The advent of next-generation sequencing technologies has revolutionized the study of genetic variation in the human genome. Whole-genome sequencing currently represents the most comprehensive strategy for variant detection genome-wide but is costly for large sample sizes, and variants detected in noncoding regions remain largely uninterpretable. By contrast, whole-exome sequencing has been widely applied in the identification of germline mutations underlying Mendelian disorders, somatic mutations in various cancers and de novo mutations in neurodevelopmental disorders. Since whole-exome sequencing focuses upon the entire set of exons in the genome (the exome), it requires additional exome-enrichment steps compared with whole-genome sequencing. Although the availability of multiple commercial exome-enrichment kits has made whole-exome sequencing technically feasible, it has also added to the overall cost. This has led to the emergence of transcriptome (or RNA) sequencing as a potential alternative approach to variant detection within protein coding regions, since the transcriptome of a given tissue represents a quasi-complete set of transcribed genes (mRNAs) and other noncoding RNAs. A further advantage of this approach is that it bypasses the need for exome enrichment. Here we discuss the relative merits and limitations of these approaches as they are applied in the context of variant detection within gene coding regions.

**Keywords:** exome • exome enrichment • next-generation sequencing • single-nucleotide variants • transcriptome

The advent of next-generation sequencing (NGS) technologies has revolutionized our approach to performing structural and functional genomics studies [1,2]. The detection and characterization of genetic variation (ranging from single-nucleotide variants [SNVs] and small insertions and deletions [indels] to larger structural rearrangements) in the human genome have been greatly facilitated by NGS technologies such as whole-genome sequencing (WGS) [3–5]. This has also driven the 1000 Genomes Project, which, upon completion, aims to provide a comprehensive map of human genetic variants. Findings from the pilot phases of this project have already provided new insights into the nature and extent of human genetic variation [6]. However, this undertaking is well beyond the technical and financial capabilities of individual laboratories. The high cost of WGS (in relation to sequencing, data storage and analysis), together with the challenges inherent in analyzing and interpreting variants detected in noncoding regions [7–9], have now made whole-exome sequencing (WES) a more popular approach in the context of variant detection [10–13]. WES has been applied to the detection of both germline and somatic variants [14,15]. As most of the disease-causing mutations in Mendelian disorders reside within gene coding regions, this has promoted the use of WES in unraveling new causal variants for these disorders [16,17]. This approach has also been widely employed in attempts to identify the somatic driver mutations within the exomes of various cancers [18]. WES has higher sensitivity and specificity for detecting SNVs than small indels [19,20]. In addition, WES also allows the detection of larger copy-number variations using depth of coverage from mapped short-sequence reads through the development of appropriate bioinformatics tools [21].

Although NGS technologies have been available since 2005, the isolation and enrichment of the entire set of all exons in the human genome (the exome) was not technically feasible until the development of commercial high-throughput exome-enrichment kits [22,23]. However, the cost of the exome-enrichment step, which constitutes a substantial proportion of the total cost of WES,
represents a ‘bottleneck’ that impedes the scale-up of WES to large sample sizes. More recently, the cost of sequencing has fallen rapidly owing to the increasing throughput of sequencing data (up to hundreds of gigabases) per instrument run by the latest sequencing platforms. As a result, multiple samples (up to tens of exomes) can now be multiplexed to avoid redundant sequencing while still achieving adequate sequencing depth. This is known as post-hybridization sample multiplexing or barcoding. By contrast, this barcoding protocol became available for the exome-enrichment steps only comparatively recently [24,25]; although it should further decrease the cost of exome enrichment and/or WES, its technical performance and effect on sequencing data from the sample barcoding in these prehybridization steps have not yet been tested experimentally by the end user.

To further optimize the cost-effectiveness of variant detection within coding regions, transcriptome or RNA sequencing (RNA-Seq) has been proposed as a potential substitute for WES [26,27]. From a theoretical standpoint, this approach represents a promising alternative since, by definition, the transcriptome comprises all transcripts for both coding RNAs (i.e., mRNAs) and noncoding RNAs in a given tissue. Hence, RNA-Seq would also be able to detect variants within the coding regions [28–30]. In addition, the use of RNA-Seq bypasses the need for exome-enrichment steps, thereby rendering this approach more cost effective than WES. It would also obviate the need for target-probe hybridization steps and the technical limitations during exome enrichment. For example, owing to the uneven capture efficiency across exons experienced when using available exome-enrichment kits, capture of all exons is incomplete. Moreover, some sequence reads map outside the targeted regions (‘off-target hybridization’), leading to the production of unusable sequence reads for downstream analysis [22,23,31,32]. However, the application of RNA-Seq in this context is not without its shortcomings and limitations. It is important to bear in mind that the transcriptome is tissue specific, so the set of genes transcribed varies between tissue types. As a result, sequencing a transcriptome from a specific tissue would not be expected to achieve a lower read base error rate as compared with NGS platforms; however, the opposite has actually been found to be the case. The read base error rate of the Helicos true single-molecule sequencing HeliScope™ is 4–5%, which is higher than any of the NGS technologies (<2%). In addition, with HeliScope, the dominant error type is indels [40]. All four of these...
sequencing platforms have been previously applied in WGS studies of the human genome. The limitations of TGS technologies (i.e., higher read base error rate and indel errors) are explicable in terms of the weak signal generated by single-molecule sequencing or due to incorporation of unlabeled nucleotides [39]. Despite these limitations, TGS technologies offer additional advantages for direct RNA-seq, such as avoiding the need to convert RNA into cDNA. Single-molecule sequencing also avoids the amplification biases inherent in measuring transcript levels [30,41].

Whole-exome sequencing

WES represents an approach to sequencing the entire set of exons in the human exome (comprising ~200,000 exons). Since this approach focuses specifically on the coding regions, exome-enrichment steps are needed before the genomic DNA can be subjected to massively parallel sequencing. The development of commercial whole human exome-enrichment kits by Agilent, Nimblegen and Illumina has been largely responsible for the popularity of this approach [22]. The relative performance of these different WES platforms has recently been compared [23]. WES has gained favor largely because gene coding regions harbor >85% of mutations in monogenic disease states [42] and it has been widely applied in the identification of the underlying germline mutations in numerous Mendelian disorders of previously unknown genetic etiology. In addition to its use as a discovery tool, WES is now increasingly being employed in diagnostic applications [43]. Furthermore, this approach has also been widely adopted in the study of somatic driver mutations within gene coding regions in various cancers [15,44,45]. When trios (parents–offspring) have been available, WES has been successfully used to identify de novo mutations in a number of different neurodevelopmental disorders [46–48]. Despite it being widely considered as a transient technology, WES studies have already generated significant new findings in the context of human genetic disease.

As noted earlier, exome enrichment is a prerequisite for WES. During enrichment, the genomic regions of interest (the exome) are captured through hybrid selection of DNA fragments using oligonucleotide probes, whereas the unwanted DNA sequences (the noncoding regions) are removed prior to sequencing. Although protein coding regions are almost invariably the main focus, exome-enrichment kits are also designed to capture sequences outside the exome. For example, the Illumina™ TruSeq™ exome-enrichment kits was designed to target a region of 62 Mb [101], more than double the size of the human exome (~30 Mb). In addition to the comprehensive coverage of the major exon/gene databases such as consensus coding sequence (CCDS) and RefSeq, this enrichment kit also provides broad coverage of noncoding DNA in exon-flanking regions (promoters and untranslated regions). Furthermore, approximately 78% of predicted miRNAs are also captured.

Although whole-exome-enrichment kits were designed to capture the exons listed in the major databases (such as CCDS and RefSeq), it is by no means complete. For example, the NimbleGen version 2 exome-enrichment kits (Roche) is predicted to provide 99.2% coverage of CCDS, but only 49.6% of RefSeq would be covered [31]. Furthermore, it is noteworthy that the human gene complement is far from being fully characterized, as demonstrated recently by Mercer et al. [49]. This study used a capture array covering 2265 contiguous regions that collectively comprised a total size of approximately 0.77 Mb and was subjected to deep RNA-seq. By focusing on regions containing well-annotated protein-coding genes, Mercer et al. identified an additional 204 unannotated isoforms of 55 protein-coding loci, representing a 2.8-fold increase over the current catalog of isoforms for these loci [49]. This suggests that considerable functional genomic complexity remains to be resolved even for quite well-characterized loci. Therefore, given that the structures of many human genes are still inadequately characterized, it is apparent that WES (unlike RNA-seq) is intrinsically incomplete and inevitably biased towards the currently known (and still limited number of) exons of protein-coding genes.

Although multiple sequence-enrichment methods are available [24], the commercial enrichment kits come in two formats that is array-based and in-solution hybrid capture. The difference between on-solid and in-solution capture methods is that the oligonucleotide probes are either tethered on microarrays or are suspended in solution (oligonucleotide probes attached to beads), respectively. The coupling of these enrichment methods with NGS technologies has made WES technically more feasible and cost effective. However, a major limitation of the exome-enrichment kits is that some exons are not captured, resulting in some variants within these regions going undetected and hence being refractory to analysis. Thus, WES may have to be supplemented with conventional PCR-based Sanger sequencing methods in order to capture these ‘missed’ exons.

Another critical limitation of the enrichment kits is the uneven capture of exons as a consequence either of the technical limitations of target-probe hybridization or the variable GC content of the regions. This, coupled with the uneven sequencing characteristic of the NGS technologies, has resulted in an inadequate sequencing depth for some of the regions. GC-rich sequence stretches can be difficult to capture and, in the worst case scenario, these GC-rich regions are not captured at all. For example, two exons without any sequencing coverage were found to contain very high GC content (76.1 and 63.6%, respectively) compared with an average GC content of 37.6% for the 50 best-covered exons [50]. Those variants called in regions characterized by poor sequencing coverage usually receive a poor accuracy score and hence are usually filtered out. As a result, a higher overall/average sequencing depth is needed to ensure that the poorly covered regions achieve the minimum sequencing depth for accurate variant detection. Despite this unevenness, an average sequencing depth of 30–50× is usually deemed sufficient for the detection of most germline SNVs. WES at this depth has high sensitivity and specificity to detect SNVs (approximately >90%), but the specificity for small indels is much lower [14,19,20,51].

Transcriptome sequencing

The transcriptome is the collection of all protein coding and non-coding transcripts (RNAs) in a given tissue. However, the coding
Table 1. Comparison of whole-exome sequencing and RNA sequencing in the context of detecting coding region variants.

<table>
<thead>
<tr>
<th>Aspects of comparison</th>
<th>Whole-exome sequencing</th>
<th>RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue specificity of exome versus transcriptome</td>
<td>Exome shared across different tissues/cell types (same set of genomic DNA and germline DNA variants, but different somatic mutational profiles)</td>
<td>Transcriptome varies between different tissues/cell types (different sets of transcribed genes)</td>
</tr>
<tr>
<td>Time-course difference</td>
<td>Same set of genomic DNA and germline DNA variants throughout life, but somatic mutations accumulate with mitotic cell divisions in a tissue-/cell type-specific manner</td>
<td>Transcriptome is ‘dynamic’ in that it changes with time in a given tissue/cell type in response to both internal and external stimuli</td>
</tr>
<tr>
<td>Application</td>
<td>Variant detection in coding regions</td>
<td>Expression analysis of coding and noncoding transcripts</td>
</tr>
<tr>
<td>Application in variant detection to identify:</td>
<td></td>
<td>Discovery of new transcripts</td>
</tr>
<tr>
<td>– Germline variants for Mendelian disorders</td>
<td>Yes</td>
<td>Studying alternative splicing patterns</td>
</tr>
<tr>
<td>– Somatic variants in cancers</td>
<td>Yes</td>
<td>Detection of transcript fusions</td>
</tr>
<tr>
<td>– De novo variants</td>
<td>Yes</td>
<td>Analysis of allele-specific expression</td>
</tr>
<tr>
<td>Sources of DNA/RNA sample:</td>
<td>DNA from any tissues (commonly DNA is extracted from a peripheral blood sample)</td>
<td>Variant detection in coding regions of transcribed genes in a given tissue</td>
</tr>
<tr>
<td>– Germline variants</td>
<td>DNA from disease tissue or tissue of interest</td>
<td>Has not been applied</td>
</tr>
<tr>
<td>– Somatic variants</td>
<td>mRNA from tissue of interest to detect variants in transcribed genes</td>
<td>Yes</td>
</tr>
<tr>
<td>Need for exome enrichment</td>
<td>Yes</td>
<td>Has not been applied</td>
</tr>
<tr>
<td>Representativeness of the entire coding region</td>
<td>Capture and sequence almost all of the ~200,000 exons in the human genome</td>
<td>Only a subset of exons from the transcribed genes (pooling of transcriptome from multiple tissues to enhance the detection of germline variants in exons is theoretically sound, but could be technically laborious)</td>
</tr>
<tr>
<td>Variant detection in protein-coding regions</td>
<td>Both transcribed and nontranscribed genes</td>
<td>Only transcribed genes in a given tissue (RNA-seq would be incapable of capturing all variants within the coding regions)</td>
</tr>
<tr>
<td>Variant detection beyond protein-coding regions</td>
<td>Whole exome-enrichment kits have also been designed to capture mRNA regions</td>
<td>Variant detection in noncoding regions is also feasible through whole-transcriptome sequencing rather than mRNA sequencing</td>
</tr>
<tr>
<td>Incomplete capture of the entire coding regions</td>
<td>Some exons failed to be captured by the exome-enrichment kits and hence variants in the missing regions cannot be detected</td>
<td>Incomplete in terms of the variants in nontranscribed genes (in a given tissue) that cannot be detected</td>
</tr>
<tr>
<td>Uneven capture and sequencing</td>
<td>Uneven capture of exons as a consequence either of technical limitations or the variable GC content of chromosomal regions</td>
<td>Different expression levels of different genes/transcripts lead to a natural ‘unevenness’ in the transcriptome</td>
</tr>
<tr>
<td>Other issues associated with variant detection</td>
<td>Inherent limitations in target–probe hybridization during exome enrichment</td>
<td>Expression imbalance of two different strands is likely to generate false-positive results in variant detection. The possibility of errors in the reverse transcription of cDNA or the existence of RNA editing must also be considered. Mutations that cause rapid mRNA degradation of the transcripts containing them may also be missed by RNA-seq</td>
</tr>
<tr>
<td>Diagnostic application</td>
<td>Widely tested to detect germline variants underlying Mendelian disorders</td>
<td>Not widely tested</td>
</tr>
</tbody>
</table>

RNA-seq: RNA sequencing.
Exome versus transcriptome sequencing in identifying coding region variants

component can be extracted from the transcriptome sample and constructed into a sequencing library before being subjected to massively parallel sequencing. As with WES, the large number of sequence reads produced are then mapped to a reference genome. RNA-seq data have previously been used in several major applications, such as the expression analysis of coding and noncoding transcripts, the discovery of new transcripts, the study of alternative splicing patterns, the detection of transcript fusions and the analysis of allele-specific expression [28–30].

In comparison to the genome (or genomic DNA), the transcriptome is both tissue- and cell-type specific. It is also dynamic in that it changes with time (within the same tissue and even within the same cell type) in response to both internal and external stimuli. Thus, the transcriptome derived from any one tissue type will not represent the entire exome (i.e., all cells may have essentially the same genome/exome, but not all genes are expressed in a specific tissue/cell type). The focus on the exome (or mRNAs) is achieved naturally through transcription and technically through mRNA extraction and library preparation methods, thereby bypassing the need for an exome-enrichment step. This also leads to the incomplete capture of the exome: variants in nontranscribed genes cannot be detected. This has important implications in the context of identifying variants in all coding regions. It has been shown experimentally that only approximately 40% of all coding SNVs can be identified by RNA-seq using peripheral blood mononuclear cells as the RNA source. However, when this analysis was focused exclusively on peripheral blood mononuclear cell-expressed genes, approximately 81% of coding SNVs were identified [27]. This suggests that RNA-seq is only a feasible alternative for identifying exonic variants in tissue-specific transcripted genes. Table 1 summarizes the comparison of WES and RNA-seq in detecting coding region variants.

A further complication of RNA-seq is that different genes have different expression levels at different times. This also leads to a natural unevenness of the transcriptome (i.e., high- vs low-abundance mRNAs), irrespective of the tissue under study. More specifically, by performing RNA-seq on an acute myeloid leukemia (AML) sample alongside the corresponding remission sample, Greif et al. found that the read depth per gene ranged from zero to over 1000 [52]. Put another way, a total of 10,152 genes had an average read depth of at least sevenfold while 6989 genes had an average read depth of 20-fold or greater in both samples [52]. This suggests that some transcripts (expressed to some levels in a tissue) will have inadequate coverage for variants to be called accurately, dependent on the sequencing depth.

From a theoretical perspective, the unevenness of the expression levels could be rectified by increasing the overall sequencing depth to ensure that the low-abundance transcripts are adequately sequenced to allow efficient variant detection. However, this will lead to highly redundant sequencing of the most abundant transcripts. This is particularly problematic in those transcriptome samples harboring the greatest variability in transcript expression levels. Since this ‘further sequencing’ approach is not commonly pursued because it is not cost effective, the accuracy of variant detection is likely to be compromised in the context of low-abundance transcripts. Increasing the sequencing depth is also not without its adverse effects. A very high sequencing depth will increase the sensitivity of variant detection in low abundance transcripts, but it also compromises the specificity [27]. This is because, at low levels of coverage, reads that could produce false-positive calls are not sufficiently abundant to pass through the quality control filters. However, as the sequencing depth increases and more reads are added to the dataset, the number of incorrect alignments and hence sequencing errors increases, resulting in more false-positive calls passing through the quality control filters. In either scenario, the critical challenges in variant detection using RNA-seq due to uneven transcripts levels are apparent. This natural variability in transcript expression levels could be even larger than the unevenness resulting from the technical limitations in exome enrichment of genomic DNA. Although the challenge of variant detection in low-abundance transcripts can be alleviated by RNA CaptureSeq enrichment for low-level transcripts using cDNA-tiling arrays prior to high-throughput sequencing [49,53], this further adds to the cost of the approach, thereby off setting the benefits of reduced cost potentially offered by this approach. However, it is also noteworthy that RNA-seq (as opposed to WES) is not biased towards currently known exons and that CaptureSeq permits a very deep RNA-seq coverage of genomic regions of interest and has a high likelihood of revealing novel gene variants.

The definition of low- versus high-abundance transcripts is somewhat arbitrary, and the existence of alternative splicing further complicates it. In any discussion of the detection of variants in low- versus high-abundance transcripts, alternative splicing must be taken into consideration. Let us take a simple hypothetical example with five alternative transcripts for a given gene; transcript ‘A’ is highly expressed; for example, fourfold higher than each of the remaining four transcripts in a given tissue. However, transcript ‘A’ does not contain exon ‘I’, although this particular exon is expressed in all four other transcripts. Hence, variants in exon ‘I’ would not be detected in the highly abundant transcript ‘A’, but would be collectively in the other alternative transcripts expressed at a lower level. It should therefore be emphasized that the ability to detect variants in a particular exon depends upon the ‘aggregate abundance’ of the alternative transcripts covering the exon rather than their ‘individual abundance’ per se.

Furthermore, the expression imbalance of two different strands is also likely to generate false-positive results in variant detection using RNA-seq [54]. For this reason, some SNVs in the genomic DNA may be missed by RNA-seq as a consequence of allele expression imbalances [55], namely when an individual is heterozygous for a given SNP (in genomic DNA) but the reference allele is much more highly expressed than the mutant allele. Other complications associated with RNA-seq include: a low percentage of uniquely aligned sequence reads; incorrect alignment at the ends of a read due to splicing; and the possibility of errors in the reverse transcription of cDNA or the existence of RNA editing [27]. Finally, beyond these technical limitations, mutations that cause rapid mRNA degradation of the transcripts containing them may also be missed by RNA-seq. These limitations
are noteworthy as some of them are specific to the RNA-seq approach.

Despite these limitations, several studies have demonstrated the successful application of RNA-seq to the discovery of novel driver mutations in cancer [52,56–60]. For example, Shah et al. identified a recurrent missense point mutation (402C/G or C134W) in FOXL2 by analyzing four ovarian adult-type granulosa-cell tumors using whole-transcriptome paired-end RNA-seq [58]. For comparison, the study also performed RNA-seq for 11 nongranulosa-cell ovarian tumors, and transcriptome variants were identified that were absent in these 11 tumors that were classified as granulosa-cell tumor-specific variants and subjected to further analysis. The specific FOXL2 mutation was further validated in the cDNA and genomic DNA of all four index samples by two additional independent methods. Furthermore, the mutation was found to be somatic in origin in two patients from whom normal constitutional tissue was available [58]. Similarly, five nonsynonymous mutations specific to the tumor sample were identified by performing RNA-seq of an AML sample (bone marrow aspirate) with the corresponding remission sample (peripheral blood). These mutations included a nonsense mutation affecting the RUNXI gene (a known mutational target in AML) and a missense mutation in the TLE4 gene (which encodes a RUNXI-interacting protein) [52]. Taken together, these studies have demonstrated that RNA-seq represents a promising tool for detecting point mutations or SNVs within coding regions of transcribed genes that could turn out to play a key role in tumorigenesis.

**Germline versus somatic mutations**

The suitability of WES and RNA-seq as mutation screening techniques is also dependent upon whether the variants in question are germline or somatic. Germline variants are heritable and hence are shared across different tissues and cell types. For the detection of germline variants, genomic DNA is required for WES and can in principle be derived from any tissue (except in the case of gonosomal mosaics). By contrast, our ability to detect germline variants from a transcriptome sample of a given tissue is confined to the transcribed genes of that tissue. However, the transcriptomes from multiple different tissues could in principle (and in practice) be combined to increase the completeness of germline variant detection in the entire coding region. Although this approach is theoretically and technically sound, it is experimentally laborious to prepare and combine mRNA samples from multiple tissues, and consequently this may also decrease the cost-effectiveness of RNA-seq. The detection of somatic variants is a somewhat different procedure as it requires comparison with ‘constitutional DNA or RNA’ from a different tissue in order to exclude germline events. In addition, detection of somatic variants requires the specific tissue of the disease of interest; for example, tissue from the primary tumor. The need to obtain the appropriate tissue for the disease of interest has essentially limited the studies of somatic mutations to cancer [52,56–59]. It may be quite challenging to obtain the appropriate tissues to study other diseases, such as schizophrenia (brain tissue), or to establish the identity of the appropriate tissue to study specific systemic diseases, such as diabetes or systemic lupus erythematosus.

WES has been widely used to detect somatic variants in primary cancer tissue in comparison to constitutional DNA derived from normal blood or skin samples from the same individual. Although RNA-seq can also be used to detect somatic mutations, finding a matched normal sample for comparison can be challenging. Normal tissue is unlikely to express exactly the same genes as the tumor sample because different tissues have different sets of transcribed genes in addition to varying in terms of transcript abundance [35,36]. However, gene expression patterns/levels have been found to be comparable between an AML sample (a bone marrow aspirate with more than 90% blasts) and a remission sample (peripheral blood with a normal white blood cell count) [52].

As with germline variants, somatic variants are likely to be detected by RNA-seq in highly transcribed genes, but combining transcriptomes from multiple tissues (to increase the abundance of low-transcribed genes) would be inappropriate in the context of somatic variants. However, it is unlikely that somatic variants (e.g., even though they may be protein truncating) detected in nontranscribed genes will be functionally important in the specific tissue under study. This therefore justifies using the RNA-seq approach for detecting somatic variants in genes that are transcribed only in the tumor tissue under study. Although somewhat speculative, it may be that somatic variants detected in genes that are not transcribed in a given tumor tissue could also be functionally important and might mediate their effects through mechanisms other than protein disruption.

In terms of detecting somatic variants, both WES and RNA-seq approaches are limited by the impurity of the primary tumor tissue (i.e., it inevitably contains a mixture of cancer and noncancer cells). Genetic heterogeneity of the cancer cells further complicates the situation, since different subclones within the same tumor tissue can harbor different mutational profiles. These issues make the detection of somatic variants even more challenging and require greater sequencing depth to be resolved [35,36].

**Diagnostic applications**

In addition to new discoveries, the arrival of NGS technologies has also created new opportunities in molecular diagnostics. WES has been shown to be a promising tool in a diagnostic setting for rare Mendelian disorders. In a pioneering study by Choi et al. the genetic diagnosis of congenital chloride-losing diarrhea in a patient was confirmed through WES by revealing a homozygous missense variant in SLC26A3 (a gene known to underlie the disease) [43]. The patient was initially diagnosed as having Barter syndrome based on superficial phenotyping [43]. Its diagnostic utility is also becoming more evident in cases with broad and previously unsuspected phenotypic heterogeneity. This has been well illustrated by the identification of a homozygous PEX1 mutation in a patient with a clinical diagnosis of Leber congenital amaurosis, which is identical to a mutation known to cause Zellweger syndrome [61].

In addition, the clinical utility of WES has also been demonstrated in the case of one patient affected by two different genetic disorders. Two different mutations in SLC45A2 and G6PC3, respectively, were identified in a single patient with an indeterminate clinical phenotype. These mutations were sufficient to
account for the two different clinical phenotypes manifested by this patient; oculocutaneous albinism Type 4 and neutropenia 
[62]. Last, WES is also a powerful tool for disorders with genetic heterogeneity, such as Charcot–Marie–Tooth disease, an inherited peripheral neuropathy characterized by extensive locus heterogeneity (mutations in more than 35 genes have been identified to date). Indeed, a WES study of two affected members in a family with Charcot–Marie–Tooth disease identified a nonsynonymous mutation in GJB1, a known Charcot–Marie–Tooth disease gene, thereby confirming the molecular diagnosis [63].

Apart from its application in diagnostics, WES also has several advantages over the conventional targeted sequencing of candidate genes by PCR-based Sanger sequencing methods, which prioritize genes for sequencing on a ‘one-by-one’ basis. Despite the potential to use WES as a diagnostic tool, the technical challenges and ethical issues involved in adopting this approach in clinical laboratories must also be appreciated. A major challenge will be to analyze the large amount of sequencing data, since WES typically generates >10,000 genetic variants per genome. Thus, a robust variant-filtering pipeline must be applied to identify the disease-causing variants. In addition, the sensitivity and specificity of WES to detect SNVs and small indels needs to be further improved to attain clinical standards. The common practice of validation of the results from WES by Sanger sequencing unnecessarily increases the cost of a diagnostic test. Furthermore, the incomplete capture of some exons and uneven sequencing depth could potentially lead to a negative result. It is therefore important to generate a report (detailing the quality of a WES run; e.g., what was not captured and what was sequenced unreliably owing to inadequate sequencing depth) for diagnostic applications. On the other hand, there are a number of unresolved and quite complex ethical issues including the disclosure of findings that might be considered incidental or unrelated to the original purpose of the diagnostic test and whether the patients have the right to demand full access to the results generated from their WES diagnostic test. Since WES is a powerful information generation tool, this also raises a concern as to whether clinicians and medical geneticists have a responsibility to sift through the list of variants to identify known pathological mutations for other diseases and what level of scrutiny should be exercised [64–67].

The discussion of the potential for RNA-seq to be used as a diagnostic tool in other applications, such as the measurement of transcript expression levels and the detection of fusion gene transcripts, is beyond the scope of this article. However, in terms of applying RNA-seq as a diagnostic tool for variant detection, it has not been widely tested empirically. It should be noted that variant detection is a minor application of RNA-seq. However, in most cases, RNA-seq is being used to examine whether the variant-harboring genes are expressed. The limited interest in applying RNA-seq in this context is probably attributable to the various technical limitations and challenges, as discussed earlier. In the context of cancer, RNA-seq may be more suitable for detecting somatic mutations in transcribed genes and in the prediction of therapeutic response. As with WES, RNA-seq is a genome-wide approach and a powerful information generation tool. Hence it would be very applicable to cancers characterized by genetic heterogeneity, rendering the ‘one-by-one’ approach inefficient. In this scenario, RNA-seq would serve multiple roles to detect genetic aberrations in a single experiment. By contrast, RNA-seq might not be an inappropriate choice as a diagnostic tool if most of the cases could be accounted for by a single genetic alteration.

Expert commentary

In our view, WES of genomic DNA is a more powerful tool to detect germline variants than RNA-seq. The WES approach is, however, reliant upon commercial exome-enrichment kits to capture the entire set of exons. As such, incomplete capture of exons in some regions due to technical limitations represents a key challenge. Nevertheless, this limitation can be remedied by conventional PCR-based Sanger sequencing methods as long as the number (or the total genomic size) of these missing exons is small, since traditional methods are laborious and not readily scaled up. By contrast, high-throughput multiplex PCR methods, such as RainDance™ and Fluidigm® technologies, could be applied if a considerable number of exons are lacking (up to hundreds of exons) [68]. The enriched genomic DNA can then be sequenced using the medium-throughput NGS machines, such as the Roche 454 Genome Sequencer Junior Sequencing System [69], the Life Technologies Ion Torrent Personal Genome Machine Sequencer [70] and the Illumina MiSeq Personal Sequencing System [102]. As the total size of genomic DNA from the ‘missing’ regions is not large (ranging from tens to hundreds of kilobases), the sequencing capabilities of these medium-throughput NGS machines, such as >35 Mb (454 Junior) to >1 Gb (Ion Torrent and MiSeq) will adequately suit this application. In addition, amplicon-sequencing protocols have also been commercially developed for these medium-throughput NGS machines. Furthermore, sample multiplexing can further optimize cost-effectiveness.

By contrast, the transcriptome from a specific tissue/cell only represents a subset of the exome. As a result, only the variants in the expressed genes or transcripts in that tissue can be detected by RNA-seq. However, where detection of germline variants is required, this can be remedied, at least in principle, by combining the transcriptome samples from multiple tissues in order to extend variant detection to the entire exome. This is feasible because germline variants are not tissue specific. However, the considerable variability in transcript expression levels presents a critical challenge to the detection of variants using RNA-seq. This will lead to either insufficient coverage of low-abundance transcripts to enable accurate variant detection or redundant sequencing of the high-abundance transcripts to ensure that the sequencing/coverage depth is sufficient for low-abundance transcripts. Both outcomes are undesirable.

In the context of somatic mutations, both WES and RNA-seq may be applied but both approaches have their pros and cons. Most of the studies that have detected somatic mutations in the cancer genome (utilizing genomic DNA extracted from cancer tissues or cell lines) have adopted either WES or WGS approaches. WES of genomic DNA detects somatic variants in all of the coding regions (both transcribed and nontranscribed genes) in the cancer tissue. It is arguable that deleterious somatic mutations,
such as nonsense/protein-truncating mutations in expressed genes, are more likely to be functionally important. However, somatic mutations located in nontranscribed genes need not necessarily be functionally inert if they coincide with regulatory elements or produce biological alterations other than protein disruptions. This does not therefore render WES redundant (in terms of generating data in nontranscribed genes) to this application. By contrast, RNA-seq only detects somatic mutations in transcribed genes in the specific tissue that is relevant to the disease of interest. Furthermore, RNA-seq possesses several distinct advantages that cannot be substituted by WES. In addition to mutation detection, RNA-seq also permits other analyses, such as the measurement of transcript expression levels, the investigation of alternative splicing patterns and the detection of fusion transcripts.

Ultimately, the choice of which approach to adopt will be dependent on both the research question posed and the original hypothesis. For example, if the study was designed to identify aberrantly expressed transcripts or fusion transcripts differentiating cancer from noncancer tissues in addition to somatic mutation detection, RNA-seq would clearly be the method of choice. On the other hand, if the aim was solely to identify somatic mutations in cancer tissues, both methods could be applied depending upon whether the detection of somatic mutations in nontranscribed genes would be important to the research question posed; if so, then WES would be the approach of choice. By contrast, if one was interested in detecting somatic mutations in transcribed genes, then RNA-seq would be the more appropriate technique to use. However, it should be noted that mutations in low-abundance transcripts might not be detected accurately in the absence of a high depth of sequencing coverage. We and others believe that WES is technically and bioinformatically less challenging for interrogating somatic mutations; this viewpoint is supported by WES being more widely applied in studies of cancer mutations than RNA-seq [15,44,45,71–74]..

Five-year view

The main reason to use RNA-seq as a means to detect variants in coding regions would be its cost-effectiveness, since this approach obviates the need for exome-enrichment steps (assuming that the sequencing cost is comparable to WES). A direct total cost comparison between these two approaches is difficult, because these technologies (and their associated costs) are rapidly changing and also differ by vendor. This weakens the justification for applying RNA-seq based on cost alone. However, even if RNA-seq is cheaper, the challenge of detecting mutations using this approach must be appreciated. At present, RNA-seq has not been applied as widely as WES in variant detection. The advent of sample barcoding protocols in the prehybridization steps may be expected to reduce the cost of exome enrichment significantly in WES.

Moving beyond the coding regions, WES is likely to be a transient technology that will eventually be replaced by WGS. However, several factors, including the total cost (sequencing costs plus other indirect costs incurred for bioinformatic analysis and data storage), analytical challenges of a large dataset and our limited ability to interpret the functional/clinical significance of variants in noncoding regions, have for the time being made WES a more popular method. Furthermore, the number of variants identified in the exome or transcriptome is much more manageable (and easier to prioritize for subsequent validation) than is the case for WGS. It is nevertheless foreseeable that these challenges will be overcome in the not-too-distant future. The cost of WGS to a sequencing depth of 50x provided as a commercial service is now below US$5000. As the total cost of WGS becomes more affordable, it is expected to become the dominant tool for many applications in structural and functional genomics studies, including variant detection in the entire genome. It is expected that the generation of excess information will then become of minor importance. Hence, our current difficulty in interpreting variants in noncoding regions will not always serve to impede the application of WGS. Similarly, RNA-seq will also benefit from further technological and bioinformatic advances. Here, RNA-seq refers specifically to the sequencing of mRNAs. However, in a wider context, the transcriptome encompasses all the transcripts, including coding (mRNAs) and noncoding RNAs, such as miRNAs and long intergenic noncoding RNAs. Future advances in sequencing technologies will enable complete transcriptome sequencing to allow variant detection in both coding and noncoding regions. More importantly, the key role of RNA-seq in interrogating noncoding RNAs in cancer has been recently demonstrated. A comprehensive analysis of long noncoding RNAs in 102 prostate cancer tissue samples and cell lines by deep RNA-seq identified 121 noncoding RNAs, termed prostate cancer-associated noncoding RNA transcripts, whose expression patterns appear to be capable of distinguishing benign, localized cancer from metastatic cancer samples, suggesting that cancer-specific functions of these noncoding RNAs may help to drive tumorigenesis [75]. This study demonstrates the utility of RNA-seq in defining functionally important (yet unannotated) elements of the genome. It must also be noted that beyond the ability to detect clinically important variants of noncoding RNAs, present data have also shown the importance of detecting variable levels of expression of these noncoding RNAs. The functional role of long noncoding RNAs in cancer and the important role of RNA-seq in identifying the relevant noncoding RNAs are increasingly being recognized [75–77]. Hence, the importance of RNA-seq in defining the complement and the abundance of protein-coding and noncoding RNAs should be appreciated.

Acknowledgements

C-S Ku contributed to the conceptualization of this article. G-S Ku, M Wu and DN Cooper contributed to the writing of the article and the preparation of the table. N Naidoo, Y Pawitan, B Pang, B Iacopetta and R Soong were involved in the discussion and critical reading. C-S Ku approved the final version and had final responsibility for this article.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.
Key issues

- The detection and characterization of genetic variations in the human genome have been greatly facilitated by next-generation sequencing technologies such as whole-genome sequencing.
- The high cost of whole-genome sequencing, together with the challenges inherent in analyzing and interpreting variants detected in noncoding regions, have made whole-exome sequencing (WES) a popular approach in the context of variant detection.
- WES has been applied to the detection of both germline and somatic variants and de novo variants in trios.
- Since WES focuses specifically on the coding regions, exome-enrichment steps are required before the genomic DNA can be subjected to massively parallel sequencing, thereby adding substantially to the total cost of WES.
- To further optimize the cost-effectiveness of variant detection within coding regions, transcriptome or RNA sequencing (RNA-seq) has been proposed as a potential substitute for WES.
- Because the transcriptome from a specific tissue/cell only represents a subset of the exome, only the variants in the expressed genes or transcripts in that tissue can be detected by RNA-seq.
- It is likely that deleterious somatic mutations, such as nonsense/protein-truncating mutations in expressed genes, are more likely to be functionally important.
- The considerable variability in transcript expression levels also presents a critical challenge to the detection of variants using RNA-seq.
- The choice of which approach to adopt will be dependent on both the research question posed and the original hypothesis.
- Unlike WES, where the major application is variant detection, RNA-seq has other applications, such as the measurement of transcript expression levels and the detection of novel fusion genes.

References

Papers of special note have been highlighted as:
- of interest
- of considerable interest

7 Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2(11), 84 (2010).
8 Shones A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 12(8), 125 (2011).
15 Single
16 -One of the recent studies that have applied WES to identify cancer somatic mutations.
20 One of the first studies to demonstrate the feasibility of WES to identify new causal mutations and genes for Mendelian disorders with previously unknown genetic etiology.
25 One of the recent studies that have applied WES to identify cancer somatic mutations.
28 Comprehensive comparison of three major commercial exome sequencing platforms from Agilent, Illumina® and Nimblegen applied to the same human blood sample.
29 Mertes F, Elsharawy A, Sauer S et al. Targeted enrichment of genomic DNA


**One of the first studies to apply the RNA sequencing (RNA-seq) technique to identify single-nucleotide variants in expressed exons from the human genome.**


**A systematic evaluation of the performance of RNA-seq to identify human coding variants by comparing variants identified through high-coverage whole-genome sequencing to those identified by high-coverage RNA-seq in the same individual.**


**The first proof-of-principle study to demonstrate the feasibility of using WES in a diagnostic context.**


47 One of the first studies to demonstrate the feasibility of WES in identifying *de novo* variants in trios by sequencing the exomes of 20 individuals with a sporadic autism spectrum disorder and their parents.


51 Sztiziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 12(9), 227 (2011).


**One of the successful studies of cancer mutations using the RNA-seq approach.**


**One of the successful studies of cancer mutations using the RNA-seq approach.**


TruSeq Exome Enrichment Kit. www.illumina.com/products/truseq_exome_enrichment_kit.ilmn

MiSeq Personal Sequencer. www.illumina.com/systems/miseq.ilmn