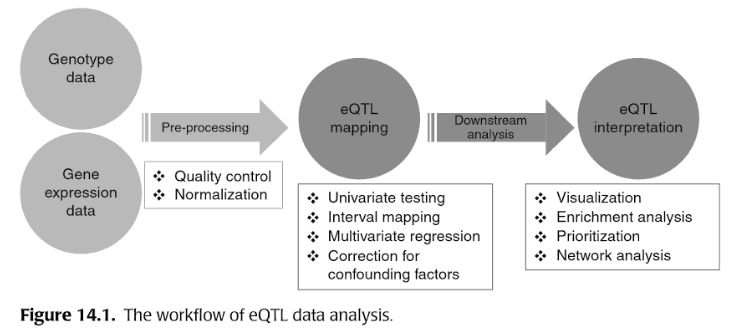
eQTL mapping

1.Introduction

With an influx of successful genome-wide association studies to identify genetic variations associated with complex diseases, an unprecedented wealth of knowledge has been accumulated for SNP-phenotype associations (McCarthy et al., 2008; Witte 2010; Manolio 2013). However, many SNP-disease associations do not lend themselves to molecular interpretations, because many of the identified loci are located outside of the coding regions. Even when a gene can be inferred to be causal, there is often a significant gap towards the understanding of the underlying molecular mechanisms (Schadt et al., 2005; McCarthy et al., 2008). Genome-wide eQTL mapping has been one effective approach to bridge this gap (Mackay et al., 2009). In eQTL studies, gene expression levels measured by high-throughput technologies, such as microarrays and RNA-Seq, are treated as quantitative traits. Marker genotypes are also collected from the same set of individuals, and statistical analyses are performed to detect associations between markers and expression traits. By simultaneously capturing many regulatory interactions, eQTLs offer valuable insights on the genetic architecture of expression regulation (Rockman and Kruglyak 2006). The ultimate goal of eQTL studies is to elucidate how genetic variations affect phenotypes by using gene expression levels as intermediate molecular phenotypes (Nica and Dermitzakis 2008). In this chapter, we provide an overview of the eQTL analysis workflow (Figure 14.1), introduce publicly available tools for analysis, and further discuss challenges and issues.

2. Data pre-processing

Genome-wide eQTL mapping considers high-density SNP genotype data and gene expression data from the same individuals in a segregating population. Both require appropriate pre-processing as described below for subsequent analysis.



2.1 Genotype data

Three quality control (QC) criteria are often used in the pre-processing of the genotype data. (1) Missing rate: individuals with a large proportion of missing SNP genotypes (e.g., 10%) should be excluded because the DNA samples of those individuals may be of poor quality. SNPs with a large missing rate (e.g., 5%) should also be filtered out. (2) Hardy-Weinberg Equilibrium (HWE): statistically significant deviations from HWE often result from genotyping errors. Therefore, SNPs that fails an exact HWE test (e.g., a P-value less than 0.001) should be filter out. The criterion does not apply to haploid organisms, such as yeast. (3) Minor allele frequency (MAF): SNPs with low MAF (e.g., 0.05) are sometimes filtered out because of the insufficient statistical power for studies with a relatively small sample size and potentially higher genotype calling error. All the QC measures mentioned above can be accomplished by PLINK (Purcell et al., 2007) for human studies and by R package qtl for animal studies (Broman et al., 2003). After the QC, the remaining missing genotype data need to be carefully handled. For study samples with a reference population panel, missing genotype data may be imputed using a phasing algorithm such as BEAGLE (Browning and Browning, 2007), or IMPUTE2 (Delaneau et al., 2012)