PROGRESS REPORT SUMMARY

A. Specific Aims

The grant proposal was funded with four original aims, which have not been modified:

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.

B. Studies and Results

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

In addition to bug fixes and basic modifications to the UCSC Genome Browser software to support new data sets, we added several new features and improvements to the browser in the past year:

- Improved the user interface and ease of use of the browser
  - Changed to a new font, implemented a new menu bar with improved navigation, and integrated a new background on the browser image page
  - Added ideograms to browsers that do not have a microscopically-derived cytology to make navigation easier
  - Added hints about how to improve browser images for publications

- Improved the display and responsiveness of the items in the genome browser image
  - Added anti-aliasing to improve the look of diagonal lines
  - Improved custom track handling by opening the bigDataUrl tracks in parallel
  - Changed the transparency of overlay wiggle tracks to support views in PDF format, and to support normalization when more cell lines are added

- Facilitated mirroring and use of command-line tools
  - Created new default compilation option: instead of separately downloading, patching and building samtools and tabix, we now use the new samtabix package (https://github.com/AngieHinrichs/samtabix) that contains combined, pre-patched samtools and tabix
  - Adopted UDR (a new package that integrates rsync with the high performance network protocol, UDT) to push data to the European mirror site, and started testing it with selected mirror sites
  - Simplified makefiles to make it easier for users to compile only the parts of the source code that they need (see 1.e)

- Improved position/search box and gene suggestion features
  - Merged the gene suggestion search pulldown menu into the position/search box
  - Included the gene description in the gene suggestion list
  - Enabled highlighting for genes chosen via the gene suggestion feature

- General cleanup
  - Retired the original text-based version of the Table Browser (hgText) and the Proteome Browser, which has been superseded by the UCSC Genes tracks and the Gene Sorter
  - Created a common library for bed and bed-plus validation that is used by multiple bed utilities and the custom-track loader, resulting in safety, correctness, consistency, and maintainability

1.a. Increasing website interactivity

To increase website interactivity, we fine-tuned several sections of source code to improve the website's responsiveness to users' clicks. We expanded the drag-reordering functionality to work at multiple levels in the track/subtrack hierarchy.

We also updated our menu system. We gathered input from users and UI experts before we decided how best to rearrange the menus into functionally grouped drop-down lists. As part of this restructuring, we switched to sans-serif fonts and a less intrusive background on the main browser image page. We have received overwhelmingly positive feedback from our users about these relatively simple changes.

1.b. Adapting to new types of data

This year, we concentrated our efforts in this area on adding more support for Variant Call Format (VCF) data. We improved display of VCF indel alleles in the main browser image and track details pages. We also improved the robustness of the haplotype clustering algorithm on sex chromosome data; for example, for VCF with genotypes for both male and female samples, chromosome X is haploid for some samples but diploid for others.

The UCSC Genome Browser currently hosts a large amount of chromatin immunoprecipitation sequencing (ChIP-seq) data on transcription factors, many of which bind to specific DNA motifs. Work is in progress to extend the display of this data type to show the location of motifs within the peak, and to show the sequence logo and matching score on the track details page for the peak. We anticipate this will be completed in Summer 2013.

1.c. Adapting to higher volumes of data

Over the years, our European users have reported that the UCSC Genome Browser can be slow when performing certain tasks. This year we built and installed a UCSC Genome Browser mirror site at the University of Bielefeld, Germany. The mirror site is fully functional, and we expect to begin automatically redirecting European users to this site in May 2013. This will reduce the load on our local servers and provide much better response time for our users in Europe.

Recently we introduced Track Data Hubs in response to the rapid increase in volume and sizes of data sets produced by next-gen sequencing. The popularity of this feature has grown considerably in the past two years: as of April 2013, we are aware of 2,303 active Track Data Hubs. Advanced users of this feature have asked us for more details about how to structure the underlying text files to take advantage of the complete set of display features. In response we created two detailed documents that explain the structure of the trackDb text files used to power data hubs:

- An internal document describing every aspect of the track database file: http://genome.ucsc.edu/goldenPath/help/trackDb/trackDbDoc.html
- An external document describing only the data hub-accessible aspects of track database files: http://genome.ucsc.edu/goldenPath/help/trackDb/trackDbHub.html

1.d. Enhancing the security of uploaded data

Because the bulk of the data in the UCSC Genome Browser is in the public domain, strong security is not generally required. However, many users upload confidential data to view alongside the native tracks in the context of the Genome Browser. We make a moderate effort to protect this data from other users.

Our main security concern is malicious attacks from hackers. To this end, we have enhanced MySQL security by removing all user privileges except SELECT for the system user (hguser) for all versions of MySQL. We have also removed website user access to the MySQL metadata database, thereby protecting user passwords from SQL injection. We are currently implementing a plan to protect all MySQL queries from SQL injections, a complicated effort that requires us to review every MySQL call in our entire code base (more than 2 million lines of code). The implementation, which places priority on those parts of the system where security is most important (the user cart, hgcentral database, custom trash, and so on), should be finished by Summer 2013.

Although we do not require a login or password to access our website, certain UCSC Genome Browser features, such as the Sessions feature that stores a browser view for sharing and later use, do require an account and password. Previously, we required users to log in through our UCSC Genome Browser Wiki site (genomewiki.ucsc.edu), which was confusing and created an unnecessary burden for our mirror sites. To rectify this, we have integrated a simple login feature to our web works that provides a seamless interface to our users, and is much easier for our mirror sites to adopt.

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

This year we re-architected the makefile software system that builds our software utilities and command-line tools, allowing mirror sites and users to now compile only those parts of the system that they actually need to produce specific tools. All of the necessary external pieces of software (e.g. samtools, tabix) are now precompiled directly in our source code. Also, the makefile system now makes assumptions about the locations of libraries on the host system instead of asking a user to puzzle through that information. We are in the process of documenting these changes, and expect to release this by June 2013.

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

This year we pushed the limits of our conservation track multiple alignment pipeline by creating and releasing a 60-species conservation track for the mouse (mm10/GRCm38). For details about the assemblies and parameters used to create this track, see this wiki page: http://genomewiki.ucsc.edu/index.php/Mm10_conservation_alignment.

In preparation for creating a 100-species conservation track on the upcoming human genome assembly, we are evaluating several new multiple genome alignment programs. The most promising so far is the Cactus program, which is being developed by Benedict Paten at UCSC. This work is called out as a specific task during the second year of the grant (see the Plans section of this document).

In our grant proposal, we planned to add nine genome browsers for new species or updated genomes in the first year. We greatly exceeded this goal by adding 29 browsers for new species and 9 updated assemblies for existing browsers (Table 1). Many of these were in support of the 60-species conservation track described above.

Table 1. Genome assemblies released in the browser July 2012 – June 2013.

Organism	Scientific name	Assembly
New Genomes		
Alligator	*Alligator mississippiensis*	allMis1
Alpaca	*Vicugna pacos*	vicPac1
Armadillo	*Dasypus novemcinctu*	dasNov3
Atlantic cod	*Gadus morhua*	gadMor1
Baboon	*Papio hamadryas*	papHam1
Bushbaby	*Otolemur garnettii*	otoGar3
Coelacanth	*Latimeria chalumnae*	latCha1
Crab-eating macaque	*Macaca fascicularis*	macFas1
Dolphin	*Tursiops truncatus*	turTru2
Ferret	*Mustela putorius furo*	musFur1
Florida manatee	*Trichechus manatus latirostris*	triMan1
Hedgehog	*Erinaceus europaeus*	eriEur1
Kangaroo rat	*Dipodomys ordii*	dipOrd1
Medium Ground Finch (Darwin Finch)	*Geospiza fortis*	geoFor1
Megabat	*Pteropus vampyrus*	pteVam1
Melodious Parrot	*Melopsittacus undulatus*	melUnd1
Mouse lemur	*Microcebus murinus*	micMur1
Nile tilapia	*Oreochromis niloticus*	oreNil2
Pika	*Ochotona princeps*	ochPri2

Rock hyrax	*Procavia capensis*	proCap1
Shrew	*Sorex araneus*	sorAra1
Sloth	*Choloepus hoffmanni*	choHof1
Squirrel	*Spermophilus tridecemlineatus*	speTri2
Squirrel monkey	*Saimiri boliviensis boliviensis*	saiBol1
Tarsier	*Tarsius syrichta*	tarSyr1
Tasmanian Devil	*Sarcophilus harrisii*	vsarHar1
Tenrec	*Echinops telfairi*	echTel1
Tree shrew	*Tupaia belangeri*	tupBel1
White Rhino	*Ceratotherium simum*	cerSim1
Updated Genome Assemblies		
Cat	*Felis catus*	felCat5
Chimp	*Pan troglodytes*	panTro4
Gibbon	*Nomascus leucogenys*	nomLeu2
		nomLeu3
Lamprey	*Petromyzon marinus*	petMar2
Naked mole rat	*Heterocephalus glaber*	hetGla2
Olive Baboon	*Papio_anubis*	papAnu2
Pig	*Sus scrofa*	susScr3
Rhesus macaque	*Macaca mulatta*	rheMac3

In the grant proposal, we discussed the development of a new type of data hub: the Assembly Data Hub. This new hub allows users to display genome assemblies that UCSC is not able to integrate into the standard Genome Browser database, and thus will allow us to serve the members of the genomics community who are working on species that we are unable to accommodate. When constructing an Assembly Data Hub, the user stores the genome sequence in a compressed, binary file format and makes it available on a remote web server along with optional data files that annotate the genome. The Assembly Data Hub and its annotations are then available for viewing in the Genome Browser.

We have made excellent progress towards this goal. We have added the software support for users to create assembly hubs and display them on the Genome Browser. However, before we publicly announce this functionality, there are a few features we would like to add, such as the ability to search for an item in assembly hub annotation tracks and technical documentation for constructing an assembly hub. We fully expect this feature to be polished and released by June 2013.

We will announce the Assembly Hubs feature in May 2013 by means of a poster at the Biology of Genomes meeting at Cold Spring Harbor Laboratory. As a proof of concept, we have built assembly hubs for 3 plant genomes including the Castor bean, *Ricinus communis*. An external user has built approximately 60 *E. coli* genome assembly hubs. So far, our small set of users has found the system robust and full-featured.

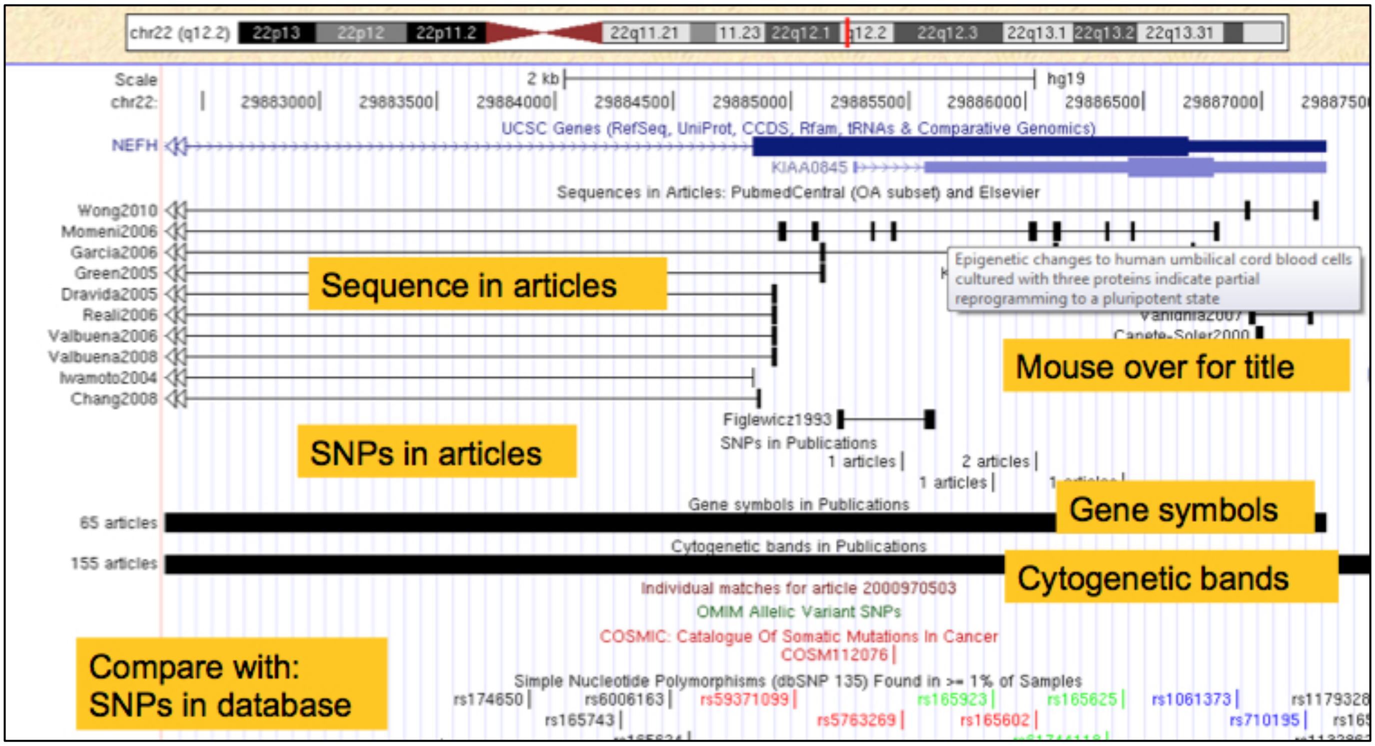
Aim 3. Import data from the scientific community that help interpret the functions of various human genome regions into the UCSC databases.

This year, to critical acclaim, we added a unique data set to our browser: the Publications track. This track is the result of the work of a postdoc in our lab, Maximilian Haeussler, who was not funded by the Genome Browser grant but received instrumental assistance from browser staff with the display of the data, the documentation, and the QA effort. Max's text-mining project scans full-text biomedical papers for genomic identifiers and maps them to the human genome and those of other model organisms. It currently recognizes DNA and protein sequences, SNPs, cytogenetic bands, and gene symbols. The current dataset consists of about 600,000 open-access main text and supplementary files from PubMed Central and more than 5 million text files from more than 30 of the major publishers. This track (Figure 1) is displayed on the hg19/GRCh37 browser, as well as 14 other browsers including most model organisms.

The Publications track has been well received by our users, and in general people are impressed by the audacity and creativity of it. As one person described it, "To scrape all those papers for GATCs and then BLAT them back to the genome with a link back to the paper is thinking outside the box. It is genome-scale thinking!"

One researcher who reacted to an initial demo with unbridled enthusiasm later wrote, "This is fantastic, I've already found two papers about my gene that I should have known about!".

Figure 1. Publications track in the Genome Browser showing publications which have sequence, SNPs, gene symbols or cytogenetic bands matching this location in the genome.



In addition to the Publications track we have added many new or update annotation tracks to the Genome Browser in the past year (Table 2). In order to stay current with of some of the quickly changing external data, we have developed a set of tools that perform automated updates of several annotation tracks similar to the method we have been using for years to update our GenBank data. These tools regularly check the downloads sites of selected data providers, and if there is new or updated data present they download it, insert it into the proper database tables, and display it on our public website. Tracks that use this process are marked “auto-updated” in Table 2.

Table 2. Annotation tracks released on the UCSC Genome Browser in 2012-13. Several of the tracks based on data obtained from external sources are now automatically updated when new data is released (“auto-updated” status).

Species	Assembly	Track Name	Status
C. elegans	ce10	Human Proteins	new
Chicken	galGal4	Chains & Nets - several species	new
Cow	bosTau7	Chains & Nets - several species	new
D. melanogaster	dm3	Human Proteins	new
Dog	canFam2	SNPs v131	new
	canFam3	Chains & Nets - several species	new
Fugu	fr3	Chains & Nets - several species	new
Gorilla	gorGor3	Chains & Nets - several species	new
Human	hg17, hg18, hg19	DGV Structural Variation	updated
	hg18, hg19	Gene Reviews	updated
		GWAS Catalog	updated
		DECIPHER – Database of Chromosomal Imbalance and Phenotype in Humans	auto-updated
		ISCA – International Standards for Cytogenic Arrays	auto-updated
		OMIM – Online Mendelian Inheritance in Man	auto-updated

	hg19/GRCh37	1000 Genomes Phase 1 Integrated Variant Calls	new
		1000 Genomes Phase 1 Paired-end Accessible Regions	new
		Affy CytoScan HD Array	new
		SNP v137: Common SNPs, Flagged SNPs, Mult. SNPs, All SNPS	new
		Coriell Copy Number Variants	new
		COSMIC - Catalog of Somatic Mutations in Cancer	auto-updated
		Denisova: Modern Human Derived, Sequence Reads, Variant Calls, Variant Calls from 11 Modern Human Genome Sequences	new
		Genetic Association Database (GAD)	updated
		GRC Patches 9 & 10	new
		miRcode Public Hub	new
		Pfam in UCSC Genes	new
		qPCR Primers	new
		UCSC Genes (*projected for release by 6/30/2013*)	new
		Chains & Nets - several species	new & updated
Lizard	anoCar2	Quality Scores	new
Marmoset	calJac3	Chains & Nets - several species	new
Medaka	oryLat2	Chains & Nets - several species	new
Mouse	mm10/GRCm38	Alternate mouse strain sequences	new
		Chromosome Band	new
		Contigs	new
		GRC Incident Database	new
		GRC Patch Release 1	new
		60-way conservation	new
		qPCR Primers	new
		SNPs (v137)	new
		UCSC Genes	new
		Chains & Nets - several species	new
Orangutan	ponAbe2	Chains & Nets - several species	new
Pig	susScr2	Nuclear mitochondrial sequences	new
Rat	rn5	Chains & Nets - several species	new
Rhesus	rheMac3	Genscan Genes	new
Tenrec	echTel1	Chains & Nets - several species	new
Turkey	melGal1	Chains & Nets - several species	new
White rhinoceros	cerSim1	Chains & Nets - several species	new
Zebra finch	taeGut1	Chains & Nets - several species	new
Several assemblies		Ensembl Genes v68	new
		Ensembl Genes v70	new
Every assembly		GenBank updates (e.g. RefSeq Genes, ESTs, mRNAs)	auto-updated

The production of the ENCODE data and subsequent creation of Genome Browser annotation tracks are funded by a separate grant. However, the browser serves as a platform for displaying the ENCODE tracks and download files. Table 3 lists ENCODE annotation tracks that have been released to the UCSC Genome Browser during the past year.

Table 3. ENCODE data tracks released during 2012-13 displayed in the UCSC Genome Browser.

Assembly	Track Name	Status
Human (hg19/GRCh37)	CSHL Small RNA-seq (Release 3)	updated
	UMass Chromatin Interations by 5C (Release 2)	updated

	UW Affymetrix Exon Array (Release 3)	updated
	UNC/BSU Proteogenomics Mapping (Release 2)	updated
	UW DNaseI Hypersensitivity (Release 6)	updated
	HAIB DNA Methylation by RRBS (Release 3)	updated
	UW Histone Modifications (Release 5)	updated
	RIKEN RNA Subcellular CAGE Localization (Release 4)	updated
	FSU Replication Timing (Release 2)	updated
	UW Affymetrix Exon Array (Release 4)	updated
	UW DNaseI Digital Genomic Footprinting (Release 4)	updated
	GIS RNA Sub-cellular Localization (Release 2)	updated
	Broad Histone Modifications (Release 3)	updated
	Caltech RNA-seq (Release 4)	updated
	SYDH Transcription Factor Binding Sites (Release 3)	updated
	UNC/BSU Proteogenomics and Gencode Mapping	new
	Duke Open Chromatin (Release 3)	updated
	UTA Transcription Factor Binding Sites (Release 2)	updated
	UNC Open Chromatin (Release 2)	updated
	HAIB Transcription Factor Binding Sites (Release 3)	updated
	CSHL Long RNA-seq (Release 3)	updated
	Uniform DNaseI Hypersensitivity	new
	Gencode Genes version 12	new
	Gencode Genes version 14	new
	ENCODE Regulation: DNase Clusters (Release 2)	updated
	ENCODE Regulation: Transcription (Release 3)	updated
Mouse (mm9/NCBI37)	FSU Replication Timing	new
	UW RNA-seq	new
	Stan/Yale RNA-seq (Release 2)	updated
	PSU Histone Modifications (Release 2)	updated
	PSU Transcription Factor Binding Sites (Release 2)	updated
	LICR Transcription Factor Binding Sites (Release 3)	updated
	PSU DNaseI Hypersensitivity	new
	Stan/Yale Transcription Factor Binding Sites (Release 4)	updated
	Stan/Yale Histone Modifications (Release 2)	updated
	CSHL Long RNA-seq (Release 3)	updated
	UW DNaseI Digital Genomic Footprinting	new
	LICR Histone Modifications (Release 3)	updated
	UW DNaseI Hypersensitivity (Release 2)	updated
	PSU RNA-seq	New

We have continued our practice of using Track Data Hubs to display data from large consortia rather than importing the full data sets ourselves. We tested this approach using the data from the Roadmap Epigenomics Project. The data sets remain on a remote server at Washington University in St. Louis, and they are available to all users of the UCSC Genome Browser through the data hub page. Based on the overwhelmingly positive response from our users, we have worked with other groups to provide links from our data hub page to their public data hubs. With the completion in September 2012 of the ENCODE-wide integrative data analysis, we assisted the ENCODE Analysis group with the completion of a Track Data Hub at the EBI. This site hosts analysis results that may be downloaded or viewed as browser tracks. Table 4 lists the public Track Data Hubs that may be accessed from the Genome Browser.

Table 4. Public Track Data Hubs available on hg19/GRCh37.

Name of Hub	Description	Link to raw Track Data Hub site
Translation Initiation Sites (TIS)	Translation Initiation Sites (TIS) track	http://gengastro.1med.uni-kiel.de/suppl/footprint/Hub/tisHub.txt
ENCODE Analysis Hub	ENCODE Integrative Analysis Track Data Hub	http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/hub.txt

miRcode microRNA sites	Predicted microRNA target sites in GENCODE transcripts	http://www.mircode.org/ucscHub/hub.txt
Roadmap Epigenomics Data Complete Collection at Wash U VizHub	Roadmap Epigenomics Data Complete Collection at Wash U VizHub	http://vizhub.wustl.edu/VizHub/RoadmapReleaseAll.txt
UMassMed ZHub	UMassMed H3K4me3 ChIP-seq data for Autistic brains	http://zlab.umassmed.edu/zlab/publications/UMassMedZHub/hub.txt
Cancer genome polyA site & usage	An in-depth map of polyadenylation sites in cancer (matched-pair tissues and cell lines)	http://johnlab.org/xpad/Hub/UCSC.txt

In order to assist researchers in annotating and prioritizing thousands of variant calls from sequencing projects, we are developing the Variant Annotation Integrator (VAI) tool that we anticipate to release to the public by June 2013. There are several existing tools that can annotate variant calls with predicted functional effects on protein-coding genes and regulatory regions, for example Ensembl's Variant Effect Predictor (VEP). However, these tools are usually restricted to one or two sources of gene annotations and a limited set of additional annotation sources. The VAI will offer much broader choices from the full UCSC database and user-provided custom tracks.

The first release of the VAI will include a simple user interface for selecting variants to annotate as well as the most commonly used annotation sources: protein-coding genes, regulatory regions, predictions from tools such as SIFT and PolyPhen2 provided by the Database of Non-Synonymous Functional Predictions (dbNSFP) and known variants from dbSNP. The interface will provide several options for filtering variants based on annotations. A link to an advanced user interface will enable sophisticated users to add annotation sources from the full database.

Because the underlying role of the VAI is to calculate intersections between data sets, similar to that of the existing Table Browser, we initially thought to extend the Table Browser to include VAI functionality. However, the architecture of the Table Browser has two limitations that make it unsuitable for this use: it does not allow for data streaming, and the intersection feature discards non-positional information from its secondary dataset rather than retaining all information from both sources. Both of these are key requirements for the VAI tool. Consequently we have created a new tool that is based on Table Browser functionality and includes a streaming interface, i.e. the data are processed row-by-row instead of via large in-memory batches, which greatly reduces memory requirements and allows results to be returned as soon as they are available rather than waiting until all computations have finished. The new infrastructure is a significant investment but we believe it will make big data operations more manageable.

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

This year we continued our practice of building our own genes set, UCSC Genes, 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. Compared to RefSeq, this gene set has generally about 10% more protein-coding genes, approximately four times as many putative non-coding genes, and about twice as many splice variants. In the past year we released a new UCSC Genes set for the most recent mouse assembly (mm10/GRCm38) and plan to generate a new set of UCSC Genes for human assembly hg19/GRCh37 by June 2013.

In our grant proposal we indicated that we would integrate Cap-Analysis Gene Expression (CAGE) data into our UCSC Genes pipeline for more accurate promoter calling. We have started work on this, but have not been able to complete it this year partly because our collaborators at RIKEN have not yet published the results of their CAGE experiments. We are hopeful that they will be published soon so we can use these data in our next generation UCSC Genes pipeline.

We typically create Ensembl Genes tracks for several assemblies shortly after Ensembl releases their updates. This past year, we updated Ensembl Genes tracks (using versions 68 and 70) for 54 of our genomes.

As part of the ENCODE project, two updates (versions 12 and 14) to the Gencode Genes track were released this year for hg19/GRCh37. Although the work to integrate the gene sets into the browser was funded by our ENCODE grant, the current grant funded work to implement extra highlighting in the display.

In this year's goals we planned to integrate and develop tools that mine 1000 Genomes and other data for gene alleles. To that end we have been working on a set of tools that discovers, compares and displays gene haplotypes. The results are displayed on the gene description pages for the UCSC Genes track on hg19/GRCh37 but could easily be extended to other assemblies and gene tracks. For any protein-coding UCSC Gene, we display gene haplotypes from the 1000 Genomes data as amino acid variants. Optionally, one can choose to see the DNA variants. By default we display 1000 Genomes common (>=1% penetrance) non-synonymous variants. Optionally one can view rare haplotypes as well. After a few small additions to the working prototype on our development server, we intend release it this feature by June 2013.

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 thousand 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, 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 battery 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

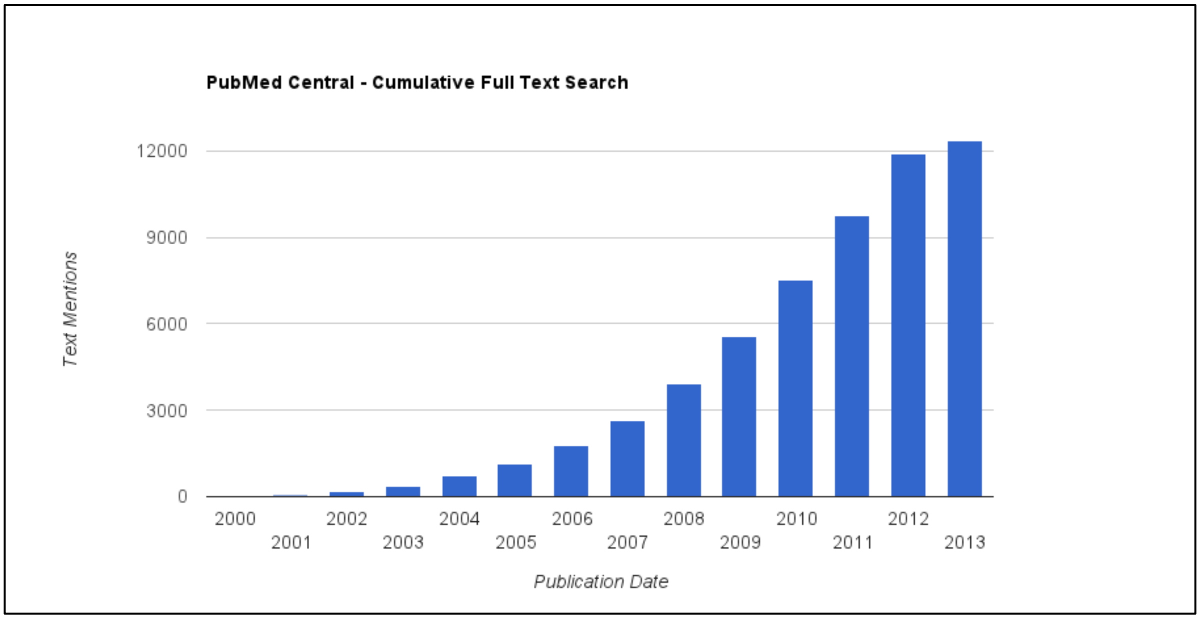
The UCSC Genome Browser has been cited thousands of times in the scientific literature (Table 5). Increasingly, however, our two most popular tools—the Genome Browser and BLAT—are taken for granted and used without citation, as may be appropriate for tools of their maturity. A search of the 15% of biomedical papers with PubMed Central full-text access reveals that the text string "UCSC + (genome OR browser OR database)" appeared in more than 2,200 papers in 2011 alone, implying a utilization rate of 14,000 publications per year (Figure 2).

Table 5. UCSC Genome Browser group paper citations tallied by Google Scholar.

Topic	Author, Year	2011	2013	Change
BLAT	Kent, 2002	2411	3395	41%
Genome Browser	Kent et al., 2002	1643	2566	56%
Browser database	Karolchik et al., 2003	969	1213	25%
Threaded Blockset Aligner	Blanchette et al., 2004	540	724	34%
Genome Browser update 2011	Fujita et al., 2011	9	510	5567%

Table Browser	Karolchik et al., 2004	302	497	65%
Genome Browser update 2008	Karolchik et al., 2008	364	448	23%
Chain/Nets (evolution's cauldron)	Kent et al., 2003	311	412	32%
Genome Browser update 2010	Rhead et al., 2010	167	405	143%
Genome Browser update 2006	Hinrichs et al., 2006	228	317	39%
Genome Browser update 2009	Kuhn et al., 2009	198	291	47%
Genome Browser update 2007	Kuhn et al., 2007	225	253	12%
Current Protocols	Karolchik et al., 2007	30	253	743%
Known Genes	Hsu et al., 2004	126	234	86%
28-way alignment	Miller et al., 2007	80	142	78%
Genome Browser extensions & updates	Dreszer et al., 2011	-	107	-
Archaeal Browser*	Schneider et al., 2006	44	71	61%
ENCODE resources 2007*	Thomas et al., 2007	26	61	135%
Gene Sorter	Kent et al., 2005	47	57	21%
Proteome Browser	Hsu et al., 2005	43	48	12%
ENCODE resources 2010*	Rosenbloom et al., 2010	24	43	79%
ENCODE resources 2012	Rosenbloom et al., 2012	-	43	-
Browser Tutorial	Zweig et al., 2008	17	34	100%
Genome Browser update 2013	Meyer et al., 2013	-	13	-
ENCODE resources 2013	Rosenbloom et al., 2013	-	6	-

Figure 2. 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.



Usage

As of April 2013 we are logging more than 4 million hits per week on our website from an average of 14,000 unique IP addresses per day (Figure 3). We know that many institutions appear to our system as a single IP address from behind a firewall; thus, the actual number of users is likely much higher than this. These figures also exclude usage on our many mirror sites. The usage of our UCSC Genome Browser Sessions’ functionality has continued to increase over the past year (Figure 4).

Figure 3. Upward trend in page hits at http://genome.ucsc.edu since inception in 2000.

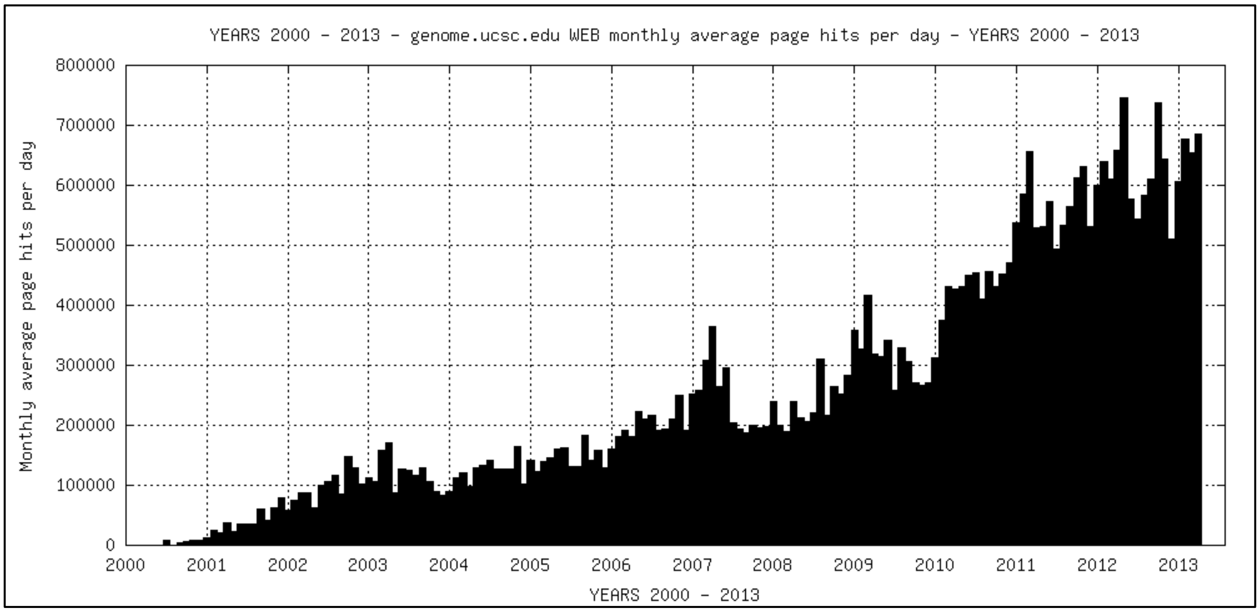
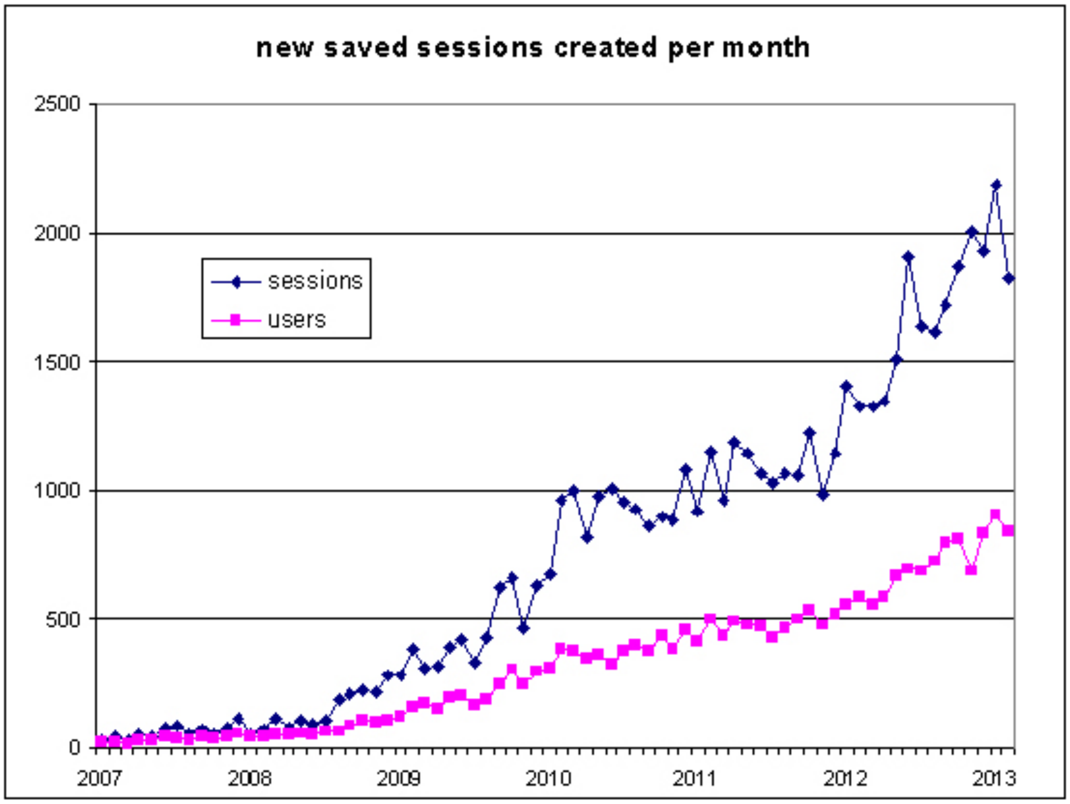


Figure 4. UCSC Genome Browser Sessions usage through April 2013.



Reliability and User Support

We maintain a very low page error rate on our website. Our public site typically logs very little downtime: in the past 10 months the site was down for about 7 hours total during peak usage hours (6AM-5PM PT, Monday-Friday). Our publicized mirror sites continued to provide Genome Browser access during these outages.

The UCSC Genome Browser staff continues to maintain two very active publicly accessible user-support mailing lists: genome@soe.ucsc.edu and genome-mirror@soe.ucsc.edu. The volume of both lists has greatly increased over the years, and the questions have generally become more sophisticated. The genome mailing list has 751 subscribers, receiving the messages in real time or as daily digests; the mirror list has 196. During the current grant period we responded to a combined total of 1,286 messages (1,121 threads) on our mailing lists to date, plus an estimated 1% additional messages that were sent directly to our staff off-list. Our mailing list staff responds to most inquiries within 1-2 business days. We also sponsor a low-volume list with 1,769 current subscribers for announcing new features and data, genome-announce@soe.ucsc.edu, and announce new features and interesting data sets on both our home page and to our more than 1,100 followers on twitter (@GenomeBrowser).

In August 2012 our mailing list was converted from a Mailman-administered list to a Google group as part of a university-wide switch to Google groups. This has permitted easier indexing and searching of previously answered user queries.

Outreach and Training

In addition to providing the UCSC Genome Browser, source code and associated tools, we also continue our outreach activities. This year we maintained our annual subcontract with OpenHelix, albeit at a reduced level compared to the previous award: we no longer support their attendance at trade shows. OpenHelix continues to provide two UCSC Genome Browser workshops per year in the United States and to produce Quick Reference Cards (QRCs) about the UCSC Genome Browser and Table Browser that we both distribute at workshops. These QRCs are used not only by individuals, but also in the classroom setting and as a quick introduction to the tools at conferences and shows.

We also provide a significant number of workshops offered by UCSC staff (Table 6), which remain quite popular. We presented a 1.5-hour workshop at the ASHG annual meeting, selling out the 200+-seat capacity in the first few days of registration. Similarly, two 1.5-hour workshops at the ESHG annual meeting filled the room of 300+ seats for each session in June 2012. We have been invited back to both meetings for 2013. The ASHG workshops are paired with sessions by the Galaxy project at Pennsylvania State University and the Ensembl project from the European Bioinformatics Institute (EBI) to provide a unified series of bioinformatics presentations.

Table 6. UCSC Genome Browser workshops presented by UCSC and OpenHelix, 2012-13.

Workshop Location	Date	Provider	Attendance (* expected)
Cytogenomics Array Group, Summer Microarray Workshop. Asheville, NC	Jul. 2012	UCSC	150
Center for Molecular Medicine. Univ. of Cologne, Germany (2)	Sep. 2012	UCSC	80
Ernst Klenk Symposium in Genetics, Cologne, Germany (2)	Sep. 2012	UCSC	50
ASHG annual meeting, San Francisco, CA	Nov. 2012	UCSC	250
Plant and Animal Genomes annual meeting. San Diego, CA	Jan. 2012	UCSC	30
Tufts University Medical Center	Jan. 2013	OpenHelix	21
Winship Cancer Center, Emory University, Atlanta, GA	Feb. 2013	UCSC	75
Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hosp, Cincinnati, OH (2)	Feb. 2013	UCSC	95
Mt. Sinai School of Medicine	Feb. 2013	OpenHelix	80
HUGO annual meeting - workshop organized by Sanger Centre. Singapore	Apr. 2013	UCSC	40
Stanford University, Palo Alto, CA	Apr. 2013	UCSC	70
Frontiers in Reproduction Course. Woods Hole, MA (2)	May, 2013	UCSC	20*
Faculté de médecine vétérinaire - Université de Montréal,	May, 2013	UCSC	80*

Saint-Hyacinthe, Que, Canada (2)			
ESHG annual meeting, Paris, France (2)	Jun. 2013	UCSC	700*

The seminars are an excellent avenue for users to give feedback on features and functions, which are then relayed to the UCSC Genome Browser development team. We also learn how our documentation could be improved. Even experienced users report that they learn about new features in our workshops.

We also give talks of a non-tutorial nature on our work and are active in local outreach, responding to requests for talks in various courses on campus and other venues involving groups not directly related to UCSC (Table 7). While all of the listed talks are about UCSC Genome Browser technology, not all were presented by staff funded by the Genome Browser grant.

Table 7. Sample of non-tutorial UCSC Genome Browser talks presented by UCSC staff, 2012-13.

Name of Talk	Conference/Location	UCSC staff	Date
Genome Browser Publications track	Biocuration conference. Washington D.C.	Maximilian Haeussler	Jul. 2012
On Building a Microarray Database	Cancer Cytogenomics Microarray Consortium (CCMC) meeting. Chicago IL	Robert Kuhn	Aug. 2012
Genome annotation with full-text articles	Biological Computation at Stanford (BCATS). Stanford CA	Maximilian Haeussler	Sep. 2012
Problems of full-text mining and copyright	Text and data mining seminar invited talk for European Commission. Bruxelles Belgium	Maximilian Haeussler	Dec. 2012
Navigating the UCSC Genome Browser	Murdoch Children's Research Institute. Melbourne Australia	Rachel Harte	Jan. 2013
Accessing ENCODE data with the UCSC Genome Browser	Joint Conference of HGM 2013 and 21st International Congress of Genetics. Singapore.	Robert Kuhn	Apr. 2013
Genome annotation with full-text articles	Biocuration conference and EBI/Hinxton invited seminar. Cambridge UK	Maximilian Haeussler	Apr.2013

OpenHelix maintains three online tutorials for the UCSC Genome Browser on their website (last updated May 2012): Introduction to the Genome Browser, Advanced Topics, and Associated Tools. In the past 12 months, OpenHelix logged 16,414 page views of their UCSC training materials landing page and recorded 1,609 downloads of their training materials, which include PowerPoint slides to match the video tutorial, slide handouts and exercises. They also expect to distribute ~500 QRCs at the Medical Library Association annual meeting, as they did last year.

The OpenHelix blog continues to be used to promote UCSC resources, including many appearances in their video tip-of-the-week (Table 8). These receive 200-500 hits each.

Table 8. UCSC Genome Browser-related posts at OpenHelix. If not directly about UCSC, they reference UCSC features or utilities.

Name of post	Date
Video Tip of the Week: ENCODE Data at UCSC (reminder)	Feb. 20, 2013
Video Tip of the Week: UCSC Genome Browser restriction enzyme display	Feb. 13, 2013
Video Tip of the Week: MotifLab workbench for TFBS analysis	Feb. 6, 2013
Video Tip of the Week: the new and improved OMIM®	Jan. 9, 2013
Video Tip of the Week: Chromohub, annotated trees of chromatin-mediated signaling	Jul. 4, 2012

Additional outreach activities include our presence on external Scientific Advisory Boards (SABs). Robert Kuhn serves on the SAB for the Ensembl project, and Kate Rosenbloom serves on the Human Proteome Project SAB. This international consortium aims to apply high-quality high-throughput proteomics methods to map the entire human proteome, and to link this data to transcriptome and genomic annotations.

Genome Browser Scientific Advisory Board

The UCSC Genome Browser project has an active SAB that currently consists of five members: Mary-Claire King (University of Washington), Tim Hubbard (Wellcome Trust Sanger Institute), Aravinda Chakravarti (Johns Hopkins University), Robert Waterston (University of Washington), and Joe Gray (Oregon Health Sciences University). Our last SAB meeting convened just prior to the start of this funding cycle. Throughout the year, we solicit advice from members of our SAB on an as-needed basis. During year two of this grant, we intend to hold a formal on-site SAB meeting.

Collaboration

In February 2013, the nine-person management team from the Ensembl project team came to UCSC for 1.5 days of cross-project consultation. Over the years, we have become as much collaborators as competitors and have found that we have much to learn from each other. We both face the challenges of rapidly expanding sequencing capability worldwide and the numerous large data sets that accompany it. Our two groups are uniquely challenged to both handle the data and to find understandable methods for displaying it.

In addition to group presentations, the site visit promoted numerous one-on-one interactions between people working on similar problems: human variation, multiple alignments and comparative genomics, infrastructure, etc. We identified areas where collaboration would benefit the two groups and the research community as a whole, and made concrete agreements to further our cooperation, including standardizing on certain data sets to reduce confusion among researchers. This builds on our previous agreement, for example, to use the same version of a genome assembly when releasing on a new organism.

A summary of our points of agreement and plans to collaborate includes:

- Communicate Ensemble's experience with their regulatory build to inform UCSC's. Consider displaying Ensembl's regulatory build as a Track Data Hub on the Browser.
- Confirm that the two groups are using the same SNP algorithm to generate subsets.
- Start using the same sequence ontology to describe variants.
- Host each other's comparative genomics alignments, and agree on the same 100 species/assemblies for the next big human-based multiple alignment.
- Port UCSC's Publications track to Ensembl.
- Solicit input from Ensembl and other external groups for the Track Data Hub mechanism.
- Help Ensembl use UCSC's gtfToGenePred converter to check their GTF format.

D. Plans

Our plans for the next year largely follow what we proposed in the grant proposal for year two. Tasks that were not completed during the first year of the grant, but will be completed during the second year, are marked with an exclamation point (**!**). Additionally, new tasks that were not listed in the original grant proposal are marked with an asterisk (*****).

Aim 1 – Software development

Yr.Qtr	Specific Task
2.1	• Combine multiple wiggles in new ways • Finish evaluation of network visualization tools
2.2	• Implement right clicks and sorts on column headers in Gene Sorter • Security expert attempts to break into site, fixes problems found • **!** Switch to cryptographic hash for user IDs to prevent URL-mangling security leaks
2.3	• Finish integration of network visualization tool • Package command-line tools with updated documentation • * Extend hubs to allow search for items within tracks, and for tracks themselves within a hub
2.4	• Dynamic conversion of genome browser tracks into Gene Sorter columns • Security expert attempts to break into site, fixes problems found • **!** Update index page with better graphics and pull-down menus

Aim 2 – Comparative genomics

Yr.Qtr	Specific Task
2.1-2.3	• Add genome browsers for three new species or updated genomes
2.4	• Add new multiple-alignment track for one set of assemblies • * Compare multiple genome alignment programs and possibly switch to new and better one • Update Assembly Hub specifications

Aim 3 – Import data (this was left as to-be-determined in the original grant proposal)

Yr.Qtr	Specific Task
2.1	• * Import selected data from Epigenomics Roadmap Project • * Import Human mutations from UniProt/SwissProt article curation project
2.2	• * Import selected data from next iteration of ENCODE project
2.3	• * Import Illumina BodyMap RNA-seq data • * Import LOVD family of locus specific databases from Leiden
2.4	• * Update data from UniProt, RefSeq, GenBank, Ensembl, GAD, OMIM, Cosmic, PFAM, and NGHRI's GWAS • **!** Add ORFeome tracks for mouse and zebrafish • **!** Add Segmental Dups tracks for several assemblies

Aim 4 – Gene sets

Yr.Qtr	Specific Task
2.1	• Update UCSC Genes for human genome
2.2	• Update UCSC Genes for mouse genome
2.3	• Update UCSC Genes for human genome • **!** Integrate CAGE data for more accurate promoter calling
2.4	• Update UCSC Genes for mouse genome • Integrate RNA-seq data to determine isoform abundance

During year two of this grant, we intend to publish a scientific paper describing the newly developed Assembly Hub technology. We plan to propose it as a standard for use across the genomics community. The Track Data Hub technology, from which Assembly Hubs have grown, has already been widely adopted by the community. We will work with the Ensembl project, NCBI, and others to make sure that the formats and features are in alignment with their needs. Ensembl has agreed to the concept, and is anxious to see it working.

E. Publications

Refereed journal papers co-authored by our group

Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. *Brief Bioinform.* 2013 Mar;14(2):144-61. PMID: 22908213; PMCID: PMC3603215
http://bib.oxfordjournals.org/content/14/2/144.long

Abramyan J, Badenhorst D, Biggar KK, Borchert GM, Botka CW, Bowden RM, Braun EL, Bronikowski AM, Bruneau BG, Buck LT et al. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.* 2013 Mar 28;14(3):R28. PMID: 23537068

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. "ENCODE data in the UCSC Genome Browser: year 5 update." *Nucleic Acids Res.* 2013 Jan 1;41(D1):D56-63. PMID: 23193274 PMCID:PMC3531152
http://nar.oxfordjournals.org/content/41/D1/D56.full

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, and Kent WJ. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*. 2013 Jan 1;41(D1):D64-9. PMID: 23155063; PMCID: PMC3531082
http://nar.oxfordjournals.org/content/41/D1/D64.long

ENCODE Project Consortium *et al*. "An integrated encyclopedia of DNA elements in the human genome." *Nature*. 2012 Sep 6;489(7414):57-74. PMID: 22955616 PMCID: PMC3439153
http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html

Lowe CB, Haussler D. "29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome." *PLoS One.* 2012 Aug;7(8):e43128. PMID: 22952639 PMCID: PMC3428314
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043128

Book Chapter co-authored by our group

Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. *Curr Protoc Bioinformatics.* 2012 Dec;Chapter 1:Unit1.4. PMID: 23255150
http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0104s40/abstract

Posters co-authored by our group

Raney BJ, Barber GP, Hinrichs AS, Goldman M, Roe G, Rhead B, Guruvadoo L, Fujita PA, Malladi VS, Zweig AS, Karolchik D, Haussler D, Kent WJ. Remote Data Track Storage for Viewing on the UCSC Genome Browser. Genome Informatics 2012. Cambridge, UK. Sept 2012.

Clawson H, Raney BJ, Barber GP, Hinrichs AS, Fujita PA, Zweig AS, Karolchik D, Kent WJ. Assembly Hubs – Display novel genome sequence using the UCSC Genome Browser. Biology of Genomes. Cold Spring Harbor, NY. May 2013.

Rhead B, Hinrichs AS, Dreszer TR, Raney BJ, Kuhn RM, Zweig AS, Karolchick D, Kent WJ. New variation resources at the UCSC Genome Browser. Biology of Genomes. Cold Spring Harbor, NY. May 2013.

F. Project-Generated Resources

The primary resources generated from this project:

- UCSC Genome Browsers for multiple assemblies of 91 species (http://genome.ucsc.edu/)
- 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 Genes sets for human (hg16, hg17, hg18, hg19/GRCh37), 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
- GenomeWiki site (http://genomewiki.ucsc.edu/)
- UCSC Genome Browser European Mirror Site hosted at the University of Bielefeld, Germany (http://genome-euro.ucsc.edu/cgi-bin/hgGateway)

2012-13 DIVERSITY ACTION PLAN (DAP) PROGRESS REPORT

The NHGRI DAP is implemented on the UCSC campus as the CBSE Research Mentoring Institute (RMI), a research education program that supports underrepresented minority (URM) students in both undergraduate and graduate (pre-doctoral) educational training and advances them toward successful careers in genomic science or its ethical, legal, and social implications (ELSI).

This progress report addresses 7 main categories of activity over the 2012-13 cycle:

- Progress of funded students
- Retention and professional development
- Research Training
- Education about genomic science
- Outreach and recruitment
- Data collection and tracking
- Evaluation

Overview

We have continued to develop and take advantage of partnerships with staff, faculty, student-run campus organizations, and leadership at local institutions to help us reach out to and recruit minority student groups, to publicize our scholarship and fellowship program, and to create greater awareness about genomic research. We have maintained a full complement of professional development workshops for our undergraduate and graduate awardees, and included some new personal development content as well. All of our awardees are engaged in mentored training in various aspects of genomic research; in the past year our students trained in the following departments: Computer Engineering, MCD Biology, Microbiology and Environmental Toxicology, Chemistry and Biochemistry.

We had six metric-driven objectives, consistent with the objectives established in our 2007 application:

1. Identify, recruit, and select 10 eligible scholars (six undergraduate awardees and four graduate fellows) each year who demonstrate a desire for doctoral education or a career in genomics or ELSI and who fit into our target group.

2. Prepare, retain, and graduate at least 90% of the scholars in our program to the next appropriate academic level (either graduate school or a career in academics or research) by providing academic support services and intensive faculty mentoring.

3. Ensure that 90% of the scholars will achieve and maintain a minimum GPA of 3.0 (on a 4.0 scale) during their active enrollment in the program.

4. Ensure that 100% of the scholars will participate in mentoring activities with a faculty member individualized to the academic area of interest.

5. Ensure that 100% of the graduating undergraduate awardees will take steps toward the next academic or career level by applying to at least one graduate program or a research-related position in an academic or industry setting.

6. Establish and maintain comprehensive program and participant files and fiscal records to effectively document the outcomes of the services provided and to facilitate periodic internal review.

We met goals 1-4, although goal 1 happened later than usual due to the late award date, we were not able to fund new awardees during summer quarter, and we diverged from the programmed number of undergraduate and graduate slots in order to work with the available pool of students. We reached an 81.3% success rate for goal 5. We increased the comprehensiveness of our record-keeping in response to goal 6. Since 2010 we also participate in the DAP data collection project initiated by the NHGRI through the Data Analysis and Coordinating Center (DACC) at Washington University, St. Louis, and are 100% current with entering all our data into that system. The DACC will maintain a centralized database for all NHGRI DAP programs and create quantitative reports that measure program success.

Composition of 2012-13 RMI Program Participant Pool

We reached out to individuals who are from underrepresented racial and ethnic groups, working particularly to identify candidates who have faced barriers to educational opportunities, such as physical or learning disabilities, have low-income backgrounds, or are the first in the family to attend college or graduate school. Table 1 shows the composition of our applicant and participant pools.

Table 1. Total number of individuals who applied, were interviewed, and who accepted and participated in the UCSC RMI program during the previous award period. Many of the accepted participants fall into more than one target group, and all meet the racial group criteria defined by NHGRI.

RMI Program	Target Group	# Applied	# Interviewed	# Accepted and Participated
Undergraduates	Target Racial Group	8	8	4
	Disability	None reported		
	Disadvantaged	6	6	4
	Total Supported 4			
Graduate Students	Target Racial Group	9	9	7
	Disability	1	1	1
	Disadvantaged	7	7	7
	Total Supported 7			

In summary, in the 2012-13 academic year, we supported 4 undergraduates and 7 graduate students.

Undergraduates Supported by the RMI Program

During the 2012-13 cycle, the UCSC RMI program has supported 4 undergraduate scholars, all of whom met the NIH inclusion criteria. All four students are currently in STEM labs receiving mentored research training, and doing well academically; 3 will be seniors next year, and one will graduate in June. The graduating student has been invited to continue to work in his lab until he leaves for grad school (he did not apply in 2012, so he will apply this fall). The senior student presented a poster at ABRCMS 2012; all 4 students will present a poster at the 2013 UCSC Undergraduate Research Symposium on June 10, 2013. Based on data from 2007-12, our success rate for students continuing to next educational level is 81.3%; however, the program director is now more closely advising the graduates on their applications for the 2012-13 graduate admissions cycle and in the long run, we expect to exceed the 90% objective for this cohort.

Graduate Students Supported by the RMI Program

During the 2012-13 cycles, the UCSC RMI program has supported 7 graduate scholars, all of whom met the NIH inclusion criteria. Four of the students are receiving full support, and 3 are receiving supplemental support in the form of research materials or travel awards for professional conferences. All but two of our graduate fellows have completed coursework and preliminary examinations, and one has advanced to candidacy. The advanced student, a computer engineer, is working on improving sequencing speed of the UCSC Genome Browser. He has already received a job offer from a local industry affiliate in Silicon Valley, but is actively exploring other postdoctoral opportunities in academic research contexts.

All RMI graduate fellows are required to formally present at least once a year as appropriate to their discipline. All graduate fellows present at the campus graduate research symposium held annually in May, plus one additional conference, either SACNAS or a discipline-specific conference. As they advance, we encourage them to present more at professional conferences to enhance career experience. Grad presentations since September 2012:

Pamela Watson (MCD Biology) Department Research Conference, September 2012
Poster title: “Regulation of respiratory arsenic reduction in *Shewanella* sp. ANA-3”

Rigo Dicochea (Computer Engineering), UCSC Multi-University Research Network Conference, November 2012; Global Foundries Research Symposium, March 2013
Presentation: “Architectural techniques for increasing the computational efficiency of genomic sequencing algorithms”

Prestina Smith (MCD Bio) Gordon Mammary Gland Conference in Lucca (Barga) Italy, June 2012
Poster title: “VANGL2 and planar cell polarity in the mammary gland”

Retention and Professional/Personal Development Programming

Over the course of the 2012-13 grant cycle, we have continued to improve an already rigorous professional development component, offering at least three programs for undergraduates and two for graduate students each quarter (Table 2). In the spirit of collaboration and partnership, these workshops are often co-hosted with other URM student success programs such as EOP, the Ethnic Resource Center, the Multicultural Engineering Program (MEP), the Graduate Division, the Career Center, and the UCSC Minority Access to Research Careers (MARC) and Initiative for Maximizing Student Diversity (IMSD) programs. We also partner with student groups such as Women in Science and Engineering (WISE), the UCSC Chapter of the Society for the Advancement of Chicanos and Native Americans in the Sciences (SACNAS), and the Graduate Student Commons. Our program director uses these events as opportunities to promote the RMI program by giving a short presentation and disseminating material at every workshop.

Table 2. RMI Professional Development Workshops (2010-11)

Undergraduate Workshops	**Attendance**
Fall 2012	
Demystifying Graduate School	21
Summer Research Programs	18
Graduate Admissions Preparation: Letters of Recommendation and Personal	30
Winter 2012	
Demystifying Graduate Research	12
Mentorship 101: How to Get the Mentoring You Want & How to Be a Good Mentor	18
Graduate School Admissions: The Campus Visit and Interview	9
Spring 2011	
Scientific Writing: Writing a Scientific Abstract	10
Poster Presentation Skills (May 10)	TBD
GRE Exam Preparation (May 17)	TBD
Graduate Workshops	**Attendance**
Fall 2012	
The Impostor Syndrome	16
Work-Life Balance (co-sponsored by WiSE UCSC)	40
Writing a Curriculum Vitae and Cover Letter (co-sponsored by Career Center)	8

Winter 2013	
Mentorship Training Seminar –day long (with CSU Monterey Bay)	37
Post-doc Roundtable	14
Spring 2013	
Skills for Success: Cultivating Positive Mental Habits (co-sponsored by MEP)	40
Career Planning: The Individual Development Plan	8

Research Training

All students—undergraduate and graduate—are in labs and receiving mentored research training from a faculty member, post-doctoral researcher, or graduate student (in the case of undergrads). The nature of training depends upon the department the student is based in and the research agenda of the PI.

Genomic Education

All students receiving financial support from the RMI program are also receiving hands-on research training. Additionally, the RMI program director alerts students to the many existing events, such as journal clubs, colloquia, and scientific talks by faculty or visiting researchers. Our co-PI, Robert Kuhn, gives workshops on the history of the Human Genome Project and how to use the Genome Browser. The RMI program also organizes an annual DNA DAY event in April of every year (April 25 for 2013). The 2013 DNA DAY activities included a poster display for non-specialists with information about the “history of DNA” (from Darwin to the completion of the assembly of the human genome in 2003); the distribution of literature, and hands-on strawberry DNA extraction.

Outreach and Recruitment

On campus, we reach out to undergraduate student groups by giving presentations in the early fall at divisional and program orientations, by tabling, and by networking with partner offices. Off-campus recruiting focused on increasing our visibility to prospective students in the Northern California region and on cultivating institutional partnerships with program managers and campus leaders at regional community colleges and four-year institutions. Most of these schools are within 30-90 minutes of the UCSC campus.

We also recruit URM graduate student applicants at the annual meetings of SACNAS, ABRCMS, NSBE, and the California Forum For Diversity in Higher Education. When applications come in, we assist graduate admissions committees in the School of Engineering and the Division of Physical and Biological Sciences in the review of applications. Our program director identifies diversity candidates for the committees. Where appropriate, she assists the department in writing nominations for the Cota-Robles Graduate Fellowship, a prestigious diversity award offered by the Chancellor’s Office and the Division of Graduate Studies. Where applicants express interest in genomic research, we offer a financial incentive to help recruit the individual (depending on budget). When URM recruits visit our campus, the director meets with them individually and shares information about diversity initiatives on the camps as well as available resources.

Data Collection and Tracking

We collect data every year by sending a survey (as email attachment) to all current and past students. The questionnaire tracks professional development (post graduate employment, publications, presentations, etc.). We are working closely to share all data with the NHGRI’s chosen data collection center (DACC) at Washington University, St. Louis, to maintain accurate records and to create survey questions that align with the database. All undergraduates meet with the program director at least once per month, and the director also tracks students’ progress by contacting the faculty mentor. All graduate students meet with the director at least twice per quarter and are required to submit a progress report to the RMI office at the end of each quarter. This report is 2-3 pages and details the progress of the student’s research as well as the status of accomplishing important benchmarks (qualifying exams, presentations, publications); the report also provides an opportunity for the student to express concerns about challenges they may encounter.

Evaluation

We solicit feedback from our students after workshops in the form of questionnaires and surveys, and the director records anecdotal information when she meets with students in person. Exit surveys are also completed when students graduate. In March of 2013 we retained the services of a professional evaluation agency, Shattuck Evaluation; a comprehensive evaluation of the program is currently underway.