Genetics and population analysis

geneAttribution: trait agnostic identification of candidate genes associated with noncoding variation

Arthur Wuster\textsuperscript{1,2,*}, Diana Chang\textsuperscript{1}, Timothy W. Behrens\textsuperscript{1} and Tushar R. Bhangale\textsuperscript{1,2}

\textsuperscript{1}Department of Human Genetics and \textsuperscript{2}Department of Bioinformatics and Computational Biology, Genentech Inc, South San Francisco, CA 94080, USA

*To whom correspondence should be addressed.

Received on August 30, 2016; revised on October 10, 2016; editorial decision October 31, 2016; accepted on November 2, 2016

Abstract

Motivation: We have developed geneAttribution, an R package that assigns candidate causal gene(s) to a risk variant identified by a genetic association study such as a GWAS. The method combines user-supplied functional annotation such as expression quantitative trait loci (eQTL) or Hi-C genome conformation data and reports the most likely candidate genes. In the absence of annotation data, geneAttribution relies on the distances between the genes and the input variant.

Availability and Implementation: The package is freely available from http://www.bioconductor.org/. A quick-start vignette is included with the package.

Contact: wustera@gene.com

1 Introduction

The majority of variants identified by genetic association studies in humans are located in noncoding regions and likely act by affecting gene expression (Maurano et al., 2012). In addition, variants typically contain multiple genes in their region of linkage disequilibrium, meaning that a naïve approach that simply designates the closest gene as the best candidate can be problematic.

Given a genomic coordinate marking a variant identified by a genetic association study, geneAttribution aims to compute the relative probability for each of the genes in the vicinity of the variant. As a first step, the algorithm considers genes that are closer to the variant to be more likely candidate genes than those that are further away. To correctly calibrate the relationship between variant-gene distance and candidate gene probability, we used eQTLs from 36 tissues identified by the Genome-Tissue Expression Project (GTEx Consortium, 2015). We noted that the distance between the eQTLs and the genes they regulate can be approximated by an exponential distribution, with more eQTLs close to the gene than further away. An exponential distribution with $\lambda = 7.61 \times 10^{-6}$ fits the distribution of eQTLs both to the 5’ and the 3’ end of genes (Fig. 1A).

The geneAttribution algorithm can also take into account other user-supplied empirical data that links genomic regions to genes. Examples include expression quantitative trait loci (eQTLs) that link variants to the expression of specific genes (Fig. 1B) and Hi-C genome conformation data linking distal genomic regions to the promoters of specific genes (Fig. 1C). geneAttribution incorporates empirical data by first determining if any of the genomic regions specified in the empirical data overlap with the input variant. If there is no overlap, the empirical data are ignored. However, if there is an overlap, the likelihood of the associated gene is multiplied by a weight associated with the empirical data, which is also user-supplied. Finally, geneAttribution normalizes candidate gene likelihoods by dividing by the sum of the individual likelihoods (Fig. 1D).

More formally, we define the candidate probability as:

$$P(g|v) = \frac{P(v|g) \cdot P(g)}{\sum_{k=1}^{n} P(v|g_k) \cdot P(g_k)}$$

$$P(v|g) = \lambda \cdot e^{-\lambda d_i}$$
the number of empirical datasets and empirical data type was based on high-resolution Capture Hi-C data sent in more than one of the 36 GTEx tissues were used. The second used. To make our approach more tissue agnostic, only eQTLs pre-

eQTL variants themselves, we defined a region around each eQTL type was based on eQTLs from GTEx. Instead of only using the

deletion sites with a recombination rate of more than 0.1 cM/Mb to de-

**3 Conclusions**

While we limited ourselves to Hi-C and eQTL data, users can easily also include any other empirical data linking genomic regions to candidate genes when using geneAttribution. These data types could either be generic or tissue or condition specific. An example of the latter would be ChIP-sequencing data for transcription factors that are relevant to the trait being studied.

In summary, geneAttribution is a tool for the identification of likely candidate genes for variants of interest. While it is not a substitute for in-depth experimental validation, it can quickly suggest candidate genes without the need to first collect trait or disease specific data.

**Acknowledgements**

We would like to thank our colleagues Brian Yaspan and William Forrest for their input and comments on the algorithm and the manuscript.

**Conflict of Interest:** All authors are employees of Genentech Inc.

**References**

GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot ana-

lysis: multitissue gene regulation in humans. Science, 348, 648–660.

International HapMap Consortium (2005) A haplotype map of the human genome. Nature, 437, 1299–1320.

Maurano,M.T. et al. (2012) Systematic localization of common disease-

associated variation in regulatory DNA. Science, 337, 1190–1195.

Mifsud,B. et al. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet., 47, 598–606.

Okada,Y. et al. (2013) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 506, 376–381.

Welter,D. et al. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 42, D1001–D1006.