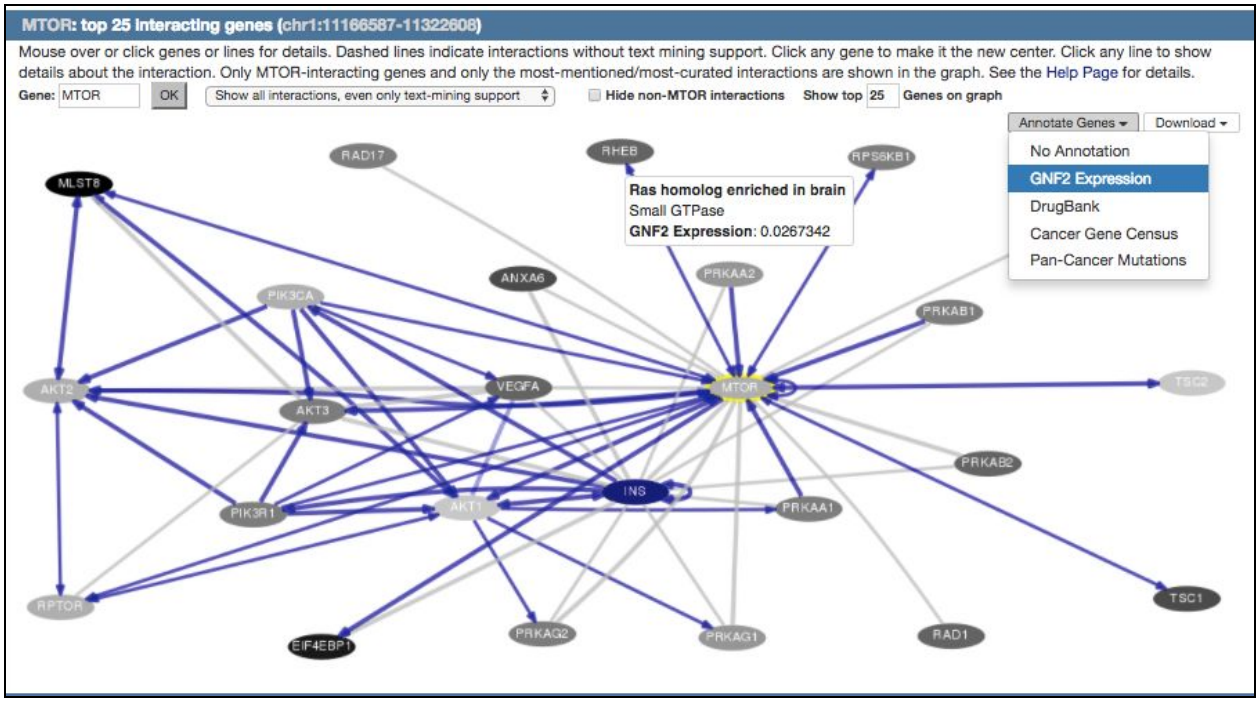
Aim 1. Develop, maintain, and extend software for web-based display and command-line-driven analysis of genomics resources.

Major new features added this year

Pathway and Gene Interactions tool and track
The new "Gene Interactions" tool combines data from 23 curated interaction and pathway databases as well as interactions text-mined from over 20 million PubMed abstracts, to provide an interactive graphical view that interoperates with the Genome Browser database and display. It comes with a companion Pathway and Gene Interactions annotation track on hg38 and hg19.

The text mining data were generated in collaboration with the Microsoft Research Project Hanover Team using Literome machine-reading. Literome is a natural-language processing system that analyzes sentences and for this application, extracts names of proteins and the type of interaction. More information about this tool is available on the help page: http://genome.ucsc.edu/goldenPath/help/hgGeneGraph.html.

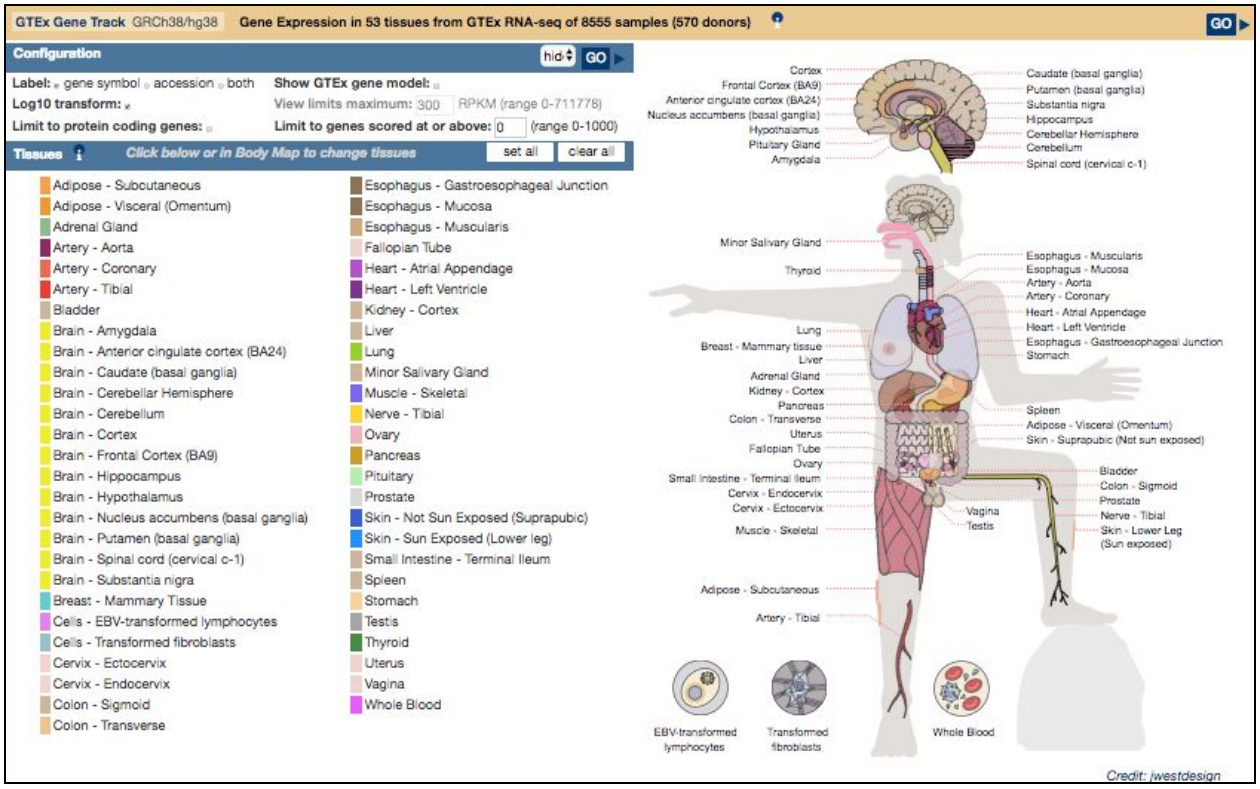
Figure B.2.1. Pathway and Gene Interactions graph for a selected gene marked in yellow (MTOR). Lines to other genes are drawn when there is support from both text mining and databases. Gene interactions can be filtered or colored on annotation details such as expression levels. Hovering over a gene shows gene details, while hovering over a connecting line displays interaction details. Clicking a line brings you to a page with external links to detailed supporting data. Navigate directly to this location in the Genome Browser: https://goo.gl/uPjin2.



Tissue body map

During this reporting period, we added several data sets from the Genotype-Tissue Expression (GTEx) project and developed a new interface to provide easier customization of GTEx tracks.The GTEx tracks reflect analysis of samples from hundreds of donors in 53 different human tissue types. The new UCSC browser interface makes tissue selection for GTEx tracks much easier by providing an anatomy graphic with tissue labels grouped near the corresponding region of the body. This interface, which we call the Tissue Body Map, allows users to identify GTEx-sampled tissues in anatomical context and select tissues from an interactive anatomy graphic or from an alphabetical tissue list of checkboxes color-coded according to GTEx project conventions.
For more details about the full set of GTEx data tracks, hubs, and tools, see **Section B3** of this RPPR.

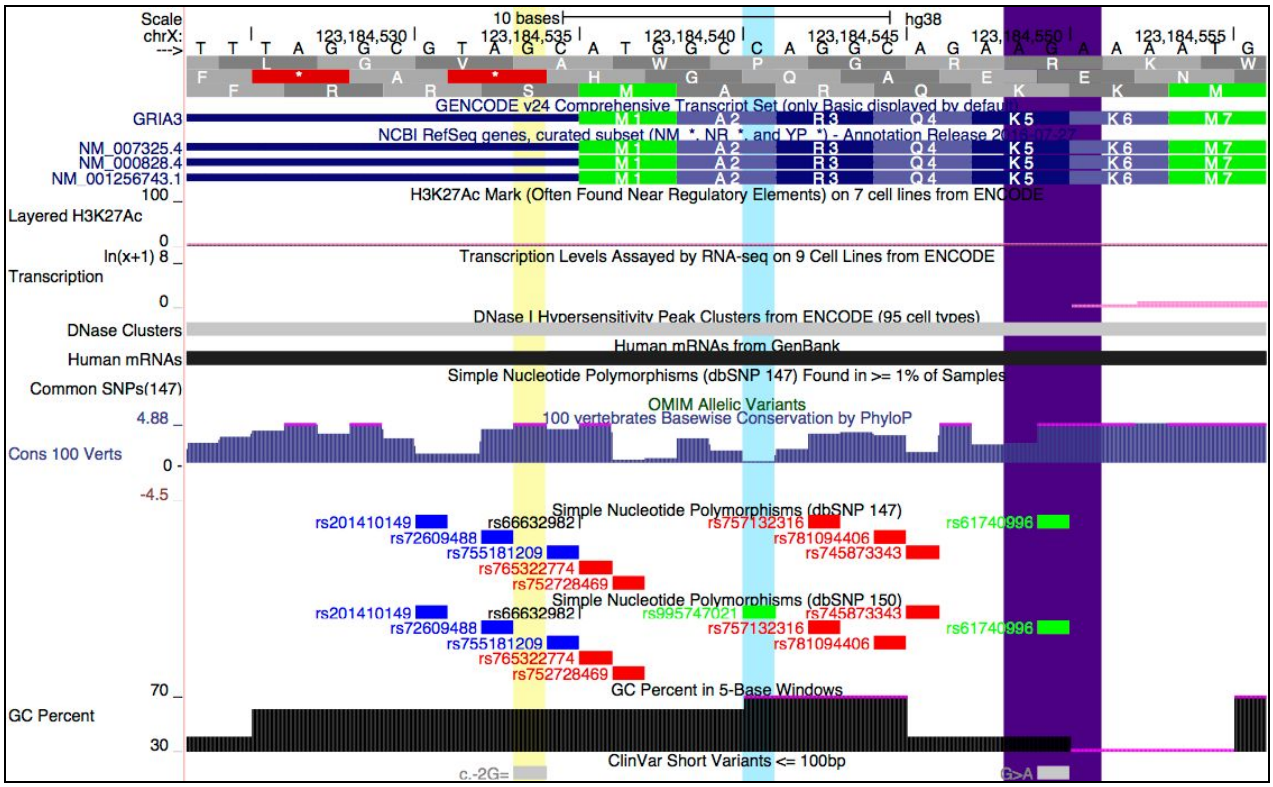
Figure B.2.2. Graphic-based "Tissue Body Map" configuration page for the GTEx Gene Expression track. Anatomical context can be used to select tissues alongside an alphabetical tissue list. Hovering over a tissue in the list highlights the corresponding anatomical region in the body map image. To show or hide tissues in the track, selections can be made from the tissue list or by clicking tissues directly in the map. Navigate directly to this location in the Genome Browser: https://goo.gl/2i88L5.



Multiple colored highlights
Three years ago, to much acclaim, we released a highlighting tool that allows users to pick a region of their view and apply a vertical highlight. During this reporting period, we added support for multiple vertical highlights and now allows user-selected colors. These highlights persist until explicitly removed or changed by the user.

The feature is activated by clicking and dragging near the top of the Browser image to select a region, by or pressing the Shift key while dragging within the image. A pop-up menu offers options to zoom in or highlight the region with a custom color selected by the user. The “Single Highlight” button removes all other active highlights before creating a new one, while the “Add Highlight” button adds the new highlight to those already present. Highlights can be cleared individually by right-clicking on them and selecting “Remove highlight” from the context menu, or by selecting “Remove all highlights” from the View menu at the top of the page.

Figure B.2.3. Multiple vertical highlights of different colors. The blue highlight results from right-clicking on an item (in this case, a SNP that is found in dbSNP track v150 but not in v147) then choosing highlight from the menu. The purple highlight results from dragging-and-dropping over a location of interest in the ClinVar track. And finally, searching for an HGVS term (in this case NM_000828.4:c.-2G>A) automatically draws a tan highlight on the resulting nucleotide. Navigate directly to this location in the Genome Browser: https://goo.gl/pFYMbR.



1.a. Increase website interactivity

- Last year one of our biggest accomplishments was implementation of ‘multiple regions’ mode. This new Browser display configuration allows users to “slice” their track-viewing experience into exon-only, gene-only, or user-defined BED coordinates. For human assemblies hg17 and later, the multi-region view also supports the replacement of a section of the reference genome with an alternate haplotype. This year, we extended this functionality by making it much easier for users to input their BED coordinates. Instead of hosting them on a web-accessible location, they can simply paste them directly into the tool.
- In a manner that was completely transparent to the user, we swapped out the backend of the UCSC Genes tracks on several human and mouse assemblies without changing any of the data itself. Now, instead of hosting the data in tables, it resides in a file of type bigGenePred. This makes the UCSC Genes track much faster for users. Access to data download via the Table Browser is unchanged.
- Upgrades to searching:
    - Traditionally our “position/search box” has been reserved for coordinate positions, gene symbols and the like, however we recently added a “DNA quick search”. If a user pastes DNA sequence into the search box, it will run our BLAT tool and show matches in the genome.
    - We added support for searching for alternate names/aliases for chromosomes.
    - We also added support for HGVS nomenclature: see **Section 1.b.** below.

- Updated the style of the static (non-software) pages to make them more consistent and modern.
- Added more keyboard shortcuts to the main Browser display page, including:
    - Typing 1, 2, 3, 4, 5, or 6 zooms the image to 50bp, 500bp, 5kbp, 50kbp, 500kbp, or 5Mbp respectively.
    - Type “h m” to mark highlights, and “h c” to clear highlights “p s” for Public Sessions.
- In an effort to make things more intuitive and easier to find, we continue to fine-tune our menu bars. This year we made the following additions to the menu bar:
    - Direct link to the mouse strains hub portal page
    - Link to the new Gene Interactions tool
    - Keyboard shortcut notations to many menu items
    - Link to Public Sessions page
    - Link to Track Search tool
    - Link to mirroring instructions
- Optimized the existing multiple alignment file for the 100-species conservation track on hg38 so that the 8 default species load much more quickly.
- Turned on codon numbering by default.
- Allowed users to set visibilities via the URL without the hub’s decoration ID for Track and Assembly Hubs. e.g. `trackName=pack instead of hub_1234_trackName=pack`.
- Added disconnect button to allow quickly removing Track Hubs from the Browser, eliminating the need to navigate back to the hubs page.
- Added a new section to many track description pages entitled: Data Access. This provides a direct link to the underlying data for that track.
- In an attempt to strike a balance between allowing fast access to our online users and supporting reasonable access to programmatic users we added bottleneck servers to more of our CGIs. Although we clearly state our usage restrictions for programmatic access to the website, sometimes users still abuse these restrictions. Bottlenecks slowly throttle abusive use of our site thereby preserving speed for the rest of our users. In other areas we loosened some existing bottleneck penalties for our more popular CGIs that are possible to query quickly by hand.

1.b. Adapt to new types of data

- Results from the BLAT sequence alignment tool on the UCSC Genome Browser have historically been stored for only 48 hours. After that period, the search results were deleted even if they were part of a saved session. An update to the BLAT alignment tool now includes a button that offers to transform the BLAT alignment results into a custom track. This custom track will display the results in exactly the same format as standard BLAT results, but will not be subject to the 48 hour expiration time. Instead, it will be treated like any other custom track on the Browser. If it is saved as part of a Browser Session, the BLAT result custom track will persist as long as the saved session exists. The custom track can also be downloaded for archival purposes using the UCSC Table Browser or used in other Browser tools such as the Data Integrator.
- Newly-supported remote data types in support of Track Hubs: bigBarChart, bigChain, and bigGenePred.
- Added support to the Custom Track tool to load a CRAM file directly from a URL.
- Added support to our VAI tool to support the updates to the VCF specification (v4.2) from the samtools group (https://github.com/samtools/hts-specs).
- Human Genome Variation Society (HGVS) nomenclature search enabled:

    - Added support for HGVS terms to all of our position/search boxes. Currently recognized HGVS variant classes include “p.” (protein), “g.” (genomic), “c.” (coding), and “n.” (non-coding). See http://genome.ucsc.edu/goldenPath/help/query.html#HGVS.
    - Added support for HGVS terms to both the input and output of the VAI tool. Options exist for generating all of the same term types that are currently recognized by the Browser. A command-line version of this function is also available in the tool vcfToHgvs.
- Changed restrictions on the maximum chromosome name length from 32 characters to 255. This allows support for Assembly Hubs such as ones in the human leukocyte antigen (HLA) typing project with lengthy accession references.
- Added support for the cytoband ideogram on Assembly Hubs, allowing quicker navigation within chromosomes.

1.c. Adapt to higher volumes of data

- Track Hubs
    - Improved the search functionality on the Track Hub Portal. Previously, a search on the public hubs listing would return a high-level terse list of hubs that matched the search terms. That list has been expanded to list all matching assemblies and tracks within each hub, arranged hierarchically. All assemblies and tracks in the search results can be right-clicked for a link to connect directly to the hub and either begin immediately viewing the assembly or configuring the track. In addition, the body of text searched for each hub has been expanded to include the text of the description pages for hubs, their assemblies (for Assembly Hubs), and their tracks.
    - The Browser is now able to connect to Track Hub data stored on Amazon AWS HIPAA compliant storage (https://aws.amazon.com/compliance/hipaacompliance/) and other similar cloud storage solutions.
    - Improved UDC redirects without sacrificing speed.
    - See **Aim 3** of this report to learn how we are working with CyVerse to provide seamless hosting for users’ Track and Assembly Hub data.
- Public Sessions
    - As a companion to the Session Gallery which was added last grant year, we have added a new Public Sessions page (http://genome.ucsc.edu/cgi-bin/hgPublicSessions) where users can share their UCSC Genome Browser snapshots with other Browser users. The Public Sessions page collects sessions that users have elected to share publicly. Sessions on this page can be filtered based on assembly, name, or a phrase from the description. Sessions can also be sorted based on their popularity or creation date.
    - The release of the new Public Sessions page also marks a change in our session expiration policy. We no longer expire sessions and associated custom tracks four months after their last use; we now keep them indefinitely.
- Genome Browser in the Cloud (GBiC)
    - Users can install user-apps via the GBiC `addTools` option.
    - Several minor bug fixes including MySQL compatibility, systemd (system daemon) support, and non-assembly database mirroring.
- Based on usage statistics, made some changes to the default set of tracks for human and mouse assemblies.
- We periodically change the default location to showcase a locus of interest. This year we changed the default location for hg38 from the ABO locus to the MTOR locus.

- Migrated the GenBank metadata to a central database table removing duplicated tables from most assembly databases.

Enhance the security of uploaded data

- In order to prevent Cross-site scripting (XSS) JavaScript injection into Genome Browser web pages generated by our CGIs, we have implemented Content Security Policy (CSP2) protocol. CSP2 is supported by and compatible with all Internet Browsers. This technique allows us to identify for the Browser the JavaScript that was legitimately generated by our site. JavaScript injected by a hackers is not allowed to run.
- Changes to logging into the Browser:
    - Allowed Apache-based (external) authentication over secure http**s**. This gives more flexibility to mirror sites; they can choose their own login method instead of defaulting to our login system. Use http**s** for Browser login unless explicitly disabled. Make the Browser login cookies more secure.
- Added support for http and http**s** and most of ftp for proxy servers. Since many modern browsers are encouraging or requiring users to use http**s**, we support it in the Browser, but we do not require it at this time.
- We aim to indefinitely maintain users' custom tracks that are saved in Sessions. However, we are not chartered to be a data storage service and so we warn users to save backup copies of their data. This year, for the first time we are aware of, we lost some user custom track data. After investigation, we pinpointed and fixed the problem (software error triggered in unusual circumstances).

1.e. Package command-line and web-services applications for broader use

- The command-line version of Variant Annotation Integrator (VAI) (also added to user-apps): `vai.pl`. This program is similar to our web-based VAI. This program is intended to be run on a Genome Browser in a box (GBiB) or server hosting a mirror of the Genome Browser. `vai.pl` forms an interface to the VAI program running on a GBiB or mirror (so private data stays local) and allows for bypassing the variant limit imposed by web-based VAI. This is especially useful for clinical users whose privacy restrictions may prevent the uploading of a patient's variant data to the UCSC Genome Browser.
- We package the updated source code and distribute it, along with command-line tools, every 3 weeks. This includes the Genome Browser in a Box (GBiB) and Genome Browser in the Cloud (GBiC) products free for non-profit academic research.
- In this reporting period, we added the following utilities to the hundreds of precompiled tools already present in our distribution. We refer to these precompiled tools as user-apps.
    - bamToPsl: converts a bam file to PSL format
    - bedJoinTabOffset: adds offset+length of items in a tab-sep file to a BED
    - bigMafToMaf: converts a bigMaf file to a MAF file
    - expMatrixToBarchartBed: converts a gene expression matrix to barChart format
    - fastqStatsAndSubsample: sanity checks and subsamples a FASTQ file
    - genePredToProt: translates genes annotations to protein sequence
    - hgBbiDbLink: creates an SQL table containing a pointer to a big* file
    - hgFakeAgp: creates an AGP file based on N's
    - mafToBigMaf: converts a UCSC MAF format file to bigMaf
    - pslMapPostChain: post-processes TransMap chained PSL files
    - `vai.pl`: command-line version of our powerful online Variant Annotation Integrator (VAI) tool (see detailed discussion above).

- In addition to the new utilities that we make available as compiled tools (see user-apps, above), we also create internal tools for use in the Browser or data processing pipelines. These utilities are also available via download from github to any interested user. Here is a list of new utilities created during this period:
    - bigMafToMaf
    - crisprKmers
    - geoIp
    - hgMergeSplitTables
    - hgvsToVcf
    - hubToMetaRa
    - markDownToHtml
    - optimalLeaf
    - pyLib
    - snpedia
    - tabRepeatedFieldsToArrayField
    - tabToHtml
    - trashRead
    - trixContextIndex
    - vcfToHgvs

- These are the new Browser software tools (CGIs) that we created during this period:
    - **hgGeneGraph**: The new Pathway and Gene Interactions tool combines and displays data from a number of curated interaction and pathway databases as well as interactions mined from over 20 million PubMed abstracts through the Literome project. It comes with a companion Pathway and Gene Interactions annotation track on hg38 and hg19.
    - **hgGtexTrackSettings**: This new graphic-based "Tissue Body Map" configuration page allows users to view GTEx-sampled tissues in an anatomical context and select tissues from the anatomy graphic as an alternative to using the alphabetical tissue list.
    - **hgLinkIn**: This new tool allows other biomedical websites to link directly to specific positions of the genome in our main tracks display. It was written specifically for UniProt, but is generic enough to work for any other databases that would like to link directly to us using their own IDs.
    - **hgCollection**: This CGI, nearly ready for release, will allow users to group tracks into collections, view the set of tracks in various display modes, and perform operations on them such as sorting, or addition/subtraction.

Aim 2. Build genome browsers and comparative genomics resources for species of biomedical interest.

New and updated genome browsers

- New assemblies:
    - Golden Snub-Nosed Monkey (rhiRox1), Novogene
    - Bison Bison Bison (bisBis1), University of Maryland
    - Malayan Flying Lemur (galVar1), Washington University

    - Chinese Pangolin (manPen1), Washington University
    - Tibetan Frog (nanPar1), BGI-Shenzhen
    - Proboscis Monkey (nasLar1), Proboscis Monkey Functional Genome Consortium
    - Golden Eagle (aquChr2), The Genome Institute - Washington University School of Medicine
    - Green Monkey (chlSab2), Vervet Genomics Consortium
- New assemblies on existing organisms:
    - C. Intestinalis (ci3), Department of Zoology, Graduate School of Science; Sakyo-ku, Kyoto
    - Turkey (melGal5), Turkey Genome Consortium
    - Chimpanzee (panTro5), Chimpanzee Sequencing and Analysis Consortium
    - Gorilla (gorGor5), University of Washington
    - Chicken (galGal5), International Chicken Genome Consortium

New multiple-alignment tracks

- 20-species Multiz Conservation track for rat (rn6)
- 60-species Multiz Conservation track for mouse (mm10)
- 17-species Cactus multiple-alignment track for mm10, rn6, and 15 mouse strains: sPRET_EiJ, pWK_PhJ, cAST_EiJ, wSB_EiJ, nZO_HlLtJ, c57BL_6NJ, nOD_ShiLtJ, fVB_NJ, dBA_2J, cBA_J, c3H_HeJ, aKR_J, bALB_cJ, a_J, lP_J

Aim 3. Import data from the scientific community that help interpret the functions of various human genome regions into the UCSC databases.

New and Updated Data

- See **Table B2.1** for a complete list of tracks released during this reporting period.
- We are implementing a new process for maintaining versions of annotation tracks that are automatically updated. Before the update, a copy of the data is posted to our download server. We provide these archived versions to allow for better reproducibility and transparency. So far, we have followed this new process for the ClinVar track, http://hgdownload.cse.ucsc.edu/goldenPath/hg38/archive/clinvar/, and we intend to gradually phase this in for the rest of the automatically updated tracks (excluding the GenBank tracks).

Table B2.1. Annotation tracks released on the Genome Browser during this reporting period. Tracks that are automatically updated when new data is released are marked as “auto-update”.

Species	Assembly	Track	Status
Human	hg38	GENCODE V24	updated
		GENCODE v26	updated
		GRCh38 Patch 9	new
		NCBI RefSeq Genes	new
		Uniprot	updated
	hg38, hg19	ClinGen Research (formerly	auto-updated

		ISCA)	
		ClinVar Variants	auto-updated
		COSMIC v81	updated
		dbSNP v150	new
		Gene Interactions	new
		Locus Reference Genomic (LRG)	new & updated
	hg38, hg19, hg18	Database of Genomic Variants (DGV): Structural Variation	updated
		Gene Reviews	auto-updated
		NHGRI Catalog of Published GWAS	auto-updated
		OMIM Genes & Phenotypes	auto-updated
	hg19	DECIPHER	auto-updated
		LOVD Variant	auto-updated
		RefSeq Acc	new
	hg18	Personal Genome Variants	updated
Human & Other	hg38, hg19, galGal5	dbSNP v147	new
	hg38, hg19, mm10, ce10, ci2, criGri1, danRer7, dm6, rn5, sacCer3	CRISPR/Cas9	new & updated
	hg38, hg19, mm10, mm9, danRer10, danRer7, galGal5	GRC Incident Database	auto-updated
	hg19, mm10	Pfam Domains in UCSC genes	updated
	most assemblies	Chains & Nets	new & updated
	many assemblies	GenBank Updates (RefSeq Genes, ESTs, RNAs)	auto-updated
	many assemblies	TransMap Alignments v4	new
Mouse	mm10	Alternate Mouse Strain Sequences	new
		GENCODE VM11	updated
		GENCODE VM14	updated
		GRCm38 Patch 4	new
Cow	bosTau8	dbSNP v148	new
Drosophilids	dm6	Pfam Domains in RefSeq Genes	new
Rat	rn6	BAC End Pairs	new
Rat & Other	rn6 & 19 others	20 Species Conservation	new

Other	anoCar2, bosTau6, calJac3, canFam3, ce11, chlSab2, danRer10, chlSab2	Ensembl Genes v86	new & updated

New Track and Assembly Hubs

- We continue to promote the use of track data hubs to display large data sets from consortia and other external labs rather than importing the full data sets ourselves. See **Table B2.2** for a list of new public hubs released during this reporting period.
- In preparation for meeting our upcoming aim to host all of our native tracks in hubs, this year we created two hubs from new external data sources:
    - Mouse Strain Assemblies hub: data from the Mouse Genomes Project at Sanger
    - GTEx Analysis hub: data from GTEx Analysis Group (Lappalainen Lab) at NY Genome Center
- Track Hubs, and related binary-indexed custom tracks such as bigBed and bigWig, require users to host their data remotely. However, many of our users remark that they do not have a web-accessible location at which to host their data for such use. The technology underlying track hubs relies on transmission over the internet of small sections -- just the data in the browser view -- from often very large data files. These small data requests (called byte-range requests) are unsupported by free storage providers such as DropBox or Google Drive. This year, we learned about CyVerse (formerly the iPlant Collaborative) (http://www.cyverse.org), an NSF-funded organization tasked with providing free storage for the life sciences research community. CyVerse has been very responsive to our requests for collaboration to ensure that we provide stable, accessible, archived storage space for our users. Together with CyVerse, we aim to provide a reliable and stable location for storing, accessing, and viewing user data on the UCSC Genome Browser.

Table B2.2. Our users continue to embrace the concept of Track and Assembly Hubs. While many thousands of hubs were created this year, this table lists the hubs whose creators requested they be made public and shared with all Browser users.

Track Hub Name	Provider	Assemblies
Mouse Strain Assemblies	Mouse Genomes Project & UCSC	Mm10, 129S1_SvImJ, WSB_EiJ, SPRET_EiJ, AKR_J, DBA_2J, CBA_J, A_J, C57BL_6NJ, CAST_EiJ, PWK_PhJ, LP_J, C3H_HeJ, NZO_HlLtJ, NOD_ShiLtJ, FVB_NJ, BALB_cJ
Human Craniofacial Epigenomics	Cotney Lab at UConn Health	hg19, mm9
GeneHancer: genome-wide integration of enhancers and target genes in GeneCards	GeneCard MalaCards	hg38
Hippocampal DNA Methylation and Gene Transcription	Rodney Johnson's lab at the University of Illinois	susScr3
EPD Viewer: Promoter specific experimental data and TSS annotation from the EPD database	Eukaryotic Promoter Database	amel5, araTha1, zm3, spo2

Uniquely mappable regions of genome and methylome	Hoffman Lab at the Princess Margaret Genomics Centre	hg38, hg19, mm10, mm9
Comparative data and new chromosome assemblies for Falcon and Pigeon	Damas et. al.	colLiv2, galGal4, falPer2
UniProtKB Features	The UniProt Consortium	hg38
ChIP-seq data from mesenchymal stem cells	Meyer et. al.	mm9
GTEx Allele-Specific Expression in 53 tissues	GTEx Analysis Consortium (NY Genome Center) & UCSC	hg19, hg38
Pancreatic islet long non coding RNA transcripts	Akerman et. al.	hg19
WormBase nematode Assembly Hub	WormBase	renneri, japonica, remanei, ovolvulus, briggsae, brugia, brugia3, elegans, s_ratti

Aim 4. Build high-quality gene sets on the human genome and selected model organism genomes.

New and updated gene sets

- See **Table B2.3** for a full list of gene sets released this year.
- We attempt to keep up-to-date with nearly every GENCODE Genes release on both human and mouse. Additionally, we often convert new GENCODE releases from to hg19 from hg38.
- On other genomes (non-human or mouse), we typically release one or two Ensembl Gene sets per year.
- Instead of updating the UCSC Genes track on hg38 this year, we devoted our resources to incorporating the new NCBI RefSeq Gene Alignments (see discussion in the next section).
- We also added two new gene-related tracks this year. The Pathways and Gene Interactions track shows a summary of gene interaction and pathway data collected from two sources: curated pathway/protein-interaction databases and interactions found through text mining of PubMed abstracts. This track is the companion track to the Gene Interactions tool discussed in **Aim 1**. The second gene-related track, Gene Reviews, is automatically created from an online collection of expert-authored, peer-reviewed articles that describe specific gene-related diseases.
- We augmented the details page of the default gene set (GENCODE) on human by adding a section showing a boxplot of RNA-seq expression data from GTEx.
- We discussed with staff from the GENCODE Genes project, the suitability of their most recent mouse genes for use as our default gene set on mm10. They assured us that although the gene set is not quite ready, it will be soon. Until it is ready, we will continue to use our UCSC Genes method to produce the default gene set for mm10.

- We added links to the Exome Aggregation Consortium from our ExAC annotation tracks.
- We added links to the Ensembl Gene Tree from our Ensembl Gene annotation tracks.

Table B2.3. New and updated gene sets released on the Genome Browser during this reporting period.

Gene Set	**Version**	**Assembly**	**Notes**
GENCODE Genes	V24	hg38	Updated
	V26	hg38	Updated
	V25	hg38	Available on genome-preview server.
	V24	hg19	Lifted down from hg38.
	V26	hg19	Lifted down from hg38. Available on genome-preview server.
	VM11	mm10	Updated
	VM14	mm10	Updated
	VM12, 13 & 15	mm10	Available on genome-preview server.
NCBI RefSeq Genes		hg38	New. See discussion below.
Ensembl Genes	v86	many	Updated
	v86	mm10	Removed from mm10 to avoid confusion with GENCODE Genes track.
TransMap Alignments	V4	many	Updated
Gene-Related Track	**Version**	**Assembly**	**Notes**
Gene Interactions		hg38, hg19	New
Gene Reviews	auto-updated	hg38, hg19, hg18	New

NCBI RefSeq Genes

- For years our users have requested that we provide RefSeq Gene annotations directly from NCBI in addition to our traditional RefSeq Genes track that shows BLAT alignments of NCBI mRNAs to the genome. In last year's progress report, we discussed the fact that together with Ensembl, we had been working closely with NCBI to obtain these files in a GFF3-format file that meets our needs. NCBI was able to produce files that satisfied both us and Ensembl. This year we released a track that shows transcript mappings taken from RefSeq release GFF3 files. Additionally, it includes predicted RefSeq Genes (XM_*, XR_*) and mitochondrial proteins (YP_* for human, NP_* for mouse). We also include other items from the GFF3 file such as non-XM/XR/NM/NR/YP/NP items e.g. pseudogenes, miRBase annotations, V_segment, and J_segment.

Other Accomplishments

Mirror Sites

- We continue to manage two full mirror sites: one in Europe (hosted in Bielefeld, Germany) and one in Japan (hosted at RIKEN in Yokohama, Japan).

- During this reporting period, we added a public MySQL server to the European mirror site. It can be accessed at genome-euro-mysql.soe.ucsc.edu. This provides faster access to MySQL for our European users and allows users to choose to use that local server to serve their GBiB instances.
- This year we replaced the machine that runs the European mirror site.
- As a nod towards testing the efficacy and efficiency of MariaDB versus the MySQL database, we have installed it on a few servers. We currently run the latest stable version of MariaDB on our European and Asian mirror sites, as well as on three failover backup machines: hgnfs1, mysqlrr and customdb.
- Although we can control which version of MySQL we use, we do not know which version(s) our users will use when they install mirror sites. Consequently, we have updated our code for compatibility with the most recent versions of MySQL servers.

Collaboration with other biomedical projects/sites

- In October 2016, six members of the Genome Browser technical staff traveled to Hinxton, England, to meet with our counterparts from the Ensembl Browser staff. These two days of meetings were fruitful and collaborative. Since this meeting, our email interactions have been more frequent and interactive. We have conducted two follow-up phone calls with Ensembl at 6 and 12 months post-visit. The two teams are uniquely qualified to work together, as we are actively trying to solve similar problems for the worldwide audience of Genome Browser users, who benefit from consistency as they move between sites. The Ensembl team has previously sent representatives to meet with us, but this is the first time we have made the trip to England to meet with them.

- In support of our GTEx supplement, Kate Rosenbloom, senior Browser software engineer, has worked closely with the GTEx consortium. She has attended meetings with the steering committee, analysis group, and user community, and shared GTEx work at UCSC with them via presentations on GTEx conference calls and invited talks to the steering committee. She has worked closely with the GTEx data portal engineers to further complementarity and cross-referencing between the GTEx portal and GTEx resources at UCSC. She has solicited feedback and input from GTEx investigators during development of annotation tracks and displays she has created for the Browser. See **Section B3** of this RPPR for details.

- Regeneron Pharmaceuticals
    - Regeneron has fully licensed the Genome Browser source code since mid-2015. After some time relying on their in-house team, they requested additional feature support and data customization from the browser group. This resulted in a 6-month service contract totaling 120 hours of work. Our Associate Director, Robert Kuhn, handled the negotiations and is the contact person at UCSC for the Regeneron team.Through biweekly phone calls, Dr. Kuhn and the Regeneron team lead determine the priority of the work based on the needs of Regeneron users of their internal Genome Browser mirror site.
    - We entered into this contract based on feedback from NHGRI encouraging us to find additional sources of revenue. This service contract has allowed us to add features and data that Regeneron is interested in, and that will also benefit a majority of our users. The terms of the contract require that we embargo some of the data sets for 6 months. Nevertheless, typically these data are not among our priorities, and the release is sooner than we would have accomplished without Regeneron's backing. While some data we work with for Regeneron (e.g. bacterial datasets) are not released to the public site, all new site features developed for Regeneron are available immediately to our users.

    - As part of this service contract, we have created and installed the following data sets to the Regeneron mirror site. Starred (*) tracks have also been released to our public website during this reporting period.
        - CRISPR tracks for Green Monkey (chlSab2), and Chinese hamster (criGri1)
        - CRISPR track with larger, 10K shoulders around genes, for mouse (mm10)
        - UniProt track for hg38*, hg19, and mm10
        - Chinese Hamster Ovary (CHO) cell line assembly browser. This is from the Beijing Genomics Institute's GenBank submission: GCA_000223145.1. It was released to our public site in November 2017, after it had been on the Regeneron server for 6 months.
        - Hepatitis B Virus (HBV) assembly browser from the NCBI Genome Project's GenBank submission: GCF_000861825.2.

- At our request, Ensembl added links to the Genome Browser directly from their public Track Hub registry. Additionally, Ensembl has a system in place to automate discovery of new public hubs at our site, so they are added to the Ensembl public hub registry on a weekly basis.

- Senior browser engineer Kate Rosenbloom who previously served on the Human Proteome Project (HPP, a project within the Human Proteomics Organization, HUPO) scientific advisory board was tapped this year to review a HUPO draft Proteomics Standards Initiative; the proBED file specification, designed for proteomics annotation in genomic context.

- Ongoing collaboration with CyVerse regarding better integration between their data storage solutions and our tools. See detailed discussion in the Track and Assembly Hubs section of **Aim 3**.

- After discussions with staff at NCBI's ClinVar, we are now displaying their star ratings, additional data fields, and more information in the mouseovers of the ClinVar tracks in the Browser.

- The National Institute of Standards and Technology (NIST) CRISPR working group has established a Genome Editing Consortium. The Consortium allows for members to work collaboratively with NIST to develop measurement solutions and standards to advance confidence in the measurements supporting the genome editing technology. Our Assistant Research Scientist, Max Haeussler, serves on this consortium.

- One of our senior engineers, Angie Hinrichs, participates in the GA4GH Variant Annotation Task Team (VATT) teleconferences. She has shared HGVS test cases with collaborators from VariantValidator.org and proposed changes to previously published test sets.

- Our Assistant Research Scientist, Max Haeussler has been in communication with staff at the UniProt Consortium. In response to consortium requests for changes to our UniProt annotation track, Dr. Haeussler implemented additional features including: splitting it into subtracks, auto-updating the data, and adding the track to more organisms.