PROGRESS REPORT SUMMARY

A. Specific Aims

The grant proposal was funded with four original aims. In Aug. 2013 we received a 2-year supplement to this grant, which is covered by newly added Aim 5.

1. Develop, maintain, and extend software for web-based display and command-line-driven analysis of genomics resources.
2. Build genome browsers and comparative genomics resources for species of biomedical interest.
3. Import data from the scientific community that help interpret the functions of various human genome regions into the UCSC databases.
4. Build high quality gene sets on the human genome and selected model organism genomes.
5. Provide training and support to the user community through online and in-person activities.

B. Studies and Results

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

We added the following major features and enhancements to the Genome Browser tools in the past year:

- *"Genome Browser-in-a-Box" (GBiB)* – In May-Jun. 2014 we plan to release a virtual machine version of the Genome Browser with a smaller memory footprint and simpler installation requirements that can be easily installed on a user's own laptop. The GBiB package uses the Oracle VirtualBox open source virtualization software, and includes the UCSC Genome Browser software, all required utilities, and a basic set of human genome annotation data. Additional annotation data can be loaded on demand from UCSC via the Internet or can be downloaded to the local machine for faster access. Individuals can also view and manipulate their own local data in the Genome Browser through the use of custom tracks, thus allowing them to view private or very large data files on their own hard disk without the need to upload the files to the UCSC server. (This feature was not part of the original grant proposal, but was developed in response to the needs of our users.)
- *Variant Annotation Integrator (VAI)* – In Jul. 2013 we released the preliminary version of the VAI software tool (http://genome.ucsc.edu/cgi-bin/hgVai), briefly mentioned in last year's progress report. The VAI annotates variant calls with predicted functional effects on protein-coding genes and regulatory regions. Unlike other variant call prediction tools, the VAI is not restricted to one or two sources of gene annotations and a fixed set of additional annotation sources, but instead offers broader choices from the UCSC database. Given a set of variants uploaded as a custom track, the VAI returns the predicted functional effect (e.g., synonymous, missense, frameshift, intronic) for each variant. It can optionally add several other types of relevant information, such as the dbSNP identifier (if applicable), protein damage scores for missense variants from the Database of Non-synonymous Functional Predictions (dbNSFP), and conservation scores computed from multi-species alignments. The VAI also offers filters to help narrow down results to the most interesting variants. Since the initial release of the VAI, we have added support for assembly hubs, additional input options (artificial example variants, rs# IDs), and large deletions, and have implemented performance optimizations.
- *Assembly data hubs* – We have expanded the functionality of track data hubs to address the increasing need for researchers to annotate sequence for which UCSC does not host an annotation database. Both the underlying reference sequence as well as the data tracks that annotate that sequence are included in the assembly data hub, which may then be viewed in the Genome Browser in a manner similar to natively hosted assemblies. As with track data hubs, we offer public access for hubs that are of general interest to the research community. As of Apr. 2014, we offer 4 public assembly hubs on the Genome Browser website featuring Drosophila, 3 plants, and a comparative assembly hub with nearly 60 species of bacteria.

- *Centromere representation* -- Debuting with the release of the hg38/GRCh38 human assembly, the large megabase-sized gaps that represented centromeric regions in previous assemblies were replaced by sequences from centromere models created by Miga et al., 2014 (see References), using centromere databases developed during her work in the Willard lab at Duke University and analysis software developed while working with the UCSC Genome Browser group. The models, which provide the approximate repeat number and order for each centromere, will be useful for read mapping and variation studies.

The following Aim 1 tasks originally listed in the grant proposal for this year were dropped as a result of our 10% funding cut:

- Implement right clicks and sorts on column headers in Gene Sorter
- Security expert attempts to break into site, fixes problems found
- Dynamic conversion of genome browser tracks into Gene Sorter columns
- Evaluate network visualization tools

1.a. Increase website interactivity

We made these improvements to the user interface, navigation, interactivity, and performance of the Genome Browser website:

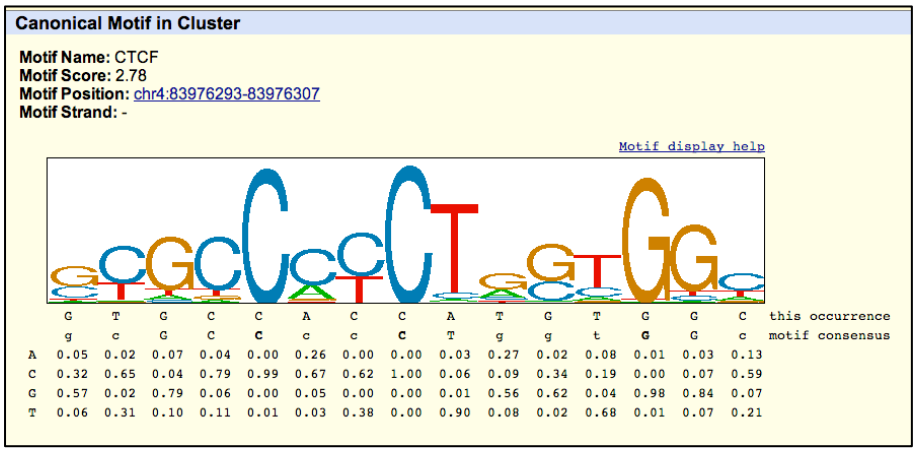
- Added region highlighting on the tracks image (including highlighting by item/gene and drag-highlighting)
- Implemented auto-scroll to top of page after a refresh on the browser tracks image page
- Improved speed and performance of track and file searches
- Fixed navigation instances on the browser tracks image in which the Back button did not correctly return the image to the previous chromosomal position
- Improved user interface on the configuration page
- Accommodated certain characters in FTP and HTTP URLs and fixed some problems with FTP URL-encoding
- Improved some network defaults for TCP connections

1.b. Adapt to new types of data

We added these enhancements and bug fixes to support new and existing data types:

- Added a browser display feature to highlight transcription factor motifs in transcription factor clusters and to show motif information, including sequence logos, on the track item details page. This display is currently featured in the human hg19/GRCh37 ENCODE Regulation Transcription Factor track (Figure 1), but may be broadly applied to several different tracks and track types.
- Added browser support for data in Hierarchical Alignment (HAL) format, which uses a graph-based structure to efficiently store and index multiple genome alignments and ancestral reconstructions. To visualize HAL alignments in the browser, we developed a new “snake” track display type that provides a way to view sets of pairwise gapless alignments that may overlap on both the chosen genome (reference) and the query genome, and shows various types of genomic variations such as insertions, substitutions, and duplications (Figure 2). The snake track type is showcased in the *E. coli* Comparative Genomes hub (Aim 3).
- Expanded wiggle tracks to support stacked wiggles, line wiggles, and optimizations (Figure 3)
- Added support for Exome Sequencing Project (ESP) exome variant VCF files (which were non-standard) on browser track details page
- Fixed Table Browser support for some hyperlink output options and VCF track intersection options
- Fixed Genome Browser color and display bugs on the tracks graphic

Figure 1. Graphic of a canonical motif from the details page for a CTCF transcription factor binding site, showing the motif alignment, position weight matrix and sequence logo. The ENCODE Regulation Transcription Factor track on human assembly hg19/GRCh37 includes this feature, using motif locations from the ENCODE Factorbook project.



Canonical Motif in Cluster

Motif Name: CTCF
Motif Score: 2.78
Motif Position: chr4:83976293-83976307
Motif Strand: -

Motif display help

	G	T	G	C	C	A	C	C	A	T	G	T	G	G	C	this occurrence
	g	c	G	C	**C**	c	c	**C**	T	g	g	t	**G**	G	c	motif consensus
A	0.05	0.02	0.07	0.04	0.00	0.26	0.00	0.00	0.03	0.27	0.02	0.08	0.01	0.03	0.13	
C	0.32	0.65	0.04	0.79	0.99	0.67	0.62	1.00	0.06	0.09	0.34	0.19	0.00	0.07	0.59	
G	0.57	0.02	0.79	0.06	0.00	0.05	0.00	0.00	0.01	0.56	0.62	0.04	0.98	0.84	0.07	
T	0.06	0.31	0.10	0.11	0.01	0.03	0.38	0.00	0.90	0.08	0.02	0.68	0.01	0.07	0.21	

Figure 2. Screenshot of a "snake" track display in the human MHC region using the DBB haplotype as reference. In this figure, the hg19/GRCh37 reference, as well as the MCF and MANN haplotypes (bottom 3 lines), show an inversion associated with a 2Kbp deletion relative to the chimp (panTro3, top), which is assumed to be the ancestral state.

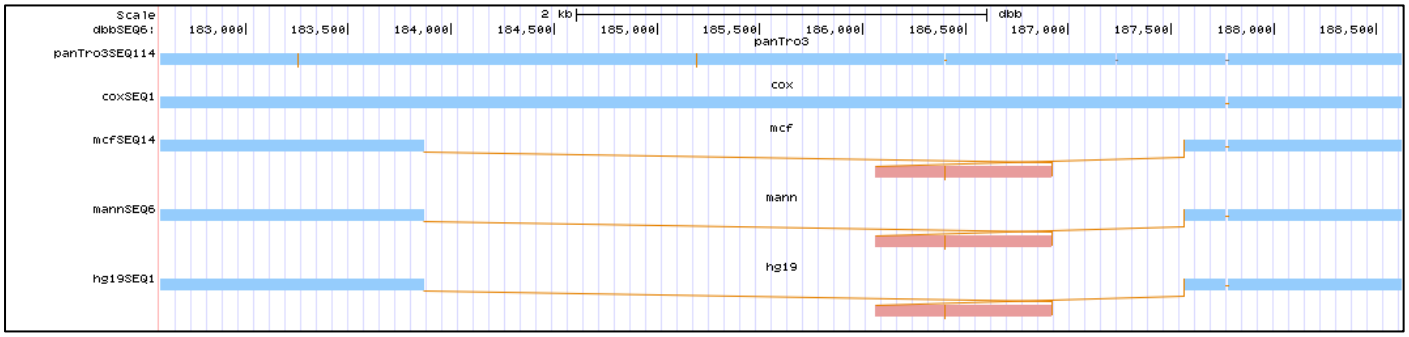
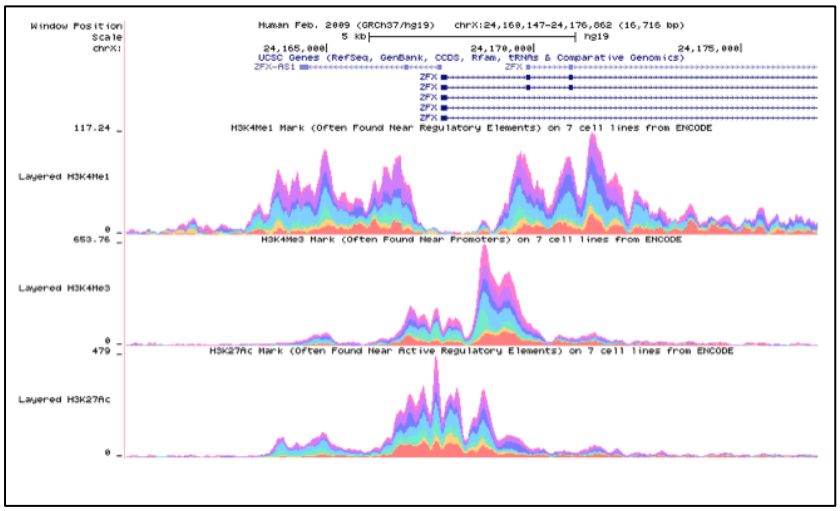


Figure 3. Genome Browser screenshot of the transcription start site of the ZFX gene. The histone marks H3K4me1, H3K4me3, and H3K27ac are shown as stacked bar wiggles in which the contribution to the sum of the signal of each of six different cell types is indicated by a different color. Note that the K3K4me1 signal is low in all cell types across the promoter and at the beginning of the transcript, but high on either side of the promoter. The other two marks show the opposite tendency, being highest at the promoter and declining with distance from the promoter.



1.c. Adapt to higher volumes of data

We invested considerable effort in fine-tuning and publicizing our track and assembly data hub functionality, improving some of our big data track formats, and promoting the use of our new Genome Browser server in Germany (Genome-euro) that improves European accessibility and performance:

- Implemented track data hub searches for tracks within a hub and for items within a track
- Added and extended our documentation on hub setup and use, and helped steer labs toward uniform presentation of hub data. We plan to release a Hub Quick Start guide by Jun. 2014.
- Improved error-reporting for BAM and bigData custom tracks
- Helped implement a new tool for bigWig manipulation: bigWigCat
- Added handling of https redirects between the UCSC and European websites

1.d. Enhance the security of uploaded data

We continued to focus our security efforts on preventing malicious hacker attacks on our website and increasing the security of our users' browser sessions:

- Implemented security measures in the Genome Browser software to prevent inappropriate access to the MySQL database and prevent exploitation through SQL injections
- Replaced sequential session IDs with cryptographically secure IDs to protect user data from unwanted access via the use of random web session IDs
- Established engineering coding standards to prevent introduction of security holes

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

We continued to improve support for individuals who build and modify our source code:

- Improved the Genome Browser source tree build process to make it easier for external users to build our most useful and popular applications
- Documented some of our recommended maintenance processes and scripts, such as our trash-cleaning procedure
- Identified and fixed the MacOS compiler errors in our source tree

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

Our most notable assembly update of the year was the release of the new human genome assembly (hg38/GRCh38) with a basic set of annotations on our public website in Mar. 2014, a few months after its completion by the Genome Reference Consortium (GRC). Because many of the annotations on our human browsers rely on data sets from external contributors (such as our popular SNPs tracks) or require massive computational effort (our comparative genomics tracks), they typically are not available for several months after the initial release of the assembly. We plan to focus on enriching the hg38/GRCh38 annotation track set in the upcoming year. Note that we have changed the UCSC internal numbering scheme for the human assemblies, starting with GRCh38, to more closely match that of the GRC.

This year we released four new genome assemblies in the browser: American alligator, Chinese hamster, elephant shark, and Minke whale. In addition to the human assembly update, we updated the assemblies of five other species (Table 1). The Genome Browser now provides access to multiple assemblies of 63 natively hosted species.

Table 1. Genome assemblies released in the Genome Browser Jul. 2013 – present. Note that we have changed the UCSC internal numbering scheme for the human assemblies to more closely match that of the GRC.

Organism	Scientific name	Assembly
New Genomes		
American Alligator	*Alligator mississippiensis*	allMis1
Chinese hamster	*Cricetulus griseus*	criGri1
Elephant shark	*Callorhinchus milii*	calMil1
Minke whale	*Balaenoptera acutorostrata scammoni*	balAcu1
Updated Genome Assemblies		
Alpaca	*Vicugna pacos*	vicPac2
Hedgehog	*Erinaceus europaeus*	eriEur2
Human	*Homo sapiens*	hg38
Sheep	*Ovis aries*	oviAri3
Tenrec	*Echinops telfairi*	echTel2
Zebra Finch	*Taeniopygia guttata*	taeGut2

In late 2013 we released a 100-species multiple alignment conservation track on the hg19/GRCh37 human genome assembly, our largest comparative alignment of genome assemblies to date. We are continuing to evaluate alternative multiple alignment programs as we construct the full conservation track for the new hg38/GRCh38 human assembly and prepare for even larger alignments in the future. We also provided access to comparative alignments of 57 species of *E. coli* and 9 species of *Shigella* through our new track assembly hub feature (Aim 3).

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

We added many new or updated annotation tracks to the Genome Browser in the past year (Table 2). To expedite the processing of tracks based on external data that are updated frequently, we have a set of automated tools that regularly checks the downloads sites of selected data providers, downloads any new or updated data, inserts the data into the proper database tables, and displays it on our public website.

The production of the ENCODE data and subsequent creation of Genome Browser annotation tracks were funded by a separate award. However, the browser continues as the primary platform for displaying the ENCODE data and download files from the initial production phase and selected data from the current analysis phase. Table 2 includes ENCODE annotation tracks that have been released in the UCSC Genome Browser during the past year.

Table 2. Annotation tracks released on the Genome Browser during 2013-14. Tracks that are automatically updated when new data is released are marked as “auto-update”.

Species	Assembly	Track	Status
Cat	felCat5	Human Chain/Net	new
Chicken	galGal4	SNPs (138)	new
Cow	bosTau6	NumtS Mitochondrial Sequence	new
	bosTau7	SNPs (138)	new
Dog	canFam3	Human Chain/Net	new
Human	hg18, hg19	Database of Chromosomal Imbalance and Phenotype in Humans (DECIPHER)	auto-update
		Database of Genomic Variants (DGV): Structural Variation	update
		International Standards for Cytogenic Arrays (ISCA)	auto-update
		NHGRI Catalog of Published Genome-Wide Association Studies (GWAS)	auto-update
		Online Mendelian Inheritance in Man (OMIM)	auto-update
		Pfam in UCSC Genes	update
		phastBias gBGC Predictions	new

	hg19	100-species Conservation	new
		Catalog of Somatic Mutations in Cancer (COSMIC)	auto-update
		Chains & Nets - several species	new
		ClinVar Variants	new
		Contigs Dropped or Changed from hg19/GRCh37 to hg38/GRCh38	new
		ENCODE Regulation: Transcription Factor ChIP-seq with Factorbook motifs	new
		ENCODE Regulation: Transcription Factor ChIP-seq Clusters	update
		GENCODE Genes v17, v19	new
		Genetic Association Studies of Complex Diseases and Disorders (GAD)	update
		Genome Segmentations from ENCODE	new
		Human Gene Mutation Database (HGMD) Public Variants	new
		Leiden Open Variation Database (LOVD) Public Variants	new
		Locus Reference Genomic (LRG) Fixed Transcript Annotations	new
		Locus Reference Genomic (LRG) Sequences Mapped to hg19/GRCh37 Assembly	new
		NHLBI GO Exome Sequencing Project (ESP) – Variants from 6,503 Exomes	new
		Personal Genome Variants	update
		Retroposed Genes	new
		SNPs (135)	new
		SNPs (137)	new
		SNPs (138)	new
		UCSC Genes	update
		UCSF Brain DNA Methylation	update
		UniProt/SwissProt Amino Acid Substitutions	new
	hg38	Standard initial track set (new assembly)	new
		NHGRI Catalog of Published Genome-Wide Association Studies (GWAS)	new
		OMIM Genes	new
		OMIM Phenotypes	new
		UCSC Genes	new
Lamprey	petMar2	Lamprey Genes	new
Mouse	mm10	Chains and Nets for several species	new
		GENCODE Genes vM2	new
		Mouse Genome Informatics (MGI) Quantitative Trait Loci (QTL)	new
		Retroposed Genes	new
		SNPs (138)	new
		UCSC Genes	update
Pig	susScr2	NumtS Mitochondrial Sequence	new
	susScr3	SNPs (138)	new
Rhesus	rheMac3	Human Chain/Net	New
Several assemblies		Accession at International Nucleotide Sequence Database Collaboration (INSDC)	new
		DNA Sequences in Web Pages Indexed by Bing.com / Microsoft Research	new
		Ensembl Genes v72	new
		Ensembl Genes v73	new
		Ensembl Genes v74	new
		Ensembl Genes v75	new
		Genscan Genes	new
Every assembly		GenBank updates (e.g., RefSeq Genes, ESTs, mRNAs)	auto-update

We 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. During the past year we worked with several groups to provide links from our public hubs page to their locally hosted data sets. Table 3 lists 13 new public track and assembly data hubs that can be accessed from the Genome Browser.

Table 3. Public track and assembly data hubs newly published on the Genome Browser website in 2013-14. We now link to a total of 19 public hubs.

Hub Description	Lab	Assembly
Blueprint Epigenomics data hub	Blueprint Epigenome Consortium	hg19
Epigenomic Data tracks	Centre for Epigenome Mapping Technologies, BCGSC, Vancouver, BC	hg19
Whole-Cell 454 Hela and K562 RNASeq tracks	Mike Snyder lab, Stanford Univ., CA	hg19
McGill Epigenomics Mapping Centre hub	McGill Epigenomics Mapping Centre, Montreal, Quebec	hg19
Ultra conserved elements in the human genome	David Haussler lab, UCSC, CA	hg19
Roadmap Epigenomics Integrative Analysis hub	Roadmap Epigenomics Consortium (hub provided by Ting Wang lab, WUSTL, MO)	hg19
DNA Methylation hub	Andrew Smith lab, USC, CA	rheMac3, mm9, hg18, hg19
Sense/antisense gene/exon expression using Affymetrix exon array	Bioinformatics Group, S. Dakota St. Univ., SD	rn4, mm9, hg19
CanFam3 improved annotation data v.1	Vertebrate Biology Group, Broad Institute, MA	canFam3
D. simulans w501 assembly hub	Kevin Thornton lab, UC Irvine, CA	*D. simulans*
E.coli comparative assembly hub	David Haussler lab, UCSC, CA	57 *E. coli* and 9 *Shigella* species
E.coli comparative assembly hub, with duplications	David Haussler lab, UCSC, CA	57 *E. coli* and 9 *Shigella* species
Plant assembly hub (CSHL Biology of Genomes 2013 demonstration)	Genome Bioinformatics Group, UCSC, CA	*A. thaliana, B. rapa, R. communis*

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

This year we released a new UCSC Genes set for the most recent mouse assembly (mm10/GRCm38) and the two latest human assemblies (hg19/GRCh37/hg19 and hg38/GRCh38). We also updated our underlying GO, UniProt, and proteome databases for the human and mouse assemblies. The UCSC Genes track is a set of gene predictions based on data from RefSeq, GenBank, CCDS, UniProt, Rfam, and the tRNA Genes track. The track is a moderately conservative set of predictions that includes both protein-coding genes and non-coding RNA genes. Transcripts of protein-coding genes require the support of one RefSeq RNA, or one GenBank RNA sequence plus at least one additional line of evidence. Transcripts of non-coding RNA genes require the support of one Rfam or tRNA prediction.

We typically create Ensembl Genes tracks for several assemblies soon after the Ensembl team releases its updates. This past year, we updated the Ensembl Genes tracks (v.72 - 75) for all of our genome assemblies on which Ensembl provided gene annotations.

Revisions

Aim 5. Provide training and support to the user community via online and in-person activities.

In Aug. 2013 we were awarded a two-year administrative supplement to the Genome Browser grant to support the hire of a new trainer to expand our outreach activities. Following a recruitment period in late 2013, we identified a qualified candidate, Dr. Pauline Fujita, in Jan. 2014. Dr. Fujita, a current member of the Genome Browser engineering staff, will formally assume the training position on Jul. 1, 2014. In the interim, she has been creating content for workshops and has administered a workshop signup questionnaire on our website

that garnered 110 applications during its initial three weeks online. Dr. Fujita is arranging an ambitious calendar of workshops, several of which may be presented before the end of this grant year.

During the past year we presented onsite Genome Browser workshops at over 20 venues worldwide (Table 4). The bulk of the workshops were provided by our Outreach and Training manager, Dr. Robert Kuhn, and others on the Genome Browser staff, although we continued to subcontract with OpenHelix (www.openhelix.com) for two workshops produced at our expense. Our Genome Browser workshops are quite popular; for example, our 1.5-hour workshop at the ASHG 2013 annual meeting once again sold out its 200+-seat capacity in the first few days of registration. Similarly, we filled the room twice for 1.5-hour workshops at the ESHG annual meeting in Jun. 2013, reaching 500 people. We have been invited back to both meetings for 2014. Many of the workshops listed in Table 4 involved a full day of instruction that included hands-on problem solving. Even experienced users reported that they learned about new features in our workshops.

The workshops are an excellent avenue for users to give feedback on features and functions, which are then relayed to the UCSC Genome Browser development team. Many of these suggestions have subsequently been incorporated into browser datasets, functionality, and documentation.

In addition to producing two Genome Browser workshops, OpenHelix again provided us with updates to the Genome Browser and Table Browser Quick Reference Cards (QRCs) that we both distribute at workshops. Together, we distributed more than 2,000 of each in the past year. OpenHelix also published 20 entries relevant to the Genome Browser in their online blog. They report nearly 40,000 page views per year of the UCSC-related content on their website, as well as more than 6,000 views of their video tutorials, which range from 30 to 60 minutes in length, about various Genome Browser tools. OpenHelix updated all three of their UCSC tutorials during the past year.

Table 4. Genome Browser workshops presented by UCSC and OpenHelix, Jul. 2013 - Jun. 2014.

Workshop	Location	Date	Provider	Attendance (* expected)
Washington Univ. St. Louis	St. Louis, MO	Jul. 2013	OpenHelix	22
Human Genetics Society of Australasia (plenary, 2 talks)	Queenstown NZ	Aug. 2013	UCSC	575
QFAB Bioinformatics, U Queensland, (2)	Brisbane, AU	Aug. 2013	UCSC	155
Centre for Clinical Genomics, Garvan Institute for Medical Research (2)	Sydney, AU	Aug. 2013	UCSC	95
Commonwealth Scientific and Industrial Research Organisation (CSIRO) (2)	Canberra, AU	Aug. 2013	UCSC	70
Victorian Life Science Computational Initiative (VLSCI), Univ. Melbourne (2)	Melbourne, AU	Aug. 2013	UCSC	80
NGS course for clinical medicine (3)	Leuven, Belgium	Sep. 2013	UCSC	200
American Society of Human Genetics (ASHG) 2013	Boston, MA	Oct. 2013	UCSC	160
Medical Informatics Conference (2)	Manipal, India	Nov. 2013	UCSC	130
UC Merced (2)	Merced, CA	Dec. 2013	UCSC	80
Plant and Animal Genomes XXII	San Diego, CA	Jan. 2014	UCSC	80
Children's Hospital	Los Angeles, CA	Jan. 2014	UCSC	60
UC San Francisco	San Francisco, CA	Jan, 2014	UCSC	70
Korean Genomics Organization (KOGO) Bioinformatics Workshop (2)	Yong Pyong, Korea	Feb. 2014	UCSC	120
European Molecular Biology Lab	Heidelberg, Germany	Mar. 2014	UCSC	20
VizBi – Visualization of Biological Data	Heidelberg, Germany	Mar. 2014	UCSC	8
EMBO Workshop	Puerto Varas, Chile	Apr. 2014	OpenHelix	50*
Emory Univ.	Atlanta, GA	Apr. 2014	OpenHelix	25*
BioIT NGS Conference (2)	Cambridge, MA	Apr. 2014	UCSC	100*
HUGO annual meeting	Geneva,	May 2014	UCSC	40*

	Switzerland			
Frontiers in Reproduction Course (2)	Woods Hole, MA	May 2014	UCSC	40*
ESHG 2014	Milan, Italy	Jun. 2014	UCSC	300*

There is a growing collection of videos of full-length workshops and talks by Dr. Kuhn available online, for example these videos on YouTube.com that have been viewed nearly 500 times:

- https://www.youtube.com/watch?v=KZxmO5x5Wqk
- https://www.youtube.com/watch?v=sEK_kdXdWYM

and this workshop video recently released by UC San Francisco for public viewing:

- http://www.library.ucsf.edu/content/watch-ucsc-genome-browser-workshop

In the next year we intend to open a UCSC Genome Browser channel on YouTube that will contain links to these videos, as well as short how-to videos that are in production at UCSC.

Our staff also present posters and talks of a non-tutorial nature on our work at scientific meetings and other venues, and are active in local outreach, responding to requests for talks in various courses on campus and from other groups not directly related to UCSC (Table 5). This year several of our staff reviewed papers for scientific journals, including Human Mutation, Bioinformatics, F1000Research, Database, and Nucleic Acids Research.

Table 5. Sample of non-tutorial UCSC Genome Browser talks presented by UCSC staff, 2013-14.

Venue	Location	Date	Presenter
UCSC COSMOS summer high school program	Santa Cruz, CA	Jul. 2013	Robert Kuhn
SciKnowMine Workshop	Los Angeles, CA	Aug. 2013	Max Haeussler
Intelligent Systems for Mol Bio (ISMB) (poster)	Berlin, Germany	Aug. 2013	Max Haeussler
Human Proteome Organization (HUPO)	Yokohama, Japan	Sep. 2013	Kate Rosenbloom
RIKEN	Yokohama, Japan	Sep. 2013	Kate Rosenbloom
UCSC BME110 class	Santa Cruz, CA	Oct. 2013	Robert Kuhn
Santa Cruz Chamber of Commerce	Santa Cruz, CA	Oct. 2013	Robert Kuhn
Genome Informatics (poster)	Cold Spring Harbor, NY	Oct. 2013	Brian Raney
Microsoft Research	Redmond, WA	Nov. 2013	Max Haeussler
UCSC campus-wide Genome Browser talk	Santa Cruz, CA	Feb. 2014	Robert Kuhn
UCSC Pediatric Cancer fundraiser	Santa Cruz, CA	Apr. 2014	Katrina Learned
Biology of Genomes (poster)	Cold Spring Harbor, NY	May 2014	Max Haeussler
UCSC DNA Day	Santa Cruz, CA	May 2014	Several
Morning Forum of Los Altos	Los Altos, CA	May 2014	Robert Kuhn

Additional outreach activities include our presence on several external Scientific Advisory Boards (SABs):

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

C. Significance

The UCSC Genome Browser website is a vital scientific resource for the biomedical research community. It provides convenient access to the sequence and annotations associated with genetic loci; integrates data from thousands of high-throughput scientific experiments; provides multiple alignments, conservation graphs, and other comparative genomics results based on dozens of vertebrate genomes; and offers a display platform where researchers can view the results of their own experiments alongside published annotations, and can share their results with others. The UCSC Genome Browser provides an informative view of any gene in the genome, including the many thousands of genes that have not been the focus of scientific papers.

The data sources integrated into the UCSC Genome Browser include human-curated and computed gene

sets, genotype-phenotype association studies, data from high-throughput sequencing of individuals and tumors, microarray-based expression data, in-situ imagery, chromatin immunoprecipitation, DNAse hypersensitivity assays, human and animal polymorphism data, the results of human gene association studies, model organism QTL studies, and a large variety of data derived from comparative genomics. More comprehensive, more accurate versions of these data are released regularly, and occasionally an entirely new type of genomic data is developed. Keeping abreast of these data is a large—but necessary—job if the UCSC tools are to be of the most use to the greatest number of research scientists. The UCSC Genome Browser staff relishes this challenge.

Citations in the literature

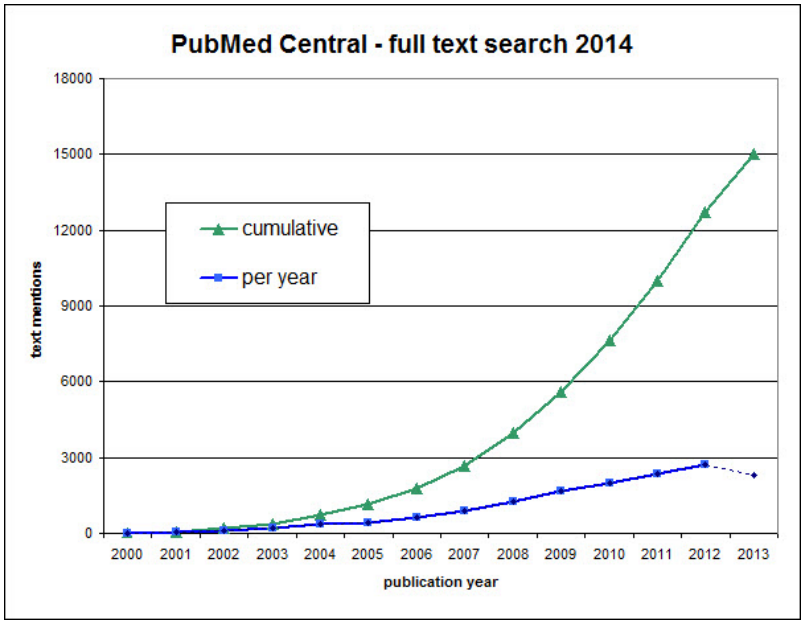
The UCSC Genome Browser tools have been cited thousands of times in the scientific literature (Table 6). Increasingly, our two most popular tools—the Genome Browser and BLAT—are used without citation, as may be appropriate for tools of their maturity. As a possible estimate of the rate at which the UCSC Genome Browser tools are being used (but not cited) for research published in the scientific literature, we have searched the approximately 15% of biomedical papers available through PubMed Central full-text access. The text string “UCSC + (genome OR browser OR database)” appeared in more than 2,700 papers in 2012 and more than 2,300 papers in 2013, implying a potential utilization rate of at least 15,000 publications a year in recent years (Figure 4).

Table 6. UCSC Genome Browser group publication citations tallied by Google Scholar.

Topic	Total citations	Increase from 2013
Kent. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002	3950	16%
Kent et al. The human genome browser at UCSC. *Genome Res.* 2002	3098	21%
Siepel et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005	1519	----
Karolchik et al. The UCSC Genome Browser database. *Nucl. Acids Res.* 2003	1270	5%
Blanchette et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 2004	814	12%
Fujita et al. The UCSC Genome Browser database: update 2011. *Nucl. Acids Res.* 2011	700	37%
Karolchik et al. The UCSC Table Browser data retrieval tool. *Nucl. Acids Res*. 2004	583	17%
Karolchik et al. The UCSC Genome Browser database: 2008 update. *Nucl. Acids Res.* 2008	477	6%
Rhead et al. The UCSC Genome Browser database: update 2010. *Nucl. Acids Res.* 2010	461	14%
Kent et al. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *PNAS.* 2003	458	11%
Hinrichs et al. The UCSC Genome Browser database: update 2006. *Nucl. Acids Res.* 2006	365	15%
Kuhn et al. The UCSC Genome Browser database: update 2009. *Nucl. Acids Res.* 2009	311	7%
Hsu et al. The UCSC Known Genes. *Bioinformatics.* 2004	290	24%
Kuhn et al. The UCSC Genome Browser database: update 2007. *Nucl. Acids Res.* 2007	267	6%
Dreszer et al. The UCSC Genome Browser database: extensions and updates 2011. *Nucl. Acids Res.* 2011	206	93%
Meyer et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucl. Acids Res.* 2013	187	1338%
Rosenbloom et al. ENCODE whole-genome data in the UCSC Genome Browser. *Nucl. Acids Res.* 2010	163	279%
Miller et al. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* 2007	159	12%
Rosenbloom et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucl. Acids Res.* 2012	108	151%
Karolchik et al. The UCSC Genome Browser. *Current Protocols in Bioinformatics.* 2009	100	----

Raney et al. ENCODE whole-genome data in the UCSC Genome Browser (2011 update). *Nucl. Acids Res.* 2011	97	----
Schneider et al. The UCSC archaeal genome browser. *Nucl. Acids Res.* 2006	82	15%
Rosenbloom et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucl. Acids Res.* 2013	70	1067%
Thomas et al. The ENCODE project at UC Santa Cruz. *Nucl. Acids Res.* 2007	68	11%
Kent et al. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* 2005	58	2%
Kent et al. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics.* 2010	58	----
Hsu et al. The UCSC Proteome Browser. *Nucl. Acids Res.* 2005	51	6%
Zweig et al. UCSC Genome Browser tutorial. *Genomics*. 2008	39	15%
Kuhn et al. The UCSC Genome Browser and associated tools. *Brief Bioinform.* 2013	39	----
Karolchik et al. The UCSC Genome Browser. *Current Protocols in Bioinformatics.* 2011	20	----
Karolchik et al. The UCSC Genome Browser: what every molecular biologist should know. *Current Protocols in Bioinformatics.* 2009	11	----
Raney et al. The UCSC Genome Browser database: 2014 update. *Nucl. Acids Res.* 2013	4	----
Total	**16083**	

Figure 4. A cumulative count of papers that mention the UCSC Genome Browser in PubMed Central (search string: "UCSC + (genome OR browser OR database)"). Note that only about 15% of papers are deposited into PubMed Central. Journal embargos and delays in making articles available in PubMed account for the dip in 2012-13 (blue dotted line).



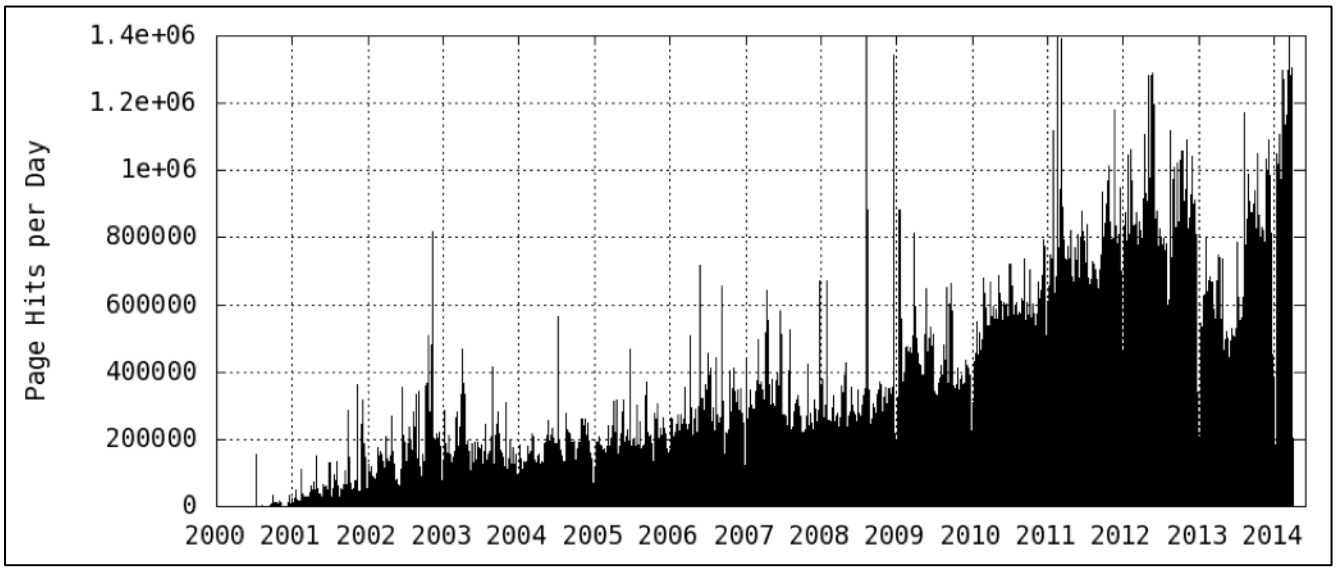
Usage

In Jun. 2013 we launched a second Genome Browser server, Genome-euro, in Germany to provide an alternate, faster access point for our European users and to reduce the load on our U.S.-based servers. During the past year, we have been automatically redirecting users to this site based on their geographic location. Genome-euro now receives approximately 260,000 hits per day. This is a small fraction of our overall traffic, but the true value of this resource is the increased speed and ease-of-use experienced for our European user community.

At present we log a combined total of approximately 8 million hits per week on our U.S. and Genome-euro websites from an average of 37,500 unique IP addresses per week (Figure 5). Many institutions appear to our

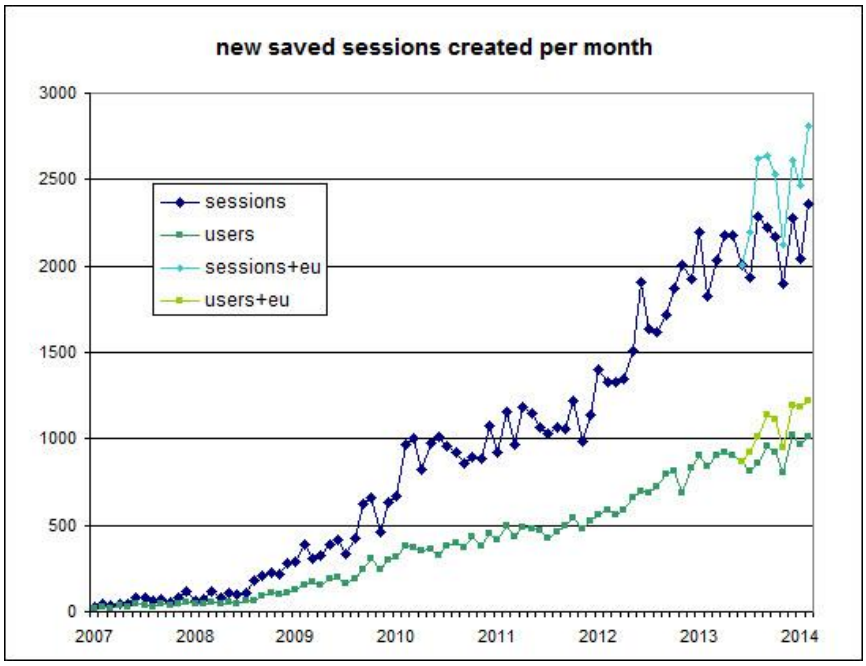
system as a single IP address from behind a firewall; thus, the actual number of users is potentially much higher than this. These figures also exclude usage on mirrors at other institutions. For example, the three most heavily used mirrors currently log a combined total of 27,600 hits per day: Denmark (20,000), Milwaukee (6,400), and Cornell (1,200). The manager of a mirror at NIH also reports heavy usage. Although these sites are intended primarily for local use, we do publish a selected list of reliable mirror sites that may be used when the UCSC and Genome-euro websites are unavailable.

Figure 5. Upward trend in page hits at http://genome.ucsc.edu since website inception in 2000.



The use of Genome Browser sessions has increased during the past year (Figure 6), consistently reaching more than 2,500 new sessions created per month by 1,000 different users. 13% of that usage is on the Genome-euro website.

Figure 6. UCSC Genome Browser Sessions usage through Apr. 2013.



Reliability and User Support

The UCSC Genome Browser site remains quite reliable, with a low page error rate and little downtime. Since Jul. 2013, our UCSC-based site has been down for about 82 minutes during peak usage hours (6am - 5pm PT, Mon. – Fri.), and Genome-euro has experienced about 75 minutes of downtime during peak usage hours (8am - 6pm UTC+1, Mon. – Fri.). Our publicized mirrors sites remained accessible during these outages, and the UCSC and Genome-euro servers have never been offline simultaneously during peak hours.

We remain committed to providing timely, thorough and clear technical support to our users through our two publicly accessible mailing lists, genome@soe.ucsc.edu (856 subscribers) and genome-mirror@soe.ucsc.edu (210 subscribers). We also provide a non-public forum for private questions from users, genome-www@soe.ucsc.edu, and an online suggestion box that netted 30 requests in the past year. Across all of the lists, our staff replied to 928 threads (many of which included additional follow-up questions and answers) during this grant period. We also maintain a twitter account (@GenomeBrowser) with 2,180 followers and a mailing list, genome-announce@soe.ucsc.edu (1,760 subscribers), for announcing feature and data releases and news items of interest to our users.

Hardware infrastructure

In the past year we made these major improvements to the hardware infrastructure of the Genome Browser project:

- Moved our development servers and compute cluster to the San Diego Supercomputer Center (SDSC), a more stable, economical facility than their previous location at UCSC
- Upgraded most of our development and public servers to the CentOS-6 platform
- Moved our hgcentral server to a solid state drive
- Upgraded the hardware of the hgnfs1 server

Genome Browser Scientific Advisory Board

The UCSC Genome Browser project has an active SAB that currently consists of six members:

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

The SAB convened with the Genome Browser management team in Oct. 2013. NHGRI Program Director Adam Felsenfeld attended remotely, and SAB members King and Gray were absent. The SAB felt that the Genome Browser project continues to do a good job of serving the needs of the research community. They offered several ideas for improving the usability and usefulness of data displayed in the browser, as well as suggestions for new features (Table 7). We hope to implement those ideas that have a reasonable ratio of usefulness vs. implementation difficulty and can be readily staffed by our current resources.

Table 7. Feedback from the Genome Browser Scientific Advisory Board meeting in Oct. 2013, categorized by the likelihood of implementation on our website.

Suggestion	Feasibility/status
Show phasing of chromosomes	Data not yet available
Show hetero/homozygous nature of variants	Somewhat addressed by existing pgSnp data format
Reduce track clutter and difficulty in finding data	We are currently preparing summary tracks that consolidate data from several tracks into one view (ENCODE datasets) or composite tracks (comparative genomics datasets). Another suggested approach is a

	session gallery to illustrate useful track combinations.
Support high-throughput sequencing technologies and display of variant data for human patients, both in research and clinical contexts	We have enhanced our support of variant analysis via software (e.g., VAI) and data (HGMD, OMIM Allelic Variants, LOVD datasets). GBiB should also facilitate this.
Show explicit splice-junction support	Supported by BAM files
Fix use of Back button on Genome Browser tracks page	Fixed
Provide better support for model organisms, such as latest *C. elegans* assembly	The assembly data hubs functionality, with links from our public hubs page, will help with this. We plan to publish Waterston lab's *C. elegans* assembly hub when it becomes available.
Label exons with mouseover	Existing feature on project todo list
VAI - browsable format, add links from table output, 1000 Genomes Project allele frequencies	Existing features on project todo list
Set up a video channel	Existing feature on project todo list
Allow simultaneous configuration of multiple tracks	Existing feature on project todo list
Implement track search on data hubs	Existing feature on project todo list
Add medical compliance for GBiB	Existing feature on project todo list
Find special-interest funding for adding non-human animals to browser (cow, etc.)	Under consideration
Create comparative pipeline for assembly hub users	Under consideration
Expand mouseover content -- pathways, knownGene content	Under consideration
Allow reordering of multiple track based on data values	Under consideration
Help users discover tracks via a "recommender", roadmap, table of contents	Under consideration
Represent diploid nature of genome when possible	Unlikely to implement soon
Improve handling of patches	Unlikely to implement soon
Indicate the uncertainty level of data (raw vs. curated vs. predicted)	Unlikely to implement soon

D. Plans

Our plans for the next year include most of the tasks originally listed in the grant proposal for Year 3, with the exception of "Circos-like display", which has been delayed to later in the grant period. We have also added some uncompleted tasks that were initially scheduled for one of the first two years of the grant, as well as some new tasks that were not listed in the original grant proposal (including selected recommendations from the Scientific Advisory Board in Table 7).

Aim 1 – Software development

Yr.Qtr	Specific Task	Status
3.1	Combine multiple wiggles in new ways (partially completed in Year 2)	Year 2 task (continued)
	Create session gallery with sample sessions for different users and use cases	new (SAB suggestion)
3.2	Start work on displaying long-distance chromatin interactions including those across chromosomes	Year 3 task
	Help users discover tracks (add functionality to Track Search)	new (SAB suggestion)
3.3	Package command-line tools with updated documentation	Year 3 task
	Add exon numbering to the gene mouseovers	new (SAB suggestion)
3.4	Update index page with better graphics and pull-down menus	Year 2 task (continued)
	Implement condensed "exon only" display	Year 3 task

Aim 2 – Comparative genomics

Yr.Qtr	Specific Task	Status
3.1	Add genome browsers for three new or updated genomes	Year 3 task

3.2	Add genome browsers for three new or updated genomes	Year 3 task
3.3	Add genome browsers for three new or updated genomes	Year 3 task
3.4	Add new multiple-alignment track for one set of assemblies	Year 3 task
	Compare multiple genome alignment programs and possibly switch to new and better one	Year 2 task (continued)

Aim 3 – Import data

Yr.Qtr	Specific Task	Status
3.1	Add Segmental Dups tracks for mouse (mm10)	Year 2 task (continued)
	Annotate centromeres on hg38/GRCh38 (will likely be a track hub)	new
3.3	Import Illumina BodyMap RNA-seq data	Year 2 task (continued)
3.4	Import selected data from ENCODE3 project	Year 2 task (continued)
3.1-3.4	Add annotation tracks to hg38/GRCh38 existing browsers as new data become available	Year 3 task

Aim 4 – Gene sets

Yr.Qtr	Specific Task	Status
3.2	Update UCSC Genes at least once for human genome	Year 3 task
3.3	Update UCSC Genes at least once for mouse genome	Year 3 task
3.4	Evaluate Gencode Genes for suitability as primary gene set on human genome assembly	Year 3 task

Aim 5 – Training and outreach (added for grant supplement)

Yr.Qtr	Specific Task	Status
3.1-3.4	Expand on-site training to approximately twice the current level	new (supplement)
	Accumulate program income to support future training	new (supplement)
	Open YouTube channel to collect video tutorials	new (supplement)
	Record and release short video clips detailing specific browser tasks	new (supplement)

E. Publications

Refereed journal papers co-authored by our group

Earl D, Nguyen NK, Hickey G, Harris RS, Fitzgerald S, Beal K, Seledtsov I, Molodtsov V, Raney B, Clawson H, Kim J, Kemena C, Chang J, Erb I, Poliakov A, Hou M, Herrero J, Solovyev V, Darling AE, Ma J, Notredame C, Brudno M, Dubchak I, Haussler D, Paten B. Alignathon: A competitive assessment of whole genome alignment methods. Submitted to *Genome Research*. http://biorxiv.org/content/early/2014/03/10/003285

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014 Jan;42(Database issue):D764-70. PMID: 24270787; PMC: PMC3964947. http://nar.oxfordjournals.org/content/42/D1/D764.long

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res*. 2014 Apr;24(4):697-707. Epub 2014 Feb 5. PMID: 24501022. http://genome.cshlp.org/content/24/4/697.abstract

Nguyen N, Hickey G, Raney BJ, Armstrong J, Clawson H, Zweig A, Kent WJ, Haussler D, Paten B. Comparative Assembly Hubs: Web Accessible Browsers for Comparative Genomics. Submitted to *Bioinformatics*. arXiv:1311.1241 [q-bio.GN] http://arxiv.org/abs/1311.1241

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC

Genome Browser. *Bioinformatics*. 2014 Apr 1;30(7):1003-5. PMID: 24227676. http://bioinformatics.oxfordjournals.org/content/30/7/1003

Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M, Hoon S, Gangu V, Roy SW, Irimia M, Korzh V, Kondrychyn I, Lim ZW, Tay BH, Tohari S, Kong KW, Ho S, Lorente-Galdos B, Quilez J, Marques-Bonet T, Raney BJ, Ingham PW, Tay A, Hillier LW, Minx P, Boehm T, Wilson RK, Brenner S, Warren WC. Elephant shark genome provides unique insights into gnathostome evolution. *Nature*. 2014 Jan 9;505(7482):174-9. PMID: 24402279; PMC: PMC3964593. http://www.nature.com/nature/journal/v505/n7482/full/nature12826.html

Other papers co-authored by our group

Miga K. Our first view into the "blackout zones" of the human genome. *Sci Am* guest blog. 2014 Mar 6. http://blogs.scientificamerican.com/guest-blog/2014/03/06/our-first-view-into-the-blackout-zones-of-the-human-genome/

Nguyen N, Hickey G, Zerbino D, Raney B, Earl D, Armstrong J, Haussler D, Paten B. Building a Pangenome Reference for a Population. (ISMB 2014) In Sharan R (ed): *Lecture Notes in Computer Science: Research in Computational Molecular Biology*, Springer International Publishing, Switzerland. 2014;8394:207-21. http://dx.doi.org/10.1007/978-3-319-05269-4_17

Posters co-authored by our group

Haeussler M, Haussler D. Genome Annotation with Large Scale Mutation Extraction from Scientific Fulltext. ISMB/ECCB 2013. Berlin, Germany. Aug 2013.

Haeussler M, Raney BJ, Hinrichs A, Clawson H, Zweig AS, Karolchik D, Casper J, Speir M, Haussler D, Kent J. Navigating protected genomics data with the UCSC Genome Browser in a box. Biology of Genomes 2014. Cold Spring Harbor, NY. May 2014.

Raney BJ, Nguyen N, Dreszer TR, Barber GP, Clawson H, Fujita PA, Karolchik D, Zweig AS, Paten B, Kent WJ. Assembly Data Hubs support viewing any sequence on the UCSC Genome Browser. Genome Informatics 2013. Cold Spring Harbor, NY. Oct 2013.

F. Project-Generated Resources

The primary resources generated from this project:

- UCSC Genome Browsers for multiple assemblies of 61 species (http://genome.ucsc.edu/)
- Downloads of all data displayed in the Genome Browser assemblies (http://hgdownload.cse.ucsc.edu/downloads.html)
- UCSC Genome Browser-in-a-Box (GBiB)
- BLAT homology search engine (http://genome.ucsc.edu/cgi-bin/hgBlat)
- UCSC Track Data Hub tool (http://genome.ucsc.edu/cgi-bin/hgHubConnect)
- UCSC Table Browser data retrieval tool (http://genome.ucsc.edu/cgi-bin/hgTables)
- UCSC Gene Sorter for 6 species (http://genome.ucsc.edu/cgi-bin/hgNear)
- UCSC Genome Graphs tool (http://genome.ucsc.edu/cgi-bin/hgGenome)
- UCSC VisiGene image browser (http://genome.ucsc.edu/cgi-bin/hgVisiGene)
- In-Silico PCR tool (http://genome.ucsc.edu/cgi-bin/hgPcr)
- UCSC Variant Annotation Integrator (http://genome.ucsc.edu/cgi-bin/hgVai)
- UCSC Genes sets for human (hg16, hg17, hg18, hg19/GRCh37, hg38/GRCh38), mouse (mm7, mm8, mm9, mm10/GRCm38), and rat (rn3)

- A large set of data manipulation tools, such as the bigWig and bigBed tools, liftOver batch coordinate conversion tool, the chaining and netting tools, and a phylogenetic tree GIF generator (http://hgdownload.cse.ucsc.edu/admin/exe/)
- GenomeWiki site (http://genomewiki.ucsc.edu/)
- European Genome Browser server (Genome-euro) hosted at the University of Bielefeld, Germany (http://genome-euro.ucsc.edu/cgi-bin/hgGateway)