Gene Expression Studies and Complex Diseases 5

Harald H.H. Göring

5.1 Introduction

The human genome contains a large number of genes, each of which may be viewed as a specific genomic segment that encodes information for one or several defined functions. Parts of the DNA sequence of a gene are transcribed into RNA that is then translated into protein. The vast majority of genes encode one or several proteins, and the intermediate RNA product is therefore referred to as messenger RNA, or mRNA. However, there are a number of genes producing non-coding RNA molecules, such as ribosomal RNA (rRNA), transfer RNA (tRNA), or micro RNA, among others, where the RNA molecules have a variety of functions by themselves. All forms of RNA molecules are the subject of gene expression studies but different technologies may be required to investigate different types of RNA.

Figure 5.1 provides a basic overview of the central information flow in biology and the various analytical techniques and areas of genetic and epidemiological investigation related to it. The main focus of human genetic epidemiological research is to identify the genes, and their variants, which influence our individual characteristics, with the most emphasis (and money) directed toward diseases and other clinically relevant traits. The statistical methods for correlating genotypes and phenotypes are referred to as linkage analysis and association analysis. For this type of analysis, genotype data must be generated. Over the last several years, aided by astounding progress in genomic and other laboratory technologies, a variety of additional approaches have gained popularity to investigate the genetic etiology of human diseases and their pathology. These approaches are complementary to genotype-based linkage and association analysis (and to each other), providing additional information that can be used to understand the biology of human conditions. These approaches include correlation analysis between a trait of interest and gene expression levels. This analysis requires quantification of gene expression levels rather than genotyping. Similar techniques include proteomic (Kooij et al. 2014; Van Eyk 2011) and metabolomic profiling (Kettunen et al. 2012; Sreekumar et al. 2009; Tukiainen et al. 2012) (or methylomic profiling (Mill and Heijmans 2013), which involves assessment of DNA methylation status; not shown in Fig. 5.1). In each case, the goal is to correlate a trait of interest to measured levels of transcripts, proteins, or metabolites, in order to identify any processes that are connected, in some manner, to the trait of interest.

H.H.H. Göring (✉)
Department of Genetics, Texas Biomedical Research
Institute, PO Box 760549, San Antonio TX
78245-0549, USA
e-mail: hgoring@txbiomedgenetics.org

H.H.H. Göring
Department of Genetics, Texas Biomedical Research
Institute, 7620 NW Loop 410, San Antonio TX
78227, USA

R. Duggirala et al. (eds.), *Genome Mapping and Genomics in Human and Non-Human Primates*,
Genome Mapping and Genomics in Animals 5, DOI 10.1007/978-3-662-46306-2_5,
© Springer-Verlag Berlin Heidelberg 2015

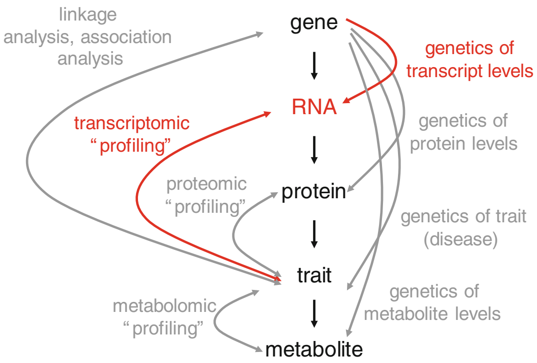

Fig. 5.1 Central information flow in biology and analytical techniques and investigations in genetic epidemiology. This chapter focuses on the investigations involving gene expression data, highlighted in *red*

It is now possible to simultaneously quantify the abundance of essentially all transcripts in a tissue sample (or even a single cell) using modern genomic technologies. The characterized transcriptome can then be used for two main purposes: First, the abundance of individual transcripts (or sets of transcripts) can be correlated with a trait of interest in order to identify those that are significantly correlated with the trait. The genes encoding these transcripts may possibly be involved in the etiology of the trait, and/or it may be that the trait in turn has an impact on the expression levels of these genes. I will refer to this type of investigation as transcriptional correlation analysis, transcriptional profiling, transcriptomic profiling, or gene expression profiling. Second, each transcript may be viewed as a trait whose genetic regulation can be investigated by statistical genetic technologies, in order to identify genomic variants that influence the transcriptional activity of the gene being examined. This second type of investigation is sometimes referred to as "genetical genomics" (de Koning and Haley 2005; Jansen and Nap 2001), a terminology which I view as confusing and which I will not use here. Both of these investigations involving gene expression data—transcriptional profiling and genetic regulation of gene expression—are the topics of this chapter.

Another way to describe the central analyses involving gene expression data in genetic epidemiological research is shown in Fig. 5.2. There are three central information sources available to us—trait phenotypes, gene expression levels, and genotypes—and these permit three types of correlations to be analyzed. The traditional association analysis (or linkage analysis) investigates the relationship between trait phenotypes and genotypes at polymorphic variants (shown on the left side of Fig. 5.2). Assuming that the genetic variants influencing a trait of interest exert their effect via modulation of gene expression, gene expression data may be viewed as an intermediate trait between genotype and clinical outcome of interest, and the overall correlation between genotype and trait phenotype may be viewed as

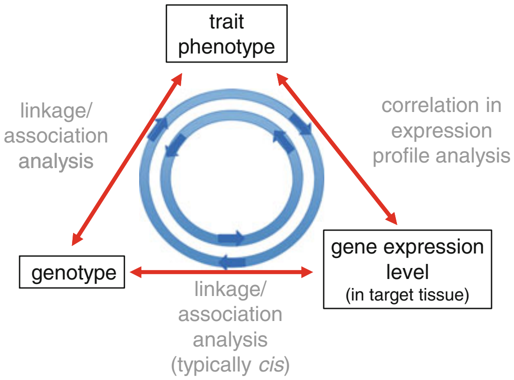

Fig. 5.2 Overview of analyses relating trait phenotypes, gene expression levels, and genotype data. Note that the circle of analyses is ideally conducted on a single sample, which requires that all three types of data are available. Alternatively, multiple different samples may be used if necessary. The analyses can be conducted in both directions

an amalgamation of the correlation between genotype and gene expression level and between gene expression level and the trait phenotype. Again, these are the two types of investigations involving gene expression data commonly undertaken, here referred to as transcriptional profile analysis and analysis of the genetic regulation of gene expression, respectively, as mentioned previously.

Perhaps the main motivation behind transcriptional investigations at the present time is the observation that most of the genetic variants that are significantly associated with complex traits (as typically identified in genome-wide association studies, or GWAS) do not alter the amino acid sequence of proteins or have any other obvious functional effect (Hindorff et al. 2009; Visscher et al. 2012; Welter et al. 2014). In many cases, the associated variants are located outside of genomic regions known to be part of a gene. This leads to the speculation that the variants underlying complex traits are often regulatory in nature. This is in contrast to the genetic defects underlying many Mendelian disorders, many of which directly impact the protein sequence, thereby altering or abolishing protein function. While the activity of genes is regulated at many different levels, including at the stage of transcription, translation, and post-translationally by modification of proteins, transcriptional regulation is a critical component, and the abundance of transcripts can be assessed more readily and more comprehensively than the amounts of proteins and their modifications, due to the pairing potential of the building blocks of DNA and RNA, which is what makes amplification via PCR possible. It is important to note that transcriptional regulation itself is a highly complex process, with multiple potential stages of regulation, including in the location, timing, and speed of transcription, in the decay of transcripts, in the usage of alternative promoters, transcriptional stops, and different splice sites (which lead to the existence of potentially many different transcripts per gene), in other RNA editing processes, and at other steps which are not yet understood. When gene expression level is measured using some molecular technique in the laboratory, we often do not know whether it is a higher transcription rate or a lower rate of decay (or both) that is behind the observed high abundance of a given transcript. Fortunately, this does not complicate the statistical analyses per se, but this uncertainty must be taken into account when interpreting the results. Note also that terms such as "gene expression level" or "transcript level" are often not well-defined in manuscripts, with sloppy wordage widespread. In most cases, what is measured is the abundance of RNA detected by some specific probe (as in microarray studies) or the number of RNA sequencing (RNAseq) reads belonging to a gene or parts of a gene (such as an exon). Quantifying specific transcripts is actually very challenging even now and is not routinely undertaken in most large-scale transcriptomic studies.

This book chapter focuses on the overall concepts related to gene expression studies. Details on the laboratory and analytical methods of these studies are beyond the scope here. This field of investigation is developing rapidly and gaining in popularity, and I will only focus on some key aspects of these types of studies. The references cited here are only a small selection of the rapidly expanding literature; the reader is encouraged to find additional, and potentially newer or better, references on his/her own.

5.2 Correlation Analysis Between Trait Phenotypes and Expression Levels

As mentioned, gene expression data permits us to correlate trait phenotypes with gene expression levels. Those genes whose expression level is significantly correlated (after appropriate multiple testing correction) with the trait of interest are presumably somehow related to that trait, and the set of genes may shed light on the biological pathways related to trait etiology or physiology. The large literature on complex trait gene mapping studies, in particular GWAS studies (Hindorff et al. 2009; Visscher et al. 2012; Welter et al. 2014), clearly shows that individual common variants associated with the risk of a complex disease can be localized. However, it should not be forgotten that this is a challenging undertaking even for common variants, typically requiring sample sizes that were unimaginable only a few years ago. The reason is that these variants individually only modestly influence the risk (of some disease) for a person, and at a population level account for only a small fraction of trait heritability. Common variants with a large influence on penetrance are very rare for complex traits. One example is ApoE4 and Alzheimer's disease (Corder et al. 1993; Strittmatter et al. 1993). It seems likely that for many complex diseases low frequency variants of high penetrance exist (Terwilliger and Göring 2009), but these variants are difficult to localize because of their rarity. Conceptually, transcriptional profile studies could be more powerful. One intuitive reason may be that many functional variants within a gene could be assessed simultaneously using this approach, as long as these variants influence the expression level that can be measured. In effect, the expression level is a read-out that combines the effects of all regulatory variants impacting transcription, and it might therefore be easier to identify genes related to a trait of interest. At least that is the argument that is often made (It may also be the case that the multiple testing burden is somewhat reduced compared to genome-wide analysis of sequence variants, but this may no longer hold as our molecular techniques for gene expression characterization become ever better, leading to discovery of many alternative transcripts per gene).

Despite the general promise and potential, correlation studies between a clinical trait and transcript data have substantial drawbacks. For illustration, let us contrast these studies with a linkage or association study that is based on genotype data. When significant evidence of linkage and/or association has been found, it is clear that a causal factor in the etiology of the trait being studied has been localized (assuming that the finding is not a false positive, the approximate risk of which can be gleaned from the obtained significance level). We may not know which gene(s) and which variant(s) in the mapped candidate region are causal, but it is certain that the identified genomic region harbors one or several variants that influence the trait being investigated. The only real concern related to study design is that the study subjects are matched for ethnicity, which really means that the genome-wide allele frequencies are very similar in cases and controls. If this concern is avoided by design from the outset, such as by ascertaining cases and controls from the same ethnically homogeneous population, or even by a variety of statistical control techniques after ascertainment, and if there is no difference in genotyping approach and quality between both groups, then a causal inference is warranted. [As an aside: One more caveat is that the relationship between genotype and trait risk may not be direct. One example is the FTO locus that was first identified as a risk locus for type II diabetes in a case-control study of that disease (Zeggini et al. 2007). It ultimately turned out that it is really an obesity locus (Frayling et al. 2007) that was mapped in the diabetes study because the cases and controls had different average body mass index levels because of the correlation between obesity and diabetes in the population.] Note that in such genetic studies it is not absolutely crucial (though it may help power and be therefore warranted) that cases and controls are matched for other characteristics, such as sex, age, socioeconomic status, or smoking habits. The key reason behind all these characteristics of linkage and association studies is that genotypes are constant throughout life. Ignoring many exceptions here for simplicity, genotypes are the same in all cells of an individual and are independent of environmental factors to which a person is exposed.

The situation is very different in studies relating a trait of interest to transcript abundance. The expression level of a given gene is not the same in all cells of a body, and the expression level is often influenced by factors of the external environment as well as of the internal environment (i.e., the body and its conditions). The ramifications of this are profound. First, if a significant correlation between disease and gene expression has been observed, the cause–effect relationship is not clear. To stay with diabetes as our example, it is possible that significant transcriptional correlates of the disease are involved in the etiology of the trait (as the loci identified in genetic association studies definitely are). It is also possible that the identified genes are themselves influenced in their expression by the disease. And both relationships could exist at the same time. The former would be useful to learn about trait etiology, while the latter would inform about pathophysiology. Upfront, it is not clear whether the identified transcripts "act upstream or downstream" of the studied disease. A prospective study design or other timeline techniques may help to clarify what is cause and what is effect. Perturbation studies, in which gene expression of a sample is taken before and after some type of manipulation, such as exposure to a chemical, may also help in addressing the ordering of the observed significant correlations.

Second, it is possible that confounder variables could explain the observed, statistically significant correlations between trait and transcripts. For example, it is possible that, say, the diabetic study participants take medications that the control individuals do not; or that the diabetics on average eat a different diet than the controls; or that they exercise less than the controls; and so on. All of these differences between cases and controls, individually or jointly, could potentially explain the observed differences in transcript levels, in which case the observed significant correlations would be artifacts caused by confounder variables. It is very difficult to exclude this possibility in transcriptional profile studies. It is advisable to match cases and controls for as many possible confounder variables as possible, or to measure known confounders and subsequently account for their effects analytically. However, the identity of these confounders is often not known or they cannot be measured accurately. A possible solution is to use the transcriptional profile data itself as a way to identify potential confounders. One example of this is to use the expression levels of "indicator genes" to infer the cell type composition of a tissue sample (such as blood) (Gaujoux and Seoighe 2013). A general approach could be implemented based on principal components analysis, and the top principal components (which tag the key sources of covariation in the expression data, many of which may be related to potential confounder variables) could be subsequently regressed out. The difficulty here is that this approach risks removing the very relationship between trait and gene expression that one seeks to identify; in essence, one may throw out the baby with the bathwater.

Third, tissue specificity of gene expression comes into play. In many cases, the appropriate target tissue is not accessible, or it may not be known. In those situations, investigators conduct their study using another source tissue, in the hope that it will serve as a suitable surrogate tissue. However, this is not a generalizable characteristic of different tissues, because it can vary from gene to gene (and potentially from genetic variant to genetic variant) whether two tissues are suitable proxies for one another. This is discussed in detail below.

We have grappled with these types of complications in a study of schizophrenia, where we contrasted the expression profiles from lymphoblastoid cell lines (LCLs) from cases with schizophrenia and controls without the disease (Sanders et al. 2013). While it is perhaps unlikely that differences in expression levels of cell lines are caused by the disease status (or related differences in environment, attributable to disease-related medication or, say, smoking)—after all, these cell lines are far removed from study subjects and their exposures—it is difficult to exclude the possibility that some aspects of the LCLs could vary between cases and controls, independent of disease. For example, could it be that the LCLs of cases and controls were generated in slightly different manners, which may be the cause of observed transcriptional differences? To guard against this, we measured various cell lines characteristics as part of the study and included these variables as covariates. However, the fact remains that using different sets of covariates leads to somewhat different findings, and one cannot know whether all relevant confounders are accounted for.

All these complications that arise in transcriptional correlation studies compared with genotypic correlation studies can be viewed as the difference between genetic epidemiology and epidemiology more generally. It is the former that is unusual. The nature of genetic inheritance endows genetic epidemiological studies with many advantages that are not shared in most areas of epidemiological research. When correlating transcript abundances with clinical phenotypes, we are no longer in the realm of genetic epidemiology, and thus face many systematic challenges that can be difficult to overcome.

5.3 Genetic Regulation of Gene Expression

Instead of correlating gene expression to a trait of interest, as discussed in the previous section, gene expression levels can also be subjected to statistical genetic dissection. The quantitative expression level of a gene, an exon, or a specific transcript may be viewed as a quantitative trait that is under the influence of genetic and environmental influence like any other trait. Linkage and association analyses can therefore be conducted on transcript abundance values in order to localize the genomic regions and variants that influence the amount of a transcript being present in a given sample.

Studies investigating the genetic regulatory machinery influencing gene expression levels have gained popularity for two main reasons: First, they permit us to study the basic biology underlying a key regulatory step in how our genes' activities are controlled. Second, knowledge about which genetic variants are significantly associated with the expression of a particular gene provides clues about the identity of likely functional variants and their regulatory potential. This sort of functional information can be used, potentially along with many other pieces of information, to prioritize which of the variants that were previously identified in a GWAS of a complex disease are most likely to be functional. This information is relevant in particular because of the hypothesis that most of the functional variants underlying complex traits are subtle regulatory variants. Ultimately, laboratory assays will generally be required to prove that a given variant causally influences a trait of interest, but by accumulating different sources of information on each variant, including whether or not (and in which direction and to which degree) it is associated with the expression of a particular gene, we can hone in on the true functional variants with a greater degree of precision, thereby reducing and speeding up the more time-consuming and much more expensive functional assays to be conducted in the laboratory. A genetic variant found to be significantly associated with some expression level is typically referred to as an expression quantitative trait locus (eQTL) or an expression quantitative trait nucleotide (eQTN) in the case of single nucleotide polymorphisms (SNPs). The term regulatory SNP, or rSNP, is also sometimes used (Guo et al. 2014).

Genetic studies of gene regulation have proven to be highly successful in many regards. This is perhaps not surprising because the expression level of a given gene is, after all, a very direct representation of gene action, and the impact of a regulatory genetic variant on gene expression thus could be pronounced. The relationship would certainly appear to be much closer than one would expect to exist between a gene's activity and a complex trait, where any one variant influences disease risk typically by only a very small amount. Thus, one might expect a priori that studying the genetic regulation of gene expression would be a fruitful undertaking. A variety of studies, conducted in families, in twins, and more recently in unrelated individuals, have shown that the vast majority of gene expression levels are significantly heritable (Göring et al. 2007; Nica et al. 2011; Price et al. 2011; Grundberg et al. 2012). This is clear evidence for the existence of genetic regulatory variants and their influence on gene expression levels in the aggregate. Note, however, that the estimated heritabilities for many expression traits are quite modest, similar to the estimates obtained for many complex diseases. This suggests that either there is substantial measurement error in quantifying gene expression levels, and/or that these traits are subject to myriad influences, including by environmental factors (both of the external environment acting upon a person as well as the internal environment, within the body, to which a given cell is exposed). Therefore, gene expression levels are best viewed as being fairly complex traits.

We conducted one of the largest genetic investigations of genome-wide gene expression at the time, measuring gene expression by Illumina microarrays in white blood mononuclear cells in 1,240 randomly ascertained Mexican American family members from around San Antonio, Texas, USA (Göring et al. 2007). A brief review of this study is provided here as an example of a genetic investigation of gene expression. After full processing, 20,413 (43 %) probes out of a total of 47,289 on the microarray detected significant expression at a false discovery rate (FDR) of 0.05. Among the autosomal probes with significant expression, the quantitative expression levels of the vast majority of probes (85 %) were significantly heritable at FDR 0.05. The median heritability estimate was 23 %, with higher heritabilities among RefSeq probes (which are much better designed and annotated, on average). These estimates support the view that gene expression traits are substantially controlled by genetic factors. We subsequently conducted linkage analysis in order to localize major loci influencing the expression traits. As described in the sections below, we broke the genome into two components—the gene locus targeted by a given probe itself and the remainder of the genome. At FDR 0.05, a large number of probes (1,345), though representing only a fairly small proportion of probes (7 %), showed significant evidence of linkage to the structural gene locus. This is clear evidence that genetic variants in and around a given gene, such as in the promoter region, have substantial influence on that gene's expression levels. For some genes, the heritability attributable to the structural gene locus explained virtually all the estimated overall heritability, suggesting that these genes' expression is largely monogenically controlled. The mean effect size of the structural locus was 5 % on average (with a median of 2 %). The structural locus thus appears to account for a substantial proportion of heritability (based on this particular study, the estimated proportion of genetic variance explained by the structural gene locus is in the range of 10–25 %). We largely failed to identify significant eQTLs elsewhere in the genome, far away from the structural gene, suggesting that these distant regulatory genetic factors, while clearly important in the aggregate, have individually very small effect sizes, making them difficult to detect. Later studies, using better molecular technologies, have largely supported our findings, refining estimates and identifying many more significant eQTLs, as described in the following paragraphs.

5.3.1 *Cis* eQTLs

Given the substantial heritabilities of most gene expression levels, the logical next step is to conduct linkage and in particular association analyses in order to localize specific variants that are significantly associated with expression traits. In contrast to studies of most clinically relevant traits, when investigating the genetic regulation of a particular gene there is an obvious genomic candidate region, namely the gene itself and its chromosomal vicinity (see Fig. 5.3). The reason for paying special attention to this small fraction of the genome is that genetic variants near a gene, and in particular in its promoter region, are quite likely to influence that gene's expression level, e.g., by interfering with the binding of proteins required for transcription. Many of these variants presumably act in *cis*. The difference between *cis* and *trans* is shown in Fig. 5.4. By *cis* (from Latin, meaning "on this side") we mean that a given gene expression regulatory variant influences the expression level only on the physical molecule—i.e., the chromosome—on which it resides, but not on the homologous sister chromosome. The reasons are likely structural, i.e., proteins and other factors bind to a particular chromosomal region to initiate, maintain, and regulate gene expression of that chromosomal molecule, and alleles on that entity thus influence the expression of the gene on only that chromosomal copy. Most

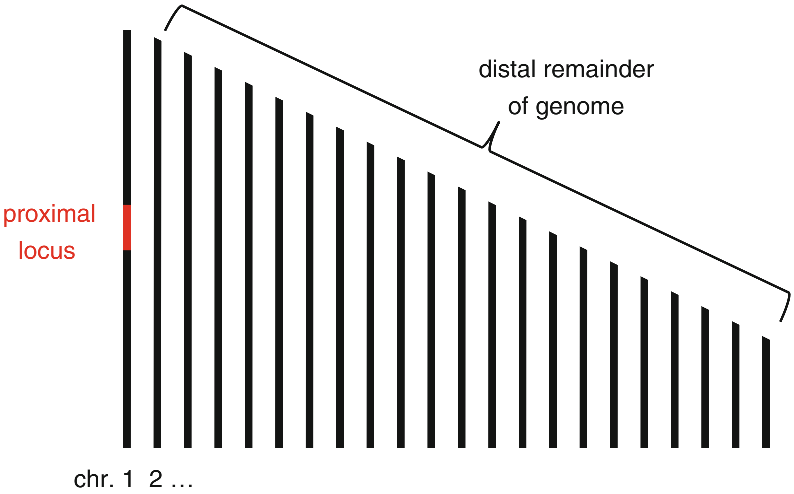

Fig. 5.3 Partitioning of the genome into the proximal locus and the distal remainder of the genome (relative to a gene). When investigating the genetic regulation of gene expression of a particular gene (shown here in *red*, located in chromosome 1), the genome-wide search may be conducted in two parts. The gene itself and its chromosomal surrounding area is a good candidate region to harbor *cis* eQTLs, while the rest of the genome may harbor *trans* eQTLs. Note the enormous differences in search area and associated multiple testing burden

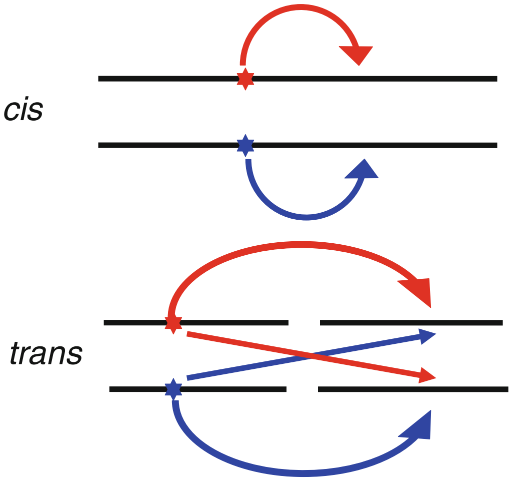

Fig. 5.4 Illustration of *cis* and *trans* effects. An allele acting in *cis* influences only the molecular molecule (chromosome) on which it resides. In contract, *trans* acting factors influence both sister chromosomes

variants located elsewhere in the genome are thought to act in *trans* (from Latin, meaning "on the other side"). These variants influence both chromosomal copies equally. For example, some variant may alter the structure of a transcription factor, which in turn alters the expression of both copies of a given gene—regardless on which sister chromosome a copy of the gene is located. In general, *cis*-acting variants are equated with those close to a given gene, sometimes referred to as proximal variants. Similarly, the term *trans*-acting variant is used for those variants located far away on the same chromosome or on a different chromosome, sometimes called distal or distant variants. While the assumption about how a variant acts based on where it is located relative to a gene will often be correct, it is nonetheless generally only an assumption until confirmed by other means (Gilad et al. 2008). Some variants close to a gene may turn out to be *trans* eQTLs. And some *cis* variants may be located far from a gene but on the same chromosome. (Could it be that *cis* variants may even be located on another chromosome, depending on how chromosomes are packed in three dimensions within the nucleus?) This caveat should be kept in mind when reading the literature, in which these terms are often used interchangeably.

Given the existence of a well-justified candidate region of interest around any gene, one may break a genome-wide search for eQTLs into two parts, one confined to a gene and its surrounding genomic region to search for *cis* variants, and one covering the remainder of the genome to localize *trans* variants (Göring et al. 2007). eQTL studies have proven to be highly successful in the search for *cis* variants (Cheung et al. 2005; Dixon et al. 2007; Göring et al. 2007; Stranger et al. 2007; Emilsson et al. 2008; Montgomery et al. 2010; Pickrell et al. 2010; Grundberg et al. 2012). With gradual improvements in measuring gene expression, and with increased sample sizes, the proportion of genes estimated to be *cis* regulated is creeping upwards, from a small number of *cis*-regulated genes observed at first to perhaps ultimately the majority of genes investigated. It appears likely that every gene is under *cis* regulation to at least some degree, with only power limiting our ability to detect significant proximal associations for all genes. *Cis* eQTLs are therefore frequent, and in fact likely universal for all genes in all tissues. Their impact on quantitative gene expression has frequently been shown to be substantial. In some cases, *cis*-regulatory variants appear to account for much of the estimated heritability of a gene, indicating that the expression level of such a gene is essentially a monogenic trait (though not necessarily influenced by only a single variant). For other genes, *cis*-regulatory effects account for only some proportion of overall genetic variation, suggesting a more complex mode of inheritance with substantial aggregate importance of *trans*-acting variants. *Cis* effects are fairly easy to detect for two reasons: Their commonly strong effect sizes, and because of the limited multiple testing correction that is required when searching for them, because only a very small proportion of the genome, namely the gene and its vicinity, needs to be screened. By now, there are catalogs listing many putative *cis* eQTLs for many genes in many tissues, and these databases are freely available http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi (Yang et al. 2010; Xia et al. 2012).

5.3.2 *Trans* eQTLs

In contrast to the local search for *cis* variants, mapping of *trans* variants requires searching systematically through the entire genome (excluding the small proximal region around a gene). *Trans* eQTL studies have proven to be quite difficult (Cheung et al. 2005; Dixon et al. 2007; Göring et al. 2007; Stranger et al. 2007; Emilsson et al. 2008; Montgomery et al. 2010; Pickrell et al. 2010; Grundberg et al. 2012), perhaps more difficult than some scientists had initially assumed. Part of the explanation is the enormous multiple testing burden incurred when searching the entire genome, especially when one does so on thousands of gene expression traits. Beyond that reason, the difficulty of finding *trans* eQTLs indicates that these regulatory variants individually only influence the expression level of a given gene modestly. In other words, the effect sizes of individual *trans* eQTLs are small. Thus, large sample sizes will be required to localize these variants robustly and comprehensively. At the present time, many studies report a small number of potential *trans* eQTLs, but the evidence is generally fairly weak and large-scale replication studies are still limited (Grundberg et al. 2012). There are some observations suggesting the existence of “master regulators”, i.e., *trans* eQTLs that influence the expression of many genes. This makes intuitive sense if one, e.g., thinks of variants within a transcription factor that is involved in the expression of a whole range of genes. At the present time, much of the supporting data for master regulators is fairly modest, and future work is required to characterize master regulatory eQTLs better.

Figure 5.5 shows an example of a genome-wide joint linkage and association study for a particular gene (*PPA2*), in this case conducted on peripheral blood mononuclear cells from randomly ascertained participants belonging to multigenerational families (Göring et al. 2007). Note that this plot looks very different from a so-called Manhattan plot from a normal GWAS on a complex disease. Here, there is an enormous peak on chromosome 4, which is centered on the exact location where the studied gene is located in the human genome. This signal almost certainly points to *cis* variants in and near the gene, and the magnitude of the signal highlights the substantial effect size of the variants in the proximal gene region. In contrast, the remaining

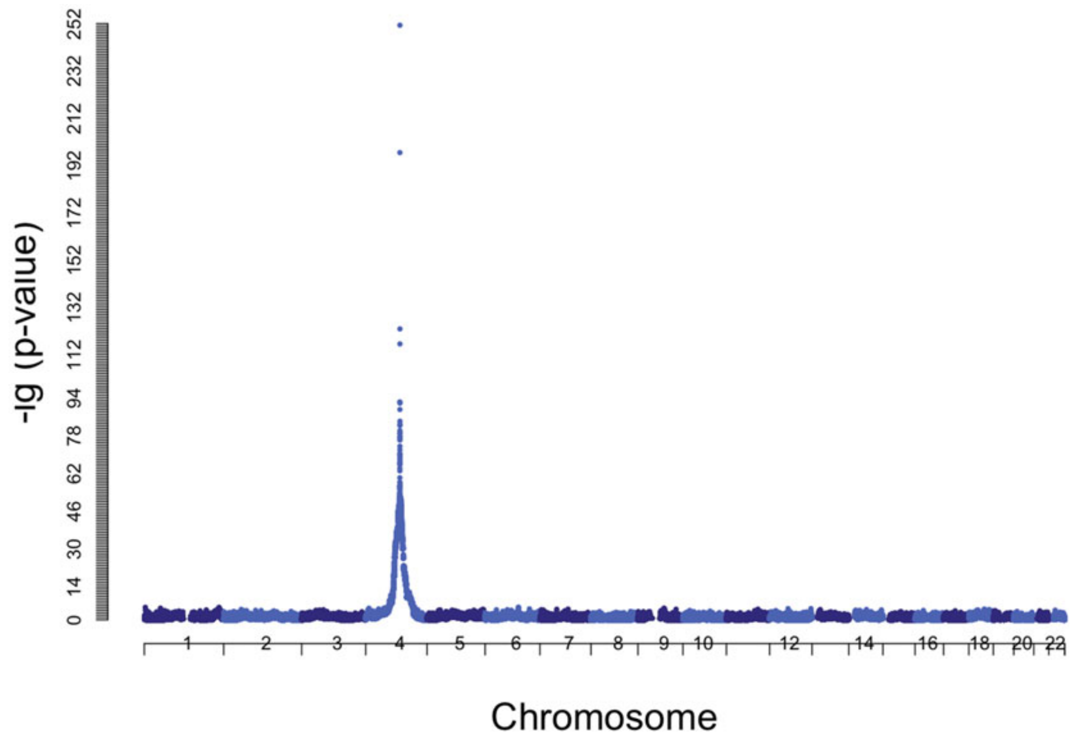

Fig. 5.5 An example of a genome-wide search for eQTLs. The particular example is the *PPA2* gene (inorganic pyrophosphatase 2 precursor), whose expression level in peripheral blood mononuclear cells was assessed with probe GI_31881619-A on a microarray. The large linkage/association peak on chromosome 4 is located at the position of the gene and demonstrates the strength of the effects of *cis* eQTLs on gene expression. In contrast, the plot does not contain outstanding peaks elsewhere in the genome, highlighting the small effect sizes of *trans* eQTLs and the associated difficulty in localizing them

genome yields a very flat pattern, without any outstanding peaks. This illustrates the small effect sizes of *trans* eQTLs and the associated difficulty in localizing them.

5.4 Integrative Genomic Studies

In the previous two sections, I have separately discussed studies correlating gene expression profiles to a trait of interest and studies investigating the genetic regulation of gene expression, respectively. These are two of the three branches of investigation shown in Fig. 5.2, with trait-genotype correlation analysis being the third, and most commonly performed investigation. Ideally, we would like to bring as many sources of information to bear to dissect the etiology of a trait and to identify the functional genetic variants. I have tried to illustrate this conceptually in Fig. 5.6. Thus, we would like to integrate the results obtained from different data sources, to comprehensively assess the evidence for genetic correlation of specific variants with a clinical trait of interest, and the likelihood that the associated variants are functional and of relevance. In the case of gene expression studies, the focus is on the three types of analyses shown in Fig. 5.2, but this is not meant as a suggestion that other sources of information, such as from proteins, metabolites, gene methylation, sequence conservation, predictions of deleteriousness of variants based on structural protein changes, etc., are unimportant or should not be used.

Such integration of the central three types of analyses shown in Fig. 5.2 is not easy, at least if

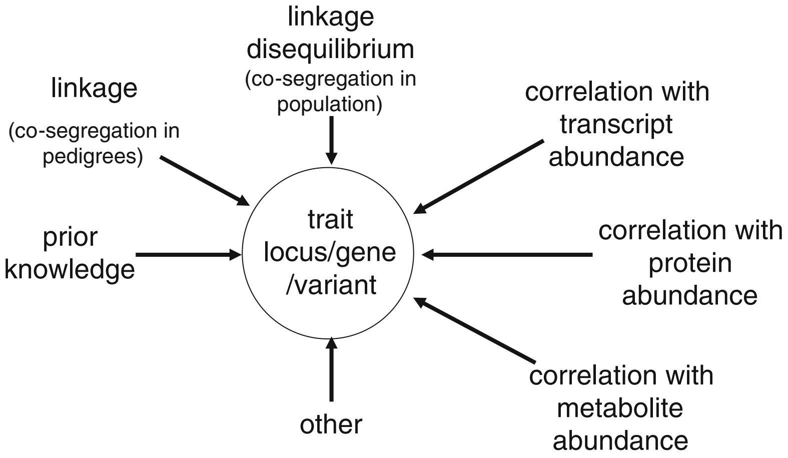
linkage
(co-segregation in pedigrees)
linkage disequilibrium
(co-segregation in population)
correlation with transcript abundance
prior knowledge
trait locus/gene /variant
correlation with protein abundance
correlation with metabolite abundance
other

Fig. 5.6 Many different information sources can be used to identify the loci, genes, and genetic variants influencing a complex trait. Integration of these disparate data sources can be difficult, and this chapter mainly focused on the use of genotype data and transcriptional profile data

the goal is to use a comprehensive analytical approach (such as a Bayesian approach in which the posterior probabilities yielded by one type of data and analysis serve as the prior probabilities for the next data source and analytical step). In most cases, studies use more of an ad hoc approach. If data on trait phenotypes, genotypes, and gene expression levels are all available in the same dataset, one may first perform regular linkage and association analysis, and then subsequently examine whether the significant (or suggestive) trait-associated variants are also significantly associated with the expression level of nearby genes. This could suggest a possible way in which a positional variant may influence the trait of interest, and may increase the interest and attention devoted to this variant. Lastly, one may then investigate whether the gene(s) regulated by the eQTLs show evidence for correlation with the trait of interest, thereby closing the circle. Note that the three analyses comprising the circle of relationships must yield results that are consistent with one another (at least if undertaken on the same dataset). Either all three correlations are positive, or one correlation is positive while the two others are negative. If different datasets are used for association analysis of the trait, for eQTL discovery, and/or for transcriptional profiling of the trait, then one may still seek to combine the results from these separate investigations, but power may be reduced (e.g., because a given variant may have a large effect size in one dataset but not in another). In this case, it is possible that the results of the three analyses are no longer consistent with one another. However, truly existing, important relationships between a trait, genotype, and gene expression level should still yield consistent results (except in unusual situations).

It is also possible to perform the analyses in the opposite direction. One may start with a transcriptional profile study, then identify *cis* eQTLs influencing trait-correlated transcripts, and lastly examine whether the identified candidate variants truly show evidence of association with the clinical trait. When the analyses are conducted in this orientation, then the eQTL and trait association steps can be used to determine whether the observed transcriptional signatures reflect genes involved in the etiology of the trait, or whether the causal connection goes in the other direction, with the trait phenotype influencing the expression levels of these genes (Schadt et al. 2005). In general, there is ample opportunity for smart approaches to be developed to integrate different information sources and to use these data to order the relationships between the different variables being examined.

5.5 Tissue Specificity

Tissue specificity is an important consideration in gene expression studies. This is a very important topic, but it is also one that is complicated and for which there are no clear-cut answers that generalize to all traits, genes, and genetic variants. This issue mainly arises because the tissue that is the primary source of a disease is often not available for study (In fact, in many cases the true source tissue is not even certain). Often the reason is that the tissue is inaccessible to a certain degree, and ethical considerations preclude invasive procedures required to obtain it. A further, related complication is that most tissues comprise many different cell types, which vary from one another in their gene expression, and the relevant cell type is often not known or cannot be readily isolated (or only in a manner which possibly impacts those cells and their behavior greatly).

When talking about tissue specificity it is important to realize that it is not important whether the absolute expression level of a gene is the same in different tissues—as long as the gene is expressed sufficiently highly so that the expression level can be accurately measured. When assessing whether one (accessible) tissue can serve as a surrogate for another (inaccessible) tissue, what matters is whether the inter-individual variation in the expression of a gene is maintained between different study subjects. In eQTL studies, the critical question is whether the same genetic variants influence gene expression, and whether the direction and effect sizes of these variants are similar. This has been empirically examined in a number of studies (Ding et al. 2010; Greenawalt et al. 2011; Nica et al. 2011; Grundberg et al. 2012). In transcriptional profiling studies relating gene expression to some disease, the central concern is whether the relationship between the trait and gene expression level is the same in both tissues. My own opinion is that the suitability of a surrogate tissue will depend on the trait being studied, the particular gene(s) involved, and their underlying eQTL(s). It seems unlikely that any given tissue, or for that matter any given cell type, can universally serve as a proxy for another tissue or cell type. The best that we can hope for is to estimate the similarity of gene expression and its genetic regulatory machinery between as many different tissues and cell types as possible, in order to identify the best overall match, with the greatest overlap in gene expression patterns and eQTLs (Göring 2012).

An important project designed to assess tissue specificity of eQTLs is the US American National Institutes of Health sponsored Genotype-Tissue Expression Roadmap project now underway (GTEx Consortium 2013). The goal is to obtain tissue samples of >1,000 fatal accident victims, gathering as many tissues as quickly as possible after death. These tissues will then be expression profiled, and large-scale eQTL studies will be undertaken in each tissue separately and multiple tissues jointly. Ultimately, this will lead to a catalog of eQTLs, and their estimated effect sizes, in a large number of human tissues. This will permit us to identify best tissue matches, potentially even on a per-gene or per-genetic variant basis. An alternative approach, which is less centralized in scientific direction but which may ultimately be even more informative, is based on attempts to recreate different cell types (ultimately all cell types?) from induced pluripotent stem cells (iPSCs). This topic is well beyond this chapter, but this technology ultimately holds the promise that many (or even all?) cell types become accessible for gene expression study in each study participant or clinical patient (Robinton and Daley 2012).

At the present time, our knowledge of the tissue specificity of eQTLs is fairly limited. It appears to be the case that strong *cis* eQTLs, in particular those close to the transcriptional start site, are fairly universal between tissues, and that those further away are increasingly tissue-specific (Dimas et al. 2009; Grundberg et al. 2012). There are some indications that *trans* eQTLs may often be tissue specific (Grundberg et al. 2012). An important caveat to keep in mind when interpreting these results is that the real effect size of a true eQTL is correlated with our ability to detect it in the first place, as well as with our certainty that the finding is real. Thus, the weaker any *cis* eQTL variants are as we move away from the transcriptional start site, the less certain we are to detect them. It is thus not surprising that less consistency is observed for those more subtle variants. This caveat is even more important in the case of *trans* eQTLs, whose effect sizes are generally smaller and where our power of localization is further weakened owing to the enormous multiple testing burden. While it seems quite plausible that more distant *cis* eQTLs and *trans* eQTLs are more tissue-specific in their influence on gene expression than strong *cis* eQTLs close to the transcriptional start site, I do not find the supporting data wholly convincing at the present time.

Note that it is fully rational and also reasonable, at least in my opinion, to use proxy tissues in many scientific examinations at this point in time. While negative results may be difficult to interpret and may even be entirely uninformative, positive correlations observed between a clinical trait and a gene's expression, or the realization that a candidate variant may be an eQTLs, provide potentially interesting clues that can then be pursued in more detail in the laboratory and/or in a more appropriate tissue that is only available in few samples.

5.6 Microarray Versus RNAseq

Microarrays containing a large number of probes and RNA sequencing (RNAseq) are the two approaches that are now being used to characterize gene expression on a whole genome basis. Older methods, such as quantitative PCR, continue to be used as well, but they are limited to specific genes rather than assess the entire genome at once. Microarray and RNAseq technologies both have advantages and disadvantages. For a review, see (Majewski and Pastinen 2011). Some of the pros and cons of both approaches are the following: The drawbacks of microarrays are that they are limited to known transcripts; they assess only the expression level of a short stretch of RNA and generally cannot distinguish between alternative transcripts; they are susceptible to polymorphisms in the sequence targeted by a probe; they require many copies of RNA molecule for robust expression detection and quantification; they are somewhat susceptible to batch effects. On the plus side, however, microarray studies are fairly cheap, fast, and require limited annotation work by the investigators.

In contrast, RNAseq is (at least conceptually) able to identify all transcripts, including alternative transcripts of a gene; RNAseq is much better suited for the study of RNA editing; since there is no probe per se, RNAseq is less susceptible to polymorphisms in specific transcript regions (though the presence of polymorphisms may interfere with alignment); and RNAseq is much more sensitive in the detection of low frequency (even single copy?) RNA molecules. Downsides of the technology include its substantial cost, and the substantial annotation work that is required. Also, RNAseq is sensitive to sequencing problems and artifacts, and it is not clear whether low copy transcripts have biological function (even if they are reproducible).

Previously, most studies used microarrays, but increasingly the field is transitioning to RNAseq as the preferred choice of technology. As the cost of sequencing comes down more, and as the length of sequencing reads and sequence accuracy improves further, the benefits of RNAseq compared to microarrays will become more pronounced.

5.7 Allele-Specific Expression Analysis

Most eQTL studies conducted to date have searched for association between the genotype of a genetic variant and the overall expression level of a gene, exon, or particular transcript. In many cases, transcripts themselves include polymorphic sites, and the allele present in a given transcript molecule tells which of the two sister chromosome produced the transcript. It is thus possible to estimate the expression level separately for each chromosome. Each "allele's" expression level can then be genetically analyzed separately. It is particularly useful to analyze the proportion of transcripts of a given gene derived from one of the two sister chromosomes, to search for eQTLs, which is referred to as allele-specific expression (ASE) analysis (Almlöf et al. 2012; Pastinen 2010). The reason why allele-specific expression analysis is so useful is that there is a built-in internal control for many factors, including most environmental influences on gene expression and also *trans* eQTLs. In general, these factors will equally influence the expression of both copies of a gene, assuming there is no interaction between the factors and proximal eQTLs. (And the very meaning of the term *trans* refers to the fact that *trans* eQTLs impact the expression of both chromosomes). By taking the relative proportions of transcripts derived from one chromosome compared to its sister chromosome, one automatically controls for these chromosome-non-specific factors. For this reason, ASE studies are extremely powerful for the detection of *cis* eQTLs. It seems likely that many more ASE studies will be conducted in the future, and there is opportunity to refine the analytical methods and to combine ASE analyses with conventional eQTL studies.

5.8 Exposome Studies and Intervention Studies

This chapter has focused on the utility of transcriptional profile studies to investigate the etiology and pathophysiology of complex diseases and to study the genetic regulation of gene expression. More generally, gene expression profiles are one kind of "deep cellular phenotype", providing a highly detailed characterization of the state of a given cell type or tissue type from a study subject at the time of sample collection. Therefore, gene expression profiling is a very general tool that can be used to address many different research questions.

One area of great interest is to use transcriptional profiles for investigating the influence of environmental factors. The totality of the environmental factors to which we are exposed is sometimes referred to as the exposome (Wild 2005). Therefore, one can attempt to correlate gene expression data to measured environmental exposures to search for significant transcriptional correlates of the exposure. This can provide information how the exposure influences cellular biology. For example, one may contrast smokers to nonsmokers. Significant differences in transcriptional profiles may conceptually include genes that influence the probability that someone is a smoker (these genes would therefore be involved in the etiology of the smoking trait). Or the differences may reflect the consequences of smoke inhalation on cellular processes, which may be useful for understanding how smoking influences our body and what the pathophysiological consequences may be. We have conducted a transcriptomic study of smoking and found substantial differences between smokers and nonsmokers in the transcriptional profiles from PBMCs (Charlesworth et al. 2010). Our interpretation was that these differences largely reflect the consequences of smoking behavior rather than modulate the probability of smoking. An example of a transcriptomic study to investigate environmental pollutants is described in a recent manuscript (De Coster et al. 2013).

Intervention studies involving transcriptional profiling are a useful way to investigate the physiological consequences of an exposure. For example, one may measure gene expression in a relevant tissue in patients with a particular disease before and after administering a relevant drug. Such an investigation may provide information about the means of drug action. In addition, it may be possible to screen the patient pool for those individuals for whom the drug is likely to be effective and for those people in whom the drug may not work. Perturbation studies can also be conducted in cell lines and fresh tissue samples, in which case gene expression levels are measured before and after the perturbation has been administered to the samples. As an example from my own research (unpublished), we are currently conducting a study in which we expose lymphoblastoid cell lines derived from schizophrenics and controls to the neurotransmitter dopamine and measure the gene expression levels before and after exposure via RNAseq, with the goal of understanding the relationship of dopamine to the disease. This type of study provides a very high level of experimental control and permits the administration of highly topical perturbations. It seems likely that expression profiling will become a more common component of such studies, in order to assess the impact of an intervention on cellular activity and processes.

5.9 Concluding Remarks

Transcriptional profile data is now generated as part of many different types of studies. This chapter has mainly focused on using gene expression data to identify genes connected to the trait of inference (either etiologically or physiologically) and for examining the genetic regulation of gene expression in detail. Comprehensive genome-wide assessment of gene expression has only been possible for less than a decade or so, and transcriptional profile studies have quickly become part of the standard repertoire of investigative tools available to researchers and clinicians. Given the enormous range of studies involving gene expression profiling, it is difficult to give a comprehensive overview. I have purposefully not focused on details of the methodology (both on the laboratory side and on the analytical side). Instead, I have sought to highlight some of the basic concepts and how transcriptional profiling studies fit into the wider context of human genetic epidemiological investigations of complex diseases. This area of research has proven to be very fruitful, and it is clear that transcriptional profiling studies will become more common in the future.

References

Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, Göring HH, Liljedahl U, Enström C, Brocheton J, Proust C, Godefroy T, Sambrook JG, Jolley J, Crisp-Hihn A, Foad N, Lloyd-Jones H, Stephens J, Gwilliam R, Rice CM, Hengstenberg C, Samani NJ, Erdmann J, Schunkert H, Pastinen T, Deloukas P, Goodall AH, Ouwehand WH, Cambien F, Syvänen AC (2012) Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. PLoS ONE 7:e52260

Charlesworth JC, Curran JE, Johnson MP, Göring HH, Dyer TD, Diego VP, Kent JW Jr, Mahaney MC, Almasy L, MacCluer JW, Moses EK, Blangero J (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. BMC Med Genomics 3:29

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

De Coster S, van Leeuwen DM, Jennen DG, Koppen G, Den Hond E, Nelen V, Schoeters G, Baeyens W, van Delft JH, Kleinjans JC, van Larebeke N (2013) Gender-specific transcriptomic response to environmental exposure in Flemish adults. Environ Mol Mutagen 54:574–588

de Koning DJ, Haley CS (2005) Genetical genomics in humans and model organisms. Trends Genet 21:377–381

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325:1246–1250

Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, Chen W, Weichenthal M, Ellinghaus E, Franke A, Cookson W, Nair RP, Elder JT, Abecasis GR (2010) Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. Am J Hum Genet 87:779–789

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nat Genet 39:1202–1207

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K (2008) Genetics of gene expression and its effect on disease. Nature 452:423–428

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett

JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894

Gaujoux R, Seoighe C (2013) CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics 29:2211–2212

Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 24:408–415

Göring HH (2012) Tissue specificity of genetic regulation of gene expression. Nat Genet 44:1077–1078

Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39:1208–1216

Greenawalt DM, Dobrin R, Chudin E, Hatoum IJ, Suver C, Beaulaurier J, Zhang B, Castro V, Zhu J, Sieberts SK, Wang S, Molony C, Heymsfield SB, Kemp DM, Reitman ML, Lum PY, Schadt EE, Kaplan LM (2011) A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res 21:1008–1016

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, Nisbett J, Sekowska M, Wilk A, Shin SY, Glass D, Travers M, Min JL, Ring S, Ho K, Thorleifsson G, Kong A, Thorsteindottir U, Ainali C, Dimas AS, Hassanali N, Ingle C, Knowles D, Krestyaninova M, Lowe CE, Di Meglio P, Montgomery SB, Parts L, Potter S, Surdulescu G, Tsaprouni L, Tsoka S, Bataille V, Durbin R, Nestle FO, O'Rahilly S, Soranzo N, Lindgren CM, Zondervan KT, Ahmadi KR, Schadt EE, Stefansson K, Smith GD, McCarthy MI, Deloukas P, Dermitzakis ET, Spector TD (2012) Multiple tissue human expression resource (MuTHER) consortium. mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44:1084–1089

GTEx Consortium (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45:580–585

Guo L, Du Y, Chang S, Zhang K, Wang J (2014) rSNPBase: a database for curated regulatory SNPs. Nucleic Acids Res 42(Database issue):D1033–1039

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362–9367

Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–391

Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikäinen LP, Kangas AJ, Soininen P, Würtz P, Silander K, Dick DM, Rose RJ, Savolainen MJ, Viikari J, Kähönen M, Lehtimäki T, Pietiläinen KH, Inouye M, McCarthy MI, Jula A, Eriksson J, Raitakari OT, Salomaa V, Kaprio J, Järvelin MR, Peltonen L, Perola M, Freimer NB, Ala-Korpela M, Palotie A, Ripatti S (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet 44:269–276

Kooij V, Venkatraman V, Tra J, Kirk J, Rowell J, Blice-Baum A, Cammarato A, Van Eyk J (2014) Sizing up models of heart failure: proteomics from flies to humans. Proteomics Clin Appl doi: 10.1002/prca.201300123 [Epub ahead of print]

Majewski J, Pastinen T (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet 27:72–79

Mill J, Heijmans BT (2013) From promises to practical strategies in epigenetic epidemiology. Nat Rev Genet 14:585–594

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464:773–777

Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman AK, Bataille V, Tzenova Bell J, Surdulescu G, Dimas AS, Ingle C, Nestle FO, di Meglio P, Min JL, Wilk A, Hammond CJ, Hassanali N, Yang TP, Montgomery SB, O'Rahilly S, Lindgren CM, Zondervan KT, Soranzo N, Barroso I, Durbin R, Ahmadi K, Deloukas P, McCarthy MI, Dermitzakis ET, Spector TD, MuTHER Consortium (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet 7:e1002003

Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet 11:533–538

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772

Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K (2011) Single-tissue and cross-tissue of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet 7: e1001317

Robinton DA, Daley GQ (2012) The promise of induced pluripotent stem cells in research and therapy. Nature 481:295–305

Sanders AR, Göring HH, Duan J, Drigalenko EI, Moy W, Freda J, He D, Shi JMGS, Gejman PV (2013) Transcriptome study of differential expression in schizophrenia. Hum Mol Genet 22:5001–5014

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC, Berger A, Shuster JR, Wei JT, Varambally S, Beecher C, Chinnaiyan AM (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457:910–914

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET (2007) Population genomics of human gene expression. Nat Genet 39:1217–1224

Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, Roses AD (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci USA 90:1977–1981

Terwilliger JD, Göring HH (2009) Update to Terwilliger and Göring's "Gene mapping in the 20th and 21st centuries" (2000): gene mapping when rare variants are common and common variants are rare. Hum Biol 81:729–733

Tukiainen T, Kettunen J, Kangas AJ, Lyytikäinen LP, Soininen P, Sarin AP, Tikkanen E, O'Reilly PF, Savolainen MJ, Kaski K, Pouta A, Jula A, Lehtimäki T, Kähönen M, Viikari J, Taskinen MR, Jauhiainen M, Eriksson JG, Raitakari O, Salomaa V, Järvelin MR, Perola M, Palotie A, Ala-Korpela M, Ripatti S (2012) Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. Hum Mol Genet 21:1444–1455

Van Eyk JE (2011) Overview: the maturing of proteomics in cardiovascular research. Circ Res 108:490–498

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001–1006

Wild CP (2005) Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev 14:1847–1850

Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, Zou F, Sun W, Sullivan PF, Wright FA (2012) seeQTL: a searchable database for human eQTLs. Bioinformatics 28:451–452

Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, Deloukas P, Dermitzakis ET (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics 26:2474–2476

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, Wellcome Trust Case Control Consortium (WTCCC), McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341