Overall Research Plan

Specific Aims

In this proposal we describe our plans to maintain, extend, and enhance the UCSC Genome Browser website, associated tools, and databases. Our overarching aims are to build useful software, load useful data, and support data exchange and analysis within the scientific community.

Aim 1. Develop, maintain, and extend software for the display and analysis of genomics resources.

Aim 2. Build and maintain genome browsers and related resources for species of biomedical interest.

Aim 3. Integrate data from the scientific community that help interpret the functions of various genome regions.

Aim 4. Develop, use, and extend data exchange standards such as file formats, APIs, and controlled vocabularies.

Research Strategy

Introduction

The primary purpose of our work is to help biomedical researchers understand the human genome. We aim to achieve this by further developing and maintaining the UCSC Genome Browser, its annotation databases, and associated tools at https://genome.ucsc.edu/. This 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 (1,2). It has since grown to encompass a broad set of data and tools widely used by the molecular biology, biomedical, and bioinformatics research communities and several consortia worldwide (Appendix 1).

Notable project achievements over the past sixteen years include the BLAT tool for quickly aligning RNA, DNA, protein, and translated sequences to genomic DNA (3); a pipeline for generating pairwise (4,5) and multiple vertebrate genome alignments (6,7); a widely used set of human gene models (8); practical techniques for handling visualization of terabyte-sized distributed next-generation sequencing data sets (9); a standardized hub mechanism for viewing remotely hosted custom track sets and assemblies alongside native browser tracks (10); the Variant Annotation Integrator tool for associating browser annotations with a custom set of variant calls (11); Genome Browser in a Box (GBiB), a standalone, personal version of the browser (12); and the multi-region display configuration for condensing a set of nonadjacent genome regions into a single browser view.

Perhaps most importantly, the work has resulted in the UCSC Genome Browser itself (13–25), 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 (Fig. 1).

In this renewal we describe our plans to maintain, extend, and enhance the Genome Browser website, tools, and databases over the next five years. At the overview level, the proposal has four specific aims that focus on building useful software, loading useful data, and supporting the exchange of data and research analysis within the scientific community. We will support this work with effective scientific, project, and personnel management; a plan for broadly disseminating the software tools, libraries, source code and data; and well-established training and outreach mechanisms. As with most long-term plans, we anticipate that we may occasionally have to accommodate the unexpected or to eliminate parts that become obsolete.

In our planning, we have taken care to constrain our costs and to find alternative external sources of funding for specific portions of the project. In addition to the Howard Hughes Medical Institute (HHMI) funding that supports Dr. Haussler and part of the computer systems administration staff, we benefit from funding provided by the California Institute for Quantitative Biosciences (QB3) and the California Institute for Regenerative Medicine (CIRM), and have been able to transfer some of our computational infrastructure costs to other funding sources. However, the project depends on NHGRI for the majority of its funding. We continue to make significant gains in automation and efficiency, but these are largely offset by an enormous increase in the volume of genomics data and Genome Browser users, and our proximity to Silicon Valley, where engineering salaries are escalating rapidly. This necessitates a modest increase in our budget request.

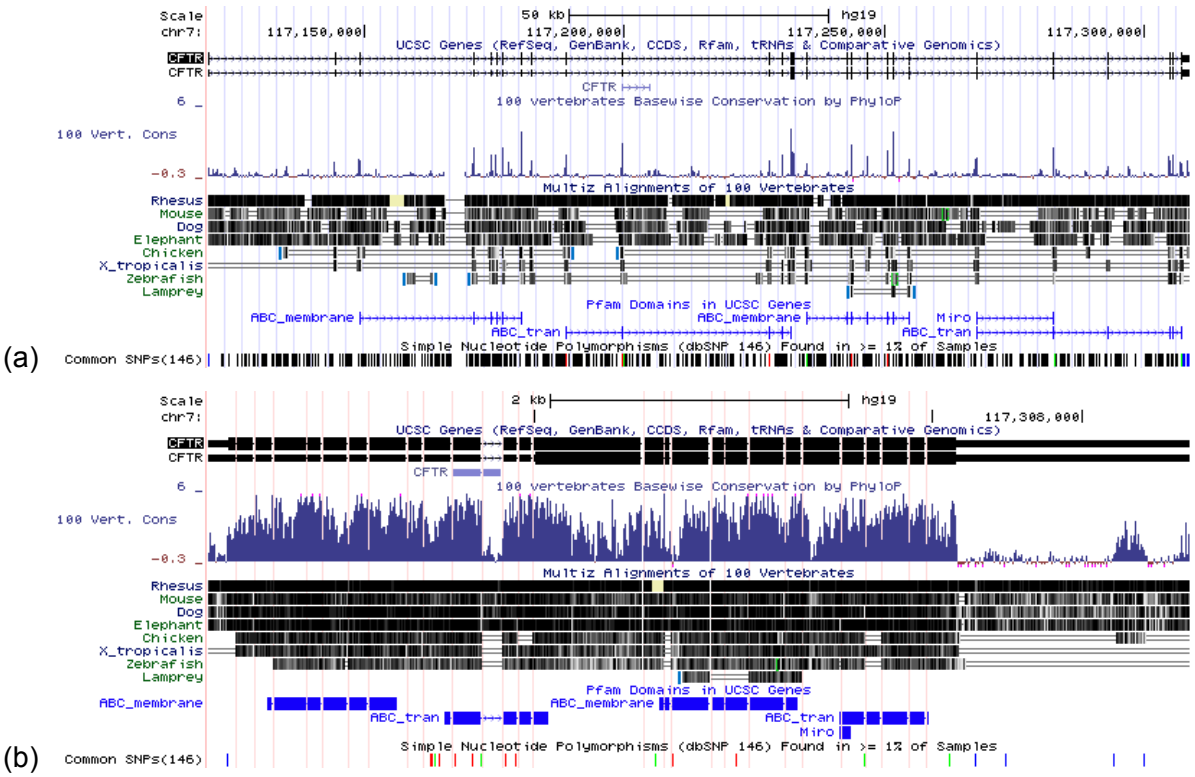


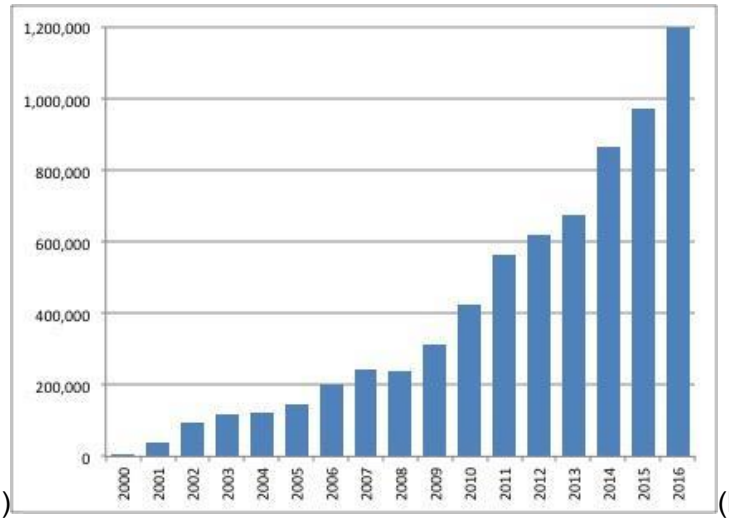
Figure 1. The main graphic of the UCSC Genome Browser showing the CFTR locus in default display mode *(a)* and in the new "exon-only" mode *(b)*. The latter configuration shows only exons (in any transcript including the small noncoding third transcript) with six flanking bases. For a large gene like CFTR, the vertebrate conservation and protein domain tracks in particular are easier to read with the introns abbreviated.

Aim 1. Develop, maintain, and extend software for the display and analysis of genomics resources

Significance

As sequencing machines and other automated laboratory devices generate increasing amounts of data, it is essential to have software tools that distill the data and present the results to the human eye. The website and tools at genome.ucsc.edu have fulfilled this need for tens of thousands of research scientists over the past sixteen years (Fig. 2). Our web-based tools 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, and our Unix command-line tools play a significant role in the analysis pipelines at institutions throughout the world.

The UCSC Genome Browser provides graphical access to a massive amount of genome sequence and annotation data, supporting the visualization of a user's own data as well as that natively hosted in our databases, in views ranging from individual bases up through entire genomes, and meeting reasonable expectations of privacy for data uploaded to the site. In addition to the Genome Browser, the genome.ucsc.edu website hosts a wide range of other useful tools. BLAT (3) rapidly aligns DNA, RNA, and protein sequence to genomic DNA. The Table Browser (26) and the Data Integrator (11) provide access to and logical analysis of the underlying data that is displayed visually on the Genome Browser. The Variant Annotation Integrator (11) produces functional-effect predictions for sequence variant calls. The In Silico PCR tool locates primer pairs on the genome or transcriptome. The Gene Sorter (27) displays sets of genes selected and sorted by various criteria such as tissue expression patterns. VisiGene displays *in situ* mRNA and fluorescently labeled gene micrographs in a display that can be zoomed much like Google maps. Genome Graphs displays information such as genome linkage and genome-wide association studies across all chromosomes simultaneously. Our popular Sessions tool allows users to save and share active screen scenarios, with more than 1,200 new sessions saved per month.

(a) 

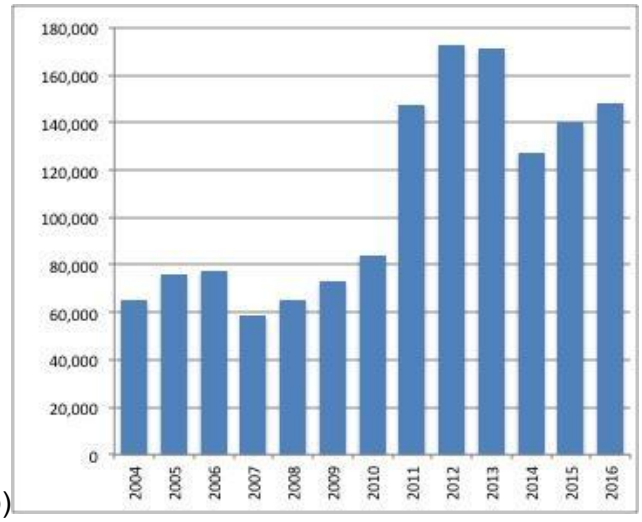
(b) 

Figure 2. *Panel (a)* - Average page hits per day on the genome.ucsc.edu website (and two mirror sites in Europe and Asia) 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 sixteen years in the public domain, our usage continues to grow. *Panel (b)* - Unique IP addresses accessing the genome.ucsc.edu website (and two mirror sites in Europe and Asia) per month, averaged over the calendar year. Years 2011-2013 include hits from randomly generated IP addresses from search engines, which have since stopped this practice.

Innovation

On a mature resource project such as the Genome Browser, some level of innovation is necessary to keep up with advances in technology and the changing needs of the scientific community. The convenience and versatility of our tool set is continually challenged by the rapid growth of new data sets and analyses, fed by new technologies and research collaborations, which in turn are built upon the critical mass of publicly accessible biological data. However, innovation must be tempered with discretion to preserve the stability and reliability of the site, and to minimize the impact on existing users. Our intention is to make our tools more accessible to novice users and new data providers, while not requiring current users to relearn the system.

The primary software innovations covered in this proposal are driven by the needs of our scientific community: the types of data they are working with, the increasing volumes of data they wish to understand and interpret, and the technologies they are using to do their work. We plan to extend our browser visualization support for several types of data, including personal genomes, single cell data, and data with non-local interactions (e.g. regions that are in close proximity in three-dimensional space, but not necessarily close in the browser's two-dimensional representation). To help scientists better understand and interpret large volumes of data, we intend to implement mechanisms that allow users to group and view summary properties over large aggregated sets of data and metadata. And finally, we intend to build a genome browser that can be used on mobile devices such as tablets and smartphones. These innovations are described in more detail in the Resource Project component.

Approach

We plan to continue our relatively conservative approach in developing and maintaining our software and databases, the bulk of which use well-established, stable technologies. In most cases, we will be able to implement the software enhancements and new features described in this proposal by incrementally expanding our existing code base. In the few cases that require the introduction of new technology, such as the support for mobile devices, we will explore our options and choose the path that allows us to best meet the needs of our users in a maintainable, sustainable fashion.

As the Genome Browser code base and databases have grown over the years, we have conscientiously applied software engineering and quality assurance best practices throughout our development and release processes. As a result, we have produced an extremely stable, reliable set of tools that is incrementally updated in a regular, timely fashion. We have also developed effective outreach and training mechanisms that address both the broad information needs of our users as well as the targeted needs of specific user groups

(Appendices 2-4). In this proposal, we plan to continue these successful approaches to creating our software and training our users, which we describe in detail in the Resource Project component and the Management, Dissemination and Training Core component.

Aim 2. Build and maintain genome browsers and related resources for species of biomedical interest

Significance

The UCSC Genome Browser database hosts a large repository of genome sequences, all of which can be accessed by the Genome Browser and by most of the tools on our website. As of August 2016, this included 176 assemblies from GenBank (28) that represent 96 different organisms across the tree of life, from vertebrates such as human, mouse, and zebrafish to insects and nematodes. We build browsers primarily on vertebrate genomes, but also support the major non-vertebrate model organisms – such as fly, worm, and yeast – in conjunction with the associated model organism databases.

All of the assemblies are annotated to varying degrees, ranging from a basic, minimal set of annotations, or "tracks", on the lesser-known species to the richly annotated human and mouse genomes. By integrating a wide variety of genomics data sets into a single display – including gene models (8,29–31), RNA alignments (3,28), epigenetic marks (32), regulation, expression levels (33), multiple alignments across related species (4,6,7,34), common human variation (35–38) and phenotype and disease associations – we provide biologists with a valuable resource for understanding the function of the genome, and offer physicians a valuable context for interpreting the medical consequences of genetic variation in patient and tumor sequences. Although several of the annotation tracks are generated by UCSC, an increasing number of our annotations are integrated from other resources, particularly for the human and mouse genomes (see Aim 3). This provides valuable visibility and access to these data sets, and allows users to view them in the context of other annotations.

To accommodate the rapidly increasing number of sequenced organisms, we have developed an "assembly hub" system that allows third parties to make their genome sequence and annotation files browsable on the genome.ucsc.edu website without the assistance or intervention of UCSC staff. Assembly hubs have been adopted by several institutions and are quickly becoming a standard.

Innovation

In upcoming years, we anticipate a continued need to accommodate the increasing number of sequenced genomes, as well as support the deepening exploration and understanding of the human genome. In this renewal, we propose ideas for improving the ease and efficiency with which new genome assemblies are integrated into our database and tool sets. We discuss our strategies for evaluating new multiple genome alignment software and creating new multiple alignments and other comparative genomics resources. We also describe our intentions to better integrate patches on the human and mouse genomes from the Genome Reference Consortium (GRC) into our tools, in an effort to make these interim updates to the reference assemblies more useful to the research community.

Approach

We will continue to strive for automation and efficiency in building the databases and files needed to add a new assembly to the browser. We also plan to automate the recomputation of UCSC-generated annotation tracks and the mapping of third-party annotation tracks to accommodate the GRC's patches and haplotype additions.

Generating multiple alignments is the most computationally intensive aspect of our project. In recent evaluations of other multiple alignment software candidates, we failed to find an alternative that outperformed our existing alignment pipeline (6,39). Therefore, we plan to continue to use our current strategy for producing multiple alignments of up to a few hundred vertebrate species while periodically evaluating other software for a superior option.

Aim 3. Integrate data from the scientific community that help interpret the functions of various genome regions

Significance

The genome sequence serves as a unifying framework where data from many scientific labs applying many techniques can be combined into one integrated view. On the human genome in particular, our annotation tracks highlight a broad range of research efforts within the science community. These include variation data from the 1000 Genomes Project (36) and other contributors to dbSNP, gene expression data from GTEx (33), epigenomics information from ENCODE (32), protein domain data from Pfam (40), Peptide Atlas (41), and UniProt (42), and information on human diseases from OMIM (43), ClinGen (44), ClinVar (45), COSMIC (46), DECIPHER (47) and NHGRI's GWAS catalog (48). In addition to these large-scale projects, we host a publications track that mines sequence information from the literature and maps it to the genome, providing a direct means for locating papers associated with a specific genomic region (24,49). We also host whole-genome data sets associated with many notable individual papers, such as the EVS Variants track that features data from the Exome Sequencing Project (ESP) (50) and the ExAC track showing variant data from Exome Aggregation Consortium studies (51).

On species other than human, we import significant information, such as gene models from the model organism databases, and in turn link back to the data sources. The model organism databases work with the Gene Ontology consortium (52) to provide functional descriptions of their gene models in a hierarchical controlled vocabulary, which we show on our genes pages and which can be used for filtering, searching, and grouping in our Gene Sorter tool.

The introduction of Genome Browser track hubs in 2011 and assembly hubs in 2012 (Aim 4) appreciably reduced the effort required by external labs and consortia to make their own data available through the Genome Browser website, and lessened the impact on our group as well. Several groups worldwide now use these mechanisms for sharing an integrated view of their data sets and analyses, e.g., Roadmap Epigenomics (53), miRcode (54), and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (55).

Innovation

We plan to continue integrating new data releases from projects that we currently support, as well as incorporating new data from selected projects and papers recommended to us by our users, our scientific advisory board, and our funding agency. When necessary, we will extend the Genome Browser display to accommodate new types of data (Aim 2). We recognize the need for scalability to accommodate increased data volumes. We will meet this challenge by automating updates for new releases when possible, and by encouraging data coordinating centers associated with major projects to use track hubs.

Approach

Over the years we have honed our approach for integrating data sets from external groups into the browser, a process we term "wrangling". Although we have developed many tools to help automate this process (and greatly reduce the need for our involvement when track and assembly hubs are used), there remain some aspects of the data integration job that resist automation, particularly those involving interactions with the labs producing the data. We plan to continue this hands-on approach for those data additions that cannot be automated or conveniently loaded through hubs.

Aim 4. Develop, use, and extend data exchange standards such as file formats, APIs, and controlled vocabularies

Significance

The Genome Browser is part of a larger ecosystem of biomedical data and resources. We build our displays and databases on the foundations of many other biomedical and computational resources, and increasingly other resources build on top of ours. Data standards and application programmer interfaces (APIs) play an important role in making it easier to build a resource that leverages the work of others. They play an even more critical role in maintaining viable connections between resources, even as the resources themselves evolve.

In the past sixteen years we have developed many file formats that have become standard within the genomics community, including MAF for multiple genome alignments, BED and BigBed for generic genome annotations, and BigWig for the rapid viewing of large sets of numerical data associated with genome coordinates (9). We have created the track and assembly hub systems for grouping together multiple genomics files with labels, colors, and other metadata to allow visualization in our browser, and have worked with other browser groups such as Ensembl (31) and Dalliance (56) to standardize the API. We provide SQL interfaces to our databases, an extensive library of C language functions, and command-line interfaces for our file formats and databases. Our entire source tree is available via the version-control system, git (57).

Innovation

We plan to continue to embrace our role in providing thoughtfully developed data exchange standards for the biomedical community, carefully extending existing models when possible and developing innovative new standards when needed. As part of this work, we intend to extend track hubs to accommodate all of the data formats displayed in the Genome Browser, including those used for comparative genomics and the new RNA expression-level displays developed for GTEx. We also plan to enrich the searchable metadata that can be kept on a hub using a flexible, hierarchical system we developed for the ENCODE project. This will allow multiple tracks to reference the same biosamples and include fields populated with controlled vocabularies.

We plan to add a JSON-based web services API to our system to support individuals developing analysis pipelines and alternative displays to the UCSC Genome Browser, as well as our own development of a mobile genome browser.

Approach

We intend to adopt existing formats and APIs where applicable, carefully construct well-documented, simple, flexible formats and APIs when working in new areas, and to advocate simplicity and harmony when participating in standardization efforts. We have demonstrated the efficacy of this approach throughout the sixteen years of our project.