BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2477–2478
doi:10.1093/bioinformatics/btt406

Genome analysis

Advance Access publication July 10, 2013

MIG: Multi-Image Genome viewer

Simon J. McGowan[1,*], Jim R. Hughes[2], Zong-Pei Han[1] and Stephen Taylor[1]

[1]Computational Biology Research Group and [2]Molecular Haemotology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Headington, Oxford OX3 9DS, UK

Associate Editor: John Hancock

ABSTRACT

Summary: Multi-Image Genome (MIG) viewer is a web-based application for visualizing, querying and filtering many thousands of genome browser regions as well as for exporting the data in a variety of formats. This methodology has been used successfully to analyze ChIP-Seq data and RNA-Seq data and to detect somatic mutations in genome resequencing projects.

Availability: MIG is available at https://mig.molbiol.ox.ac.uk/mig/

Contact: simon.mcgowan@imm.ox.ac.uk

Received on May 15, 2013; revised on June 14, 2013; accepted on July 8, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) technologies are producing vast amount of data for analyses. It is usual for a typical ChIP-Seq, RNA-Seq or genome resequencing experiment to produce millions of sequence reads that are subjected to various iterations of analysis. Once the reads are mapped to a reference genome, they can be visualized in a genome browser such as the UCSC genome browser (Kent *et al*., 2002), GBrowse (Stein *et al*., 2002), IGV (Robinson *et al*., 2011) or IGB (Nicol *et al*., 2009), often in combination with a number of associated tracks such as gene structures, expression graphs, conservation plots and other quantitative data. Depending on the experiment, the data are then stored in a variety of file formats, queried and filtered using a variety of bespoke tools and languages to produce spreadsheets, web pages and overview plots and visualizations. Bioinformatics tools for analysis of such large amounts of data are still immature, have a wide range of parameters and often produce confusing results. A key issue is to verify the findings by linking filtered results to the original mapped data. For example, in a typical ChIP-Seq experiment, one would map the reads and isolate regions of enrichment using a peak finding algorithm such as MACS (Zhang *et al*., 2008). Peak finding is often dependent on antibody quality and experimental conditions. Choosing suitable cutoff parameters such as *P*-values, false discovery rate, density of sequences and fold enrichment is difficult; hence, verification of the results is often done by visual inspection. Once a set of satisfactory peaks are identified, the next stage is to look for overlaps with various genomic features, such as CpG islands, genes, promoters and regions of conserved sequences. Many biologists are overwhelmed by the volume of the data and find it difficult to query and manipulate the output. An alternative use case might be in filtering somatic variant calls following NGS analysis of matched samples. Here, variants might be filtered according to total coverage at the variant position, coverage supporting the reference and variant in both the affected and unaffected datasets, variant effects (synonymous versus non-synonymous) and intersection with other datasets (e.g. known cancer mutations).

Although it is possible to visualize and compare multiple tracks using the aforementioned genome browsers, comparison of multiple features at different locations within these tracks is slow and requires the user to repeatedly scroll along a chromosome or manually enter genomic locations before reloading the browser. To facilitate this process, we present the Multi-Image Genome viewer (MIG), which allows the rapid filtering of thousands of metadata pertaining to genomic regions in a simple web interface combined with genome views of the region of interest.

2 IMPLEMENTATION

MIG stores information concerning a set of genomic features—for example, multiple peaks called in a ChIP-Seq analysis—in a MySQL 'project' database. For ChIP-Seq data, this meta-information can include *P*-values, fold enrichment and gene annotations, but may also include a wide variety of other types of data more appropriate to the analysis (e.g. read coverage or dbSNP annotation). A web-based interface has been written to allow the genomic features to be filtered by the user, using a combination of drop-down menus, checkbox selections and filtering by keywords. The filtered set of features can be presented as a sorted list with user-defined annotations, or as a set of plots generated 'on-the-fly'. Clicking on any of the filtered features causes the import and display of a configurable view of the feature from a pre-defined genome browser session. Finally, the filtered set of features can be exported in a variety of formats including GFF3 and FASTA. In addition, a given set of filters within MIG can be saved as a session to be used later. This feature has proved extremely useful in allowing an analyst to identify an initial dataset and then share this set with other biologists (who are then able to refine this starting set of features).

In addition to the 'project' databases, MIG also has a master database to store administration data concerning each of the projects (project titles, user permissions, etc.). When logging into MIG, a user is presented with projects they have loaded as well as those they have been granted access to by other users. Projects are created by uploading GFF3-formatted datasets. GFF3 is a tab-delimited, nine-column format with the added advantage of allowing metadata to be associated in the ninth column for each genomic feature. Each name/value pair in

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

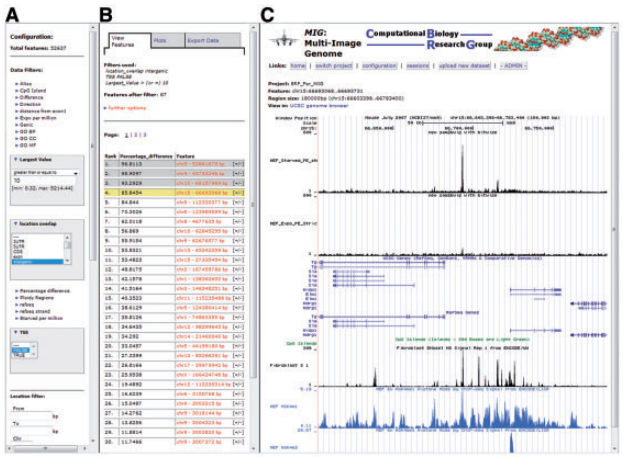

Fig. 1. MIG view of a sorted list of 87 of the most differentially bound peaks with the two ERF ChIP-seq datasets. Control facets in panel A are used to refine the peaks based on maximum signal in either datasets to select for strongly bound regions, lack of overlap with annotated transcription start sites (TSS from UCSC Known Gene annotation build mm9) and intergenic localization to select for potential distal regulatory elements. The resultant list is viewed in panel B ranked by the calculated difference between the normalized binding signal in the two ChIP-seq experiments. An example peak in panel B is maximized to show the other annotated data associated with each peak. The genomic view from the predefined UCSC genome browser session for the example peak is shown in panel C. This genome view allows the visual inspection of the primary data in conjunction with relevant publicly available data such as Dnase-seq, ChIP-seq and gene annotation

the ninth column defines a 'facet' for the feature that can subsequently be queried using MIG's web interface.

Accessory scripts have been written to aid the construction of the input GFF3 files. Facets may be added using a perl script, *annotate.pl*, which uses CisGenome (Hongkai, 2008) for annotating region overlaps with various 'standard' gene feature classes such as introns, exons, intragenic and intergenic regions. Gene ontologies (The Gene Ontology Consortium, *2000*) are also mapped using the nearest gene. Additionally, a second script, *intersectappend.pl*, can be used to integrate data from other resources (e.g. CpG islands, areas of histone modification, peak calls from other ChIP-Seq experiments). This script wraps BEDTools (Quinlan and Hall, 2010) and allows regions to be classified as overlapping (TRUE/FALSE) each of the features in the set. Both *annotate.pl* and *intersectappend.pl* append a name/value pair to the ninth column of the source GFF3 file. These tools are available for download on the web site.

MIG has been used to analyze a combined DNase-seq, ChIP-seq and RNA-seq dataset (Kowalczyk *et al.*, 2012). As an example of MIG's use with ChIP-Seq data, we use a dataset generated with an antibody recognizing the transcriptional activator, ERF (Twigg *et al.,* 2013). Examples of data querying using the MIG interface can be seen in Figure 1 and also within MIG itself at https://mig.molbiol.ox.ac.uk/mig. Tutorials on the web site are also available based on this dataset.

In summary, we see MIG as a compelling method to investigate NGS data and other large datasets by bringing a powerful and intuitive web-based interface to the biologist. MIG allows the user to generate these collections in an automated way, and store multiple projects in the MIG database. These collections can also be shared with other MIG users so that whole groups can perform analysis on the same data.

Funding: The Computational Biology Research Group is funded by the Medical Research Council and Weatherall Institute of Molecular Medicine.

Conflict of Interest: none declared.

REFERENCES

Hongkai,J. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kowalczyk,M.S. *et al.* (2012) Intragenic enhancers act as alternative promoters. *Mol. Cell*, **45**, 447–458.

Nicol,J.W. *et al.* (2009) The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, **25**, 2730–2731.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotech.*, **29**, 24–26.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Twigg,S. *et al.* (2013) Reduced dosage of ERF causes complex craniosynostosis in humans and mice and links ERK1/2 signaling to regulation of osteogenesis. *Nat. Genet.*, **45**, 308–313.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016