	sorAra2	5	JH798149
Dolphin_DNA	Marmoset	calJac3	4	chr12
Dolphin_DNA	Dog	canFam3	4	chr5
Dolphin_DNA	Green monkey	chlSab2	4	chr22
Dolphin_DNA	Painted turtle	chrPic1	4	JH584482

Figure 4. The BLAT All search results for a chunk of dolphin DNA show that the majority of potential hits apply only to the source turTru2 (*Tursiops truncatus*) dolphin assembly, but also suggest interesting matches to minke whale.

Session Custom-Track backups

We have also made improvements to the Sessions ('My Data → My Sessions') tool. Primarily, users can now download all of the custom tracks currently loaded in their browsing session by visiting the Sessions page and using the 'Save Custom Tracks' button in the 'Save Settings' section. The custom track data will all be combined into one archive file for download, and then, if needed, immediately reloaded back into the Genome Browser. We hope this will be convenient for users who experiment with large numbers of custom tracks before finally deciding on a session to share for a publication or send to a colleague. For more information on this backup feature, see the 'Creating a session' section of the Sessions help page (https://genome.ucsc.edu/goldenPath/help/hgSessionHelp.html). In addition to the backup feature, sessions can now be loaded via significantly shorter links that are more informative for publications and for emailing to colleagues. The new URL format is https://genome.ucsc.edu/s/username/sessionname.

OUTREACH AND CONTACT INFORMATION

In the last year, the Genome Browser's training team provided 30 seminars and workshops to help users take advantage of the latest features. We have now presented trainings in 27 countries and 34 states and Canadian provinces (http://bit.ly/ucscMap). Outreach is supported by a blog (http://genome.ucsc.edu/blog) and by updates to the training documentation (https://genome.ucsc.edu/training/) with links to videos and in-depth descriptions of new Browser features. The training page also includes information on how to submit a request for a workshop and where future workshops may be scheduled.

General contact information for the UCSC Genome Browser can be found at https://genome.ucsc.edu/contacts.html, including information on accessing our email support list and an archive of previously answered mailing list questions. UCSC also maintains mirrors in Germany and Japan with the gracious assistance of Bielefeld University, Germany, and RIKEN, Japan. Those sites can be found at https://genome-euro.ucsc.edu and https://genome-asia.ucsc.edu.

FUTURE PLANS

The UCSC Genome Browser team has several major goals for the coming year. We plan to continue development of the single-cell browser and to add more support for single-cell sequencing data. We also plan to develop two new visualizations, a heatmap display for Hi-C chromatin interactions and a phased trio display for personal genomics data. On the data side, we plan to add annotations onto the mm10 and hg19 patch sequences and to support visualization of data from the NCBI Sequence Read Archive.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank all the users for continuing to use their resource and suggesting improvements. The authors also thank all the data providers responsible for the annotations that appear in the Browser; it would be an empty shell without their contributions. Finally, the authors thank the support staff who keep the computers running and the lights on.

FUNDING

National Human Genome Research Institute [5U41HG002371 to G.P.B., H.C., J.N.G., M.H., D.H., A.S.H., W.J.K., R.M.K., B.T.L., C.M.L., L.R.N, C.C.P., B.J.R., K.R.R., D.S., M.L.S., A.S.Z.; 2U41HG007234 to M.D., D.H.]; National Institute of Mental Health, National Institutes of Health [5U01MH114825 to W.J.K.]; California Institute for Regenerative Medicine [GC1R-06673-C to J.C., M.H., W.J.K., B.T.L., M.L.S., A.S.Z.]; Howard Hughes Medical Institute [090100 to D.H.]; Regeneron [CBJCHBCAABAA-nBGXMxFrE0WG7VvHOtm8RAzv674WMNt to H.C., J.N.G., R.M.K.]; St. Baldricks Foundation [427053 to D.H.]; Silicon Valley Community Foundation [2018-182809 to J.C., M.H., W.J.K., M.L.S.; 2017-171519 to G.P.B., J.C., M.D., M.H., W.J.K., B.T.L., M.L.S., A.S.Z.; 2018-194093 to M.D.]. Funding for open access charge: National Human Genome Research Institute [5U41HG002371].

Conflict of interest statement. G.P.B., J.C., H.C., M.D., J.N.G., M.H., D.H., A.S.H., W.J.K., R.M.K., B.T.L., C.M.L., B.J.R., K.R.R., M.L.S., and A.S.Z. receive royalties from the sale of UCSC Genome Browser source code, LiftOver, GBiB and GBiC licenses to commercial entities. W.J.K. owns Kent Informatics.

REFERENCES

1. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
2. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
3. Kent,W.J. (2002) BLAT—the BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.
4. Haeussler,M., Raney,B.J., Hinrichs,A.S., Clawson,H., Zweig,A.S., Karolchik,D., Casper,J., Speir,M.L., Haussler,D. and Kent,W.J. (2015) Navigating protected genomics data with UCSC Genome Browser in a Box. *Bioinformatics*, **31**, 764–766.
5. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
6. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M., Lancet,D. and Cohen,D. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**, bax028.
7. ENCODE Project Consortium, Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
8. Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T. and Flicek,P.R. (2015) The Ensembl regulatory build. *Genome Biol.*, **16**, 56.
9. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
10. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
11. Khan,A. and Zhang,X., (2016) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, **44**, D164–D171.
12. Dreos,R., Ambrosini,G., Cavin Périer,R. and Bucher,P. (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.*, **41**, D157–D164.
13. Dimitrieva,S. and Bucher,P. (2013) UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, **41**, 101–109.
14. Grossman,R.L., Heath,A.P., Ferretti,V., Varmus,H.E., Lowy,D.R., Kibbe,W.A. and Staudt,L.M. (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.
15. TabulaMuris Consortium, Overall coordination, Logistical coordination, Organcollection and processing, Library preparation and sequencing, Computational dataanalysis, Cell type annotation, Writing group, Supplemental text writing group and Principal investigators (2018) Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. *Nature*, **201**, 367–372.
16. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
17. Wolf,F.A., Angerer,P. and Theis,F. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
18. MacDonald,J.R., Ziman,R., Yuen,R.K., Feuk,L. and Scherer,S.W. (2013) The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
19. Stenson,P.D., Mort,M., Ball,E.V., Evans,K., Hayden,M., Heywood,S., Hussain,M., Phillips,A.D. and Cooper,D.N. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.
20. Gheorghe,M., Sandve,G.K., Khan,A., Chèneby,J., Ballester,B. and Mathelier,A. (2019) A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Res.*, **47**, e21.
21. Kuksa,P.P., Amlie-Wolf,A., Katanic,Z., Valladares,O., Wang,L. and Leung,Y.Y. (2019) DASHR v2.0: database of small human noncoding RNAs in human tissues and cell types. *Bioinformatics*, **35**, 1033–1039.
22. Nguyen,T.T., Grimm,S.A., Bushel,P.R., Li,J., Li,Y., Bennett,B.D., Lavender,C.A., Ward,J.M., Fargo,D.C., Anderson,C.W. *et al.* (2018) Revealing a human p53 universe. *Nucleic Acids Res.*, **46**, 8153–8167.
23. Field,A.R., Jacobs,F.M.J., Fiddes,I.T., Phillips,A.P.R., Reyes-Ortiz,A.M., LaMontagne,E., Whitehead,L., Meng,V., Rosenkrantz,J.L., Olsen,M. *et al.* (2019) Structurally conserved primate lncRNAs are transiently expressed during human cortical differentiation and influence cell type specific genes. *Stem Cell Rep.*, **12**, 245–257.
24. Koepfli,K.P., Paten,B. and Genome 10K Community of ScientistsGenome 10K Community of Scientists and O'Brien,S.J. (2015) The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.*, **3**, 57–111.