BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2059–2061
doi:10.1093/bioinformatics/btt318

Gene expression

Advance Access publication June 4, 2013

Whole-Genome rVISTA: a tool to determine enrichment of transcription factor binding sites in gene promoters from transcriptomic data

Inna Dubchak[1,2,*], Matthew Munoz[3], Alexandre Poliakov[2], Nathan Salomonis[4], Simon Minovitsky[2], Rolf Bodmer[5] and Alexander C. Zambon[3,*]

[1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA, [3]Departments of Pharmacology and Medicine, University of California at San Diego, La Jolla, CA 92093, USA, [4]California Pacific Medical Center Research Institute, San Francisco, CA 94107, USA and [5]Development and Aging Program, Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: We have developed a web-based query tool, Whole-Genome rVISTA (WGRV), that determines enrichment of transcription factors (TFs) and associated target genes in sets of co-regulated genes. WGRV enables users to query databases containing pre-computed genome coordinates of evolutionarily conserved transcription factor binding sites in the proximal promoters (from 100 bp to 5 kb upstream) of human, mouse and *Drosophila* genomes. TF binding sites are based on position-weight matrices from the TRANSFAC Professional database. For a given set of co-regulated genes, WGRV returns statistically enriched and evolutionarily conserved binding sites, mapped by the regulatory VISTA (rVISTA) algorithm. Users can then retrieve a list of genes from the query set containing the enriched TF binding sites and their location in the query set promoters. Results are exported in a BED format for rapid visualization in the UCSC genome browser. Flat files of mapped conserved sites and their genomic coordinates are also available for analysis with stand-alone software.

Availability: http://genome.lbl.gov/cgi-bin/WGRVistaInputCommon.pl.

Contact: azambon@ucsd.edu or ildubchak@lbl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 19, 2012; revised on April 18, 2013; accepted on May 25, 2013

1 INTRODUCTION

Methods for transcriptome analysis (e.g. microarray and RNA sequencing) have revolutionized our ability to identify global changes in mRNAs during cellular remodeling. A continuing challenge for such analyses is identifying the key signaling events that mediate the observed gene expression changes. Transcription of mRNA is a dynamic process controlled by many factors, including chromatin states, cofactors, RNA polymerase and the dynamic regulation of transcription factors (TFs). The Gene Ontology Consortium has classified ~1030 human TF genes that encode for proteins with sequence-specific DNA-binding activity.

TFs, by their very nature, are key regulators of the observed transcriptional changes during cellular remodeling. TFs can be regulated post-transcriptionally (e.g. by phosphosphorylation and/or changes in intracellular location), and thus may not be altered in expression alongside their target genes. The enrichment of TF binding sites (TFBS) near target genes, however, can be used as a surrogate for predicting altered TF activity. Thus, we developed Whole-Genome rVISTA (WGRV), a web-based tool for finding overrepresented evolutionarily conserved TFBS within the proximal promoter regions of a set of differentially expressed genes.

Searching proximal promoters for TFBS often results in many false positives. This is because the positional weight matrices that are used to computationally predict TFBS are often short in length and degenerate at specific positions. The use of phylogenetic footprinting to filter TFBS that are conserved in gene promoters between two relatively closely related species has been shown to significantly reduce false positive sites (Loots *et al.*, 2002).

Chromatin immunoprecipitation followed by high-content sequencing (ChIP-seq) has emerged as the method of choice for determining bona fide TFBS (Mardis, 2007). ChIP-seq studies, however, require a priori knowledge of the TF for which binding sites are to be determined. Moreover, such studies require a validated antibody and can be costly and time-consuming. WGRV queries are fast compared with ChIP-seq experiments, and the queries can be carried out simultaneously for all TFBS stored in the database at no cost. We believe WGRV is an excellent method to computationally identify potentially relevant TFs that drive tissue and cellular remodeling, which can then be followed up with ChIP-seq for further validation.

2 METHODS AND IMPLEMENTATION

We initially developed the WGRV algorithm to identify the enrichment of cyclic AMP-responsive elements in clustered murine transcripts altered by protein kinase A activation (Brudno *et al.*,

*To whom correspondence should be addressed

2007; Zambon *et al*., 2005). The capabilities of WGRV have now been significantly enhanced. WGRV is now available for human, mouse and *Drosophila* TF target analysis. It incorporates mappings for all TFBS matrices in the most current TRANSFAC Professional database (Matys *et al*., 2006). As described later, we have incorporated a number of enhanced input and output features, and we compare WGRV performance with existing tools using publicly available ChIP-seq and microarray data.

WGRV uses the VISTA/LAGAN alignment pipeline (Dubchak *et al*., 2009) to first align whole-genome assemblies of two related species (Table 1).

The rVISTA algorithm (Loots *et al*., 2002) is then applied to identify all the conserved TFBS within the 5 kb proximal promoter regions of every orthologous gene mapped between the two species. Conserved TFBS (cTFBS) must exhibit flanking sequence conservation between species (see details in the Supplementary Methods). These results are stored in a database that can be queried over the web using a list of genes (e.g. genes significantly up- or down-regulated between two conditions in a microarray or high-content sequencing experiment). Alternatively, flat files are available (http://pipeline.lbl.gov/data/conserved_tfbs/) that can be incorporated in overrepresentation analysis software (e.g. GO-Elite; Zambon *et al*., 2012).

WGRV users can query lists of co-regulated genes for enrichment of cTFBS via a web interface (Supplementary Fig. 1). Once genes are submitted, the corresponding cTFBS are retrieved and a binomial test is performed to determine whether any cTFBS are overrepresented (at a user defined *P*-value cutoff) in the promoters of the submitted genes. Statistical enrichment is computed by comparing the enrichment in the query set against all aligned promoters in the genome or against a user-defined background set (e.g. all the genes surveyed by a microarray).

Results include a list of enriched cTFBS sorted by *P*-value of enrichment. The user can also view lists of predicted target genes for each of the enriched TFs, the location of cTFBS on the genome and VISTA conservation plots for each mapped gene promoter that depicts the positional location of the enriched cTFBS.

Overrepresented cTFBS can be exported to BED format for easy upload and viewing on the UCSC genome browser. This enables sites to be viewed alongside a wide variety of publicly available data, such as histone modifications and ChIP-seq data that are available at the UCSC genome browser. It also facilitates retrieval of flanking genomic sequences for designing primers for ChIP-quantitative PCR validation studies. In addition to the submission of gene lists, WGRV allows users to search the database for all the mapped locations for any TF in the database. More detailed analysis of TFBS, including non-conserved sites and genomic regions outside of the 5 kb proximal promoters, can be easily conducted using the rVISTA tool that is linked to WGRV (Loots *et al*., 2002).

Table 1. Whole-genome alignments used in WGRV

Reference species	Aligned species
Homo sapiens	*Mus musculus*
Mus musculus	*Homo sapiens*
Drosophila melanogaster	*Drosophila pseudoobscura*
Drosophila melanogaster	*Drosophila virilis*

3 PERFORMANCE COMPARISON

There are several web-based tools for identifying enrichment of cTFBS in gene lists, such as oPOSSUM (Ho Sui *et al*., 2007), DiRE (Gotea and Ovcharenko, 2008) and CONFAC (Karanam and Moreno, 2004). Significant differences between WGRV and these tools include (i) genome alignment methods, (ii) database of TF matrices used, (iii) raw data access and (iv) data visualization. These features and others are summarized in Supplementary Table 1. We compared the capabilities of WGRV with oPOSSUM, the most similar tool in terms of features and output, to predict significant enrichment of the TF used in a number of published ChIP studies (Linhart *et al*., 2008). Lists of TF target genes (Supplementary Data File 1) were input into both the tools and identical query settings were used (2000 bp upstream of the transcriptional start site and $P<0.005$). WGRV correctly identified the significant enrichment of TF sites in 9 of the 10 datasets compared with 3 of the 10 experiments correctly predicted by oPOSSUM (Supplementary Table 2). Moreover, WGRV was able to rank the correct site in the top five most significant enriched TFs in 7 of the 10 datasets.

As a secondary performance comparison, we examined the ability of WGRV, oPOSSUM and DiRE (Gotea and Ovcharenko, 2008) to detect hypoxia-inducible factor (HIF1) binding site enrichment in the promoters of 960 genes significantly up-regulated (fold >1.3, $P<0.05$, Supplementary Data File 2) in cells cultured in hypoxia (1%O_2) for 8 h (GSE27975) (Perman *et al*., 2011). We hypothesized that genes up-regulated in this dataset would contain an enrichment in HIF1 binding sites. WGRV identified HIF1 as being the most significantly enriched TFBS in the query set promoters (Table 2). DiRE analysis also identified HIF1 enrichment, but HIF1 was listed the 28th out of 98 sites positively contributing to the enhancer model. Using these input genes, oPOSSUM was not able to identify significant enrichment of HIF1 sites.

If an increased fold cutoff (>2, $P<0.05$) is used to define the query set, thereby decreasing the number of genes to 162, all three tools were able to identify HIF1 enrichment (Supplementary Table 3). WGRV results, however, were not significantly different in terms of HIF1 ranking or the total number of enriched cTFBS. This is in stark contrast to oPOSSUM and to a lesser extent to DiRE results.

In conclusion, WGRV is a robust tool to predict altered TF activity and associated target genes from large lists of

Table 2. Comparison of the TFBS enrichment methods

Tool	Number of enriched TFBS	HIF1 rank
WGRV	4	1
oPOSSUM	1	Not found
DiRE	98	28

differentially expressed or co-regulated transcripts. It provides a number of options for data visualization and is publically available at http://genome.lbl.gov/vista.

Funding: National Institutes of Health (R01 HL091495 to I.D., 1U54HL08460 to M.M. and A.Z., 8UL1TR000100, P01HL098053 to A.Z., R01 HL54732, P01 HL0980539, P01 AG033561 to R.B.) and American Heart Association (10SDG2630130 to A.Z.). The work, conducted by the U.S. Department of Energy Joint Genome Institute (I.D., A.P., S.M.), is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors would like to thank Ivan Ovcharenko for helpful discussions.

Conflict of Interest: none declared.

REFERENCES

Brudno,M. *et al*. (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. *Nucleic Acids Res*., **35**, W669–W674.

Dubchak,I. *et al*. (2009) Multiple whole-genome alignments without a reference organism. *Genome Res*., **19**, 682–689.

Gotea,V. and Ovcharenko,I. (2008) DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res*., **36**, W133–W139.

Ho Sui *et al*. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res*., **35**, W245–W252.

Karanam,S. and Moreno,C.S. (2004) CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Res*., **32**, W475–W484.

Linhart,C. *et al*. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res*., **18**, 1180–1189.

Loots,G.G. *et al*. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*., **12**, 832–839.

Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.

Matys,V. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.

Perman,J.C. *et al*. (2011) The VLDL receptor promotes lipotoxicity and increases mortality in mice following an acute myocardial infarction. *J. Clin. Invest*., **121**, 2625–2640.

Zambon,A.C. *et al*. (2012) GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, **28**, 2209–2210.

Zambon,A.C. *et al*. (2005) Gene expression patterns define key transcriptional events in cell-cycle regulation by cAMP and protein kinase A. *Proc. Natl Acad. Sci. USA*, **102**, 8561–8566.