Review

The study of eQTL variations by RNA-seq: from SNPs to phenotypes

Jacek Majewski and Tomi Pastinen

Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield Avenue, Rm 7210, Montreal, Quebec, H3A 1A4, Canada

Common DNA variants alter the expression levels and patterns of many human genes. Loci responsible for this genetic control are known as expression quantitative trait loci (eQTLs). The resulting variation of gene expression across individuals has been postulated to be a determinant of phenotypic variation and susceptibility to complex disease. In the past, the application of expression microarray and genetic variation data to study populations enabled the rapid identification of eQTLs in model organisms and humans. Now, a new technology promises to revolutionize the field. Massively parallel RNA sequencing (RNA-seq) provides unprecedented resolution, allowing us to accurately monitor not only the expression output of each genomic locus but also reconstruct and quantify alternatively spliced transcripts. RNA-seq also provides new insights into the regulatory mechanisms underlying eQTLs. Here, we discuss the major advances introduced by RNA-seq and summarize current progress towards understanding the role of eQTLs in determining human phenotypic diversity.

Complex traits and common variants in humans: noncoding DNA takes center stage

The majority of mutations underlying monogenic disease traits alter protein structure. As a consequence of this observation, protein-coding variants were the primary candidates in the early search of susceptibility alleles for multifactorial, complex disease traits [1]. However, the first 5 years of genome-wide association studies (GWAS) for complex disease have shown that if mapping had been restricted to coding variants alone, only approximately 5% of the currently validated disease associations would have been discovered [2]. Thus, the dissection of the genetic architecture of human disease is now focused on variants residing outside of coding regions; that is variants that potentially affect regulatory elements. A well-known example, the lactase persistence phenotype in European populations, was mapped 5′ distal to *LCT*, a gene coding for the lactase enzyme in the small intestine [3]. This noncoding stretch of DNA within an intron of an adjacent gene and with no known function was subsequently shown to contain a distal enhancer specific to enterocytes producing lactase in the digestive track [4]. The replication of this association in other ethnicities [5] confirmed the role of nucleotide substitutions within this element in regulating *LCT* expression and explaining population differences in their abilities to digest milk sugar (lactose tolerance). The more widespread importance of noncoding or regulatory DNA alterations in disease as implied by GWAS now calls for approaches to characterize such a variation and its links to disease phenotypes.

Genome-wide identification of loci controlling gene expression

The parallel assessment of thousands of transcripts using DNA microarrays is clearly one of the revolutionary technologies that launched the 'genomic' era. The genome-wide association of genetic and transcriptome variations was first achieved in yeast [6], where expression traits of the progeny were shown to be largely correlated with the genetic contribution of parental genotypes. The excitement of observing thousands of quantitative traits, or eQTLs, in a technically straightforward experiment quickly spread to studies in more complex genomes [7] including the human genome [8]. Several eQTL studies in humans

Glossary

Expression quantitative trait loci (eQTL): Term most commonly used to describe a statistically significant genotype–gene expression level correlation. Expression is detected by using microarrays or RNA-seq, and genotypes can be collected at a high density (typical for association-based mapping) or lower density (utilized in family- or model organism-based linkage or eQTL mapping).

Genome-wide association studies (GWAS): GWAS use large case-control cohorts of individual- or population-based samples with quantitative phenotypic data (such as height or lipid levels), which are characterized for genetic variation at a high density, e.g. 500 000 to 1 000 000 genotypes collected across the genome. The links between polymorphisms and disease risk or quantitative traits are then observed by the point-wise assessment of genetic marker alleles for enrichment among cases or among tails of distribution for quantitative phenotypes.

Linkage disequilibrium (LD): The nonrandom association of alleles at different loci. LD is usually the result of a close physical location and lack of recombination between loci. One of the consequences of high LD in the human genome is the presence of haplotype blocks consisting of large numbers of polymorphic markers that can be grouped into a limited number of haplotypes.

Next-generation sequencing (NGS): Techniques based on the amplification and sequencing of short stretches of DNA in parallel for millions of individual target molecules using randomly ordered arrays or the suspensions of sheared target molecules.

Paired-end reads: NGS targets are generally short sheared DNA fragments that can be sequenced from one or both ends of the fragment. The latter approach allows the collection of physically linked, or paired-end reads, facilitating the mapping and understanding of polynucleotide sequences beyond the read length of a single NGS read.

RNA sequencing (RNA-seq): NGS application for RNA species present in a sample. Typically, mRNA is isolated from a tissue of interest, converted into cDNA and sheared into smaller fragments, millions of which can be sequenced in parallel using one of the NGS technologies. Aligning these short fragments to the genome can explain the sequence composition in mRNA, expression level (based on the number of overlapping sequences to a specific gene) and gene structure (based on splice junctions).

Corresponding author: Majewski, J. (jacek.majewski@mcgill.ca).

0168-9525/$ – see front matter doi:10.1016/j.tig.2010.10.006 Trends in Genetics, February 2011, Vol. 27, No. 2

Box 1. The study of eQTL in humans

In humans, identifying eQTLs is usually carried out by analyzing the linkage or association [8,11,42,82] between gene expression levels and genetic markers in *cis* (within a preselected interval close to the gene) or in *trans* (distant or located on different chromosomes). In the genomic era, many screens for eQTLs have been carried out by measuring the expression levels of a large number of genes and testing them for linkage (in families) or association (in populations) with a large number of genetic markers. Although such approaches have enjoyed some success, as in any whole-genome analysis false positive results can be introduced due to multiple testing problems – for example, when testing tens of thousands of genes against millions of SNP markers – and systematic errors related to the specific genomic technologies used. Hence, the choice of a most accurate gene expression assay is a crucial component of eQTL surveys. A recent suggestion is that RNA-seq can provide the more accurate assessment of expression, and extending this technology to studies of population variation could potentially provide refined information at the isoform, transcript and allelic expression levels. With this approach even minute changes in the levels of the expression of genes are detectable, but detecting variations in relative isoform abundance or allelic expression can need substantially higher coverage than for observing population variation in full transcript expression [25,26,59]. In RNA-seq, establishing optimal correction for known and hidden technical biases in experiments [26] as well as modeling for isoform structures based on short-read data [25] are crucial. However, a generalizable approach has yet to emerge given the rapid progression of NGS technology in terms of throughput and read length, which are both influencing the choice of raw data processing.

have subsequently been conducted using family-based samples of cultured or purified cells [9–12] (Box 1), as well as population-based samples of cultured cells [13–15], purified cell populations [16,17] or complex tissues [18–20]. Similarly, the microarray platforms used in these studies have evolved from arrays querying only well-known transcripts [9] to ones assessing all known and predicted exons [21]. Despite diverse study designs and expression microarray platforms, some common observations have emerged. First, local genetic associations, which are assumed to act directly or in *cis* on target gene regulation, have a strong influence and can often be observed across studies and validated by independent methods. Second, distal genetic associations acting in *trans* have more subtle effects, appear more numerous in the genome and are considerably more difficult to validate [10,22]. Consequently, *trans*-eQTLs might not be replicable across studies and their validity is a much more contentious issue compared with *cis*-eQTLs. Furthermore, although studies in mice have established the biological reproducibility of eQTL mapping on the same microarray platform [23], similar studies have not been performed in human tissues or across microarray platforms. Therefore, despite reports of thousands of genetic associations contributing to population variation in gene expression, there is currently no consensus as to what is the optimal approach to explore gene expression variation comprehensively, or a set of most robust and replicable eQTLs in humans. Some of these issues could be solved by applying a yet undetermined 'gold standard' methodology to transcriptomic data, which would avoid biases related to the sparse sampling of complex gene structures [21], be resistant to spurious signals due to related gene sequences or genetic variants [24] and have the ability to measure transcripts not known in the public domain.

RNA-seq could provide a platform-independent and objective standard compared with the microarray approach (Box 2). Recent RNA-seq-based eQTL studies have both confirmed and further clarified previous microarray results [25,26]. The first comparisons of *cis*-eQTL detection by RNA-seq compared with microarray-based approaches are promising; when disagreeing results between two approaches have been detected, the RNA-seq data have more frequently matched the allelic biases observed by the Sanger sequencing-based validation method [27]. Importantly, sequencing technologies are advancing so rapidly that even the most recently published RNA-seq studies have used already outdated platforms, with low sequencing coverage and short reads. Technologies available today allow much higher coverage and longer reads at a reduced cost. In parallel with these incremental improvements, the introduction of 'third generation sequencing' [28] promises to allow simple sample preparation (without the need for amplification) and longer read lengths (thousands of bases), resulting in a more direct assessment of RNA abundance and mRNA isoforms.

In the following sections, we describe in detail the major advances brought about by RNA-seq that have allowed us to identify the eQTLs responsible for variations at the transcript, isoform and allele levels. We also outline the

Box 2. A comparison of microarray and RNA-seq approaches

The genome-wide profiling of gene expression has traditionally been carried out using microarrays. The most common 3′ targeted microarrays contained probes predominantly located in the 3′ UTRs of genes. This design, implemented in popular Affymetrix and Illumina expression microarrays, attempted to target areas common to most or all isoforms produced by a genomic locus to represent the total mRNA output of a locus, regardless of how the transcript was spliced and polyadenylated. A more comprehensive design, as used by the Affymetrix Exon Array, targeted probes to individual exons. Combining the exons into transcriptional units still allowed monitoring expression at a whole 'gene' level, whereas analyzing exons individually identified changes at the level of distinct, alternatively spliced isoforms. Several groups made inroads into the genomic analysis of splicing using exon arrays [21,83]; however, the analysis was complicated and the results noisy due to the limitations imposed by small probe design regions (exons). A higher level of resolution at the splicing level could be obtained with arrays composed of both exon and junction-targeted probes [84,85]. Such approaches were sometimes implemented in custom-designed platforms, but required *a priori* knowledge of the isoforms to be targeted, and, largely because of elevated costs, complexity of design and analysis, junction microarrays have not achieved mainstream popularity. RNA-seq combines all the advantages of the different array designs mentioned above, but overcomes many of the shortcomings [35]. Large numbers, typically tens to hundreds of millions of random sequence tags, are produced and can be used to monitor both the expression and isoform structure of each gene (Figure 1). No selection of target regions is needed prior to the experiment because the random sequence tags can be mapped to all known annotated transcripts – to exons and across exon–exon junctions. The sequence data can also be used to identify yet unknown isoforms and new genes [86]. False positive results are generally a consequence of sequence alignment artifacts, particularly in the case of gene families and repetitive sequences. However, many of the artifacts can be removed after carrying out the experiment using rapidly improving bioinformatic approaches [86] – a strategy that was not possible for analog, hybridization-based techniques.

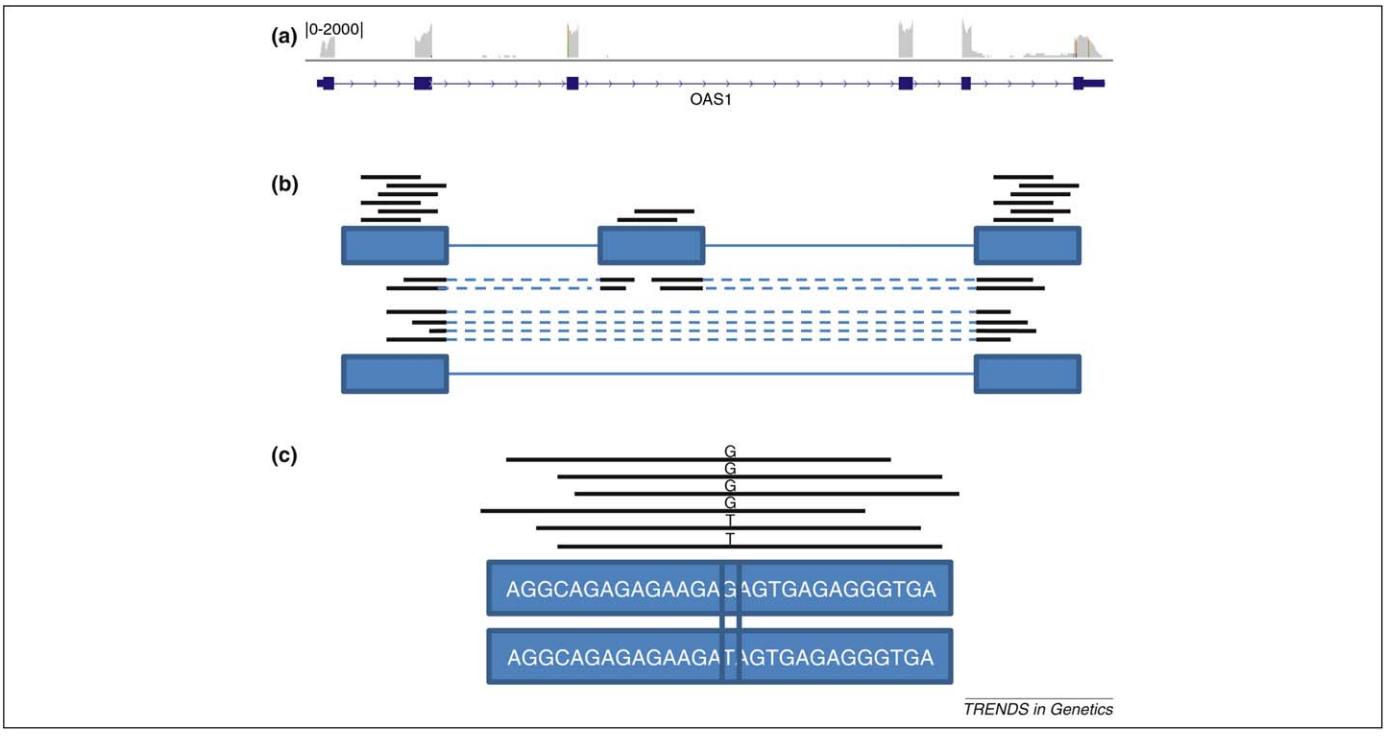

Figure 1. The use of RNA-seq to profile gene expression. (**a**) A gene-level view illustrating sequencing reads (light gray track) mapped to the genome (dark gray line). The height of the upper track represents the number of sequences covering any given region. Most of the reads will map to exons (blue rectangles), and averaging the number of reads over all exons provides an estimate of the total transcriptional output of a given locus. (**b**) An exon-level view showing individual reads (black lines) mapping to exons (exon body reads) and across junctions (junction reads, with gaps in the alignment indicated by dotted lines). A combination of exonic and junction reads can be used to estimate the frequency of alternative splicing. In this example, the middle exon appears to be alternatively spliced and included in one out of three transcripts. (**c**) At heterozygous positions, reads containing each allele of a SNP can be used to detect allelic imbalance. Here, the G allele appears to be expressed at a level twice as high as that of the T allele.

remaining hurdles that must be overcome to connect the increased understanding of eQTLs with the phenotypic traits that they potentially influence.

Investigating eQTLs affecting alternative splicing using RNA-seq

The human genome contains surprisingly few, in fact fewer than 21 000, distinct protein-coding genomic loci [29]. This number is only slightly higher than the number of genes present in much less complex organisms, such as the fruit fly and the nematode, and is lower than that of most plants. One of the mechanisms for creating complexity using a limited number of loci is through alternative pre-mRNA splicing (AS) [30], which is the variety of ways exons within a gene are joined together to form mRNA molecules. Recent studies have estimated that the vast majority of mammalian genes (more than 95%) produce more than one isoform [31]. Such isoforms might be produced via the alternative initiation of transcription, usage of alternative polyadenylation sites or alternative splicing of internal exons.

Variation in pre-mRNA splicing is known to be responsible for introducing phenotypic diversity within human populations. It is estimated that 10% of all mutations involved in human Mendelian disorders affect splicing [32]. Even subtle differences in the ratio of AS isoforms might lead to visible phenotypic effects, as illustrated by splicing mutations causing frontotemporal dementia [33]. Hence, it is important to identify genetic variants influencing the quantitative expression of individual alternatively spliced isoforms. Such variants have been referred to as isoform eQTLs or sQTLs (where the s stands for splicing).

RNA-seq can be used to profile individual alternatively spliced isoforms of a gene. The most useful feature of RNA-seq data is the presence of sequence reads that map to splice junctions. Mutually exclusive splice junctions (i.e. those joining alternatively spliced exons) can then be quantified to estimate the relative abundance of alternative isoforms. Novel spliced isoforms can be detected by joining together all annotated exons, and novel exons can be discovered by the use of *de novo* spliced alignment algorithms [34]. Several approaches already exist that combine junction- and exon-mapping reads (Figure 1) to fully optimize the reconstruction and quantification of individual isoforms [26,35–37]. Current sequencing reads, which are commonly shorter than 100 nucleotides, allow only the monitoring of a single splice junction at a time. However, as the read lengths increase, profiling splice variants at a single molecule level, which will allow the detection of combinatorial splicing patterns, will become possible.

In previous microarray approaches to the genome-wide analysis of AS, the nature of each alternative event had to be deduced from the hybridization levels of individual probes [38] and RT-PCR reactions had to be designed to verify the expected products, which in turn had to be quantified and sequenced [39]. RNA-seq data provide quantification and sequence information in one fast and convenient, albeit still expensive, step. False positive results still occur – most often because of sequence alignment artifacts – however, at a rate much lower than in

microarray studies. A recent study benchmarking the efficacy of RNA-seq in detecting AS obtained a greater than 85% validation rate of alternative events using RT-PCR [35].

Microarrays, along with RT-PCR validation, have been previously used to detect dozens of genetically regulated alternatively spliced isoforms [21,40]. RNA-seq data raise this number into the hundreds, confirming that the genetic regulation of pre-mRNA splicing is indeed prevalent in human populations. In addition, the new data show that some previous microarray results confounded splicing and whole-transcript changes, particularly in experiments where microarrays targeting specific transcript regions were used [26]. Hence, by clearly identifying the precise location of isoform level changes, RNA-seq will also facilitate the identification of nearby causative polymorphisms.

Investigating eQTLs affecting transcription using RNA-seq

The regulation of gene expression has been postulated to be achieved at three levels: transcriptional, cotranscriptional and post-transcriptional [41]. Transcription levels might be affected by modifying the efficiency of initiation, as well as the speed of transcription. Cotranscriptional regulation refers to mRNA processing, such as splicing, 5′ capping or polyadenylation. Post-transcriptional regulation refers to processes affecting the mRNA molecule after the transcript has been completed and has fully dissociated from the DNA strand; these processes include mRNA stability, antisense RNA-mediated degradation and nonsense-mediated decay. It should be noted that all these processes are to some extent coupled and interdependent; however, the regulatory variants underlying eQTLs will most likely affect only one of the processes at a time. Understanding the regulatory mechanisms underlying eQTLs and being able to identify the precise regulatory variants is essential if we are to use eQTLs as diagnostic markers of disease or to design therapeutic approaches targeting the causative polymorphisms.

Very few regulatory single nucleotide polymorphisms (SNPs) discovered by eQTL mapping have been experimentally validated, and there is some controversy as to what their predominant regulatory mechanisms are. Initially, it was assumed that regulatory SNPs act by affecting transcription factor binding sites. This hypothesis was supported by early results obtained for the chitinase 3-like 2 gene (*CHI3L2)* [42], where a SNP in the promoter was implicated in recruiting variable amounts of RNA polymerase, implying regulation at the level of transcription. However, because of the ubiquitous presence of extended regulatory haplotypes – comprised of many SNPs often in perfect linkage disequilibrium that are all associated with the expression level of a given gene – and difficulty in demonstrating causation, this result has not yet been generalized. More recently, an alternative hypothesis to transcriptional regulation was suggested, proposing that a large fraction of eQTLs can be regulated by post-transcriptional mechanisms [43]. The authors observed an excess, over the null expectation, of regulatory SNPs in both the 5′ and 3′ regions of target genes. Hence, they proposed that upstream SNPs might act by regulating transcription, whereas a nearly equal number of the downstream regulatory variants – located in 3′ UTR regions of genes – might act post-transcriptionally by affecting RNA stability. This could take place by altering microRNA binding sites or polyadenylation patterns. One example of such a 3′ UTR variant has been detected in large-scale studies [21] and validated in the laboratory [44]: a SNP affecting a polyadenylation recognition element in the interferon regulatory factor 5 (*IRF5*) resulted in a longer, less stable isoform associated with slightly reduced expression levels of the protein [21].

However, recent RNA-seq data do not support 3′, post-transcriptional regulation as a major determinant of eQTL action. The more precise quantification of expression at the individual isoform level suggests that the number of 3′ regulatory SNPs is much lower than previously believed [25,26]. It is therefore possible that the effect observed in earlier studies was an artifact of probe placement in 3′ targeted microarrays, and that many of the presumed eQTLs are actually splicing differences occurring in the 3′ genic regions, rather than whole gene expression changes.

RNA-seq results also highlight a general relationship between eQTLs and sQTLs. It is probable that some SNPs that affect splicing patterns result in the production of less stable isoforms, which in turn affect the expression of the entire transcripts. At least one example of such a mechanism has been experimentally dissected so far [45]: *ERAP2*, encoding an endoplasmic reticulum-specific aminopeptidase, contains a SNP affecting an intronic donor splice site, resulting in alternative splice site usage and the insertion of a premature stop codon in the mRNA. This isoform has been shown to be a target for nonsense-mediated decay, and the causative splicing SNP is associated with both the splicing pattern and gene expression levels [40]. Because the level of mRNA of this gene is regulated post-transcriptionally, RNA-seq can be used to demonstrate that RNA expression differences cannot be detected in reads that map to intronic sequences (which represent the primary, unprocessed product of transcription), but can be readily seen in exonic reads (which represent the mature mRNA) [26].

Regulation at the level of pre-mRNA splicing is generally controlled by polymorphisms located in the vicinity of the splice sites [46]. Although it is still not feasible to confidently predict *a priori* the regulatory potential of a randomly selected SNP on splicing [47,48], once an sQTL is identified there has been a good level of success pinpointing the nearby regulatory variant [40]. The effects of those SNPs can be predicted *in silico* by determining the effect of each allele on the strength of consensus splice sites [49], of accessory splicing control elements (such as the polypyrimidine tract) or of exonic and intronic splicing enhancer and silencer elements. Downstream *in vivo* validation can be carried out using transient transfection of minigene constructs [50], which in most cases unequivocally confirms the function of the causative polymorphism.

Until now, because microarray probes were targeted to exons, it has been difficult to distinguish between the transcriptional, cotranscriptional or post-transcriptional regulatory mechanisms. By monitoring expression levels

across the entire transcript, without the need for a targeted design, RNA-seq allows us to deconvolute the expression levels of individual isoforms and significantly clarifies our understanding of eQTL regulation [25]. In addition, because the information contained within RNA-seq data is not only quantitative but also 'digital' – providing the identity of every single nucleotide within a transcript – this technique allows us to monitor not only expression levels but also allelic expression; that is the transcriptional output of each individual allele at polymorphic loci.

Direct detection of *cis*-regulatory variation by allele counting in RNA-seq data

The strongest genetic effects observed for the expression of individual genes predominantly localize close to the gene itself. These local associations are assumed to alter the function of regulatory elements directly controlling expression in *cis*. Such *cis*-regulatory variants should give rise to the unequal regulation of alleles in samples heterozygous at the site. To detect allelic expression, the expression profiling method needs to be able to distinguish the expressed alleles quantitatively. The test for differential allelic expression [51] can be used to independently validate local eQTLs and provide a specific approach to confirm *cis*-regulatory variation [52]. The relative expression of two alleles in autosomes (or the female X-chromosome) can be carried out genome-wide using genotyping microarrays [53]. The promise of allelic expression is the dissection of the genetic (or epigenetic) control of gene expression to its *cis*-and *trans*-acting components [54]. In principle, the allelic expression test provides greater power to detect genetic variants acting in *cis* compared with eQTL mapping [53] because allelic differences are measured within rather than between samples.

The quantitative assessment of allelic effects by genotyping methods is complicated by unequal signals from two alleles even when they are equally represented in the sample, thereby requiring the careful normalization of allele ratios [53]. Methods based on single molecule detection allow, in principle, the unbiased assessment of true allele ratios based on simple allele counting. The first measurements of allelic expression were realized even before next-generation sequencing (NGS) [55], and early applications of NGS in allelic expression used allele counting for specific amplified cDNAs [56]. The latter approach is not scalable so more recently the large-scale capture of specific fragments of cDNAs harboring genetic variation have been coupled with NGS [57,58], which can be used to detect allelic biases across thousands of polymorphic-expressed sites. The expansion of RNA-seq studies to numerous individuals and coupling genotype information with transcriptome analysis (i.e. population-based RNA-seq studies) will, in principle, also allow the assessment of population variation in the allelic expression and mapping of its genetic determinants. The first explorations of RNA-seq across human populations did not use allelic expression patterns as a tool to map *cis*-regulatory variation, but showed that this should be feasible, because the prevalence [59] of differential expression as well as concordance with *cis*-eQTLs were shown to be high [25,26]. At this early stage, the optimal approach for population RNA-seq studies is unclear, but paired-end sequencing coupled with counting alleles has been shown to be beneficial in deciphering genetically controlled alternative splicing patterns [25]. It has also been suggested that allelic expression states detected by RNA-seq in only a few individuals would be sufficiently powerful to observe effects exerted by rare *cis*-regulatory alleles [25]. To date, studies have shown that the investigation of allelic expression in addition to eQTLs in population RNA-seq data can enhance the detection of genetic variation [25,26], although this would optimally require higher sequence depth than that applied to date because of the relatively sparse distribution of variation in human coding regions [60]. This technical limitation will easily be overcome with the decreasing cost of NGS. Similarly, longer reads [28] will render a high proportion of reads informative for the allelic state, thereby improving the ability to study allelic and total expression levels in parallel.

The allele specificity of RNA-seq has also been used to study the epigenetic mechanisms of *cis*-regulation in inbred mice [61,62]. The results of two early studies were however discordant, yielding strikingly different estimates of imprinting prevalence in the developing mouse brain. One of the studies suggested that there could be over a thousand imprinted genes based on paternal or maternal allelic expression biases observed in RNA-seq data [61]. By contrast, the other study concluded that less than a hundred of the already known imprinted genes would represent a relatively complete list [62]. Such discrepancies between studies with seemingly similar designs highlight the early stage of the interpretation of allelic information from RNA-seq datasets. Allelic data from RNA-seq can be technically biased by steps in the sample (library) preparation or the actual base incorporation rates in sequencing [59] as well as analytical differences that can arise depending on alignment approach [63]. These systematic biases are in addition to stochastic variation at individual regions inherent to large-scale experiments. These issues will prompt the development of common guidelines to filter, validate and normalize RNA-seq data, similar to the guidelines established for microarrays [64].

Hurdles in the large-scale translation of population expression variation into biological insight

One of the greatest challenges in eQTL analysis is to prove the causal relationship between disease processes and gene regulatory changes. An excess of eQTLs among disease-associated SNPs has been reported [65,66]. The interpretation of these enrichments is complicated because they tell little about the role of individual variants. Moreover, the observed excesses might be biased towards non-causal links simply because disease variants and eQTLs might both be more commonly observed in genomic regions harboring expressed genes. Direct evidence of common regulatory alleles in human disease phenotypes, similar to the *LCT* variants involved in lactose tolerance, has been found for only a few genes. A change in TATA-binding protein binding to dinucleotide repeat in the promoter of *UGT1A1,* coding for a UDP-glucuronosyl transferase crucial for bilirubin metabolism in the liver [67], explains the strong association of Gilbert's syndrome with this variant [68]. The same variant has since turned out to be involved

in the pharmacogenetics of multiple drugs [69]. In the case of malaria resistance, a SNP in the *FY* promoter altering GATA-1 binding abolishes the expression of the Duffy antigen, a glycoprotein present on the erythrocyte surface in the erythroid cell lineage [70]. The first 'modern' example exploiting the overlap between GWAS and eQTL data to identify and molecularly dissect a disease-coding variant showed that a common SNP upstream of sortilin (*SORT1*) alters the binding of the CEBP transcription factor in hepatocytes, leading to differences in circulating LDL levels [71]. These successes in linking gene expression-altering polymorphisms to complex phenotypes share some features: in all cases, there is strong genetic evidence for the involvement of a particular variant in the phenotype observed, and each regulatory variant has been shown to exert its function primarily in the tissue or cell-type of interest, whereas many eQTLs show effects across tissues [72]. Therefore, establishing the links between regulatory variants and the molecular processes directly linked to the phenotypes is possible.

Therefore, can we expect an increase in the identification of mechanistic links between genotypes and phenotypes in the coming years with the large numbers of disease associations already characterized and the emergence of eQTL surveys? Describing molecular processes underlying gene expression variation in metabolic phenotypes in mice has allowed the identification of a large gene network, where new candidate regulators for metabolism have emerged by perturbing key genes [73]. By contrast, the use of a similar approach to study human metabolic phenotypes, although showing an enrichment of obesity- or diabetes-associated alleles among SNPs associated with expression profiles in adipose or liver tissues, did not identify new genes or provide mechanistic insights into the function of individual variants [18,74]. These latter studies highlight the challenge of correlating the relatively weak marginal effects of individual human complex disease-associated SNPs with eQTL associations; both often exist in a genomic region with multiple correlated variants and provide only weak evidence for causality without extensive follow-up experiments [71].

The serendipitous overlap between regulatory and disease-associated variants can be expected to increase with the higher resolution and higher discovery rates of RNA-seq-based eQTL studies [25,26]. Consequently, in cases where eQTLs have been characterized in tissues with known relevance to disease processes, a necessary step in establishing the link between regulation and phenotype is the isolation of regulatory variants underlying the eQTLs. Causal variants can be identified by combining independent methods measuring both gene expression and transcription in the same samples to assess allelic function at individual loci [56] or genome-wide [75]. Given the multitude of potential connections in complex gene networks between heritable variants governing gene expression and complex disease association data [74] finding causal variants in the appropriate tissues will be insufficient. Links between disease phenotypes and common regulatory variants are facilitated for traits where subsets of patients harbor highly penetrant, deleterious coding variants unequivocally linking the gene product with disease pathogenesis [76]. Depending on the architecture of complex traits and the success of sequence-based complex trait studies [77], such a strategy of first establishing links between rare variants in specific genes and complex phenotypes can provide stronger *a priori* evidence for making causal connections between common regulatory variation and phenotypic variation in populations. In addition to ensuring that phenotype and expression traits map to the same variants and to establishing the role of genes in the disease trait, the downstream effects of regulatory SNPs could be observed and verified by other high-throughput methods linking protein levels [78] or metabolites [79] to the same variants. However, these methods currently do not have adequate coverage or sensitivity to provide genome-wide measurements of downstream consequences of eQTLs (Box 3).

Box 3. Transcriptomics, proteomics or metabolomics to measure population variation at a genome-wide scale

eQTLs have been considered an intermediate phenotype that might connect regulatory genetic variants and phenotypes. However, and particularly in view of the increasingly large number of small effect eQTLs, it would be interesting to consider the next level of intermediate phenotype – namely the protein expression level. Currently, it is unclear to what extent compensatory mechanisms at the level of translation might overcome mRNA-level differences and stabilize potential functional variations. Such compensatory mechanisms are suggested by the observation of higher evolutionary correlations among expressed proteins than among expressed transcripts [87]. In the case of sQTLs and other isoform changes, many of the observed splicing variants might actually not result in the production of stable proteins. Some, or many, might represent an increased level of background noise, and be subject to correction by processes such as nonsense-mediated decay [88]. However, although protein abundance in yeast shows a strong correlation with corresponding mRNA abundance [89], less than 10% of the yeast protein products could be measured with sufficient accuracy to carry out genome-wide eQTL mapping [90]. Despite great strides in quantitative proteomics [91], the application of shotgun (non-targeted) proteomics in complex organisms still suffers from low dynamic range and suboptimal sensitivity. Early attempts to associate human protein production [78] or metabolites [79] with genetic variants, while providing a proof-of-principle, still await a more comprehensive demonstration of the utility of proteomics in population studies to decipher functional variation at a genome-wide scale.

Concluding remarks

NGS technologies offer functional genomic data at a higher resolution than DNA microarrays. In parallel, these same technologies are enabling studies of virtually all common genetic variants in association studies [80] as well as the high-throughput identification of regulatory elements in our genomes [81]. The integrative analysis of variation in transcriptome, sequence and regulatory sites using population-based samples is consequently expected to result in unprecedented accuracy in both mapping as well as understanding the consequences of regulatory variation at single base resolution. The next challenge is the large-scale translation of these functional variants into insights into the molecular pathogenesis of complex traits. The pace of discovery for links between regulatory variation and disease will increase, particularly with parallel improvements in disease variant fine mapping and the development of genome-wide post-transcriptomic techniques. Therefore, early isolated successes in characterizing the roles of

individual regulatory variants in disease [71] are likely to become more frequent and provide us with the mechanistic basis of genetic disease associations and human gene regulation.

References

1 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.* 33 (Suppl.), 228–237
2 Hindorff, L.A. *et al.* (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
3 Enattah, N.S. *et al.* (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30, 233–237
4 Lewinsky, R.H. *et al.* (2005) T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity *in vitro*. *Hum. Mol. Genet.* 14, 3945–3953
5 Ingram, C.J. *et al.* (2009) Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J. Mol. Evol.* 69, 579–588
6 Brem, R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755
7 Schadt, E.E. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302
8 Morley, M. *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747
9 Cheung, V.G. *et al.* (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33, 422–425
10 Cheung, V.G. *et al.* (2010) Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* 8
11 Goring, H.H. *et al.* (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 39, 1208–1216
12 Monks, S.A. *et al.* (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–1105
13 Dixon, A.L. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–1207
14 Grundberg, E. *et al.* (2009) Population genomics in a disease targeted primary cell model. *Genome Res.* 19, 1942–1952
15 Stranger, B.E. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.* 39, 1217–1224
16 Murphy, A. *et al.* (2010) Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* 19, 4745–4757
17 Zeller, T. *et al.* (2010) Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5, e10693
18 Emilsson, V. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature* 452, 423–428
19 Heinzen, E.L. *et al.* (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* 6, e1
20 Myers, A.J. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.* 39, 1494–1499
21 Kwan, T. *et al.* (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* 40, 225–231
22 Smirnov, D.A. *et al.* (2009) Genetic analysis of radiation-induced changes in human gene expression. *Nature* 459, 587–591
23 van Nas, A. *et al.* (2010) Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics* 185, 1059–1068
24 Benovoy, D. *et al.* (2008) Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res.* 36, 4417–4423
25 Montgomery, S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777
26 Pickrell, J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772
27 Babak, T. *et al.* (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. *BMC Genomics* 11, 473
28 Schadt, E.E. *et al.* (2010) A window into third generation sequencing. *Hum. Mol. Genet.* DOI: 10.1093/hmg/ddq416
29 Clamp, M. *et al.* (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19428–19433
30 Nilsen, T.W. and Graveley, B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463, 457–463
31 Wang, E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476
32 Cooper, D.N. (2010) Functional intronic polymorphisms: buried treasure awaiting discovery within our genes. *Hum. Genomics* 4, 284–288
33 Faustino, N.A. and Cooper, T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.* 17, 419–437
34 Trapnell, C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111
35 Griffith, M. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods* 7, 843–847
36 Jiang, H. and Wong, W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25, 1026–1032
37 Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515
38 Hiller, D. *et al.* (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* 25, 3056–3059
39 Kwan, T. *et al.* (2007) Heritability of alternative splicing in the human genome. *Genome Res.* 17, 1210–1218
40 Coulombe-Huntington, J. *et al.* (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* 5, e1000766
41 Pandit, S. *et al.* (2008) Functional integration of transcriptional and RNA processing machineries. *Curr. Opin. Cell Biol.* 20, 260–265
42 Cheung, V.G. *et al.* (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–1369
43 Veyrieras, J.B. *et al.* (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4, e1000214
44 Graham, R.R. *et al.* (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U. S. A.* 104, 6758–6763
45 Pinyol, M. *et al.* (2007) Inactivation of RB1 in mantle-cell lymphoma detected by nonsense-mediated mRNA decay pathway inhibition and microarray analysis. *Blood* 109, 5422–5429
46 Majewski, J. and Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12, 1827–1836
47 ElSharawy, A. *et al.* (2009) Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum. Mutat.* 30, 625–632
48 Hull, J. *et al.* (2007) Identification of common genetic variation that modulates alternative splicing. *PLoS Genet.* 3, e99
49 Yeo, G. and Burge, C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11, 377–394
50 Singh, G. and Cooper, T.A. (2006) Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques* 41, 177–181
51 Yan, H. *et al.* (2002) Allelic variation in human gene expression. *Science* 297, 1143
52 Pastinen, T. and Hudson, T.J. (2004) Cis-acting regulatory variation in the human genome. *Science* 306, 647–650
53 Ge, B. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* 41, 1216–1222
54 Wittkopp, P.J. *et al.* (2004) Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88
55 Dressman, D. *et al.* (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8817–8822
56 Verlaan, D.J. *et al.* (2009) Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res.* 19, 118–127
57 Lee, J.H. *et al.* (2009) A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet.* 5, e1000718
58 Zhang, K. *et al.* (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* 6, 613–618

59 Heap, G.A. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* 19, 122–134
60 Fontanillas, P. *et al.* (2010) Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol. Ecol.* 19 (Suppl. 1), 212–227
61 Gregg, C. *et al.* (2010) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* 329, 682–685
62 Wang, X. *et al.* (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* 3, e3839
63 Degner, J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25, 3207–3212
64 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371
65 Nica, A.C. *et al.* (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895
66 Nicolae, D.L. *et al.* (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888
67 Hsieh, T.Y. *et al.* (2007) Molecular pathogenesis of Gilbert's syndrome: decreased TATA-binding protein binding affinity of UGT1A1 gene promoter. *Pharmacogenet. Genomics* 17, 229–236
68 Bosma, P.J. *et al.* (1995) The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N. Engl. J. Med.* 333, 1171–1175
69 Strassburg, C.P. (2010) Gilbert-Meulengracht's syndrome and pharmacogenetics: is jaundice just the tip of the iceberg? *Drug Metab. Rev.* 42, 162–175
70 Tournamille, C. *et al.* (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat. Genet.* 10, 224–228
71 Musunuru, K. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719
72 Dimas, A.S. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–1250
73 Chen, Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452, 429–435
74 Zhong, H. *et al.* (2010) Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* 6, e1000932
75 McDaniell, R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328, 235–239
76 Emison, E.S. *et al.* (2010) Differential contributions of rare and common, coding and noncoding ret mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* 87, 60–74
77 Altshuler, D. *et al.* (2008) Genetic mapping in human disease. *Science* 322, 881–888
78 Melzer, D. *et al.* (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* 4, e1000072
79 Gieger, C. *et al.* (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* 4, e1000282
80 Liu, J.Z. *et al.* (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* 42, 436–440
81 Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680
82 Stranger, B.E. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1, e78
83 Sandberg, R. *et al.* (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647
84 Calarco, J.A. *et al.* (2007) Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.* 21, 2963–2975
85 Castle, J.C. *et al.* (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat. Genet.* 40, 1416–1425
86 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
87 Schrimpf, S.P. *et al.* (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 7, e48
88 Lareau, L.F. *et al.* (2007) The coupling of alternative splicing and nonsense-mediated mRNA decay. *Adv. Exp. Med. Biol.* 623, 190–211
89 Lu, P. *et al.* (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25, 117–124
90 Foss, E.J. *et al.* (2007) Genetic basis of proteome variation in yeast. *Nat. Genet.* 39, 1369–1375
91 Domon, B. and Aebersold, R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* 28, 710–721