## RESEARCH PLAN

## Specific Aims

This is a proposal to maintain, extend, and enhance the UCSC Genome Browser website and associated tools and databases. Our aims are to build useful software, load useful data, support comparative genomics, and build comprehensive gene sets.

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

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

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

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

## Research Strategy

#### Overview

The primary purpose of our work is to facilitate an understanding of the human genome by biomedical researchers worldwide through the use of the UCSC Genome Browser, its annotation databases, and associated tools at <http://genome.ucsc.edu>. This now mature informatics resource originated in the year 2000 for use by the International Human Genome Sequencing Consortium while assembling the first drafts of the human genome ([1](#_ENREF_1)). It has since grown to encompass a broad set of data and tools widely used by the consortium, its numerous successors, and the molecular biology and biomedical research communities at large.

Notable project achievements over the past eleven years include the Blat tool for quickly aligning RNA, protein, and translated sequences to genomic DNA ([2](#_ENREF_2)); a pipeline for generating pairwise ([3](#_ENREF_3)) and multiple vertebrate genome alignments ([4](#_ENREF_4),[5](#_ENREF_5)); the development of a widely used set of human gene models ([6](#_ENREF_6)); and the development of practical techniques for handling visualization of terabyte-sized distributed next-generation-sequencing data sets ([7](#_ENREF_7)).

Perhaps most importantly, the work has resulted in the UCSC Genome Browser itself ([8-15](#_ENREF_8)), a web-based tool that integrates the genomics annotations developed at UCSC with the efforts of hundreds of other genomics scientists into a fast, robust, reliable display that is driven by one of the most comprehensive databases in the field (Figure 1).

Major additions and enhancements to the genome browser tools and underlying data support during the current grant period include these:

* Significant navigation and configuration improvements to the annotation tracks page through support for clicking and dragging on the image: drag-and-zoom to highlight a region and zoom to it, drag-scroll to pan the image horizontally, drag-reorder to quickly reposition tracks or groups of tracks within the image, and right-click to display context menus to directly target and configure track features in the browser image.
* A track search tool that allows a user to find data sets by entering a few key words or controlled vocabulary terms.
* Gene search functionality on the gateway and annotation tracks pages that allows the user to enter a gene name to jump to and suggests names as the text is typed.
* A configurable navigation option that allows the user to easily jump between features or exons in a track.
* A reverse button that changes the view of negative-strand genes to the 5’-to-3’ direction.
* Multi-view display, with a multiple wiggle-track overlay that displays multiple data sets in a single visualization track for focused comparisons.
* Support for several new compressed, indexed file formats that allow remote hosting of very large data sets: BigBed, BigWig, binary alignment/map (BAM), variant call format (VCF) and genome variation format (GVF).
* Support for remote Track Data Hubs.
* Performance enhancements to speed up the track display, especially when viewing large chromosomal regions.

Description: basicBrowserShot.tiff

Figure 1. The UCSC Genome Browser displaying the region that includes the superoxide dismutase 1 (SOD1) gene with the default set of 10 visible tracks. The user can select additional tracks to display from a list of 156 tracks on this assembly of the human genome (GRCh37/hg19). The tracks can be displayed at various degrees of information density. The UCSC Genes track is shown in a “packed” mode where all of the items in the track are visible, while the RefSeq Genes track underneath it is in a “dense” mode where all items are drawn on top of each other. Clicking on a dense track will open it so that each item is visible. Clicking on an individual item brings up another page with detailed information. Tracks can be further filtered and configured through a context menu that is displayed by right-clicking (or ctrl-clicking on a Mac) on a track.

A notable recent addition to the genome browser, Track Data Hubs, holds great promise for further decentralizing our data model, an effort that started with support for remotely hosted BigWig, BigBed, and BAM tracks. Data hubs allow collaborators to manage their own data sets on their local servers while being able to configure, view, and share them with all genome browser users alongside native annotation tracks. The hubs are a key part of our efforts to control costs in the face of an exponentially increasing demand from scientific consortia and individual labs wishing to display their data on our browser.

From its original focus on the early drafts of the human genome, the genome browser database now offers genomic data for 53 organisms, many with multiple assemblies. The latest mouse assembly, NCBI37 (mm9), was added in 2007, followed by a new human genome assembly, GRCh37 (hg19), in 2009. The human and mouse genomes are the most heavily annotated of the entire genome set. On the human genome, annotation data sets (“tracks”) cover conservation and evolutionary comparisons, gene models, regulation, expression, epigenetics and tissue differentiation, variation, phenotype, and disease associations.

During the current grant cycle, we added 93 new tracks to the newer human assemblies, updated 33 existing tracks, and “lifted” approximately 30 tracks from the hg18 assembly to hg19 coordinates to increase the robustness of the annotation set on the latest human assembly. During this same time period, the mouse browsers acquired 36 new tracks and 18 updates. We incorporate the GRCh37 patches, haplotypes, and alternate sequence from NCBI as they become available as a separate track on top of the main human assembly. In addition, the browser serves as a platform for a host of tracks from the ENCODE project, which is funded by another mechanism. Table 1 lists selected annotation track highlights during the period from 2007 to the present.

Table 1. Significant track releases and updates during the 2007-present grant period.

|  |
| --- |
| **Gene models for human and model organism genomes** |
| 4 new UCSC Genes sets for human (2 each on hg18 and hg19), 2 for mouse |
| Regularly updated CCDS, Ensembl, N-SCAN, Vega, RefSeq, tRNA, and TransMap annotations on human and other organisms |
| New links from UCSC genes to KEGG, Hinv, LS-SNP, Chimera, GeneNetwork, HuGE, GeneTest, GeneReviews |
| Enabled isPCR on UCSC Genes to amplify exons for human and mouse assemblies |
| New software tool for generating FASTA-format amino acid alignments for UCSC Genes |
| **Large-scale sequence alignments** |
| Conservation tracks for 17 assemblies, including multiple alignment and conservation of 46 vertebrates relative to the human genome |
| Addition of phyloP to augment phastCons as a measure of sequence conservation |
| Downloadable quality-annotated MAF files and FASTA amino acid alignments using coding frames from a gene track |
| **Human genetic polymorphism and disease association** |
| Notable new and updated tracks: GWAS Catalog, SNPs, DGV, HGSV, Genome Variants, HGDP, Segmental Duplications, DECIPHER, OMIM |
| Neanderthal and Denisova tracks aligned to the human genome |
| Enhanced display of dbSNP data with three new data subsets: Common SNPs, Flagged SNPs, and Multiple SNPs |
| **Model organism databases** |
| Interaction with modENCODE to share fly and worm data |
| Links from UCSC Genes details page to most model organisms supported by NHGRI |
| **Transgenic studies in model organisms** |
| Regular updates of International Gene Trap Consortium data on mouse genome |
| Regularly updated International Knockout Mouse Consortium data mapped to mouse and orthologous position in human |

We collaborate with many high-throughput data producers and computational biologists worldwide to incorporate their data. Our partner groups and consortia include the ENCODE Consortium, the Roadmap Epigenomics Mapping Consortium, the 1000 Genomes Project, the FaceBase Consortium, the International Standards for Cytogenomic Arrays (ISCA) Consortium, and GenomeSpace.

Our site is heavily used, and the usage is continuing to grow, more than doubling over the last four years (Figure 2). In the past six months, our website logged an average of 16.9 million hits per month, more than double the average of 8.3 million hits per month logged during the same six months of 2007. Additionally, our website currently logs visits from about 152,000 unique IP addresses per month, as compared to about 61,000 unique IP addresses per month in 2007. 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. Also, these figures exclude the usage on our many mirror sites. As attested to by our letters of support, the genome browser is one of the most widely used sources of genomics information within the scientific community.



Figure 2. Average page hits per day on the genome.ucsc.edu website graphed over the lifetime of this resource. These hits include simply the page access itself, and not the images, style sheets, and other objects that constitute the page. After ten years in the public domain, our usage continues to grow, perhaps even exponentially.

The genome browser is heavily cited in the scientific literature. Google Scholar notes 28,230 citations of 63 papers relating to the work supported by this grant (Table 2). 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. For example, a parsing of the 15% of biomedical papers with PubMed Central full-text access reveals that the text string “UCSC Genome” appeared in more than 1,400 papers in 2009 alone (the last year for which PMC is fully populated), implying a utilization rate of 10,000 publications per year.

In this renewal proposal we describe our plans to maintain, extend, and enhance the genome browser website, tools, and databases over the next five years. The proposal is based on four specific aims that encompass building useful software, loading useful data, supporting comparative genomics, and building a good gene set. As with most long-term plans, we anticipate that we may occasionally have to accommodate the unexpected or to eliminate parts that become obsolete.

Table 2. Citations of the UCSC Genome Browser and associated work, as reported by Google Scholar.

| **Type of paper** | **# Papers** | **# Citations** |
| --- | --- | --- |
| Large consortium papers in which UCSC participated materially | 12 | 13,412 |
| Papers written specifically about the web works that include the UCSC Genome Browser | 23 | 7,504 |
| Comparative genomics papers | 17 | 5,129 |
| Transcribed genes papers | 7 | 1,100 |
| Papers related to identifying regulatory regions | 4 | 1,085 |
| **Total** | **63** | **28,230** |

In our planning, we have taken care to constrain our costs and find alternative external sources of funding for some portions of the project. In addition to the Howard Hughes Medical Institute (HHMI) funding that has traditionally supported Dr. Haussler and part of the computer systems administration staff, we have been able to transfer some of the computational infrastructure costs to other funding sources, resulting in a sharp drop in hardware costs. This savings has partly offset the expiration of our supplemental ARRA funding. We have also made significant efforts at automation and efficiency. However, both the number of genome browser users and the number of data sources have more than doubled over the period of our current grant, entailing larger costs for user support and data integration that will necessitate a modest increase in our core genome browser staffing.

#### Research Resource Production

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

We will continue to develop and maintain useful software tools for viewing and analyzing genomic data resources. This includes adding visualization and search capabilities for new types of data and accommodating the growing number and sizes of data sets on the UCSC Genome Browser website. In addition to the website, we will continue to develop command-line tools for alignment, mapping, and statistical analysis. We will develop web services as well as file-download approaches to distributing our data. When appropriate, we will integrate externally produced tools to seamlessly interface with our web-based and command-line tools. We also plan to focus on security efforts to protect the privacy of the data uploaded by users. Our goal is to produce tools that are easy to learn and use, that work reliably and rapidly, and that address the needs of a broad base of users. We have a talented staff of software engineers and a well-honed software development process to help us achieve these goals.

##### ***Significance***

As automated laboratory devices enable scientists to generate increasing amounts of data, it is essential to have strong software tools that distill the data and present the results to the human eye. Ideally these tools should support the visualization of the user’s own data as well as that which is already present in existing databases, and they should meet reasonable expectations of security when the data is uploaded to the site. Databases that respond quickly and reliably to a wide range of queries and graphical browsers that can display genomic views ranging from individual bases up through entire genomes are in many ways as important to the current generation of biomedical scientists as well-stocked libraries and well-crafted microscopes were to the scientists of previous years. Our software contribution is not limited to what is visible on the genome browser website; the Unix command-line interface to our tools plays a significant role in the analysis pipelines at institutions throughout the world.

##### ***Innovation***

On a mature resource project such as the UCSC Genome Browser, innovation is necessary to keep up with advancements in technology and the changing needs of the science community, but it should be done with discretion. We intend to innovate in five aspects of this aim during the next grant cycle: a) increasing the interactivity of the website, b) adapting to new types of data, c) accommodating higher volumes of data, d) enhancing the security of user-uploaded data, and e) creating command-line and web services tools for other bioinformaticians. Our overarching goal is to make our tools more accessible to novice users and new data providers, while not requiring current users to relearn the system.

#### a. Increasing website interactivity

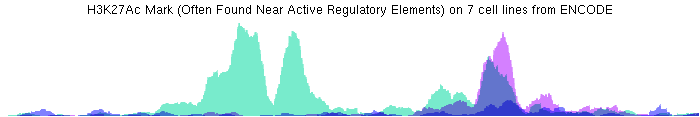
When the genome browser was initially designed in 2000, the prevailing web technology limited the amount of interactivity possible between the user and the website: each click was a discrete event that was followed by a refresh of one or two seconds. We were able to provide a lot of functionality within this framework, as long as we paid special attention to performance tuning.

Web users are now accustomed to the broader range of interactivity provided by modern web technologies, such as being able to click and drag with the mouse and access context menus through a right click (or ctrl-click). In response to this, we recently added support for drag-zooming, drag-scrolling, drag-reordering, and right-click context menus in the annotation tracks interface. We intend to expand the drag-reordering functionality to work at multiple levels in the track–subtrack hierarchy. We also plan to incorporate other web interface conventions that have become standard in recent years. For example, we intend to replace the current menu bars with drop-down menus offering an expanded range of options that are functionally grouped to facilitate easy access. We would like to add hierarchical list controls to help manage our continually growing list of tracks, allowing users to dynamically collapse and expand lists according to their research interests.

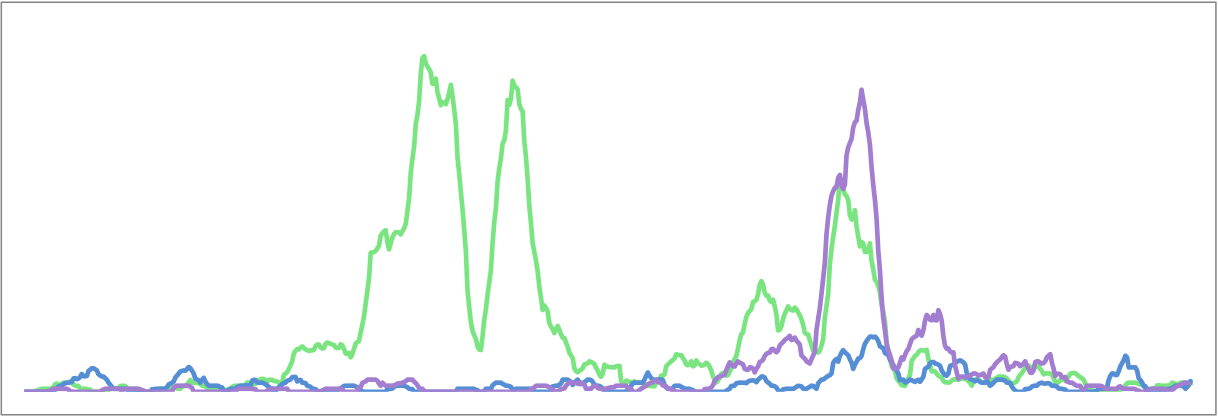
We also plan to provide our users with more options for combining data. The genome browser is hosting increasing amounts of continuous-valued “wiggle” data. Recently we developed a method to display transparent overlays of multiple wiggle tracks, which has proven popular. We would like to support other ways of combining graphs in the same display area, such as stacking graphs and displaying line graphs with different colors and line patterns (Figure 3). We also want to give our users the option to combine graphs selected from diverse sources, rather than being limited to simply toggling graphs off or on within defined container track.

For many research purposes it is more useful to think of the genome as a large collection of genes rather than as long stretches of DNA arranged into chromosomes. This gene-centric view has become increasingly important as the set of human genes becomes better defined. Within the genome browser tool set, the Gene Sorter provides a gene-set view, essentially a large dynamic table in which each row corresponds to a gene (or transcript) and the columns show information about the gene, such as its HUGO name, a short description, a list of Gene Ontology (GO) annotations, expression levels across various tissues, and so on. We plan to modernize and extend the Gene Sorter in a number of ways. We will support sorting within columns in response to a click on the column header, and we will implement a pop-up context menu when the user right-clicks (or ctrl-clicks) on an item, similar to the context menus recently implemented in the genome browser tracks display. We will also make it easier for users to convert genome browser tracks into Gene Sorter columns. For example, the Gene Sorter could display the average signal value inside a gene’s exons or list the transcription factor peaks that may be associated with a gene by proximity.

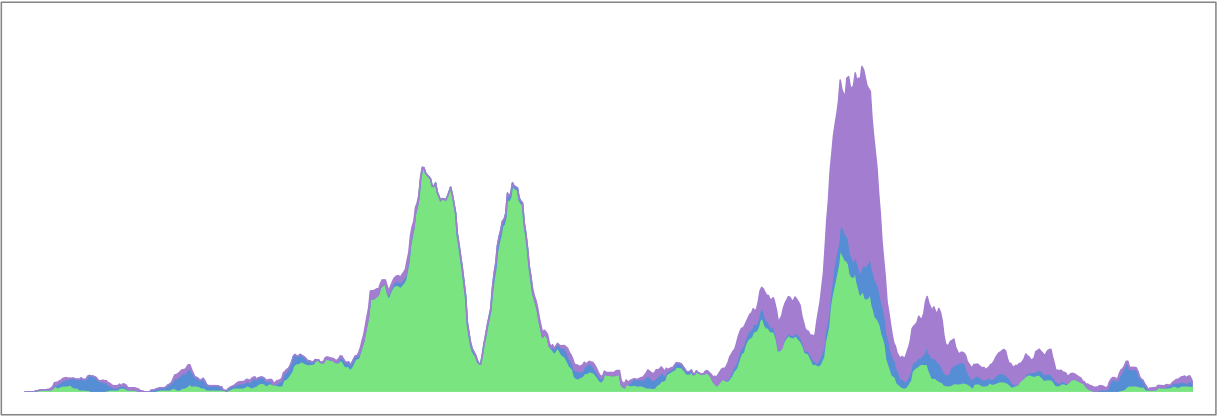
Another useful approach for visualization and analysis is to draw pathway graphs in which genes are the nodes and connections between genes are the edges. Due to the large size of such graphs, these displays are most useful when they are interactive, allowing the user to easily control which subsets are visible. Because we lack network visualization expertise within our staff, we plan to evaluate several web-based network visualization tools, and then integrate a suitable tool as seamlessly as possible into our own web works.



**a**



**b**



**c**

Figure 3a. Existing transparent overlay display of multiple data sources (in this case H3K27Ac ChIP-seq data from ENCODE on three cell lines), with two proposed alternatives: 3b. a simple line overlay, and 3c. a stacked area chart.

###### **b. Adapting to new types of data**

As genomics technology advances, it produces new data types, many of which require new visualization techniques for optimal display in the genome browser. Though it is difficult to predict exactly what new types of data will emerge in the next five years, we can make educated guesses, while reserving some additional bandwidth to address the inevitable surprises as well.

Technology has reached the point where it is easier to distinguish the two nearly identical versions of each chromosome carried in diploid species such as human. The emerging single-molecule sequencing technologies will facilitate this, but even in the absence of long single-molecule reads, “phasing” techniques for separating out maternal and paternal alleles are progressing rapidly. At UCSC, we have developed some preliminary phased displays in our Personal Genome Variants track and are adding display support for VCF, as shown in the display of 1000 Genomes project pilot data in Figure 4. It is important that we continue to develop techniques for displaying diploid genomes and alternative haplotypes both within and outside the reference human genome.

We plan to improve our data visualization in 5-C, ChIA-PET, and related displays where two regions, potentially on different chromosomes, are connected. The current browser display—similar to that of a two-exon gene—works well only when the connections are sparse and relatively local. We would like to implement diagrams similar to Circos plots ([16](#_ENREF_16)) to provide a high-level view of these types of data. Circos-style plots can also be useful for displaying structural rearrangements of the genome.

Many classes of data, such as mass spectrometry proteomics data, RNA-seq data, and protein domain information, are available primarily or exclusively over exons. A long-standing request from many genome browser users is a method for omitting or condensing introns and intergenic regions within the browser display. Following a recent round of performance optimizations, we have determined that it will be feasible to implement this option with reasonable performance.

We currently host a large amount of chromatin immunoprecipitation sequencing (ChIP-seq) data on transcription factors, many of which bind to specific DNA motifs. We would like to extend the display of this data type to show not only the ChIP-seq peak, but also the location of the motifs (if any) within the peak. On the associated details pages, we plan to show a sequence logo ([17](#_ENREF_17)) of the motif and a score indicating how well the particular sequence matches the motif.

Description: pilotVariants.pdf

Figure 4. Experimental display of variation in Japanese and Chinese populations, including a “local phasing on the fly” approach. The top track shows the UCSC Genes. The middle track shows a single-pixel row for each phased haplotype (two per person). The order of the haplotypes is determined by a hierarchical agglomerative clustering approach that expands out from the center SNP, which is surrounded by a black box. The local haplotype blocks tend to become evident from the clustering. Blue is used where the SNP allele matches the reference genome; red indicates where it differs. The bottom track shows all the SNPs observed in the population.

###### **c. Adapting to higher volumes of data**

In the last few years, the success of high-volume short-read sequencing machines has resulted in the emergence of several significant new data types. The sheer volume of data produced by this technology has made it impractical for users to upload their data for display in the genome browser, and has made it problematic for us to import data sets from large consortia, such as the 1000 Genomes Project or the Roadmap Epigenomics Project, into our databases at UCSC.

In response to this problem, we have developed methods that allow the data to reside at its location of origin, with only a summarized, compressed version of the current browser view transmitted to UCSC, where it is then cached to eliminate the need to resend. As part of this decentralization effort, we have developed the BigWig and BigBed file formats, and have added support for the BAM format using an extension of the samtools library developed by Heng Li at the Wellcome Trust Sanger Institute. Short reads are typically aligned to the genome and the results stored in BAM format. In many applications, such as the sequencing of RNA, chromatin immunoprecipitation, or DNase products, these alignments are converted into depth-of-coverage graphs over the genome that are stored in BigWig format. Higher-level analysis often produces a list of regions, with or without gaps such as introns, which can be stored in BigBed format along with confidence scores and additional user-defined data.

While BAM, BigWig, and BigBed formats make it possible to browse an individual high-volume sequencing experiment, there is a growing trend in which the smaller labs produce tens of experiments, the larger labs produce hundreds, and the scientific consortia generate thousands of experiments. Since the human eye cannot easily interpret thousands of tracks simultaneously, we have developed techniques to display summaries of many tracks simultaneously and to search large sets of tracks for those of particular interest. These visual summary techniques include the transparent overlay of BigWig coverage graphs on top of one another and the clustering of features from different experiments.

The recently introduced data hubs allow data producers to organize their own data into a hierarchy of tracks in a more stable, flexible, and integrated manner than the existing custom track mechanism that was designed with smaller data sets in mind. Tracks on remote hubs display somewhat more slowly than locally hosted tracks, but we have invested substantial effort in engineering the data hub to minimize performance issues. Popular remote tracks are cached, and therefore display as quickly as native tracks. We can fetch uncached data from 10 or 20 remote data tracks in just a second or two through parallelization of the network requests.

We aware of the problems associated with the needless proliferation of genomics file formats, and are therefore careful to introduce a new format only when absolutely necessary. For instance, in our migration toward a compressed, indexed remote file strategy, we opted to use the existing BAM format for short-read alignments. We also helped establish the standards for the VCF format, ensuring it is robust enough to encompass all possible chromosome-rearrangement scenarios. When we find it necessary to develop a new file format, we also provide public domain C libraries with documented APIs and command-line tools to use with that format. For example, the BigWig command-line tools include wigToBigWig and bigWigToWig that convert to and from the earlier text-based wig format, bigWigMerge to combine signals from multiple BigWigs, bigWigAverageOverBed to extract the average signal over each item in a browser extensible data (BED) file, and bigWigSummary to quickly extract the signal at various scales for a portion of the genome.

In the upcoming years we hope to expand our track data hub support and the compressed file formats that can be queried remotely to include multiple genome alignments and other data types that are currently restricted to local files and databases. We would also like to extend the indexing scheme to support fetching of particular track items by name as well as by chromosome position.

###### **d. Enhancing the security of uploaded data**

Security is a necessary element of almost any software endeavor. Because the bulk of information available on the genome browser website has typically been in the public domain, our security efforts to date have focused primarily on ensuring that the web servers and the data behind them are not compromised by hacking efforts pervasive on the web. We have also made a modest effort to maintain the privacy of custom tracks and other confidential data sets uploaded by users.

In recent years, use of the genome browser has grown within the medical research community. To adequately serve these users, even in a non-clinical application, we must upgrade our security efforts. They are accustomed to a more competitive research environment and want to secure their data from competitors before publication. They are also more likely to upload and view data from human subjects, and they expect these data sets will be secure from unauthorized viewing by others.

We currently do not target the genome browser website toward clinical use, and therefore have not had to meet HIPAA compliance requirements. However, we would like to secure the website from SQL injections, URL mangling, and other common hacking techniques. We have identified several places where the site is potentially vulnerable to such attacks, which we would like to fix. To accomplish this, we would like to employ a security consultant who periodically attempts to break into our site and then fixes any detected vulnerabilities.

We would also like to upgrade our authentication mechanisms. Although we do not require a login to access our website, certain personalized genome browser features, such as the Sessions feature that stores a browser view for sharing and later use, do require a user account and password. Our existing mechanism for handling these logins redirects users to the login page on our genome browser Wiki (genomewiki.cse.ucsc.edu). We would like to replace this awkward and confusing implementation with one that requires fewer clicks and context switches, and that can take advantage of federated login systems, potentially allowing logged-in, authorized users to gain access to private data not available to the general public, such as the dbGAP data at NCBI.

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

During the ten years of software development on the genome browser project, we have produced a large number of command-line tools. These tools have grown in use and popularity as the level of computational expertise within biology labs has increased. In our last ENCODE DCC survey (May 2011), almost 30% of our users identified “bioinformatician” as one of their roles. If the survey results are representative of the entire genome browser community, this translates to over 10,000 potential command-line users. Many of these users already take advantage of our open-source tools, most of which are written in the C language and are therefore typically both faster and more portable across different computing platforms than those written in other languages. Some examples of our most popular tools include liftOver, for moving annotations from one genome assembly to another; isPcr, for determining where primer pairs map in a genome; featureBits, for figuring out the relationship between two sets of annotations in any of a wide variety of formats; and dozens of utilities used to convert files from one genomics file format to another.

In addition to command-line tools, we are beginning to develop web services interfaces to our software to increase access to the genome browser data by other research tools and websites. Web services interfaces have the advantage of working well with programs written in any computer language and can combine data and software into the same package. A web services interface is essentially a URL that refers to a program rather than an HTML file. In this respect it is very similar to the CGI scripts that constitute the bulk of the genome browser tool. For example, a URL for a web services program that fetches the DNA in a particular region of chromosome 1 might be constructed as "http://genome.ucsc.edu/services/fetchDna?chrom=chr1&start=123000000&end=124000000", where the chrom, start, and end are all CGI-encoded input parameters. But, whereas a normal CGI is invoked in a web browser and the output read by a human, a web services application is invoked (using the same internet protocols as a CGI) by a computer program, with the output read by that same program. The output is typically in tabular text, XML, or JSON format, the latter of which is particularly well-suited to JavaScript programs, and is increasingly being adopted by other programming languages as a less verbose, more flexible alternative to XML.

Through this grant we plan to continue our development of new command-line tools and web services interfaces that are well-documented and available for download in both source and executable formats by external bioinformaticians and by our own staff. When appropriate we will publish articles on new tools in peer-reviewed journals.

###### **Approach**

Commercial software products are generally larger than most academically developed software and are expected to function more robustly in a wider range of circumstances. However, the genome browser software’s large user base dictates that we aspire to this higher level of robustness.

Our approach to software development is influenced in part by the software industry background of Dr. Kent and many of the individuals on the genome browser staff. Whenever possible we follow an incremental approach to development, in which software is developed in small pieces that incorporate testability into their design. This ensures that we don’t invest too much bandwidth on a project that turns out to be intractable, and it prevents simple improvements from being delayed while complex improvements are implemented. We maintain a three-week software release cycle. Changes that require more time to implement are implemented on a separate branch in our source code control system (currently git) and then merged into the main branch when ready, accompanied by additional testing. Our software development process is capable of building and sustaining the very large code base (currently over two million lines) of the UCSC Genome Browser and associated tools.

Two of our software engineering practices merit particular discussion: software testing and code review. We expend significant resources on these and feel that the payback is well worth the investment in terms of quality and long-term maintainability and productivity, where development is not hindered by the difficulties of building on an unstable foundation.

We have found, as IBM did 40 years ago, that it is extremely effective to maintain a quality assurance (QA) group separate from our development group and staffed by people who are motivated to find bugs—in contrast to software developers who tend to be less enthusiastic about finding problems in their code. Having a separate QA group allows us to target our hiring at skills specific to software testing and to somewhat reduce our personnel costs, since testers tend to impact the budget less than programmers.

A second practice we rigorously maintain is ubiquitous code review: each line of code that a programmer commits to our source code control system is reviewed by another programmer. This helps catch certain classes of erratic bugs that are not easily found by the QA staff, most of whom do not have significant programming expertise. It also helps diffuse the knowledge of the code throughout the organization, making it more likely that a programmer will reuse a colleague’s piece of code rather than reinvent it, and making it easier for one programmer to extend and maintain another programmer’s code. Code reviews also help ensure that style, commenting, and naming conventions are applied uniformly through the code, making it more readable and understandable particularly to people getting started on the project. If a new software task includes user interface elements, we require the programmer to review the initial designs with at least two other people charged with representing the user’s perspective and maintaining the website’s consistency.

We take care and pride in our software development process, but we also try to make use of existing software when possible. For example, we are currently using the jQuery JavaScript library for developing advanced user interface features on our web applications in a manner that is portable across web browsers. Our recent drag-reordering and drag-zoom features make use of this library, and jQuery modules exist for the drop-down menus and hierarchical lists we plan to add during the next grant cycle. Rather than developing our own interactive tool for viewing gene-gene interaction graphs, we plan to evaluate several existing ones that can work well in the context of a web page, such as VisAnt ([18](#_ENREF_18)), for integration into the browser. We also will ensure that interaction data we host will import easily into tools that are not web-based, such as Cytoscape ([19](#_ENREF_19)).

The breadth of software expertise within our engineering team is generally sufficient to accomplish our project goals, except in the area of security, and in particular for interfaces to security mechanisms that protect data at NIH and other medical- and research-oriented networks. Because this is likely to be a growth area in the coming years, we plan to recruit a new engineer with this expertise who will be shared with and partially funded by the UCSC Cancer Genomics Browser group.

In summary, our approach to developing and maintaining the genome browser software development goals outlined in this aim includes continuing to use the time-tested practices that have served us well during previous grant cycles, leveraging existing high-quality libraries in our own code development whenever possible, and expanding the security expertise within our development group through a targeted hire.

###### **Quarterly milestones**

Table 3. Quarterly milestones for Aim 1.

| **Yr.Qtr** | **Software development components** |
| --- | --- |
| 1.1 | * Update index page with better graphics and pull-down menus * Patch known vulnerability to SQL injection attacks |
| 1.2 | * Add pull-down menus on genome browser application * Finish support for Variant Call Format (VCF) in track hubs * Security expert attempts to break into site, fixes problems found |
| 1.3 | * Add pull-down menus on other applications * Update documentation on track hubs * Package command-line tools with updated documentation |
| 1.4 | * Add motif hits and sequence logos to transcription factor binding tracks * Switch to cryptographic hash for user IDs to prevent URL-mangling security leaks * Security expert attempts to break into site, fixes problems found |
| 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 |
| 2.3 | * Finish integration of network visualization tool * Package command-line tools with updated documentation |
| 2.4 | * Dynamic conversion of genome browser tracks into Gene Sorter columns * Security expert attempts to break into site, fixes problems found |
| 3.1 | * Improve phased haplotype displays |
| 3.2 | * Initiate Circos-like display on test site * Security expert attempts to break into site, fixes problems found |
| 3.3 | * Release Circos-like display on main site * Package command-line tools with updated documentation |
| 3.4 | * Initiate condensed "exon only" display * Security expert attempts to break into site, fixes problems found |
| 4.1 | * Switch from Wiki-based to more robust authentication system |
| 4.2 | * Release condensed "exon-only" display |
| 4.3 –5.4 | * TBD |

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

As of Sept. 2011, our website provides active genome browsers for 94 genome assemblies of 53 distinct species. These browsers typically include tracks based on GenBank mRNA and EST data, mappings of the UCSC Genes set (or where appropriate, a curated gene set from a model organism database), and multiple genomic alignments of selected related species. The genome browser primarily focuses on vertebrates and other Metazoa, with some exceptions. Browsers for organisms that have an active research community generally showcase a number of tracks contributed by that community. The human genome annotation set is the richest and includes a track featuring a 46-vertebrate multiple alignment that we plan to expand to 50 vertebrates during the current grant cycle.

###### **Significance**

By building a basic genome browser for a species, we support basic molecular biological research on that organism. We enable targeted access to the genomic sequence with Blat and in silico PCR search capabilities. We provide several browser tracks, such as TransMap and Human Proteins, that are useful for determining which parts of the genome are likely to be genes. The GenBank mRNA alignments also help identify areas under investigation by other labs, often providing an entry point into the scientific literature and preventing redundant efforts. Through the diverse range of species it offers, the genome browser helps connect research communities focused on different organisms. In particular, it alerts researchers working primarily on the human genome to related work on model organisms. The broad range of organisms covered by the genome browser and its database facilitates the research of scientists working with multiple organisms, who can access their data within a consistent environment rather than learning a different database interface for each organism.

Alignments of multiple species serve several important purposes. They allow users to identify regions in other species that correspond to a particular region in the human genome or other well-studied organism. They also serve as a basis for studying the process of evolution itself at a molecular level ([20](#_ENREF_20)). The patterns of evolution at a particular base or region can shed light on the function of the region as well, with protein-coding regions, RNAs with secondary structure, and to a certain extent even regulatory regions having their own evolutionary signature ([21](#_ENREF_21)).

###### **Innovation**

In the next grant cycle we intend to expand the vertebrate multiple alignment on our human genome reference assembly to at least 100 species, and considerably beyond that, if feasible. We also plan to expand the number of species for which we build genome browsers to at least 100. We intend to restrict this browser collection to a selected set of species. The number of available assemblies is increasing so rapidly that it is cost-prohibitive for us to build genome browser databases on an unlimited set of genomes using our current approaches. The curational aspects, including documentation and quality assurance, scale linearly with the number of species. Our Scientific Advisory Board (SAB) and the participants in our user surveys have conveyed relatively little interest in genome browsers on most species other than the human and the model organisms. Therefore, it seems a better investment to enrich the annotations on the genomes we currently support, notably the human, rather than scaling our genome browser build process to 1000 or more species. We do plan to expand our browser collection to include additional primates and other high-quality genome assemblies that have an active research community or belong to a clade that is underrepresented in the current vertebrate collection.

Although we cannot curate every genome, we intend to develop a new type of data hub, an Assembly Data Hub, which will enable the genomics community to easily extend the genome browser to display genome assemblies we are unable to integrate into our own database. The assembly data hub will be similar in concept to the track data hub, which is described in detail in Aim 3. The data provider will store the genome sequence in a compressed, binary, indexed file format and make it available on a remote web server along with a list of track hubs that annotate that genome. The assembly and its annotations will then be available for view in the genome browser. Because so much of the hub infrastructure has already been implemented during the development of track data hubs, we estimate that the assembly hub concept can be developed within a year by slightly less than 2 FTEs. This new technology will greatly enhance our ability to serve members of the genomics community working on species we are currently unable to accommodate.

###### **Approach**

Our current multiple-species alignment pipeline was developed more than five years ago from a synthesis of base-level alignment tools developed at Penn State University ([22](#_ENREF_22)) and genomic region chaining tools developed at UCSC ([20](#_ENREF_20)). Since that time we have made incremental improvements to the pipeline, such as the substitution of the lastz program for the blastz pairwise base aligner, which has increased the throughput of the pipeline to the extent that we are confident it can be scaled up to 100 species on modern hardware. It will be difficult to extend the current pipeline past that level; however, separately funded research efforts, notably within the Ensembl group ([23](#_ENREF_23)) and the Haussler lab at UCSC ([24](#_ENREF_24)), plan to scale multiple-species alignments of 10,000 genomes. We plan to monitor the progress of the multiple-alignment research community and adopt appropriate new tools that can reliably scale to large numbers of genomes.

###### **Quarterly milestones**

Table 4. Quarterly milestones for Aim 2.

| **Yr.Qtr** | **Comparative genomics components** |
| --- | --- |
| 1.1-1.3 | * Add genome browsers for three new species or updated genomes |
| 1.4 | * Add new multiple-alignment track for one set of assemblies * Initiate development of assembly hubs |
| 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 * Update assembly hub specifications |
| 3.1-3.3 | * Add genome browsers for three new species or updated genomes |
| 3.4 | * Add new multiple-alignment track for one set of assemblies |
| 4.1-4.3 | * Add genome browsers for three new species or updated genomes |
| 4.4 | * Add new multiple-alignment track for one set of assemblies |
| 5.1-5.3 | * Add genome browsers for three new species or updated genomes |
| 5.4 | * Add new multiple-alignment track for one set of assemblies |

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

Many of the annotations in the UCSC database are imported from external scientific groups. These include gene, allele, and expression information from WormBase, FlyBase, SGD, ZFIN, MGD, RGD, and Online Mendelian Inheritance in Man (OMIM); data from many different NCBI databases; the entire UniProt database; and data associated with dozens of individual scientific publications. In some cases, such as in our role as the Data Coordinating Center (DCC) for the ENCODE project, we receive separate funding to import, curate, and display specific data sets. However, in general, we consider the integration of data from a large variety of complementary sources to be a core part of the work that is funded on this grant. This includes mining high-quality data sets from the literature, scientific consortia, and other databases; communicating with the data producers; reformatting the data; integrating it into our database and website; and performing quality checks. It also includes generating documentation to facilitate understanding of the data by the biomedical community at large.

###### **Significance**

When trying to interpret a particular region of the genome, it is extremely useful to have access to data from many sources aligned to the same genomic coordinates in the same graphic display. Likewise, having access to diverse data sets that have been carefully checked for quality, documented, and converted into common formats makes work much easier for scientific analysts working at the genome level. The power of the genome browser to display each data set quickly at any scale and to show how that data relates to other data sets, particularly to available gene sets, greatly facilitates data analysis.

###### **Innovation**

The scientists whose work we import generate most of the innovation in this area. The innovation from UCSC is mostly in the area of developing new displays and support for larger data sets, both described in Aim 1. Additionally, we have introduced some innovation in our approach to the data import process.

Historically most of our data import work has been done by our software engineers, whose knowledge of the genome browser database and code enables them to determine whether a new data set can use existing browser infrastructure or instead requires the development of new table types in the database and new display types in the browser. With the growing maturity of our site, the majority of incoming data can now be represented without new engineering, although a substantial amount of integration effort still remains. Through our experience with the ENCODE DCC, we have determined that it is more efficient and practical for a separate data integration staff, informally known as “data wranglers”, to handle this work. Although data wranglers need database skills comparable to those of our software engineers, they do not need strong C programming skills. On the other hand, they generally require stronger organizational and communication skills than software engineers. Based on our successful use of data wranglers on the ENCODE DCC project, we plan to extend this model to our genome browser organization, adding more data wrangler expertise to our staff to handle our data import and integration needs in the future.

###### **Approach**

The genome browser data import process varies relative to the level of general scientific interest in the data and the size of the data set. Data sets are targeted for import based on the recommendations of our users, advisors, staff members, and the data producers themselves. We review publications associated with the data set and talk with the data producers to understand the data. Then we decide whether the data should become an integrated part of our website or should be instead be remotely hosted as a data hub or custom track.

If we decide to import a data set, we fully integrate it into the genome browser databases hosted on our local development server, reformat it if necessary to fit into genome browser conventions, and create documentation pages that describe the data at the comprehension level of a first-year graduate student in a biomedical field. We ensure that the displayed data matches the documentation and other communications with the contributor, and that it overlaps as expected with other genome browser annotation tracks (e.g., overlap of RNA and gene models with our gene tracks and overlap of regulatory regions with DNAse hypersensitivity sites). We perform automated checks to ensure that both strands are covered where appropriate, that there are no large gaps in coverage, that the coordinates are not off by one, and that the data contains none of the processing errors we have encountered in the past. Our quality-checking process has occasionally uncovered critical problems in data provided by a contributor or sequencing center that we were able to report back to them for correction at the source.

In some cases we extensively filter imported data sets to make them more useful to our browser community. Our Common SNPs track is a good example of this. This track is derived from dbSNP, but it shows only variants where the population frequency information is believable and at least a 1% of the overall population. While dbSNP provides a field for frequency information, the data have some misleading quirks, such as an SNP observed in only a single individual who is heterozygous being labeled as having a 50% incidence. Fortunately, in many cases we can infer the total number of genomes examined for the SNP and filter for only those in which the total is large enough for the population numbers to be believable.

We determine the display visibility of a data set based on its usefulness to the general research community. If a track is of high interest to a broad range of users, we set it to automatically display in one of our visible display modes in the default browser view. Tracks of slightly less interest to the overall community are hidden by default, which requires the user to discover it through our track search mechanism or by casual browsing.

If we decide that a track is of interest to only a small segment of our user community, we ask the data producer to create a custom track or track data hub hosted on their own server, and we link to it from our custom tracks or make it available on our data hub page. We provide the contributor instructions for linking to our site to view and share the data within the context of the genome browser. We invest only minimal effort in checking the quality of these remotely hosted data sets, primarily ensuring that the URLs are active and that the data is in the correct format. For data of even narrower interest (for instance, of interest to only a few individuals or small labs), we simply instruct the contributor on how to format the data as a custom track or data hub, and then provide the appropriate self-loading links to our website.

At the request of the OMIM group, we recently started hosting their website ([omim.org](http://omim.org/)) on our servers in addition to producing tracks based on their annotations. This was a relatively easy transition for us, since we already had the web server infrastructure in place. Because OMIM is such an important biomedical resource with relatively modest computational infrastructure needs, we agreed to do this under the auspices of our genome browser grant. We estimate that it will cost less than 5% of an FTE on an ongoing basis. We have also worked closely with the OMIM staff to produce a set of three tracks that enhances their value to our users.

We are considering the use of track data hubs for the display of data from large consortia rather than importing the full data sets ourselves, even in cases where we consider the data to be of wide interest. We recently experimented with this approach for the Roadmap Epigenomics Project data, with favorable results. The data sets are hosted on a remote server at Washington University in St. Louis, and they are accessible to all genome browser users through the Hub Import page. Given the large size of data sets produced by present-day consortia, our efforts to fully integrate the data as built-in browser tracks can be time- and cost-prohibitive, and even the disk and bandwidth requirements can be substantial. The data hub technology provides a potential solution for making large data collections accessible with minimal overhead in a timely manner. Although data sets added as track hubs have less visibility, we believe we can compensate for this through well-placed announcements and links to the new consortium data hubs, which will also receive visibility from the consortium web pages and publications.

As an alternative to custom tracks and data hubs, we offer support for users to create their own databases and mirror our software on their own servers for private use. More than 200 different groups currently take this approach. We provide complete instructions for setting up a mirror site, and we publish a mailing list specific to mirror site technical issues. Once a mirror is set up, it is relatively straightforward to add data of particular interest to the local installation. Before track data hubs, a mirror site was the only alternative available to users who wished to visualize data sets that exceeded the capabilities of custom tracks. Because data hubs are much easier to set up than a mirror (1-2 days of work rather than 2-4 weeks), we anticipate sites will prefer hubs to modified mirrors in the future. However, for track types not yet supported in hubs, and for those who need their data to remain strictly behind firewalls, a modified mirror is still the only option, and we therefore plan to continue our support for these users.

In summary, we plan to continue maintaining our current pipeline for importing high quality data sets from the scientific community into the genome browser, while at the same time improving our efficiency through the use of data integration specialists on our staff and increasing our scalability and accessibility through the use of track data hubs.

###### **Quarterly milestones**

Table 5. Quarterly milestones for Aim 3.

| **Yr.Qtr** | **Data import components** |
| --- | --- |
| 1.1 | * Update Recombination Rate track for hg19 * Add ORFeome tracks for mouse and zebrafish * Prioritize data suggested by users, funding agencies, consortia, and our scientific advisors |
| 1.2 | * Add Segmental Dups tracks for several assemblies * Add selected personal genomes to Genome Variants track * Prioritize data suggested by users, funding agencies, consortia, and our scientific advisors |
| 1.3 | * Include alternate assembly data in mouse browser * Prioritize data suggested by users, funding agencies, consortia, and our scientific advisors |
| 1.4 | * Reconcile LiftOver differences between UCSC's and Ensembl's GRCh37 * Add1000 Genomes Coverage Masks * Prioritize data suggested by users, funding agencies, consortia, and our scientific advisors |
| 2.1-5.4 | * TBD |

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

UCSC has been building gene models based on alignments of mRNA to the genome for the past ten years. We produce an internally generated UCSC Genes set for human and mouse. We import gene sets from the model organism databases for the rat, *C. elegans*, *D. melanogaster*, *S. cerevisiae,* and zebrafish genomes. The mRNA collection is not currently rich enough for us to produce a quality gene set on other organisms using our process; however, where available we do import the Ensembl gene set ([25](#_ENREF_25)) and provide de novo gene predictions, a Human Proteins track, and a TransMap track, which maps vertebrate mRNAs to the other assemblies.

Our initial gene models were based simply on Blat alignments of RefSeq mRNA to the reference genome. Later, recognizing a need for a well-defined connection between RefSeq ([26](#_ENREF_26)) and the most popular protein-oriented database, SwissProt/trEMBL (which has since been absorbed into UniProt ([27](#_ENREF_27))), we integrated the latter into an improved gene set, UCSC Genes. We further modified the design of the UCSC Genes model in response to feedback from our SAB and users, who stressed the importance of including non-coding genes because of their relationship to significant peaks in linkage and genetic association studies. Our current pipeline includes significantly more non-coding and alternatively spliced transcripts than RefSeq, but still requires at least two independent lines of evidence for non-RefSeq transcripts, one in addition to the GenBank mRNA itself. This gene set encompasses the so-called linc-RNAs and the subset of small RNAs that have good sequence available in GenBank. We are currently adding additional small RNAs that have good Rfam ([28](#_ENREF_28)) hits to our gene set.

In addition to providing transcribed coding and non-coding genes in the UCSC Genes track and to importing high quality gene sets (as described in Aim 3), we also plan to add associated regulatory regions and allelic differences to the genes set as part of this aim.

###### **Significance**

Genes are both the major functional units and the major landmarks within the huge expanse of DNA that constitutes the genome. From the DNA sequence alone it is not generally possible to determine the locations of many of the major classes of genes. The large introns typical of many vertebrate genes and the presence of many interleaved alternative transcripts from the same region of DNA add to the difficulty of locating genes.

Currently the UCSC gene set contains an intermediate level of alternative transcripts—significantly more than RefSeq, but significantly less than the Gencode and Acembly gene sets. This balance was driven by a pragmatic desire to avoid overwhelming users with a large number of transcripts, not all of which are likely to have biological significance, while still alerting users to the fact that alternative transcripts are a regular feature of vertebrate genomes that cannot be ignored. The National Cancer Institute standardized on our gene set for The Cancer Genome Atlas (TCGA) project largely because the level of alternative splicing seemed appropriate. However, as the databases grow richer, more transcripts are passing our criterion of two independent lines of evidence, demanding us to develop a more nuanced transcript-ranking system to prevent our display from becoming too cluttered.

###### **Innovation**

Scientific consensus on the human gene set is growing, resulting in more consistency among the gene sets produced by various expert groups than has existed in the past, but substantial variation among gene sets still does exist. Different groups tend to excel in specific areas: Ensembl in protein modeling, Gencode and RefSeq in expert curation, and UCSC in automated RNA-based pipelines. Although there are benefits in continuing the development of the UCSC Genes set, we are aware of the problems caused by the proliferation of too many gene sets. To address this, we ensure that our gene set is a proper superset of the Consensus Coding Sequence (CCDS) gene set, which attempts to standardize protein-coding genes, and of RefSeq, which is the most popular conservative gene set among our users. We will consider standardizing on the Gencode gene set when it is completed, but it is unclear when the first high-confidence full-genome set will be released. The Gencode set tends to include by default more variant transcripts than we want, which can overwhelm the display and obscure the main transcript. We currently plan to complete our new transcript-ranking system for the UCSC Genes set, then apply it to the Gencode set, which should be relatively stable at that point. We will then decide whether to adopt the Gencode genes as our main gene set and whether the UCSC Genes set still adds significant value.

In the coming years we would like to incorporate information from cap-analysis gene expression (CAGE) tags ([29](#_ENREF_29)), short-read RNA sequencing (RNA-seq), polymorphism studies, and data on regulatory regions into our gene set. The CAGE tags will primarily refine our transcription start sites. We will use RNA-seq to identify the relative abundance of various alternative transcripts and rank them. We will then allow the user to choose how many transcripts to display.

Genome browser users currently have the option to limit the display to what we term the “canonical” transcript, which is the transcript with the longest coding region or (for non-coding genes) the longest transcript. However, this is an unsatisfactory solution, since some of the sequencing libraries behind the GenBank data are highly normalized, some of the transcripts we see are exceedingly rare, and it is therefore possible that the longest transcript will be non-representative. A better solution would be to calculate abundances from deep RNA-seq data on a large number of cell types and then identify the one most abundant in the widest number of cell types as the canonical transcript. We would also like to offer users the ability to rank transcripts both globally and within the context of a particular cell type and to display only the top-ranked transcripts for any gene up to a particular user-selected ranking.

There is more to a gene than its transcripts. In particular, the regulatory regions associated with a gene and the alternative alleles of a gene circulating in the population are both important pieces of information. To some degree it is possible to see regulatory regions and alternative allele information by consulting the DNAse, ChIP-seq, SNP, and copy number polymorphism tracks within the genome browser, but it is often more productive to work at the gene level rather than in terms of diverse tracks in chromosome coordinates. While the regulatory regions and alternative alleles associated with a gene are less defined than the transcripts, tremendous progress has been made in this area in the past five years, and it now appears possible to include these in the gene set without the signal being dominated by the noise. We plan to integrate transcriptional regulation information from the ENCODE and Roadmap Epigenomics Projects and allelic diversity information from the 1000 Genomes Project into the gene sets we present on the genome browser, leveraging the analytical and raw data-generating aspects of those projects when possible.

###### **Approach**

The work described in this aim will be built on top of our existing gene set pipeline (Figure 5). In brief, we align mRNAs from GenBank with Blat ([2](#_ENREF_2)), and filter the resulting alignments to keep only the best for each transcript. The alignments from multiple mRNAs are merged into splicing graphs, the splicing graphs from human and mouse are compared ([30](#_ENREF_30)), and the parts of the graph observed in both species are given additional weight. Additional lines of evidence are added to the graph, including EST evidence from multiple independent clones and conservation signatures consistent with a coding exon. The graph is then turned into a collection of non-redundant transcripts. To avoid a combinatorial explosion, we include only intron–exon combinations seen in one of the mRNAs used to construct the graph. To reduce noise, we include only transcripts supported by at least two lines of evidence, with the exception that RefSeq ([26](#_ENREF_26)) and CCDS ([31](#_ENREF_31)) transcripts are always included. For transcripts that do not have a protein determined by CCDS or RefSeq, we predict the protein-coding region using a process with five inputs: the length of the open reading frame (ORF) in the species, the length of the ORF in related species, the presence or absence of the Kozak sequence, the presence or absence of upstream ORFs, and whether the resulting transcript would become a target for nonsense mediated decay; for RefSeq and CCDS transcripts, we use the existing annotation instead.

The transcription graph framework is very flexible. Evidence from many sources—such as information from CAGE tags obtained from the ENCODE Project and elsewhere—can be added to the graph if it can be expressed in genomic coordinates. Currently we calculate the start of a transcript by sorting the start sites of all mRNAs compatible in intron–exon structure with that transcript. Since mRNA sequences are often fragmentary, we take the start site that is at the boundary between the third and fourth quartiles rather than taking the middle start site (i.e., the median), and are thus somewhat biased toward the longer end. We omit the largest, since this can often be an outlier. By including CAGE tags, we have a much larger collection of start sites, which will improve the robustness of our current process that in many cases has just one or a few mRNAs to work with.

Description: txGraph1.tiffDescription: txGraph2.tiff

Figure 5. Construction of a transcription graph, and the resulting non-redundant and defragmented transcripts. This graph represents the heart of the UCSC Genes build. Not shown is the stage where additional evidence is added to the graph. Only non-redundant transcripts with at least two lines of evidence to support them are kept.

Adding RNA-seq data will be somewhat more difficult than CAGE data, largely due to the challenge of mapping short reads that span exons. Although CuffLinks ([32](#_ENREF_32)) and Scripture ([33](#_ENREF_33)) both attempt this, they produce substantially different results ([34](#_ENREF_34)). We plan to augment these with programs that map reads to existing transcripts, including a recent program from the CuffLinks group ([35](#_ENREF_35)).

To determine the variants of a gene in the population, we initially plan to simply map data from existing SNP and copy number variation tracks to the exons in our gene models. In this phase we won’t consider the haplotype structure (i.e., which SNPs tend to migrate together), but will instead continue to integrate this information as the 1000 Genomes Project and other efforts define sets of common haplotypes. We will display this information in a new section on our UCSC Genes details page, and we will highlight changes that affect the protein produced by the transcript.

We plan to rely on the ENCODE Project consortium to produce a list of regulatory regions. As a member of this consortium, we are aware of work in progress (to be published by late 2011) that integrates information from DNAse hypersensitivity, chromatin immunoprecipitation of modified histones, chromatin immunoprecipitation of transcription factors, and assays of DNA methylation. Similar work is ongoing within the Roadmap Epigenomics consortium and other groups. In general we plan to take a relatively conservative subset of these regulatory annotations and add them to our UCSC Genes details page along with evidence for associating a region with the gene. In some cases we will have good evidence from chromatin capture conformation assays, but in the majority of cases, at least initially, the only evidence will be that the promoter of the transcript is the closest promoter to the regulatory region.

In summary we plan to use a combination of existing pipelines, work from scientific consortia, and work described in the literature to produce a gene set that is increasingly refined and well annotated.

###### **Quarterly milestones**

Table 6. Quarterly milestones for Aim 4.

| **Yr.Qtr** | **Gene set components** |
| --- | --- |
| 1.1 | * Update UCSC Genes for human genome |
| 1.2 | * Update UCSC Genes for mouse genome |
| 1.3 | * Update UCSC Genes for human genome |
| 1.4 | * Update UCSC Genes for mouse genome * Integrate CAGE data for more accurate promoter calling |
| 2.1 | * Update UCSC Genes for human genome |
| 2.2 | * Update UCSC Genes for mouse genome |
| 2.3 | * Update UCSC Genes for human genome |
| 2.4 | * Update UCSC Genes for mouse genome * Integrate RNA-seq data to determine isoform abundance |
| 3.1 | * Update UCSC Genes for human genome |
| 3.2 | * Update UCSC Genes for mouse genome |
| 3.3 | * Update UCSC Genes for human genome |
| 3.4 | * Update UCSC Genes for mouse genome * Evaluate Gencode to decide if it will become our main gene set on human |
| 4.4 | * Evaluate Gencode as replacement for primary gene set on mouse and human |
| 5.4 | * Evaluate Gencode as replacement for primary gene set on mouse and human |

###### **Access and Dissemination**

Our resource is accessed primarily through the Internet. Our web site at http://genome.ucsc.edu provides graphical tools for visualizing and manipulating the genomic data on our site along with supporting documentation and access to downloadable data. Our public MySQL server (http://genome-mysql.cse.ucsc.edu) and a variety of Internet file transfer protocols provide access to the data in bulk or computer-readable form. Many external mirror sites have set up scripts that sync with our ftp site on a regular basis.

The presence of a web application or a large set of files on the Internet is not by itself sufficient for people to find the resource. However, awareness of our resource is widespread throughout the scientific community, evidenced by its presence in the literature, links from collaborating websites, and frequent mention in presentations at scientific meetings. We ensure that the genome browser website is accessible to web search robots in a regulated manner that avoids performance degradation. Currently our website is the first item returned by Google for the search term “genome browser,” and it is in the first page of results returned by a search on “human genome."

Despite widespread awareness of the genome browser within the genomics community, it is important that we simultaneously reach new users while encouraging our long-time users to explore new data sets and take advantage of new features. We support all our users in learning more about our website in three major ways: documentation, user support, and user training.

The user support and training functions provide us with an excellent opportunity to collect feedback from our users. We also seek user feedback through regular surveys posted on the website and through lab visits where we observe people using our research tools during the course of their normal work.

###### **Website and data access infrastructure**

The genome browser public website is backed by 8 web servers that have access to a central file server and a central MySql database server. Blat web access is provided by 15 additional servers. We also provide a genome-preview server that allows the public to access our raw data before it has been quality-checked, a public MySql server that hosts the browser’s MySql data, a custom track server to store user-generated custom tracks, and a Wiki server that holds public information and tracks named sessions. Finally, a local download server that allows users to download our data serves nearly 2 TB of data every day. All machines serving the genome browser data are housed in a data center that is designed to function 24/7, 365 days a year. We also house one identical download server and one additional fileserver at the University of California, San Diego Supercomputer Center for load-balancing and redundancy. During the upcoming years, we plan to provide servers in different geographical areas to reduce the access penalty; the first European system (in Germany) will be deployed by 2012.

For detailed specifications of the hardware and networking infrastructure that supports the genome browser website and development environment, see the Resources section.

###### **Technical support**

We provide technical assistance to our user community through three mailing lists. Our primary interactive list, [genome@soe.ucs.edu](mailto:genome@soe.ucs.edu), currently has over 700 subscribers and provides a discussion forum for general topics about the genome browser and data. Over the years, the questions directed to this list have become increasingly complex, reflecting the growing sophistication of our user base. A second interactive list with nearly 200 subscribers, [genome-mirror@soe.ucsc.edu](mailto:genome-mirror@soe.ucsc.edu), allows our mirror sites to discuss mirroring issues. We also provide a low-volume non-interactive list (approximately 1650 subscribers) for users interested only in project announcements ([genome-announce@soe.ucsc.edu](mailto:genome-announce@soe.ucsc.edu)) and a mail alias ([genome-www@soe.ucsc.edu](mailto:genome-www@soe.ucsc.edu)) for reporting technical problems or making confidential inquiries.

During this grant period, we have answered more than 10,000 questions through our mailing list interface, and nearly 2000 more through direct email. The mailing list volume currently requires the full-time attention of at least one staff person each day, seven days per week. We recently tied in our mailing lists with our project-tracking tool (Redmine), which has improved our efficiency in administering the lists and allowed us to archive our messages in a way that provides better searching and tracking capabilities than the Mailman archives. Most questions on the lists are answered within one or two business days of posting, and many within a few hours. We have received recognition from our users for the quality and promptness of our responses.

###### **User feedback**

We frequently receive requests for new software features and data through our mailing list. Reasonable requests are logged in our Redmine issue-tracking system for prioritization and implementation. During the current grant cycle we conducted two user surveys, publicized through the genome browser website and mailing list. In addition to providing us with useful demographics about our user community, the surveys yielded several good suggestions for website improvements, some of which were subsequently implemented. These included adding genome browsers and BLAT support for more vertebrate genomes, adding additional species to the conservation track, adding more gene expression, protein–protein interaction, and ChIP/CHIP data, and improving the overall speed of the site.

##### **Training**

###### **Seminars and workshops**

Hand-in-hand with the outreach and dissemination efforts described in the previous section, it is more important than ever that we invest in training to help our users effectively apply UCSC’s tools and data to their work. As the volume and complexity of genomic data continues to increase, we regularly introduce new data types, visualization features, and tools to help users browse and analyze the data. It is necessary to hide the full array of available functionality from the casual user so that potential new users are not turned away by the complexity. As a result, even seasoned browser users have told us at workshops, “I’ve been using your browser for years and didn’t know it could do that!” The usage of certain browser features continues to climb as people discover them. For example, use of the browser’s Session feature increased dramatically since its introduction in early 2007 to 100 users who were saving and sharing a new session per month in early 2010 and to nearly 500 active users by mid-2011. This shows both that the feature is found to be very useful and that we could do a better job of getting the word out early. We are not sure whether the sudden increase in usage is due to its prominent mention in our workshops, but the coincidence is suggestive.

For many years, we have contracted with OpenHelix to conduct four annual workshops and feature the genome browser in their booth at two trade shows. In the past few years, we have expanded our outreach efforts to include workshops presented by UCSC staff at major annual meetings of scientific societies (Table 7). Workshops at major meetings are an excellent way to reach people from a large number of institutions and disseminate information widely. There is a great unmet demand for bioinformatics expertise. The genome browser workshops are in high demand, and are invariably well attended. For example, the UCSC Genome Browser workshops at American Society of Human Genetics meetings sell out within the first two days of registration.

Table 7. UCSC Genome Browser workshops presented by UCSC and our training partner OpenHelix, 2008 to present.

| **Year** | **# Workshops** | | **# Locations** | | **Est. attendance** | |
| --- | --- | --- | --- | --- | --- | --- |
|  | UCSC | OpenHelix | UCSC | OpenHelix | UCSC | OpenHelix |
| 2008-2009 | 5 | 16 | 3 | 8 | 1500 | 320 |
| 2009-2010 | 2 | 16 | 1 | 8 | 440 | 312 |
| 2010-2011 | 32 | 8 | 22 | 4 | 2070 | 375 |
| Total | 39 | 40 | 26 | 20 | 4010 | 1007 |

In August 2009, we offered cost-free workshops to institutions, intending to schedule a small number of workshops as a pilot project. We instead collected more than 200 requests in a two-week period. We subsequently delivered workshops to several institutions on this list, with 50-80 attendees at each, and transferred several other requests to OpenHelix. More recently, we have introduced a new training model (which we call “host-pays”) in which we schedule workshops at geographical locations near scientific meetings we are attending, and then ask the workshop hosts to fund local transportation costs and lodging. While this model still costs us the trainer’s salary, it extends our reach at modest expense and brings informatics training to many more potential users. For example, in the week before the 2011 European Society of Human Genetics (ESHG) meeting in Amsterdam we gave two workshops each in Nijmegen and Rotterdam, the Netherlands, and Leuven, Belgium. Typical attendance included scientists from several departments and individuals from nearby universities and hospitals. Thus, for the cost of flight and lodging for the ESHG meeting, we trained approximately 400 people in four cities.

In addition to seminars and conferences, both OpenHelix and UCSC Genome Bioinformatics group personnel have contributed to genome browser outreach through speaking engagements, attendance at professional meetings, and student outreach activities.

We plan to increase our outreach efforts in the next few years, transitioning most of the training from OpenHelix personnel to our own staff and using the host-pays model. The funds saved by eliminating the OpenHelix workshops will be used to cover approximately 25% of a genome browser FTE to focus on training and outreach. OpenHelix will continue to provide genome browser training materials, described below.

##### ***Genome browser documentation***

We provide an extensive set of online documentation for genome browser users. Each of the annotation tracks displayed in the genome browser has an associated description page that describes the annotation, the methods used to produce it, acknowledgements of the annotators and data sources, and related published references. We provide a general User’s Guide (<http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>) covering the primary genome browser tool, as well as guides specific to the use of the Table Browser, Gene Sorter, and Proteome Browser. Our FAQ (<http://genome.ucsc.edu/FAQ/>) covers eleven main subject areas, highlighting common questions asked by our users. We provide an automatically updated release log of new genome assemblies, annotations, and data downloads, and a large set of information pages on data downloads, database structure and contents, Browser-related utilities, licensing, mirroring, custom tracks, data and collaboration acknowledgements, training materials, publications, project staff, and website contacts. We also offer links to free online tutorials and training materials for the genome browser developed through a contract with our genomics outreach partner, OpenHelix (<http://www.openhelix.com>), and available at <http://openhelix.com/ucscmaterials.shtml>. Supplementing the main genome browser documentation pages, we provide a set of online information pages specific to the ENCODE browser (<http://genome.ucsc.edu/ENCODE/>) and a gallery of sample genome browser sessions showcasing interesting ENCODE data tracks.

To assist individuals who download our source code or mirror our site, we have an extensive technical documentation set in our source tree that describes how to download, compile, and debug our source code, set up a genome browser mirror, create and load a new genome browser track or database, and so on. We also provide information on many topics of interest on our public Wiki (http://genomewiki.cse.ucsc.edu/).

##### **Administration and management**

##### ***Organizational structure and staff responsibilities***

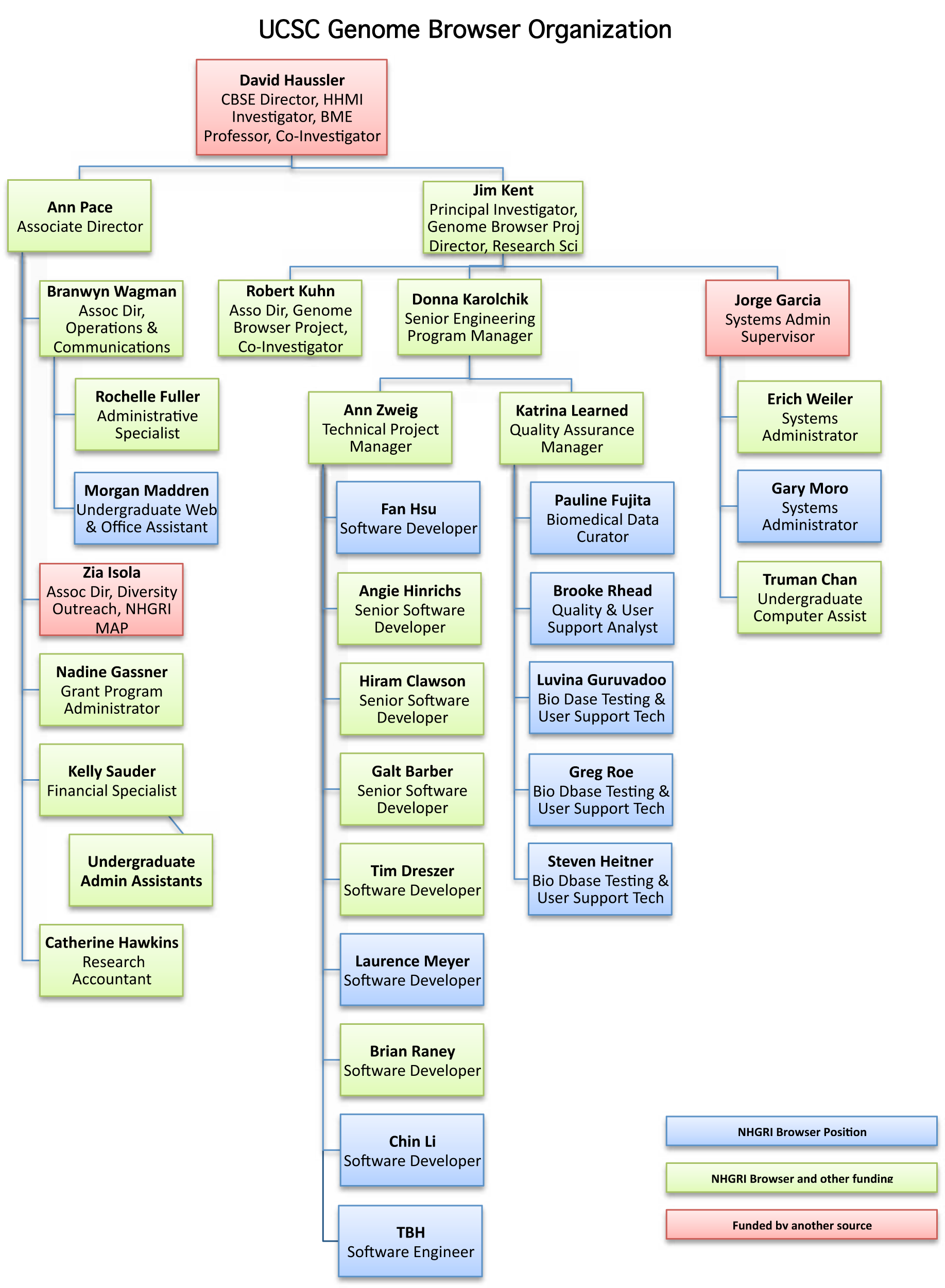
Reflecting a trend in our organizational structure over the past few years, the role of principal investigator will transition from David Haussler, who will remain as co-investigator providing high-level guidance, to Jim Kent, who has directed the UCSC Genome Browser since its inception. An additional co-investigator, Robert Kuhn, has assumed increasing responsibility for scientific direction of the project in recent years. These changes should not be apparent to UCSC Genome Browser users.

Figure 6 shows the organizational chart.

***Management***

*W. James (Jim) Kent*, Principal Investigator. Sets overall direction and priorities for the project with advice from co-PIs and Scientific Advisory Board. Trains software engineers and managers. Responsible for overall software architecture. Estimates amount of work required for new features. Writes grant proposals and scientific papers. Represents project to funding agency and outside scientific community.

*David Haussler*, Co-Investigator. Advises PI on overall direction and priorities of project, particularly on aspects involving comparative genomics, interfacing with scientific consortia, interfacing with the University of California as a whole, fundraising, and the use of externally developed genome analysis tools. Represents project to outside scientific community.



**Figure 6.** UCSC Genome Browser organization chart. Positions in blue are fully funded by this award; positions in green are partially funded by this award; positions in red are funded elsewhere.

*Robert Kuhn*, Co-Investigator. Advises PI on overall direction and priorities of project. Responsible for programs to train genome browser users and reach new users. Interfaces with OMIM, Ensembl, NCBI, and other databases that have substantial human genetic and genomic data. Represents project to outside scientific community.

*Donna Karolchik,* Senior Engineering Program Manager and Senior Writer. Manages and trains other managers of technical staff, including software project managers for this and other projects and the manager of quality assurance. Writes and edits documentation for the website, tutorials, grants, progress reports, and scientific papers.

*Ann Zweig,* Technical Project Manager. Organizes overall work flow into discrete tasks and assigns tasks to individual staff. Tracks progress on tasks. Organizes staff meetings and promotes communication between the staff. Organizes code reviews. Performs UCSC supervisory management duties for software engineers and data wranglers, including conducting performance reviews and handling personnel issues.

***Software development***

*Fan Hsu*, Senior Software Developer and Data Integrator. Develops protein-oriented aspects of gene model pipeline. Integrates third-party data into our database.

*Angie Hinrichs*, Senior Software Developer and Data Integrator. Imports human variation data from dbSNP, copy number variation databases, and 1000 Genomes project. Develops new displays for haplotype and other new variation data. Integrates our code (primarily in C) with the Perl code for that programming community. Interfaces with FlyBase.

*Hiram Clawson*, Senior Software Developer and Data Integrator. Generates genome browsers for non-human organisms. Supports and trains users of mirror sites and cloud. Runs the multiple alignments and conservation scoring pipelines. Gathers usage statistics.

*Galt Barber*, Senior Software Developer. Responsible for network-oriented programming and programming involving parallel execution, such as the code that fetches data from hubs into the cache at UCSC. Develops data integration tools. Administers and optimizes the MySql database.

*Tim Dreszer*, Software Developer and Data Integrator. Responsible for JavaScript development of advanced user interface features, metadata representation, and file search. Integrates third-party data, particularly ChIP-seq, into the database.

*Larry Meyer*, Software Developer. Responsible for JavaScript development of advanced user interface features, full text track search, speed optimizations of the website, and data importing on regulatory regions. Develops web services interfaces.

*Brian Raney*, Software Developer and Data Integrator. Develops new comparative genomics displays, Circos development, user interface to data hubs. Integrates third party data, particularly RNA-seq, comparative genomics, and chromatin conformation capture into the database.

*Chin Li*, Software Developer and Data Integrator. Creates genome browser databases for non-human organisms. Integrates third party data into the database.

*TBH*, Software Engineer. Responsible for data security, particularly of uploaded data. Develops improved authentication interface. Works with other biomedical databases on federated authentication mechanisms.

***Quality assurance, user support, and user training***

*Katrina Learned*, Quality Assurance Manager. Hires, trains, and manages performance of quality assurance staff for this and other UCSC biomedical software projects.

*Pauline Fujita*, Biomedical Data Curator. Quality assurance of gene build. Primary user support person two days/week.

*Brooke Rhead*, Quality and User Support Analyst. Develops database-oriented test scripts. Primary user support person two days/week.

*Luvina Guruvadoo*, Quality Assurance and User Support. Tests software and validates third-party data. Primary user support person three days/week.

*Greg Roe*, Quality Assurance and User Interface. Develops web-oriented test scripts. Designs and reviews user interface.

*Steve Heitner*, Quality Assurance and User Training. Tests software and validates third-party data. Develops and gives user training courses one week/month.

***Systems administration***

*Gary Moro*, Computer Systems Administrator. Installs, configures, and maintains compute, storage, and network systems. Integrates UCSC campus and genome browser network and services. Diagnoses and fixes computer systems problems. Provides data loss prevention via backups, redundant power systems, etc. Supports staff workstations and laptops.

*Erich Weiler*, Computer Systems Administrator. Designs, installs, and maintains compute, storage, and network systems. Designs remote cluster. Diagnoses and fixes computer systems problems.

***Administration***

*Ann Pace*, Associate Director, Center for Biomolecular Science and Engineering. Oversees or coordinates all administrative functions associated with the UCSC Genome Browser, including proposals, reporting, financial management, resources and personnel. Supervises HR, communications, financial and other administrative support for this project. Serves as liaison with agency on administrative matters.

*Branwyn Wagman*, Associate Director, Operations and Communications. Manages UCSC human resources processes for the center. Coordinates funding extensions for personnel during gaps in funding (due to late awards and increments). Oversees communications and public and media relations, including web site that includes information about UCSC Genome Browser and staff.

*Kelly Sauder*, Financial Specialist. Conducts purchasing, reimbursements, travel advances, and other financial transactions. Negotiates service contracts and quotes and serves as liaison between campus business contracts office, browser staff, and vendors. Educates browser personnel on appropriate use of funds in compliance with federal and UC guidelines.

*Catherine Hawkins*, Research Accountant. Performs official campus accounting function for this award, maintaining up-to-date detailed and summary financial reports, coordinating internal campus processes, and reporting to divisional business office.

*Nadine Gassner*, Grant Program Administrator. Provides analytical support in personnel, financial, and resource management, ensuring this very large project effectively dovetails with other large CBSE grant projects, particularly where personnel and resources are shared between awards.

*Rochelle Fuller*, Administrative Specialist. Provides technical support on administrative software and prepares final versions of reports, letters, forms, etc. Supports special projects such as SAB meeting coordination, large internal meeting coordination, and internal documentation projects.

*Work-study students*. Perform basic clerical functions associated with genome browser staff and space, such as mail sorting, copying, scanning, signature collection, hand-delivery of documents, electronic filing, meeting room preparation, audio/visual set up, document preparation, administrative data entry and retrieval, and other supportive tasks. Fill in for permanent administrative staff as needed, when possible. Some perform web support and systems administration support.

##### ***Progress reporting***

Progress will be reported as required by the funding agency. At a minimum, we will provide an annual report of accomplishments and plans.

##### ***Scientific Advisory Board***

The genome browser project has an active Scientific Advisory Board (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).

Key personnel from the project team meet with the SAB and our grant agency contact, currently Adam Felsenfeld, every 12-15 months to present major accomplishments and seek the SAB’s advice on priorities, problems and future directions. Efforts are made to make the meetings as inclusive and efficient as possible: agendas and supplemental materials are provided to attendees in advance, and videoconferencing facilities accommodate members who are unable to attend in person. Following each meeting, a summary of the key decisions and action items is distributed to attendees, and items that directly affect the priorities of the genome browser project team are incorporated into the project planning. For example, our recent efforts to incorporate non-coding genes into the gene set and to distinguish common SNPs from the rest of the dbSNP data set were a response to suggestions from our SAB team.

At each SAB meeting, we review the areas of expertise of our board members, identify areas in which we’d like more representation, and then target potential new board members in these fields. For example, we recently invited Joe Gray to join our SAB to add more human disease focus.

## Bibliography and References Cited/Progress Report Publication List

#### Bibliography and References Cited

1. Kent, W.J. and Haussler, D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res*, **11**, 1541-1548. PMC311095.

2. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664. PMC187518.

3. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*, **13**, 103-107. PMC430961.

4. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. and Miller, W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**, 708-715. PMC383317.

5. Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S.L., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., Lindblad-Toh, K., Gibbs, R.A., Lander, E.S., Siepel, A., Haussler, D. and Kent, W.J. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, **17**, 1797-1808. PMC2099589.

6. Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M. and Haussler, D. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036-1046.

7. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. and Karolchik, D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204-2207. PMC2922891.

8. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006. PMC186604.

9. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, **31**, 51-54. PMC165576.

10. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., Hillman-Jackson, J., Kuhn, R.M., Pedersen, J.S., Pohl, A., Raney, B.J., Rosenbloom, K.R., Siepel, A., Smith, K.E., Sugnet, C.W., Sultan-Qurraie, A., Thomas, D.J., Trumbower, H., Weber, R.J., Weirauch, M., Zweig, A.S., Haussler, D. and Kent, W.J. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, **34**, D590-598. PMC1347506.

11. Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pedersen, J.S., Hsu, F., Hinrichs, A.S., Harte, R.A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G.P., Baertsch, R., Haussler, D. and Kent, W.J. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res*, **35**, D668-673. PMC1669757.

12. Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Kober, K.M., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D. and Kent, W.J. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, **36**, D773-779. PMC2238835.

13. Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A.S., Harte, R.A., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G.P., Haussler, D. and Kent, W.J. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*, **37**, D755-761. PMC2686463.

14. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., Pohl, A., Pheasant, M., Meyer, L.R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R.A., Giardine, B., Dreszer, T.R., Clawson, H., Barber, G.P., Haussler, D. and Kent, W.J. (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, **38**, D613-619. PMC2808870.

15. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D. and Kent, W.J. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, **39**, D876-882. PMC Journal - In Process.

16. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639-1645. PMC2752132.

17. Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097-6100. PMC332411.

18. Hu, Z., Snitkin, E.S. and DeLisi, C. (2008) VisANT: an integrative framework for networks in systems biology. *Brief Bioinform*, **9**, 317-325. PMC2743399.

19. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431-432. PMC3031041.

20. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, **100**, 11484-11489. PMC208784.

21. Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., Ruby, J.G., Brennecke, J., Hodges, E., Hinrichs, A.S., Caspi, A., Paten, B., Park, S.W., Han, M.V., Maeder, M.L., Polansky, B.J., Robson, B.E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D.G., Eastman, D.A., Rice, M., Weir, M., Hahn, M.W., Park, Y., Dewey, C.N., Pachter, L., Kent, W.J., Haussler, D., Lai, E.C., Bartel, D.P., Hannon, G.J., Kaufman, T.C., Eisen, M.B., Clark, A.G., Smith, D., Celniker, S.E., Gelbart, W.M. and Kellis, M. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219-232. PMC2474711.

22. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*, **13**, 103-107.

23. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. and Birney, E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*, **18**, 1814-1828. PMC2577869.

24. Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D. and Haussler, D. (2011) Cactus: Algorithms for genome multiple sequence alignment. *Genome Res*, **21**, 1512-1528. PMC Journal - In Process.

25. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S., Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J., Parker, A., Proctor, G., Vogel, J. and Searle, S.M. (2011) Ensembl 2011. *Nucleic Acids Res*, **39**, D800-806. PMC3013672.

26. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-65. PMC1716718.

27. Magrane, M. and Consortium, U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, **2011**, bar009. PMC3070428.

28. Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R. and Bateman, A. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*, **39**, D141-145. PMC3013711.

29. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, **100**, 15776-15781. PMC307644.

30. Sugnet, C.W., Kent, W.J., Ares, M., Jr. and Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput*, 66-77.

31. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., Hart, E., Suner, M.M., Landrum, M.J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J.L., Curwen, V., Dicuccio, M., Kellis, M., Lee, J., Lin, M.F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B.L., Mudge, J., Murphy, M.R., Murphy, T., Rajan, J., Rajput, B., Riddick, L.D., Snow, C., Steward, C., Webb, D., Weber, J.A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R. and Lipman, D. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, **19**, 1316-1323. PMC2704439.

32. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-515. PMC3146043.

33. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S. and Regev, A. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, **28**, 503-510. PMC2868100.

34. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. PMC Journal - In Process.

35. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325-2329. PMC Journal - In Process.

#### Progress Report Publication List

##### **Refereed journal papers**

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. Retrocopy contributions to the evolution of the human genome. *BMC Genomics*. 9(1):466. 2008. PMCID: PMC2584115.

Cline MS, Karchin R. Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*. 27:441-8. 2011. PMCID: PMC3105482.

Cline MS, Kent WJ. Understanding genome browsing. *Nat Biotechnol*. 27(2):153-5. 2009. http://www.nature.com/nbt/journal/v27/n2/full/nbt0209-153.html

Dreszer TR, Wall GD, Haussler D, Pollard KS. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res.* 17:1420-30. 2007. PMCID: PMC1987345.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*. 39(Database issue):D876-82. 2011. PMC Journal - In Process.

Karolchik D, Kuhn, RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*. 36:D773-9. 2008. PMCID: PMC2238835.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. Human genome ultraconserved elements are ultraselected. *Science*. 317(5840):915. 2007. <http://www.sciencemag.org/content/317/5840/915.full>

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed data sets. *Bioinformatics*. 26(17):2204-7. 2010. PMCID: PMC2922891.

Kern AD and Haussler D. A population genetic hidden Markov model for detecting genomic regions under selection. *Mol Biol Evol*. 27:1673-85. 2010. PMCID: PMC2912474.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res. 37:D755-D761. 2009. PMCID: PMC2686463.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darré F, Farré D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Della Valle G, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AF, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordoñez GR, López-Otín C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A, Roychoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA, Wilson RK. Comparative and demographic analysis of orang-utan genomes. *Nature*. 469(7331):529-53. 2011. PMCID: PMC 3060778.

Ma J, Ratan A, Raney BJ, Suh BB, Miller W, Haussler D. The infinite sites model of genome evolution. *Proc Natl Acad Sci U S A*. 105:14254-61. 2008.

Ma J, Ratan A, Raney BJ, Suh BB, Zhang L, Miller W, and Haussler D. DUPCAR: Reconstructing contiguous ancestral regions with duplications. *J Comput Biol*. 15:1007-27. 2008.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, Kosakovsky Pond SL, Nekrutenko A, Giardine B, Harris RS, Tyekucheva S, Diekhans M, Pringle TH, Murphy WJ, Lesk A, Weinstock GM, Lindblad-Toh K, Gibbs RA, Lander ES, Siepel A, Haussler D, Kent WJ. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*. 17:1797-808. 2007.

Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, Haussler D. Cactus graphs for genome comparisons. *J Comput Biol*. 18(3):469-81. 2011. PMC Journal - In Process.

Pohl AA, Sugnet CW, Clark TA, Smith K, Fujita PA, Cline MS. Affy Exon Tissues: Exon Levels in Normal Tissues in Human, Mouse, and Rat. *Bioinformatics*. 25:2442-3. 2009. PMCID: PMC2735668.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*. 38(Database issue):D613-9. 2010. PMCID: PMC2808870.

Sanborn JZ, Benz SC, Craft B, Szeto C, Kober KM, Meyer L, Vaske CJ, Goldman M, Smith KE, Kuhn RM, Karolchik D, Kent WJ, Stuart JM, Haussler D, Zhu J. The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Res.* 39(Database issue):D951-9. 2011. PMCID: PMC3013705.

Siepel A, Diekhans M, Brejová B, Langton L, Stevens M, Comstock CL, Davis C, Ewing B, Oommen S, Lau C, Yu HC, Li J, Roe BA, Green P, Gerhard DS, Temple G, Haussler D, Brent MR. Targeted discovery of novel human exons by comparative genomics. *Genome Res*. 17:1763-73. 2007. PMCID: PMC2099585.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 24:637-44. 2008.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J; Harvard FlyBase curators; Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*. 450(7167):219-32. 2007. PMCID PMC 2474711.

Twigger SN, Pruitt KD, Fernández-Suárez XM, Karolchik D, Worley KC, Maglott DR, Brown G, Weinstock G, Gibbs RA, Kent J, Birney E, Jacob HJ. What everybody should know about the rat genome and its online resources. *Nat Genet.* 40:523-7. 2008. PMCID: PMC 2505193.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A*. 104:18613-8. 2007. <http://www.pnas.org/cgi/content/abstract/0703637104v1>

Yeang CH and Haussler D. Detecting coevolution in and among protein domains. *PLoS Comput Biol.* 3:e211. 2007. <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0030211>

Yeang CH, Darot JF, Noller HF, Haussler D. Detecting the coevolution of biosequences--an example of RNA interaction prediction. *Mol Biol Evol*. 24:2119-31. 2007. <http://mbe.oxfordjournals.org/cgi/content/abstract/24/9/2119>

Zheng J, Yan J, Wang T, Mosbrook-Davis D, Dolan K, Christensen R, Stormo G, Haussler D, Lathrop R, Brachmann R, Burgess SM. Genome wide screens in yeast to identify potential binding sites and target genes of DNA binding proteins. *Nucleic Acids Res*. 36(1):e8. 2008. PMCID: PMC2248728.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle T, Haussler D. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol*. 3:e247. 2007. <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0030247>

Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. UCSC genome browser tutorial. *Genomics*. 92(2):75-84. 2008. <http://www.sciencedirect.com/science/article/pii/S0888754308000451>

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 467(7319):1061-73. 2010. PMCID PMC3042601.

##### **Non-refereed papers**

Harte RA, Diekhans M, Kent WJ, Haussler D. Guide to the UCSC Genome Browser. Cambridge, MA: NPG Education, 2010. <http://www.nature.com/scitable/ebooks/guide-to-the-ucsc-genome-browser-16569863>

Lathe W, Williams J, Mangan M, Karolchik D. Genomic Data Resources: Challenges and Promises. *Nature Education*. 1(3). 2008. <http://www.nature.com/scitable/topicpage/Genomic-Data-Resources-Challenges-and-Promises-743721>

##### **Conference papers**

Hsu F, Perez-Losada M, Kent J, Zweig A, Kuhn R, Crandall K, Sinangil F. GSID HIV sequence database: a powerful new tool for HIV vaccine research that relates HIV sequence information to clinical specimens and data. *AIDS Vaccine 2008*, Cape Town, South Africa.

Li G, Ma J, Zhang L. Selecting genomes for reconstruction of ancestral genomes. 5th Annual RECOMB Satellite Workshop on Comparative Genomics (2007). In McLysaght and A, Huson, DH. *Lecture Notes in Computer Science: Comparative Genomics*, 4751: 110-21, Springer, Berlin/Heidelberg, 2007.

Ma J, Ratan A, Raney BJ, Suh BB, Miller W, Haussler D. Reconstructing ancestral genomes using the infinite sites model of genome evolution. *The Biology of Genomes. CSHL*. May 2008.

Ma J, Ratan A, Zhang L, Miller W, Haussler D. A heuristic algorithm for reconstructing ancestral gene orders with duplications. 5th Annual RECOMB Satellite Workshop on Comparative Genomics (2007). In McLysaght and A, Huson, DH. *Lecture Notes in Computer Science: Comparative Genomics*, 4751:122-35, Springer, Berlin/Heidelberg, 2007.

##### **Book chapters**

Blanchette M, Diallo AB, Green ED, Miller W, Haussler D. Computational reconstruction of ancestral DNA sequences. In Murphy WJ. (ed): *Methods in Molecular Biology: Phylogenomics*, Springer, New York, 2007.

Harte RA, Karolchik D, Kuhn RM, Kent WJ, Haussler D. Databases and Genome Browsers. In Antonarakis S and Speicher M. (ed): *Human Genetics – Principles and Approaches – 4th Edition*, Springer, New York. 2010.

Harte RA, Karolchik D, Kuhn RM, Kent WJ, Haussler D. Databases and Genome Browsers. In Speicher M, Antonarakis S, Motulsky AG. (eds): *Vogel and Motulsky's Human Genetics, Principles and Approaches, 4th Edition*, Springer, New York, 2009.

Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, Rosenbloom KR, Zweig AS, Haussler D, Kent WJ. Comparative genomic analysis using the UCSC Genome Browser. In Bergman NH. (ed): *Methods in Molecular Biology: Comparative Genomics, Volume 1*, Springer, New York, 2007.

Ma J. Reconstructing ancient DNA sequences in silico. In Cooper DN and Kehrer-Sawatzki H. (ed): *Handbook of Human Molecular Evolution*, John Wiley & Sons, Inc., New Jersey, 2008.

Mangan ME, Williams JM, Kuhn RM, Lathe WC. [The UCSC Genome Browser: what every molecular biologist should know](http://www.currentprotocols.com/protocol/mb1909#_blank) In *Curr. Protoc. Mol. Biol.* 88:19.9.1-19.9.28. John Wiley & Sons, Inc., New York, 2009. <http://www.currentprotocols.com/protocol/mb1909>

Mangan ME, Williams JM, Lathe SM, Karolchik D, Lathe WC III. UCSC Genome Browser: Deep support for molecular biomedical research. In El-Gewely MR (ed): *Biotechnology Annual Review Volume 14*:63-108, Elsevier B.V., Amsterdam, 2008. [http://www.sciencedirect.com/science/bookseries/13872656](http://www.sciencedirect.com/science/bookseries/13872656" \t "_blank)

Rosenbloom K, Schaeffer S, Miller W. Phylogenomic Resources at the UCSC Genome Browser. In Murphy WJ. (ed): *Methods in Molecular Biology: Phylogenomics*, Springer, New York, 2007.

Rosenbloom K, Taylor J, Schaeffer S, Kent J, Haussler D, Miller W. Phylogenomic Resources at the UCSC Genome Browser. In Murphy WJ. (ed): *Methods in Molecular Biology: Phylogenomics*, Springer, New York, 2008. <http://genome.ucsc.edu/goldenPath/rosenbloom.pdf>

Wang, T. Using PhyloCon to Identify Conserved Regulatory Motifs. In Baxevanis, AD. (ed): *Current Protocols in Bioinformatics*, 19:2.12.1-2.12.29, John Wiley & Sons, Inc., New Jersey, 2007.

##### **Posters**

Harte R, Rosenbloom K, Trumbower H, Barber G, Karolchik D, Haussler D, and Kent , WJ. New Features on the UCSC Genome Browser. ISMB 2007.

Hinrichs AS, Kuhn RM, Karolchik D, Zweig AS, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Harte RA, Giardine B, Fujita P, Dreszer TR, Diekhans M, Clawson H, Barber GP, Haussler D, Kent WJ. Interpreting Genomic Data with the UCSC Genome Browser Tool Suite. GCD 2008.

Hinrichs AS, Hsu F, Kuhn RM, Raney BJ, Barber G, Zweig AS, Roe G, Haussler D and Kent WJ. Visually integrating genomic data in the UCSC Genome Browser. HGV 2011.

Kent WJ. Deeper, more distributed, more interactive annotations at UCSC. Biology of Genomes. CSHL 2010.

Kuhn R, Hinrichs AS, Zweig AS, Barber G, Karolchik D, Kent WJ, Haussler D. Display of High-throughput genomic data in the UCSC Genome Browser. European Society of Human Genetics annual meeting, Gothenburg, Sweden. 2010.

Kuhn R, Hinrichs AS, Hsu F, Zweig A, Barber G, Haussler D, Kent WJ. Using the UCSC Genome Browser to Interpret Genetic Disease. Genomic Disorders Meeting, Hinxton, UK. 2011.

Zweig AS, Raney BJ, Barber G, Hinrichs AS, Karolchik D, Haussler D, and Kent WJ. UCSC Genome Browser Track Data Hubs. Biology of Genomes, CSHL. 2010.

## Protection of Human Subjects

The UCSC Genome Browser displays three main classes of data from human subjects: data in peer-reviewed journals, data downloaded from other publicly accessible databases, and data uploaded by our users themselves. Our approach is to assume that data from these sources is either de-identified or has been properly consented for public access, since human data should not be publicly available except under these conditions. This is generally a valid assumption, but sometimes mistakes are made by journals or other databases, or changes are made in the perception of what is identifiable. For example, a few years ago the Genome Browser featured a set of pooled, high-resolution SNP data released by several groups doing genome-wide association studies (GWAS). Initially, this data set was considered de-identified because of its pooled nature; however, advances in the forensic uses of high-resolution SNP technology resulted in the data becoming identifiable. We subsequently joined the Wellcome Trust and several other data producers in removing the data from our site. For reasons such as this, we conduct an independent review of all data before posting it on the browser.

To safeguard users who upload their own data, we regard our site as merely a visualization tool and do our best to ensure that only the users themselves have access to their data, unless they make explicit efforts to share it with others. However, since these data sets are potentially sensitive, this proposal addresses efforts to ensure that uploaded data is indeed private. These include storing user-uploaded data on encrypted file systems and taking precautions against SQL injections and other standard website attacks.

We are in the process of submitting a request to the UCSC campus to review our UCSC Genome Browser research plan and determine whether our work is considered human subjects research and whether it qualifies for exemption.

## Consortium/Contractual Plan

We have a service agreement with the private company OpenHelix (<http://www.openhelix.com>) to provide free online UCSC Genome Browser tutorials and training materials to the public (see Training section). The service agreement is renegotiated annually, and payment is made through quarterly invoices.

## Letters of Support

See the following pages.