Published online 25 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W615–W621***
doi:10.1093/nar/gks427

Savant Genome Browser 2: visualization and analysis for population-scale genomics

Marc Fiume[1], Eric J. M. Smith[1], Andrew Brook[1], Dario Strbenac[2], Brian Turner[3], Aziz M. Mezlini[4], Mark D. Robinson[5], Shoshana J. Wodak[2,6] and Michael Brudno[1,4,*]

[1]Department of Computer Science, University of Toronto, Ontario, Canada M5S 2E4, [2]Epigenetics Laboratory Cancer Research Program, Garvan Institute of Medical Research, Sydney, New South Wales 2011, Australia, [3]Molecular Structure and Function Program, [4]Centre for Computational Medicine, Hospital for Sick Children, Toronto, Ontario, Canada M5G 1L7, [5]Institute of Molecular Life Sciences, University of Zurich, 8057 Zürich, Switzerland and [6]Department of Biochemistry, University of Toronto, Ontario, Canada M5S 1A8

Received March 7, 2012; Revised April 23, 2012; Accepted April 24, 2012

ABSTRACT

High-throughput sequencing (HTS) technologies are providing an unprecedented capacity for data generation, and there is a corresponding need for efficient data exploration and analysis capabilities. Although most existing tools for HTS data analysis are developed for either automated (e.g. genotyping) or visualization (e.g. genome browsing) purposes, such tools are most powerful when combined. For example, integration of visualization and computation allows users to iteratively refine their analyses by updating computational parameters within the visual framework in real-time. Here we introduce the second version of the Savant Genome Browser, a standalone program for visual and computational analysis of HTS data. Savant substantially improves upon its predecessor and existing tools by introducing innovative visualization modes and navigation interfaces for several genomic datatypes, and synergizing visual and automated analyses in a way that is powerful yet easy even for non-expert users. We also present a number of plugins that were developed by the Savant Community, which demonstrate the power of integrating visual and automated analyses using Savant. The Savant Genome Browser is freely available (open source) at www.savantbrowser.com.

INTRODUCTION

High-throughput sequencing (HTS) technologies have revolutionized the speed and economy with which genomic information can be obtained. The unprecedented data-generating capacity of these technologies has stimulated their use throughout molecular biology: in genomics, in transcriptomics (e.g. RNA-seq) and proteomics (e.g. ChIP-seq). The large influx of raw sequencing data and resulting read alignment and genetic variant data has increased the need for efficient data storage, analysis and exploration. Although many data analysis steps can be automated through computational pipelines, a significant number of tasks still require human interpretation.

Visualization tools such as the UCSC Genome Browser (1) and IGV (2) can aid in the interpretation of large and complex data sets. While the former is hosted on a webserver with access to a large backend database, the latter is run on client computers, and can access both local data on the user's hard drive and remote data sets via the internet (if the data is stored in some standard format such as BAM or Tabix). Genome browsers have become integral parts of HTS analysis pipelines, where they are used to assess the reliability of computational predictions, validate findings in specific regions and guide refinement of automated tools. Despite the proliferation of programs that deal with HTS data, most have been developed for either automated (e.g. read mapping, genotyping) or visualization (e.g. genome browsing) purposes, but not both. Yet visual and automated approaches are most powerful when used together, such that users can seamlessly inspect and perform computation on their data, iteratively refining their analyses.

In this article we present the second version of the Savant Genome Browser (3). The original version of Savant, like the UCSC Genome Browser and IGV, enabled fast visualization and navigation of reference genomes and corresponding genomic data sets—such as HTS read alignments, gene annotations and other tracks. The new version advances upon its predecessor and existing genome browsers by introducing a number

*To whom correspondence should be addressed. Tel: +1 416 9782589; Fax: +1 416 9781455; Email: brudno@cs.toronto.edu; support@savantbrowser.com

of innovative visualizations and navigation interfaces, and allowing for seamless integration of diverse external data sets. We have also significantly expanded plugin functionality, and this article describes a number of analytic, visualization and datasource plugins. These plugins, developed by the core Savant Development Team and by the Savant User Community, help synergize visual and automated analyses of genomic data.

SAVANT FEATURE SUMMARY

Within this section we briefly summarize important Savant features, while in the following ones we describe in more detail major recent enhancements to visualizations and browser extensibility. A full set of features is described in the User Manual, available on the Savant website.

Formats

Savant supports a wide range of file formats, including the common standards for read alignment (BAM), genetic variants (VCF), interval (BED, GFF, Tabix) and continuous-valued (WIG, BigWIG, TDF) data. When necessary, Savant automatically indexes and compresses all these datatypes.

Datasources

In addition to working with local files, Savant supports the use of remote (through the internet) files and datasources. All remote resources are cached locally, to enable rapid visualization upon re-loading of a previously visited region. While many commonly used data sets (e.g. reference genomes, genes) are accessible through a public repository, tracks can also be quickly loaded directly from the UCSC Genome Database (4), without a need for manual download, as described below.

Navigation

Navigation to genomic regions of interest is assisted through textual search (e.g. seek to region by gene name), as well as through bookmarks that the user can add to and edit. User sessions can be saved for later use, or for sharing among users, ensuring that collaborators have identical views of the same data.

VISUALIZING HTS DATA SETS

The size of HTS and related genomic data sets challenges their interactive visualization: a single sequenced genome can yield billions of reads resulting in read alignment and variant prediction files that are each many gigabytes in size. Savant applies a multi-resolution visualization principle for each of these datatypes. For example, for read alignment tracks Savant seamlessly switches from showing individual reads at basepair-resolution to showing an alternate, coverage representation at lower resolution ranges (e.g. when viewing millions of basepairs). This helps maintain interactive yet informative views of the data at all resolutions.

Furthermore, within a given resolution alternative views of the data, or a subset of the data, may be required. For example, each read alignment record typically contains the predicted fragment sequence, several quality metrics and information about the location of the read's mate. It is useful to consider some but not all of these aspects, depending on the analytic task at hand. Savant includes several new visualization modes for HTS and genetic variant data sets, each designed to focus on specific aspects of the data that are relevant to common analytic tasks like assessing data quality, identifying genetic variants and discovering population trends.

Visualizations for SNP and small indel discovery

In the simplest terms the determination of whether a position is variable depends on the percentage of reads with differing nucleotides at this position. Instead of requiring the user to infer this percentage from raw sequenced reads, this can instead be visualized as per-nucleotide histograms in the SNP visualization mode. However HTS platforms also exhibit unique and reproducible biases that arise due to imperfect sequencing chemistries and/or library preparation. For example, strand-specific errors are common, and support from reads sequenced from both strands of DNA is often a requirement to separate true from false variants. In Savant, a user can choose to separate coverage profiles by strand within the Strand-SNP mode, making the identification of SNPs with support from both strands straightforward (Figure 1B) while also highlighting issues pertaining to strand-specific coverage biases.

Simultaneously just looking at percentages is insufficient, as the processes of sequencing and alignment are prone to error. In order to highlight reads and nucleotides most informative for variation detection Savant users can choose to shade whole reads or individual positions based on their mapping quality or base quality, respectively. This visualization mode makes the disentanglement of high and low confidence supporting data simpler, as shown in Figure 1A.

Visualizations for structural variant discovery

Detecting large-scale structural variants (insertions, deletions and inversions of multiple kilobases) using HTS reads requires emphasizing very different aspects of the data than the high-resolution coverage maps described above. Duplications or deletions in a sequenced genome manifest in increased or decreased coverage over the corresponding region in the reference (5), and breakpoints can be identified more precisely via clusters of read pairs that are discordant, that is they fall outside the expected distance and orientation constraints (6). When one views read alignment tracks over large regions in Savant, the browser automatically switches to the low-resolution coverage view, enabling the coarse identification of such variants. Alternatively, a user can choose to visualize read pairs as arcs, which are scaled vertically according to the distance between the component reads, and colored according to relative orientations and size discordance. A distinct representation is also used for paired

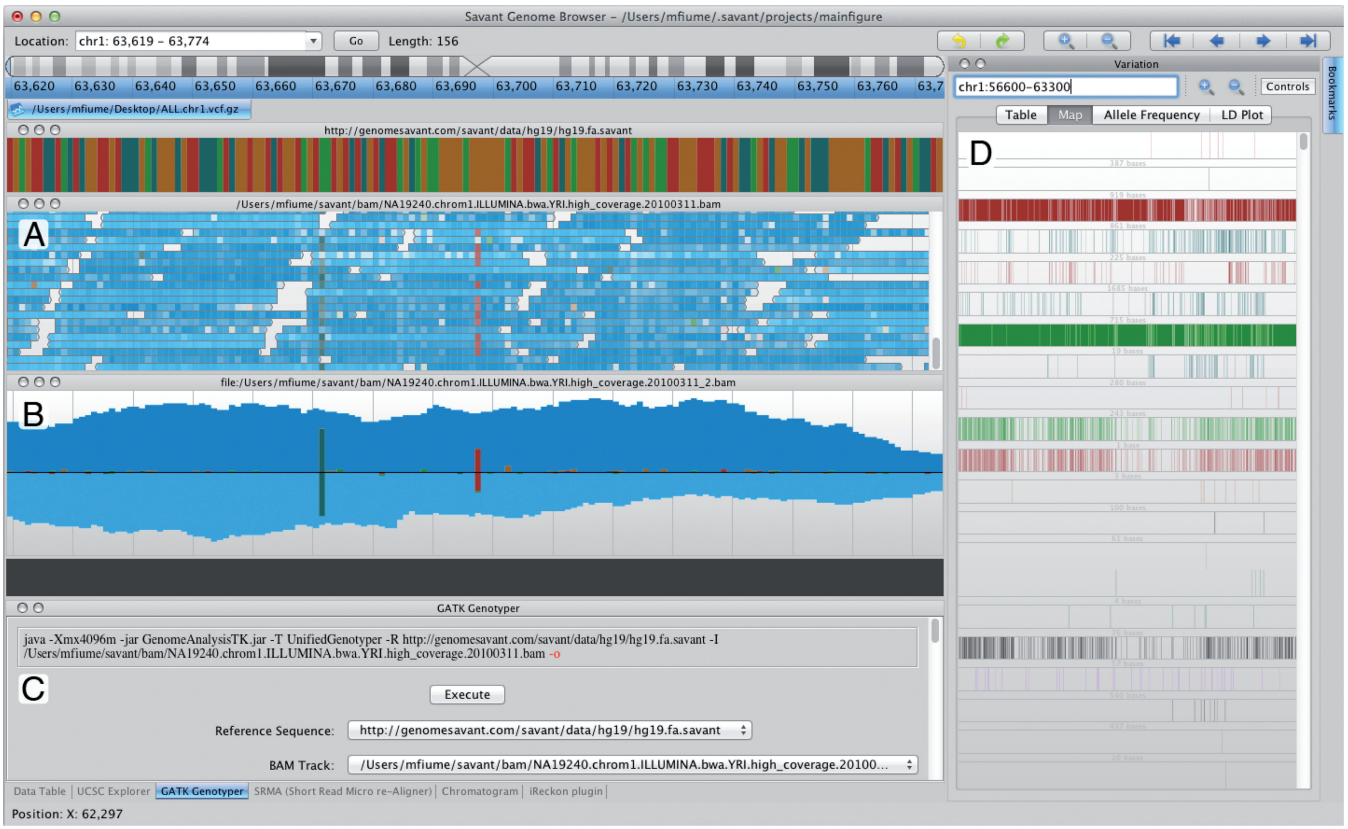

Figure 1. The Savant Genome Browser interface. (**A**) Read alignment track in Standard mode. Each position in a read has an intensity proportional to the Base Quality. Mismatches within reads are denoted by colors. (**B**) Read alignment track in Strand-SNP mode. This is the same data set as in A, but this mode shows coverage and allele support partitioned by strand, with positive strand support above the black line and negative strand support below. (**C**) Plugin panel. The opened GATK plugin can be used to compute genotypes from read alignment tracks within the browser. (**D**) Variant Navigator panel. The Variant Navigator visualizes and guides the navigation of genetic variant data. The map page of the Variant Navigator displays a matrix where each column represents an individual or sample from the file and each row represents a variant position; each cell in the matrix is colored according to the allele possessed by the corresponding sample and position, or is transparent if no allele is predicted there. The genomic range displayed in the Variant Navigator is a superset of the range for tracks, and users can click within the Variant Navigator to navigate to subranges of the variant range.

reads whose mate was not mapped by the aligner. These representations, especially when combined, make structural rearrangements easily interpretable as shown in Figure 2, where absent read coverage indicates a deletion in the genome, which is confirmed by discordant (over-stretched) read-pairs joining the two ends of a deletion.

Visualizations for population sequencing studies

The decreased costs and increased throughput of HTS have enabled the sequencing of large cohorts, both from specific disorders and general populations (7,8). Genome browsers, in turn, need to support visualization of data that has been agglomerated from many genomes. Savant has comprehensive support for multi-individual genotype data sets in the standard VCF format (9). However the visualization of variation data introduces additional resolution complexities, as SNPs typically appear every 100–1000 bp, depending on the number of individuals and the locus, but can also appear at adjacent nucleotides. Thus, viewing SNP data over a 10 kb region would require drawing each SNP as less than 1 pixel, to maintain the genome scale and prevent overlaps between adjacent SNPs. Savant introduces original visualizations and navigation interfaces tailored for the efficient perusal of genotyped cohorts. Data from a genetic variant data set is shown in two areas of the Savant interface: in the main genome-scale track browser, as well as in a new navigation component that agglomerates data from larger areas of the genome.

Variant tracks are visualized as a matrix, where each row represents an individual or sample from the file and each column represents a genomic position that is vertically aligned with the rest of the tracks. Each cell in the matrix is colored according to the non-reference nucleotide in the corresponding sample and position, or is transparent if the allele is reference. Because it is difficult to identify trends from bands of mixed colors when data sets contain data for many samples, as in the 1000 Genomes Project, a summative view is also provided by an Allele Frequency mode, which shows the frequency of each allele per position in a way that is analogous SNP mode for HTS read alignments.

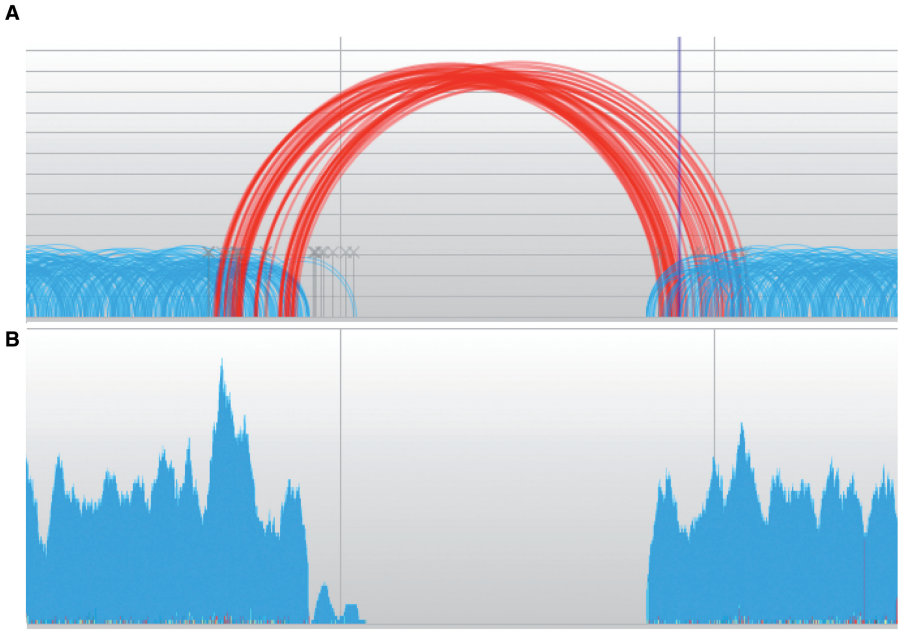

Figure 2. Visualization modes for structural variation detection. (**A**) Paired read alignments displayed in Arc mode. The taller arcs represent pairs that are identified by Savant as being discordant (red) and are colored differently from the concordant (blue) ones. This indicates a deletion event in the sequenced genome. (**B**) The same data set as the top track but displayed in coverage mode. The lack of coverage in the region within the bounds of the discordant pairs confirms the deletion event.

The Variant Navigator interface, located on the right side of the browser (see Figure 1) is used to display variant data on an independent scale, allowing the visualization of SNPs over larger segments of the genome. In the simplest mode, the Variant Navigator lists variants in text or visual form (Figure 1D). Within the same window users can also perform case–control analysis by assigning samples to one of two cohorts, and visualizing the respective allele frequencies distributions at the variable positions (Figure 3A). Linkage disequilibrium, a measure of allelic correlation across variant positions, can also be computed and visualized with this component as done in Haploview (10) (Figure 3B).

PLUGINS

The Savant Genome Browser is further extensible by plugins. These plugins can conduct computations using currently loaded data, visualize results and navigate the browser to regions of interest, all while utilizing external data sets. The Savant Application Programming Interface (API) provides plugins with extensive

- 'Visualization functionality' to display graphics that are either superimposed on top of tracks or in a separate reserved space, whose visibility can be toggled.
- 'Analytic functionality' to perform computation on and manipulation of the data. If the computations are fast, plugins can visualize the results in realtime alongside track navigation. Otherwise, they may load the results as a track upon completion.
- 'Navigation functionality' to provide interfaces for quickly loading genomic regions of interest. This is particularly useful for using external data sets (e.g. a list of genes) for guiding genome navigation.
- 'Datasource functionality' to enable retrieval of track data from alternate data sources. This functionality is useful for loading tracks directly from public or private databases, or from external programs.

The Savant Software Development Kit (SSDK) includes source code for sample plugins and a full documentation of the rich API enables the development of such plugins by the user community. New plugins can be contributed to a public repository, and are made available for download to all users through Savant's built-in Plugin Manager. Table 1 summarizes a selection of Savant Plugins, some developed by the core Savant Development Team, and others by external users. Several of these are explained in more detail in the following sections.

UCSC Plugin

Since Savant runs on client machines instead of on a centralized server it has the important advantage of maintaining complete privacy of sensitive data, such as the genomes of specific patients. Nevertheless, it is often necessary to examine this data in the context of the wealth of publicly available genomic information. The UCSC Genome Database is perhaps the most extensive repository of this type and contains the underlying data for all available tracks displayed on the UCSC browser. The UCSC Explorer is a plugin that makes it possible to open UCSC tracks within Savant without downloading the raw data files—only the relevant data is downloaded via a direct connection to the UCSC Database, and is

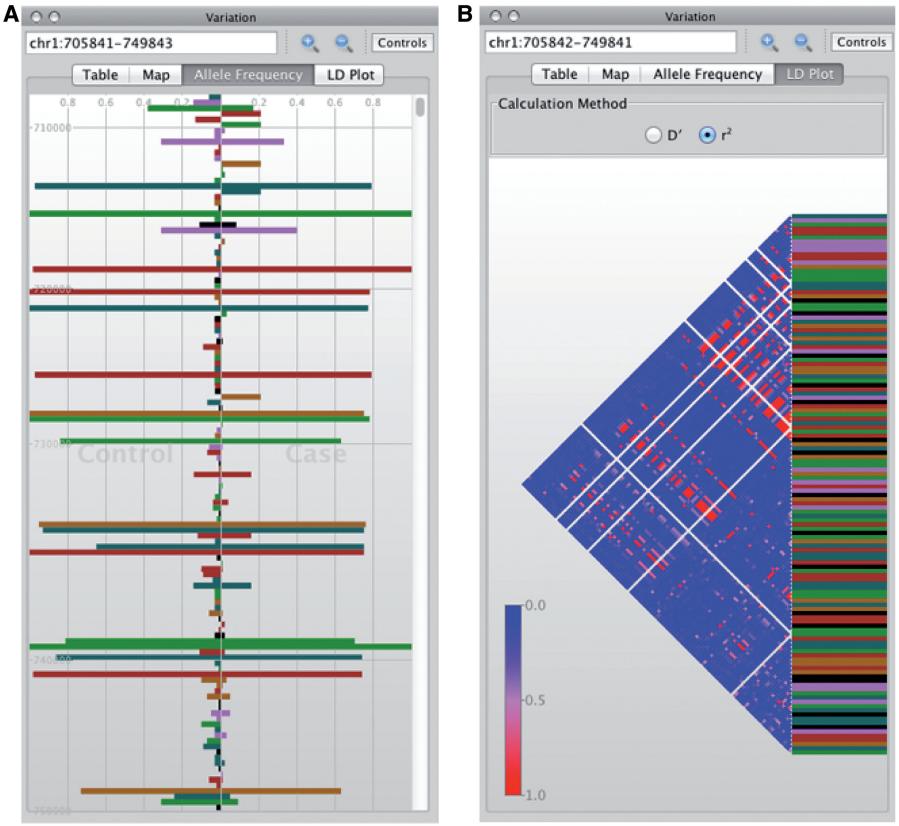

Figure 3. Visualizations of genetic variant data. (**A**) A view of the Allele Frequency page of the Variant Navigator, which compares allele frequencies of genetic variants from two cohorts from the 1000 Genomes Project. At most positions the frequencies are similar between cohorts, though there are positions that exhibit different frequencies. (**B**) An LD plot of variants in the same range as on the left. Blue and red cells represent low and high correlation between variant positions, respectively.

Table 1. List of selected Savant plugins

Plugin	Description
Chromatogram	Shows Sanger sequencing chromatograms overlaid on the reference genome
edgeR	Detects and visualizes differential enrichment from RNA-seq or ChIP-seq data
Data table	Shows textual data from track records in tabular form
GATK	Predicts and visualizes genotypes from read alignment tracks
Gene ontology	Guides navigation based on ontology terms
PING	Guides navigation based on protein–protein interaction databases
Remote commander	Issue navigation and other commands remotely through external tools
Ribosome	Shows the amino acid translation of gene tracks
RNA-seq analyzer	Reconstructs and estimates isoforms from RNA-seq data
SimpleSNP	Predicts and visualizes SNVs from read alignment tracks
Snapshot	Exports track images at every bookmarked genomic region
SRMA	Realigns HTS read alignments
UCSC explorer	Provides a graphical interface for loading UCSC tracks
WikiPathways	Guides navigation based on biological pathways

Plugins are available for installation directly through the built-in Plugin Manager.

immediately presented to the user. The tracks are categorized in a manner that mimics their presentation within the popular web-based UCSC Genome Browser.

RNA-seq analyzer plugin

RNA sequencing has been transformative in transcriptomics by simplifying the process of determining the identity and abundance of isoforms within the cell. The plugin accepts input from previously reconstructed isoforms, through programs such as Cufflinks (11), or alternatively performs isoform reconstruction and abundance estimation from the set of read alignments directly using the iReckon algorithm (Mezlini and Brudno, unpublished data). The plugin overrides the default coloring scheme in Savant and instead colors each read according to the most probable isoform from which it was generated. For any gene of interest, a multi-coverage profile is provided for comparing read support for each isoform. A pie chart summarizing the relative proportions of the isoforms is also provided. Finally, the plugin can incorporate two data sets simultaneously, and allows for their comparison by visualizing the differences in expression.

edgeR plugin

The analysis of quantitative HTS data (e.g. from RNA-seq or ChIP-seq) relies on statistical procedures that highlight

differential regions. For example, the density of mapped reads in a particular genomic region may represent enrichment level of a protein–DNA interaction (ChIP-seq), or gene expression level (RNA-seq). The edgeR plugin is a wrapper for software written in the R statistical programming language for the detection of significantly differentially enriched regions or expressed genes, relative to observed biological variation, directly within Savant (12). The plugin computes on multiple BAM tracks, some designated as Case and others as Control, and provides a table of ranked results, including the region locations, log-fold-changes, *P*-values and estimated false discovery rates of the change between conditions.

WikiPathways plugin

WikiPathways is an open collaborative platform for the curation of biological pathways (13). The WikiPathways plugin provides an interface to search, browse and visualize the over 1500 pathways available from this platform, and to use pathways to guide navigation to relevant genomic locations within Savant. The use of functional annotations for navigation through large genomes represents a significant departure from existing navigation techniques, which are almost entirely based on linear scanning.

PING: protein interaction network to genome plugin

To better understand the functional consequences of sequence variants it is necessary to look beyond the genome, to gene products and their interactions, especially when dealing with complex (i.e. non-monogenic) diseases. Given a query gene, the PING plugin allows one to view the partner genes (interactors) that engage in known protein–protein interactions, mapped across the genome. The program provides hyperlinks to further information: to Entrez for information about each interactor, to iRefWeb (14) for query-interactor information and to DAnCER (15) for gene annotations including associations to disease and GO (16).

Application plugins

A large number of computational tools have been developed for analyzing HTS data but nearly all require command-line invocation and have file-based input and output. These programs can be easily chained together and run on large data sets, however such convenience hampers the ability to efficiently fine-tune their parameters and severely restricts their use mainly to computational specialists. The SSDK has been expanded to support the incorporation of a wide array of genomic tools within Savant, thereby unlocking opportunity for performing many computational analyses within a powerful visual environment. Now, with a minimal amount of effort—namely, specification of the program's input, output and parameters—virtually any command-line tool that computes on genomic data can be incorporated as a plugin within Savant, and its results rendered as a track immediately upon completion. This functionality is similar to that provided by the Galaxy Track Browser (GTB) (17) as part of the larger Galaxy framework (18), the key distinction being that Galaxy is a web based, server-side package, whereas Savant is client-side. To illustrate this ability we have built wrappers for two popular applications: the Unified Genotyper of the Genome Analysis Toolkit (GATK) and the Short Read Micro Realigner (SRMA).

GATK plugin

GATK predicts SNPs and indels from HTS read alignments (19). While this genotyper is modeled to account for technology-specific biases automatically, it is still highly tuneable, allowing users to carefully adjust the sensitivity of the underlying detection algorithm. The GATK plugin is an XML specification of the input, output and parameters of the Unified Genotyper. These parameters are specified within Savant, which invokes the genotyper, and subsequently visualizes the resulting VCF file. The plugin can quickly compute and visualize genotypes for a segment of the genome, allowing for rapid and dynamic experimentation with program parameters, prior to invoking the tool on a whole-genome scale.

SRMA plugin

Read alignment tools consider each read independently. In the absence of additional information the precise positioning of indels within a mapped read is difficult, particularly towards the ends of reads where sequencing quality tends to deteriorate. SRMA is a tool that performs realignment of previously mapped reads based on a local consensus (20), with the aim of sharing information across reads so as to remove false positives and properly place aberrantly positioned variants. Like the GATK plugin, the SRMA plugin is a wrapper around this command-line tool that facilitates its running within the visual environment of Savant, making possible real-time realignment of track reads and further deployment on a whole genome scale.

DISCUSSION AND FUTURE WORK

Despite the proliferation of automated tools for computation on HTS data, human interpretation is still necessary for its analysis. A lack of tools that support visual analysis of HTS data has precluded efficient interpretation of data, especially for biologists without significant informatics expertise. The Savant Genome Browser has been designed to meet the demands of even large population sequencing efforts. In addition to having extended data access and support, the current version of Savant delivers creative visualization representations for HTS data and a significantly upgraded plugin architecture that provides the opportunity to incorporate any computational tool within a visual environment.

We believe that synergizing the processes of visualization and analysis of genomic data significantly improves on the capabilities of current genome browsers. To enable this, we will work to further develop the Savant plugin environment and to work with developers to incorporate new visualizations, analytics and tools into the browser. We welcome developers to join the Savant Community, download the SSDK (which includes source code for sample plugins and the API) and contribute to the project.

ACKNOWLEDGEMENTS

We thank the Savant User Community for feedback on our tool, and the Savant Developers for submitted plugins.

FUNDING

CIHR Tools, Techniques and Innovation grant and MITACS Seed (to M.B.); the OGI SPARK (to M.B. and M.F.); an NSERC Graduate Fellowship (to M.F.); Google Summer of Code. Funding for open access charge: CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
2. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nature Biotechnol.*, **29**, 24–26.
3. Fiume,M., Williams,V., Brook,A. and Brudno,M. (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, **26**, 1938–1944.
4. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) University of California Santa Cruz The UCSC Genome Browser database. *Nucleic Acids Res.*, **31**, 51–54.
5. Yoon,S., Xuan,Z., Makarov,V., Ye,K. and Sebat,J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
6. Medvedev,P., Fiume,M., Dzamba,M., Smith,T. and Brudno,M. (2010) Detecting copy number variation with mated short reads. *Genome Res.*, **20**, 1613–1622.
7. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
8. O'Roak,B.J., Deriziotis,P., Lee,C., Vives,L., Schwartz,J.J., Girirajan,S., Karakoc,E., MacKenzie,A.P., Ng,S.B., Baker,C. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.*, **43**, 585–589.
9. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* and 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
10. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
11. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, **28**, 511–515.
12. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
13. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184+.
14. Turner,B., Razick,S., Turinsky,A.L., Vlasblom,J., Crowdy,E.K., Cho,E., Morrison,K., Donaldson,I.M. and Wodak,S.J. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, **2010**. October 12 (doi: 10.1093/database/baq023; epub ahead of print).
15. Turinsky,A.L., Turner,B., Borja,R.C., Gleeson,J.A., Heath,M., Pu,S., Switzer,T., Dong,D., Gong,Y., On,T. *et al.* (2011) DAnCER: disease-annotated chromatin epigenetics resource. *Nucleic Acids Res.*, **39**, D889–D894.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
17. Goecks,J., Li,K., Clements,D., Team,T.G. and Taylor,J. (2011) The Galaxy Track Browser: transforming the genome browser from visualization tool to analysis tool. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on IEEE*, pp. 39–46.
18. Goecks,J., Nekrutenko,A., Taylor,J. and Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86+.
19. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
20. Homer,N. and Nelson,S. (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol.*, **11**, R99+.