Genet. Res., Camb. (2010), **92**, *pp.* 361–369.
doi:10.1017/S001667231000056X

Insights from GWAS into the quantitative genetics of transcription in humans

JINHEE KIM AND GREG GIBSON*
School of Biology, Georgia Institute of Technology, Atlanta GA 30332, USA

(*Received 4 October 2010 and in revised form 4 November 2010*)

Summary

Human gene expression profiles have emerged as an effective model system for the dissection of quantitative genetic traits. Peripheral blood and transformed lymphoblasts are particularly attractive for their ready availability and repeatability, respectively, and the advent of relatively inexpensive genotyping and microarray analysis technologies has facilitated genome-wide association for transcript abundance in numerous settings. Thousands of genes have been shown to harbour regulatory polymorphisms that have large local effects on transcription, explaining 20% or more of the variance in many cases, but the focus on such results obscures the reality that the vast majority of the genetic component of transcriptional variance remains to be ascertained. This mini-review surveys the inferences derived from genome-wide association studies (GWAS) for gene expression to date, and discusses some of the issues we face in finding the remainder of the heritability and understanding how environmental and genetic regulatory factors orchestrate the highly structured architecture of transcriptional variation.

1. Scope

Genome-wide association studies of gene expression (GWAS-GE) serve a variety of purposes. On the one hand, they are a window on the genetics of disease, serving as a potential bridge between statistical association and functional mechanisms of the influence of polymorphisms on phenotypic variation. On the other hand, they provide fundamental insight into the genetic architecture of complex traits, namely the abundance measures of tens of thousands of transcripts. This mini-review will focus on the latter application. It aims to briefly survey the basic insights that have been gained from almost a decade of research into the genetics of transcriptional variation, to highlight major gaps in our understanding and suggest areas that are now ripe for investigation.

We focus on four main conclusions and insights. First, how much of the heritability of transcription is explained by genome-wide associations, and is the missing heritability problem of the same magnitude as that observed for disease liability and visible variation? Second, how additive is the genetic regulation of transcription and is there evidence for epistasis and/or genotype-by-environment interactions? Third, what do GWAS-GE have to say about pleiotropy, and can results obtained in the study of one tissue be used to make inference about genetic regulation in other tissues? Fourth, how do we proceed to dissect the mechanisms by which hundreds and sometimes thousands of genes are co-regulated?

* Corresponding author: School of Biology, Georgia Institute of Technology. Tel: (404) 835-2343. e-mail: greg.gibson@biology.gatech.edu

2. Brief survey of progress

The initial studies of the genetics of human transcription measured the heritability of transcript abundance, measured as relative fluorescence intensity on DNA microarrays, in transformed lymphoblast cell lines (LCL) from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (Cheung *et al.*, 2003; Monks *et al.*, 2004; Morley *et al.*, 2004). Environmental variation is minimized in such studies, and so the genetic contribution to transcriptional variation is likely to be enhanced relative to gene expression in living people; yet, the basic insight that the majority of transcripts display significant heritability, often with over 50% of the variance explained by familial relatedness, laid the foundation for the field. The result was quickly confirmed in peripheral blood samples from the extensive cross-sectional cohort study of Icelanders (Schadt *et al.*, 2003). In this case,

conclusions must be tempered by the realization that the peripheral blood is a complex mixture of cell types, heritable variation in the abundance of which may contribute to observed correlations among relatives. However, it is now broadly accepted that transcription is strongly affected by segregating genetic variation (Gilad *et al*., 2008; Skelly *et al*., 2009) and that heritability estimates are not noticeably different from estimates of visible phenotypes, a conclusion that was not *a priori* obvious.

Subsequent efforts have illuminated the extent of environmental contributions, and begun to dissect the genetic ones. Regarding the environmental component, a large number of studies point to influences across transient and life-long timescales (Gibson, 2008). The stress of taking exams, for example, alters student peripheral blood profiles, suggesting that immediate changes in cytokine activity associated with hypothalamic activation are mirrored in genome-wide transcriptional changes (Kawai *et al*., 2007). Dietary intervention also quickly modifies transcription profiles (Camargo *et al*., 2010), and there is much interest in determining whether anti-oxidant diets or other interventions can reduce the pro-inflammatory activity of the immune system in chronically stressed contemporary adult populations. People living in different villages or environments display different transcriptome profiles (Idaghdour *et al*., 2008), with geographic factors apparently having a greater influence than genetic divergence, and certainly more than gender differences (although these results need to be confirmed with tissues other than blood). Furthermore, early life exposure, including low socio-economic status, appears to have a lasting effect on gene expression (Miller *et al*., 2009), again influencing pro-inflammatory responsiveness, possibly involving epigenetic modification of the chromatin.

Genetic contributions have been dissected both by linkage and association studies. Studies of LCL expression in the CEPH pedigrees quickly established that for a large number of transcripts, major quantitative trait loci (QTL) map to the chromosomal vicinity of the gene itself (Cheung *et al*., 2005; Stranger *et al*., 2005). This gave rise to the concept of local eQTL, namely expression quantitative trait loci that map to the location of the target gene (Rockman & Kruglyak, 2006). Subsequent research has confirmed that local effects typically map not just within the several cM of a QTL peak, but actually within 100 kb or less of the transcript's promoter (Veyrieras *et al*., 2008). Statistical power to detect linkage is a function of sample size, and so it is not particularly meaningful to estimate what fraction of genes have local eQTL effects since no published eQTL studies have yet exceeded much more than a thousand individuals, but it is noteworthy that the initial studies detected genome-wide significant linkages with fewer than 100 samples. The reason is that many of the effects explained 20% or more of the transcript abundance. More controversial is the question of the prevalence of distant linkages, and whether these fall into so-called hotspots where one locus regulates the expression of dozens or hundreds of targets (de Koning & Haley, 2005; Hubner *et al*., 2005). A few groups have observed an excess of distant eQTL linkages in the CEPH families (Cheung *et al*., 2005; Duan *et al*., 2008) and these increased in number after exposure of the LCL to radiation, accounting for over 90% of the eQTL effects (Smirnov *et al*., 2009). However, with the transition to association mapping, the consistent observation of many groups has been that local eQTL effects are far more numerous and only a handful of robust and repeatable distant eQTL have been described.

The most recent advances have been in the application of GWAS to gene expression, first in LCL (Cheung *et al*., 2005; Stranger *et al*., 2005; Dixon *et al*., 2007; Cookson *et al*., 2009), but subsequently in a diversity of tissues from various peripheral blood samples (Göring *et al*., 2007; Emilsson *et al*., 2008; Heap *et al*., 2009; Idaghdour *et al*., 2010) to adipose biopsies (Göring *et al*., 2007), and liver (Schadt *et al*., 2008) and brain (Myers *et al*., 2007; Webster *et al*., 2009) samples from cadavers. A typical study involving 200 samples will detect several hundreds of independent local eSNP associations at the genome-wide significance threshold of 10^{-8}, most supported by multiple associations in the linkage disequilibrium (LD) block. The term eSNP replaces eQTL to denote that the effect is an association between a single nucleotide polymorphism (SNP) in a population of unrelated individuals, rather than a linkage signal among relatives. There are no systematic surveys of how many of the eSNPs are likely to be causal, and while most presumably simply tag the functional SNP, careful follow-up has established that some disrupt binding sites for transcription factors and influence transcription directly (e.g. Musunuru *et al*., 2010). eSNPs are also notably enriched within several kilobases of annotated promoters (Veyrieras *et al*., 2008), although it is also clear that they can sometimes exert their influence over hundreds of kilobases and across intervening genes (Kleinjan & van Heyningen, 2005).

3. Magnitude and distribution of eSNP effects

Lost in the astonishing conclusion that eSNPs can be detected efficiently in samples as small as 200 or fewer participants is the recognition that most of these cases represent the exception rather than the rule. As shown in Fig. 1, for 450 independent genome-wide significant associations ($P<10^{-8}$) in our Moroccan data (Idaghdour *et al*., 2010), effect sizes range from 0·5 to over 2 standard deviation units (median 0·85),

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press

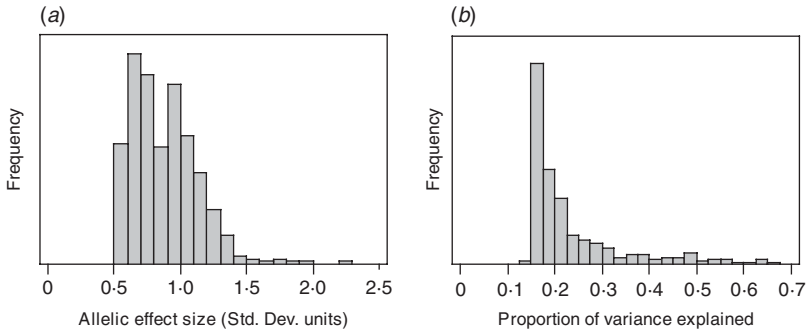

Fig. 1. Allelic contributions to gene expression variation. (*a*) The allelic effect, estimated as the absolute value of the slope of the regression of genotype against transcript abundance, divided by the standard deviation of transcript abundance. Data are for the 450 most significant eSNPs in a study of healthy adults in Morocco described in Idaghdour *et al.* (2010). (*b*) The proportion of variance explained (*R*-squared) for the same eSNP effects.

while the proportion of variation explained by the associations range from 15% to just under 70% (median 20%). These values are likely inflated by Winner's Curse, but clearly are far larger than the typical effects of disease or visible trait associations. Presumably effect sizes covering the full range from less than 1% to 15% will be detected as sample sizes increase to tens of thousands, and the majority of genes will likely eventually be shown to harbour rare and/or common variants that influence transcript abundance. For now, though, we only have estimates for the upper fifth percentile of effects.

One evolutionary study (Kudaravalli *et al.*, 2008) has suggested that these cases may represent genes that have experienced recent positive selection associated with the migration of modern humans across the globe, although formal significant proof of this was only provided for a handful of loci. What is clear is that eSNP frequencies do vary substantially among human populations, which would imply that GWAS-GE surveys in different ethnic groups may detect different associations (Rosenberg *et al.*, 2010), also providing a window on the genetic basis for differences in disease susceptibility among populations. Here too, though, it is important to recognize that only 15% of the variation in LCL expression is partitioned among African American, Asian and Europeans (Storey *et al.*, 2007), and that the majority of eSNPs detected so far show similar magnitudes of effect across groups (Spielman *et al.*, 2007; Stranger *et al.*, 2007).

As mentioned, far fewer distant eSNP effects have been described in the GWAS-GE literature. This partly reflects a statistical power issue, since in a GWAS-GE, three orders of magnitude more tests of association are performed for *trans* than *cis* associations, and the formal genome-wide significance threshold rests lies close to 10^{-11}. The fact that several replicated associations have been observed at this level in samples of a few hundred individuals is remarkable, particularly since many diseases produce no such associations in samples of several thousands, again highlighting the fact that genetic effects on transcription are often much larger than genetic influences on disease liability. On the other hand, when false discovery rate criteria are applied, GWAS-GE continue to demonstrate an excess, somewhere in the range of 20 to 1, of local over distant effects (Skelly *et al.*, 2009; Idaghdour *et al.*, 2010). It is not difficult to explain this difference as a consequence of the very close molecular relationship between regulatory variants in a promoter (or intron) and transcript initiation, relative to the biochemical separation between distant acting regulatory factors and target genes. Yet, it also seems highly likely that common polymorphisms influencing the expression and/or activity of key regulatory proteins and microRNAs (miRNAs) exert a strong influence on the abundance of multiple targets. As larger studies are published and methods that borrow power from analysis of the co-regulation of multiple targets are employed, we expect that weaker distant effects will come into focus.

For now, though, the inescapable conclusion from GWAS-GE is that the regulation of gene expression for the most part suffers from the same 'missing heritability problem' that plagues the genetic dissection of visible phenotypes, including disease susceptibility (Manolio *et al.*, 2009; Eichler *et al.*, 2010). That is to say, there is a considerable gap between the proportion of variance expected to be attributable to genetic polymorphism on the basis of heritability estimates, and the proportion actually detected in GWAS scans. Even in those cases where a major eSNP is detected, the majority of the genetic variance for the abundance of that transcript is unexplained, as it is quite rare for a single transcript to be associated with two or more high-confidence eSNPs. Since the vast majority of transcripts are not associated with eSNPs that explain more than 10% of their variance, the bulk of the genetic variance for individual transcripts remains unexplained. Even in situations where gene expression is bimodal, the genetic regulation can

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press

be inferred to be complex (Hsieh *et al.*, 2007). Rare polymorphisms that impact expression in a small number of individuals are likely to occur, and may be enriched in certain disease situations, but there is little evidence that gene expression is generally regulated by common polymorphisms of moderate effect. For the most part, we expect that hundreds of variants of small effect will turn out to contribute to the genetic variance of most transcripts, just as they influence variation for such traits as height (Yang *et al.*, 2010) and blood lipids (Teslovich *et al.*, 2010).

4. Additivity of gene expression

A second strong conclusion to be taken from GWAS-GE to date is that the detected eSNP effects tend to be additive. Individual cases where heterozygotes for local eSNPs have transcript abundance measures that are greater or less than either homozygote class may be missed in genome-wide scans that typically employ allele trend tests, but they seem to be rarities. In fact, there is also not much evidence for dominance at the level of transcription, at least not for local regulatory effects. Genotype-based tests of association that make no assumptions about heterozygotes relative to homozygotes, and rather test for difference in abundance among the three genotype classes, do not produce more significant associations than trend tests that assume additivity. More impressively, when significant eSNP associations are examined in detail, it is the norm for heterozygotes to have transcript abundance intermediate between the homozygote classes, as expected of additive effects.

Furthermore, one study that directly tested for genotype–environment interactions demonstrated that these are also not a major influence on transcription (Idaghdour *et al.*, 2010). Despite highly significant differences between a rural village and an urban sample in Morocco, allelic effects were observed to be constant across the sub-populations for each of almost 400 *cis* eSNPs, and there were no more significant $G \times E$ interaction effects in a statistical model testing for these, than expected by chance. If generalized to other tissues and other populations, this result would imply that environmental effects are largely additive with respect to locally acting regulation of individual transcripts. Note that this does not exclude the possibility of $G \times E$ interactions at the phenotype level, and in fact suggests a plausible mechanism whereby the combination of strong genetic effects and strong environmental ones, although additive, could see the extreme genotype in a particular environment having transcription above or below a threshold that is disease promoting. Additional studies of this nature seem warranted.

Genotype-by-genotype (namely 'epistatic') interactions have also not been shown to make a substantive contribution to gene expression variation in humans. This may be largely a power issue, because the number of possible interactions is so large that achieving genome-wide significance with relatively small sample sizes would require very large interaction effects that are relatively insensitive to the diverse environments and genetic backgrounds encountered in human samples. The strong covariance of gene expression also complicates efforts to document specific instances of genotype-by-genotype interaction, but possibly the best place to look for epistasis is modification of the magnitude of effect of locally acting polymorphisms. Even if this cannot be attributed to individual interactions between SNP pairs, heterogeneity of allelic effects among individuals is a distinct possibility.

The two recently published RNASeq-based GWAS-GE studies (Montgomery *et al.*, 2010; Pickrell *et al.*, 2010) suggest a plausible mechanism for the additivity of eSNP effects. Both groups observed a correlation between eSNP type and allele-specific expression levels. Since RNASeq provides the sequence of polymorphisms that lie within transcripts, it is straightforward to estimate the relative contribution of each chromosome to transcript abundance in each individual. Given the known LD between an observed eSNP and the transcript haplotype, it is possible to estimate whether the variation among individuals can be attributed to eSNPs regulating the transcription of expression solely from the same chromosome. The general result is that it can be, and while this does not exclude the possibility that regulation also influences transcription from the paired chromosome on occasion, the direct conclusion is that regulatory effects tend not only to be local but also exert their effect on the same DNA molecule. The term *cis*-eQTL should, strictly speaking, be reserved for such chromosome-specific interactions (Rockman & Kruglyak, 2006; Skelly *et al.*, 2009), and these provide a simple explanation for the observed additivity of local effects: were regulatory variants to confer their effects on the paired homologue, dominance would be more common.

5. Pleiotropy and tissue specificity

One of the major questions in the field currently is whether eSNPs commonly exert their influence across multiple tissues. The question is directly relevant to the issue of whether GWAS-GE is a reliable tool for identifying which genes mediate disease SNP effects. It is for example not uncommon for disease SNPs to fall in either a gene desert or a locus with multiple transcripts, and it has been proposed that the demonstration that the SNP is also associated with transcription of one or more of the genes in the vicinity provides evidence that those genes are causal in the condition. Such a conclusion must be treated with

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press

caution – for example, the often observed association in the SUOX locus with Type 1 diabetes and *RPS26* transcription may suggest the wrong gene, partly because transcript abundance does not correlate with disease (Heinzen *et al*., 2008). In several instances, GWAS-GE papers have reported a specific example of an eSNP effect that mirrors a disease association in the public database (Emilsson *et al*., 2008; Schadt *et al*., 2008; Cookson *et al*., 2009; Idaghdour *et al*., 2010), but the expression profiling was performed in a very different tissue – say, peripheral blood expression suggesting an influence on expression of a gene involved in heart failure. It is thus important to know how often eSNPs operate across tissues.

The superficial answer is that transitivity across tissues of eSNP effects is not common, with perhaps not more than 30% of associations in one tissue also being detected in another (Petretto *et al*., 2006; Dimas *et al*., 2009). This is a difficult estimate to make because it depends on which thresholds of significance are adopted. There are also trivial reasons why it is often the case that SNP effects fail to replicate across tissues, most notably where the gene is not actually expressed in the second tissue. However, regulatory enhancers are typically tissue specific, and if an eSNP influences enhancer activity, as opposed to basal promoter activity, then there is no reason to expect transitivity. That is to say, it may be common for different regulatory SNPs to influence expression of the same transcript in different tissues – and in general there is no reason to expect that the same tagging SNP would capture both effects. The other major explanation for failure to detect eSNP effects in two or more tissues is statistical power: in fact, the overlap in eSNP detection across studies of the same tissue is also only in the vicinity of 10–20% (Heap *et al*., 2009; Idaghdour *et al*., 2010). This is in accordance with the power to detect association at the stringent GWAS threshold for variants, but if criteria are relaxed to control the false discovery rate, or even simply to use a nominal significance threshold of $P<0{\cdot}05$ for rediscovery, the proportion of replicated associations increases markedly. The authors of the RNASeq GWAS studies, for example, reported very significant overlap, with over 90% of nominal associations also being consistent with effects on transcription in the same direction (Montgomery *et al*., 2010; Pickrell *et al*., 2010). Given the diversity of statistical and experimental methods adopted in GWAS, including genotyping and expression profiling platforms and methods of tissue growth or sampling, it is difficult to accurately measure repeatability within, let alone between tissues. It is however safe to conclude that GWAS-GE in one tissue will very often miss eSNP effects on other tissues, and conversely that an effect observed in peripheral blood, for example, does not guarantee one in liver or brain.

A different approach to this problem has been to ask whether there is enrichment for eSNP associations in the database of Genotypes and Phenotypes (at www.ncbi.nlm.nih.gov/gap) (dbGaP) database of disease associations. The clear answer is that there does seem to be one, and hence it can be concluded that regulatory variants make a significant contribution to disease, quite likely accounting for more associations than coding variants that disrupt protein structure and function (Nica *et al*., 2010; Nicolae *et al*., 2010). The latter are, *a priori*, more likely to have pleiotropic effects across tissues and hence to explain the association of individual genes with multiple diseases that involve expression in different organs. Heterogeneity in regulatory eSNP effects might contribute to loss of genetic correlation between diseases, yet allow the same gene to contribute, for example, to age-related macular degeneration and susceptibility to meningococcal septicemia (Davila *et al*., 2010; complement factor H in this case). Unfortunately, access to diseased tissues in large population samples, certainly from living donors, is often not possible, slowing progress in relating the genetics of expression variation to disease. However, the GEN-EX consortium will soon provide data on eSNP effects measured in multiple tissues from the same donors, and studies of differentiated pluripotent stem cells may also be informative in the near future.

6. Distant effects and the co-regulation of gene expression

Arguably the most pressing challenge for those specifically focused on the quantitative genetics of gene expression is identifying the genetic mediators of co-regulation. There are multiple challenges. First of these is overcoming the statistical power issues that plague detection of what are expected to be relatively small influences of *trans* eSNPs (each explaining <1% of the variance), although ongoing studies of cohorts of 10 000 or more may resolve this.

Just as pressing is the issue of dealing with the complex covariance structure of gene expression, including the pervasive influence of technical factors. Authors are starting to normalize gene expression data by measuring the residuals (or ranks) after fitting the most significant principal components to the data, on the assumption that if these do not correlate with any known biological factors then they are likely to represent technical confounders (batch effects, hybridization or RNA amplification artefacts and transient biological noise) that should be removed (Leek & Storey, 2007; Pickrell *et al*., 2010). If this improves the ability to detect associations with transcript abundance, or with known biological factors, it may be well warranted, albeit with unknown consequences. Other authors prefer to perform aggressive

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press

quantile normalization that forces distributions to be similar, but almost certainly throws out a meaningful biological signal where thousands of genes are affected by biological factors, as is common (Dabney & Storey, 2007). Mecham *et al*. (2010) have recently introduced a supervised normalization approach that jointly fits biological and technical sources of variance prior to testing of individual transcripts, likely indicating an optimal approach to appropriately defining the covariance structure in population-scale gene expression data sets.

Even given appropriate normalization and sufficient power to detect distant effects, there remains the problem that important regulatory factors may themselves be monomorphic (hence will never contribute to the genetic variance, yet be responsible for mediating environmental responses), or the allelic effects will be too small to detect. If regulation of transcript abundance is as complex as that of serum lipids, where a GWAS in excess of 100 000 people only uncovered 30% of the genetic (and $<15\%$ of the phenotypic) variance (Teslovich *et al*., 2010), there is little prospect for ever describing the major sources of distantly acting regulatory variance for any, let alone the majority, of transcripts. On the other hand, it may be possible to perform joint analyses on multiple transcripts to increase power. The possibility that non-genetic factors also account for covariance (for example, variation in cell-type abundance in tissues) will have to be accounted for. Other promising approaches include focused analyses of sub-networks of genes that are known to be regulated by common transcription factors, and these may include procedures that adjust analyses for binding site enrichment for such regulatory molecules, and/or structured equation modelling that attempts to predict causal pathways. Certainly novel approaches that go beyond straightforward association testing are called for.

7. Concluding remarks

GWAS-GE studies have attracted much attention in the context of systems biology, as they promise to open the black box between genotype and phenotype (de Koning & Haley, 2005; Gilad *et al*., 2008; Skelly *et al*., 2009). Ultimately the heritability of visible phenotypes must be understood in terms of the genetic influence of polymorphisms on gene activity, which includes a substantial component of transcript abundance. A minor component of the transcriptome displays eSNP effects that are an order of magnitude greater than most observed phenotype associations, accounting for as much as half of the variance of the transcript. We do not yet know whether such associations make a disproportionate contribution to disease. Studies that directly consider GWAS in the context of disease are just beginning to appear, with obesity (Emilsson *et al*., 2008) and celiac disease (Heap *et al*., 2009) notable examples, and many more are needed.

Technology continues to drive advances. RNASeq holds great promise since it can discern allele-specific effects, and is well suited for detecting alternate transcript abundance. Several studies have reported splicing SNPs that influence alternate splicing (Heinzen *et al*., 2008; Zhang *et al*., 2009), with effects as large as eSNPs, but it remains to be seen whether the depth of coverage of RNASeq is sufficient to make this approach practical for all but the major splice variants. New third-generation sequencing technologies will facilitate even deeper sequence coverage, providing digital readouts of transcript abundance, instead of the relative measures that are obtained with hybridization-based methods. Another issue that we have not discussed is epigenetic regulation (Serre *et al*., 2008): studies of the genetic regulation of methylation are starting to appear (Zhang *et al*., 2010), and should be very powerful when combined with the genetics of gene expression. Similarly, GWAS for metabolites measured with high-throughput mass spectrometry are now feasible (Illig *et al*., 2010), and systems analyses jointly of the transcriptome and metabolome will be an exciting new development.

At the same time, we would emphasize the need for targeted analyses that supplement the genome-wide view across populations with pathway-based profiles of genetics at the level of families. The really interesting and important aspect of the population structure of gene activity is that it may contribute to the observation that diseases tend to be enriched in particular pedigrees. Family studies have fallen out of favour in the GWAS era because linkage prevents resolution to the individual gene, but research that bridges the gap between epidemiological genomics and individual risk will require efforts targeting the traditional units of inheritance.

References

Camargo, A., Ruano, J., Fernandez, J. M., Parnell, L. D., Jimenez, A., Santos-Gonzalez, M., Marin, C., Perez-Martinez, P., Uceda, M., Lopez-Miranda, J. & Perez-Jimenez, F. (2010). Gene expression changes in mononuclear cells in patients with metabolic syndrome after acute intake of phenol-rich virgin olive oil. *BMC Genomics* **11**, 253.

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M. & Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics* **33**, 422–425.

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M. & Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369.

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**, 184–194.

Dabney, A. R. & Storey, J. D. (2007). Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biology* **8**, R44.

Davila, S., Wright, V. J., Khor, C. C., Sim, K. S., Binder, A., Breunis, W. B., Inwald, D., Nadel, S., Betts, H., Carrol, E. D., de Groot, R., Hermans, P., Hazelzet, J., Emonts, M., Lim, C., Kuijpers, T., Martinon-Torres, F., Salas, A., Zenz, W., Levin, M. & Hibberd, M. L. (2010). Genome-wide association study identifies variants in the *CFH* with susceptibility to meningococcal disease. *Nature Genetics* **42**, 772–776.

de Koning, D. J. & Haley, C. S. (2005). Genetical genomics in humans and model organisms. *Trends in Genetics* **21**, 377–381.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez-Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J., Deloukas, P., Dermitzakis, E. T. & Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R. & Cookson, W. O. (2007). A genome-wide association study of global gene expression. *Nature Genetics* **39**, 1202–1207.

Duan, S., Huang, R. S., Zhang, W., Bleibel, W. K., Roe, C. A., Clark, T. A., Chen, T. X., Schweitzer, A. C., Blume, J. E., Cox, N. J. & Dolan, M. E. (2008). Genetic architecture of transcript-level variation in humans. *American Journal of Human Genetics* **82**, 1101–1113.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**, 446–450.

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

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature Reviews Genetics* **9**, 575–581.

Gilad, Y., Rifkin, S. A. & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics* **24**, 408–415.

Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B., Abraham, L., Rainwater, D., Comuzzie, A., Mahaney, M., Almasy, L., CacCluer, J., Kissebah, A., Collier, G., Moses, E. & Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics* **39**, 1208–1216.

Heap, G. A., Trynka, G., Jansen, R. C., Bruinenberg, M., Swertz, M., Dineson, L., Hunt, K., Wijmenga, C., Vanheel, D. & Franke, L. (2009). Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics* **2**, 1.

Heinzen, E. L., Ge, D., Cronin, K. D., Maia, J. M., Shianna, K. V., Gabriel, W., Welsh-Bohmer, K., Hulette, C., Denny, T. & Goldstein, D. B. (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biology* **6**, e1000001.

Hsieh, W. P., Passador-Gurgel, G., Stone, E. A. & Gibson, G. (2007). Mixture modeling of transcript abundance classes in natural populations. *Genome Biology* **8**, R98.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., MacIver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Müller, A., Cook, S., Kurtz, T., Whittaker, J., Pravenec, M. & Aitman, T. J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* **37**, 243–253.

Idaghdour, Y., Czika, W., Shianna, K. V., Lee, S. H., Visscher, P. M., Martin, H. C., Miclaus, K., Jdallah, S. J., Goldstein, D. B., Wolfinger, R. D. & Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nature Genetics* **42**, 62–67.

Idaghdour, Y., Storey, J. D., Jadallah, S. J. & Gibson, G. (2008). A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genetics* **4**, e1000052.

Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B., Mewes, H., Meitinger, T., de Angelis, M., Kronenberge, F., Soranzo, N., Wichmann, H. E., Spector, T. D., Adamski, J. & Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics* **42**, 137–141.

Kawai, T., Morita, K., Masuda, K., Nishida, K., Shiishima, M., Ohta, M., Saito, T. & Rokutan, K. (2007). Gene expression signature in peripheral blood cells from medical students exposed to chronic psychological stress. *Biological Psychology* **76**, 147–155.

Kleinjan, D. A. & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics* **76**, 8–32.

Kudaravalli, S., Veyrieras, J.-B., Stranger, B. E., Dermitzakis, E. T. & Pritchard, J. K. (2008). Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution* **26**, 649–658.

Leek, J. T. & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, 1724–1735.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Mecham, B. H., Nelson, P. S. & Storey, J. D. (2010). Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315.

Miller, G. E., Chen, E., Fok, A. K., Walker, H., Lim, A., Nicholls, E., Cole, S. & Kobor, M. S. (2009). Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. *Proceedings of the National Academy of Sciences of the USA* **106**, 14716–14721.

Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J. W., Sachs, A. & Schadt, E. E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics* **75**, 1094–1105.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. & Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

Musunuru, K., Strong, A., Frank-Kamenesky, M., Lee, N. E., Ahfeldt, T., Sachs, K., Li, X., Li, H., Kuperwasser, N., Ruda, V., Pirruccello, J., Muchmore, B., Prokunina-Olsson, L., Hall, J., Schadt, E. E., Morales, C., Lund-Katz, S., Phillips, M. C., Wong, J., Cantley, W., Racie, T., Ejebe, K., Orho-Melander, M., Melander, O., Koteliansky, V., Fitzgerald, K., Krauss, R. M., Cowan, C. A., Kathiresan, S. & Rader, D. J. (2010). From non-coding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719.

Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., Zismann, V. L., Joshipura, K., Huentelman, M. J., Hu-Lince, D., Coon, K. D., Craig, D. W., Pearson, J. V., Holmans, P., Heward, C. B., Reiman, E. M., Stephan, D. & Hardy, J. (2007). A survey of genetic human cortical gene expression. *Naure Genetics* **39**, 1494–1499.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I. & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6**, e1000895.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, E. & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLS: Annotation to enhance discovery from GWAS. *PLoS Genetics* **6**, e1000888.

Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N. & Aitman, T. J. (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics* **2**, e172.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772.

Rockman, M. V. & Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* **7**, 862–872.

Rosenberg, N. A., Huang, L., Jewett, E. M., Spziech, Z. A., Jankovic, I. & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics* **11**, 356–366.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M., Johnson, J., Rohl, C., van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S., Schuetz, E., Rushmore, T. H. & Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6**, e107.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B. & Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse, and man. *Nature* **422**, 297–302.

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D., Dickinson, T., Fan, J. B. & Hudson, T. J. (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS Genetics* **4**, e1000006.

Skelly, D. A., Ronald, J. & Akey, J. M. (2009). Inherited variation in gene expression. *Annual Reviews of Genomics and Human Genetics* **10**, 313–332.

Smirnov, D., Morley, M., Shin, E., Spielman, R. S. & Cheung, V. G. (2009). Genetic analysis of radiation-induced changes in human gene expression. *Nature* **459**, 587–591.

Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J. & Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics* **39**, 226–231.

Storey, J. D., Madeoy, J., Strout, J. L., Wurfel, M., Ronald, J. & Akey, J. M. (2007). Gene-expression variation within and among human populations. *American Journal of Human Genetics* **80**, 502–509.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Kahl, B., Antonarakis, S. E., Tavaré, S., Deloukas, P. & Dermitzakis, E. T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**, e78.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C., Beazley, C., Ingle, C., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P. & Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics* **39**, 1217–1224.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Ioannis, M. & 208 others (2010). Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713.

Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. & Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* **4**, e1000214.

Webster, J. A., Gibbs, R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., McCorquodale, D. S. 3rd, Cuello, C., Leung, D., Bryden, L., Nath, P., Zismann, V., Joshipura, K., Huentelman, M., Hu-Lince, D., Coon, K., Craig, D., Pearson, J.; NACC-Neuropathology Group, Heward, C. B., Reiman, E., Stephan, D., Hardy, J. & Myers, A. J. (2009). Genetic control of human brain transcript expression in Alzheimer disease. *American Journal of Human Genetics* **84**, 445–448.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A., Nyholt, D. R., Madden, P., Heath, A., Martin, N. G., Montgomery, G. W., Goddard, M. E. & Visscher, P. M. (2010). Common SNPs explain a large

proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.

Zhang, D., Cheng, L., Badner, J. A., Chen, C., Chen, Q., Luo, W., Craig, D., Redman, M., Gershon, E. & Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. *American Journal of Human Genetics* **86**, 411–419.

Zhang, W., Duan, S., Bleibel, W., Wisel, S., Huang, R., Wu, X., He, L., Clark, T., Chen, T., Schweitzer, A., Blume, J., Dolan, M. E. & Cox, N. J. (2009). Identification of common genetic variants that account for transcript isoform variation between human populations. *Human Genetics* **125**, 81–93.

https://doi.org/10.1017/S001667231000056X Published online by Cambridge University Press