European Journal of Human Genetics (2013) 21, 134–142

www.nature.com/ejhg

REVIEW

RNA-Seq and human complex diseases: recent accomplishments and future perspectives

Valerio Costa*,[1], Marianna Aprile[1], Roberta Esposito[1] and Alfredo Ciccodicola*,[1]

The availability of the human genome sequence has allowed identification of disease-causing mutations in many Mendelian disorders, and detection of significant associations of nucleotide polymorphisms to complex diseases and traits. Despite these progresses, finding the causative variations for most of the common diseases remains a complex task. Several studies have shown gene expression analyses provide a quite unbiased way to investigate complex traits and common disorders' pathogenesis. Therefore, whole-transcriptome analysis is increasingly acquiring a key role in the knowledge of mechanisms responsible for complex diseases. Hybridization- and tag-based technologies have elucidated the involvement of multiple genes and pathways in pathological conditions, providing insights into the expression of thousand of coding and noncoding RNAs, such as microRNAs. However, the introduction of Next-Generation Sequencing, particularly of RNA-Seq, has overcome some drawbacks of previously used technologies. Identifying, in a single experiment, potentially novel genes/exons and splice isoforms, RNA editing, fusion transcripts and allele-specific expression are some of its advantages. RNA-Seq has been fruitfully applied to study cancer and host-pathogens interactions, and it is taking first steps for studying neurodegenerative diseases (ND) as well as neuropsychiatric diseases. In addition, it is emerging as a very powerful tool to study quantitative trait loci associated with gene expression in complex diseases. This paper provides an overview on gene expression profiling of complex diseases, with emphasis on RNA-Seq, its advantages over conventional technologies for studying cancer and ND, and for linking nucleotide variations to gene expression changes, also discussing its limitations.
European Journal of Human Genetics (2013) **21,** 134–142; doi:10.1038/ejhg.2012.129; published online 27 June 2012

Keywords: RNA-Seq; noncoding RNA; epigenetics

INTRODUCTION

The sequencing of the human genome is a milestone in the scientific landscape and a springboard for genetic studies.[1] In the 'post-genomic era' considerable effort has been done to understand the genome content, whose knowledge was limited until 2001. Predictions about the number of protein-coding genes were far from correct, as well as the role of noncoding RNAs (ncRNAs) was very limited and confined to few processes, such as X-inactivation.[2] Introns, interspersed repeated sequences and transposable elements were considered junk DNA and evolutionary debris, and alternative splicing was an exception rather than the rule.

The availability of the entire euchromatic sequence (GRCh37/hg19) has allowed researchers to easily identify disease-causing mutations in more than 2850 genes responsible for a huge number of Mendelian disorders, and to detect statistically significant associations of about 1100 loci to more than 165 complex diseases and traits.[2] Nonetheless, studying human genetic disorders is a complex task – especially for multifactorial diseases – due to the small contribution of multiple genes to the resulting phenotype, and often to yet unknown gene–gene and gene–environment interactions. In addition, although for most of Mendelian disorders the causal variant has been described, for complex traits and common diseases, such as metabolic (type 2 diabetes, obesity), cardiovascular (atherosclerosis, hypertension) or neurological (Alzheimer, Parkinson) diseases as well as for cancer, these findings are far from complete.

About 88% of the genetic variants (single-nucleotide polymorphisms (SNPs)) currently associated to complex diseases and traits by genome-wide association studies (GWAS) rely within intronic or intergenic regions.[3] This evidence strongly suggests these nucleotide variations are likely to have causal effects by influencing gene expression rather than affecting protein function. Loci with such a property are referred to as expression quantitative trait loci (eQTL). A growing number of studies has unequivocally shown that such inherited polymorphisms account for gene expression variation in the population[4,5] and that global gene expression studies – not requiring *a priori* hypothesis – provide a large-scale way to investigate complex traits and the pathogenesis of common disorders.[6]

Thus, despite a deep genetic knowledge for many human genetic diseases, to date most of the studies do not provide relevant clues about the real contribution, or the functional role, of such DNA variations to disease onset. In this scenario, whole-transcriptome analysis is increasingly acquiring a pivotal role as it represents a powerful discovery tool for giving functional sense to the current genetic knowledge of many diseases.

The introduction of hybridization- (microarray) and sequencing-based (Serial Analysis of Gene Expression (SAGE), and Cap Analysis of Gene Expression (CAGE)) technologies has started to elucidate the involvement of multiple genes, or entire gene networks, in physiological and pathological conditions.[7] Until recently, microarrays have represented the more rapid, cost-effective and reliable technology able

[1]CNR, Institute of Genetics and Biophysics 'A. Buzzati-Traverso' (IGB), Naples, Italy
*Correspondence: Dr V Costa or Professor A Ciccodicola, CNR, Institute of Genetics and Biophysics 'A. Buzzati-Traverso' (IGB), Via P. Castellino 111, 80131 Naples, Italy. Tel: +39 081 6132 259; Fax: +39 081 6132 617; E-mails: valerio.costa@igb.cnr.it (VC) or alfredo.ciccodicola@igb.cnr.it (AC)
Received 6 December 2011; revised 10 May 2012; accepted 16 May 2012; published online 27 June 2012

to analyze, in a single experiment, the gene expression patterns of cells/tissues/organs/organisms. However, despite the rapidity and the affordable cost, its low computational complexity and the large availability of software for data analysis, some crucial tasks are not feasible with microarray platforms. *A priori* knowledge of sequences to interrogate is a limitation for *de novo* identification of splice isoforms or novel exons/genes. In addition, allele-specific expression, RNA editing and fusion transcripts represent some of the missing information, which may be crucial when comparing samples for disease-related studies. Moreover, hybridization-based platforms, which indirectly quantify gene expression suffer from background and cross-hybridization issues, and the limited dynamic range makes difficult to confidently detect and quantify low-abundance transcripts, as well as very high-abundance ones.[8,9]

Sequencing-based approaches, SAGE and CAGE, allow quantitative analysis of gene expression by counting the number of tags (corresponding to the number of mRNA transcripts) rather than measuring signal intensities as in hybridization-based approaches.[10] These technologies have been successfully employed to simultaneously study the expression levels of thousand genes, leading to promising results for Down syndrome (DS),[11] cardiovascular diseases[12] and diabetes.[13] However, the laborious concatenation and cloning of such tags, and the high costs of automated Sanger sequencing, have thus far limited their use.

Of note, undoubtedly, the recent development of a less expensive, faster and massive NGS technology and the wide use of short reads has taken its cue by the original SAGE and CAGE methods. Indeed, the widespread diffusion of NGS platforms – able to analyze hundreds of millions (up to billions) fragments of DNA or RNA – and of its applications, particularly RNA-Seq, has brought a significant qualitative and quantitative improvement to transcriptome analysis,[9] offering an unprecedented level of resolution and a unique tool to simultaneously investigate different layers of transcriptome complexity. It gives the possibility to detect even low-expressed genes, to accurately quantify their expression levels in each condition (pathology, drug treatment, different developmental stages), a more accurate estimate of sense/antisense transcription of genes, and also to analyze transcription starting sites (TSS) of genes. However, it does not allow – unlike CAGE – to get the exact positions of all TSS for a given gene, even though an innovative approach based on a combination of NGS and Oligo-capping (TSS-tag sequencing) has been recently developed to overcome this limitation.[14] Nonetheless, RNA-Seq provides more information than SAGE and CAGE in terms of splicing, post-transcriptional RNA editing and SNPs expression across the entire length of (virtually) all expressed transcripts in a cell. Indeed, it allows to analyze at a single-nucleotide resolution, the allele-specific expression and the post-transcriptional RNA editing, to examine known splice junctions- or to discover novel splicing events and to detect fusion transcripts, crucial especially in cancer research.[15] In addition, methodological refinements (ribodepletion, small- and microRNA isolation and purification) allow to select specific RNA species before RNA-Seq experiments, providing a more comprehensive view of the transcriptional landscape. However, along with the undoubted progress made by the introduction of NGS, not previously encountered issues have been also raised (reviewed in Costa *et al*[8]).

In the present review we describe how – and to what extent – human genetic research is gradually shifting toward the massive employment of RNA-Seq for a more comprehensive and detailed transcriptome analysis, also considering the current RNA-Seq limitations. In particular, here we discuss three classes of human disorders to date commonly investigated by this innovative NGS approach: (1) neurodegenerative disorders (ND) and neuropsychiatric disorders, (2) cancer and (3) complex traits/diseases (through the analysis of eQTL). Moreover, given the well-documented key role of epigenetic changes in the regulation of gene expression, we will also briefly discuss this interplay, describing some relevant findings and the current NGS approaches employed to study this complex interaction.

NEURODEGENERATIVE AND NEUROPSYCHIATRIC DISORDERS

ND result from the gradual and progressive loss of neural cells, and lead to nervous system dysfunction. The pathogenesis of ND is complex and remains mostly unknown.[16]

Because of the inaccessibility of human brain, a growing number of studies have been performed in animal models.[17,18]

In the 'pre-genomic era', only a small subset of causative genes for ND had been identified by linkage analyses followed by positional cloning. Further analyses of SNPs and copy number variations (CNVs) have revealed the existence of more than 200 distinct disease-causing mutations.[19,20]

More recently, GWAS have revealed the association of many common polymorphisms to ND sporadic cases, providing in about 3 years more reproducible and consistent findings than 2 decades of candidate-gene-driven research.[21] However, despite the step forward, the identification of potential causative loci associated to ND by GWAS explained only a little percentage of the cases, and the 'missing heritability' issue (ie, the contribution of epigenetic modifications on gene expression)[22] is still a limitation. As proof-of-concept, all previously cited approaches aimed to mutation discovery, do not provide any relevant clue about the contribution of such genetic alterations on ND onset. Therefore, transcriptome analysis has become central to functionally correlate the genetic variations to disease phenotypes.

In transcriptomic studies so far performed for ND and neuropsychiatrics disorders, the primary source has been the mRNA isolated from transgenic animal models and, more recently, patient-derived cell lines, although post-mortem brains have been frequently reported as the 'gold standard'.[23,24] However, although promising, the clear difficulties in obtaining brain tissue and the fragile nature of isolated RNA render transcriptome studies quite difficult.[25,26] Microarray analysis, widely used for ND and neuropsychiatric disorders, provided much information about the transcriptional profiles in pathological states,[27–29] although discordant results have been often reported. The lack of convergence could be attributed to microarray drawbacks (discussed in[8,9,15]), as well as to the variable quality/integrity of RNAs strictly influenced by pH,[25] which may dramatically alter the binding to the nucleotide probes, affecting the measure of gene expression levels. As ND patients have prolonged agonal state in brain tissue (strongly correlated with pH alterations), differences in RNA integrity may – at some extent – account for aberrant gene expression profiles.[30] This could be partially overcome by using a sequencing-based technology, less – if at all – sensitive to the fragmentation issue (but not to complete degradation). Indeed, SAGE technology has been successfully applied for studying DS,[11] Parkinson[31] and Alzheimer diseases.[32] Moreover, CAGE, and more recently nano-CAGE, have enabled to investigate brain-specific transcription,[33] whilst the deep-CAGE – combining standard CAGE method with NGS – has provided a detailed analysis of the hippocampus-specific core promoters.[34] However, although the excellent results of CAGE analyses, the central role that cell-specific alternative splicing has in the differentiation of neurons, and the emerging role of

ncRNAs – particularly of miRNAs, long intergenic and long ncRNA (lincRNAs and lncRNA, respectively) – in neurogenesis, strongly support the usage of RNA-Seq in brain transcriptome analysis.[35]

Some recent papers have pointed out the great advantages of using RNA-Seq to profile the transcriptome of brain tissue affected by ND. Nonetheless, to date only one published work has described the use of RNA-Seq on AD patients' brains, whereas another has employed similar approach to profile the transcriptome of human neurons derived from induced pluripotent stem cells proposing an ideal system for further studies on defective neurogenesis in patients.[6,35] The study of Twine *et al*[6] has provided, for the first time, an extensive transcriptome analysis of post-mortem frontal and temporal lobes of AD patients, highlighting a differential expression of known causative genes and also of previously unannotated expressed regions. It should be considered that given the high-level complexity of the human brain, achieved with the same number of genes as those of less evolved organisms, some of its complexity may probably be due to alternative splicing and alternative promoter usage. Such events have been described[6] in this study and possibly associated to the progression of neurodegeneration in patients.

Another crucial aspect to be reckoned with is the emerging driving role of ncRNAs, and particularly of miRNAs. Their pivotal role in regulating expression levels of genes involved in mental retardation and AD has been partially elucidated.[36,37] A recent work[36] has demonstrated in a mouse model of AD, the abnormal expression of miR-34a affecting the expression of bcl2 and contributing to AD pathogenesis. However, expression studies do not allow establishing of whether differential expression is the consequence or the cause of the disease and drawing any conclusion may be misleading. Despite this consideration, the miRNA-based deregulation of gene expression is one of the main etiologic factors underlying human diseases,[37] as recently highlighted by the revolutionary ceRNA theory.[38]

RNA-Seq have revealed that the expression of lincRNAs – another class of ncRNAs – dramatically changes during the transition from pluripotent stem cells to early differentiating neurons.[35] As these previously unexplored RNA molecules map to non-exonic regions (intergenic or intronic), these results indicate that RNA-Seq is very relevant also to assess the biological meaning of nucleotide variants falling outside annotated genes, associated with ND by GWAS. However, in order to confirm the role of these ncRNAs in the etiology of ND and neuropsychiatric disorders, functional assays are needed. Aging, due to a progressive accumulation of changes in an organism over time, is a strong risk factor in the onset of ND and represents another factor to consider in ND research. The incidence of AD increases from 0.6% at 65–69 y.o., to 2% between 75 and 80 y.o. and to 8.4% above 85 y.o.[39] As the cognitive decline is strictly associated with age in humans, it would be crucial to explore the association between gene variants, differential expression, disease-specific splicing and human longevity, as well as to understand the common mechanism underlying aging and neurodegeneration.

Important results in the identification of age-related changes in gene expression have been achieved using microarrays.[40] Particularly, Cao *et al*[41] showed that brains from fronto-temporal lobar degeneration and AD patients exhibit prematurely aged gene expression profiles. Nonetheless, much remains to be discovered about transcriptomic and epigenetic changes occurring during an organism lifetime. Therefore, in the next future it would be desirable to couple RNA- and ChIP-Seq experiments for studying epigenetics in ND and neuropsychiatric disorders.

Neurocognitive function has been also explored in DS by microarray on DS fetal and adult post-mortem human tissues,[42] or in animal models. However, most of the published studies revealed conflicting results highlighting the need of (almost) unbiased technologies and platforms for analyzing gene expression. In this context, our recently published work,[43] even though focused on the endothelial/immune aspects of DS, revealed the great potential of using RNA-Seq for human genetic diseases. Indeed, by using ribominus RNA-Seq we analyzed the global transcriptome of DS cells, also investigating ncRNAs.[43] We believe it would be desirable to apply this approach to profile DS brain tissue, in order to explore some pathogenic mechanisms underlying the defective neurocognitive behavior of DS patients.

RNA-SEQ IN CANCER

Cancer encompasses more than 100 distinct human malignancies[44] and is highly heterogeneous in its genetic and molecular aspects. Several classes of DNA alterations – nucleotide substitutions, indels, chromosomal rearrangements, such as CNVs – may give rise to human cancers, or DNA variations may just represent a consequence of the global cancer-induced genomic instability. Thus, establishing the relative contribution of genetic changes to cancer onset or progression (ie, 'driver' or 'passenger' mutations) is often difficult.[44] To further complicate the picture, some crucial alterations may not be detected by commonly used DNA analysis as they affect the gene expression levels and/or the DNA methylation status.

Cancer research has been focusing for more than 25 years on the identification of 'candidate' genes by using cytogenetic techniques, mutational screening and low-resolution genome-wide approaches, only providing limited results.[45] After the completion of the Human Genome Project,[1] cancer cells have been investigated by hybridization-based technologies, at both the genomic and the transcriptomic level. Array comparative genome hybridization – combining the genome-wide coverage of chromosome banding and the high resolution of fluorescent *in situ* hybridization (FISH) – has allowed to detect a large number of microscopic and submicroscopic chromosomal abnormalities with clear advantages over conventional analyses. In contrast, standard FISH requires *a priori* knowledge of the genomic sequence to interrogate, and thus it may fail to identify some duplications.[46]

SNP arrays have been widely used for genotyping cancer cells and to investigate the structural alterations frequently occurring in cancer genomes, even though qualitative and quantitative RNA analysis (of both coding and noncoding) has gradually acquired a central role in cancer research.[47] Indeed, gene expression profiling allows a deeper understanding of disease contribution providing a more dynamic view of the genome. Microarrays have significantly helped to profile tumors (at different stages and under different conditions), detecting clinically relevant markers associated with tumor subtypes.[44,48] Oncotype DX and MammaPrint, specific gene expression-based prognostic tests, have been developed to predict tumor behavior, prognosis and the response to drug treatment.[49,50]

In more recent years, the introduction of NGS platforms has largely and positively impacted cancer research. Particularly, RNA-Seq to investigate cancer transcriptomes may be the answer to a multitude of questions about carcinogenesis in humans. The possibility to simultaneously analyze by RNA-Seq several classes of alterations, frequently co-occurring in the genomes of cancer cells allows discovering previously unrecognized – or not yet fully characterized – pathogenic mechanisms.

Many RNA-Seq studies have suggested that detrimental fusion transcripts and alternative splicing may be involved in the

carcinogenesis of different tissues and organs such as breast,[51] prostate,[52,53] soft tissue,[54] melanocytes[55] and lymphoid tissue and organs (Table 1).[56–58] Most of them have discovered a considerable fraction of fusion transcripts – that is chimeric mRNAs that may alter cell's functionality – commonly produced by genomes rearrangement and critically involved in the pathogenesis of several types of malignancies. However, it should be noted that some of the newly identified rearrangements may not be the molecular cause of the aberrant phenotypes, and that using RNA-Seq solely allows detecting expressed fusion genes giving no information about other kind of structural rearrangements.

Sequencing of paired-end, rather than fragment libraries, has recently proved to be the most suitable approach to discover with high efficiency and sensitivity gene fusions and other chimeric transcripts, allowing the simultaneous analysis of gene expression, splicing and expressed nucleotide variations.[15] The use of paired-end libraries helps to reduce the bias in mapping reads to the reference genome, particularly to repeated regions and splice junctions, and is a 'gold standard' for the detection of breakpoints. Different computational methods and software for the detection of fusion transcripts in tumors have been developed.[68–69] To this purpose, a novel computational method, deFuse, has allowed to discover for the first time gene fusions in ovarian cancer specimen, also showing novel chimeric mRNAs in sarcoma.[66] Novel fusion transcripts have been also discovered, especially in breast cancer (Table 1).[51,61] RNA-Seq revealed that the occurrence of chimeric transcripts in melanoma is a frequent event, also highlighting novel genes and pathways previously not associated to its pathogenesis.[55]

Precisely defining the specificity and occurrence of some rearrangements may help clinicians to discern the molecular subtypes of the same cancer, such as in B-cell lymphomas and breast cancer. In a recent study on B-cell lymphomas, MHC class II transactivator (*CIITA*) has been identified as a novel partner of various fusions transcripts, suggesting a possible novel intriguing genetic mechanism underlying the onset of lymphoid cancers.[58] Moreover, the application of RNA-Seq to breast cancer samples has allowed to detect alternative splicing events associated with epithelial–mesenchymal transition (EMT), suggesting the classification of cancer cell lines into basal and luminal subtypes, based on their EMT-associated splicing pattern.[62]

Furthermore, the integration of multiple levels of analysis has allowed the identification of fusion genes associated with CNVs, suggesting that fusion events may contribute to the selective advantage provided by DNA amplifications and deletions, or may mediate the activation of a dormant gene. Moreover, RNA-Seq revealed a valuable resource to identify new ERBB2-mediated events and private fusions in some BRCA1-mutated transcriptomes, novel potential biomarkers for diagnosis and treatment.[60,61]

Another main advantage of NGS is the ability to detect ncRNA species, now emerging as potential contributors to different pathogenic mechanisms, also in human cancer. In this regard, a regulatory role of ncRNAs has been suggested by a recent analysis performed in Myelodysplastic syndromes (MDS), in which differences in miRNAs' expression were associated to early and later stages of the disease.[56] A very recent paper of Prensner *et al*[63] described previously unannotated prostate cancer-associated ncRNAs and one of them,

Table 1 RNA-Seq experiments in cancer

Cancer type	*Analysis type*	*Results*	*Ref.*
Hodgkin lymphoma	PE WT	Identification of gene fusions, among which fusions *CIITA*-involving	58
Non-Hodgkin lymphoma	PE poly-A$^+$	Detection of 109 genes with multiple somatic mutations, including those involved in histone modifications	59
MDS	FR small RNA	Discovery of novel miRNA differentially expressed in tumor	56
Breast cancer	FR poly-A$^+$	Alternative splicing and alterations in gene expression (ie, *LOX*, *ATP5L*, *GALNT3* and *MME*) have been identified in modulated ERBB2 overexpressing mammary cells	60
	PE poly-A$^+$	Identification of 3 known and 24 novel fusion transcripts (including *VAPB-IKZF3*)	51
	SE, PE poly-A$^+$	Discovery of gene fusions in breast cancer transcriptomes with BRCA1 mutations, including novel in-frame *WWC1-ADRBK2* fusion in HCC3153 cell line and ADNP-C20orf132 in a primary tumor	61
	FR poly-A$^+$	Investigation of EMT-associated alternative splicing events regulated by different classes of splicing factors (RBFOX, MBNL, CELF, hnRNP or ESRP)	62
Prostate cancer	SE poly-A$^+$	Detection of transcription-induced chimeras in prostate adenocarcinoma	52
	PE WT	Discovery and characterization of seven novel cancer-specific gene fusions (four involving non-ETS)	53
	PE poly-A$^+$	Identification of 121 unannotated prostate cancer-associated ncRNA transcripts, including the characterization of *PCAT-1*	63
	FR poly-A$^+$	25 Previously undescribed alternative splicing events involving known exons, and high-quality single-nucleotide discrepancies, have been detected in prostate cancer cell line LNCaP	64
Melanoma	PE poly-A$^+$	Identification of 11 novel gene fusions, 12 readthrough transcripts, somatic mutations and unannotated splice variants	55
	FR poly-A$^+$	Somatic CNVs affecting gene expression and new potential genes and pathways involved in tumorigenesis have been identified in seven human metastatic melanoma cell lines	65
Ovarian cancer	PE poly-A +	Discovery of the first gene fusions in ovarian cancer through a novel computational method	66
Sarcoma	PE poly-A +	Detection of novel gene fusions in sarcoma through a novel computational method	66
	FR ribodepletion	Evidence of a closer relationship between gene expression levels and protein expression in a human osteosarcoma cell line	54
Oral carcinoma	MP WT	Association of allelic imbalance with copy number mutations and with differential gene expression	67
Hepatocellular carcinoma	SE WT	Characterization of HBV-related HCC transcriptome, including identification of exon-level expression changes and novel splicing variants	57

Abbreviations: CNVs, copy number variations; EMT, epithelial–mesenchymal transition; FR, fragment library; HCC, hepatocellular carcinoma; MDS, Myelodysplastic syndrome; PE, paired-end; SE, single-end; WT, whole-transcriptome.

PCAT-1, has been described as a prostate-specific regulator of cell proliferation, targeted by the polycomb repressive complex 2.

Moreover, the advantage offered by RNA-Seq over hybridization-based approaches in studying role of allelic imbalance in allele-specific changes has been fruitfully employed to investigate cancer transcriptome.[67,70] Finally, the previously unexplored 'RNA editome' has been very recently proposed as contributor in cancer, even though only in a human glioblastoma cell line (U87MG).[71]

Reported evidences strongly suggest RNA-Seq will have an increasingly leading role in cancer research for both the diagnosis, prognosis and also to improve surgical and therapeutic interventions. However, it is clear that combining RNA-Seq with other NGS applications – as well as other platforms (ie, SNP and CGH arrays) – will help to detect somatic CNV affecting gene expression and potentially new candidate genes involved in tumorigenesis.[65]

eQTL, EPIGENETICS AND RNA-SEQ

The spectrum of nucleotide variations predisposing to, or responsible for, human genetic diseases ranges from very rare mutations (MAF, minor allele frequency $<<0.01$) – in Mendelian disorders – and rare variants (MAF <0.01) to very common SNPs (MAF 0.01–0.05) with weak effects on complex traits and common diseases. In the latter case, a small fraction of them falls in the coding regions and affecting the protein. GWAS have revealed that most of disease- and trait-associated SNPs (about 90%) are intronic or intergenic, suggesting these variants may affect gene expression.[3] The undeclared dispute among the 'classical geneticists' and the 'proponents of gene expression analysis'[72] has reached a compromise by systematically integrating such theories toward a genome-wide analysis of gene expression variations between healthy and affected individuals.

Gene expression is a heritable trait, amenable to genetic mapping, and its variation is one of the main driving mechanisms underlying complex diseases' susceptibility.[73] The association between nucleotide variants in a regulatory element of *LCT* gene and the lactase persistence phenotype in European population, identified about 10 years ago,[74] is one of the first – and perhaps better-known – demonstration of this hypothesis. Since then, GWAS have unequivocally shown that SNPs affect gene expression.[4,5,75] A common finding of eQTL studies is that *cis*-acting SNPs (ie, in close proximity to a gene) have a strong influence on gene expression and a greater replicability in different populations and by independent detection methods. On the opposite, *trans*-acting variations[76] with subtle effects on expression are less replicable and their causal association to traits/diseases is not trivial. However, it is clear that using a 'less-biased' experimental approach or technology is crucial for such analyses.

Recent studies have shown RNA-Seq may represent a 'gold standard' for high-resolution eQTL analysis, allowing a joint analysis of variation in gene expression levels, splicing and allele-specific expression across individuals.[77,78] Convincing evidence for allelic imbalance in *CD6* gene was shown by RNA-Seq at a multiple sclerosis-associated SNP (rs17824933), confirming previous GWAS, and linking a polymorphism to *CD6* gene expression changes.[79] Coupling RNA-Seq to other NGS applications (ChIP-Seq and exome sequencing), may reveal in the same sample different layers of complexity, showing the interplay among them (Figures 1 and 2). Gene expression may be affected at a transcriptional, co- and post-transcriptional level and the choice of combining RNA- and ChIP-Seq for the analysis of methylation and histone modifications will provide higher resolution giving a more comprehensive view of the transcriptome. Indeed, integrating data from such NGS applications may

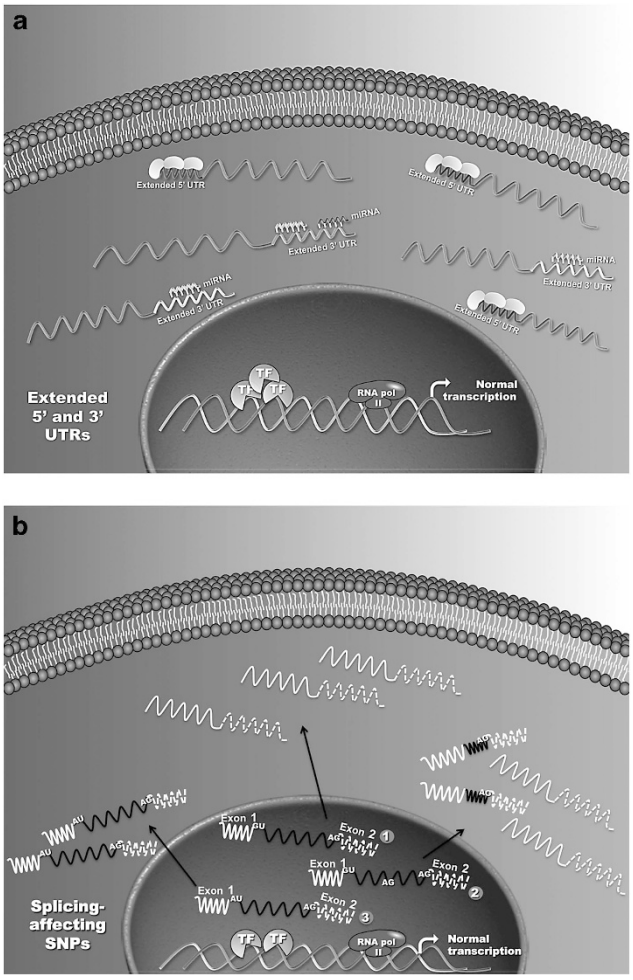

Figure 1 Nucleotide variations altering gene expression and splicing. (**a**) Graphical representation of nucleotide variations potentially affecting the binding of transcription factors (TFs) and/or RNA polymerase II, thus altering gene expression, detectable by integration of RNA-Seq and ChIP-Seq experiments. (**b**) SNP possibly occurring within the introns (black lines) affecting donor and acceptor splice sites (GU and AG) altering the splicing of the coding exons. In detail, in (1) a canonically spliced pre-mRNA following the GU-AG rule; (2) an example of nucleotide variation/s occurring within the introns and generating a novel acceptor 'cryptic' splice site. In this case, two different mRNAs are produced, depending on the different used acceptor splice site; (3) SNPs within the donor site (GU to AU change), leading to intron retention.

reveal, at the same time, SNPs that abolish or (just) partially affect the binding of RNA polymerase II and/or of transcription factors and complexes (both co-activators and -repressors), altering the initiation and progression (in terms of speed and stability) of transcription at specific loci (Figure 1a). Nucleotide variations may also be responsible of pre-mRNA splicing modifications, generating cell-, tissue- and developmental stage-specific transcripts, all potentially detectable by RNA-Seq (Figure 1b).

In addition, mRNA stability, antisense or miRNA-mediated degradation of a transcript are other relevant post-transcriptional processes possibly accounting for gene expression variability in humans.[80] RNA-Seq studies, and our recent work among them,[43] revealed that many genes annotated in currently available databases (ie, RefSeq, UCSC and Ensembl) have extended 3' UTRs, containing putative

miRNA binding sites, suggesting a previously undescribed miRNA-mediated regulation of such transcripts. This would also help to understand the impact of SNPs falling within these regions considered as 'non-genic' until now (Figure 2a). In addition, CNVs, insertions/deletions, short tandem repeats (di-, tri- and tetranucleotide expansion) and large genomic rearrangements can affect gene expression at some specific loci even up to several kb from the breakpoints.[81] Their impact on transcriptome is not limited to a quantitative regulation of the expression levels at some loci, but it also affects the timing of gene expression.[82]

Finally, despite our knowledge there are no conclusive studies directly linking epigenetics to complex traits and diseases, the involvement of an epigenetic framework as 'unifying principle' in the etiology of common diseases has been hypothesized.[83] Epigenetic contribution may explain the age-dependence of common diseases and the quantitative nature of complex traits, representing a possible direct link between environmental stimuli and gene expression (discussed in detail in Petronis *et al*[83]).

DNA methylation status of CpG islands is crucial in the epigenetic control of gene expression (Figure 2b) and is related to environmental factors, some of them we are continuously exposed to, such as the nutrients (reviewed in Costa *et al*[84]). Histone modifications and nucleosome positioning are not only responsible for what portions of the genome are expressed, but they also contribute to determine how they are (alternatively) spliced.[85]

It is evident that to better understand the interplay between epigenetic modifications and gene expression, as well as to assess their impact on human complex traits and common diseases, further combined studies (RNA-Seq and other NGS applications) are needed. To this purpose, a growing number of studies is currently showing that the integration of data derived from ChIP- (and its subapplications such as MeDIP-Seq or Methyl-Seq) and RNA-Seq analyses is the way forward.[86] Systematically profiling epigenome and transcriptome in multiple cell types and stages – in both physiological and pathological states – will improve the understanding of developmental processes and disease onset.[86,87]

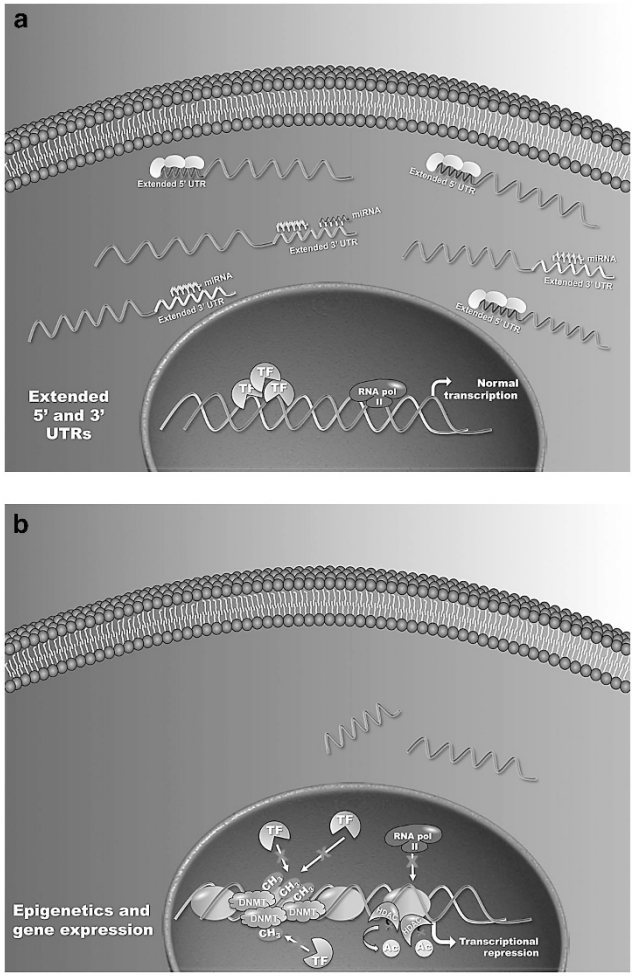

Figure 2 Extended UTRs and epigenetics in gene expression regulation. (**a**) Graphical representation of mRNAs with putative extended untranslated regions (UTRs). RNA-Seq may reveal new unannotated extended 5' UTRs, potentially involved in the binding of previously unexplored stabilizing protein complexes, whereas in extended 3' UTRs there may be new putative binding sites for miRNAs. (**b**) Schematic representation of some epigenetic mechanisms, regulating gene expression, possibly investigated by combining RNA-Seq to other NGS applications (ie, ChIP-Seq). TF, transcription factor; miRNA, microRNA; DNMT, DNA methyltransferase; HDAC, histone deacetylase; CH_3, methyl groups; Ac, acetyl groups.

RNA-SEQ LIMITATIONS AND ISSUES

After the 'early days enthusiasm' RNA-Seq has revealed its pitfalls, from sample preparation to data analysis, showing an obscuring variability.[88] Criticism about the experimental design and the validation issues in RNA-Seq experiments are now emerging in the literature, and different strategies to avoid – or at least to control – some unwanted effects have been proposed.[89]

RNA-Seq sample preparation includes multiple procedures (RNA extraction, fragmentation, reverse transcription and amplification), susceptible to experimental bias introducing nonlinear effects. One of the first sources of bias is fragmentation. It has the advantage of reducing the formation of secondary structures, particularly in ncRNAs, allowing higher sequence coverage across the transcript length, above all for long RNAs. However, the secondary structure itself, as well as the length of the transcript (as fragmentation is not random in short RNAs), affect the ability of RNA to be fragmented. The presence of 'susceptibility fragmentation sites' can dramatically alter the representation of that sequence within the library, leading to a 'pile-up' of reads, very common for short RNAs, such as snoRNAs (details in Sendler *et al*[90]). Moreover, locally, the GC percent may alter the probability of random fragmentation, leading in turn to a 'fragmentation model'.[88–90] This affects the 'counting efficiency' providing a severe bias in gene expression measurement, as certain RNA fragments are preferentially detected compared with others.[88] Other than affecting fragmentation, GC content has a relevant impact on cDNA amplification efficiency.[91] GC-rich RNA fragments undergo base pairing and often form double-strand or highly-paired secondary structures that affect – or impede – reverse transcription of such fragments, leading to a dramatic unbalance in PCR products.[90]

Furthermore, RNA-to-cDNA conversion (retrotranscription) before sequencing may introduce biases and artifacts interfering with the characterization and quantification of transcripts.[92] Furthermore, cDNA synthesis is not suitable to analyze short RNAs, degraded and/or small quantity RNA samples. After RT, a PCR amplification of cDNAs is needed for sequencing on most NGS platforms, which require clonally amplified templates. Insertion of confounding mutations in cDNA templates as well as overrepresentation or underrepresentation bias of fragments due to AT- and GC-rich sequences have been reported in this phase. Other effects, such as the choice of PCR enzyme or instruments have been also raised, and

globally the PCR amplification has been identified as the most discriminatory step with some relevant hidden factors still to be examined.[91] To overcome the previously cited limitations of RT and amplification, direct single molecule RNA sequencing approach has been developed,[92] in which PCR amplification is no more required. However, the higher error rate compared with other reversible terminator chemistries is a severe issue even for this technology (discussed in Metzker *et al*[93]).

Even though the accuracy in base sequencing is rapidly growing, systematic biases still exist. False-positive results, usually due to a misalignment of reads deriving from gene families and repetitive sequences may affect both the quantitative measure of gene expression and the analysis of allele-specific expression, as well as the detection of expressed SNPs in RNA samples. By analyzing the sequence of reads that overlap a given (heterozygous) SNP, it is possible to determine whether (and where) the transcription in a specific locus is allele-specific,[77] even though this is a challenging analysis. For instance, mapping the reads on a reference genome may not be the right way to study allele differences, due to biases in reference sequences. Although most of the analyses so far performed on human genomic data have used the reference genome for comparison, aligning the reads against a diploid sequence of the same analyzed individual is a more suitable solution to assess allele-specific behavior.[94]

RNA-Seq issues and concerns do not limit to experimental/technical procedures, but are also present in downstream computational analysis as well as in the informatics infrastructures, needed to support high-quality data generation and interpretation. NGS has shifted the bottleneck from the generation of large-scale experimental data to their management and computational analysis.[95] As discussed in Costa *et al*,[8] all NGS downstream analyses are difficult, if not impossible, without an appropriate information technology infrastructure. Indeed, the handling of terabytes of sequencing data – not huge in general for today's standards and not a serious problem for large sequencing centers and core facilities – is a novel problem to deal with for most of the research groups. In particular, permanent storage of such data, as well as keeping them available for quick online access and browsing, or sharing them among research groups worldwide, or submitting such data to public repositories (ie, Short Read Archive, European Genome-phenome Archive and Gene Expression Omnibus), still represent crucial limitations for RNA-Seq experiments.

CONCLUSIONS

In the last decade, human genetic research has made significant advances toward the understanding of many molecular aspects underlying human-inherited disorders, including the identification of 'disease-causing' mutations. However, particularly for complex diseases, the road ahead is still long, and 'the deep we investigate, the more it gets complicated'. Nonetheless, several evidences have unequivocally demonstrated that SNPs, identified by GWAS, and falling outside the coding regions of genes, may account for gene expression perturbation, pointing out to a crucial role of transcriptome studies for several complex diseases.[4–6,43]

Human genetics research has drawn particular benefit by the introduction of NGS platforms and, particularly of RNA-Seq, which has significantly improved the way of looking at cell transcriptome in physiological and pathological conditions.[8,9] It is reasonable to believe that massive analysis of transcriptomes, as well as large-scale NGS studies, will become a routine in the next future, within just a few years, and that not only cancer and ND research will benefit this technology. However, as previously discussed there are still challenges to face.[8]

Defining appropriate protocols for massive RNA sequencing and developing novel methodological procedures to isolate, select and target specific RNAs of interest, such as ncRNAs – emerging as new disease contributors – is a crucial task. Moreover, analyzing, validating, interpreting the large amount of data and finally translating them into potentially useful treatments for diseases may not be trivial. On the contrary, there is the risk of generating tons of 'under-used' information that in few months may become unused because new ones are massively produced. Indeed, to date, we are more capable at producing data rather than at analyzing them. In addition, there is urgent need for the development of novel computational strategies to deal with the high volumes of sequencing data created by RNA-Seq and other NGS applications, and integrating the results derived from different platforms and NGS applications will become an essential process in the next future. Indeed, most of the commonly used approaches usually handle each experiment independently. Instead, by integrating the vast amount of often complementary data, produced through the different NGS applications, we will surely gain more significant biological insights toward a complete understanding of the mechanisms driving gene expression changes in human genetic pathologies, rather than limiting to the interpretation of single data sets.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by COST Action BM1006: Next Generation Sequencing Data Analysis Network (SEQAHEAD), from European Cooperation in the field of Scientific and Technical Research, and 'Legge 5', Regione Campania.

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
2 Lander ES: Initial impact of the sequencing of the human genome. *Nature* 2011; **470**: 187–197.
3 Freedman ML, Monteiro AN, Gayther SA *et al*: Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 2011; **43**: 513–521.
4 Göring HH, Curran JE, Johnson MP *et al*: Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007; **39**: 1208–1216.
5 Morley M, Molony CM, Weber TM *et al*: Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004; **430**: 743–747.
6 Twine NA, Janitz K, Wilkins MR, Janitz M: Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One* 2011; **6**: e16266.
7 Horan MP: Application of serial analysis of gene expression to the study of human genetic disease. *Hum Genet* 2009; **126**: 605–614.
8 Costa V, Angelini C, De Feis I, Ciccodicola A: Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol* 2010; **2010**: 853916.
9 Shendure J: The beginning of the end for microarrays? *Nat Methods* 2008; **5**: 585–587.
10 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. *Science* 1995; **270**: 484–487.
11 Sommer CA, Pavarino-Bertelli EC, Goloni-Bertollo EM, Henrique-Silva F: Identification of dysregulated genes in lymphocytes from children with Down syndrome. *Genome* 2008; **51**: 19–29.
12 Gnatenko DV, Dunn JJ, McCorkle SR, Weissmann D, Perrotta PL, Bahou WF: Transcript profiling of human platelets using microarray and serial analysis of gene expression. *Blood* 2003; **101**: 2285–2293.
13 Misu H, Takamura T, Matsuzawa N *et al*: Genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in livers of patients with type 2 diabetes. *Diabetologia* 2007; **50**: 268–277.
14 Tsuchihara K, Suzuki Y, Wakaguri H *et al*: Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res* 2009; **37**: 2249–2263.

15 Ozsolak F, Milos PM: RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011; **12**: 87–98.
16 Bertram L, Tanzi RE: The genetic epidemiology of neurodegenerative disease. *J Clin Invest* 2005; **115**: 1449–1457.
17 Watson JB, Hatami A, David H *et al*: Alterations in corticostriatal synaptic plasticity in mice overexpressing human alpha-synuclein. *Neuroscience* 2009; **159**: 501–513.
18 Walsh DM, Selkoe DJ: Deciphering the molecular basis of memory failure in Alzheimer's disease. *Neuron* 2004; **44**: 181–193.
19 Bertram L, Lill CM, Tanzi RE: The genetics of Alzheimer disease: back to the future. *Neuron* 2010; **68**: 270–281.
20 Belin AC, Westerlund M: Parkinson's disease: a genetic perspective. *FEBS J* 2008; **275**: 1377–1383.
21 Klein C, Ziegler A: From GWAS to clinical utility in Parkinson's disease. *Lancet* 2011; **377**: 613–614.
22 Eichler EE, Flint J, Gibson G *et al*: Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**: 446–450.
23 Sutherland GT, Janitz M, Kril JJ: Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics? *J Neurochem* 2011; **116**: 937–946.
24 Soldner F, Hockemeyer D, Beard C *et al*: Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* 2009; **136**: 964–977.
25 Monoranu CM, Apfelbacher M, Grünblatt E *et al*: pH measurement as quality control on human post mortem brain tissue: a study of the BrainNet Europe consortium. *Neuropathol Appl Neurobiol* 2009; **35**: 329–337.
26 Atz M, Walsh D, Cartagena P *et al*: Methodological considerations for gene expression profiling of human brain. *J Neurosci Methods* 2007; **163**: 295–309.
27 Courtney E, Kornfeld S, Janitz K, Janitz M: Transcriptome profiling in neurodegenerative disease. *J Neurosci Methods* 2010; **193**: 189–202.
28 Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW: Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc Natl Acad Sci USA* 2004; **101**: 2173–2178.
29 Colangelo V, Schurr J, Ball MJ, Pelaez RP, Bazan NG, Lukiw WJ: Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and proinflammatory signalling. *J Neurosci Res* 2002; **70**: 462–473.
30 Papapetropoulos S, Shehadeh L, McCorquodale D: Optimizing human post-mortem brain tissue gene expression profiling in Parkinson's disease and other neurodegenerative disorders: from target "fishing" to translational breakthroughs. *J Neurosci Res* 2007; **85**: 3013–3024.
31 Noureddine MA, Li YJ, van der Walt JM *et al*: Genomic convergence to identify candidate genes for Parkinson disease: SAGE analysis of the substantia nigra. *Mov Disord* 2005; **20**: 1299–1309.
32 Xu PT, Li YJ, Qin XJ *et al*: A SAGE study of apolipoprotein E3/3, E3/4 and E4/4 allele-speciWc gene expression in hippocampus in Alzheimer disease. *Mol Cell Neurosci* 2007; **36**: 313–331.
33 Salimullah M, Sakai M, Plessy C, Carninci P: NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb Protoc* 2011; doi:10.1101/pdb.prot5559.
34 Valen E, Pascarella G, Chalk A *et al*: Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* 2009; **19**: 255–265.
35 Lin M, Pedrosa E, Shah A *et al*: RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PlosOne* 2011; **6**: e23356.
36 Wang X, Liu P, Zhu H *et al*: miR-34a, a microRNA up-regulated in a double transgenic mouse model of Alzheimer's disease, inhibits bcl2 translation. *Brain Res Bull* 2009; **80**: 268–273.
37 Weinberg MS, Wood MJ: Short non-coding RNA biology and neurodegenerative disorders: novel disease targets and therapeutics. *Hum Mol Genet* 2009; **18**: 27–39.
38 Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP: A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011; **146**: 353–358.
39 Bowen RL, Atwood CS: Living and dying for sex. A theory of aging based on the modulation of cell cycle signaling by reproductive hormones. *Gerontology* 2004; **50**: 265–290.
40 Lu T, Pan Y, Kao SY *et al*: Gene regulation and DNA damage in the aging human brain. *Nature* 2004; **429**: 883–891.
41 Cao K, Chen-Plotkin AS, Plotkin JB, Wang LS: Age-correlated gene expression in normal and neurodegenerative human brain tissues. *PLoS One* 2010; **5**: e13098.
42 Esposito G, Imitola J, Lu J *et al*: Genomic and functional profiling of human Down syndrome neural progenitors implicates S100B and aquaporin 4 in cell injury. *Hum Mol Genet* 2008; **17**: 440–457.
43 Costa V, Angelini C, D'Apice L *et al*: Massive-scale RNA-Seq analysis of non ribosomal transcriptome in human trisomy 21. *PLoS One* 2011; **6**: e18493.
44 Stratton MR, Campbell PJ, Futreal PA: The cancer genome. *Nature* 2009; **458**: 719–724.
45 Wood LD, Parsons DW, Jones S *et al*: The genomic landscapes of human breast and colorectal cancers. *Science* 2007; **318**: 1108–1113.
46 Pollack JR, Sorlie T, Perou CM *et al*: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional programs of human breast tumors. *Proc Natl Acad Sci USA* 2002; **99**: 12963–12968.
47 Mardis ER, Wilson RK: Cancer genome sequencing: a review. *Hum Mol Genet* 2009; **18**: R163–R168.
48 Berns A: Cancer: gene expression in diagnosis. *Nature* 2000; **403**: 491–492.
49 De Rienzo A, Dong L, Yeap BY *et al*: Fine-needle aspiration biopsies for gene expression ratio-based diagnostic and prognostic tests in malignant pleural mesothelioma. *Clin Cancer Res* 2011; **17**: 310–316.
50 Kim C, Paik S: Gene-expression-based prognostic assays for breast cancer. *Nat Rev Clin Oncol* 2010; **7**: 340–347.
51 Edgren H, Murumagi A, Kangaspeska S *et al*: Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* 2011; **12**: R6.
52 Nacu S, Yuan W, Kan Z *et al*: Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics* 2011; **4**: 11.
53 Pflueger D, Terry S, Sboner A *et al*: Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res* 2011; **21**: 56–67.
54 Klevebring D, Fagerberg L, Lundberg E, Emanuelsson O, Uhlén M, Lundeberg J: Analysis of transcript and protein overlap in a human osteosarcoma cell line. *BMC Genomics* 2010; **11**: 684.
55 Berger MF, Levin JZ, Vijayendran K *et al*: Integrative analysis of the melanoma transcriptome. *Genome Res* 2010; **20**: 413–427.
56 Beck D, Ayers S, Wen J *et al*: Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in Myelodysplastic Syndromes. *BMC Med Genomics* 2011; **4**: 19.
57 Huang Q, Lin B, Liu H *et al*: RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS One* 2011; **6**: e26168.
58 Steidl C, Shah SP, Woolcock BW *et al*: MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* 2011; **471**: 377–381.
59 Morin RD, Mendez-Lago M, Mungall AJ *et al*: Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 2011; **476**: 298–303.
60 Carraro DM, Ferreira EN, de Campos Molina G *et al*: Poly (A) transcriptome assessment of ERBB2-induced alterations in breast cell lines. *PLoS One* 2011; **6**: e21022.
61 Ha KC, Lalonde E, Li L *et al*: Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Med Genomics* 2011; **4**: 75.
62 Shapiro IM, Cheng AW, Flytzanis NC *et al*: An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* 2011; **7**: e1002218.
63 Prensner JR, Iyer MK, Balbin OA *et al*: Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 2011; **29**: 742–749.
64 Bainbridge MN, Warren RL, Hirst M *et al*: Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 2006; **7**: 246.
65 Valsesia A, Rimoldi D, Martinet D *et al*: Network-guided analysis of genes with altered somatic copy number and gene expression reveals pathways commonly perturbed in metastatic melanoma. *PLoS One* 2011; **6**: e18369.
66 McPherson A, Hormozdiari F, Zayed A *et al*: DeFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* 2011; **7**: e1001138.
67 Tuch BB, Laborde RR, Xu X *et al*: Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* 2010; **5**: e9317.
68 Sboner A, Habegger L, Pflueger D *et al*: FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* 2010; **11**: R104.
69 Kim D, Salzberg SL: TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* 2011; **12**: R72.
70 Parisi F, Ariyan S, Narayan D *et al*: Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. *BMC Genomics* 2011; **12**: 230.
71 Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X: Accurate Identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* 2011; **22**: 142–150.
72 Darvasi A: Genomics: gene expression meets genetics. *Nature* 2003; **422**: 269–270.
73 Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009; **10**: 184–194.
74 Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I: Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 2002; **30**: 233–237.
75 Emilsson V, Thorleifsson G, Zhang B *et al*: Genetics of gene expression and its effect on disease. *Nature* 2008; **452**: 423–428.
76 Majewski J, Pastinen T: The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* 2011; **27**: 72–79.
77 Pickrell JK, Marioni JC, Pai AA *et al*: Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010; **464**: 768–772.
78 Montgomery SB, Sammeth M, Gutierrez-Arcelus M *et al*: Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010; **464**: 773–777.
79 Heap GA, Yang JH, Downes K *et al*: Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet* 2010; **19**: 122–134.
80 Veyrieras JB, Kudaravalli S, Kim SY *et al*: High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008; **4**: e1000214.
81 Stranger BE, Forrest MS, Dunning M *et al*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007; **315**: 848–853.
82 Chaignat E, Yahya-Graison EA, Henrichsen CN *et al*: Copy number variation modifies expression time courses. *Genome Res* 2011; **21**: 106–113.
83 Petronis A: Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010; **465**: 721–727.

84 Costa V, Casamassimi A, Ciccodicola A: Nutritional genomics era: opportunities toward a genome-tailored nutritional regimen. *J Nutr Biochem* 2010; **21**: 457–467.
85 Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T: Epigenetics in alternative pre-mRNA splicing. *Cell* 2011; **144**: 16–26.
86 Costa V, Gallo MA, Letizia F, Aprile M, Casamassimi A, Ciccodicola A: PPARG: gene expression regulation and next-generation sequencing for unsolved issues. *PPAR Res* 2010 pii: 409168.
87 Hawkins RD, Hon GC, Ren B: Next-generation genomics: an integrative approach. *Nat Rev Genet* 2010; **11**: 476–486.
88 Hansen KD, Irizarry RA, Wu Z: Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 2012; **13**: 204–216.
89 Fang Z, Cui X: Design and validation issues in RNA-seq experiments. *Brief Bioinform* 2011; **12**: 280–287.
90 Sendler E, Johnson GD, Krawetz SA: Local and global factors affecting RNA sequencing analysis. *Anal Biochem* 2011; **419**: 317–322.
91 Aird D, Ross MG, Chen W-S *et al*: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011; **12**: R18.
92 Ozsolak F, Platt AR, Jones DR *et al*: Direct RNA sequencing. *Nature* 2009; **461**: 814–818.
93 Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 2010; **11**: 31–46.
94 Rozowsky J, Abyzov A, Wang J *et al*: AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011; **7**: 522.
95 McPherson JD: Next-generation gap. *Nat Methods* 2009; **6**: S2–S5.