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:

- **Data Integrator** (http://genome.ucsc.edu/cgi-bin/hgIntegrator)**:** a new Genome Browser CGI that provides a simple and flexible interface for combining data from up to 5 tracks from the Genome Browser database as well as custom tracks and hub tracks (Figure B.2.1). The CGI was implemented with a new javascript UI framework that uses the ReactJS and ImmutableJS libraries.
- **Ebola Portal and Browser:** In Sep. 2014, in response to the Ebola epidemic in West Africa, we released a Genome Browser and information portal (http://genome.ucsc.edu/ebolaPortal/) for the Jun. 2014 assembly of the Ebola virus submitted by the Broad Institute (Figure B.2.2). We worked closely with the Pardis Sabeti lab at the Broad Institute and other Ebola experts throughout the world to incorporate annotations useful to those studying Ebola. We received retroactive permission to use Genome Browser funds for this work.
- **GTEx data support:** We received additional funding effective Jul. 2014 to integrate the NIH Common Fund's Genotype-Tissue Expression (GTEx) data into the UCSC Genome Browser. See sections B.2.1.b and B.3 for descriptions of work accomplished during this grant period.

Figure B.2.1. Screenshot of the Data Integrator showing several selected tracks: a custom track, genes, conserved TFBS, and ENCODE Transcription Factor ChIP peaks. The text output can be saved to a local file or displayed in the web browser window. The "Choose fields" button pops up a dialog in which columns of each track can be selected or deselected, so that only the desired columns appear in the output.

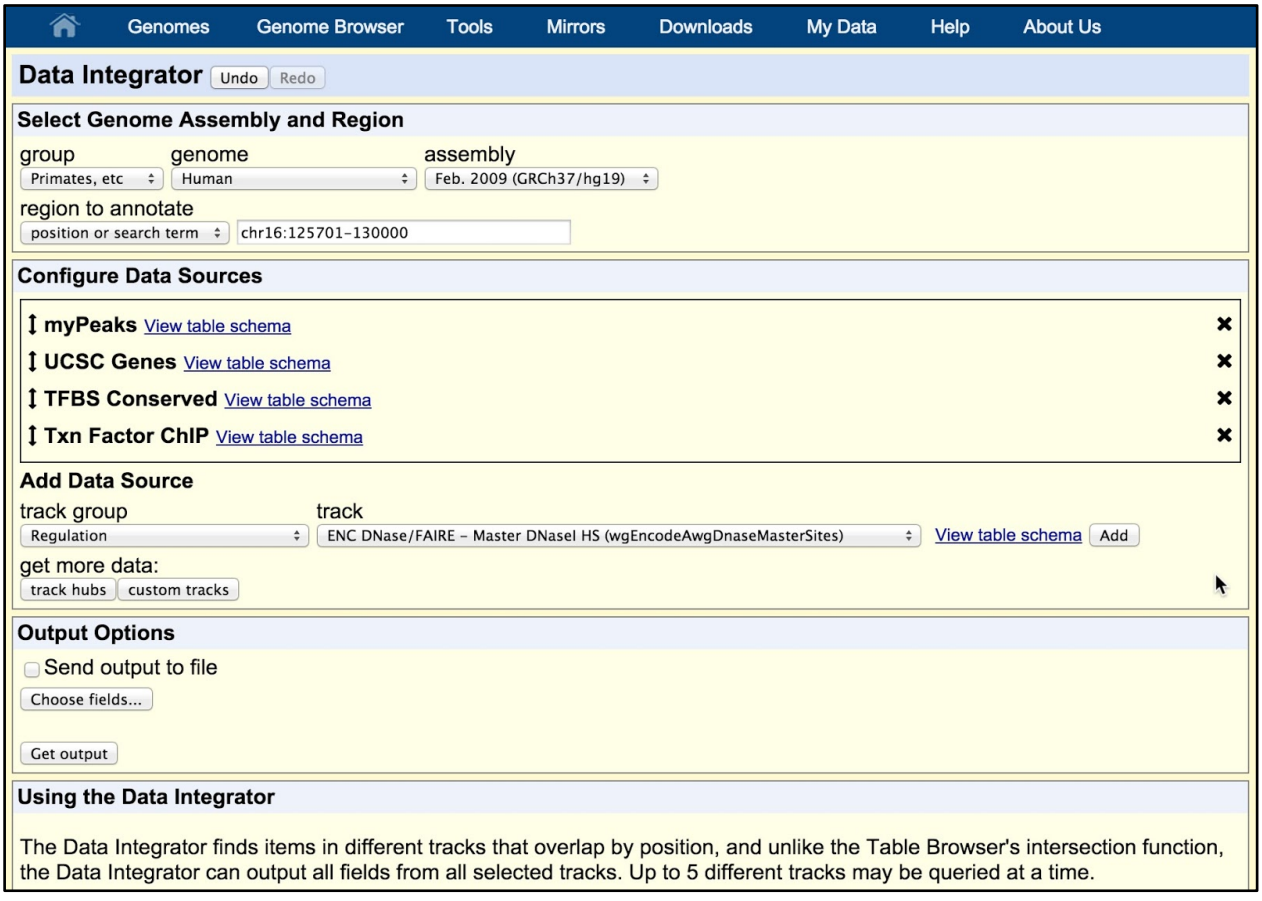
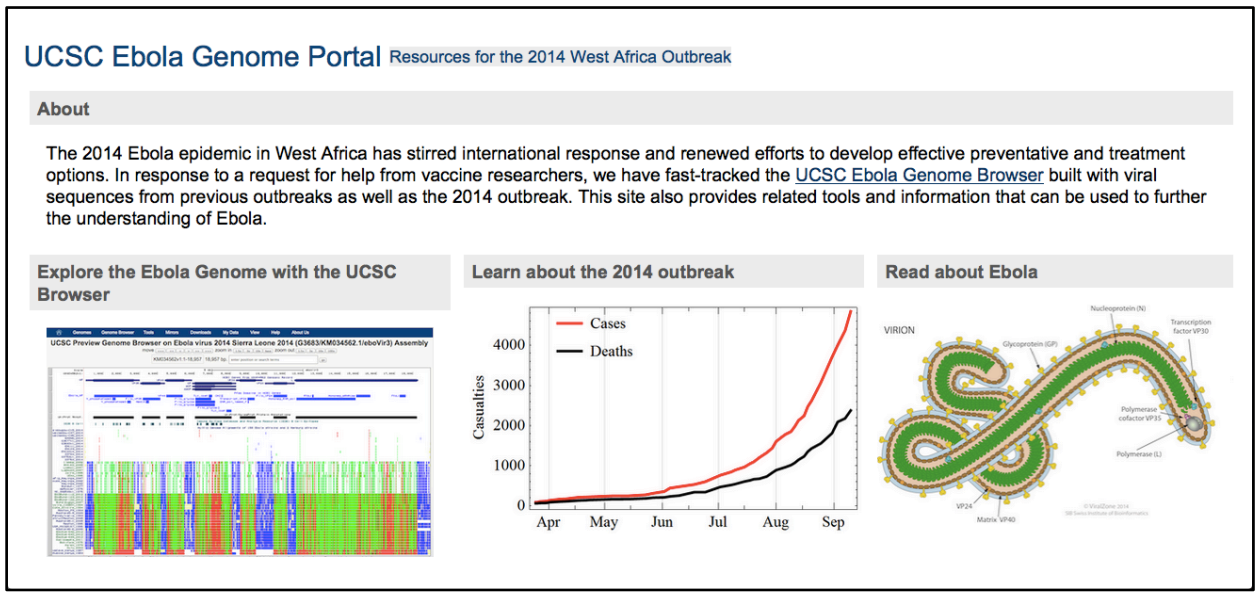


Figure B.2.2. Screenshot of top half of the UCSC Ebola Genome Portal page. In addition to linking to the Ebola Genome Browser, this page also contains links to related publications and data resources, Ebola biology, antibody resources, and related external web pages of general interest.



1.a. Increase website interactivity

- Completed the technical design, implementation, code profiling and performance tuning of infrastructure to support condensed exon-only display in the browser. We have not yet determined the UI for this feature, but intend to make it completely transparent to those who do not wish to use it (i.e. “regular” users will not be inconvenienced by this feature).
- Added a “Watson-Crick strand implementation” that allows the display of wiggle files containing values on both the positive and negative strands. The associated trackDb setting, “negateValues”, inverts the values in the wiggle file, changing the positive values to negative and vice versa. This is useful for wiggle tracks representing transcription or other activities on the Crick strand.
- Implemented the display of exon number when a user hovers the mouse over an exon in the UCSC Genes track.
- Updated the menu bar on the browser home page to use the javascript-enabled style already implemented on browser CGIs. By 6/30/15, all static pages on the Genome Browser website will display the new menu style.
- Improved image-only reload speed on FireFox browsers when many (up to 50) tracks are displayed.
- Set up logging mechanisms for analyzing track usage metrics, to determine how best to improve our user track search capabilities.
- Ran jshint code analysis tool on JavaScript source codebase, and corrected the errors and coding compliance problems detected.

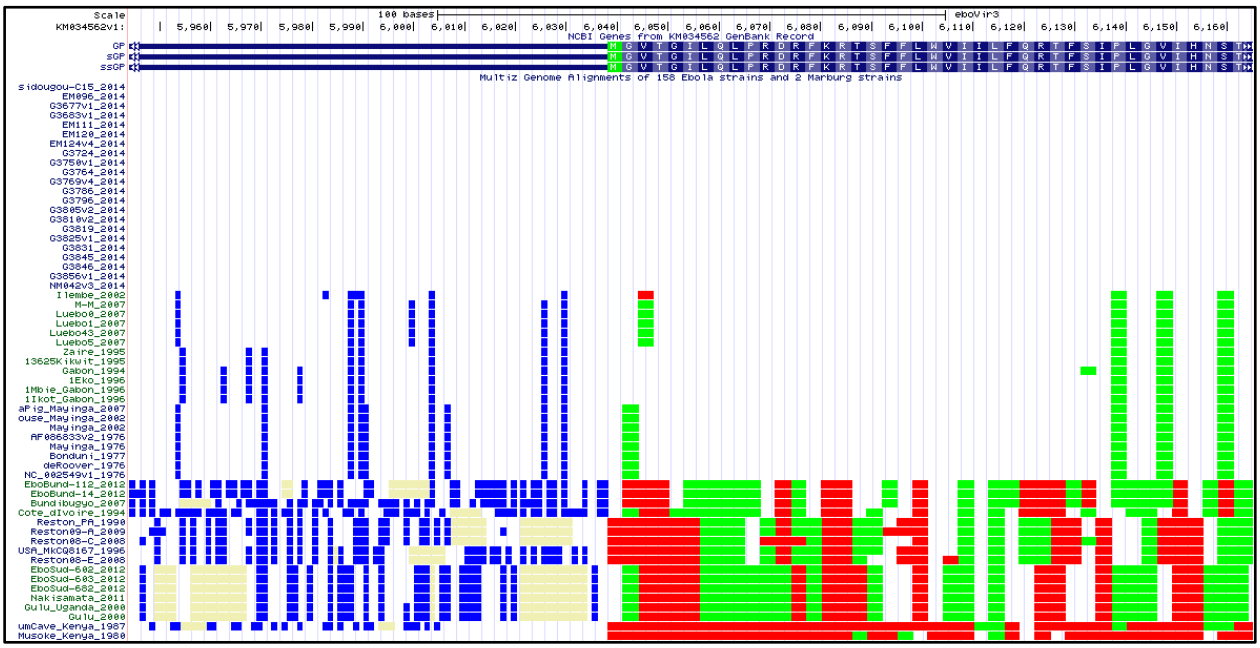
1.b. Adapt to new types of data

- Completed draft implementation to display long-distance chromatin interactions, including those across chromosomes (final implementation will be released on public website in Year 4).
- Improved and extended support for Variant Call Format (VCF) data:
    - added support for several track line settings to reduce the need for additional user configuration

  - improved the readability of the details display of certain types of VCF tabular data, such as VEP annotations

- Started adding support for Genotype Tissue Expression (GTEx) data (supplement to existing grant): initial design work on schema and loader for GTEx data tables, and initial implementation of schema and display code for GTEx tracks.
- Added new RepeatMasker visualization in collaboration with Robert Hubley.
- Added SNP-oriented MAF display with coloring by function (featured in the Ebola browser comparative alignment of 160 virus strains) (Figure B.2.3).
- Added SNP track display option to include alleles in the hgTracks item label even when there are no orthologous alleles (e.g. chimp etc. for human) to display.

Figure B.2.3. The MAF SNP mode is an easy way to quickly see differences between species in a multiple species alignment. This figure displays a portion of the gene GP in Ebola virus that includes a segment of the 5' UTR and the beginning of the protein-coding region. Red blocks are drawn when a polymorphism in a coding region results in a change in the amino acid that is generated. Green blocks are drawn when a polymorphism in a coding region results in no change to the amino acid that is generated. Blue blocks are drawn when a polymorphism is outside a coding region. This mode is currently available only on the 160-way alignment on the Ebola virus genome, but will be integrated into conservation tracks for other genomes as feasible.



1.c. Adapt to higher volumes of data

- Added BLAT support for assembly hubs and GBiB.
- Extended the functionality of track and assembly hubs:
  - Modified HubCheck program to improve early error detection in hubs before they are published
  - Added "retry hub" error message to hgHubConnect to allow disconnect of broken hubs
  - Added hubClear URL variable for dynamic hub builders
  - Added support for bigGenePred data format in track hubs
- Enabled bigBed item search on native bigBed tracks.

1.d. Enhance the security of uploaded data

- Continued to implement security best practices in codebase.
- Employed standard security measures in setting up new online web store (e.g. https, security certificate), which passed a complete security audit by the UCSC ITS security team before it was released to the public site.

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

- Extended command-line tools packaging to include new tools with updated documentation as they became available. Added 11 new tools this year: faAlign, genePredFilter, genePredToBigGenePred, getRna, getRnaPred, hgLoadOutJoined, hgSpeciesRna, hubPublicCheck, mafToSnpBed, pslPosTarget, pslScore.
- Made browser source code available via github: https://github.com/ucscGenomeBrowser.
- Created an easy-to-use package of browser bioinformatics tools (as part of the productization of Genome Browser in a Box) that are automatically installed and can be invoked from the VM (virtual machine) command line.
- Added a file of pre-calculated chromosome sizes for each genome assembly to the downloads server, for use by the many kent command line utilities that require the information, eliminating the need for utility users to calculate the sizes themselves.

Other updates to the Genome Browser website and software

Genome Browser in a Box (GBiB):

- Implemented several modifications to the code base to support the varying needs of GBiB users and mirrors.
- Fixed several bugs.
- Added a new CGI to GBiB, hgMirror, which allows users to easily download of tracks and data of interest.

Table Browser:

- Added export to GenomeSpace.
- Expanded export to GREAT to include GRCh38/hg38 data.
- Fixed several bugs.

Variant Annotation Integrator (VAI):

- Added regulatory consequences from ENCODE summary tracks.
- Added option to intersect with COSMIC.
- Added proper HGNC gene symbols for predictions based on UCSC Genes.

Beacon tool (http://genome.ucsc.edu/cgi-bin/hgBeacon):

- Added in response to a request from the Beacon Project (http://genomicsandhealth.org/our-work/current-initiatives/beacon-project) at the Global Alliance for Genomics and Health (GA4GH). The Beacon Project tests the willingness of international sites to share genetic data in the simplest of all technical contexts. This open web service is designed to be technically simple, easy to implement and to not return privacy violating information. GA4GH provides a Beacon of Beacons that queries all Beacons around the world, including ours.
- The Genome Browser Beacon tool serves answers for queries about data contained in the Leiden Open Variation Database (LOVD) and Biobase's Human Gene Mutation Database (HGMD).

Session gallery:

- Collected ideas for popular sample sessions derived from commonly asked questions on the browser user mailing list and feedback from onsite browser training workshops (sample sessions will be released on the public website in Year 4).

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

Added genome browsers for 2 new and 7 updated genomes

- New genomes:
  - bonobo (panPan1)
  - Ebola virus (eboVir3)

- Updated genome assemblies:
  - cow (bosTau8)
  - D. melanogaster (dm6)
  - pika (ochPri3)
  - rat (rn6)
  - shrew (sorAra2)
  - tarsier (tarSyr2)
  - zebrafish (danRer10)

Added 6 new multiple-alignment tracks for 5 different genome assemblies

- Human Netherlands (GRCh38/hg38): 20 species, mostly primates (human, chimp, bonobo, gorilla, orangutan, gibbon, crab-eating macaque, golden snub-nosed monkey, baboon, proboscis monkey, rhesus, green monkey, squirrel monkey, marmoset, mouse lemur, tarsier, bushbaby, mouse, dog, tree shrew)
- Human (GRCh38/hg38): 7 species (human, chimp, rhesus, mouse, rat, dog, opossum)
- Tarsier (tarSyr2): 20 species, same list as for GRCh38/hg38
- Ebola virus (eboVir3): 160 strains
- Mouse (mm10): 60 species
- Rat (rn5): 13 species (rat, mouse, guinea pig, human, chimp, rhesus, cow, dog, panda, opossum, chicken, turkey, zebrafish)

Compared multiple genome alignment programs with the possibility of switching to new, better one

- Conducted an alignment tools “bake-off” in which we compared a 100-species alignment produced using two pipelines: our current multiz method and the new cactus tool. A third method (a modified multiz pipeline) was unable to build the alignment and was eliminated. Other invited groups (such as Ensembl) were invited to participate, but were unable to or declined.
- Bake-off results were evaluated by a human curator who checked 20 randomly chosen regions on both alignments, and examined how well open reading frames were retained with an automatic genome-wide process.
- The currently used multiz process clearly excelled on both criteria, primarily due to the poor performance of cactus outside the placental mammals group, and its inability to separate paralogous regions and orthologous regions or indicate which should be considered orthologous.

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

Annotation tracks added to GRCh38/hg38 and other browsers

Table B.2.1 lists the annotation tracks added to the Genome Browser this year. We further standardized and automated the pipeline that builds and releases automatically updated tracks: it downloads the data from the source, builds the track, performs quality assurance checks, and releases it to the public site.

- Tracks specifically called out in last year's progress report and SAB meeting:
  - Segmental Dups tracks for mouse (mm10) added.
  - Centromeres annotation on GRCh38/hg38: added a Centromere Location track, but no longer intend to create track hub.
  - Illumina BodyMap RNA-seq data: dropped in favor of GTEx data.
  - Exome variant server data on GRCh37/hg19 (http://evs.gs.washington.edu/EVS/) added.
  - New version of NCBI RefSeq Genes on 4 assemblies (GRCh38/hg38, mm10, rn6, and danRer10) built directly from the NCBI mappings, rather than using the UCSC remapping pipeline. Will include RefSeq Genes with the accession XM_*, frequently requested by browser users.
- Added a new Peptide Atlas track that displays high-coverage, high-quality peptide identifications from mass spectrometry providing an additional evidence source for evaluating genomic annotations of gene activity. The track displays peptide identifications from the Aug. 2014 Human Build (433) processed by the PeptideAtlas project at the Institute for System Biology. This project collects raw mass spectrometry proteomics datasets from laboratories around the world and reprocesses them in a uniform bioinformatics workflow. The 2014 Human Build identified 1,021,823 distinct peptides covering 15,136 canonical proteins, resulting from over 400 million spectra from 971 samples. (Figure B.2.4)
- Imported selected data from the ENCODE project:
  - Added selected ENCODE2 data (the Integrated Regulation tracks) to GRCh38/hg38. The DNAse data were reprocessed from fastQs using the UCSC implementation of the analysis pipeline; the other tracks were lifted from GRCh37/hg19 to GRCh38/hg38. We also added schema and code support for generic controlled vocabularies to the browser to supplant the ENCODE-specific cv.ra.
  - Asked the ENCODE3 DCC to make a track hub of all ENCODE3 data that we can make publicly available via our Track Hub portal.

Table B.2.1. Annotation tracks released on the Genome Browser during 2014–15. Tracks that are automatically updated when new data is released are marked as "auto-update".

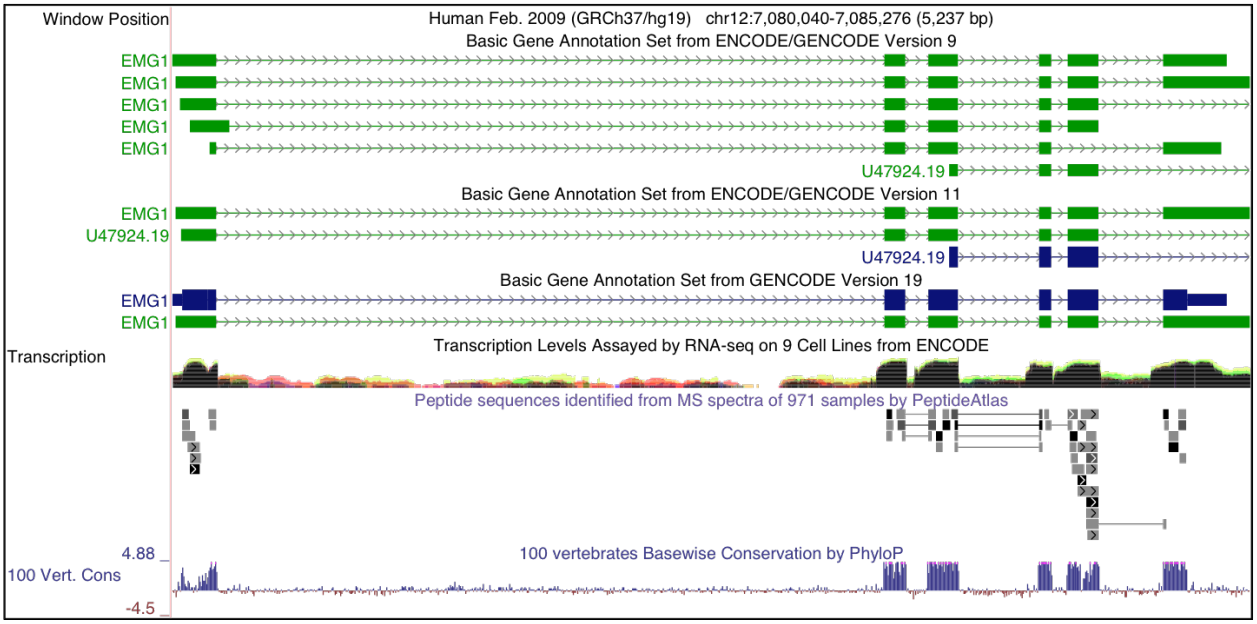
Species	Assembly	Track	Status
Human	hg38 (GRCh38)	20-species Conservation	new
		7-species Conservation	new
		Alignments of Affymetrix Consensus/Exemplars from GNF1H	new
		Alignments of Affymetrix Consensus/Exemplars from HG-U133	new
		Alignments of Affymetrix Consensus/Exemplars from HG-U95	new
		Alternative Splicing Graph from Swiss Institute of Bioinformatics	new
		Centromeres	new
		Chromosome Band (Ideogram)	new
		ClinVar Variants	new, auto-update

		Common SNPs(141), Flagged SNPs(141), Mult. SNPs(141), All SNPS(141)	new
		Common SNPs(142), Flagged SNPs(142), Mult. SNPs(142), All SNPS(142)	new
		Database of Genomic Variants: Structural Variation	new
		ENCODE Regulation	new
		ENCODE Regulation: DNase Clusters	new
		ENCODE Regulation: DNase HS	new
		ENCODE Regulation: DNase Signal	new
		GENCODE Version 20	new
		GeneReviews	new, auto-update
		GNF Atlas 2	new
		GRC Incident Database	new, auto-update
		GRC Patch Releases 1, 2 & 3	new
		Locus Reference Genomic (LRG) Regions & Transcripts	new
		NCBI RefSeq Genes	new
		NHGRI Catalog of Published Genome-Wide Association Studies (GWAS)	new, auto-update
		Online Mendelian Inheritance in Man (OMIM) Genes & Phenotypes	auto-update
		Retroposed Genes V9	new
		Segmental Dups	new
		Swiss Institute of Bioinformatics Gene Predictions	new
		Zebrafish Chain/Net	new
	hg19 (GRCh37)	Clinical Genome Research (ClinGen) – was ISCA	auto-update
		ClinVar Variants	auto-update
		ClinVar Variants	update
		Common SNPs(141), Flagged SNPs(141), All SNPS(141)	update
		Common SNPs(142), Flagged SNPs(142), Mult. SNPs(142), All SNPS(142)	update
		Database of Genomic Variants: Structural Variation	update
		DECIPHER: Chromosomal Imbalance and Phenotype in Humans	auto-update
		DNase Clusters V3	update
		Exome Aggregation Consortium (ExAC) Variants and Calling Regions	new
		GeneReviews	auto-update
		GRC Incident Database	auto-update
		Human Gene Mutation Database	update
		Leiden Open Variation Database (LOVD) Public Variants	auto-update
		Locus Reference Genomic (LRG) Regions & Transcripts	update
		Master DNaseI HS	new
		Neandertal Methyl, Denisova Methyl	new
		NHGRI Catalog of Published Genome-Wide Association Studies (GWAS)	auto-update
		Online Mendelian Inheritance in Man (OMIM) Genes & Phenotypes	auto-update
		Peptide Atlas	new
		Pika Chain/Net	new
		Ribosome Profiling from GWIPS-viz	new
		Shrew Chain/Net	new
		UniProt	new

	hg18	Database of Genomic Variants (DGV): Structural Variation	update
		GeneReviews	auto-update
		NHGRI Catalog of Published Genome-Wide Association Studies (GWAS)	auto-update
		Online Mendelian Inheritance in Man (OMIM) Genes & Phenotypes	auto-update
Mouse	mm10	60-species Conservation	update
		Cow Chain/Net	update
		FaceBase 24 Sample Types Averaged	new
		GENCODE Version M3	update
		GENCODE Version M4	update
		NCBI RefSeq Genes	new
		Retroposed Genes V6	update
		Shrew Chain/Net	update
		UCSC Genes	new
	mm9	FaceBase 24 Sample Types Averaged	new
Chicken	galGal4	Quantitative Trait Loci from animalQTLdb	new
Cow	bosTau7	Quantitative Trait Loci from animalQTLdb	new
Drosophila	droPer1	D. melanogaster Chain/Net	new
	droSec1	D. melanogaster Chain/Net	new
	droSim1	D. melanogaster Chain/Net	new
Ebola Virus	eboVir3	160-species conservation phastCons/phyloP	new
		Immune Epitope Database and Analysis Resource (IEDB) B-Cell Epitopes	new
		Immune Epitope Database and Analysis Resource (IEDB) B-Cell Epitopes with Negative Assay Result	new
		Immune Epitope Database and Analysis Resource (IEDB) Curated T-Cell Epitopes, MHC Class I	new
		Immune Epitope Database and Analysis Resource (IEDB) Curated T-Cell Epitopes, MHC Class II	new
		Immune Epitope Database and Analysis Resource (IEDB) HLA binding predictions Tracks	new
		muPIT - Mapping Genomic Positions on Protein Structures	new
		NCBI Genes from KM034562 GenBank Record	new
		Pfam Domains in NCBI Genes	new
		Protein Data Bank (PDB) Sequence Matches	new
		UniProt/SwissProt Annotations	new
Horse	equCab2	Cow Chain/Net	new
		Quantitative Trait Loci from animalQTLdb	new
Mosquito	anoGam1	D. melanogaster Chain/Net	new
Pig	susScr3	Quantitative Trait Loci from animalQTLdb	new
Rat	rn5	13-species Conservation	new
	rn6	NCBI RefSeq Genes	new
Sheep	oviAri3	Chromosome Band (Ideogram)	new
		Quantitative Trait Loci from animalQTLdb	new
Tarsier	tarSyr2	20-species Conservation	new
Zebrafish	danRer10	NCBI RefSeq Genes	new
Several assemblies		Ensembl Genes v76, v78, v79	update
		Unmasked CpG	new
Every assembly		GenBank Updates (e.g. RefSeq Genes, ESTs, RNAs)	auto-update

Figure B.2.4. New PeptideAtlas track. In this example from the EMG1 locus, one can see successive improvements to the GENCODE gene annotation attributable to incorporation of peptide evidence. In 2012 Banfai et al. identified, using peptide evidence, 69 misannotated loci. The 2012 GENCODE (V11) made use of these results with the inclusion of a partial

transcript U47924. The most recent GENCODE (V19) corrected the locus once again, adding a first translated exon. Here you can see the confirmatory peptide evidence along with RNA-seq and evolutionary conservation.



New Track and Assembly Hubs added to public hubs page

- Have 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.
- Added links to 12 new hubs that can be accessed from the Genome Browser (Table B.2.2).

Table B.2.2. Public track and assembly hubs newly released on the Genome Browser website in 2014-15. As of May 2015 we linked to a total of 30 public hubs.

Hub Description	Lab	Assembly
RNAseq data across human brain development by age group	Lieber Institute for Brain Development, Baltimore, MD	hg19
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	Fenyo Lab, NYU School of Medicine	hg19
Evidence summaries and provisional results for the new Ensembl Regulatory Build	Ensembl, Hinxton, UK	hg38, hg19
RIKEN FANTOM5 CAGE clusters by RECLU	RIKEN Center for Life Science Technologies, Japan	hg19, hg18
DNA methylation and low methylation annotations for human tissues and cultured cells	LaSalle Lab, UC Davis	hg19, hg18
RIKEN FANTOM5 Phase 1 data	RIKEN Center for Life Science Technologies, Japan	hg19, mm9
Evolutionary protein-coding potential as measured by PhyloCSF	Kellis Lab, MIT/The Broad	hg38, hg19, mm10
FaceBase Hub	FaceBase Consortium	hg19, hg18, mm10, mm9
Genome issues and other features	Sanger & Genome Reference Consortium	hg38, hg19, mm10, mm9, danRer7
Non-coding RNAs from the Rfam database	EMBL-EBI	hg38, mm10, ce10, galGal4, ci2, danRer7, dm6, sacCer3

Burgess Lab Zebrafish Genomic Resources	Burgess Lab, NHGRI	danRer7
Croc, Bird, and Archosaur	Haussler Lab, UCSC	45 different species

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

UCSC Genes

- Human genome: rather than update the UCSC Genes set for the human genome, we refocused our efforts on transitioning to GENCODE Genes as our default gene set (below).
- Mouse genome: the updated gene set on the mm10 assembly will be released to the public site in Jun. 2015.

GENCODE Genes

- Evaluated GENCODE Genes for suitability as primary gene set on human genome assembly, and have decided to switch to this annotation set (GENCODE v22) for our next release of a default gene set on the GRCh38/hg38 human assembly. Targeted public release: Jun. 2015.
- In conjunction with the switch to GENCODE Genes, we have updated several of the underlying databases, including the UniProt, proteome, and GO databases.

Training and Outreach (supplement)

Expanded on-site training to approximately double the level of the 2013-14 program

We hired Dr. Pauline Fujita on 7/1/14 for the position funded by the administrative supplement. After an initial ramp-up period, she took on the extra on-site workshops we had the bandwidth to add with a second trainer, and was active in implementing the YouTube channel while the workshop logistics were scheduled. Dr. Fujita is now fully engaged in our training program.

During the period 4/30/14 to 4/30/15, our onsite training program reached more than 2,500 scientists at 30 talks in 25 locations, and OpenHelix gave 2 additional browser-related workshops. We have 13 more workshops scheduled before 6/30/15, including a UCSC campus visit by a group of PhD students from Wageningen University, Netherlands (Table B.2.3). The typical presentations are technical, focusing on the best use of our software in the laboratory and the clinic, although some are directed at non-scientists interested in a wide range of topics, including genomics. Surveys indicate that our audiences are dominated by graduate students and postdocs, but we also attract a significant number of PIs.

By 6/30/15, we will have given eight presentations in the UCSC local area that reached 370 people, e.g. our participation in the Epic Genetics Days at the Tech Museum of San Jose, in conjunction with the NHGRI traveling exhibit, *Unlocking Life's Code*.

Our training seminars have been well received, frequently leading to invitations for future workshops. Our post-workshop surveys indicate that even regular users discovered new information about the Browser's functionality, and workshop feedback helps determine topics for our videos. Onsite training classes 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 develop a browser that works locally, eventually leading to the creation of Genome Browser in a Box.

Our expanded training page (http://genome.ucsc.edu/training/index.html) includes links to videos, an announcement about our on-site workshop program, and a list of upcoming

workshops. We now list the locations of upcoming appearances to facilitate attendance by people from other institutions.

OpenHelix continues to maintain 3 online tutorials about the Genome Browser, 2 of which were updated this year. Views of their Genome Browser videos have increased by nearly 30% in the past year. They also maintain and distribute Quick Reference Cards (QRCs) on the Genome Browser and Table Browser.

Table B.2.3. Genome Browser on-site training workshops for 2014-15, including Apr.-Jun. 2014 workshops not covered in previous progress report.

Date	Location	Attendees
Apr. 2014	Bio-IT, Boston, MA	20 + 55
May 2014	Los Altos Morning Forum, CA	400
	Frontiers in Reproduction. Woods Hole, MA	23
Jun. 2014	ESHG Milan encode workshop, Italy	70
	ESHG Milan Browser workshop, Italy	280
	Medical Genetics, Universita Cattolica del Sacre Coure, Rome, Italy	40
Jul. 2014	IRCM, Montreal, Canada	32
Sep. 2014	QUT, Brisbane, Australia	45
	U New South Wales, Sydney, Australia	95
	Peter MacCallum Cancer Centre, Melbourne, Australia	70
	Harry Perkins Cancer Centre, Perth, Australia	90
	Telethon Kids, Perth, Australia	20
	U of Adelaide, Australia	65
Oct. 2014	ASHG, San Diego, CA	120
	San Jose State University, CA	40
Nov. 2014	IRDiRC, Shenzhen, China	100
Dec. 2014	U Pittsburgh, PA	75
	U Pittsburgh, PA	40
	U Cincinnati, OH	550
Jan. 2015	PAG, San Diego, CA	50
	Mt. Sinai Medical Center, NYC (presented by OpenHelix)	30+30
Feb. 2015	Bio-IT Tri-Conference, San Francisco, CA	10
	Bio-IT Tri-Conference, San Francisco, CA	75
	Bio-IT Tri-Conference, San Francisco, CA	32
Mar. 2015	Cedars-Sinai, LA, CA	48
	Cedars-Sinai, LA, CA	21+14
	Radboud U Med Center, Nijmegen, Netherlands	100
Apr. 2015	Erasmus U Med Center, Rotterdam, Netherlands	16
	Erasmus U Med Center, Rotterdam, Netherlands	15
	JAX Lab, Bar Harbor, ME	40
	U Maine, Orono, ME	15
	Bio-IT, Boston, MA	15
	Bio-IT, Boston, MA (presented by OpenHelix)	7
May 2015	UCSC (Wageningen visit)	TBD
	Sanford-Burnham Institute, Orlando, FL	TBD
	Spelman College, Atlanta, GA	TBD
	U Georgia, Athens, GA	TBD
Jun. 2015	U Nevada, Reno, NV	TBD
	U North Dakota, Grand Forks, ND	TBD
	ESHG, Glasgow, Scotland	TBD
	Haukeland Hospital, Bergen, Norway	TBD
	H Lee Moffitt Cancer Center, Tampa, FL	TBD
	ENCODE Users' Meeting, Potomac, MD	TBD

Recorded and released short video clips detailing specific browser tasks on YouTube

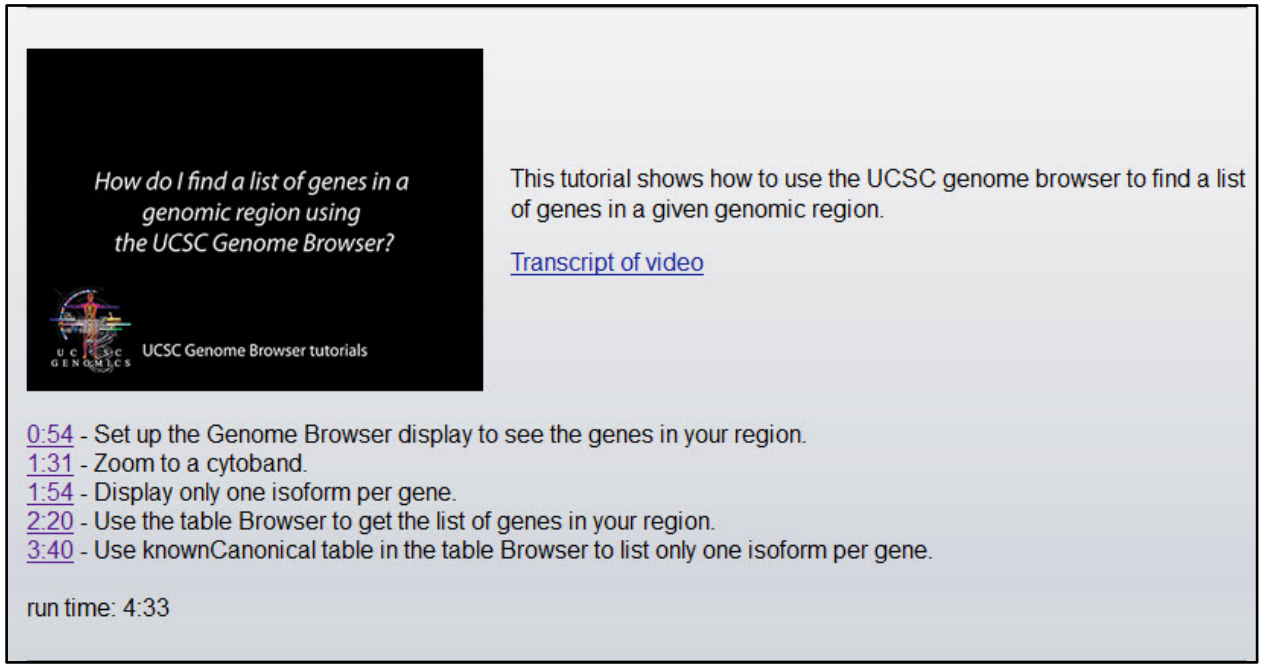
We released the Genome Browser YouTube channel (http://bit.ly/genomebrowserYoutube) in Jan. 2015. Currently it offers six videos, 4-9 minutes in length, with 121 subscribers and 3100 views, for more than 9000 minutes of viewing time.

The videos supplement the existing OpenHelix online introductory tutorials, offering instruction in specific tasks not immediately obvious to the casual user. Topics are guided by user experiences and feedback from workshops, and frequently asked questions on the browser mailing list:

- Finding a list of genes in a region
- Finding exon numbers
- Finding SNPs in a gene
- Finding SNPs upstream from a gene
- Finding which table belongs to a data track
- Identifying codon numbers in a gene

Videos are indexed with 5-10 signposts, outlining the steps required to accomplish the stated goal and providing links to let users quickly jump to the desired section of the video. The indexed steps summarize the solution to reach the goal. (Figure B.2.5). Each video is accompanied by a transcript for the deaf and for those with impaired ability to comprehend spoken English.

Figure B.2.5. The index screen for a video explaining how to find a list of genes in a region. Each of the key steps is listed, along with a link into the video.



Accumulated program income to support future training

Since the advent of our signup system for workshops, we have collected more than expenses and have accumulated $2000 in program income. Workshops continue to operate primarily on a “host pays” basis, which occasionally generates program income. Because of the limited budgets of many organizations, it is often infeasible to bill host institutions for the cost of the trainer salary; some institutions are unable to cover even the modest cost of plane, hotel, and a

flat fee of up to $1200. Workshops at national meetings are typically given at our expense though we often obtain a waiver of registration fees, limiting expenses to cost of travel.

Genome Browser Usage

During the past year, we averaged 56 million **hits** per month for the main site (44 million) and genome-euro (12 million) combined. In Mar. 2015 we turned on Google Analytics on the public Genome Browser site, and are gathering statistics (minus robots) that we will report next year.

Citations in the literature

The browser continues to be cited in the formal reference section of scientific papers, but it is increasingly mentioned only in the text. Table B.2.4 shows formal references to Genome Browser-related papers, a 22% increase in citations (as indexed by Google Scholar) over the previous year.

Table B.2.4. Formal references to our papers in the past year, as indexed by Google Scholar.

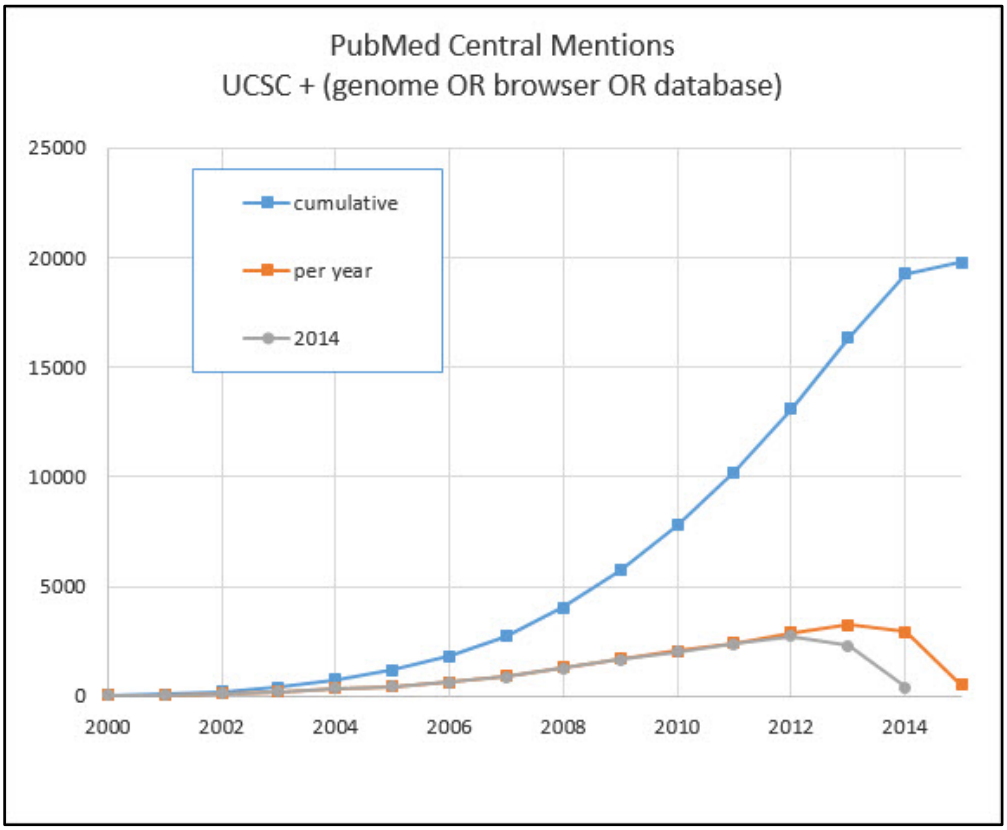
Topic	Author, Year	Google 2015
BLAT	Kent, 2002	4539
Genome Browser	Kent et al., 2002	3658
Conservation	Siepel et al., 2005	1763
Browser database	Karolchik et al., 2003	1343
Threaded Blockset Aligner	Blanchette et al., 2004	910
Genome Browser update 2011	Fujita et al., 2011	824
Table Browser	Karolchik et al., 2004	714
Genome Browser update 2014	Karolchik et al., 2014	682
Chain/Nets (evolution's cauldron)	Kent et al., 2003	526
Genome Browser update 2010	Rhead et al., 2010	505
Genome Browser update 2008	Karolchik et al., 2008	477
Genome Browser update 2006	Hinrichs et al., 2006	417
Genome Browser update 2013	Meyer et al., 2013	406
Known Genes	Hsu et al., 2004	336
Genome Browser update 2009	Kuhn et al., 2009	322
Genome Browser update 2007	Kuhn et al., 2007	280
Genome Browser extensions & updates	Dreszer et al., 2011	256
ENCODE resources 2010*	Rosenbloom et al., 2010	189
28-way alignment	Miller et al., 2007	179
ENCODE resources 2013 - 5-yr update*	Rosenbloom et al., 2013	175
ENCODE resources 2012*	Rosenbloom et al., 2012	146
Current Protocols	Karolchik et al., 2009	125
ENCODE resources 2011*	Raney et al., 2011	118
BigWig and BigBed	Kent et al., 2010	102
Archaeal Browser*	Schneider et al., 2006	91
Browser and associated tools	Kuhn et al., 2013	86
ENCODE resources 2007*	Thomas et al., 2007	75
Gene Sorter	Kent et al., 2005	61
Proteome Browser	Hsu et al., 2005	50
Browser Tutorial	Zweig et al., 2008	38
Current Protocols	Karolchik et al., 2011	32
Track Hubs	Raney et al., 2013	26
Understanding Genome Browsing	Cline et al., 2009	21
Biotechnology Annual Review	Mangan et al., 2008	17

Comparative Genomics with Browser	Karolchik et al., 2008	17
Current Protocols 2009	Mangan et al., 2009	15
Genomic Data Resources	Lathe, et al., 2008	14
Genome Browser update 2015	Rosenbloom et al., 2015	4
Comparative Assembly Hubs	Nguyen et al., 2014	3
GBiB	Haeussler et al., 2014	1
Current Protocols 2014	Mangan et al., 2014	1
Ebola Portal	Haeussler et al., 2014	1

* Resources not directly or fully funded by main Genome Browser grant, but featuring Genome Browser

As of Apr. 2015, the number of mentions in literature indexed in PubMed Central (PMC) is just under 20,000. These are estimated by mining PMC for the text string (UCSC AND (genome OR browser OR database). When collecting these statistics, it is necessary to account for the fact that PMC contains only papers released for full-text public access, and there is often a significant lag time between when a paper is released and when it is archived in PMC. For example, an additional 950 papers published in 2013 that mention our papers reached PMC in the past year (Figure B.2.6).

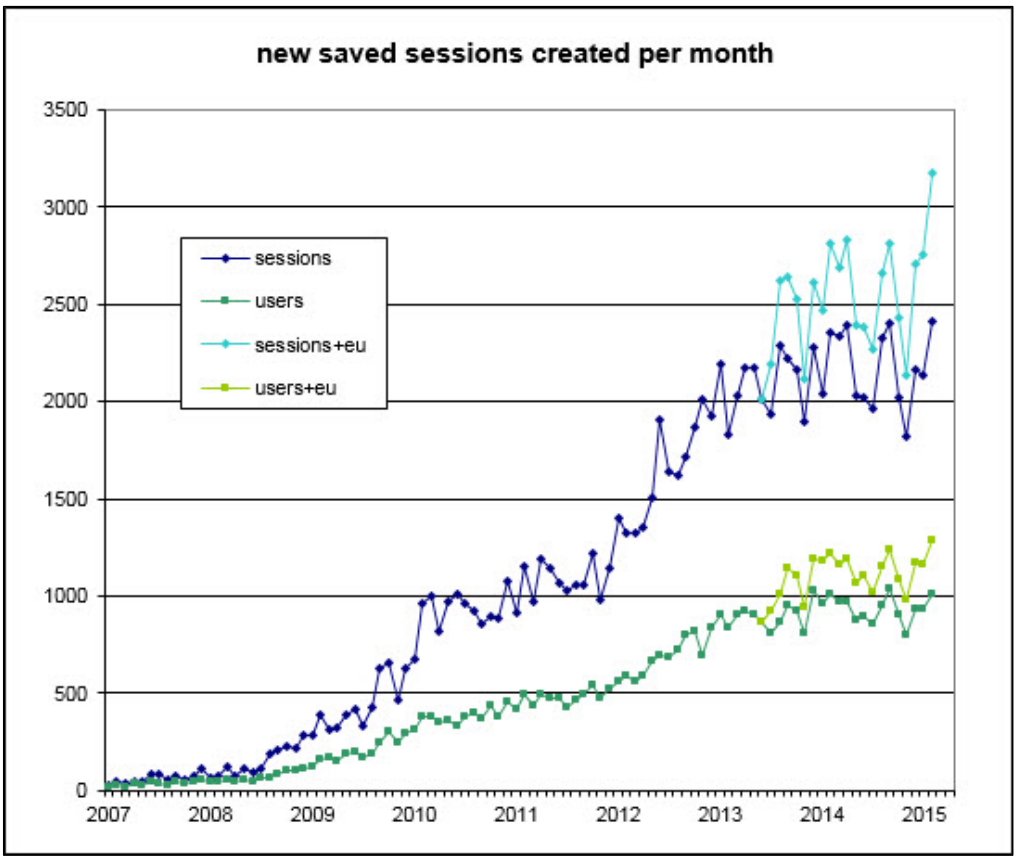
Figure B.2.6. Mentions of Genome Browser in the literature, mined from PubMed Central. Journal embargos and delays in making articles available in PubMed account for the dip in 2013-15.



UCSC Genome Browser Sessions usage

The sessions utility is a popular tool that is growing in usage, with consistently more than 1000 users creating 2500 fresh sessions per month (Figure B.2.7). Because this utility is such a powerful way to capture a view of the Browser and share it with others, we use it extensively in our trainings, which may add to its visibility among our users.

Figure B.2.7. The number of new sessions created per month has grown steadily since the inception of the sessions utility in 2007.



UCSC Genome Browser Track Hub usage

Track hub usage has increased steadily. The ENCODE portal at Stanford generates track hubs on the fly. Excluding those, more than 1000 new track hubs were created per month (including our genome-euro mirror site) from approximately 200 independent IP addresses in 2015.

We encourage users to host their own track hubs to reduce the impact on our storage and to allow them greater control of their own content. Many of our more biologically oriented users either do not have access to http accessible space, or do not realize that they do. Only about 10% of workshop attendees claim to have access, even in places where we know access is available.

Website and hardware infrastructure

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

- Upgraded the hardware for our development file system (the hive), in part with the goal of allowing us to add more space.
- Applied ongoing security updates to our systems in light of various revealed vulnerabilities (heartbleed, glibc, etc.).
- Added space to our customdb backup server, and are bringing in some additional hardware to further extend it.
- Turned on Google Analytics to better assess usage patterns and bottlenecks.
- Applied more restrictive bot policies via robots.txt files.
- Added bottleneck service to the genome-euro machine to stop programmatic usage.

This year, for the first time, we recorded zero downtime for the UCSC Genome Browser site, as well as genome-euro! This was due, in large part, to the skills of our systems administrators, who put various systems into place that can be switched in/out while the site is still live.

Genome Browser Scientific Advisory Board

The UCSC Genome Browser SAB members:

- Aravinda Chakravarti -- Johns Hopkins University
- Joe Gray -- Oregon Health Sciences University
- Tim Hubbard -- University College, London
- Mary-Claire King -- University of Washington
- Robert Waterston -- University of Washington
- Barbara Wold -- Caltech University

The SAB convened with the Genome Browser management team in Mar. 2015, with NHGRI Program Director Adam Felsenfeld attending remotely. They were generally pleased with the Genome Browser team's efforts to keep up with new data, add new features, and accommodate the needs of our users, but admonished us to maintain "constant vigilance": the world of genomics is changing rapidly, and consequently we must keep our eyes on the new tools being built and integrate those that make sense into the browser.

The SAB gave specific feedback in these areas:

- **New feature in development: multiple vertical slices view.** Our preliminary use cases are: exon-only view, patches/haplotypes, and user-selected regions. The SAB requested that we allow users to collapse together chromatin-interaction sites (e.g. results of 3C, Hi-C data). We will certainly be able to accommodate this use case, as it is essentially a special case of the user-selected regions in the preliminary spec. The SAB suggested that we consider differentiating slices by color to make it easier to spot discontinuities in genome sequence.
- **4D views.** Increasingly more data sets include a time sequence. UCSC should monitor how others are displaying these data, and adopt solutions that are feasible in the browser. For example, we should keep an eye on the 4D Nucleome Network project. A stack of time points is useful, but we should consider an animated gif track that scrolls through time points in one data track. We may want to invite someone from the visualization community to join our SAB, if we can find the right person.
- **GTEx.** The SAB found our preliminary draft of a browser display model confusing, saying it breaks the current browser paradigm of gene location as the x-axis. It displayed a series of bar graphs, one bar for each tissue showing the expression level of the entire gene at that location, but the graphs extended left and right with the x-axis no longer representing the coordinates on the genome. Based on SAB input, we are doing another design iteration to make the display more intuitive and maintain a connection to the start coordinate of the gene. The SAB recommended that UCSC get visualization input from subject-matter experts on these types of data, and consider attending GTEx meetings and jamborees. Since the SAB meeting, Jim & Kate have worked on the design for the display, and hope to have addressed most of the SAB's concerns.
- **Data hubs.** The SAB was surprised by how many data hubs are being loaded into the browser and noted that they are being embraced by many other projects. However, the SAB emphasized that -- as the data hub inventors and "owners" -- we must become the standards group and provide a stable, predictable format. All hubs will then have to comply if they wish to be displayed on the various visualization tools that support the standards (currently the UCSC, Ensembl, Dalliance, and Roadmap Epigenomics browsers). The SAB requested that we make hubs queryable by everyone, and suggested that we put a wrapper around track hubs to serve them through an API. We would have to convert the formats of much of the browser data (for example, from BED to bigBed) to make the data track-hub accessible in this way.

- **Migrating users from hg19 to hg38.** The SAB feels we should strongly encourage our users to migrate to the newest human assembly, hg38 (GRCh38). We should consider allowing users to dynamically jump from hg19 to hg38, with the data they were viewing made available in the new assembly. The computing challenges of this are formidable, however, especially as the number of browser tracks grows.
- **Training & outreach.** The SAB liked the YouTube browser training videos, especially the transcripts and indexing that allow users to gain access to key points within. They recommended that we create more videos and more online resources, striving to get at least 50% of the training in commonly used operations online. We might also consider creating a Coursera-like long course. The SAB noted that onsite training is not scalable, and therefore should be targeted at people who have already gone through a basic set of training videos, although they acknowledged the value of onsite training. Developing a robust online training curriculum will entail a high cost, and online videos will become outdated as the browser evolves. Ross pointed out that a Grateful Dead concert is still more enriching than a recorded session, and similarly there is a place for in-person workshops. Also, we learn from being in the room with our users – what they need, and what we can improve.

Table B.2.5. Summarizes the key recommendations of the SAB and our estimated abilities to implement them within this grant

Recommendation	**Implement?**
Track Hubs: (a) set the standards, (b) build validation tools, (c) put a wrapper around them to serve them through API (d) create better search mechanism	Year 4
Allow users to post their sessions for others to see	Year 4
Implement 4D views (with 'time' as 4th dimension)	Year 5
Allow user to choose which isoform they want to see or hide. Implement voting mechanism to vote transcripts up/down.	Year 5
Enable dynamic liftOver of annotation track data from one assembly to another	No plans to implement

Training and outreach program

See section B.2 for a description of our Training and Outreach program accomplishments for this year.

Additional staff presentations

Supplementing the outreach program, our staff gave talks at several venues and presented posters at scientific meetings. (Table B.4.1)

Table B.4.1. Presentations made by Genome Browser staff during 2014-15, in supplement to our Training and Outreach program.

Date	Location	Attendees	Presenter
Talks			
Jul. 2014	COSMOS program, UC Santa Cruz, CA	180	Robert Kuhn
Aug. 2014	ENCODE Analysis Working Group (online call)	40	Jim Kent
Sep. 2014	UC San Francisco, CA	30	Jim Kent
Nov. 2014	CNRS Gif-sur-Yvette, France	12	Max Haeussler
Mar. 2015	EFOR, Paris, France	100	Max Haeussler
Apr. 2015	Epic Genetics, Tech Museum, San Jose, CA	100	Kate Rosenbloom
Posters			
Sep. 2014	Genome Informatics, Hinxton, UK	----	Galt Barber
Oct. 2014	ASHG, San Diego, CA	----	Brian Raney
Mar. 2015	Genome 10K, UC Santa Cruz, CA	----	Brian Raney
Mar. 2015	ACMG, Salt Lake City, UT	----	Robert Kuhn
May 2015	Biology of Genomes, CSHL, NY	----	Angie Hinrichs
Jun. 2015	UC-Wide Bioengineering Symposium, UCSC	----	Matt Speir

Journal peer reviews completed by staff

- BMC Bioinformatics (3)
- Nucleic Acids Research (3)
- Oxford Bioinformatics (1)
- PeerJ (1)

Scientific advisory board members among staff

- Jim Kent, WormBase
- Kate Rosenbloom, Human Proteome Project
- Brian Raney, EBI project team constructing track hubs from proteomic data

Student opportunities and participation

During the past year, the Genome Browser project provided research training and professional development opportunities for two undergrad students:

- Chris Eisenhart: Ebola Genome Portal, lifted several tracks from GRCh37/hg19 to GRCh38/hg38, developed several small command-line utilities programmed in C.
- Parisa Nejad: Ebola Genome Portal, various other small tasks.