BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 11 2011, pages 1589–1590
doi:10.1093/bioinformatics/btr185

Data and text mining

Advance Access publication April 19, 2011

VIZ-GRAIL: visualizing functional connections across disease loci

Soumya Raychaudhuri[1,2,3]

[1]Divisions of Genetics and Rheumatology, Brigham and Women's Hospital, [2]Partners Center for Personalized Genomic Medicine, Boston, MA 02115 and [3]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: As disease loci are rapidly discovered, an emerging challenge is to identify common pathways and biological functionality across loci. Such pathways might point to potential disease mechanisms. One strategy is to look for functionally related or interacting genes across genetic loci. Previously, we defined a statistical strategy, Gene Relationships Across Implicated Loci (GRAIL), to identify whether pair-wise gene relationships defined using PubMed text similarity are enriched across loci. Here, we have implemented VIZ-GRAIL, a software tool to display those relationships and to depict the underlying biological patterns.

Results: Our tool can seamlessly interact with the GRAIL web site to obtain the results of analyses and create easy to read visual displays. To most clearly display results, VIZ-GRAIL arranges genes and genetic loci to minimize intersecting pair-wise gene connections. VIZ-GRAIL can be easily applied to other types of functional connections, beyond those from GRAIL. This method should help investigators appreciate the presence of potentially important common functions across loci.

Availability: The GRAIL algorithm is implemented online at http://www.broadinstitute.org/mpg/grail/grail.php. VIZ-GRAIL source-code is at http://www.broadinstitute.org/mpg/grail/vizgrail.html.

Contact: soumya@broadinstitute.org

Supplementary Information: Supplementary methods and data are available at *Bioinformatics* online.

Received on March 5, 2011; revised on April 2, 2011; accepted on April 4, 2011

1 INTRODUCTION

As genome-wide association studies rapidly identify genetic loci for a broad range of phenotypes, investigators are critically focused on identifying key pathways and biological processes suggested by genetic findings (Iossifov *et al.*, 2008; Lage *et al.*, 2007; Perez-Iratxeta *et al.*, 2007; Rossin *et al.*, 2011; Wang *et al.*, 2007). We have separately described a computational strategy, Gene Relationships Across Implicated Loci (or GRAIL), that uses statistical text-mining strategy to rapidly identify genes across multiple loci that are similar to each other, and to then assess if that degree of similarity is more than might be expected by chance (Raychaudhuri *et al.*, 2009a). The approach depends on pairs of related genes using 525 000 PubMed article abstracts identified using word similarity metrics. GRAIL has now been applied to prioritize SNPs for replication or to demonstrate common function among genes near associated SNPs across a wide range of phenotypes including height (Lango Allen *et al.*, 2010), rheumatoid arthritis (Raychaudhuri *et al.*, 2009b), Crohn's disease (Franke *et al.*, 2010) and cancer (Beroukhim *et al.*, 2010). While the GRAIL statistical approach calculates the statistical significance of the number and strength of functional similarity across loci, it does not concisely illustrate functional similarities in an intuitive fashion that reveals the underlying biology. Our goal was to produce a visualization that allowed users to see more clearly the underlying genes and biological functionality driving the GRAIL statistical scores.

2 IMPLEMENTATION

In order to make the GRAIL algorithm accessible online, we have implemented it at http://www.broadinstitute.org/mpg/grail/grail.php. The online interface is implemented with a PHP script. Users enter genetic loci, typically as a list of SNPs or coordinates of genomic segments, and select a gene similarity metric. Those inputs are then passed onto backend software implemented in Perl and MATLAB®, operating on an LSF cluster farm, and the user is emailed when the job is completed with links to the online results page. Gene similarity metrics can be based on word vector similarity of PubMed text of abstracts referencing genes (Raychaudhuri, 2006), of gene expression vector similarity in a gene expression database (Su *et al.*, 2004) or similarity in gene ontology terms (Ashburner *et al.*, 2000). The infrastructure is flexible and can allow for defining alternative similarity metrics in the future.

In order to effectively display results, we implemented the VIZ-GRAIL software in Perl that interacts with the GRAIL online site to download the results of user-defined analyses jobs, and to construct high-quality display graphics of interactions between genes across phenotypically associated loci. The different loci are arranged in a circle, similar to Circos genome plots (Krzywinski *et al.*, 2009). Genetic loci are arranged around an outer circle, while genes within them are grouped together in an inner circle. Lines are drawn between functionally similar genes that are within different loci. The VIZ-GRAIL program determines the thickness of lines between two genes to be proportional to the relative similarity of the two genes, and inversely proportional to the number of genes within the loci that the genes are derived from, in order to account for the possibility of spurious connections (see Supplementary Material).

Critical to displaying connections between loci clearly is the particular arrangement of the loci and genes around the circle. Often lines connect specific subsets of loci together. If those subsets are not carefully arranged around the circle, then many intersecting connections obscure biologic intuition. Therefore, as a key part of our software we have implemented an optimization procedure that minimizes the total burden of intersecting connections in

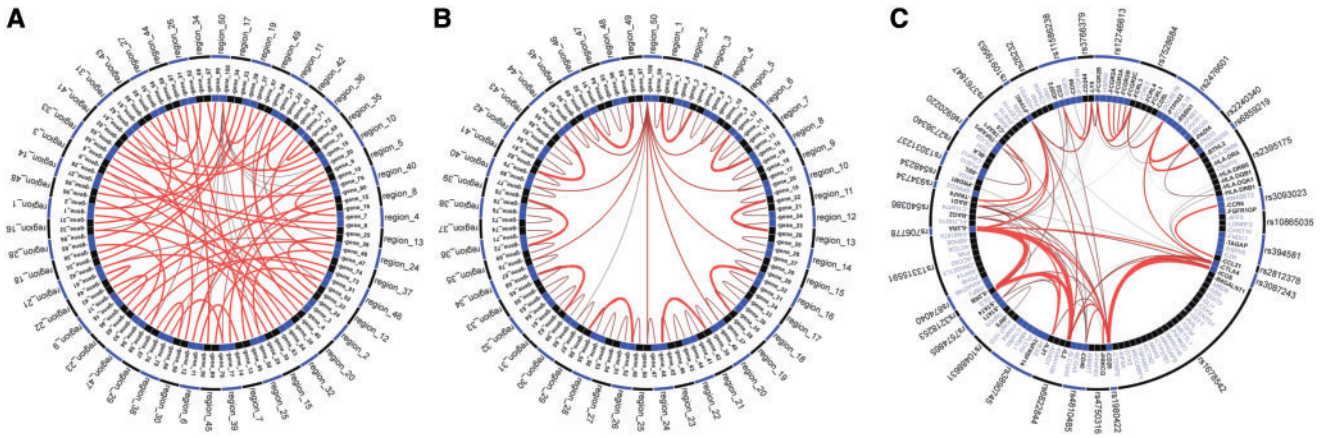

Fig. 1. Examples of GRAIL-VIZ plots. For each plot, genomic regions are arranged along the outer circle alternating colors. Inner circle represents the individual genes. The redness and thickness of lines connecting pairs of genes represents the strength of the connections. (**A**) A plot of 50 regions and 100 genes randomly scrambled. (**B**) The same plot as in (A) but GRAIL-VIZ has been used to optimize ordering of genes and loci to minimize any intersecting of lines. (**C**) Published RA risk loci plotted using the GRAIL-VIZ (Raychaudhuri, 2010).

the figure (see Supplementary Material). Briefly, we define an objective function that calculates the total burden of intersections, weighing intersections between thicker connections more heavily. Then we iteratively chose random loci with at least one gene with an intersecting connection, and then we try manipulating the arrangement by either (i) moving the locus to each of the different positions in the circle, (ii) swapping the locus with every other locus in the circle or (iii) inverting different segments of the circle starting from that locus and ending at other positions. At each iteration, we chose the manipulation that most reduces the total number of intersecting connections and update the arrangement iteratively. Once the loci have been arranged, then genes within each of the loci are permuted to reduce the number of total intersections.

3 EXAMPLES

We present examples of VIZ-GRAIL runs in Figure 1. The files used to create this figures are provided in the Supplementary Material. Figure 1A and B presents an illustrative example. In this case, there is a single optimal solution without any intersecting connections. In Figure 1A, the regions and genes are plotted without arranging to minimize intersections—the display looks jumbled and it is difficult to see any clear patterns. VIZ-GRAIL is able to find the optimal arrangement in 184 iterations run on a personal laptop in <1 h (Fig. 1B); the connections between genes and loci are much more clear. As a realistic example, we plot the literature-based similarity across 34 known rheumatoid arthritis (RA) risk loci, implicating a total of 132 genes (Raychaudhuri, 2010). After using VIZ-GRAIL to arrange genes and loci to minimize intersections, related genes are clearly seen, for example the genes involved in the *IL2* pathway (*IL2*, *IL2RA* and *IL2RB*), as well as the *CD28-CTLA4* pathway.

ACKNOWLEDGEMENTS

We acknowledge the generous institutional support from Brigham and Women's Hospital and the Broad Institute. We also acknowledge the feedback from Mark Daly, Andre Frank and the GRAIL user community.

Funding: National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH-NIAMS) Development Award (1K08AR055688).

Conflict of Interest: none declared.

REFERENCES

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

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Iossifov,I. *et al.* (2008) Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res.*, **18**, 1150–1162.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lango Allen,H. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Perez-Iratxeta,C. *et al.* (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res.*, **35**, W212–W216.

Raychaudhuri,S. (2006) Computational Text Analysis for Functional Genomics and Bioinformatics. Oxford University Press, Oxford.

Raychaudhuri,S. (2010) Recent advances in the genetics of rheumatoid arthritis. *Curr. Opin. Rheumatol.*, **22**, 109–118.

Raychaudhuri,S. *et al.* (2009a) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.

Raychaudhuri,S. *et al.* (2009b) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat. Genet.*, **41**, 1313–1318.

Rossin,E.J. *et al.* (2011) Proteins encoded in genomic regions associated to immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**, e1001273.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. US A*, **101**, 6062–6067.

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