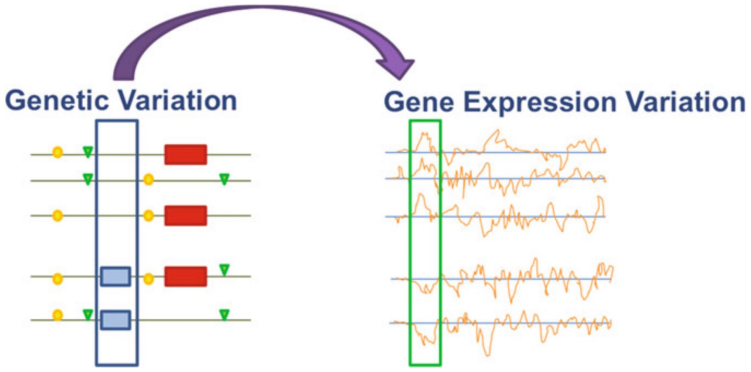
eQTL

Contents

1 Introduction

Expression quantitative trait loci (eQTL) are genomic loci that explain variation in expression levels of mRNAs. By using the expression of mRNA transcripts for genes as a phenotype of interest, one may quantify on a continuous scale the effect of regulatory variation by calculating the correlations between genetic differences and respective transcript abundance or probe intensity values.



Typically, eQTLs may be annotated as single nucleotide polymorphisms (SNPs), large structural variants (SVs), or copy number variants (CNVs).

Trans-eQTLs If genome-wide all markers are tested for association with a gene, this is called trans-eQTL. *Cis-eQTLs* In case of a more regional test (using only markers that are up to 2 megabases (MB) away from the gene of interest) the analysis is called cis-eQTL. *Linkage-based eQTL mapping* strategy will utilize either inbred strains or families to search for eQTLs, relying on recombination events between genes and genetic loci with potential regulatory effects. *Association eQTL mapping* strategy involves a large and heterozygous population-level analysis.

1.1 Software for Confounder Correction and Normalization

Normalization addresses concerns of the technical differences in the experimental design and execution among samples.

EMMA: EMMA provides a computationally efficient method to reduce the effect of population substructure when performing an analysis on using inbred strains. **ICE:** ICE similarly may be used to correct for the heterogeneous makeup of the samples used for an eQTL study. **SVA:** This package can correct for the population structure of the datasets being studied by derived surrogate variables. The SVA package provides ComBat as an additional functionality, a method to adjust for batch effect using an empirical Bayes' framework. **AffyGG:** AffyGG is a package that checks for and eliminates deviating probes in Affymetrix microarrays.

2 Performing QTL and eQTL Analyses

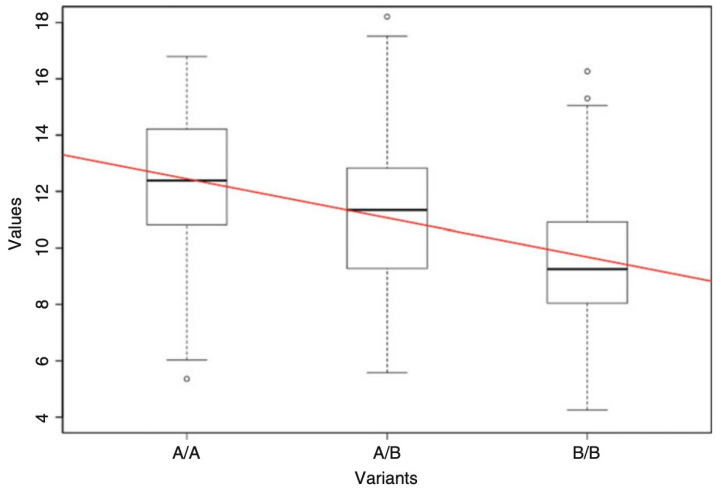
2.1 The Linear Model Approach

One approach to link an eQTL back to a disease (or any other trait) is to perform first a differentially gene expression analysis to identify genes that have different expression profiles between two trait groups and then perform an eQTL analysis for these genes in order to identify possible SNPs that are affecting the variation of the observed expression profile. Take the following table:

SNP	Loci	Sequence	Allele A	Allele B
rs234534	Chr15:97,028,311	CTGAGGA[A/G]GAAAAAT	A	G

[A/G] indicates that at this SNP, there are two different Alleles, namely A (Allele A) or G (Allele B). In case that either Allele A or Allele B is observed on both strands, the variant is called homozygous and is labelled as A/A or B/B, respectively. In case Allele A is present on one strand and Allele B on the other, the variant is called heterozygous and is labelled as A/B. These three different groups are also often marked as 0 (A/A), 1 (A/B), and 2 (B/B). These three groups are called here "variant group".

If we consider now a study population and we are interested in correlating an SNP with either a phenotype (QTL analysis) or a gene expression (eQTL analysis), we assign our study population for each SNP into one of the three corresponding variant groups. As some coding inventions, for example, if p(3,2) = 23 and g(1,45) = 1.32 it means that phenotypes value for phenotype #2 for individual #3 is measured as 23 and for individual #1 the gene expression for gene #45 is measured as 1.32. Finally, v(3,2) = 2000 means that the genomic loci of the variant #2 for individual #3 is at 2000.



In the above figure, in addition to the raw data, we also added in red a regression line:

$g = av + b$

where a and b are the coefficients to be estimated using the Best Linear Unbiased Estimator (BLUE), which under certain assumptions, is the Ordinary Least Square (OLS) estimator. The R2 value gives the goodness-of-fit of this line with respect to the data. In case of a perfect match (all points lay on the line) it is R2 = 1, and 0 if the line has no explanatory power. The goodness-of-fit value R2 is equal to the squared Pearson correlation between g and v, called P(g,v). In an eQTL analysis, the following hypotheses are tested:

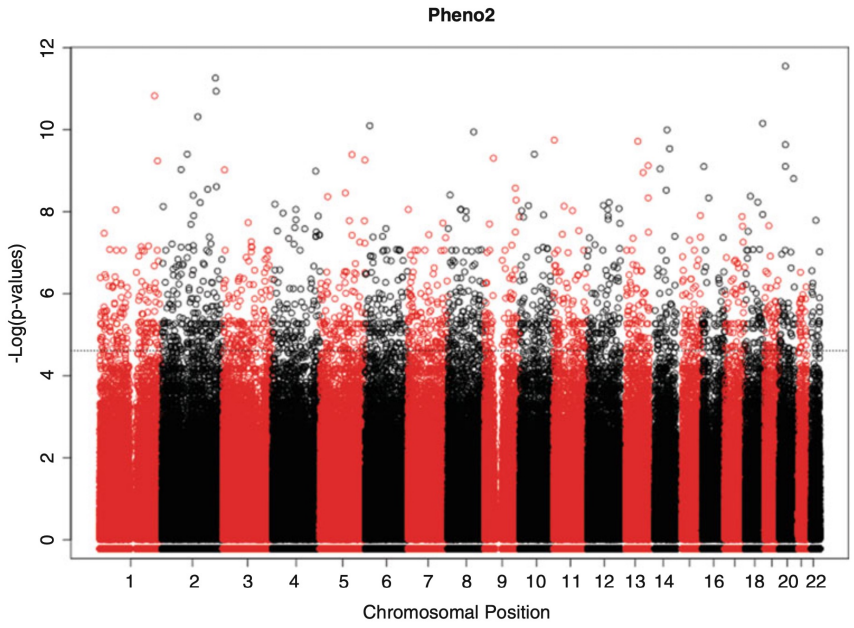
H0 :a = 0 vs: H1 :a != 0

2.2 Directional Test

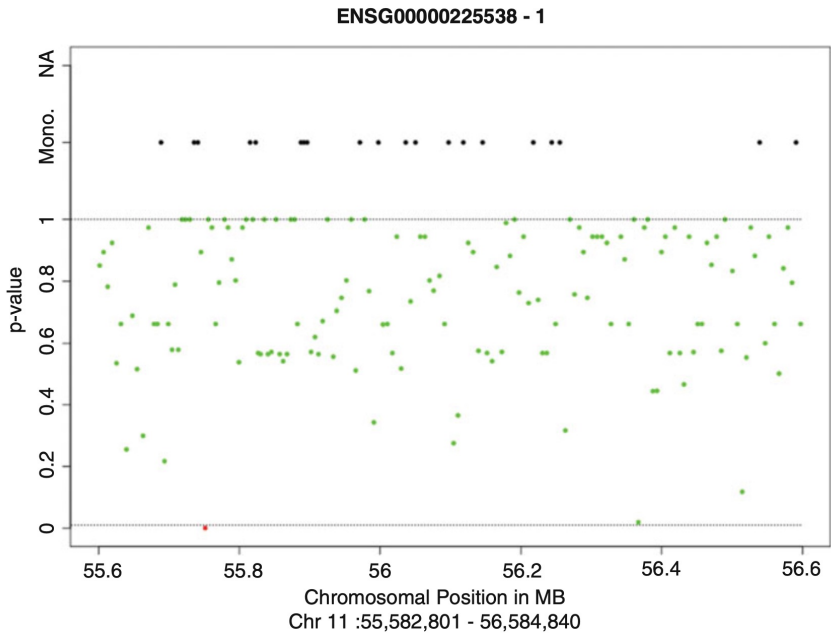
This method is a non-parametrical approach based on a generalization of the Mann–Whitney test. In analogy to the connection between the linear model and the Pearson correlation, we could in that case also test if the Kendall tau correlation between the variant and the phenotype/gene expression equals zero or not.

2.3 From Single Tests to eQTL Analyses

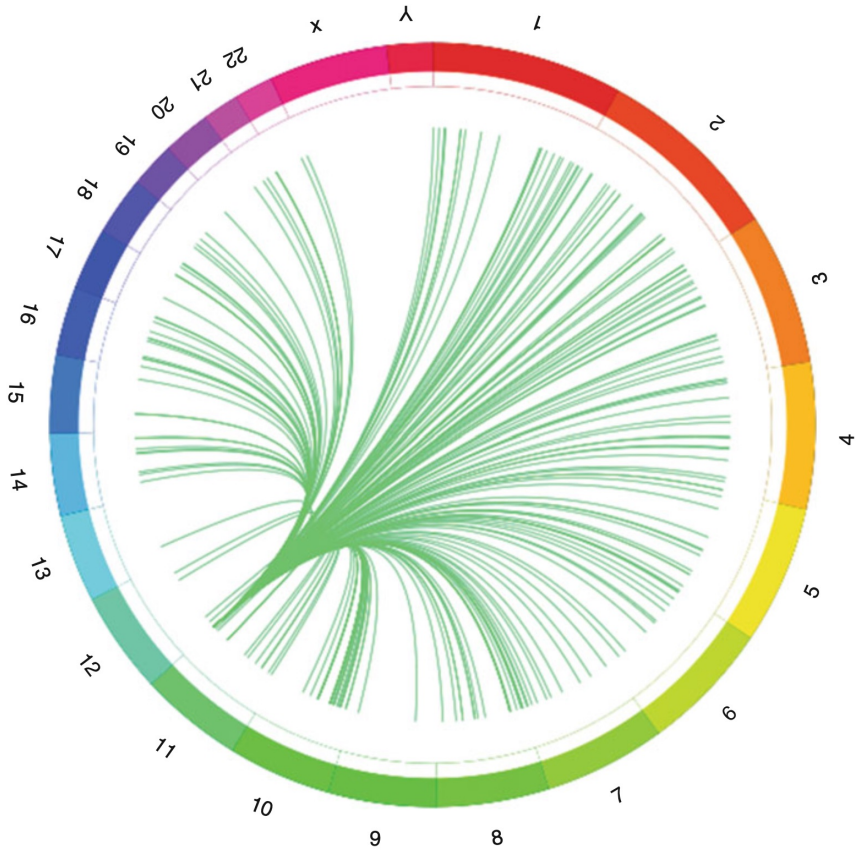
In case of a QTL, usually all available loci are tested against all available phenotypes and the level of association is then typically visualized in a Manhattan plot, where the x-axis represents the loci on the genome and the y-axis represents the level of association:



The typical visualization for cis-eQTL is also a Manhattan-like plot:



The most common trans-eQTL visualization is done using a circular representation:



2.4 Multiple Testing

Here, the familiar candidates like the FDR/Benjamini–Hochberg or the Bonferroni correction can be applied. However, as these adjustment methods are maybe too conservative, in practice it might be better to not adjust for multiple testing and identify interesting hotspots and then confirm these with an independent experiment in the lab.