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

Multiple Regions display mode: This new Genome Browser display configuration allows users to "slice" their track-viewing experience into a variety of different modes that focus the display on certain features: exon-only, gene-only, or user-defined BED coordinates. Only the portions of track annotations that fall within these displayed regions are shown; extraneous intergenic, intronic and otherwise unwanted regions are hidden from view (Figures B.2.1a and b). 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 chromosome. This allows the user to view annotations upstream and downstream of the haplotype sequence, and visualize the haplotype in the general context of the reference chromosome.

Figure B.2.1a. Multi-region exon-mostly mode discards introns and intergenic regions. This view, which uses the browser's standard display mode, shows a 2.1 Mb region of chromosome X with a long poly-A+ RNA mapped to both strands in duplicate (from Cold Spring Harbor Lab's ENCODE phase 2 data). Each cell line is shown in a different color. Several genes are seen to be differentially expressed in various cell lines, while the ATP11C gene is expressed in all cells lines.

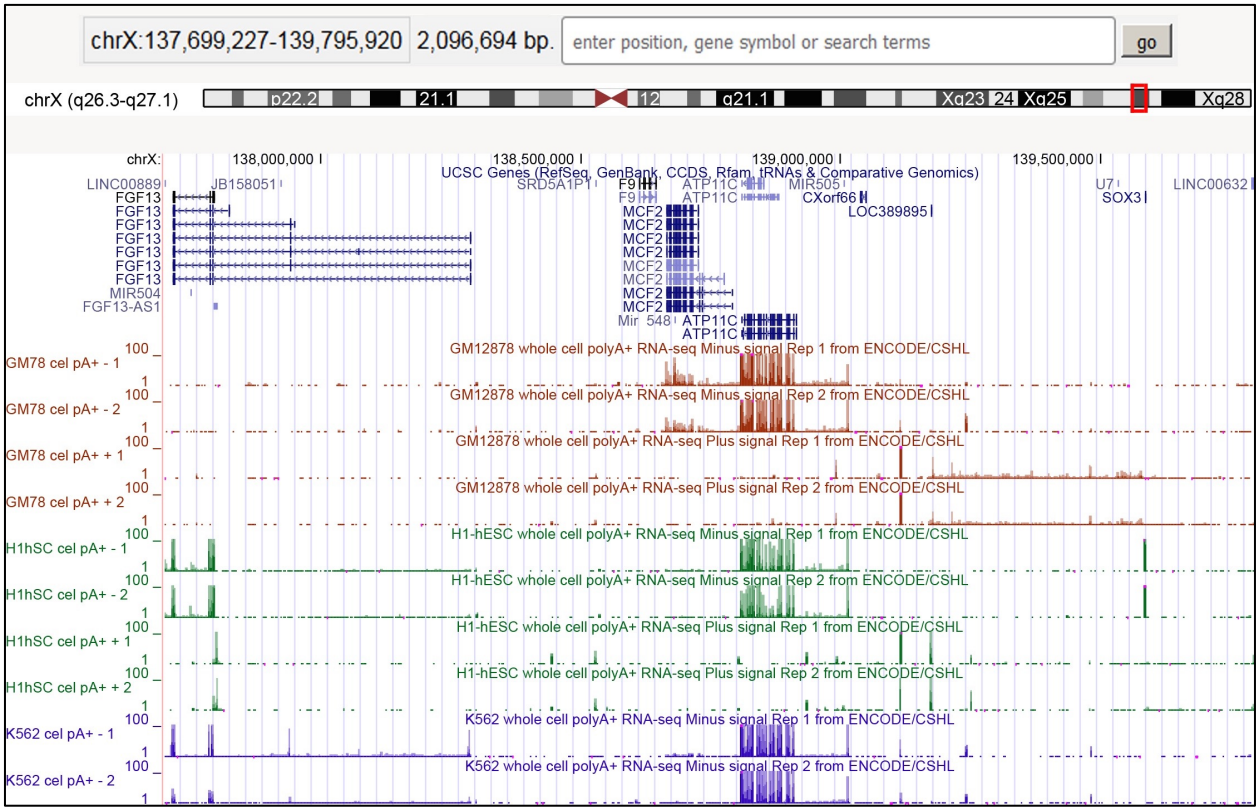
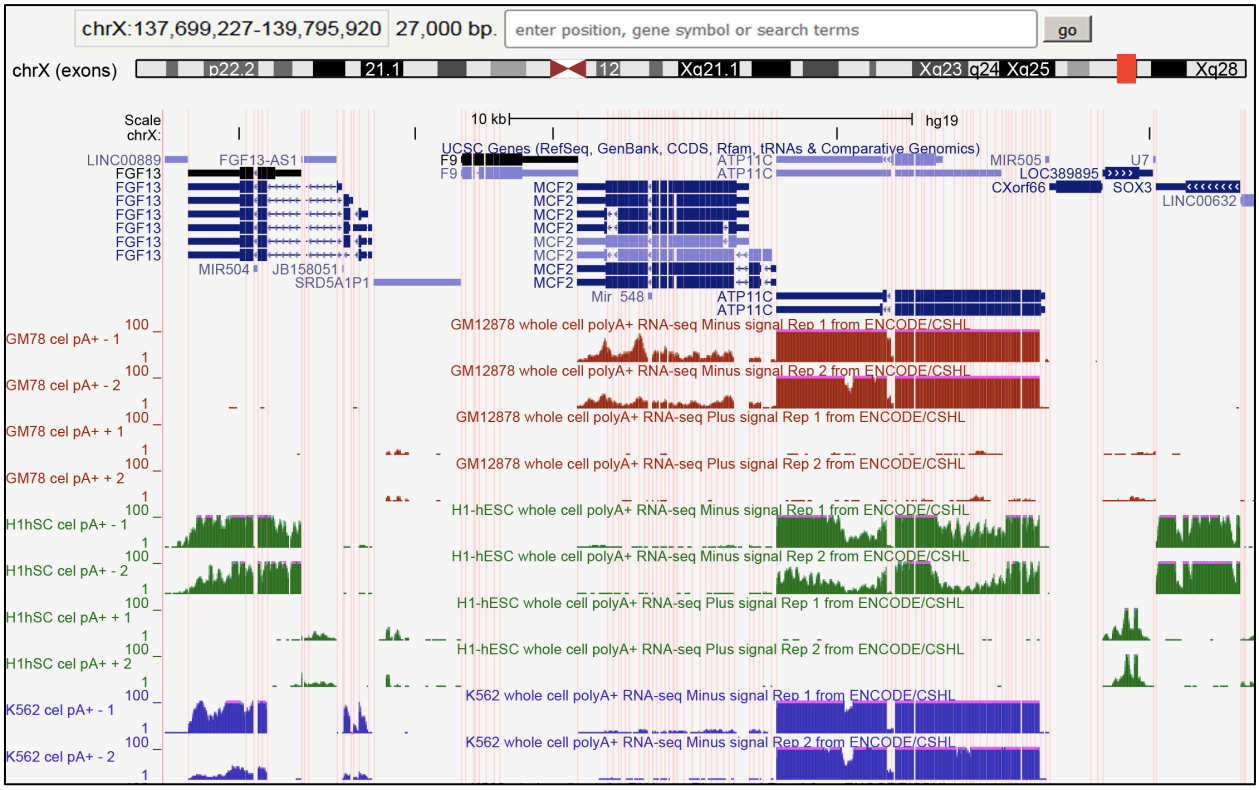


Figure B.2.1b. The same region of chromosome X, this time displayed using the browser's exon-only mode. Only 27 Kb of genome is displayed, and the expressed regions are much more evident.



Support for Genotype-Tissue Expression (GTEx) data: We have developed a new gene expression track display to showcase RNA-seq expression data from the GTEx Consortium. The new display depicts a horizontally oriented bar graph for each gene. Within the graph, each bar corresponds to a specific tissue: the color indicates the tissue type and the height reflects the median expression level of the gene in that tissue (Figure B.2.2). Each bar graph in the track has an associated details page that displays a box and whiskers plot showing distribution of the per tissue sample values (quartiles and outliers) and provides links to affiliated information (Figure B.2.3). The horizontal layout of this new format allows the browser to support a denser display of expression annotations alongside other tracks, in contrast with the older display format of earlier expression tracks (such as GNF Atlas) in which the annotations were laid out vertically on the tracks display.

The GTEx track can be configured to filter the genes by overall expression value or to show only protein-coding genes. A sortable tissue selector, which lets the user pick the tissues to show in the track, can be displayed in a separate popup window that can be viewed in conjunction with the main track display. While the annotation track display is a summary across samples, the full set of expression calls per gene per sample is available for data mining, along with the publicly available sample and donor metadata.

Figure B.2.2. A 90 Kb region of chromosome 9 where GENCODE annotates eight protein-coding genes with GTEx data showing differing patterns of tissue-specificity in gene expression. Each gene is represented by a fixed-width bar graph, where each bar represents a tissue. By hovering the mouse cursor over a bar of the graph, the user can view the tissue name and median expression level in RPKM (reads per kilobase of transcript per million mapped reads).

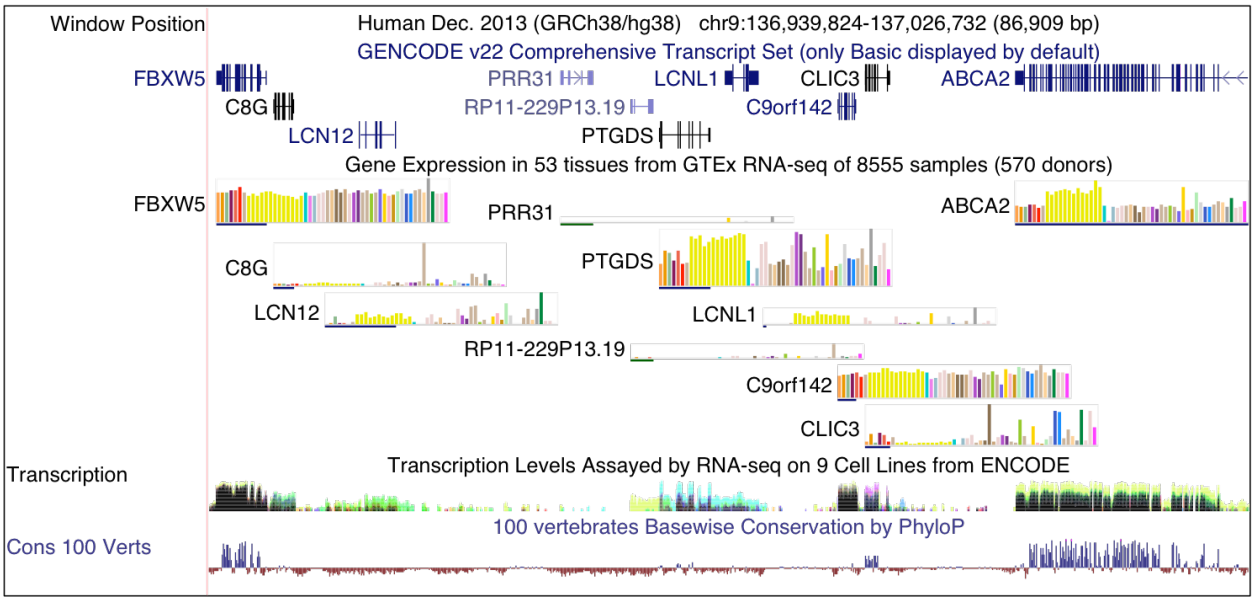
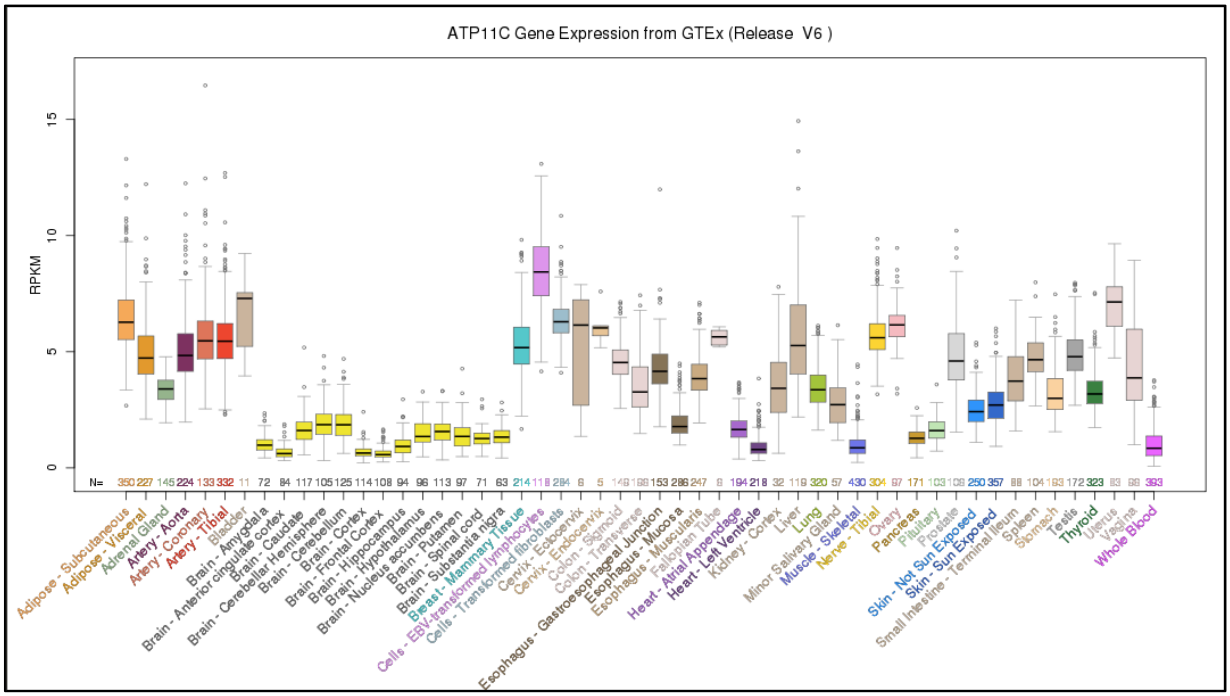


Figure B.2.3. Box and whiskers plot on the track details page for the ATP11C gene, displayed by clicking on the gene's bar graph in the track display. The plot shows the distribution of the per tissue sample values (quartiles and outliers). The details page also provides links to the affiliated UCSC Genes page and the corresponding page at the GTEx portal.



Genome Browser in the Cloud (GBiC): This new program sets up the UCSC Genome Browser on a user's virtual machine or native Linux server. The virtual machine can run in a data center of one of the many cloud service providers. GBiC has been tested on Amazon, Microsoft, and Google cloud services, as well as the on-site cloud solution OpenStack. GBiC detects the Linux distribution, installs the appropriate Apache server, MySQL database servers and UCSC Genome Browser software, downloads genomic data for selected assemblies from UCSC, and sets up file and database permissions and default passwords, similar to GBiB (Genome Browser in a Box). GBiC is useful for individuals who are unable to use our GBiB virtual machine solution due to technical limitations, or who have already migrated IT services to external cloud providers.

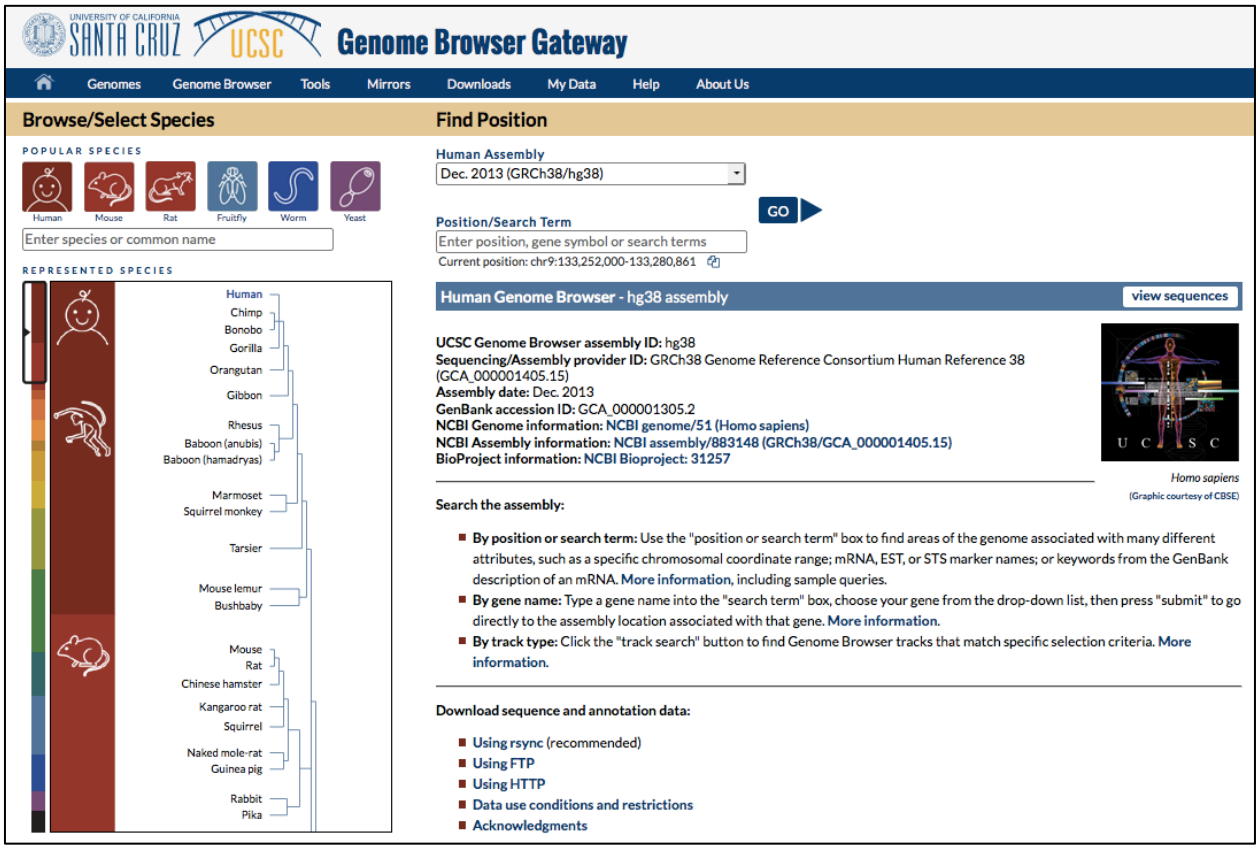
1.a. Increase website interactivity

- Redesigned the interface of the Genome Browser gateway page (http://genome.ucsc.edu/cgi-bin/hgGateway) to be more interactive and intuitive (Figure B.2.4).
- Began redesigning the Genome Browser home page (http://genome.ucsc.edu/) to include more graphics, less text, and better access to Genome Browser tools and content. Release planned for Summer 2016. [carried over from Yr3]
- Made several improvements to the Genome Browser navigation menu bar to facilitate access to popular genome assemblies and configuration options.
- Added a new menu option to allow users to export the currently selected data region to several different external tools, such as Ensembl, NCBI Map-View, CRISPOR, etc.
- Added keyboard shortcuts to the Genome Browser tracks image and menus for common tasks.
- Added “Apply” button to all pop-up configuration boxes to allow real-time visualization of the configuration changes in the tracks image.
- Added a link back to the Genome Browser (at previous position) from BLAT results.
- Added position retention functionality when user returns to a previously visited assembly.
- Added support for automatically converting the display from BAM to wiggle format, to allow users to view large regions of BAM alignments that previously could be displayed only in dense mode.

1.b. Adapt to new types of data

- Added support for Genotype-Tissue Expression (GTEx) data (see section “Major new features added this year”, above).
- Added support for new remotely hosted data types: bigPsl, bigChain, bigMaf.
- Added display for long-distance chromatin interactions including those across chromosomes. [carried over from Yr3 goals]
- Added support for CRAM files (compressed alignment files, similar to BAM).
- Added Table Browser access to the OMIM data when a user is not in whole-genome mode, in compliance with OMIM’s request to restrict access to no more than one chromosome at a time.

Figure B.2.4. The redesigned Genome Browser Gateway page offers the same easy access to the browser's genome collection and datasets, but features an updated style and color scheme, a browsable phylogenetic order tree-style menu, quick access shortcuts to popular genomes, and autocomplete searching for genomes available for browsing.



1.c. Adapt to higher volumes of data

- Built the foundation for establishing Track Hubs as an industry standard:
  - Collaborated with Ensembl, Dalliance, and the Epigenomics Roadmap teams to identify the core set of trackDb settings that must be present in a valid track hub.
  - Defined the various levels of support (required, base, deprecated, new, full) and categorized the settings included in each (http://genome.ucsc.edu/goldenPath/help/trackDb/trackDbHub.html).
  - Added versioning to trackDb settings and documentation.
  - Created a tool for users to validate their hubs against trackDb versions, which Ensembl has incorporated into their new automated hub tool.

  See section B.6 for future collaboration plans with Ensembl on our new automated public track hub upload and validation tool. [new goal added in Yr3 report]

- Continued our ongoing support for assembly hubs.
- Expanded the Data Integrator to access a wider range of browser database tables.
- Added code to prevent new default browser tracks from automatically appearing in existing saved sessions. [new goal added in Yr3 report]

- Updated the custom tracks management “view in” interface, and added Data Integrator to the view options.
- Increased the capacity for handling larger browser URLs.

1.d. Enhance the security of uploaded data

- Added SSL (Secure Sockets Layer) connectivity from the browser to MySQL, and added SSL support in several parts of the browser.
- Added Server Name Indication (SNI) support for HTTPS with certificates for wild-card domains.
- Maintained and updated the content, look and security of the Genome Browser online store (https://genome-store.ucsc.edu/).
- Added support for redirects at the URL data cache (UDC) level.
- Patched potential point-of-attack from user’s track hub description pages.

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

- Continued to package the source code and command-line tools every 3 weeks and distribute them for public use.
- Added the following utilities to the hundreds of utilities already present in our distribution:
  - bigPslToPsl: converts a bigPsl file to PSL format
  - chainToPslBasic: converts a chain file to PSL format
  - hgLoadChain: loads a generic chain file into a database
  - hgLoadMaf: loads a MAF file index into a database
  - hgLoadMafSummary: loads a summary table of pairs in a MAF file into a database
  - hgLoadNet: loads a generic net file into a database
  - newPythonProg: makes a skeleton template for a new Python program
  - pslToBigPsl: converts a PSL file into bigPsl input (BED format with extra fields)

Other updates to the Genome Browser website and software

- Continued our ongoing technical support for Genome Browser mirror sites, and added the ability for a mirror site to specify the location of a Galaxy instance to the Table Browser.
- Improved the search mechanisms for track hubs.
- Added several enhancements to the Variant Annotation Integrator (VAI):
  - offer transcript status info when available
  - offer gene tracks from assembly hubs
  - updated underlying dbNSFP3.1a data for human assembly hg38
- Enabled KeepAlive in httpd.conf to improve static content retrieval for users with slow TCP connect speeds.
- Removed obsolete 8-bit color support.
- Provided ongoing support for Genome Browser in a Box (GBiB).

- Added the ability to specify the genome within assembly hub in URL.
- Provided mechanism that allows users to post their sessions for others to see. [carried over from Yr3]
- Improved the method by which GenBank data is updated on Genome Browser servers.
- Added performance-tracking statistics to UDC layer.

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

New and updated genome browsers

- Added 2 new genomes:
  - brown kiwi (aptMan1)
  - crab-eating macaque (macFas5)
- Updated 7 genome assemblies:
  - cat (felCat8)
  - C. elegans (ce11)
  - frog (xenTro7)
  - gorilla (gorGor4)
  - mouse lemur (micMur2)
  - platypus (ornAna2)
  - rhesus (rheMac8)
- Generated many new assembly hubs: To determine the robustness of our assembly hub system, we created a script that automatically generates assembly hubs for every complete genome in GenBank and RefSeq, and then adds 2-3 simple annotation tracks, including a gene track if available. As of April 2016, we have automatically created approximately 70,000 assembly hubs on our engineering development server.

New multiple-alignment tracks

- Human (GRCh38/hg38): 100 species
- C. elegans (ce11): 26 species

Reevaluation of alternative alignment tools

This year we did not conduct a formal evaluation of alignment tools, although we did maintain close contact with the developers of the Cactus alignment tool. Although Cactus has been used to successfully create smaller multiple alignments (15-30 species) within clades, the pipeline has not yet scaled up to the 100s. For the foreseeable future, we intend to continue using our current multiz alignment pipeline.

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

New default human genome assembly: GRCh38/hg38

We switched the human assembly displayed by default in the Genome Browser to the newest version in conjunction with the release of the 100-species Conservation track on the GRCh38/hg38 assembly in mid-Sept. 2015.

New and updated annotations for GRCh38/hg38 and other recent human assemblies

Table B.2.1 lists the annotation tracks added to the Genome Browser during this reporting period. Some GRCh38/hg38 annotations were "lifted" from the previous GRCh37/hg19 assembly (as noted in the table), but most data sets were remapped or obtained from the original data providers. The GTEx Gene Tissue Expression track was added in conjunction with the release of our new gene expression track display (see section "Major new features added this year" in Aim 1, above). This work was funded by the GTEx supplemental funding.

- In addition to the tracks listed in Table B.2.1, several other annotation tracks are under development, many of which will be released by the end of this reporting period. For example, we will soon release a Publications track on GRCH38/hg38 that pulls in data from millions of scientific papers.
- Several of the data sets planned for this year have been deferred to Yr5 (see section B.6, Aim 3). One planned set -- the UMD.be variant database -- was dropped from the list after collaboration attempts with the data provider failed.

Table B.2.1. Annotation tracks released on the Genome Browser during 2015–16. Tracks that are automatically updated when new data are released are labeled "auto-updated".

Species	Assembly	Track	Status
Human	GRCh38/hg38	100 species conservation	new
		BAC End Pairs	new
		ClinGen CNVs	new
		Clone Ends	new
		Coriell deletion/duplication cell line	lifted from hg19
		FISH clones	lifted from hg19
		Fosmid end pairs	lifted from hg19
		GENCODE Genes V24 (UCSC Genes)	new
		GENCODE v23	new
		GeneID Genes	new
		GRC Patch 6	new
		IKMC Genes Mapped	lifted from hg19
		MalaCards	new
		Non-coding RNA	lifted from hg19
		Ribosome profiling -- GWIPS-viz	updated
		SGP evidence-based gene predictions	new
		sno/miRNAs	lifted from hg19

		STS Markers	lifted from hg19
	GRCh38/hg38, GRCh37/hg19	GTEx Gene Tissue Expression	new
		tRNAs	new
		ClinGen Benign Aggregate	new
		ClinGen Research (formerly ISCA)	auto-updated
		ClinVar Variants	auto-updated
		CNV Developmental Delay	new
		GRC Incident Database	auto-updated
		NCBI dbSNP Build 142, 144, 146	new
	GRCh37/hg19	1000 Genomes Phase 3 Variants	updated
		COSMIC	auto-updated
		DECIPHER: Chromosomal Imbalance & Phenotype	auto-updated
		Lens Patents	new
	GRCh38/hg38, GRCh37/hg19, NCBI36/hg18	DVG Structural Variation	new & updated
		Gene Reviews	auto-updated
		NHGRI Catalog of Published GWAS	auto-updated
		OMIM	updated
		OMIM Genes & Phenotypes	auto-updated
		ORegAnno	new & updated
	GRCh37/hg19, NCBI36/hg18	Leiden Open Variation Database (LOVD) Public Variants	auto-updated
Mouse	GRCm38/mm10	Clone Ends	new
		GENCODE VM7	new
		GeneID Genes	new
		Lens Patents	new
		NCBI dbSNP Build 142	new
		SGP evidence-based gene predictions	new
		UCSC Genes	updated
	GRCm38/mm10, NCBI37/mm9	Pfam in UCSC Gene	new
C. elegans	ce11	26 species Conservation	new
Cow Hedgehog Rat	bosTau8 eriEur2 rn6	SGP evidence-based gene predictions	new
Ebola Virus	eboVir3	Lens Patents	new
Dog Drosophilids, Yeast	canFam3 dm2, dm3, dm6 sacCer1, sacCer2, sacCer3	ORegAnno	new & updated
15 more species	----------	GeneID Genes	new
Most species/ assemblies	----------	GenBank updates (e.g. RefSeq Genes, ESTs, RNAs)	auto-updated
		Chains & Nets	new
		Augustus Genes	new & updated
		Ensembl Genes v81	new
		Microsatellites	new & updated

New Track and Assembly Hubs

We added links to 12 new hubs that can be accessed from the Genome Browser (Table B.2.2), and continued 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.

Table B.2.2. Public track and assembly hubs newly released on the Genome Browser website in 2015-16. As of April 2016 we linked to a total of 42 public hubs.

Hub Description	Lab	Assembly
GTEx RNA-Seq Signal	GTEx Consortium, UCSC Computational Genomics Lab, UCSC Genome Browser group	hg38, hg19
Peptide evidences CNIO	Spanish National Cancer Research Centre (CNIO) and Spanish Institute of Bioinformatics	hg38, hg19
Principal Splice Isoforms APPRIS	Spanish National Cancer Research Centre (CNIO) and Spanish Institute of Bioinformatics	hg38, hg19, mm10, danRer10, rn6, susScr3, panTro4, dm6, ce10
DASHR small ncRNA	Li-San Wang Lab, University of Pennsylvania	hg19
Cancer Genomics data from TCGA, ICGC, Immune Epitope Database and Cancer Epitopes	UCSC Genome Browser Group	hg19
lncRNA in Breast Cancer	Kraus Lab at UT Southwestern	hg19
Vista Enhancers	VISTA Enhancer Browser	hg19, mm9, mm10
Cohesin(Smc1)-associated chromatin interactions in murine embryonic stem cells	Young Lab, Whitehead Institute, MIT	mm9
Promoterome CAGE and nucleosome positioning	Computational Regulatory Genomics Group, MRC Clinical Sciences Centre	danRer7
Porcine DNA methylation and gene transcription	Laboratory of Comparative Genomics at the University of Illinois & Animal Breeding and Genomics Centre at Wageningen University, The Netherlands	susScr3
Peterhof Yeasts	Saint Petersburg State University	sacCer3, and strains
Human exons mapped by CESAR	Hiller Lab, Max Plank Institute of Molecular Cell Biology and Genetics	99 assemblies

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

New and updated gene sets

Table B.2.3 shows the list of gene sets (a subset of Table B.2.2) that have been added or updated in the past year.

The default gene set on the GRCh38/hg38 human assembly has been switched to GENCODE Genes, following last year's favorable evaluation of its suitability for this role.

UCSC Genes

- Updated UCSC Genes set for GRCm38/mm10 mouse assembly (scheduled for release in June 2016).
- Evaluated GENCODE Genes as a candidate for replacing UCSC Genes as the primary gene set on mouse (as we have done for the latest human assembly). We decided that the annotation was not yet adequate; thus will reevaluate next year.
- Worked with GENCODE to add piRNAs to the gene set. GENCODE has not completed this work; therefore, this is deferred to Yr5 (see section B.6, Aim 4).
- Updated links from the UCSC Genes details pages to Lynx project (hg38) and MalaCards database (GRCh38/hg38, GRCh37/hg19).

NCBI RefSeq Genes

Many browser 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. We have been working closely with NCBI for multiple years to acquire this information, but have been unsuccessful in obtaining the annotations in a GFF3-format file that meets our needs. Ensembl, who has been experiencing the same difficulties, has joined our collaboration with NCBI, and we are now in agreement on the details of the GFF3 data file. NCBI intends to include this new data in their genome annotation release scheduled for summer 2016. We have added these annotations to our plans for Yr5 (see section B.6, Aim 4).

Table B.2.3. New and updated gene sets released on the Genome Browser during 2015–16.

Gene Set	Assembly	Status
UCSC Genes (based on GENCODE V24)	GRCh38/hg38	updated
UCSC Genes	GRCm38/mm10	updated
Augustus Genes	most assemblies	new & updated
Ensembl Genes v81	most assemblies	new & updated
GENCODE Genes VM7	GRCm38/mm10	updated
GENCODE Genes V23	GRCh38/hg38	updated
GeneID Genes	GRCh38/hg38, GRCm38/mm10, rn6, and 14 additional assemblies	new
IKMC Genes Mapped	GRCh38/hg38	new
Pfam in UCSC Gene	GRCm38/mm10, GRCm37/mm9	new
SGP evidence-based gene predictions	GRCh38/hg38, GRCm38/mm10, eriEur2, rn6, bosTau8	new

Training and Outreach (supplement)

The training and outreach program has two main objectives: Reaching new users and supporting existing users. We provide both in-person workshops, as well as online resources such as help pages and video tutorials. During a 90-minute meeting workshop or a half- or two-day workshop onsite at an institution, a bioinformatics or genomics novice can learn the scope of the Browser toolset and become quite adept at using it. Furthermore, we learn from being in the room with our users – what they need, and what we can improve.

Workshop trainings and meeting appearances

This year we continued to offer trainings in the use of the Genome Browser and its attendant tools, including, where appropriate, instruction in command-line approaches to sequencing-file manipulation, the display of user-generated sequencing results in the Browser, and comparative genomics datasets. The trainings were conducted in a variety of venues, including international meetings of scientific societies, and at individual institutions over the course of 42 full or partial days of instruction (Table B.2.4).

During the funding period, presentations were made to more than 500 attendees at national or international meetings of the American Society of Human Genetics, European Society of Human Genetics, Korean Genetics Organization and Plant and Animal Genomics. The large meeting format, while shorter than that of institutional workshops, allows us to reach people from a large number of institutions at once. These workshops afforded an opportunity to learn what our users need and how we can make our tools more intuitive and easy to use.

Our presentations at large meeting have often resulted in invitations to visit individual institutions for longer workshops, where we can seed a core local user group with deeper instruction. Such intensive workshops, of half- to two-days duration, were given to 1000 scientists at 20 locations this year. Our training seminars are well received, and our post-workshop surveys indicate that even regular users discovered new information about the browser's functionality. Our workshop hosts are frequently contacted by people from outside the institution who wish to attend; these individuals sometimes travel several hundred miles to attend a workshop. Onsite training classes also provide valuable insight into the data and software/display needs of the user community. For example, interactions with users at workshops provided some of the stimulus for development of the new exon-only display mode and of the Genome Browser in a Box.

OpenHelix continues to maintain 3 online tutorials about the Genome Browser, one of which was updated this year. Their online Genome Browser videos have been viewed thousands of times. OpenHelix also maintains and distributes Quick Reference Cards (QRCs) on the Genome Browser and Table Browser.

Browser training video clips

Due to personnel turnover within the training group this year, we were unable to meet our goal to record and release 8-10 short video clips detailing specific browser tasks for display on our YouTube channel; we released only a single video in this reporting period. We recently hired an experienced undergraduate assistant to help with video content editing, and plan to produce more video clips during Yr5 (see section B.6). [new goal added in Yr3 report]

Table B.2.4. Genome Browser training workshops presented in 2015-16.

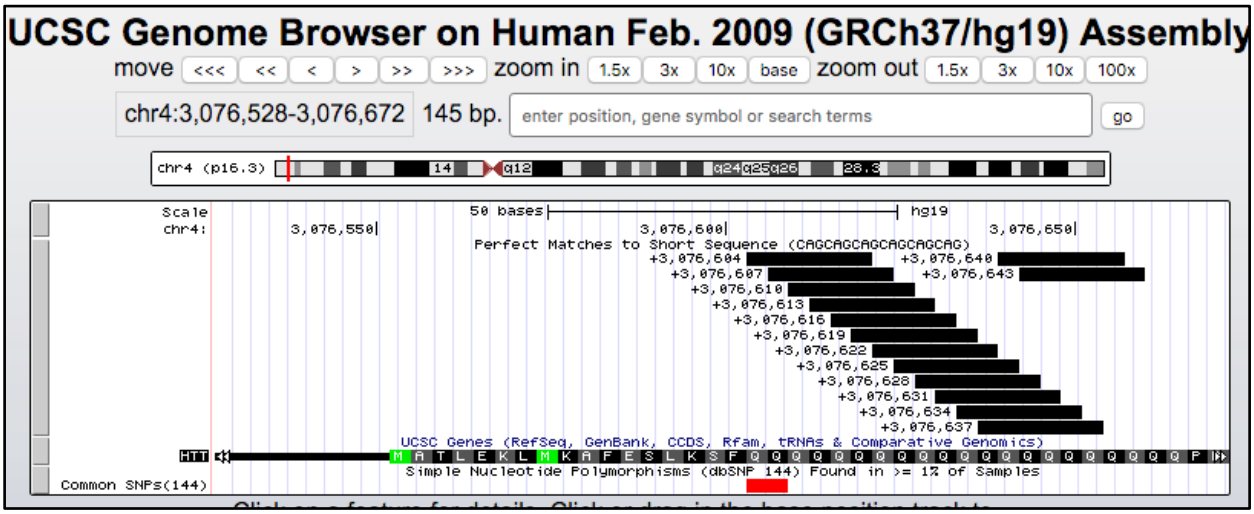
Date	Location
July 2015	Texas A&M University, College Station
	JAX Lab Mammalian Genetics course, Bar Harbor, ME
	James Madison University, Harrisonburg, VA
September 2015	Max Planck, Bad Nauheim, Germany
	Helmholtz, Munich, Germany
October 2015	Johns Hopkins University, Baltimore, MD
	ASHG, Baltimore
	University of Louisville, KY

November 2015	Xmeeting Bioinformatics meeting, Sao Paolo, Brazil
	University of Minas Gerais, Belo Horizonte, Brazil
December 2015	University of California Davis Veterinary School
January 2016	City of Hope, Duarte, CA
	Plant and Animal Genomics, San Diego, Equine session, San Diego
	Plant and Animal Genomics, San Diego, San Diego
	Mt. Sinai Medical School, NYC
	Regeneron, Inc., Tarrytown, NY
February 2016	Korean Genome Organization symposium, Hongchun
March 2016	Cancer Informatics for Cancer Centers (CI4CC) meeting, Napa, CA
	California State University Monterey Bay, Seaside, CA
April 2016	ICHG, Kyoto, Japan
	UT-KBRIN Tennessee/Kentucky bioinformatics summit, Lake Barkley, KY
May 2016	Karolinska Institute, Stockholm, Sweden
	ESHG Barcelona, Spain
June 2016	University of Indiana (tentative)
	Aarhus University, Denmark
	University of Copenhagen

Session gallery highlighting Genome Browser features

This year we added a session gallery (http://genome.ucsc.edu/goldenPath/help/sessions.html) featuring sample Genome Browser sessions that highlight topics of interest in molecular biology education. Sessions include the display of coding and wobble bases, alt-splicing, evolution, variation and disease, as well as topics derived from commonly asked questions on the browser user mailing list and feedback from onsite browser training workshops (Figure B.2.5). [new goal added in Yr3 report]

Figure B.2.5. In this session from the Genome Browser session gallery, the Sequence Match track is configured to match $(CAG)_6$, illustrating the region of the huntingtin gene where the polyglutamine motif is subject to expansion in Huntington's disease.



Genome Browser mirrors

European mirror (genome-euro) in Bielefeld, Germany: This year we planned to double the server's disk capacity to accommodate increases in our data footprint and to site a downloads server. Unfortunately, international logistical roadblocks prevented us from purchasing equipment through Dell for shipment to Germany. We recently found a solution that required us to purchase the equipment in the U.S. and then ship it to Germany ourselves. As of April 2016, the equipment has arrived in Germany, and installation and data transfer has begun. We expect to be operational by the end of this reporting period.

Asian mirror (genome-asia) at RIKEN in Japan: This is a fully functional copy of the Genome Browser currently operational as genome-asia.ucsc.edu as it undergoes final testing. At the ICHG meeting in Kyoto, we announced our intention to publicly release the mirror in May 2016, and plan to begin automatically redirecting browser traffic originating in Asia to this mirror by June 2016. All hardware for this mirror site was purchased directly by RIKEN, and RIKEN and UCSC have signed a Memorandum of Understanding in which RIKEN agrees to provide electricity and cooling for the machines. We will maintain the mirror and provide software and browser expertise in exchange for RIKEN making the site accessible to all of our users.

Program income to support future trainings

Workshops at national meetings are typically given at our expense, though we obtain a waiver of registration fees that limits our expenses to the cost of travel. Institutional workshops continue to operate primarily on a "host pays" basis for plane and accommodation, and also generate program income.

Our program income for this grant year totaled $25,000: $11,800 in collected income, $8,600 invoiced but uncollected, and $5,100 in commitments for scheduled workshops. We may consider raising the price of the workshop fee once the current trove of requests has been completed, as we are easily able to fill the schedule at the current rates. However, we do not feel it is possible to recover the full cost of on-site trainings, which would entail raising the rate by at least five-fold.

Genome Browser citations in the literature

A survey of formal citations in the literature using Google Scholar indicates that the papers describing the Genome Browser and its tools have received more than 22,000 formal citations in the literature, 15% of which occurred in the past year (Table B.2.5). This list does not include papers that our staff have co-authored on topics other than the browser, such as papers about the Genome 10K project, the release of human or other genome assemblies, or bird evolution.

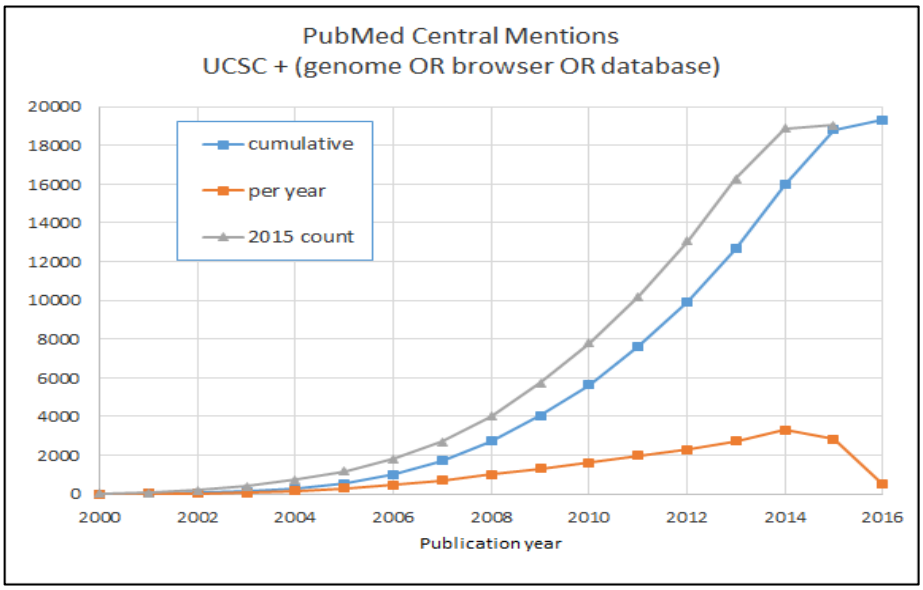
Many researchers who use the Genome Browser or database mention it in the text of the paper, but do not include a formal reference. By querying the text of papers in PubMed Central (PMC), we have identified nearly 20,000 mentions of "UCSC AND (genome OR browser OR database)" in the portion of literature available as full text to PMC (Figure B.2.6). It is not clear how many of these PMC mentions overlap the Google Scholar metric. The trend of increasing usage is clear: each year more papers are published that mention the Genome Browser.

Table B.2.5. Formal citations of the Genome Browser and tools in the literature as of April 2016. Resources marked with asterisk are not directly or fully funded by the Genome Browser grant, but feature the Genome Browser in the subject matter.

Topic	Author, Year	Google Scholar
BLAT	Kent, 2002	5149
Genome Browser	Kent et al., 2002	4252
Conservation	Siepel et al., 2005	1968
Browser database	Karolchik et al, 2003	1394
Threaded Blockset Aligner	Blanchette et al., 2004	970
Genome Browser update 2011	Fujita et al., 2011	906
Genome Browser update 2008	Karolchik et al., 2008	888
Table Browser	Karolchik et al., 2004	851
Genome Browser update 2014	Karolchik et al., 2014	682
Chain/Nets (evolution's cauldron)	Kent et al., 2003	552
Genome Browser update 2010	Rhead et al., 2010	538
Genome Browser update 2013	Meyer et al., 2013	500
Genome Browser update 2006	Hinrichs et al., 2006	470
Known Genes	Hsu et al., 2004	369
Genome Browser update 2009	Kuhn et al., 2009	337
Genome Browser update 2007	Kuhn et al., 2007	291
ENCODE resources 2013 - 5-yr update*	Rosenbloom et al., 2013	291
Genome Browser extensions and updates	Dreszer et al., 2011	281
ENCODE whole-genome data*	Rosenbloom et al., 2010	212
28-way alignment	Miller et al., 2007	196
ENCODE resources 2012*	Rosenbloom et al., 2012	175
ENCODE resources 2011*	Raney et al., 2011	149
Current Protocols	Karolchik et al., 2009	145
BigWig and BigBed	Kent et al., 2010	136
Browser and associated tools	Kuhn et al., 2013	122
Genome Browser update 2015	Rosenbloom et al., 2015	116
Archaeal Browser*	Schneider et al., 2006	105
ENCODE resources 2007*	Thomas et al., 2007	83
Gene Sorter	Kent et al., 2005	65
Proteome Browser	Hsu et al., 2005	51
Track Hubs	Raney et al., 2013	45
Browser Tutorial	Zweig et al., 2008	42
Current Protocols	Karolchik et al., 2011	33
Understanding Genome Browsing	Cline et al., 2009	22
Biotechnology Annual Review - deep support	Mangan et al., 2008	19
Comparative Genomics with Browser	Karolchik et al., 2008	19
Current Protocols 2009	Mangan et al., 2009	15
Genomic Data Resources	Lathe et al., 2008	15
Comparative Assembly Hubs	Nguyen et al., 2014	5
Genome Browser-in-a-Box	Haeussler et al., 2014	5
Current Protocols 2014	Mangan et al., 2014	5

Ebola Portal	Haeussler et al., 2014	4
Genome Browser update 2016	Spier et al., 2016	1
TOTAL		22474

Figure B.2.6. Mentions of the Genome Browser and tools in the text of papers indexed by PubMed Central, shown by publication year (brown) and cumulative (blue). Note the similarity between the same query last year (gray) and this year (blue). Because PMC contains only those papers released for full-text public access, there is often a significant lag time between a paper's release and its archival in PMC. For example, a significant number of papers published in 2014 did not reach PMC until 2015. Because of this, the last two time points in the graph are incomplete.



Genome Browser usage

General website usage

Statistics from this grant year (gathered from our website logs):

- We averaged ~51 million hits per month for the UCSC-based site (45 million) and genome-euro (6 million) combined. Our website logs count a "hit" as any request from a single user's computer whether it's to a CGI, an image, an HTML page, etc.
- These hits are from ~167,500 IP addresses per month averaged over the grant year.
- 144 unique sites successfully created and used a mirror site.
- 320 users successfully downloaded, installed and used the GBiB product.

Since activating Google Analytics (GA) for our CGIs (software web tools) on the Genome Browser website, we have collected the following statistics:

- We have ~120,000 users per month with no upward or downward trend; a very constant stream of users. Note that GA is activated only on our CGIs; not on the rest of our

pages. This explains the slight difference in count between the GA “users” and the “IP” count above.

- We experience the following active user rates: ~11,500 users per day, ~40,000 users per week, ~127,000 users per month.
- We receive ~5 million page views per month with no upward or downward trend. GA counts a “hit” as a single visit to a CGI by a single user. Automated robots and spiders are not included in this value.
- 43% of all browsing sessions are from the USA, 10% from China, 5% from Japan, 4% from the UK, and 4% from Germany.

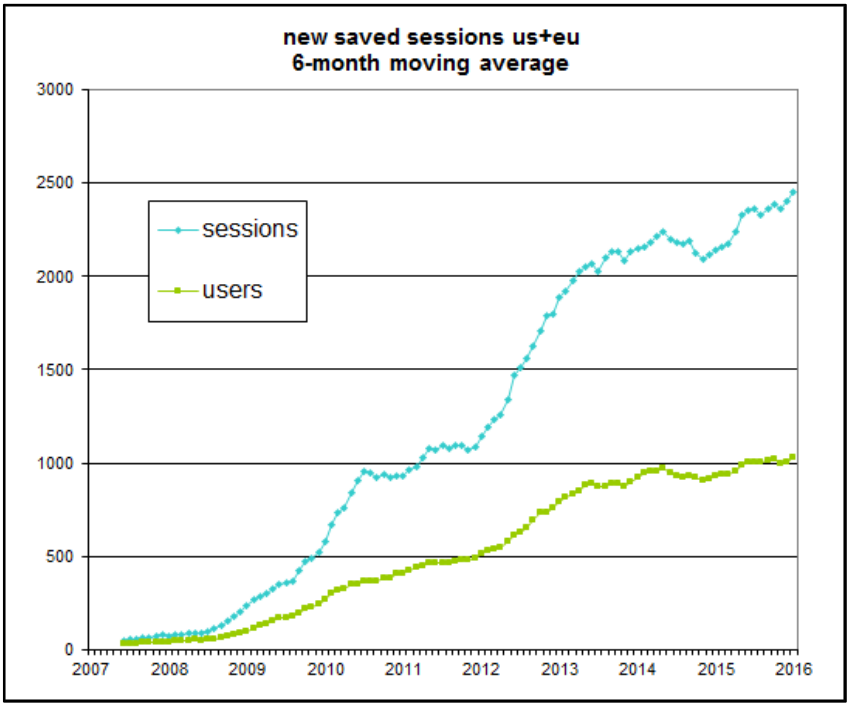
Saved sessions usage

Figure B.2.7 shows the steady increase in the use of the saved sessions feature since its introduction in 2007. In the past year, more than 1000 users each month have created a new saved session, a good measure of active engagement with the browser. 15% of session activity is from the European mirror, and an unknown amount of activity takes place on private mirrors.

Track Hub usage

Track hub usage has increased steadily over the past year. More than 1020 new track hubs were created per month (including our European mirror site) from some 160 independent IP addresses in 2015-16. Additionally, the ENCODE portal at Stanford University generates dynamic track hubs on the fly and displays them on the Genome Browser (excluded from the above count).

Figure B.2.7. New sessions created, and number of users creating them, since 2007 on the US and European servers combined (15% takes place on the European mirror). As the month-to-month variation in the creation of sessions is quite variable, it is presented here as a six-month moving average to make the trend more obvious.



Custom Track usage

- Approximately 189 million custom track files were uploaded to the Genome Browser servers during the past year. Over 73,000 custom track files were moved into sessions for long-term use.
- Nearly 28 TB of data was uploaded to the Genome Browser servers 3/1/15 - 3/1/16. This provides a rough metric of custom track use on the browser.
- 579 TB of data were downloaded from our servers during the same timeframe.

Website and hardware infrastructure

Improvements to the infrastructure of the Genome Browser project during the past year:

- Provisioned two new virtual machine (VM) servers in our hardware pool at the San Diego Supercomputer Center (SDSC).
- Moved our secure online store to one of these new VM servers.
- Installed a complete official mirror site in Japan: genome-asia.
- Updated BLAT servers.
- Implemented access denial to CGIs for Baidu Spider.
- Upgraded the Redmine (issue-tracking system) server and moved to it to the SDSC.

Consistent with past years, the UCSC Genome Browser sites continue to perform with high reliability and stability. Since April 2015, our UCSC-based website has been down for about 10 hours during peak usage hours (6AM - 5PM PT Mon–Fri). When the main server was offline, the genome-euro server in Germany (which experienced no downtime this year) and our other publicized mirrors remained accessible to provide consistent service to our users.

Genome Browser Scientific Advisory Board

The Scientific Advisory Board consists of prominent scientists in genetics and bioinformatics:

- Aravinda Chakravarti, Johns Hopkins University
- Ross Hardison, Penn State University
- Tim Hubbard, King's College, London
- Mary-Claire King, University of Washington
- Robert Waterston, University of Washington Genome Sciences
- Barbara Wold, California Institute of Technology

The Board meets annually in Santa Cruz to offer guidance. The 2016 meeting was held April 21-22. SAB input that will influence our Genome Browser work in the upcoming year is tagged with "[SAB]" in section B.6.