bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Apollo: Democratizing genome annotation

Nathan Dunn[1,*], Deepak Unni[1], Colin Diesh[2], Monica Munoz-Torres[3], Nomi L. Harris[1], Eric Yao[2], Helena Rasche[4], Ian H. Holmes[2], Christine G. Elsik[5] and Suzanna E. Lewis[1]

[1] Lawrence Berkeley National Laboratory, Berkeley, California, United States of America
[2]University of California, Berkeley, California, United States of America
[3]Oregon State University, Corvallis, Oregon, United States of America
[4]University of Freiburg, Freiburg, Germany
[5]University of Missouri, Columbia, Missouri, United States of America

*Corresponding author
Email: nathandunn@lbl.gov (ND)

Abstract

Genome annotation is the process of identifying the location and function of a genome's encoded features. Improving the biological accuracy of annotation is a complex and iterative process requiring researchers to review and incorporate multiple sources of information such as transcriptome alignments, predictive models based on sequence profiles, and comparisons to features found in related organisms. Because rapidly decreasing costs are enabling an ever-growing number of scientists to incorporate sequencing as a routine laboratory technique, there is widespread demand for tools that can assist in the deliberative analytical review of genomic information. To this end, Apollo is an open source software package that enables researchers to efficiently inspect and refine the precise structure and role of genomic features in a graphical browser-based platform.

In this paper we first outline some of Apollo's newer user interface features, which were driven by the needs of this expanding genomics community. These include support for real-time collaboration, allowing distributed users to simultaneously edit the same encoded features while also instantly seeing the updates made by other researchers on the same region in a manner similar to Google Docs. Its technical architecture enables Apollo to be integrated into multiple existing genomic analysis pipelines and heterogeneous laboratory workflow platforms. Finally, we consider the implications that Apollo and related applications may have on how the results of genome research are published and made accessible.

- Source: https://github.com/GMOD/Apollo
- License (BSD-3): https://github.com/GMOD/Apollo/blob/master/LICENSE.md
- Docker: https://hub.docker.com/r/gmod/apollo/tags/, https://github.com/GMOD/docker-apollo
- Requirements: JDK 1.8, Node v6.0+
- User guide: http://genomearchitect.org; technical guide: http://genomearchitect.readthedocs.io/en/latest/
- Mailing list: apollo@lists.lbl.gov

Introduction

Apollo's design is based on the premise that the best genomic descriptions ('annotations') can be produced by starting with automatically-generated sequence features and then providing expert researchers with interactive editing tools to examine these multiple sources of evidence and collaboratively refine the genomic annotations. The first version of Apollo was a standalone desktop application (1). As software technologies advanced, each new generation of Apollo took advantage of these to improve the user experience. The most fundamental change occurred circa 2010 when Apollo

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

transitioned to running inside a web browser (2). Once Apollo became a web-based application that permits real-time collaboration, the user base grew to include research and teaching environments studying a wide variety of species. Our most recent version of Apollo (3) offers a broad range of researchers an accessible way to improve the biological precision of their genomic feature descriptions.

Organizations that use Apollo include Echinobase (4), Hymenoptera Genome Database (5), i5k Workspace (6), PhytoPath (7), TreeGenes (8), Vectorbase (9) and XenBase (10). To date, the i5K Workspace has supported publication of seven insect genomes that were manually curated with Apollo (11–17). Other projects that have used Apollo include genomes of additional insects (18–20), human parasites (21–23), birds (24,25), the sea lamprey (26), plants (27–29), fungi (30–35) and a plant pathogenic nematode (36). Projects such as the re-annotation of the whipworm genome by hundreds of high school students in the UK, supported by the Institute for Research in Schools (IRIS) (37), and the curation of 33,044 gene loci in the kiwifruit genome by 93 annotators, are evidence of Apollo's robust support for large dispersed projects.

The ease of setting up Apollo makes it appealing to small projects as well as large. For example, one small group used Apollo to annotate 14 genes of a fungal mitochondrial genome (32). Other reported Apollo use cases include annotating gene loci that pose challenges in automated gene prediction, such as the MHC-B region in the genome of the Mikado pheasant (25) and the effector complement of the flax rust pathogen Melampsora lini (33). Through the process of gene model curation, the use of Apollo can reveal species-specific genome characteristics that can be used to improve automated gene prediction. For example, curation of some gene models of the yellow potato cyst nematode, *Globodera rostochiensis*, using RNAseq alignments as evidence, revealed a high frequency of non-canonical splice sites. Subsequent use of these manually curated genes as a training set markedly improved the automated gene predictions (36).

Thanks to its ability to simplify and accelerate annotation efforts for both large and small projects, Apollo's user base continues to grow. Since 2015, Apollo has had an annual growth rate of roughly 70% for returning users, peaking at over 2,700 unique users one day in late 2017, with a current average of around 1,000 unique users per month.

Apollo's integrated graphical environment allows users to browse and modify the location(s) and other information for a variety of feature types and streamlines common editing tasks by providing built-in calculations for features such as predicted proteins, splice sites, and gene set membership. An overview of the interface is shown in Figure 1.

To briefly describe the basic capabilities, Apollo's Genomic Editing Workspace (bottom left of Figure 1) displays tracks of information gathered from upstream pipelines and individual users' analyses. These provide the evidence (predictions and alignment) for refining genomic annotations. Any combination of features can be dragged from the evidence area into the editable area, where researchers carry out their edits without affecting the features from the evidence area. When evidence features are dropped into the editable track, they are assigned a default feature type of "protein coding transcript" and the longest open reading frame is automatically calculated, as well as its gene membership based on overlap with the CDS in the same reading frame as existing transcripts. Exon boundaries can be set either by dragging them upstream or downstream, or by using a menu option to set them to the nearest upstream or downstream splice junction (these are automatically calculated based on the configured donor and acceptor dinucleotides).

Apollo provides several ways to customize the display. From the track tab, in the information and administration panel on the right, users can select the specific evidence tracks they want to view, categorize and filter tracks, and change the track order. The annotation tab lists every annotation across the genome, and can be searched by scaffold, identifier, researcher, or biological type. Information such

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

as the gene symbol, description, cross-references, Gene Ontology functional class, links to publications, or general comments on each annotation may also be added from this tab. The reference sequence tab provides a sortable and searchable list of every scaffold, including the length, name, and number of annotations on each, for navigation across the genome.

Design and Implementation

Apollo's design has always been driven by its users; their engagement in the development process has been a critical factor in Apollo's success. Over time the demographic of Apollo users has changed, with concomitant changes to Apollo's requirements. Notably, as sequencing costs have fallen, there are now a burgeoning number of projects targeting specific organisms, clades, or populations that frequently lack the funds or expertise to create their own software tools from scratch and are therefore reliant on available open source applications. Because members of these projects may be geographically distributed, they need tools that enable **real-time collaborative editing**. Additionally, annotating the effects different **variants** have on known genes has become a high priority research focus. And finally, particularly for collaborative projects, tracking the complete **annotation history** is crucial, not only for undo/redo operations but also to review the changes that have been made over time by different individuals.

Real-time Collaborative Editing

Apollo was designed with a standard client-server architecture (Figure 2) that can be run within a servlet container (e.g., Tomcat) and works with most relational database engines (e.g., PostgreSQL). The architecture provides a uniform authorization layer for external applications using its web services. For example, the i5K's project management software leverages Apollo web services to register new users and set appropriate user and group membership. The newly added users then have the necessary credentials to perform manual edits or utilize the same web services, allowing them to perform operations such as uploading bulk annotations.

Apollo's client interface is built as a JBrowse (40) plugin, a popular genome browser written in JavaScript. It provides the ability to import and export standard genomic data formats, flexible display of multiple types of genomic features, and fast scrolling and zooming. The primary editing client is a single-page application that embeds JBrowse. The server is built using Grails (38), an open source framework for developing web applications using Groovy (39) and other JVM languages. The Grails framework enables us to leverage well established technologies such as Spring (https://spring.io) for event control, the Grails Object Relational Mapper (GORM), Hibernate (http://hibernate.org) for efficiently mapping data objects to a backend persistent store, Ivy and Maven for build and plugin dependencies, and Grails plugins for security and navigation. Communication between the client and the server is provided through a REST API secured by the Apache Shiro library (https://shiro.apache.org/). To support integration into larger workflows, the web services that support user-interface functionality are fully exposed.

Concurrent editing by multiple users is implemented via WebSockets. WebSockets are well-supported in most recent web browsers and are an ideal technology to support push operations to all connected clients efficiently in real-time. Once a user's client connects to the server, WebSockets keep the line open for subsequent communication, including any structural and functional editing operations. This makes every annotation update in one client instantaneously visible in every other client. Apollo uses the STOMP (Streaming Text Oriented Messaging Protocol) protocol which uses a publish and subscribe communication style, minimizing communication overhead. WebSockets provide a robust and performant solution for pushing updates to multiple web clients that can fall back to a more traditional long-polling approach when client support is lacking as in older browsers.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Variants

In addition to allowing genomic features to be viewed and edited, Apollo provides the ability to annotate alterations in the underlying genomic sequence and visualize their impact (Figure 3). These may be assembly error *corrections*, to correct errors introduced in the sequencing and/or assembly process (a common issue when dealing with low-coverage genome sequences). Or these may be naturally occurring *variants,* genomic differences found among different members of a population. The effect of the annotated variants are reflected within the annotated genomic features they intersect with.

Annotation History

As researchers progressively refine the sequence features on a genomic region, information is automatically recorded for every change they make: what change was made; the time and date of the change; and the username (or email) of the editor. This edit history was a key design requirement, ensuring that all changes made are captured in a revertible, visual history of structural edits (Figure 4), which lets users graphically navigate through the different versions and roll back if necessary.

Results

Apollo's wide appeal across research projects of various sizes that focus on various organisms owes much to the many years of engagement between Apollo developers and its user community. In working with its users to maximize Apollo's utility for their breadth of organisms and purposes, it became clear to the development team that successful widespread uptake of Apollo depends on ensuring 1) reliable **scalability** so it can transparently handle very large genomes, a large number of genomes, and multiple users; 2) smooth **integration** into each group's technical environment; 3) a range of **customization** to accommodate different biological situations and project arrangements; and 4) direct engagement with users to encourage feedback and support **community contributions.**

Scalability

One of the major objectives when designing the architecture of the current version of Apollo was the ability for a single server to handle different dimensions of scale, whether it is thousands of genomes or large numbers of simultaneous users. We have encountered situations where a research group is studying many species in a particular clade; large, geographically distributed teams focused on a particular genomic region; and many students in a class working on team projects. Minimal requirements for Apollo are at least 500 MB of memory, or as much as several GB for optimal performance. However, with that allocation, we have optimized Apollo such that a single server can be successfully scaled to support several hundred genome projects and researchers. We tested and improved Apollo's performance and reliability via a combination of improved algorithms, optimized I/O requests, and more efficient database queries. As part of the testing process, we used a test suite that utilized the Apache JMeter load test tool, allowing the tool to simulate extraordinarily heavy read and write load over a sustained period. Additionally, we were able to scale up by modeling all organisms and users in a single database instance.

Ease of Integration

Biological data and tools do not exist in a vacuum. To enjoy wide use, bioinformatics environments such as Apollo need to be able to smoothly integrate with multiple analysis tools and user interfaces.

Web Services

Documented and secure web services are key to integrating any software into different bioinformatics ecosystems. Apollo exposes the methods used to drive the user interface as a web service, as well as providing services that support integration into different laboratories' existing environments. All methods are secured and require the same user permissions they would from the interactive browser application. Web services documentation is automatically generated from annotations within the software. There are many workflow environments that Apollo has been integrated into, typically after multiple alignment, filtering, and automated genome annotation steps. These environments include Galaxy (40) via the G-

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

OnRamp project (41), GenSAS (42,43), DNA Subway (44), and the i5K workspace (6). The i5K project leverages the user registration services, and the Galaxy Genome Annotation (GGA) project (45) automatically generates new projects in Apollo from data created via its biological workflow. The GGA project also provides a Python library for interacting with the Apollo API (46) and is used by projects such as BioInformatics Platform for Agroecosystem Arthropods (BIPAA) (47) and Texas A&M University Center for Phage Technology (TAMU-CPT) (48).

Import and Export

Importing new information as it becomes available is essential for revealing additional genomic insights. Likewise, exporting the curated annotations provides corrected information for downstream analysis, such as protein motif profiling. In either direction, a variety of standard genomic data formats, such as GFF3, BAM, GTF, GVF, GenBank, VCF, BED, BigWig, or the Chado database (49) are supported. These import/export capabilities are also available via a REST endpoint for direct programmatic use in other applications. Additionally, JBrowse has a large number of other input/output plugins, and associated visualization widgets, (https://gmod.github.io/jbrowse-registry/), which can be made available within Apollo.

Customization

Apollo's collection of configuration options enable it to meet the unique biological and organizational needs of individual projects. Options include: which organism genomes the server will host; the appropriate codon translation table to use for each genome; organism-specific acceptor and donor sites; how deep the '`undo`' stack should be; which algorithm to use when determining if transcripts are isoforms of the same gene; and many others.

In addition to the particular biological configuration, each project can specify the permissions granted to specific users or user-groups that may correspond, for example, to a laboratory or organism within a larger project. For more information about configuring Apollo, see http://genomearchitect.org/users-guide/.

Community contributions

As it has evolved, Apollo has greatly benefited from community contributions via bug reports, comments, feature suggestions, as well as directly from code changes submitted by external developer via pull requests. Many of Apollo's newer features are based on contributions from or joint development projects with members of the bioinformatics community. One recent example was the creation of the Genome Feature Widget (https://www.npmjs.com/package/genomefeaturecomponent) to provide a lightweight overview of genomic features in order to embed them within a web page. Working with external developers at the Human Phenotype Ontology (50) the Mouse Genome Database (51) and Wormbase (52), we expanded the Apollo web services to serve pieces of genomic evidence as JSON snippets that can be digested by the widget. The Genome Feature Widget is now being used by the Monarch Initiative (53) and the Alliance of Genome Resources (AGR) (54) in some of their web pages (Figure 5a), as well as to embed Apollo visualizations in other platforms such as Jupyter Notebooks (Figure 5b). Other examples of community contributions include addition of an "Instructor" administrator role to allow a teacher who does not handle the administration of the the project to more easily use Apollo in classes. Additionally, users have added web services , the ability to select tracks, and numerous build improvements.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Availability and Future Directions

Availability

- **Main website**: http://genomearchitect.org/. Source and documentation materials are linked from there.
- **Source:** https://github.com/GMOD/Apollo
- **License (BSD-3):** https://github.com/GMOD/Apollo/blob/master/LICENSE.md .
- **Local installation requirements**: Java 8+ JDK and Node.js 6+. Other requirements, such as Grails or Gradle, can be automatically installed if not present. Installing, running, and testing are all accomplished using a provided `bash` script.
- **Docker installation**: We provide a complete Docker implementation (55). Additionally, after every Apollo release, an Amazon Web Services EC2 public image is provided.
- **Feedback and code contributions**: We welcome improvements submitted as GitHub pull requests by the community.

Future Directions

As we work to increase Apollo's repertoire of visual exploration and visual analytics tools, several major enhancements are currently under development. First is improving the visualization of variants and their predicted effects to help in identifying disease-causing variants across diverse groups. Second is sequence coordinate transforms, which will combine different sequence regions into a single, synthetic region. This will allow the visualization of two or more genomic regions, from the length of entire chromosomes to just a few exons, within a single artificially constructed genomic region. Artificially joining scaffolds facilitates annotation of genomic features that were split in a fragmented assembly, or it can hide intra- and intergenic regions to provide a more densely information-rich visualization of the genome. Additionally, we plan to simplify the annotation workflow by eliminating the need for manual server-side preprocessing of genomes and genomic evidence during initial installation and allowing all configuration to be done via the web interface. Finally, we are hoping to further improve Apollo's performance by using graph databases.

Graph databases for performance improvement

Apollo relies on a traditional relational database, a well-established and performant technology that provides schema enforcement and transaction support, which are both requirements for a reliable curation tool. However, this is problematic if a user wants to promote an entire evidence track to the editing window, which vastly simplifies downstream merging of evidence. Genomic features are represented using a nested data model similar to Chado (49) and thus require multiple joins in order to retrieve them from the database, which is inherently inefficient, especially over larger sections of the genome. While denormalization is possible, the data is constantly changing due to edits, requiring a cascade of changes to ensure consistency. A coming solution, and one which improves the modeling of the data, will be to replace the relational database with a graph database. Experiments have suggested that they offer an order of magnitude speedup while still providing schema enforcement, transaction support, and a more adaptive schema.

Genome publication

The plummeting price of sequencing is leading to an explosion of genomic sequencing. This in turn is producing a growing trove of information from which to gain insights into each new genome's encoded features. Projects such as the joint Wellcome Trust Sanger Centre and Beijing Genome Institute project to sequence every vertebrate genome (56) are the tip of the iceberg. While large genomic resource centers may have funding for staff members to maintain genome curation efforts for a handful of organisms, this will not scale to the annotation effort needed to cover the rapidly accumulating genomes of other organisms or strains. Annotation on this larger scale requires contributions from a much wider community

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of researchers, who have the biological expertise to improve annotations, but require an efficient user interface that is collaborative and accessible through a web browser. Apollo provides a free, open source annotation platform that these researchers can integrate into their workflow, thereby helping to democratize the process of genome annotation.

Frequently, when a genome analysis project is completed, gene annotations and metadata generated during the life of the project become inaccessible to other researchers unless they are integrated into a stably supported central resource (57). To overcome this, annotations could be saved to a central track hub registry (such as Ensembl or UCSC), as a read-only JBrowse snapshot of the annotations. This would not only preserve the data in a GFF3 file, but would also offer a means of viewing it. A JBrowse registry hub, where indexed snapshots are listed, would ensure the long-term preservation of the evidence trail that supports each annotation and its micro-attribution. This archive methodology has been shown to be successful within the G-OnRamp group's Galaxy workflow (https://github.com/goeckslab/jbrowse-archive-creator).

Expanding on the idea of the track hub 'publication' of a genome, Apollo establishes a new data capture and dissemination paradigm that can benefit the individual researcher as well as the wider community. By recording their genome annotations precisely, Apollo makes it possible for researchers to claim professional credit for their work when it is utilized in subsequent research. Citable contributions could derive from creation, structural changes, and for enriching an annotation with additional information such as the biological function associated with a gene. The annotations produced by a particular author, identified in Apollo by their Open Researcher and Contributor ID (ORCID, https://orcid.org/), would become citable micro-publications, and could be included in data exports to show the provenance of the annotations. A 'genome press release' in which the contributors release a summary of their genome annotation set findings would bring the annotations of new organisms and clades to the attention of the wider community and provide appropriate credit to the authors.

Acknowledgements

Thanks to the Apollo and JBrowse communities for bringing issues to our attention, requesting new features, contributing code, integrating and using our product. Some notable contributors, in addition to those in the author list: Yating Liu, Luke Sargent, and Antony Bretaudeau.

We also thank Chris Childers and Monica Poelchau at the National Agricultural Library for use cases, bug reports, feedback and stress-testing.

This work has been supported by a National Institutes of Health grant R01-GM080203 from the National Institute of General Medicine Sciences.

References

1. Lewis SE, Searle SMJ, Harris N, Gibson M, Lyer V, Richter J, et al. Apollo: a sequence annotation editor [Internet]. Vol. 3, Genome Biol. 2002. p. research0082.1. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0082

2. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol [Internet]. 2013 Aug 30;14(8):R93. Available from: http://dx.doi.org/10.1186/gb-2013-14-8-r93

3. Unni D, Dunn N, Yao E, Buels R, Li Y, Holmes I, et al. GMOD/Apollo: Apollo2.1.0(JB#d3827c) [Internet]. 2018. Available from: https://zenodo.org/record/1295754

4. Kudtarkar P, Cameron RA. Echinobase: an expanding resource for echinoderm genomic information. Database [Internet]. 2017 Jan 1;2017. Available from: http://dx.doi.org/10.1093/database/bax074

5. Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, et al. Hymenoptera Genome Database: integrating genome annotations in

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HymenopteraMine. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D793–800. Available from: http://dx.doi.org/10.1093/nar/gkv1208

6. Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee C-Y, et al. The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res [Internet]. 2015 Jan;43(Database issue):D714–9. Available from: http://dx.doi.org/10.1093/nar/gku983

7. Pedro H, Maheswari U, Urban M, Irvine AG, Cuzick A, McDowall MD, et al. PhytoPath: an integrative resource for plant pathogen genomics. Nucleic Acids Res [Internet]. 2016 Jan 4;44(D1):D688–93. Available from: http://dx.doi.org/10.1093/nar/gkv1052

8. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol [Internet]. 2014 Mar 4;15(3):R59. Available from: http://dx.doi.org/10.1186/gb-2014-15-3-r59

9. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res [Internet]. 2015 Jan;43(Database issue):D707–13. Available from: http://dx.doi.org/10.1093/nar/gku1117

10. James-Zorn C, Ponferrada VG, Burns KA, Fortriede JD, Lotay VS, Liu Y, et al. Xenbase: Core features, data acquisition, and data processing. Genesis [Internet]. 2015 Aug;53(8):486–97. Available from: http://dx.doi.org/10.1002/dvg.22873

11. Poynton HC, Hasenbein S, Benoit JB, Sepulveda MS, Poelchau MF, Hughes DST, et al. The Toxicogenome of Hyalella azteca: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. Environ Sci Technol [Internet]. 2018 May 15;52(10):6009–22. Available from: http://dx.doi.org/10.1021/acs.est.8b00837

12. McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, et al. Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome Biol [Internet]. 2016 Nov 11;17(1):227. Available from: http://dx.doi.org/10.1186/s13059-016-1088-8

13. Linnen CR, O'Quin CT, Shackleford T, Sears CR, Lindstedt C. Genetic Basis of Body Color and Spotting Pattern in Redheaded Pine Sawfly Larvae (Neodiprion lecontei). Genetics [Internet]. 2018 May;209(1):291–305. Available from: http://dx.doi.org/10.1534/genetics.118.300793

14. Schoville SD, Chen YH, Andersson MN, Benoit JB, Bhandari A, Bowsher JH, et al. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). Sci Rep [Internet]. 2018 Jan 31;8(1):1931. Available from: http://dx.doi.org/10.1038/s41598-018-20154-1

15. Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, et al. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome Biol [Internet]. 2016 Sep 22;17(1):192. Available from: http://dx.doi.org/10.1186/s13059-016-1049-2

16. Kanost MR, Arrese EL, Cao X, Chen Y-R, Chellapilla S, Goldsmith MR, et al. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. Insect Biochem Mol Biol [Internet]. 2016 Sep;76:118–47. Available from: http://dx.doi.org/10.1016/j.ibmb.2016.07.005

17. Benoit JB, Adelman ZN, Reinhardt K, Dolan A, Poelchau M, Jennings EC, et al. Unique features of a global human ectoparasite identified through sequencing of the bed bug genome. Nat Commun [Internet]. 2016 Feb 2;7:10165. Available from: http://dx.doi.org/10.1038/ncomms10165

18. Fu Y, Yang Y, Zhang H, Farley G, Wang J, Quarles KA, et al. The genome of the Hi5 germ cell line from Trichoplusia ni, an agricultural pest and novel model for small RNA biology. Elife [Internet]. 2018 Jan 29;7. Available from: http://dx.doi.org/10.7554/eLife.31628

19. Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury J-M, Duvic B, et al. Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda, Noctuidae) with different host-plant ranges. Sci Rep [Internet]. 2017 Sep 25;7(1):11816. Available from: http://dx.doi.org/10.1038/s41598-017-10461-4

20. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci U S A [Internet]. 2015 Nov 3;112(44):E5907–15. Available from: http://dx.doi.org/10.1073/pnas.1516410112

21. Zhu Y, Engström PG, Tellgren-Roth C, Baudo CD, Kennell JC, Sun S, et al. Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of Malassezia sympodialis. Nucleic Acids Res [Internet]. 2017 Mar 17;45(5):2629–43. Available from: http://dx.doi.org/10.1093/nar/gkx006

22. Ifeonu OO, Simon R, Tennant SM, Sheoran AS, Daly MC, Felix V, et al. Cryptosporidium hominis gene catalog: a resource for the selection of novel Cryptosporidium vaccine candidates. Database [Internet]. 2016 Oct 19;2016. Available from: http://dx.doi.org/10.1093/database/baw137

23. Ifeonu OO, Chibucos MC, Orvis J, Su Q, Elwin K, Guo F, et al. Annotated draft genome sequences of three species of Cryptosporidium: Cryptosporidium meleagridis isolate UKMEL1, C. baileyi isolate TAMU-09Q1 and C. hominis isolates TU502_2012 and UKH1. Pathog Dis [Internet]. 2016 Oct;74(7). Available from: http://dx.doi.org/10.1093/femspd/ftw080

24. Colquitt BM, Mets DG, Brainard MS. Draft genome assembly of the Bengalese finch, Lonchura striata domestica, a model for motor skill variability and learning. Gigascience [Internet]. 2018 Mar 1;7(3):1–6. Available from: http://dx.doi.org/10.1093/gigascience/giy008

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

25. Lee C-Y, Hsieh P-H, Chiang L-M, Chattopadhyay A, Li K-Y, Lee Y-F, et al. Whole-genome de novo sequencing reveals unique genes that contributed to the adaptive evolution of the Mikado pheasant. Gigascience [Internet]. 2018 May 1;7(5). Available from: http://dx.doi.org/10.1093/gigascience/giy044

26. Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ, et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet [Internet]. 2018 Feb;50(2):270–7. Available from: http://dx.doi.org/10.1038/s41588-017-0036-1

27. Pilkington SM, Crowhurst R, Hilario E, Nardozza S, Fraser L, Peng Y, et al. A manually annotated Actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics [Internet]. 2018 Apr 16;19(1):257. Available from: http://dx.doi.org/10.1186/s12864-018-4656-3

28. Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, et al. Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets. DNA Res [Internet]. 2017 Sep 23; Available from: http://dx.doi.org/10.1093/dnares/dsx038

29. Xu Z, Luo H, Ji A, Zhang X, Song J, Chen S. Global Identification of the Full-Length Transcripts and Alternative Splicing Related to Phenolic Acid Biosynthetic Genes in Salvia miltiorrhiza. Front Plant Sci [Internet]. 2016 Feb 5;7:100. Available from: http://dx.doi.org/10.3389/fpls.2016.00100

30. Chen L, Gong Y, Cai Y, Liu W, Zhou Y, Xiao Y, et al. Genome Sequence of the Edible Cultivated Mushroom Lentinula edodes (Shiitake) Reveals Insights into Lignocellulose Degradation. PLoS One [Internet]. 2016 Aug 8;11(8):e0160336. Available from: http://dx.doi.org/10.1371/journal.pone.0160336

31. Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, et al. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics [Internet]. 2018 May 22;19(1):381. Available from: http://dx.doi.org/10.1186/s12864-018-4750-6

32. Jelen V, de Jonge R, Van de Peer Y, Javornik B, Jakše J. Complete mitochondrial genome of the Verticillium-wilt causing plant pathogen Verticillium nonalfalfae. PLoS One [Internet]. 2016 Feb 3;11(2):e0148525. Available from: http://dx.doi.org/10.1371/journal.pone.0148525

33. Nemri A, Saunders DGO, Anderson C, Upadhyaya NM, Win J, Lawrence GJ, et al. The genome sequence and effector complement of the flax rust pathogen Melampsora lini. Front Plant Sci [Internet]. 2014 Mar 24;5:98. Available from: http://dx.doi.org/10.3389/fpls.2014.00098

34. Schuelke TA, Westbrook A, Broders K, Woeste K, MacManes MD. De novo genome assembly of Geosmithia morbida, the causal agent of thousand cankers disease. PeerJ [Internet]. 2016 May 2;4:e1952. Available from: http://dx.doi.org/10.7717/peerj.1952

35. Syme RA, Tan K-C, Hane JK, Dodhia K, Stoll T, Hastie M, et al. Comprehensive Annotation of the Parastagonospora nodorum Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. PLoS One [Internet]. 2016 Feb 3;11(2):e0147221. Available from: http://dx.doi.org/10.1371/journal.pone.0147221

36. Eves-van den Akker S, Laetsch DR, Thorpe P, Lilley CJ, Danchin EGJ, Da Rocha M, et al. The genome of the yellow potato cyst nematode, Globodera rostochiensis, reveals insights into the basis of parasitism and virulence. Genome Biol [Internet]. 2016 Jun 10;17(1):124. Available from: http://dx.doi.org/10.1186/s13059-016-0985-1

37. Genome Decoders: The Human Whipworm [Internet]. 2017 [cited 2018 Sep 25]. Available from: https://www.sanger.ac.uk/news/view/uk-students-working-scientists-help-prevent-childhood-parasite-infection

38. Smith G, Ledbrook P. Grails in Action [Internet]. Manning; 2014. 545 p. Available from: https://market.android.com/details?id=book-ZyCdmwEACAAJ

39. The Apache Groovy programming language [Internet]. 2018 [cited 2018 Sep 25]. Available from: http://groovy-lang.org/

40. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res [Internet]. 2018 Jul 2;46(W1):W537–44. Available from: http://dx.doi.org/10.1093/nar/gky379

41. G-OnRamp – Create Genome Browsers for Genome Annotation [Internet]. 2018 [cited 2018 Sep 25]. Available from: http://gonramp.wustl.edu/

42. Lee T, Peace C, Jung S, Zheng P, Main D, Cho I. GenSAS — An online integrated genome sequence annotation pipeline. In: 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI) [Internet]. 2011. p. 1967–73. Available from: http://dx.doi.org/10.1109/BMEI.2011.6098712

43. Humann JL. GenSAS v5.1: A Web-Based Platform for Structural and Functional Annotation and Curation of Genomes. In: PAG - Plant and Animal Genome XXVI Conference (January 13 - 17, 2018) [Internet]. Washington State University; 2018 [cited 2018 Sep 25]. Available from: https://pag.confex.com/pag/xxvi/meetingapp.cgi/Paper/28336

44. Hilgert U, McKay S, Khalfan M, Williams J, Ghiban C, Micklos D. DNA Subway: Making Genome Analysis Egalitarian. In: Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment [Internet]. ACM; 2014 [cited 2018 Sep 25]. p. 70. Available from: http://dl.acm.org/citation.cfm?doid=2616498.2616575

45. Bretaudeau A, Dunn N, Gladman S, Grüning B, Rasche H, Seemann T. Galaxy Genome Annotation project: Integrating Galaxy and GMOD for genome annotation. F1000Res [Internet]. 2018 Oct 3 [cited 2018 Oct 3];7. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

http://dx.doi.org/10.7490/f1000research.1116180.1

46. Rasche H. Apollo Python Integration [Internet]. 2017 [cited 2018 Sep 25]. Available from: https://pypi.org/project/apollo/

47. Bretaudeau A. Deployment of genome databases for insects using Galaxy Genome Annotation [Internet]. F1000Research; 2017 Jul 11 [cited 2018 Sep 25]. Available from: http://dx.doi.org/10.7490/f1000research.1114390.1

48. Rasche H, Grüning B, Dunn N, Bretaudeau A. GGA: Galaxy for genome annotation, teaching, and genomic databases. F1000Res [Internet]. 2018 Oct 3 [cited 2018 Oct 3];7. Available from: http://dx.doi.org/10.7490/f1000research.1116181.1

49. Mungall CJ, Emmert DB, FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics [Internet]. 2007 Jul 1;23(13):i337–46. Available from: http://dx.doi.org/10.1093/bioinformatics/btm189

50. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res [Internet]. 2017 Jan 4;45(D1):D865–76. Available from: http://dx.doi.org/10.1093/nar/gkw1039

51. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome Database Group. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D836–42. Available from: http://dx.doi.org/10.1093/nar/gkx1006

52. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, et al. WormBase 2017: molting into a new stage. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D869–74. Available from: http://dx.doi.org/10.1093/nar/gkx998

53. McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, et al. Navigating the Phenotype Frontier: The Monarch Initiative. Genetics [Internet]. 2016 Aug;203(4):1491–5. Available from: http://dx.doi.org/10.1534/genetics.116.188870

54. Alliance of Genome Resources [Internet]. [cited 2018 Nov 22]. Available from: https://www.alliancegenome.org/

55. Dunn N, Rasche H, Paulini M. GMOD/docker-apollo: Apollo 2.1.0 Docker+PostgreSQL [Internet]. 2018. Available from: https://zenodo.org/record/1296537

56. Researchers reboot ambitious effort to sequence all vertebrate genomes, but challenges loom [Internet]. Science | AAAS. 2018 [cited 2018 Nov 19]. Available from: http://www.sciencemag.org/news/2018/09/researchers-reboot-ambitious-effort-sequence-all-vertebrate-genomes-challenges-loom

57. Gibney E, Van Noorden R. Scientists losing data at a rapid rate. Nature News [Internet]. 2013 Dec 19 [cited 2018 Oct 8]; Available from: http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

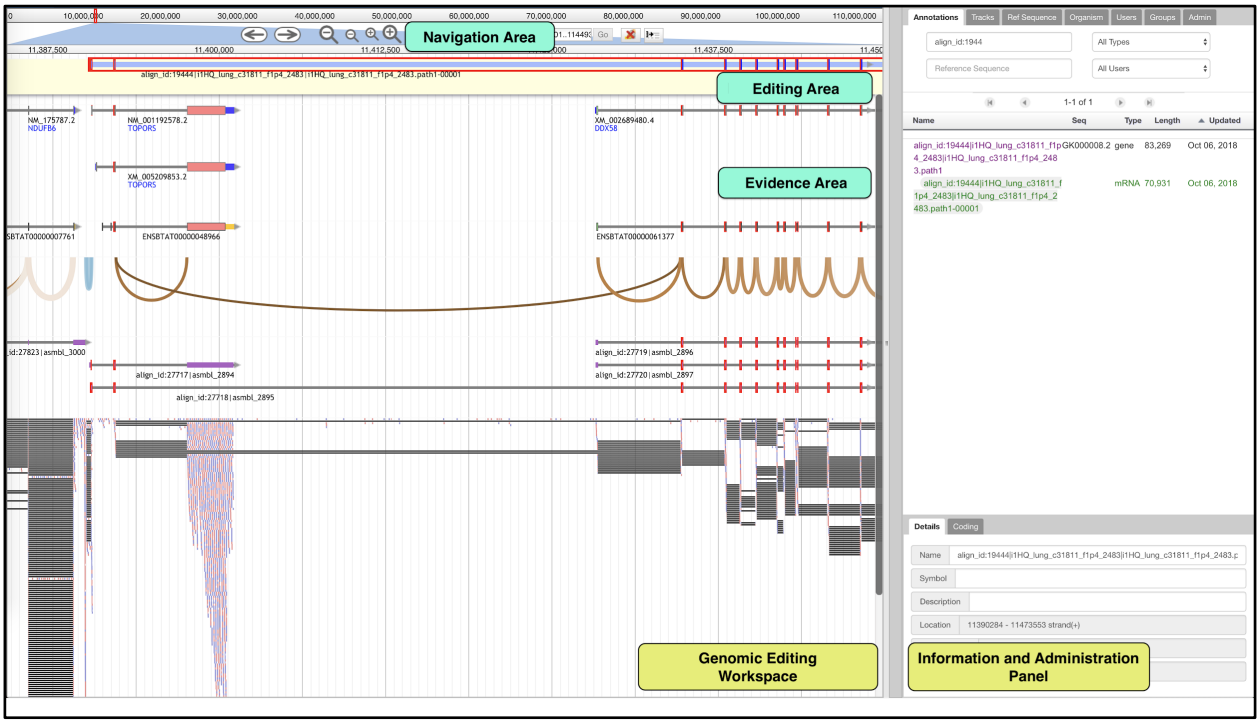


Figure 1: The Apollo Genome Editor has two main panels: a **Genomic Editing Workspace** and a closeable **Information and Administration Panel** which contains a range of configurable tabs. Within the **Genomic Editing Workspace,** the **Navigation Area** offers several ways to move to a region of interest. A user may: move upstream or downstream in fixed units; zoom in or out; or enter the coordinates or feature identifier to center on an exact genomic location. The **Evidence Area** contains data imported from local or remote files. In the **Editing Area** users can create annotations by dragging up evidence to create editable features of various types: coding and non-coding transcripts, pseudogenes, repeat regions, transposons, variant calls, transcription and translation start sites, and others. In the above example, the evidence area shows that there are reads spanning across exons that belong to two separate, previously known transcripts - NM_001192578.2 and XM_002689480.4. In the editing area, we add these known transcripts as annotations and then merge the two transcripts to create a single transcript. This newly created annotation then goes through additional refinements, to ensure that the transcript is a faithful representation of the evidence observed.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

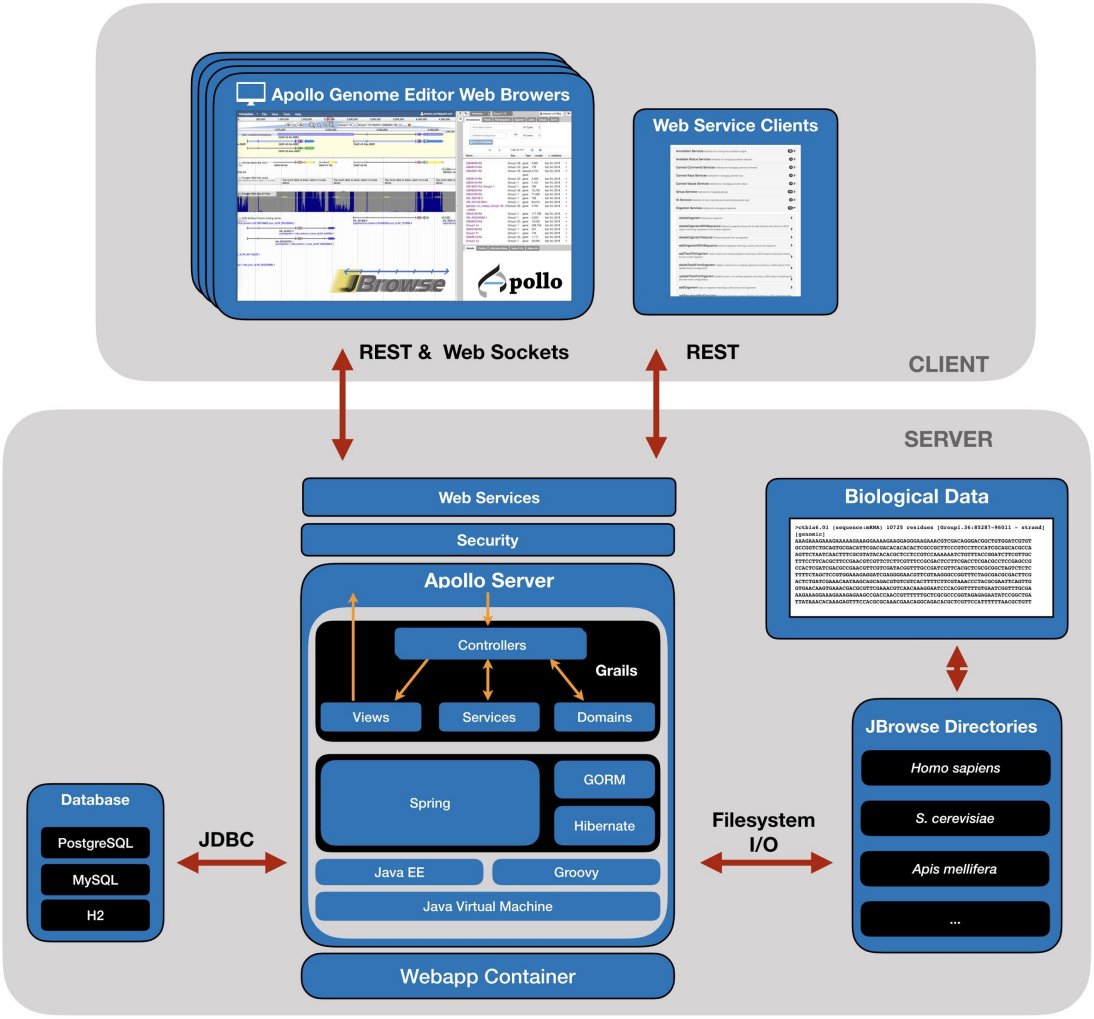


Figure 2: Apollo's client-server architecture. The server is built over the Grails framework using a relational database backend (RDBMS), for example PostgreSQL, MySQL or H2. Genomic evidence is provided by pointing to existing JBrowse directories that contain processed biological data. The Apollo Genome Editor and Web Services clients communicate with the server via REST and WebSockets.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

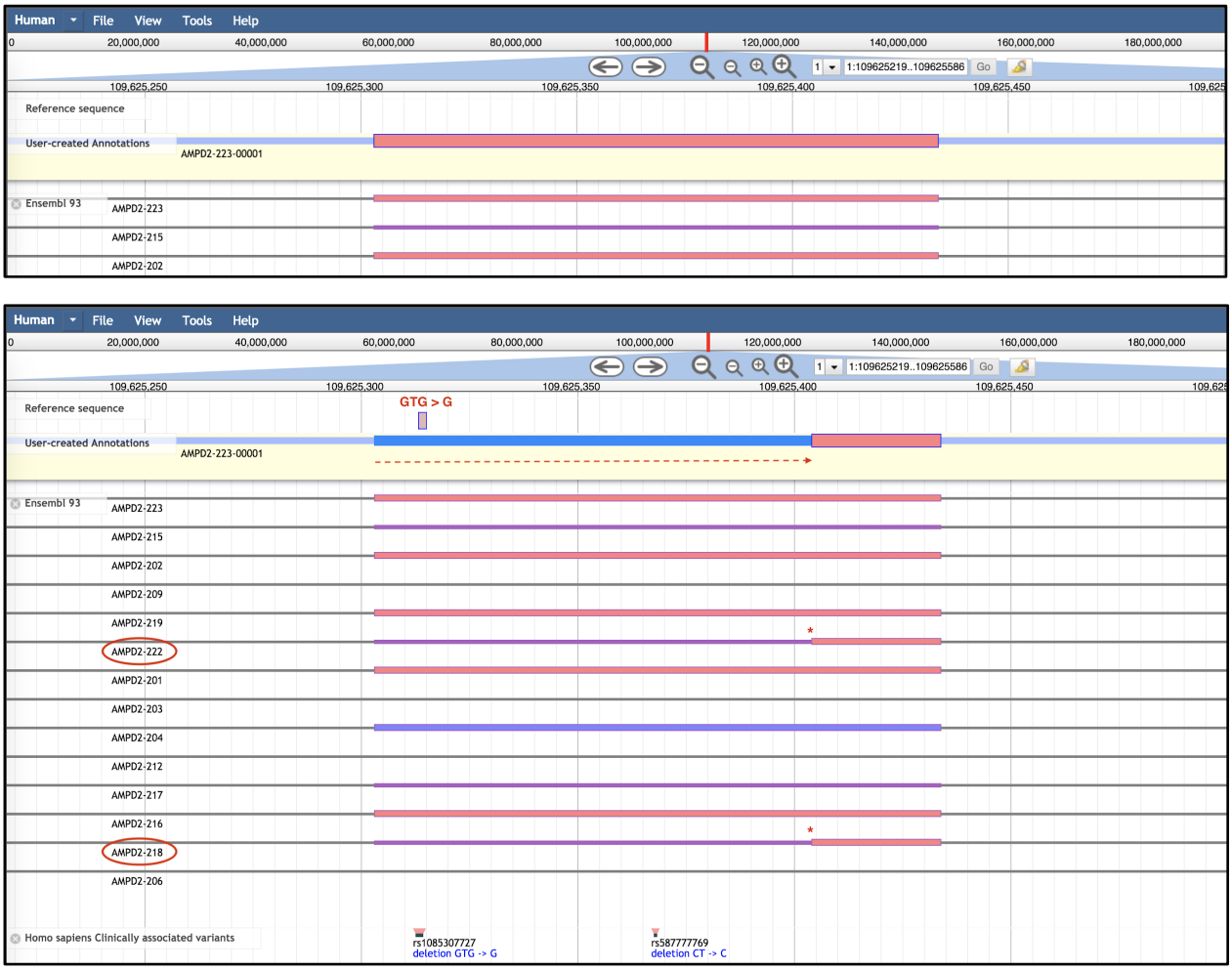


Figure 3. Example of variant annotation in Apollo. A) AMPD2-223, an isoform of gene AMPD2 as seen from the Evidence Area (truncated for space). From AMPD2-223, an isoform is dragged into the Editing Area. B) The deletion variant rs1085307727, from the 'Homo sapiens Clinically associated variants' track, overlaps with AMPD2-223-00001. Creating a corresponding deletion in the Editing Area of the Sequence Track allows visualization of the effect of the variant on transcript AMPD2-223-00001. Here, the transcription start site has moved further downstream, as indicated by the red dashed line. In this case, the altered form of the transcript recapitulates other alternate isoforms for this gene (AMDP2-218 and AMPD2-222), which are circled and starred for clarity.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

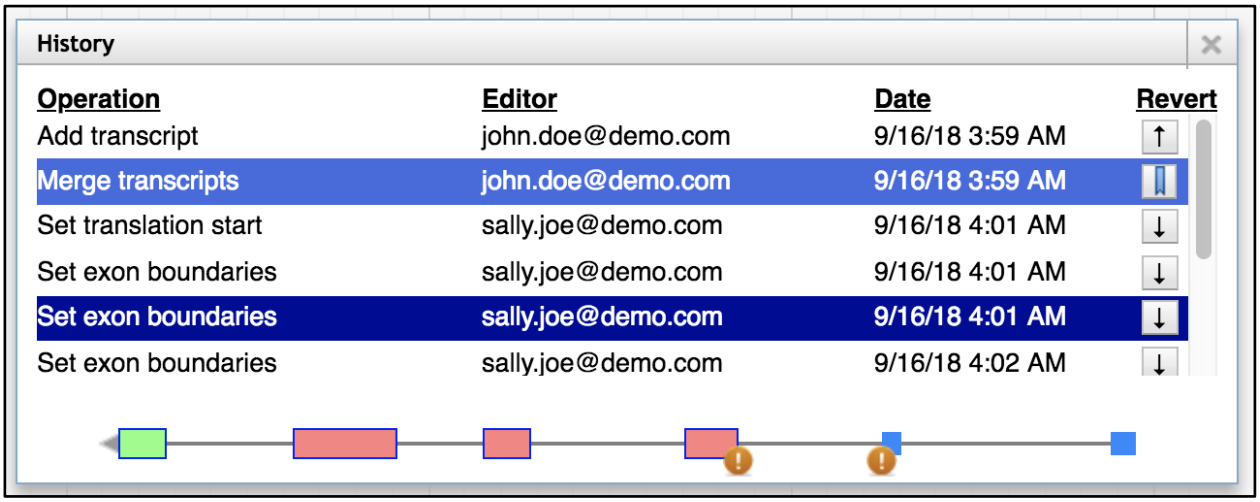
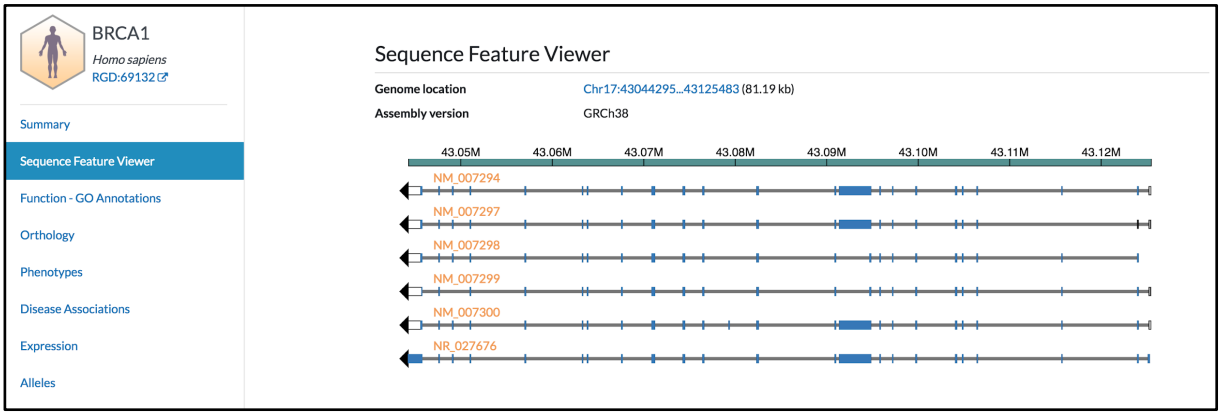
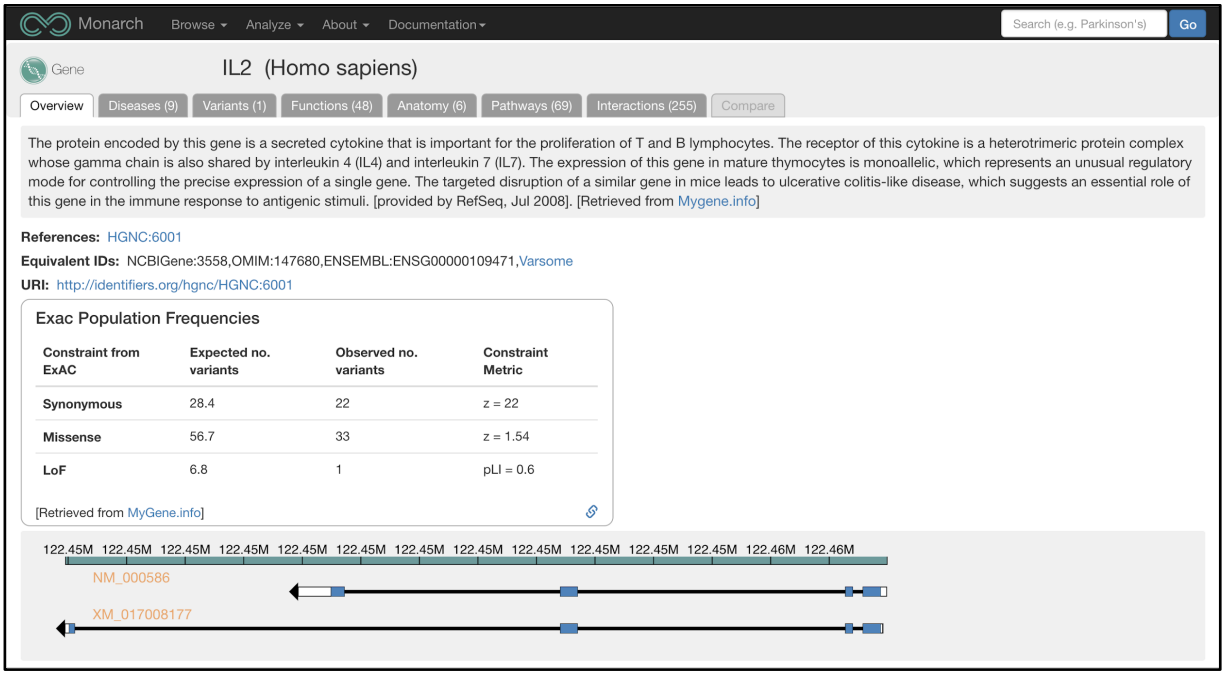


Figure 4. The history navigator allows visual navigation of genomic edits as well as the ability to return to previous versions. The current version is indicated with a bookmark icon (in the Revert column). Users can select any version from the history, and make edits starting from that version if desired. The orange circles with an exclamation point indicate non-canonical splice sites.

bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.





bioRxiv preprint doi: https://doi.org/10.1101/512376; this version posted January 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

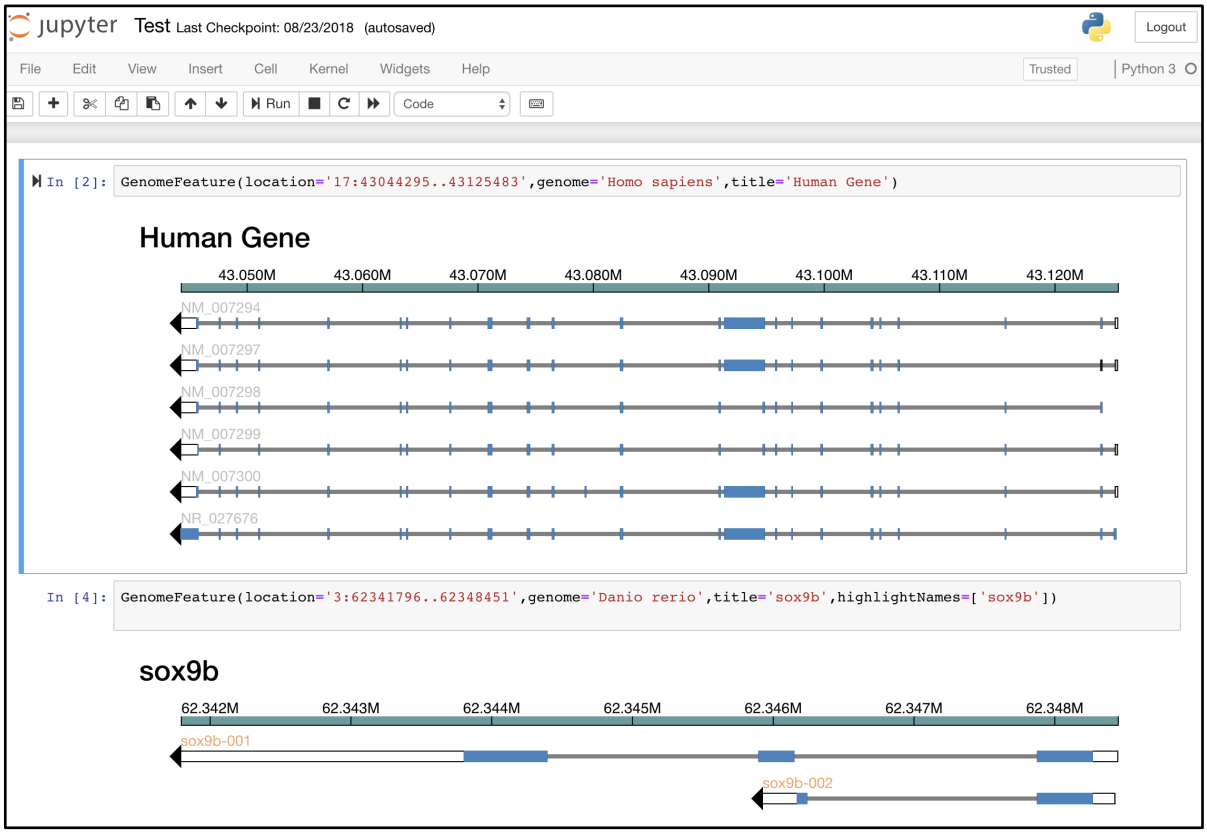
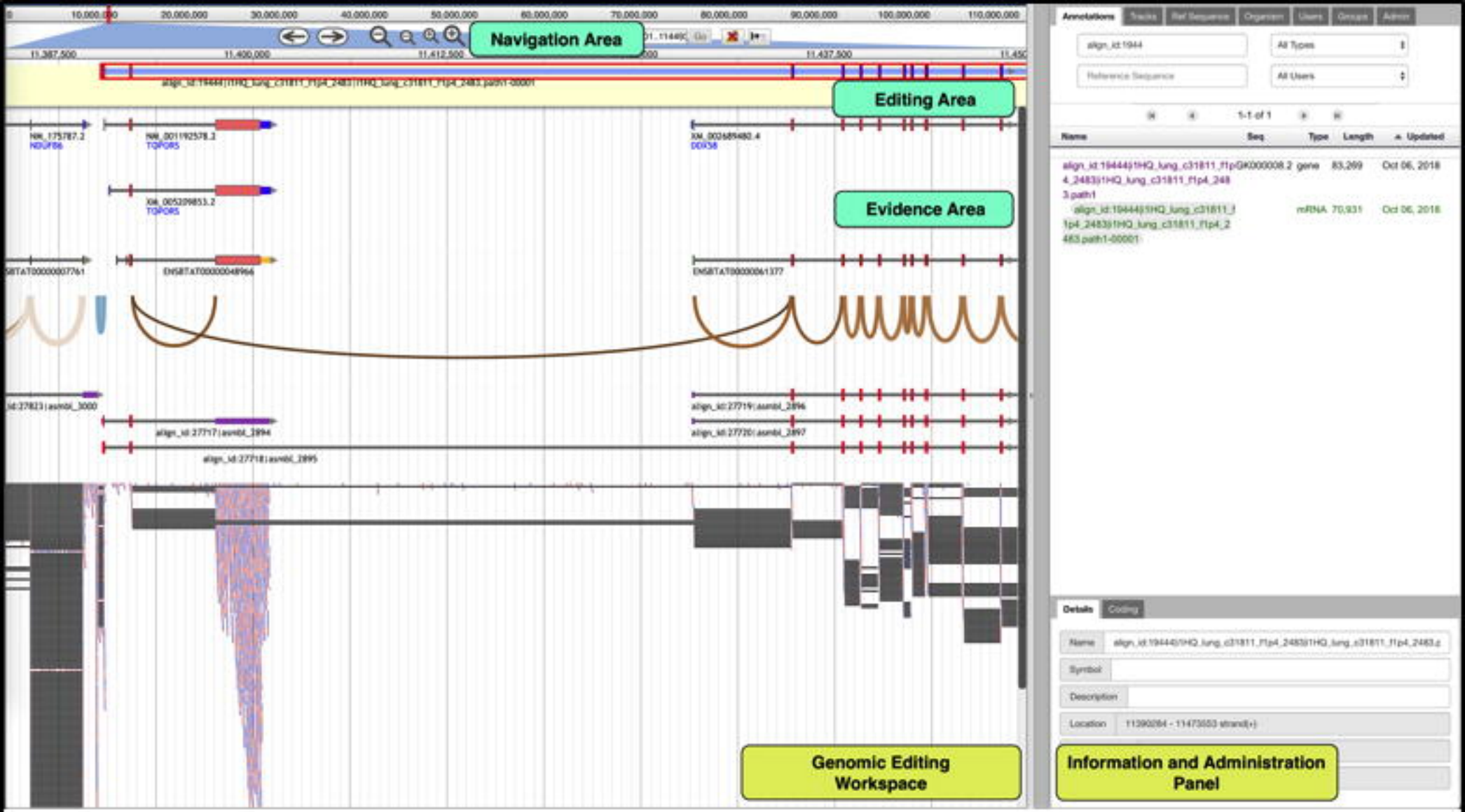
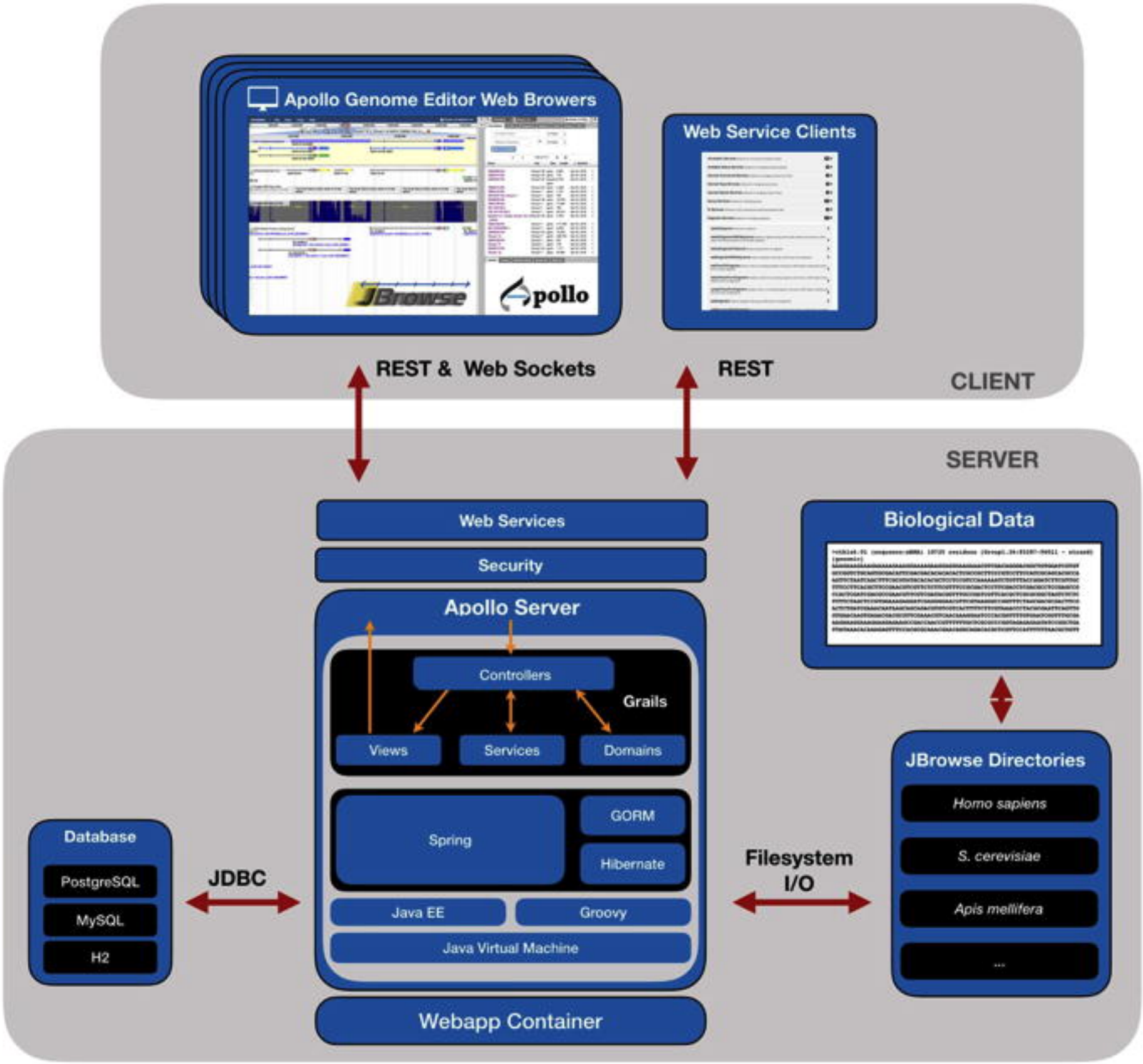
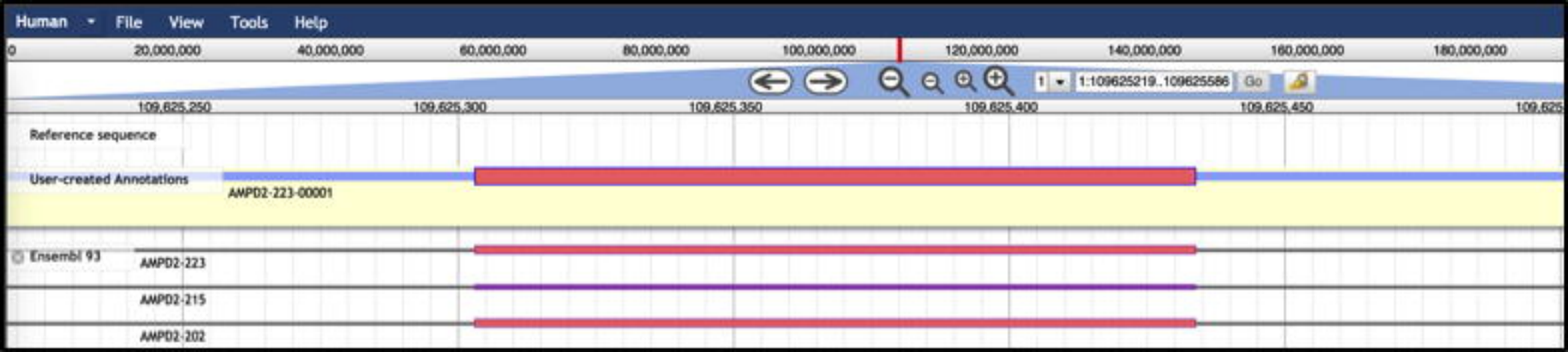
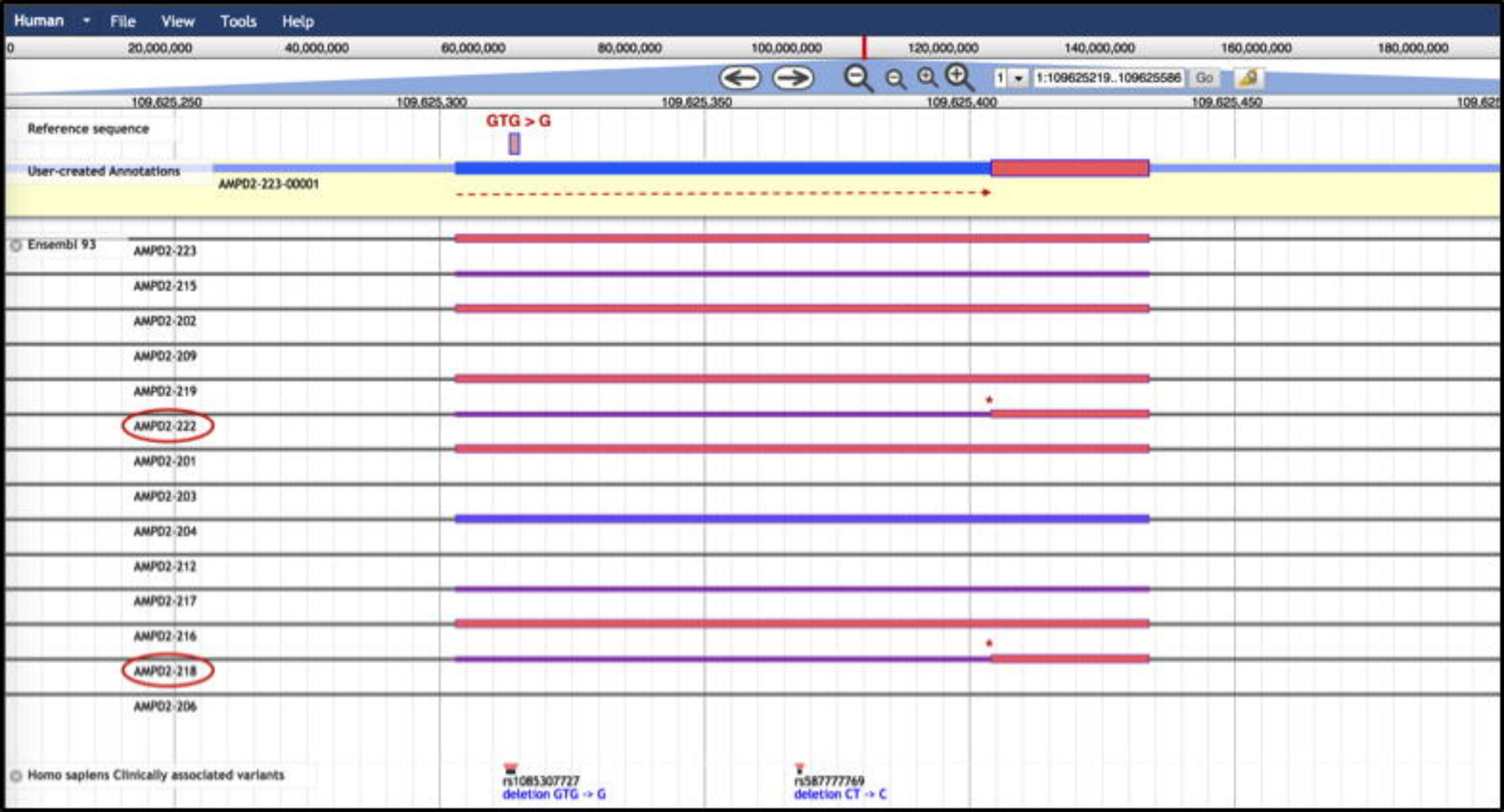


Figure 5. Three demonstrations of the Genome Feature Component npm widget (https://www.npmjs.com/package/genomefeaturecomponent) show examples that leverage Apollo's web services by consuming snippets of data for particular regions. A) The Alliance for Genome Resources web page (https://www.alliancegenome.org/gene/HGNC:1100) visualizes the Human BRCA1 gene. B) the Monarch Initiative (https://monarchinitiative.org/) web page visualizes the human IL2 gene. C) We embed the npm widget within a Jupyter Notebook widget to be called directly from a Python command-line script.


Navigation Area
Editing Area
Evidence Area
Genomic Editing Workspace
Information and Administration Panel

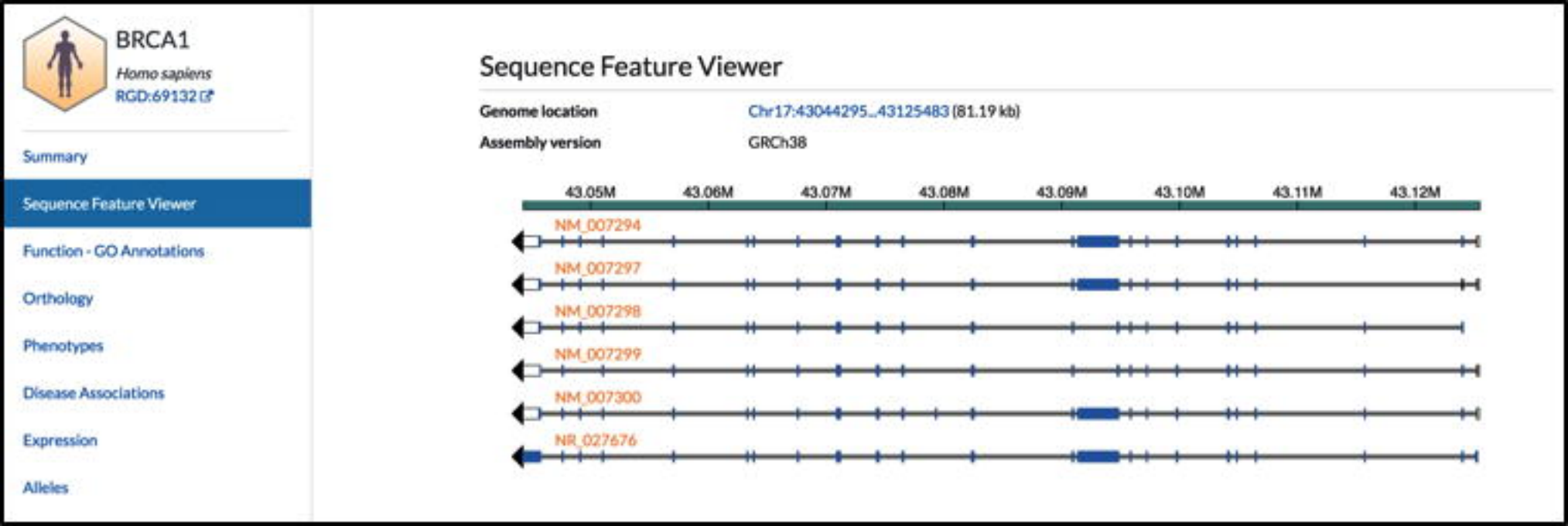

Apollo Genome Editor Web Browers
JBrowse
Apollo
Web Service Clients
REST & Web Sockets
REST
CLIENT
SERVER
Web Services
Security
Apollo Server
Controllers
Grails
Views
Services
Domains
Spring
GORM
Hibernate
Java EE
Groovy
Java Virtual Machine
Webapp Container
Biological Data
JBrowse Directories
Homo sapiens
S. cerevisiae
Apis mellifera
...
Database
PostgreSQL
MySQL
H2
JDBC
Filesystem
I/O


Human
File
View
Tools
Help
0
20,000,000
40,000,000
60,000,000
80,000,000
100,000,000
120,000,000
140,000,000
160,000,000
180,000,000
1
1:109625219..109625586
Go
109,625,250
109,625,300
109,625,350
109,625,400
109,625,450
109,625
Reference sequence
User-created Annotations
AMPD2-223-00001
Ensembl 93
AMPD2-223
AMPD2-215
AMPD2-202


Human
File
View
Tools
Help
0
20,000,000
40,000,000
60,000,000
80,000,000
100,000,000
120,000,000
140,000,000
160,000,000
180,000,000
1:109625219..109625586
Go
109,625,250
109,625,300
109,625,350
109,625,400
109,625,450
GTG > G
Reference sequence
User-created Annotations
AMPD2-223-00001
Ensembl 93
AMPD2-223
AMPD2-215
AMPD2-202
AMPD2-209
AMPD2-219
AMPD2-222
AMPD2-201
AMPD2-203
AMPD2-204
AMPD2-212
AMPD2-217
AMPD2-216
AMPD2-218
AMPD2-206
Homo sapiens Clinically associated variants
rs1085307727
deletion GTG -> G
rs587777769
deletion CT -> C

History

Operation	Editor	Date	Revert
Add transcript	john.doe@demo.com	9/16/18 3:59 AM	↑
Merge transcripts	john.doe@demo.com	9/16/18 3:59 AM	
Set translation start	sally.joe@demo.com	9/16/18 4:01 AM	↓
Set exon boundaries	sally.joe@demo.com	9/16/18 4:01 AM	↓
Set exon boundaries	sally.joe@demo.com	9/16/18 4:01 AM	↓
Set exon boundaries	sally.joe@demo.com	9/16/18 4:02 AM	↓


BRCA1
Homo sapiens
RGD:69132
Summary
Sequence Feature Viewer
Function - GO Annotations
Orthology
Phenotypes
Disease Associations
Expression
Alleles
Sequence Feature Viewer
Genome location
Chr17:43044295...43125483 (81.19 kb)
Assembly version
GRCh38
43.05M
43.06M
43.07M
43.08M
43.09M
43.10M
43.11M
43.12M
NM_007294
NM_007297
NM_007298
NM_007299
NM_007300
NR_027676

Gene

IL2 (Homo sapiens)

The protein encoded by this gene is a secreted cytokine that is important for the proliferation of T and B lymphocytes. The receptor of this cytokine is a heterotrimeric protein complex whose gamma chain is also shared by interleukin 4 (IL4) and interleukin 7 (IL7). The expression of this gene in mature thymocytes is monoallelic, which represents an unusual regulatory mode for controlling the precise expression of a single gene. The targeted disruption of a similar gene in mice leads to ulcerative colitis-like disease, which suggests an essential role of this gene in the immune response to antigenic stimuli. [provided by RefSeq, Jul 2008]. [Retrieved from Mygene.info]

References: HGNC:6001

Equivalent IDs: NCBIGene:3558,OMIM:147680,ENSEMBL:ENSG00000109471,Varsome

URI: http://identifiers.org/hgnc/HGNC:6001

Exac Population Frequencies

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	28.4	22	z = 22
Missense	56.7	33	z = 1.54
LoF	6.8	1	pLI = 0.6

[Retrieved from MyGene.info]

jupyter Test Last Checkpoint: 08/23/2018 (autosaved)

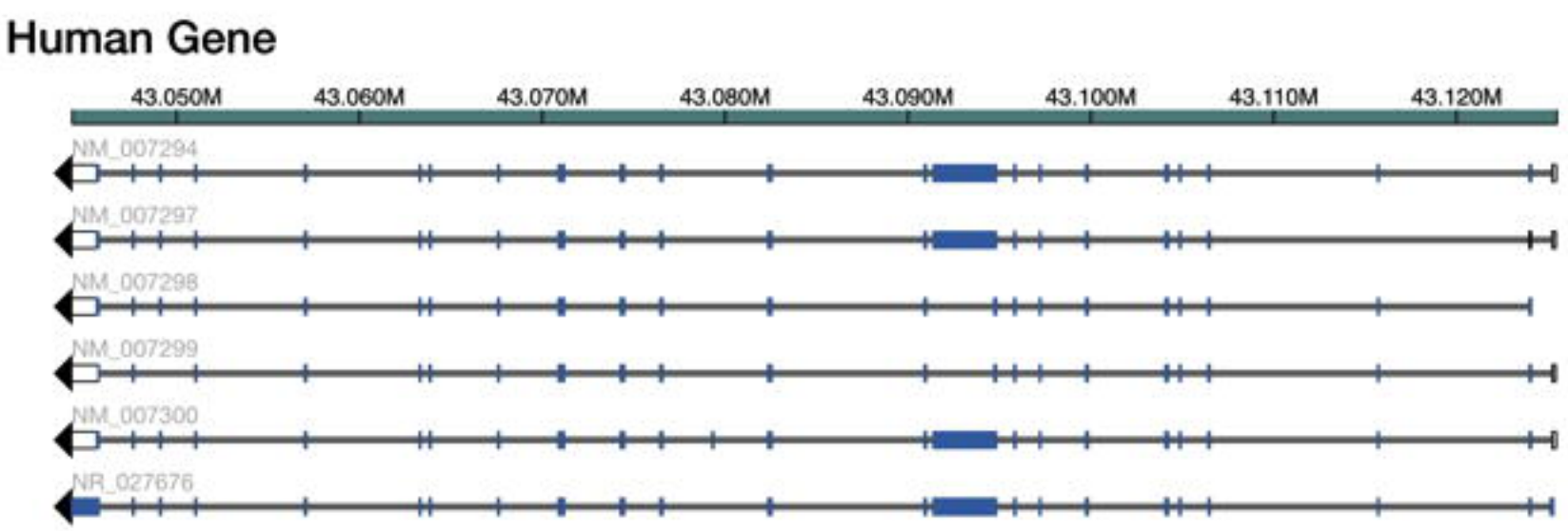
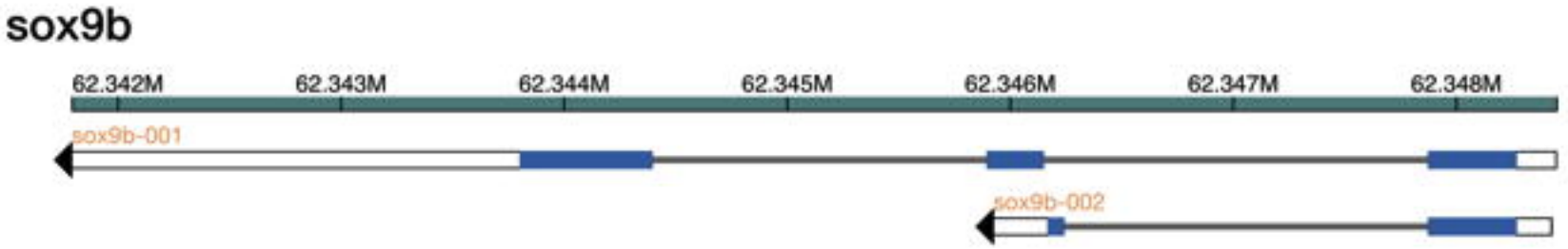
Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [2]:
```
GenomeFeature(location='17:43044295..43125483',genome='Homo sapiens',title='Human Gene')
```



In [4]:
```
GenomeFeature(location='3:62341796..62348451',genome='Danio rerio',title='sox9b',highlightNames=['sox9b'])
```