Human Molecular Genetics, 2009, Vol. 18, Review Issue 2 **R211–R215**
doi:10.1093/hmg/ddp400

The resolution of the genetics of gene expression

Stephen B. Montgomery[1,2] and Emmanouil T. Dermitzakis[1,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva CH-1211, Switzerland and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1HH, Cambridge, UK

Received August 17, 2009; Revised and Accepted August 18, 2009

Understanding the influence of genetics on the molecular mechanisms underpinning human phenotypic diversity is fundamental to being able to predict health outcomes and treat disease. To interrogate the role of genetics on cellular state and function, gene expression has been extensively used. Past and present studies have highlighted important patterns of heritability, population differentiation and tissue-specificity in gene expression. Current and future studies are taking advantage of systems biology-based approaches and advances in sequencing technology: new methodology aims to translate regulatory networks to enrich pathways responsible for disease etiology and 2nd generation sequencing now offers single-molecular resolution of the transcriptome providing unprecedented information on the structural and genetic characteristics of gene expression. Such advances are leading to a future where rich cellular phenotypes will facilitate understanding of the transmission of genetic effect from the gene to organism.

PAST AND CURRENT STUDIES OF GENETICS OF GENE EXPRESSION

Gene expression is a fundamental cellular function. The pattern and properties of gene expression in any organism or cell is an indicator of the cellular state and can also influence the function of other cells. For the last 20 years, technology has allowed us to measure levels of gene expression for many or all genes of an organism, and this has revolutionized our ability to screen the effect of genetic and environmental perturbations. The accuracy by which one can now measure mRNA levels has allowed the use of such measurements in the context of genetic variation within species.

Understanding the effects of genetic variation in basic cellular processes such as gene expression is key to the dissection of the genetic contributions to whole organism phenotypes. The effects of genetic variants can be quite simple and easily interpretable at the cellular level but may be hard to dissect at the whole organism level owing to the large number of direct and indirect interactions occurring between the DNA variant and the phenotype (1).

In the recent years, the method of choice for the study of complex phenotypes and diseases in humans is to perform genome-wide association studies (GWAS) in large samples of cases and controls and/or cohorts with disease-related traits, such as lipid levels or body mass index, or other anthropometric traits such as height (2). One of the key features of GWAS is that one can detect common genetic variants that statistically explain a fraction of the variance of the phenotype, but quite frequently such signals of association are found in the regions of the genome with no apparent function or the correlation structure of variants in the genome (linkage disequilibrium) does not allow firm conclusions about what the functional effect is (i.e. which gene has its function modified due to the genetic variant). The ability to interrogate and study the genetics of phenotypes that are intermediate between the DNA variant and the phenotype of interest can provide substantial additional power in inferring the true biological effect, which is essential for the development of medical interventions. Gene expression is one of these key intermediate phenotypes and there have been a number of studies that have shown its value in the disease context.

Studies on the genetics of gene expression were first performed only in the last decade, which acted as proof of principle. Some of the initial experiments were performed in yeast, which showed extensive genetic variation for gene expression (3). Two key papers that followed, which looked at the genetics of gene expression in lymhoblastoid cell lines (LCLs) from the CEPH samples from Utah, have both shown that gene expression phenotypes are heritable in family pedigrees and therefore there is genetic variation to be mapped in outbred populations (4–6). A number of papers have followed that attempted mapping of genetic variants that affect expression levels of a limited number of

*To whom correspondence should be addressed. Emmanouil.Dermitzakis@unige.ch

genes primarily in *cis* but to some extent in *trans* (7,8). These studies have been facilitated by the availability of the HapMap data, which provided 1 million and, subsequently, more than 3 million of genotyped SNPs for each of the 270 individuals from four global populations.

MULTIPLE-TISSUE STUDIES

Gene expression in higher eukaryotes is well recognized to be cell-type specific; different gene repertoires are intrinsic to a cell's function and the developmental processes of cell differentiation. Analyses in the brain, kidney and liver in mouse strains have highlighted the complexity of genetic influences across multiple tissues; genes were subjected to complex *trans*-genetic influences, and only 2% of genes with genetic differences were shared among all three tissues (9,10). In contrast, in hippocampus, lung, and liver tissues collected from heterogeneous stock mice, it was reported that two-thirds of *cis*-acting eQTLs and one-half of *trans*-acting eQTLs are shared (11). In humans, comparison of adipose and blood from two separate Icelandic cohorts identified sharing between 50% of the *cis*-eQTLs (12). Low correlation was reported between cortical tissue and LCLs between European-descendent samples (13). Fewer than 50% of eQTLs were shared in a comparison of autopsy-derived cortical tissue with peripheral blood mononucleated cell samples from living donors (14). Furthermore, recent efforts by our laboratory have also begun to decipher genetic variation in the context of cell-type specificity (15). We investigated eQTLs in primary fibroblasts, LCLs and primary T-cells detected from 85 Swiss individuals and found that 69–80% of all discovered regulatory variants were cell-type specific. Cell-type-specific variants were identified to have lower effect sizes and were broadly distributed around transcription start sites, suggesting their role on tissue-specific enhancer elements. These results suggest that cellular context will play a fundamental role in our ability to attribute expression variation to higher level phenotypes.

POPULATION DIFFERENTIATION OF GENE EXPRESSION

Understanding the genetic basis of gene expression variability is a fundamental component in building our understanding of the etiology of complex traits within populations. A principal consideration has been identifying the extent to which gene expression is a heritable trait. Early studies on a limited number of genes suggested reduced genetic variability in gene expression among monozygotic twins compared with siblings (16). Examination of 2726 and 2340 expressed genes from 4 and 15 CEPH reference families, respectively, identified that 29–31% had significant heritability (4,6). Heritability estimates in 30 CEPH and 30 Yoruban reference families have demonstrated 10 and 13% of 47 294 assayed probes having heritability greater than 0.2 with 958 genes shared (17). Similar estimates in 333 Icelandic families identified that 26% of 23 720 transcripts could be identified as heritable at 5% FDR after adjusting for sex, age, cell count and BMI (12). Heritability in 30 recombinant inbred rat strains demonstrated in four tissues that >20% of transcripts had heritability greater than 0.2 and of those with *cis*-eQTLs an even higher proportion (0.31–0.51) (18).

Given that gene expression is heritable, the question remains to what degree are populations differentiated due to regulatory variation. Analysis of 16 individuals of African and European ancestry estimated that 17% of genes are differentially expressed among populations (19). Comparison of European- and Asian-derived populations in 4197 genes observed that 1097 showed population differentiation (20). Similar comparison across 120 European- and African-derived populations observed a mean value of 0.2 and a median of 0.12 for the proportion of gene expression variation attributable to population differences (21). Detection of eQTLs from approximately 2.2 million SNPs in four worldwide populations for 270 individuals as part of the phase II HapMap for 13 643 genes reported cross-population replication of *cis*-eQTLs as 37% and that between 17 and 29% of genes have significant gene expression differences between any two populations (17,22).

Such observations of heritability and population differentiation highlight the dynamics of regulatory evolution and the importance of genetics in human phenotypic variation.

HIGH-DIMENSIONALITY OF GENE EXPRESSION DATA: ASSESSING EFFECTS BY FACTOR AND PCA ANALYSIS

The effects of non-genetic variation such as environment, population stratification and cellular stage invariably dampen the ability to detect true genetic associations with gene expression. The scope of these effects can be biological such as perturbation of multiple genes within a regulatory cascade or, technical, through false association with latent experimental factors. However, in a similar vein, true genetic effects can also propagate through multiple genes and thereby exhibit similar patterns. Discriminating non-genetic factors from genetic factors as topological features of gene expression has characteristically involved assessing the higher dimensionality of gene expression data sets with respect to hidden or observed covariates of interest using statistical techniques such as factor analysis and regression. Factor analysis on the Phase II HapMap eQTL study by Stranger *et al*. (23), correcting for a maximum of 40 unobserved latent variables, tripled the number of statistically significant associations detected. Surrogate variable analysis, which removes through regression orthogonal vectors which are considered to have significantly more variation than expected by chance, has been demonstrated to find ~20% more *cis*-linkages compared with those originally detected in Brem *et al*. (3,24). Principal component analysis has been used to correct genotype and phenotype data in samples with mixed ancestry (25). Supervised principal component analysis, which removes irrelevant genes in advance of principal component analysis by using an *a priori* defined gene set, has been used in mice to characterize sexual dimorphism of aortic lesions (26). Further approaches have explicitly corrected against what have been identified as *trans*-eQTL hotspots, bands of statistically significant *trans*-associations and technical confounding factors (27). However,

the extent to which *trans*-eQTL hotspots are non-genetic is still undetermined, as analysis of *trans*-eQTL hotspots in radiation hybrid cell lines demonstrated GO enrichment for terms such as transcription (28). Our own work has seen a significant excess of transcription factors associated with *trans*-eQTL hotspots in humans, suggesting that factor analysis is best considered with respect to biologically relevant posteriors (S.B.M. and E.T.D., unpublished data).

NETWORK INFERENCE AND SYSTEM GENETICS USING GENE EXPRESSION

Gene regulatory networks are well sought for their predictive ability to determine gene–gene interaction outcomes in a transcriptional network. The utility of gene expression data to uncover the topology of GRNs has been well studied with many different statistical models (reviewed in 29). The advantage of such models from a genome-wide association standpoint is that they immediately reduce the impact of multiple testing corrections and further support the detection of eQTL effects through joint modeling. Such an approach has been demonstrated in a recent analysis of human adipose tissue where genetic variants were associated with a macrophage-enriched metabolic network, and within this class of eQTLs, statistical significance for association with obesity was thereby enriched in genetic associations to obesity (12). Another analysis aimed to reproduce likely networks from available functional genomics data in yeast observed that subnetworks are significantly enriched for genes sharing common eQTLs (30). Furthermore, several network modeling algorithms have also been proposed and developed for GWAS in lieu of actual network or extensive functional genomics data being available (31–33). However, GRNs are context dependent, and harmonization between network priors and gene expression context will likely continue to enrich the correlation between expression and trait.

Network-based approaches expand on power to associate single variants with expression phenotypes; however, increasing relevance of many variants to one or many genes associations is highlighted by low-effect sizes seen in many GWAS (34). Most GWAS assume additive model of effects. Our laboratory has explored an epistatic model where regulatory variation modifies the effect of coding risk alleles predicting from 210 unrelated individuals that 18% of non-synonymous SNPs are differentially expressed among individuals (35). These results highlight that not only the spectrum of genes in a trait-related network but the intrinsic variant–variant interactions of that network will influence phenotype.

ENVIRONMENTAL EFFECTS ON GENE EXPRESSION

Although genetics is always an important component of variation in gene expression, we should not forget that gene expression is also highly sensitive to the environment in which the cell is found. Environmental effects in this context are several variables of the individual for which we measure gene expression (diet, smoking, age) as well as effects that are caused by the treatment of cells, which affect the levels and patterns of gene expression. In a recent study (36), it was shown that environmental effects and lifestyle can have very strong effects on gene expression patterns that may be stronger than the genetic signal. In a different study, it was shown that intrinsic properties of LCLs as well as properties that are imposed to them by the experiment can also influence gene expression levels (37). The ability to measure *a priori* such effects or to infer them using statistical methodologies (24) is bound to have a large effect in our ability to evaluate the effect of genetic factors in cellular processes such as gene expression.

HIGH-RESOLUTION GENETICS (ARRAYS VERSUS SEQUENCING)

Recent advances in sequencing have enabled a more detailed resolution of the transcriptome landscape. Recent uses of RNA sequencing (RNA-seq) have improved by-transcript quantification, assessment of alternative splicing and detection of novel gene structure. Initial sequencing of adult mouse brain, liver and skeletal muscle reported 3500 different genes with alternative splice sites and high levels of technical reproducibility ($r^2 = 0.96$) (38). Furthermore, comparison of splicing detected in this experiment with splicing arrays has demonstrated that there are non-biological splicing array biases that do not exist in the sequencing data (39). Similarly, analysis of RNA-seq from six tissues highlighted that $\sim$95% of multiexon genes are alternatively spliced and technical comparison of 1548 cassette-type alternative exon-splicing events in microarrays reported correlation r of 0.80 (40). Furthermore, a general assessment of reproducibility of sequencing quantification against microarrays in two tissues showed Spearman correlations of 0.73 and 0.75, respectively (41). Likewise, when these authors compared genes identified as differentially expressed, they found 81% were shared between platforms and, through followup with qPCR, suggested that many of those genes called differentially expressed in the sequencing data were likely true positives. Further resolution of RNA-seq suggests that it is better at discriminating low-level expression from background noise such as that found in microarrays; one study reported that RNA-seq captured 25% more known transcripts (42). We have investigated the ability to impute results obtained on arrays into RNA-seq experiments both to enable better tissue-specific array designs and to facilitate hybrid sequencing/array experimental designs. Preliminary results suggest that mean Spearman correlation between exons across 48 individuals is >0.5 (S.B.M. and E.T.D., unpublished data).

Although increasingly affordable, it still remains technically challenging and relatively costly to perform RNA-seq on large cohorts. The advantages, however, are that a broader spectrum of quantitative phenotypes compared with array-based studies are now accessible.

SECOND GENERATION SEQUENCING IN DNA AND RNA: PERSPECTIVES

Next generation sequencing in DNA and RNA has ushered in a new era of genetic analysis with respect to regulatory com-

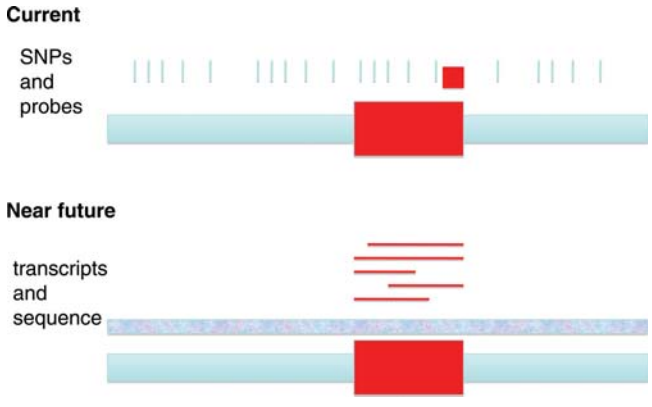

Figure 1. Current genome-wide studies make use of genetic variation and transcript abundance acquired from arrays. In the near future, advances in sequencing will provide access to full variation and transcript information.

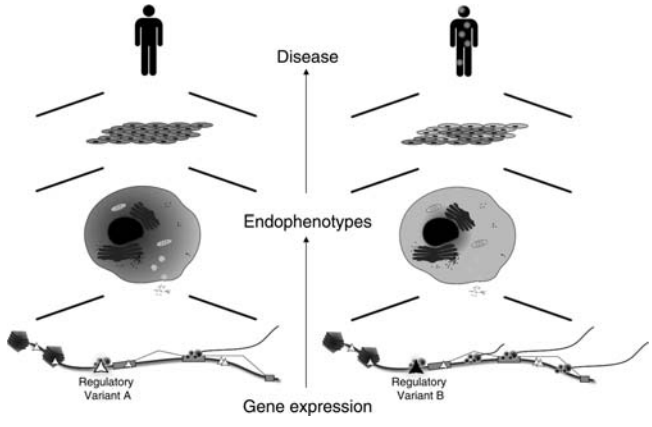

Figure 2. The understanding of the transmission of the genetic effect from the gene to the cell to whole organisms is a key priority for medical science and molecular biology studies.

plexity. The 1000 Genomes Project (www.1000genomes.org) promises to uncover a deeper spectrum of rare variants and lesser studied variants such as indels and CNVs which when coupled with expression technology will increase our understanding of rare variant effects and allow for richer quantification of allele-specific expression (Fig. 1). The increasing affordability of bisulfite sequencing (43,44), strand-specific sequencing (45), Chip-Seq (46), GLO-seq (47), MeDIP-seq (48) and future sequencing methodologies when surveyed in families and populations will increase our understanding of genetic influence on the regulatory genome. It is these intermediate cellular endophenotypes that offer the clearest translation to the understanding of the molecular basis of human phenotypic variation.

IMPLICATIONS FOR DISEASE AND CONCLUSION

Ultimately, we would like to understand the biochemical and molecular basis of disease susceptibility and risk. The current genetic studies provide the framework to pinpoint the genomic location and statistical properties of the genetic factors involved, but provide little insight into the specific functions in the cell or the body that are predisposing an individual. What one would like to know is what is the first cellular effect that is different between an individual who carries the predisposing allele and an individual who does not, and what are the reasons and means by which the predisposition is realized to a disease state. Gene expression is a critical phenotype that reveals such biochemical properties and allows us to dig into the cellular functions. Combining sophisticated statistical methods with relevant sample collections of tissues and cell types from well-phenotyped individuals enables the integrated treatment of biological and epidemiological information in an iterative way. This provides us with the highest possible resolution and will reveal the real causes for disease predisposition. Such collections are becoming a reality now through new sample collections. Finally, dissecting the genetics of cellular processes will not only revolutionize medical sciences but also will provide very important clues for basic biology and understanding of biological systems since we will have in our hands a tremendous number of natural 'mutants' that perturb the cellular processes and we can measure their effect from the cell to the whole organism (Fig. 2). It is probably not a stretch to say that there are great opportunities for human cellular systems to become the basis for the progress of systems biology.

Conflict of Interest statement. None declared.

FUNDING

Funding was provided by the Wellcome Trust and the Louis-Jeantet foundation.

REFERENCES

1. Dermitzakis, E.T. (2008) From gene expression to disease risk. *Nat. Genet.*, **40**, 492–493.
2. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
3. Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
4. Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A. and Schadt, E.E. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.
5. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
6. 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. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
7. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. and Burdick, J.T. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.
8. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
9. Cotsapas, C.J., Williams, R.B., Pulvers, J.N., Nott, D.J., Chan, E.K., Cowley, M.J. and Little, P.F. (2006) Genetic dissection of gene regulation in multiple mouse tissues. *Mamm. Genome*, **17**, 490–495.
10. Cowley, M.J., Cotsapas, C.J., Williams, R.B., Chan, E.K., Pulvers, J.N., Liu, M.Y., Luo, O.J., Nott, D.J. and Little, P.F. (2009) Intra- and

inter-individual genetic differences in gene expression. *Mamm. Genome*, **20**, 281–295.
11. Huang, G.J., Shifman, S., Valdar, W., Johannesson, M., Yalcin, B., Taylor, M.S., Taylor, J.M., Mott, R. and Flint, J. (2009) High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.*, **19**, 1133–1140.
12. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
13. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
14. Heinzen, E.L., Ge, D., Cronin, K.D., Maia, J.M., Shianna, K.V., Gabriel, W.N., Welsh-Bohmer, K.A., Hulette, C.M., Denny, T.N. and Goldstein, D.B. (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.*, **6**, e1.
15. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M.G., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. doi:10.1126/science.1174148. Epub ahead of print 30 July 2009.
16. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.
17. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
18. Petretto, E., Mangion, J., Dickens, N.J., Cook, S.A., Kumaran, M.K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M. *et al.* (2006) Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.*, **2**, e172.
19. Storey, J.D., Madeoy, J., Strout, J.L., Wurfel, M., Ronald, J. and Akey, J.M. (2007) Gene-expression variation within and among human populations. *Am. J. Hum. Genet.*, **80**, 502–509.
20. Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., Ewens, W.J. and Cheung, V.G. (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.*, **39**, 226–231.
21. Price, A.L., Patterson, N., Hancks, D.C., Myers, S., Reich, D., Cheung, V.G. and Spielman, R.S. (2008) Effects of *cis* and *trans* genetic ancestry on gene expression in African Americans. *PLoS Genet.*, **4**, e1000294.
22. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
23. Stegle, O., Kannan, A., Durbin, R. and Winn, J. (2008) *Accounting for non-genetic factors improves the power of eQTL studies*. Lecture Notes in Computer Science (RECOMB) 2008, pp. 411–422 (ISBN 978-3-540-78838-6).
24. Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.
25. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
26. Chen, X., Wang, L., Smith, J.D. and Zhang, B. (2008) Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, **24**, 2474–2481.
27. Kang, H.M., Ye, C. and Eskin, E. (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.
28. Park, C.C., Ahn, S., Bloom, J.S., Lin, A., Wang, R.T., Wu, T., Sekar, A., Khan, A.H., Farr, C.J., Lusis, A.J. *et al.* (2008) Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids. *Nat. Genet.*, **40**, 421–429.
29. Bolouri, H. (2008) *Computational Modeling of Gene Regulatory Networks—A Primer*. Imperial College Press, London, UK.
30. Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E. and Schadt, E.E. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, **40**, 854–861.
31. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D. and Daly, M.J. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.
32. Wang, K., Li, M. and Bucan, M. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
33. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C. and Craddock, N. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.
34. Hardy, J. and Singleton, A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.
35. Dimas, A.S., Stranger, B.E., Beazley, C., Finn, R.D., Ingle, C.E., Forrest, M.S., Ritchie, M.E., Deloukas, P., Tavare, S. and Dermitzakis, E.T. (2008) Modifier effects between regulatory and protein-coding variation. *PLoS Genet.*, **4**, e1000244.
36. Idaghdour, Y., Storey, J.D., Jadallah, S.J. and Gibson, G. (2008) A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet.*, **4**, e1000052.
37. Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C. *et al.* (2008) Genetic analysis of human traits *in vitro*: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, **4**, e1000287.
38. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
39. Gaidatzis, D., Jacobeit, K., Oakeley, E.J. and Stadler, M.B. (2009) Overestimation of alternative splicing caused by variable probe characteristics in exon arrays. *Nucleic Acids Res*. doi:10.1093/nar/gkp508. Epub ahead of print 15 June 2009.
40. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
41. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
42. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
43. Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W. and Shi, H. (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, **67**, 8511–8518.
44. Korshunova, Y., Maloney, R.K., Lakey, N., Citek, R.W., Bacher, B., Budiman, A., Ordway, J.M., McCombie, W.R., Leon, J., Jeddeloh, J.A. *et al.* (2008) Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res.*, **18**, 19–29.
45. Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J. *et al.* (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet.*, **5**, e1000569.
46. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein–DNA interactions. *Science*, **316**, 1497–1502.
47. Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
48. Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.