Enrichment in the genome

Jeff Leek

@jtleek

www.jtleek.com

Enrichment, not just for genes!

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888

RESEARCH ARTICLE

Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS

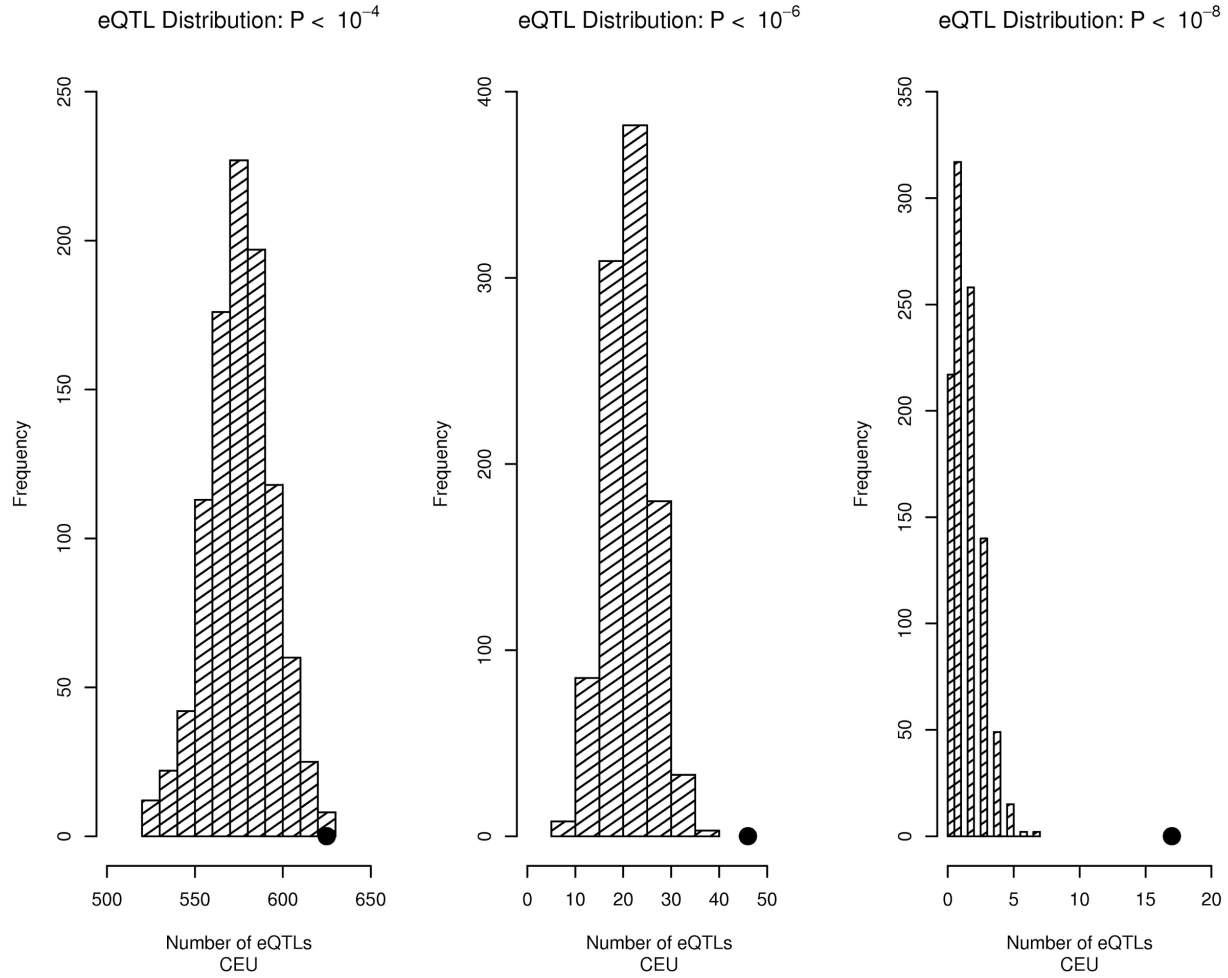
Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, Nancy J. Cox

Published: April 1, 2010 • DOI: 10.1371/journal.pgen.1000888

Abstract

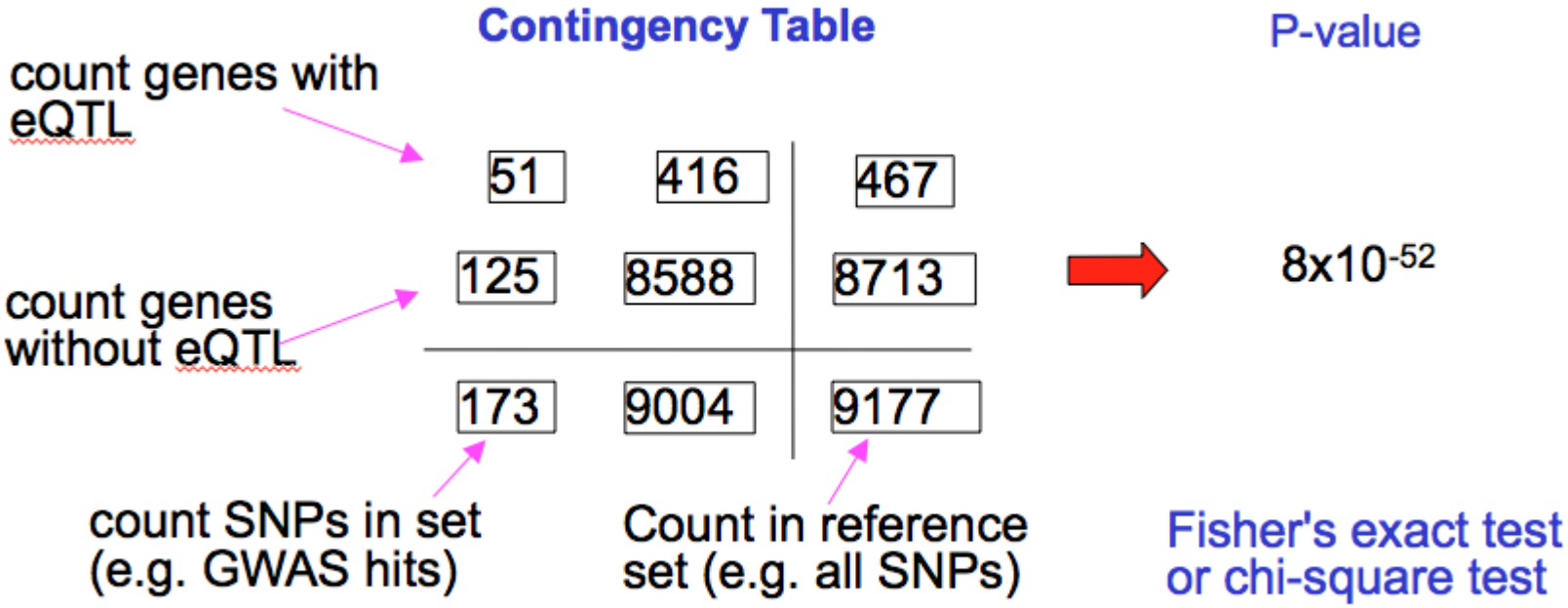
Although genome-wide association studies (GWAS) of complex traits have yielded more reproducible associations than had been discovered using any other approach, the loci characterized to date do not account for much of the heritability to such traits and, in general, have not led to improved understanding of the biology underlying complex phenotypes. Using a web site we developed to serve results of expression quantitative trait locus (eQTL) studies in lymphoblastoid cell lines from HapMap samples (http://www.scandb.org), we show that single nucleotide polymorphisms (SNPs) associated with complex traits (from http://www.genome.gov/gwastudies/) are significantly more likely to be eQTLs than minor-allele-frequency–matched SNPs chosen from high-throughput GWAS platforms. These findings are robust across a range of thresholds for establishing eQTLs (p-values from 10^{-4}–10^{-8}), and a broad spectrum of human complex traits. Analyses of GWAS data from the Wellcome Trust studies confirm that annotating SNPs with a score reflecting the strength of the evidence that the SNP is an eQTL can improve the ability to discover true associations and clarify the nature of the mechanism driving the associations. Our results showing that trait-associated SNPs are more likely to be eQTLs and that application of this information can enhance discovery of trait-associated SNPs for complex phenotypes raise the possibility that we can utilize this information both to increase the heritability explained by identifiable genetic factors and to gain a better understanding of the biology underlying complex traits.

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888



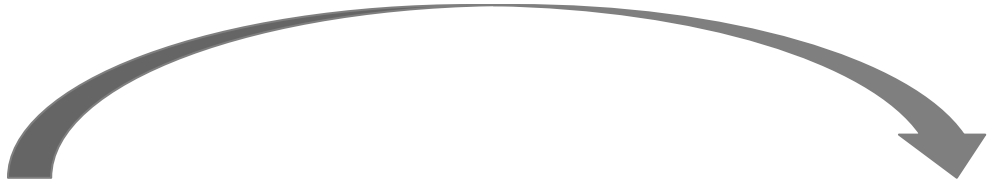
Association between genomic features

http://cals.arizona.edu/~anling/MCB516/lecture21.pdf



Permutation

- This time permute genomic features
- Hard to get right!



Response	R	R	•••	NR	NR
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

Response	NR	R	•••	NR	R
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

To get the null right

Account for genomic correlations

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Exploring Massive, Genome Scale Datasets with the GenometriCorr Package

Alexander Favorov, Loris Mularoni, Leslie M. Cope, Yulia Medvedeva, Andrey A. Mironov, Vsevolod J. Makeev, Sarah J. Wheelan

Published: May 31, 2012 • DOI: 10.1371/journal.pcbi.1002529 • Featured in PLOS Collections

Abstract

We have created a statistically grounded tool for determining the correlation of genomewide data with other datasets or known biological features, intended to guide biological exploration of high-dimensional datasets, rather than providing immediate answers. The software enables several biologically motivated approaches to these data and here we describe the rationale and implementation for each approach. Our models and statistics are implemented in an R package that efficiently calculates the spatial correlation between two sets of genomic intervals (data and/or annotated features), for use as a metric of functional interaction. The software handles any type of pointwise or interval data and instead of running analyses with predefined metrics, it computes the significance and direction of several types of spatial association; this is intended to suggest potentially relevant relationships between the datasets.

Availability and implementation: The package, GenometriCorr, can be freely downloaded at http://genometricorr.sourceforge.net/. Installation guidelines and examples are available from the sourceforge repository. The package is pending submission to Bioconductor.

Figures

Notes and further reading

- Getting the null right is very hard
- It is easy to “tell stories” if you aren’t careful
- Incurs a second multiple testing problem