Genomics Made Easier: An Introductory Tutorial to Genome Datamining

Peter Schattner

Center for Biomolecular Science and Engineering

University of California, Santa Cruz

1156 High Street, Santa Cruz, CA 95065

schattner@soe.ucsc.edu

Abstract:

Integrated genome databases - such as the UCSC, Ensembl and NCBI MapViewer databases – and their associated data querying and visualization interfaces (e.g. the genome browsers) have transformed the way that molecular biologists, geneticists and bioinformaticists analyze genomic data. Nevertheless, because of the complexity of these tools, many researchers take advantage of only a fraction of their capabilities. In this tutorial, using examples from medical genetics and alternative splicing, I describe some of the biological questions that can be addressed with these techniques. I also show why doing so typically is more effective than using alternative methods and indicate some of the resources available for learning more about the advanced capabilities of these powerful tools.

Keywords and key phrases:

genomes; genome databases; genome browsers; datamining; UCSC Genome Browser; Ensembl; Mapviewer; GMOD; GBrowse; Galaxy

Introduction

Genome browsers including those from UCSC [1], Ensembl [2] and NCBI [3] – have greatly eased the task of analyzing and correlating the large amounts of data associated with genomic "regions of interest", such as disease-associated polymorphisms, transcribed regions of unknown function [4] or highly conserved genomic regions located far away from any known gene [5]. Prior to the advent of the genome browsers, retrieving the experimental data available about a genomic region required accessing multiple databases, each with its own user interface and data format. Then one often had to develop custom tools for integrating the data that had been obtained from these different sources. Moreover, some of the most useful data, such as multi-species sequence alignment and conservation data were almost completely unavailable. In contrast, with a genome browser, it is easy to obtain a unified picture of a genomic region, integrating information that was originally available only in multiple, disparate databases.

In addition to offering genome browsing, the UCSC and Ensembl systems provide tools that enable users to directly query the databases that underlie their genome browsers. These tools include application programming interfaces (APIs) that facilitate the coding of computer programs to query the genome databases, as well as web based tools allowing genome database querying by researchers with little or no programming experience. The objective of the present work is to illustrate some of these resources, showing how they can be used to address realistic biological questions. It is not our intention to describe in detail the techniques needed to use these tools. Doing so is not

feasible within the scope of a brief introductory tutorial. Rather we seek to present a flavor of the capabilities of these tools and to point the reader toward the on-line and print resources that explain how to master these tools in detail.

Genome Browsing

In a recent publication, Morrow et al [6] identified numerous regions in the human genome that are associated with autism when they occur as homozygous deletions. Although each of these deletions is extremely rare and accounts for only a small fraction of the cases of autism, they are potentially important because they may lead to the identification of other, possibly more common, autism-associated, genetic variations and because they may provide clues as to the molecular pathways involved in the etiology of the disease.

However, it is not trivial to identify the specific sequence features within a deletion – which may be over a megabase in extent –that are the causal factors of a phenotype such as autism. Often several genes are deleted or truncated and one needs to identify which gene underlies the phenotype. Moreover, it is possible that none of the deleted genes are responsible and that, instead, the phenotype is the result of the deletion of a distal control region of a non-deleted gene. To address these issues, one typically considers multiple biological questions relating to the deleted region, such as:

* Are there known SNPs in the deleted region that could be investigated for correlation with increased prevalence of autism-like phenotypes?

* Are any of the deleted or nearby genes annotated as having central nervous system (CNS) related function?

* Do any of the deleted or nearby genes exhibit expression patterns specific to the CNS?

* Are any homologs of the deleted genes known in mouse or other model organisms, which could be exploited in animal studies of the biological functions of the deleted region?

* Are there known transcription factor binding sites (TFBS) or other regulatory regions in the deleted region?

* Has the deleted region been associated with any disease phenotype in any genome-wide association (GWA) studies?

* What subregions of the deleted region are highly conserved in other mammals, suggesting that they are of functional importance?

As a specific example, we will consider one of the regions identified in the work of Morrow et al, namely the deleted region near c3orf58 on chromosome 3. This region includes exactly one gene, c3orf58 – which Morrow et al renamed DIA1 for "deleted in autism 1" – as well as some surrounding non-coding sequence. In order to investigate this region, we first point our web browser to a genome-browser web site (we will use the Ensembl and UCSC Genome Browser sites for this example) and select the appropriate species and genome assembly. Then we determine the coordinates for our region of interest, for example by inserting c3orf58 in the "position or search" field of the genome browser display.

Finally we need to select and configure the annotation tracks corresponding to the types of data that we need, since the default browser displays may not include the annotation data for the questions we want to address. Specifically, on the UCSC Browser, we will select the Refseq, RNA-gene, human and non-human mRNA, SNP, genome-association, transcription-factor and miRNA binding site, regulatory region, gene expression, mammalian conservation and mouse alignment tracks. We will also add a UCSC "custom track" (described in more detail below) indicating the region of the deletion identified by Morrow et al. The track-selection procedure on Ensembl is similar, where we select the Ensembl transcript, SNP, conservation, regulatory region and mouse alignment tracks. In addition, in order to obtain the displays shown in Figures 1 and 2 we will need to reconfigure some of the track display options. So, for example in the UCSC display, we will configure the non-human mRNA track to only display mouse data and configure the gene expression track so that it groups together expression data from related cell types.

The techniques for selecting and configuring tracks are straightforward. However, because of the large number of configuration options, mastering the different display choices takes a little time to get used to. To aid the new user in navigating among the possible display configurations, documentation and tutorial information is provided on the Ensembl and UCSC websites. In addition, examples of how to navigate among the various browser configuration options are presented in detail in the on line tutorials provided by OpenHelix (http://www.openhelix.com/) as well as in Chapters 2 and 3 of [7]. Once we have made the necessary track selection and configuration choices, we can

submit our request to the browser, which responds to our query with the displays shown in Figures 1 and 2 for the UCSC and Ensembl Genome Browsers, respectively.

DIA1 in the UCSC Genome Browser

We can now answer several of our questions about DIA1 by simple inspection of the UCSC Browser display in Figure 1. For example, from the Conservation track in Figure 1, we see that in addition to well-conserved coding exons, DIA1 has highly conserved regions within one of its introns as well as in its 3' untranslated region (UTR). From the "TFBS conserved" and "TS miRNA sites" tracks, we see that the region includes numerous conserved transcription factor binding sites [8] and motifs that have been predicted to be miRNA target sites [9]. We can use the mouse mRNAs in the nonhuman mRNA track, as well as the "mouse chained alignment" track, to identify potential regions of homology to DIA1 in the mouse genome that might be appropriate for in an experimental study.

The (empty) "GAD View" genome-association track [10] indicates that the region has not been previously associated with any known disease phenotypes. The SNP track indicates the locations within the region of previously identified SNPs, which have been entered in the DbSNP database [11]. In addition, the color coding of the SNP track indicates that at least one SNP has been previously detected in DIA1's coding region.

The "GNF Expression Atlas" track displays tissue-specific, mRNA expression data developed by the Genomics Institute of the Novartis Research Foundation [12]. Somewhat surprisingly, the dark green color of the "brain" subtrack of GNF track

indicates that DIA1 was found to have somewhat lower expression in the brain than in other tissues. Additional evidence regarding the tissue specific expression of DIA1 can be inferred from the expression data of the mRNAs annotated on in the "Human mRNAs from Genbank" track (as well as from the Human EST track, which is not shown in Figure 1 to save space).

We can learn more regarding possible biological functions of DIA1 by using the auxiliary Proteome Browser and Gene Sorter Tools of the UCSC Genome Browser. In particular, the Proteome Browser [13] annotates predicted properties of the protein(s) derived from the DIA1 gene. In contrast, the Gene Sorter Tool [14] identifies genes (possibly including ones with known functions) which are in some ways "similar" to DIA1, in the sense of having similar amino-acid sequence, PFAM domains or expression patterns, or by being close to one another in a protein-interaction network.

DIA1 in the Ensembl Browser

We can obtain similar information about the DIA1 region from the Ensembl Browser (see Figure 2). Looking at Figure 2, we see Conservation, EST, SNP and mMus-blastz tracks with annotations similar to those we found in the UCSC Browser. (The name "Mmus-blastz refers to the BLASTZ genomic pairwise-alignment program [15] used by both UCSC and Ensembl.) In the Conservation track, we again see the highly conserved intronic and 3'UTR subregions. Since Ensembl and UCSC use different multi-species alignment and conservation algorithms (Ensembl uses the PECAN (http://www.ebi.ac.uk/~bjp/pecan/) alignment tool and the GERP [16] sequence conservation algorithm, whereas UCSC uses the multiz program [17] for sequence

alignment and the phastCons program [18] to estimate sequence conservation), seeing the same regions annotated in both browsers increases our confidence that the observed conservation is not dependent on the specific alignment or conservation algorithm.

Although Ensembl's annotations are similar to UCSC's, Ensembl's track formats and user interface are somewhat different from UCSC's. In particular, Ensembl uses some thirty different data "Views" to display its data, with each View optimized for a specific type of annotation. Ensembl's display views include chromosome views, alignment views, transcript views, SNP views and many more. In contrast, the UCSC interface is configured to rely on a single view (shown in Figure 1) for most data annotations.

As a result of the different strategies for data presentation, navigating through the data may sometimes be simpler in one system than in the other. For example, Figure 2 illustrates two appealing features of the Ensembl interface. First, switching to the homologous region in the mouse genome from the human DIA1 region is particularly easy. We just click on the "Mmus blastz" track and then select the option to jump to the homologous region in the mouse genome (see Figure 2). Second, Ensembl displays four different levels of genomic resolution simultaneously. Consequently we can see a sequence feature together with its genomic context. For example, the exon-intron structure of the DIA1 gene is shown in Ensembl's "Detailed view" (Figure 2, third section from the top), while the gene is displayed in its genomic context as the open red rectangle in the "Overview" component of the display (Figure 2, second section from the top). The

reader is again referred to the Browser's on line documentation as well as to the detailed tutorials (http://www.openhelix.com/ and [7]) for step-by step descriptions for navigating among the multiple display modes and options available within the Ensembl Browser.

MapViewer and other Genome Browsers

Some genome sequences and annotations are currently only available in either Ensembl or the UCSC Genome Browser. Consequently, one should check the other website if the annotation one needs is not found in the browser one tried initially. Moreover some genomes and annotations are not available in either the UCSC or Ensembl systems, but are available in other genome databases. For example, NCBI's MapViewer Genome Browser has annotations for over sixty plant and fungal genomes most of which are not included in Ensembl or UCSC. In addition, MapViewer is quite useful for applications involving comparisons of genomic maps, or analyses that require tight integration with other NCBI tools. On the other hand, MapViewer does not currently support multispecies sequence alignments, nucleotide-level resolution, custom tracks or genomic batch querying, as described in the following section. As a result, MapViewer is less suitable for the type of genomic datamining described here. Other genome databases that can be helpful if one needs data not found in the Ensembl or UCSC databases include the Gramene Database [19], for comparative plant genomics, as as well as the single-organism genome databases, such as the Saccharomyces Genome Database [20], the Mouse Genome Database [21], Flybase [22] and Wormbase [23].

Genomic Batch Querying

In the previous example, we queried the genome databases about a single genomic region (i.e. the deleted region surrounding DIA1). However, the genome databases enable data analyses that are much more powerful than the querying of a single genomic region. In particular, many important biological questions can only be addressed by simultaneously querying multiple genomic regions or even entire genomes. We refer to such querying of multiple genomic regions as genomic "batch querying". For example, in the paper of Morrow et al [6], numerous deleted or otherwise modified regions were identified, in addition to the one surrounding DIA1. Using a genome browser to individually examine each of those regions would quickly become very tedious and time consuming. Instead one would like to be able to annotate and analyze all of these regions using a single query. To address such needs, the UCSC and Ensembl systems, as well as "third party" websites such as Galaxy [24] [25] and Taverna [26, 27], provide tools with which one can analyze multiple genomic regions with a single set of commands.

With batch querying one can not only more easily characterize multiple genomic regions, but one can answer biological questions that cannot be addressed with genome browsers at all. Many applications of such genomic batch querying can be envisioned – ranging from genome-wide searches for RNA-editing [28] to detection of transposon-mediated exon generation [29] to genomic screens for "nonsense mediated decay" [30]. For detailed descriptions of how to apply the UCSC and Ensembl genome databases to these biological questions as well as to numerous others, the reader is referred to reference [7].

Here we will illustrate this approach with an example involving the detection of "tandem-site" or "NAGNAG" alternative splicing events[31].

NAGNAG alternative splicing may occur when a preMRNA transcript includes the subsequence "NAGNAG" at one of its acceptor splice sites (in this context, "N" refers to any one of the four ribonucleotides: A, C, G or U) [31]. Such transcripts may produce two different spliced mRNAs, differing in length by exactly three nucleotides. This situation is depicted schematically in Figure 3. If the splice site is in the mRNA's coding sequence, the translated proteins differ by exactly one amino acid. It is still unknown to what extent these small transcript variations are a way for the cell to "fine tune" protein structure by adding or deleting a single amino acid [32] or are simply a form of splicing "noise" with no biological function [33].

Here, we will not address the possible functions of NAGNAG splicing, but rather consider the question of simply how to screen a genome for putative cases of NAGNAG alternative splicing. We can search for such examples, by looking for pairs of transcripts (mRNAs or ESTs) for which an exon of one transcript is exactly 3 nucleotides (nt) longer or shorter at its 5' end than the overlapping exon of the other transcript. If the sequence surrounding such a splice site is NAGNAG, then the transcripts are most likely the result of NAGNAG alternative splicing. Specifically, we need to:

1. Extract all exons of all mRNAs from the genome database

2. Extract all exons of all ESTs from the genome database. (Note that there is nothing essential here about comparing mRNAs with ESTs. We could instead compare mRNAs with mRNAs for a test with higher specificity, or ESTs with ESTs for a test with higher sensitivity, since there are more EST sequences available, but EST sequences are often incomplete and have more sequencing errors.)

3. Pair each mRNA exon with each same-strand EST exon with which it overlaps and select the pairs for which the mRNA exon is exactly three nt longer or shorter at the exon's 5'-end.

4. For each such exon, retrieve the sequence surrounding the splice-site.

5. Keep only those exon pairs for which the splice-site sequence matches NAGNAG.

The data extraction necessary for steps 1, 2, and 4 can be directly carried out with the UCSC's Table Browser[34] or Ensembl's Biomart Tool [35]. However, performing the data set filtering described in steps 3 and 5 requires either writing a computer program or using a data-analysis toolset such as Galaxy[24].

Galaxy

Galaxy is a suite of data analysis tools for handling genomic sequences and annotations that have been downloaded from the UCSC, Ensembl or other genome databases. These tools include data converters (e.g. MAF to FASTA conversion) and data manipulation tools such as data "joining" and "filtering" tools, as well as a some widely used bioinformatics data-analysis program suites, such as EMBOSS[36] and HyPhy [37].

Figure 4 shows a screenshot of a Galaxy "workflow" implementing a search for NAGNAG alternative splicing sites. In Figure 4, the "Join" tool implements the initial pairing of mRNA exons with overlapping EST exons and the first "Filter" tool selects only those exon pairs where both exons are on the same strand. The first "Compute" tool and the second " Filter" tool are used to select those transcript pairs where the mRNA exon is three nt longer at its 5'-end. The subsequent "Compute" and "Cut" tools are used to specify the region around the splice site for which one needs to obtain sequence data. The sequence data is then retrieved with the "Extract Genomic DNA" tool, and subsequently reformatted with the "FASTA-to-Tabular" tool, so that it is in a format suitable for further analysis with Galaxy. Finally, the "Select" tool extracts all exon pairs for which the sequence surrounding the splice site is of the form NAGNAG. Figure 5 shows one example of a transcript pair identified by this screen in the UCSC Genome Browser.

We should note that we have glossed over some important details that must be addressed for a practical implementation on Galaxy. First, we need slightly different workflows for cases where the mRNA exon is three nt shorter than EST exon rather than three nt longer, as well as for searches for positive and negative strand NAGNAG transcripts. This latter issue is not entirely trivial since negative strand transcripts are stored in the UCSC database in "strand coordinates", which require an additional coordinate conversion step (see Appendices 1 and 2 of [7] for further discussion of strand coordinates in the UCSC system). Next, we would need to remove duplications arising when multiple ESTs overlap the same mRNA splice site. Last but not least, we need to address the fact that

EST tables are very large (e.g. the EST tables in the UCSC Human Genome Database have millions of records). Consequently, transferring an entire EST table to Galaxy is extremely slow at best and may fail altogether, as a result of system time-out errors. As a result, when querying large genome database tables, one typically first performs table intersections directly on the UCSC Table Browser or Ensembl Biomart so that only the intersected (and consequently, smaller) data sets need to be loaded onto Galaxy(see [7] chapter 5). For data analyses in which such initial table intersection is not possible, it may be necessary to perform the analysis multiple times on smaller data sets, e.g performing the analysis separately for each chromosome.

Taverna

Galaxy is not the only computational platform designed for the non-programmer biologist to analyze genomic data. The Taverna toolkit [26, 27] is also intended to assist biologists in executing genome-scale data-analysis. However, Taverna uses a very different approach. Whereas Galaxy contains a suite of prepackaged tools installed on a single server, Taverna does not explicitly include any data-processing or computational tools at all. Instead, Taverna provides a graphical user interface (similar to that used by the Galaxy Workflow Tool) for building a workflow or pipeline consisting of any combination of data analysis programs available as "web services" [38]. Since many widely used genomic data analysis tools such as BLAST, ClustalW, Repeatmasker and EMBOSS are currently available as web services, one can create flexible and varied data-analysis pipelines with Taverna, often without needing to do any computer programming. Moreover, since all of the data-analysis programs are invoked over the internet via Web Services protocols, which are handled by Taverna, the user needs neither to install any

programs locally (other than Taverna itself) nor to be concerned about the protocols required for remote program execution.

However, Taverna – at least in its current implementation – also has significant limitations. First, Taverna doesn't include data joining, filtering or reformatting tools, such as those provided by Galaxy. Instead such tools need to be provided by the user. Although these tasks are simple conceptually, they are tedious to write, and must be implemented carefully, if they are to be performed in an error-free manner. In addition, if any of the web servers in one's Taverna pipeline are "down" or overloaded, one's entire workflow will stop. Similarly, if any program in one's pipeline has been modified or upgraded by its host system, the results of one's pipeline analysis may change. Now, to be sure, similar issues will arise with Galaxy if the Galaxy system is down or is modified. However, with Galaxy, one is dealing with only a single server. Consequently, if one's workflow suddenly fails or produces a different answer, there is only one system to consider in determining what has changed. Moreover, one can install a local mirror of the entire Galaxy server without too much difficulty. In this case, one will have complete control of any changes in the data analysis system. In contrast, with a Taverna pipeline, it may be difficult to identify which server in the pipeline is down or has changed if one's data-analysis results change.

Programmed genome database querying

Interactive web-based tool sets such as Galaxy and Taverna have made it possible to execute analyses of genomic data without needing to write any computer code. Although this capability is often very attractive, as one's biological analyses become more complex,

the lack of a conventional programming framework for them increasingly becomes a mixed blessing.

First, some components of the UCSC and Ensembl databases can currently *only* be accessed via direct computer querying. For example, data that has not been mirrored by Ensembl to its Biomart database (e.g. Repeatmasker data) is not accessible via Galaxy. Similarly, some UCSC data, such as Genbank mRNA and EST sequence data, can only be accessed from the UCSC databases by computer querying.

In addition, computer languages, such as C or Perl, have many powerful features, including subroutines, command line arguments and complex logical branching operations to facilitate creating flexible analysis workflows. With these features it is possible to write a single program that can handle multiple types of data having different formats or modified data-processing requirements. In contrast, implementing such workflow flexibility in a prepackaged environment such as Galaxy is typically more challenging. Consequently, both Ensembl and UCSC provide API's (in Perl for Ensembl and in C for the UCSC database) that greatly facilitate the programmatic querying of their underlying databases. Moreover, public mirrors of the entire Ensembl database system (located at ensembldb.ensembl.org) and a large part of the UCSC databases (genome-mysql.cse.ucsc.edu) can be accessed programmatically over the internet, often eliminating the need to mirror the databases locally.

I will not describe programmed querying of the UCSC and Ensembl databases further here, as I have already written about this topic in detail elsewhere (reference [7], chapters 7 - 10). A brief overview of these methods is also available [39]. Suffice it to say that the experienced Perl or C programmer may sometimes find direct programmed querying using the Ensembl or UCSC APIs more straightforward or flexible than using a packaged tool kit such as Galaxy.

Analyzing custom data

Genomic data mining often involves combining newly acquired data from a local experiment or sequencing project with publicly available data located in the genome databases. In some cases, integrating local and public data may be as simple as adding an annotation track, containing the locations of newly identified genes or other genomic features, to one of the public genome browsers. With the UCSC and Ensembl Databases such data integration is particularly simple, as both systems provide tools for creating "custom tracks" for this purpose. Data for custom tracks can be uploaded to the UCSC or Ensembl Database and viewed alongside all the conventional browser tracks by the user (and generally only by the user in order to ensure data privacy and security).

For example, if we had a list of the coordinates of the autism-associated, genomic deletions identified by Morrow et al. [6], we could make a custom track of these regions. Such a custom track would consist of a single header line plus one line for each region to be annotated; in particular, a custom track that annotated just the single autism region at DIA1, in UCSC format, would be:

```
track name="autism deletions" description="Morrow et al autism deletions"
chr3    145091098    145977477    DIA1_deletion    0    +
```

(The custom track format for Ensembl is similar.) Once the custom track has been uploaded to the UCSC website by selecting the "add custom tracks" button in the browser interface, it would appear in the UCSC Browser as the "Morrow et al autism deletions" track shown in Figure 1. Now, in the Browser display shown in Figure 1, this custom track is not particularly informative, since we have "zoomed in" the display to be completely within one of the deleted regions. However, if we zoomed out to a larger genomic field of view, the custom track could be helpful in visualizing what other genomic features are in the vicinity of the deletions. More importantly, if we were to upload a custom track that included all of the deletions onto Galaxy or the UCSC Table Browser, we could ask global questions regarding the properties of the entire set of autism-associated deletions. In this way, we could, for example, identify all the nonsynonymous SNPs that are located within one of the deletions, or determine whether the GC content of these regions varied from that of the overall genome, or search for deletions that are near regions with high recombination rates. (Note that although Ensembl does not directly support custom tracks within its Biomart batch-querying tool, we could also compare our custom data with Ensembl annotations, by uploading both our custom data and the desired Ensembl annotations to Galaxy.)

For some types of locally generated data, simply adding a custom track to an existing genome database is insufficient. An obvious, but important, example is the assembly and

annotation of a previously unsequenced genome. Indeed, in this case one needs to create an entirely new database and browser for the new genome. Moreover, considering the accelerating pace at which genome sequencing projects are being carried out, this sort of application is becoming increasingly common. Needless to say, creating a genome database and browser from scratch for a newly sequenced genome is not trivial. To facilitate this task, the GMOD (Generic Model Organism) Project has developed a suite of free open source software tools [40]. These tools include software to build and access the database, as well as a genome browser, called GBrowse for displaying the data[41]. Now a genome database system implemented with GMOD tools will not be as full-featured as the UCSC or Ensembl systems. In particular, GMOD systems are designed principally to be single-organism databases and offer little support for multi-species annotations such as genomic alignment or conservation tracks. However, in return, it is far easier to implement a GMOD database than to clone the UCSC or Ensembl architecture, and the GMOD architecture does provide most of the browser and querying features one would expect in a modern genome database. In fact, several of the widely used model organism databases such as FlyBase [22], WormBase [23] and the Mouse Genome Database [21] were created using GMOD tools.

Climbing the learning curve

The reader should be, by now, convinced that genome browsers and their associated genome databases and support tools can significantly simply the tasks of integrating and analyzing genomic data. Indeed, the reader who is not yet convinced is encouraged to attempt the analyses of the DIA1 region and the identification of potential NAGNAG alternative splicing sites described above *without* the use of a genome database.

That said, we should emphasize again that there are definite learning curves associated with the genome browsers and the genome databases. Although using the UCSC, Ensembl or MapViewer Browsers in their basic manner is easy and intuitive, knowing how to find and configure the correct "tracks" or "views" or "maps" which are needed to address one's specific query – or even to determine whether the data one wants is available in the browser at all – typically requires a certain amount of experience. And if one wants to use batch-querying tools, such as Galaxy or Taverna, or the programmer APIs provided by Ensembl and UCSC, the necessary learning curves are steeper.

Fortunately, all of the resources described here (Table 1) provide detailed on line documentation and, typically, tutorial support as well. Of particular utility for the genome browser novice are the on line tutorials from Open Helix. For learning how to use the Galaxy toolset, the on line videos available at the Galaxy web site are highly recommended. In addition, a book is now available that describes how to use all of the resources covered here [7].

In summary, hopefully I have persuaded the reader that genome browsers and integrated genome databases, such as those found at Ensembl and UCSC, provide comprehensive sources of genomic data in standardized formats, making data acquisition and subsequent analysis substantially simpler than using multiple specialized databases. Further, I have presented examples of how emerging web-based tools such as Galaxy can enable biologists, even without programming skills, to perform quite sophisticated data analyses

on this genomic data. Finally, I have noted that, although a certain level of effort is required to master these tools, the recent emergence of detailed, on line and print references and tutorials can ease this learning task, and, moreover, that one's effort in mastering these tools will be amply repaid by one's enhanced ability to integrate and analyze the ever-growing collection of genomic data.

Acknowledgements

I am grateful to Deanna Church, Hiram Clawson, Mark Diekhans, Xose Fernandez, Jim Kent, Anton Nekrutenko and Lincoln Stein for numerous helpful discussions on the topics presented here. I would also like to thank Winston Hide for encouraging me to write this tutorial.

Figure Captions:

Figure 1. DIA1 on the UCSC Genome Browser. Display of the region surrounding the DIA1 / c3orf58 gene in the UCSC Browser, showing mRNA, SNP, expression, regulatory region and conservation annotations. A custom track indicating deleted regions in autism is also included in the display.

Figure 2. DIA1 on the Ensembl Genome Browser. View of the DIA1 / c3orf58 gene region in ContigView on the Ensembl Browser. The display includes views of the DIA1 genomic region at four distinct genomic resolutions. These are referred to in Ensembl as (starting from the top of the display) Chromosomal view, Overview, Detailed view and Basepair view. In this screen shot, the computer mouse has been placed over the "Mmus blastz" track, resulting in the display of the coordinates of the homologous region in the

mouse genome in the lower right hand corner of the figure. Note that Ensembl indicates this region as chromosome 9:94429692-94430213, whereas in Figure 1, the UCSC mouse chain annotation specifies chromosome 9:90208000. This is not a disagreement since, by convention, UCSC displays the 5' coordinate of the entire syntenic region, whereas Ensembl displays the coordinates of the homologous subregion, as limited to the current display window.

Figure 3. NAGNAG alternative splicing. Schematic cartoon of NAGNAG alternative splicing. Since there are two adjacent splice-acceptor sequence motifs, two distinct alternatively spliced transcripts are possible.

Figure 4. Galaxy workflow for NAGNAG detection. Galaxy workflow diagram showing the steps required to identify potential NAGNAG alternative splicing events from mRNA and/or EST data. See the text for description of the various data processing blocks.

Figure 5. An example of a possible NAGNAG alternative splice site in the human genome identified by the screen for NAGNAG sites described in the text.

Table 1. Principal internet resources for genome browsers and databases. A listing of web addresses for the extensive tutorial and documentary material associated with each of these resources can be found in Appendix 7 of [7].

References

1. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update*. Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.
2. Flicek, P., et al., *Ensembl 2008*. Nucleic Acids Res, 2008. **36**(Database issue): p. D707-14.
3. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 2008. **36**(Database issue): p. D13-21.
4. Kapranov, P., et al., *Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays*. Genome Res, 2005. **15**(7): p. 987-97.
5. Bejerano, G., et al., *Ultraconserved elements in the human genome*. Science, 2004. **304**(5675): p. 1321-5.
6. Morrow, E.M., et al., *Identifying autism loci and genes by tracing recent shared ancestry*. Science, 2008. **321**(5886): p. 218-23.
7. Schattner, P., *Genomes, Browsers and Databases*. 2008: Cambridge University Press.
8. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes*. Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
9. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*. Cell, 2005. **120**(1): p. 15-20.

10. Becker, K.G., et al., *The genetic association database*. Nat Genet, 2004. **36**(5): p. 431-2.

11. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.

12. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes*. Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

13. Hsu, F., et al., *The UCSC Proteome Browser*. Nucleic Acids Res, 2005. **33**(Database issue): p. D454-8.

14. Kent, W.J., et al., *Exploring relationships and mining data with the UCSC Gene Sorter*. Genome Res, 2005. **15**(5): p. 737-41.

15. Schwartz, S., et al., *Human-mouse alignments with BLASTZ*. Genome Res, 2003. **13**(1): p. 103-7.

16. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence*. Genome Res, 2005. **15**(7): p. 901-13.

17. Miller, W., et al., *28-way vertebrate alignment and conservation track in the UCSC Genome Browser*. Genome Res, 2007. **17**(12): p. 1797-808.

18. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res, 2005. **15**(8): p. 1034-50.

19. Liang, C., et al., *Gramene: a growing plant comparative genomics resource*. Nucleic Acids Res, 2008. **36**(Database issue): p. D947-53.

20. Christie, K.R., et al., *Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related*

sequences from other organisms. Nucleic Acids Res, 2004. **32**(Database issue): p. D311-4.

21. Eppig, J.T., et al., *The mouse genome database (MGD): new features facilitating a model system*. Nucleic Acids Res, 2007. **35**(Database issue): p. D630-7.
22. Crosby, M.A., et al., *FlyBase: genomes by the dozen*. Nucleic Acids Res, 2007. **35**(Database issue): p. D486-91.
23. Schwarz, E.M., et al., *WormBase: better software, richer content*. Nucleic Acids Res, 2006. **34**(Database issue): p. D475-8.
24. Blankenberg, D., et al., *A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly*. Genome Res, 2007. **17**(6): p. 960-4.
25. Taylor, J., et al., *Using galaxy to perform large-scale interactive data analyses.* Curr Protoc Bioinformatics, 2007. **Chapter 10**: p. Unit 10 5.
26. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows*. Bioinformatics, 2004. **20**(17): p. 3045-54.
27. Hull, D., et al., *Taverna: a tool for building and running workflows of services*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W729-32.
28. Levanon, E.Y., et al., *Systematic identification of abundant A-to-I editing sites in the human transcriptome*. Nat Biotechnol, 2004. **22**(8): p. 1001-5.
29. Lev-Maor, G., et al., *The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons*. Science, 2003. **300**(5623): p. 1288-91.

30. Green, R.E., et al., *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes*. Bioinformatics, 2003. **19 Suppl 1**: p. i118-21.

31. Hiller, M., et al., *Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity*. Nat Genet, 2004. **36**(12): p. 1255-7.

32. Hiller, M., et al., *Alternative splicing at NAGNAG acceptors: simply noise or noise and more?* PLoS Genet, 2006. **2**(11): p. e207; author reply e208.

33. Chern, T.M., et al., *A simple physical model predicts small exon length variations*. PLoS Genet, 2006. **2**(4): p. e45.

34. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool*. Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.

35. Kasprzyk, A., et al., *EnsMart: a generic system for fast and flexible access to biological data*. Genome Res, 2004. **14**(1): p. 160-9.

36. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite*. Trends Genet, 2000. **16**(6): p. 276-7.

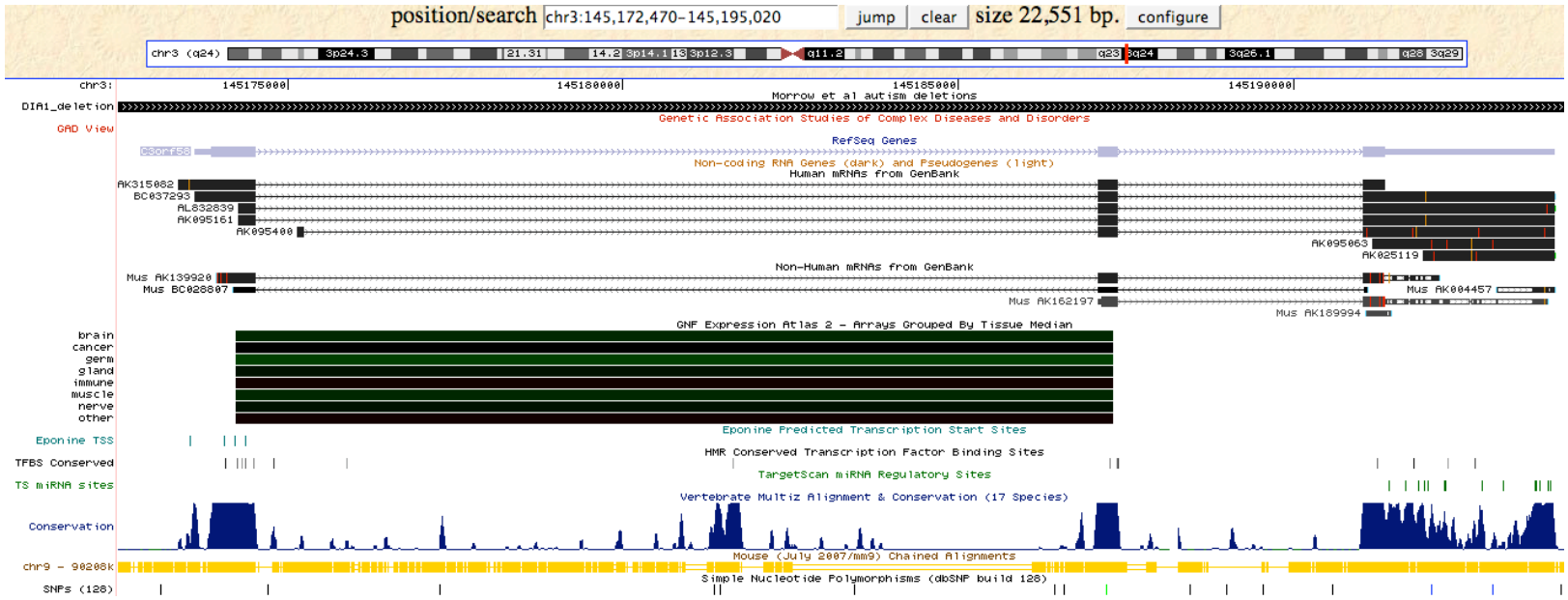
37. Pond, S.L., S.D. Frost, and S.V. Muse, *HyPhy: hypothesis testing using phylogenies*. Bioinformatics, 2005. **21**(5): p. 676-9.

38. Zimmerman, O., M. Tomlinson, and S. Peuser, *Perspectives on Web Services*. 2005: Springer.

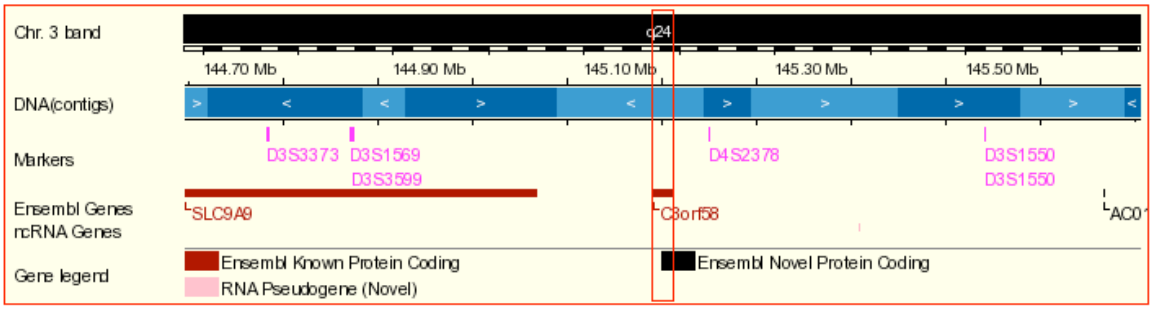
39. Schattner, P., *Automated querying of genome databases*. PLoS Comput Biol, 2007. **3**(1): p. e1.

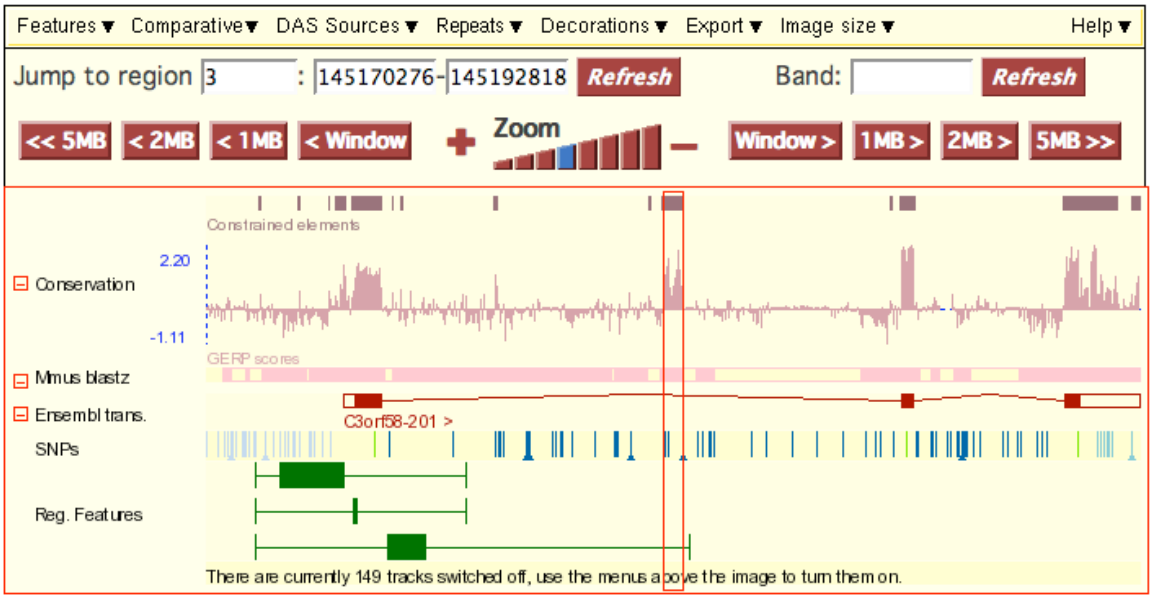
40. O'Connor, B.D., et al., *GMODWeb: a web framework for the Generic Model Organism Database*. Genome Biol, 2008. **9**(6): p. R102.

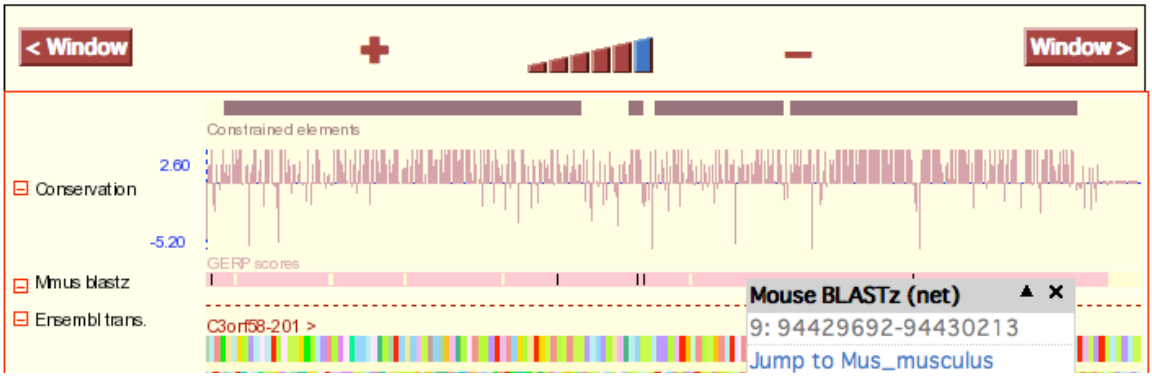
41. Stein, L.D., et al., *The generic genome browser: a building block for a model organism system database*. Genome Res, 2002. **12**(10): p. 1599-610.

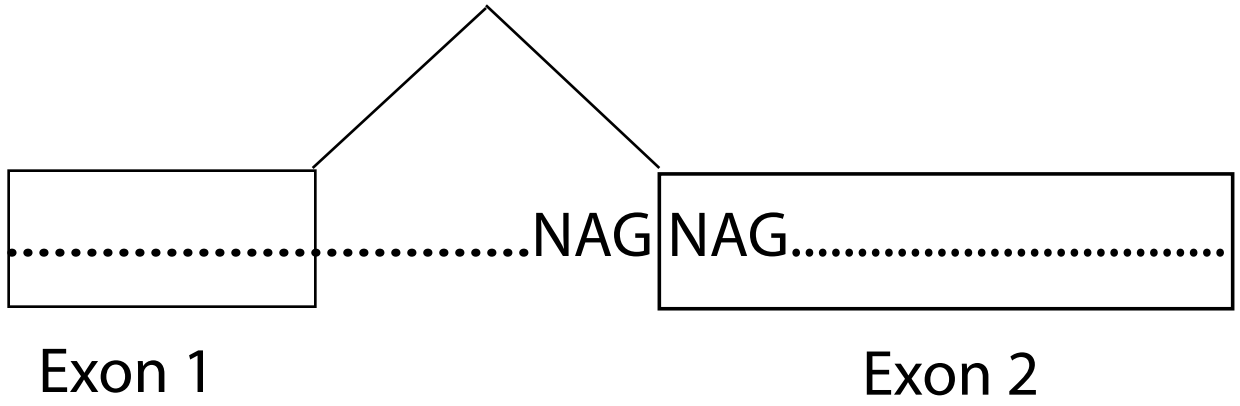

position/search
chr3:145,172,470-145,195,020
jump
clear
size 22,551 bp.
configure
chr3 (q24)
chr3:
145175000
145180000
145185000
145190000
Morrow et al autism deletions
DIA1_deletion
Genetic Association Studies of Complex Diseases and Disorders
GAD View
RefSeq Genes
C3orf58
Non-coding RNA Genes (dark) and Pseudogenes (light)
Human mRNAs from GenBank
AK315082
BC037293
AL832839
AK095161
AK095400
AK095063
AK025119
Non-Human mRNAs from GenBank
Mus AK139920
Mus BC028807
Mus AK004457
Mus AK162197
Mus AK189994
GNF Expression Atlas 2 - Arrays Grouped By Tissue Median
brain
cancer
germ
gland
immune
muscle
nerve
other
Eponine Predicted Transcription Start Sites
Eponine TSS
HMR Conserved Transcription Factor Binding Sites
TFBS Conserved
TargetScan miRNA Regulatory Sites
TS miRNA sites
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
Mouse (July 2007/mm9) Chained Alignments
chr9 - 90208k
Simple Nucleotide Polymorphisms (dbSNP build 128)
SNPs (128)

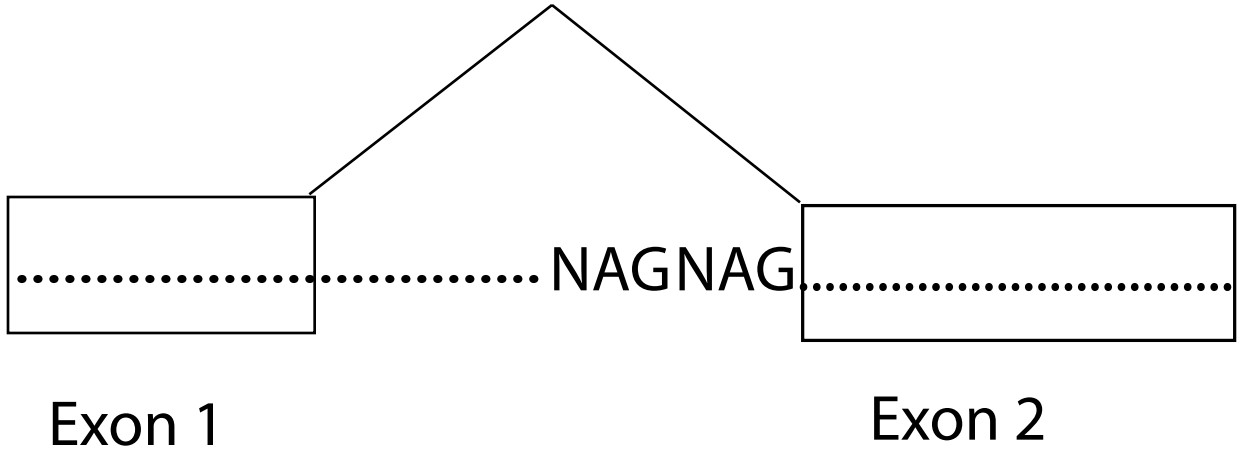

Chr. 3

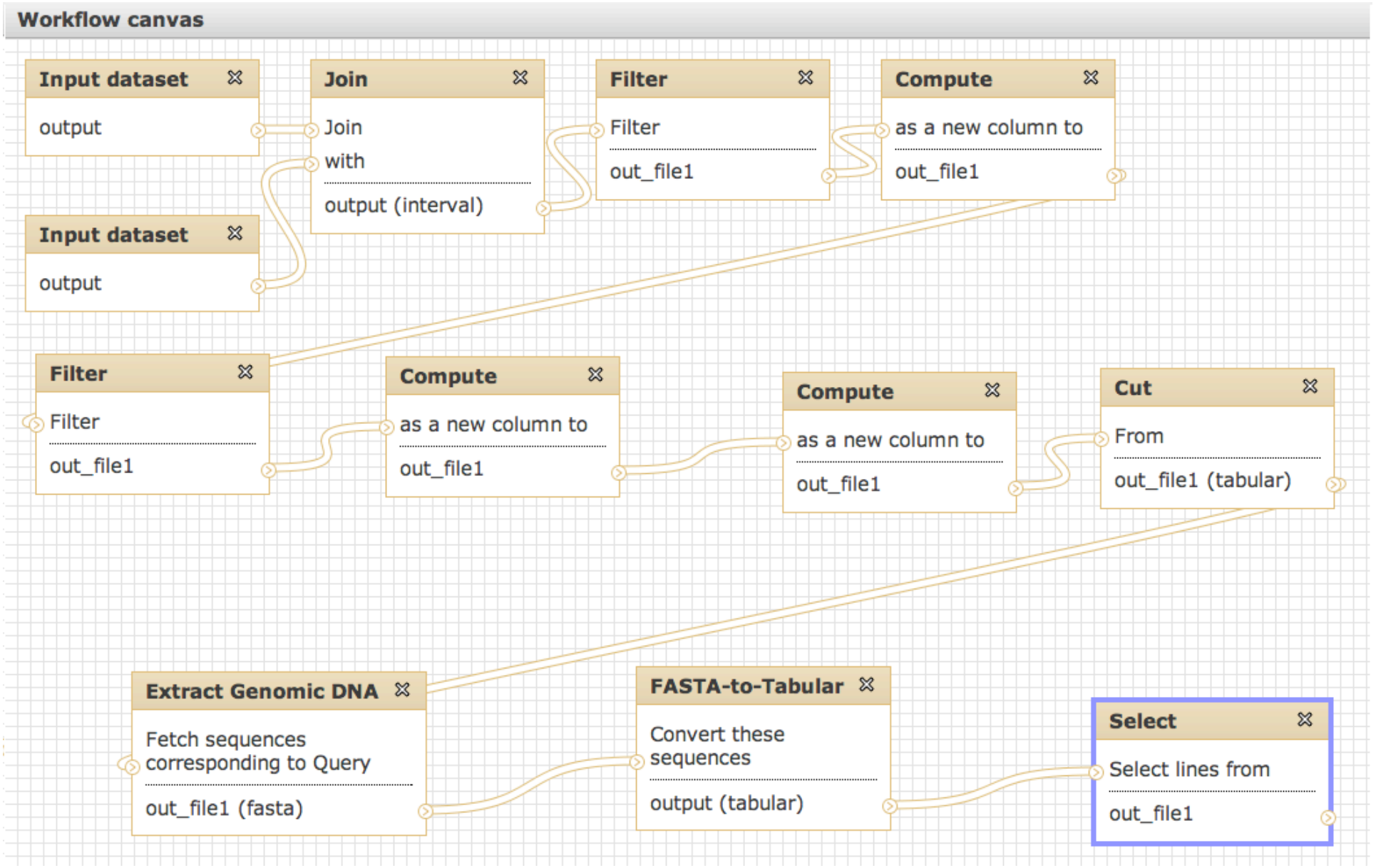

Chr. 3 band
q24
144.70 Mb
144.90 Mb
145.10 Mb
145.30 Mb
145.50 Mb
DNA(contigs)
Markers
D3S3373
D3S1569
D3S3599
D4S2378
D3S1550
D3S1550
Ensembl Genes
ncRNA Genes
SLC9A9
C3orf58
AC0
Gene legend
Ensembl Known Protein Coding
Ensembl Novel Protein Coding
RNA Pseudogene (Novel)


Features
Comparative
DAS Sources
Repeats
Decorations
Export
Image size
Help
Jump to region
3
145170276-145192818
Refresh
Band:
Refresh
<< 5MB
< 2MB
< 1MB
< Window
Zoom
Window >
1MB >
2MB >
5MB >>
Constrained elements
Conservation
2.20
-1.11
GERP scores
Mmus blastz
Ensembl trans.
C3orf58-201 >
SNPs
Reg. Features
There are currently 149 tracks switched off, use the menus above the image to turn them on.


< Window
Window >
Constrained elements
Conservation
2.60
-5.20
GERP scores
Mmus blastz
Ensembl trans.
C3orf58-201 >
Mouse BLASTz (net)
9: 94429692-94430213
Jump to Mus_musculus


NAG NAG
Exon 1
Exon 2


NAGNAG
Exon 1
Exon 2


Workflow canvas
Input dataset
output
Input dataset
output
Join
Join
with
output (interval)
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Cut
From
out_file1 (tabular)
Extract Genomic DNA
Fetch sequences corresponding to Query
out_file1 (fasta)
FASTA-to-Tabular
Convert these sequences
output (tabular)
Select
Select lines from
out_file1

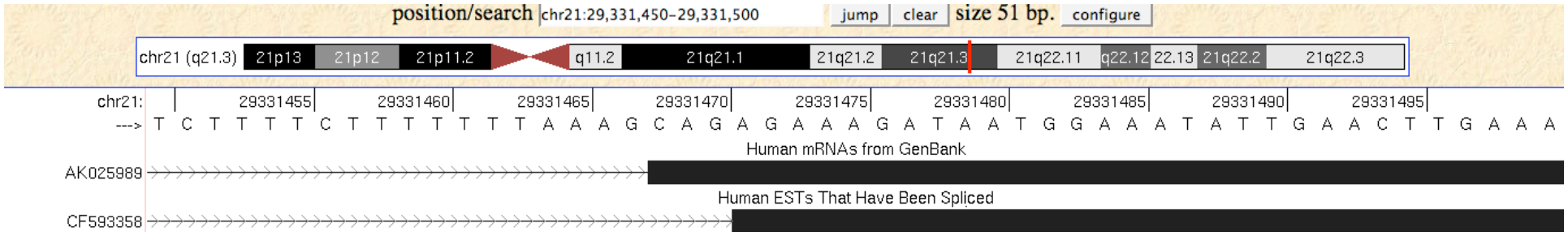

position/search
chr21:29,331,450-29,331,500
jump
clear
size 51 bp.
configure
chr21 (q21.3)
21p13
21p12
21p11.2
q11.2
21q21.1
21q21.2
21q21.3
21q22.11
q22.12
22.13
21q22.2
21q22.3
chr21:
29331455
29331460
29331465
29331470
29331475
29331480
29331485
29331490
29331495
---> T C T T T T C T T T T T T T A A A G C A G A G A A A G A T A A T G G A A A T A T T G A A C T T G A A A
Human mRNAs from GenBank
AK025989
Human ESTs That Have Been Spliced
CF593358

Table 1:

Resource	Web address	Description	Sponsoring Organizations
Open Helix	http://www.openhelix.com/tutorials.shtml	On-line tutorial material for all of the genome databases.	OpenHelix, LLC
UCSC Genome Browser	http://genome.ucsc.edu	Comprehensive, multi-species genome database providing genome browsing and batch querying.	Genome Bioinformatics Group, University of California, Santa Cruz
Ensembl Browser	http://www.ensembl.org	Comprehensive, multi-species genome database providing genome browsing and batch querying.	European Bioinformatics Institute (EBI) and the Sanger Center
NCBI MapViewer	http://www.ncbi.nlm.nih.gov/mapview	Multi-species genome browser focusing especially on genome mapping applications.	National Center for Biotechnology Information (NCBI)
Biomart	http://www.biomart.org/	Genome-database, batch-querying interface used by Ensembl and several single-genome databases.	Ontario Institute for Cancer Research and European Bioinformatics Institute
Galaxy	http://main.g2.bx.psu.edu	Integrated toolset for analyzing genome batch-querying data.	Center for Comparative Genomics and Bioinformatics. Penn State University
Taverna	http://taverna.sourceforge.net	Toolset for creating pipelines of bioinformatics analyses implemented via the Web services protocol	Open Middleware Infrastructure Institute, University of Southampton (OMII-UK)
GMOD	http://www.gmod.org	Repository of software tools for developing generic genome databases	A consortium of organizations operating as the Generic Model Organism Database project