BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 3 2012, pages 419–420
doi:10.1093/bioinformatics/btr666

Genome analysis
Advance Access publication December 6, 2011

Genome*Runner*: automating genome exploration

Mikhail G. Dozmorov*, Lukas R. Cara, Cory B. Giles and Jonathan D. Wren*
Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104-5005, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: One of the challenges in interpreting high-throughput genomic studies such as a genome-wide associations, microarray or ChIP-seq is their open-ended nature—once a set of experimentally identified regions is identified as statistically significant, at least two questions arise: (i) besides P-value, do any of these significant regions stand out in terms of biological implications? (ii) Does the set of significant regions, as a whole, have anything in common genome wide? These issues are difficult to address because of the growing number of annotated genomic features (e.g. single nucleotide polymorphisms, transcription factor binding sites, methylation peaks, etc.), and it is difficult to know *a priori* which features would be most fruitful to analyze. Our goal is to provide partial automation of this process to begin examining associations between experimental features and annotated genomic regions in a hypothesis-free, data-driven manner.

Results: We created Genome*Runner*—a tool for automating annotation and enrichment of genomic features of interest (FOI) with annotated genomic features (GFs), in different organisms. Besides simple association of FOIs with known GFs Genome*Runner* tests whether the enriched FOIs, as a group, are statistically associated with a large and growing set of genomic features.

Availability: Genome*Runner* setup files and source code are freely available at http://sourceforge.net/projects/genomerunner.

Contact: mikhail-dozmorov@omrf.org; Jonathan-Wren@omrf.org; jdwren@gmail.com

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

Received on June 1, 2011; revised on November 10, 2011; accepted on November 29, 2011

1 INTRODUCTION

Genomes encode the instructions for life, and within genomes are many different features with different roles (e.g. genes, CpG islands, microRNAs). High-throughput (HT) technologies have enabled us to experimentally examine whether or not genomic variations associate with certain conditions (e.g. disease). These variations in features of interest (FOI) such as single nucleotide polymorphisms (SNPs) come from various technologies like genome-wide association (GWA) studies, peaks called from ChIP-on-Chip and ChIP-seq experiments and other data from deep sequencing experiments. Regardless of the type of data, each FOI has genomic coordinates that uniquely identify it within the genome.

Tools to prioritize FOIs usually analyze local genomic features (e.g. evolutionary conservation) of individual FOIs and do not consider the set of experimentally identified FOIs as a whole. Functional interpretation of FOI sets is often performed using gene- and pathway-based tools adopted from microarray data analysis (Dennis *et al.*, 2003; Ji *et al.*, 2008; Subramanian *et al.*, 2005; Wang *et al.*, 2007). These tools perform mapping of FOIs to genes, prioritize lists of FOIs and calculate gene- and pathway-enrichment statistics. They, however, rely upon a gene's role being known and, for humans, approximately one-third of genes have no known function (Wren, 2009). Plus, as projects like ENCODE are showing, there are many non-coding regions of interest in the genome (Birney *et al.*, 2007). As such, we're lacking a way to automatically explore the genome and associate our FOIs with genomic features beyond well-annotated gene regions.

The idea of automating genome exploration is not new (Wren *et al.*, 2005), but automated exploration of correlations is complicated by the heterogeneous nature of the HT data versus annotated genomic features (GF) types. Many efforts have been devoted to developing tools for prioritizing individual SNPs (Cline and Karchin 2011; Gauderman *et al.*, 2007), but analysis of what annotated genomic features other than well-annotated gene regions may be associated with FOI sets has not received as much attention, particularly with regards to data-driven exploration. The GREAT tool, for example, associates genomic regions with their putative target genes and calculates Gene Ontology enrichment statistics (McLean *et al.*, 2010), but is gene-centric, whereas Genome*Runner* is designed as a much more general-purpose association tool. Galaxy has a variety of tools for operating on genomic intervals and SNP prioritization (Goecks *et al.*, 2010), and GenGen (Chen *et al.*, 2010; Wang *et al.*, 2007) provides a set of Perl scripts for the association of FOIs with genes, transcription factor binding sites, microRNAs, EvoFold regions and other user-provided data but requires programming skills and data formatting. Genome*Runner* differs from these tools in that it annotates SNPs and regions, and automatically searches for any statistically significant enrichment of FOIs with multiple GFs.

2 FEATURES AND METHODS

2.1 The interface

Genome*Runner* has an intuitive point-and-click interface for querying of FOIs against a database of GFs. It accepts lists of FOIs in a tab-delimited format where each region is represented by chromosome name and a start and end coordinates. A user has an option to annotate and calculate enrichment of a set of FOIs with >750 GFs (hg19 database), including genes/exons/introns, upstream

*To whom correspondence should be addressed.

promoter regions, DNAse clusters, empirically validated and predicted conserved transcription factor binding sites, epigenetics marks, empirical and predicted microRNAs and regions conserved across organisms and more.

3 THE DATABASE

Tracks in UCSC genome browser are represented by tables (Karolchik *et al*., 2004) available for download. As such they represent an ideal mechanism for assembling genome annotation into a single database and for querying associations of FOIs with a GF of interest. We host selected UCSC tracks (Supplementary Material) in a MySQL database available to all Genome*Runner* users, but there is also an option for local installation of this database. Genome*Runner*'s database is monthly synchronized with UCSC data and new GFs are added for Genome*Runner*'s analysis as they become available.

3.1 Monte-Carlo simulations

Besides simple associations of individual FOIs with GFs, Genome*Runner* estimates the likelihood of observing associations with individual GFs for a set of FOIs versus a set of random regions with the same characteristics. Genome*Runner* runs enrichment tests against whole genome as a background by default; options for loading user-defined background are available. Each simulation is done by selecting the same number of random points from within a background and correlating them with the same GF. Parameters of random simulations are evaluated and an observed number of associations of FOIs with a GF are compared with a Gaussian's distribution of random simulations. A user has the option to generate a file with random features/SNPs and run them as an input to evaluate the performance of Monte-Carlo simulations.

3.2 Implementation

Genome*Runner* was developed using Visual Basic 2010 and the ALGLIB add-on (http://www.alglib.net), and is distributed as an executable program along with source code.

4 RESULTS

As a proof of principle of the potential of Genome*Runner* to discover associations, we ran an enrichment analysis of all 5190 probes recognizing non-coding RNAs on Affymetrix's Human Gene 1.0 ST array. Genome*Runner* found them enriched within genes ($P = 3.04$E-05) but significantly underrepresented in exon regions ($P = 5.25$E-21), confirming recently published findings (Rearick *et al*., 2011). Another positive control was validation of the H3K4me2 -DNAse hypersensitive site association (Birney *et al*., 2007), ($P < 1.00$E-32). More examples of Genome*Runner* analyses are shown in Supplementary Material.

5 CONCLUSION

Genome*Runner* is designed to scan a large set of genomic features, including non-gene regions such as ncRNAs and epigenetic marks, in search of correlations with genomic FOI. The open-ended structure of Genome*Runner* can and will be adapted to be included in the Galaxy suite. Currently, human and mouse genome annotations are available for Genome*Runner*; and new organisms will be added in the near future. Besides automatic correlation of FOIs with known genomic features, Genome*Runner* will calculate enrichment against either the whole genome or a user-defined background. In summary, we provide the scientific community with a tool for automated exploration of statistically significant associations between experimental FOIs and annotated genomic features.

Funding: National Institutes of Health (grant # 5P20RR020143-08 to J.D.W.) and a pilot project subaward (to M.G.D.) to develop GenomeRunner.

Conflict of Interest: None declared.

REFERENCES

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Chen,L.S. *et al.* (2010) Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. *Am. J. Hum. Genet*. **86**, 860–871.

Cline,M.S. and Karchin,R. (2011) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*, **27**, 441–448.

Dennis,G. Jr *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Gauderman,W.J. *et al.* (2007) Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.*, **31**, 383–395.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

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

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Rearick,D. *et al.* (2011) Critical association of ncRNA with introns. *Nucleic Acids Res.*, **39**, 2357–2366.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wang,K. *et al.* (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.

Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–1701.

Wren,J.D. *et al.* (2005) Automating genomic data mining via a sequence-based matrix format and associative rule set. *BMC Bioinformatics*, **6** (Suppl. 2), S2.