Current Genetic Medicine Reports (2021) 9:13–21
https://doi.org/10.1007/s40142-021-00199-x

CLINICAL GENETICS (D MORRIS-ROSENDAHL, SECTION EDITOR)

The Role of RNA-Sequencing as a New Genetic Diagnosis Tool

Philippa D. K. Curry[1,2] · Krystyna L. Broda[1,3] · Christopher J. Carroll[1]

Accepted: 12 March 2021 / Published online: 23 April 2021

Abstract

Purpose of Review Whole exome sequencing (WES) and whole-genome sequencing (WGS) are frontline approaches for the genetic diagnosis of rare diseases. However, WES/WGS fails in up to 75% of cases. Transcriptomics via RNA-sequencing (RNA-Seq) is a novel approach that aims to increase the diagnostic yield in rare diseases.

Recent Findings Recent publications focus on the success of RNA-Seq for increasing diagnosis rates in WES/WGS-negative patients in up to 36% of cases, across a range of different diseases, sample sizes, and tissue types.

Summary RNA-Seq is beneficial for aiding prioritisation of causative variants currently not detected or often overlooked by WES/WGS alone. An improvement in diagnostic yields has been demonstrated using multiple source tissues, with muscle and fibroblasts being the most representative, but the more accessible blood still demonstrating diagnostic success, particularly in neuromuscular disorders. The introduction of RNA-Seq to the genetic diagnosis toolbox promises to be a useful complementary tool to WES/WGS for improving genetic diagnosis in patients with rare disease.

Keywords RNA-sequencing · Rare disease · Transcriptomics · Next-generation sequencing (NGS) · Whole exome sequencing (WES)

Introduction

Rare diseases affect over 350 million individuals worldwide, with ~8% of live births having a Mendelian genetic disorder recognisable by early adulthood [1, 2]. Accurate identification of causal variants in rare diseases is imperative, not only providing the patient with a genetic diagnosis and ending years of diagnostic odyssey but also permitting more accurate prognosis, improving family risk planning, and enabling research for novel therapeutic interventions. Improvement of genetic sequencing techniques, from Sanger sequencing through to next-generation sequencing (NGS), has seen a vast expansion in the ability to identify variation in individuals. This review briefly describes the history of causal variant identification, before exploring the advantages and disadvantages of using RNA-sequencing (RNA-Seq) as a pioneering method in the genetic diagnosis of a variety of rare diseases. We emphasise its success in improving diagnostic yields, across an array of tissue sources and cohort sizes, when used independently and as part of a larger multi-omics approach.

This article is part of the Topical Collection on *Clinical Genetics*

✉ Christopher J. Carroll
ccarroll@sgul.ac.uk

Philippa D. K. Curry
philippa.curry@postgrad.manchester.ac.uk

Krystyna L. Broda
k.broda19@imperial.ac.uk

[1] Genetics Research Centre, Molecular and Clinical Sciences Research Institute, St. George's, University of London, Cranmer Terrace, London SW17 0RE, UK

[2] Centre for Musculoskeletal Research, School of Biological Sciences, University of Manchester, Manchester, UK

[3] Centre for Blast Injury Studies, Department of Bioengineering, Imperial College London, London, UK

Next-Generation Sequencing (NGS)

Next-Generation Sequencing (NGS) is used in various applications such as examining DNA sequences (whole-genome sequencing (WGS), whole-exome sequencing (WES)), investigating DNA and histone modifications (ChIP-Seq, Methyl-Seq, ATAC-Seq), and RNA-Sequencing, with the commercial sequencing technologies currently being dominated by Illumina [3]. WGS allows sequencing of the genome with near complete coverage, and typically will identify ~3 million

single nucleotide variants (SNVs) [4, 5]. WES focuses only on the protein-coding region of DNA, which constitutes ~2% of the genome, and reduces the average number of SNVs to 23,000 [4, 5].

WES is currently a first-tier approach for genetic diagnosis of Mendelian disorders, due to the vast majority of disease variants reported being located in coding regions. Improved diagnostic yields have prompted decreased healthcare costs and improvement in patients' quality and outcomes [6, 7]. However, despite an improvement in bioinformatic approaches, the diagnostic yield remains low and variable, at only 25–50%, leaving a large proportion of individuals awaiting a genetic diagnosis [4]. The limitations of WES include the ever-emerging challenge of interpreting variants of unknown significance (VUS), including variants in genes not yet associated to any disease, and the limited efficacy of WES to identify structural rearrangements, copy number variants and tandem repeat expansions, as well as G-C and A-T rich regions of the genome. Furthermore, despite previously being overlooked due to research bias and their exclusion through use of stringent filtering techniques, synonymous, intronic, and noncoding variants could also be pathogenic. Evidence suggests that 9–30% of disease-causing variants are in non-coding regions and that there are synonymous variants associated with over 50 different diseases [8]. This demonstrates the increased importance of scrutinising these overlooked variant classes.

As NGS gets cheaper and more available to diagnostic labs, the far more comprehensive technique of WGS will rise in prominence as it does sequence non-coding regions. However, the sheer amount of data poses the large bioinformatic challenge of interpreting 3 million SNVs per sample, followed by the additional functional validation. One particular type of variant that causes bioinformatic difficulties with current approaches is splicing variants. These are reported to account for up to 10% of disease causing variants, a figure speculated to be underestimated [9]. Consequences of these mutations can lead to exon creation, skipping, truncation, extension, and intron retention [10]. These pose analytical challenges when investigating WES and WGS data. Although canonical splice mutations can be identified using these techniques, both cryptic and enhancer sites still provide interpretation challenges. Moreover, the vast majority of intronic variants are not identified at all by WES, whilst they are too abundant to prioritise from WGS data and most often remain as VUSs. Thus, there is a mass shift in the challenges faced by clinicians and scientists today, away from being able to sequence genomes comprehensively and more towards interpreting the large datasets generated by them and translating this information into actionable results. Additionally, the high number of VUSs identified highlights the importance of re-filtering NGS data as new disease genes emerge. Recently, RNA-sequencing-based transcriptomics is starting to become a major player in genetic diagnostics for aiding annotation and interpretation of these types of variants.

RNA-Sequencing

Since the discovery of reverse transcriptases and their use to create cDNA in 1971, followed by the ability to measure RNA on a singular gene basis using RT-PCR in 1990, analysis of RNA has been available to observe the effect of pathogenic variants [11, 12]. RNA-Seq allows the entire transcriptome (~12,800 transcripts) to be analysed in a single run [13]. Initially, this technology was clinically applied for determination of viral gene expression, monitoring of immune responses, or for the most part, in cancer diagnostics to readily detect gene fusions [14–17]. More recently, RNA-Seq has been used successfully to genetically diagnose those with Mendelian disorders as the primary method or as part of a larger multi-omics approach. Whereas RT-PCR has been previously utilised for its abilities to confirm candidate variants, the ever-reducing costs of RNA-Seq make it an equally viable option not only for validation purposes but also in providing additional transcriptome annotation [5, 18].

Briefly, the protocol used for RNA-Seq is total RNA isolation and purification and library preparation; enrichment for RNA of interest, e.g. filtered for poly(A) tails to capture mRNA and reverse transcription to cDNA; fragmentation of sample, addition of adaptors and barcodes/indexes; and amplification. Next automatic sequencing occurs, followed by reassembly of transcriptomic data, where the results are bioinformatically analysed (Fig. 1) [19]. There are numerous tools available to fully analyse the data, though comparison of RNA data with that of a comprehensive control dataset, such as data provided by the GTEx consortium atlas of genetic regulatory effects across human tissues (GTEx), is recommended. GTEx is an ongoing effort towards building a public resource for scientists to use, which is comprised of data from 54 human tissues from approximately 1000 donors [20]. Alternatively, independent study control data sets have also been utilised instead of GTEx in multiple studies of varying cohort sizes [4, 21]. RNA-Seq technology enables detection of transcriptome defects, a functional consequence of deleterious variants that usually cannot be predicted from the WES/WGS data alone. As underlined in a previous review by Kremer et al. [5], it enables the identification of aberrant gene expression levels between control and experimental samples, reveals allele-specific expression, and provides information on aberrant splicing events (*vide infra*).

Studies Utilising RNA-Seq in Genetic Diagnostics

The quantity of papers that have started utilising transcriptomic techniques to improve diagnostic levels in WES/WGS-negative patients has increased over the past few years (Table 1). Primary articles of note on this subject include

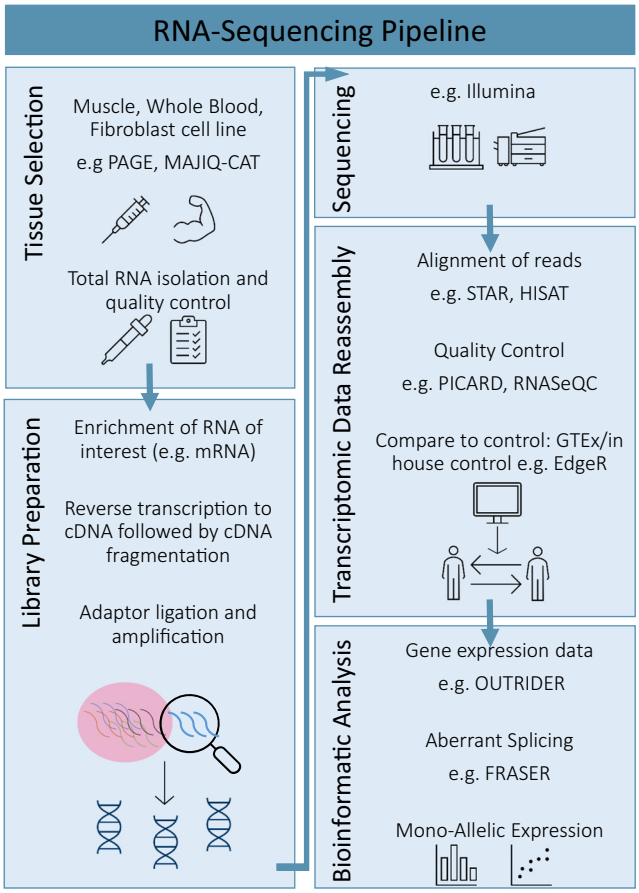

Fig. 1 RNA-sequencing protocol. RNA-sequencing offers the opportunity to add additional variant annotation to individuals who have previously undergone other forms of genetic testing either unsuccessfully or resulting in variants of unknown significance (VUS). Additionally, it can be used as a completely independent diagnosis technique for novel patients. RNA-Seq protocol steps include the isolation of RNA from chosen tissue; library preparation; sequencing; reassembly of transcriptomic data; and bioinformatic analysis of results. Resources often utilised during sequencing include PAGE [23], MAJIQ-CAT [43], Illumina [45], STAR [46], HISAT [47], PICARD [48], RNASeQC [49], EdgeR [50], OUTRIDER [31], and FRASER [40]. Data achievable from RNA-Seq encompasses gene expression, aberrant splicing, and mono-allelic expression

Cummings et al. [22] who yielded a genetic diagnosis in 35% of their cohort of patients with previously undiagnosed muscle disorders; Kremer et al. [4] attained a genetic resolution of 10% in their cohort of patients with diverse phenotypes of mitochondrial disease; Kernohan et al. [21] explored the potential use of whole blood for RNA-Seq in a single patient; as well as Gonorazky et al. [23] who described a 36% genetic diagnosis in their cohort of patients with previously undiagnosed neuromuscular disorders. Furthermore, even more recent papers include Frésard et al. [24], who reported a 7.5% diagnostic rate utilizing multiple source tissues for a diverse range of diseases, and Maddirevula et al. [25] who support the use of blood as a source of RNA material in neurological disorders. In contrast, Rentas et al. [26] explore the option of using B-lymphoblastoid cell lines (LCL) from patients to analyse expression and splicing. Additionally, Lee et al. [27] find an 18% diagnostic yield in a heterogenous cohort, while Murdock et al. [28] suggest a novel method of integrating RNA-Seq data that contrasts the commonly used candidate gene approach used in most studies. Instead, Murdock et al. suggest starting with RNA-Seq and, thus excitingly, advocate its role in research and gene discovery. These papers will be further discussed in order to demonstrate the success of RNA-Seq thus far, specifically in aberrant expression levels, splicing events and allele specific expression, as well as to highlight difficulties still present in this continually developing approach and discuss its place in both the diagnostic and research setting.

Aberrant Gene Expression Levels

Aberrant gene expression levels can be described as a gene expression level outside its normal physiological range. This process is tightly regulated by chromatin packing, histone modification, transcription initiation, RNA polyadenylation, splicing, and translation initiation, with both exonic and intronic variants disrupting these functions. Approximately 1% of disease-causing variants are estimated to be located within promoter regions, thus disrupting gene activation and transcriptional initiation [29]. Furthermore, splicing mutations can also lead to changes in gene expression and will be discussed within the context of the limitation of WES and WGS in further detail (*vide infra*).

Success of utilising RNA-Seq to investigate gene expression has been reported with an additional diagnostic yield of 10% by Kremer et al. [4]. In particular, RNA-Seq enabled identification of reduced expression of a nuclear encoded mitochondrial protein (*MGST1 and TIMMDC1)* in three patients (*MGST1 (n = 1) and TIMMDC1 (n = 2)* from a mitochondriopathy cohort ($n = 48$). Here, quantitative proteomics validated severe loss-of-function in these patients. Interestingly, this included a deeply intronic variant in *TIMMDC1*, which could never be identified in WES whilst overlooked in WGS, but RNA-Seq helped identify it as a novel disease–associated gene with mitochondrial disease. This emphasises the use of RNA-Seq not only in genetic diagnosis but also in gene discovery, as well as the importance of intronic variants in disease.

Frésard et al. [24] highlights the need for stringent filtering criteria when investigating gene expression, with 343 expression outliers averaged per sample ($n = 94$). Following the prioritisation of the Human Phenotype Ontology (HPO) matching genes, containing deleterious (CADD ≥ 10), rare (minor allele frequency (MAF) $\leq 0.1\%$) variants within 20 bp, ~1% of identified outliers remained, with at least 80% of samples having at least one candidate gene remaining. Novel

Table 1 RNA-Seq as a diagnostic tool in Mendelian disease

	Article	Phenotypes	Tissue	Prior investigations	RNA-Seq analysis	RNA-Seq diagnostic yield
1	Cummings *et al.* 2017 [22]	Rare monogenic muscular disorders	Skeletal muscle	WES and WGS	Aberrant splicing	35% (n = 50) 17 solved, 2 VUS, 8 candidate genes identified
2	Kremer *et al.* 2017 [4]	Mitochondrial disorders	Fibroblast cell lines	WES and WGS	Aberrant expression Aberrant splicing MAE	10% (n = 48)
3	Kernohan *et al.* 2017 [21]	Spinal muscular atrophy (SMA)	Whole blood	*SMN1* deletion testing Chromosomal microarray SMA gene panel testing	Splicing	100% (n = 1)
4	Gonorazky *et al.* 2019 [23]	Neuromuscular disorders	Muscle Fibroblasts T-myotubes	WES Gene panels	Aberrant expression Aberrant splicing MAE	36% (n = 25)
5	Frésard *et al.* 2019 [24]	80 different rare diseases Including: neurological, musculoskeletal, haematology and ophthalmology as the most frequent	Whole blood	WES and WGS	Aberrant expression Aberrant splicing MAE	7.5% (n = 94) 16.7% (n = 94) candidate genes identified
6	Lee *et al.* 2020 [27]	Wide range of undiagnosed Mendelian diseases Including Neurological, musculoskeletal, gastroenterological, endocrinological and dermatology	Whole blood Skin fibroblasts Muscle	WES and WGS	Aberrant splicing	18% (n = 100)
7	Rentas *et al. 2020* [26]	Neurodevelopmental disorders	B-lymphoblastoid cell lines	WES	Aberrant expression Aberrant splicing	60% (n = 5)
8	Murdock *et al.* 2020 [28]	Wide range of undiagnosed Mendelian diseases Including: neurological , musculoskeletal, immune phenotypes as the most frequent	Whole blood Fibroblasts	WES and WGS	Aberrant expression Aberrant splicing	17% (n = 82)
9	Maddirevula *et al.* 2020 [25]	Rare Mendelian phenotypes	Whole blood Skin fibroblasts Urine-derived renal epithelial cells	WES	Aberrant expression Aberrant splicing	13.5% (n = 155)

genetic causes were discovered in four patients following gene expression analysis, including two brothers experiencing delayed milestones and hypotonia for whom a large number of variants remained following WGS analysis ($n = 245$ and 302). The addition of expression data substantially reduced the number of candidate variants ($n = 11$ and 15), including a heterozygous pathogenic variant in *MECR* in both siblings. This provided a MEPAN disorder diagnosis for both patients [30] and demonstrated the advantage of combining RNA-Seq with WGS to answer the challenges of variant prioritisation.

Murdock et al. [28] support the idea of using a multi-omics approach rather than solely WGS or WES for intronic variants after a deeply intronic variant was observed in a 3-yr-old male patient with multiple congenital abnormalities. Here, a 50% reduction in the expression of *PQBP1* from whole blood was associated to a hemizygous variant in *PQBP1*, resulting in an activated cryptic splice donor, abnormal splicing pattern and intron retention. This had been previously overlooked by WES and WGS due to the poor reliability of bioinformatic *in silico* prediction tools concerning splice-site mutations. Identification of this maternally inherited hemizygous variant resulted in a diagnosis of Renpenning syndrome. This discovery allowed recurrent risk discussions within the family, highlighting the importance of an accurate genetic diagnosis.

Furthermore, Gonorazky et al. [23] developed PAGE (Panel Analysis of Gene Expression), a webtool which enables visualisation of expression levels across multiple tissues. Even more interestingly, it identifies the best tissues to study when performing RNA-Seq, as "disease" tissue is not always easily attainable. Excitingly, statistical methods such as OUTRIDER have also been recently developed to further aid the identification of abnormal gene expression [31]. Altogether, this emphasises the increasing usefulness of RNA-Seq in genetic diagnosis. Whilst WES data analysis could potentially identify a synonymous VUS, it would most likely have been excluded during variant prioritisation, whereas an intronic variant would likely not even be sequenced. Thus, quantifying global RNA expression by RNA-Seq has proved to be invaluable in interpreting VUSs for genetic diagnosis.

Aberrant Splicing

Splicing is estimated to occur in ~94% of human genes, with alternative splicing occurring in the vast majority of these [32]. Both intronic and exonic variants can affect splice sites, or cis-regulatory sites, and can therefore have drastic consequences and lead to aberrant splicing. These include exon skipping, exon extension, exon truncation or deletion, intron retention, pseudoexon creation and alteration in isoform abundance [10]. The effects of these variants can range in severity and can trigger loss of function or nonsense-mediated decay and affect gene expression, and therefore accurate evaluation of these mutations is vital.

As mentioned above, splice variants are not only highly detrimental, accounting for at least 15% and potentially up to 60% of genetic diseases [33, 34], but also are located in both exons and introns. An obvious limitation of WES for identifying individuals with these types of variants is that splice-modifying variants that lie beyond the exome are not captured. Furthermore, it is commonly seen that splicing variants lying within the sequencing range are still often overlooked, thus highlighting downfalls of current prioritisation protocols, which also extends to WGS techniques. Kernohan et al. [21] emphasise that despite the development of many *in silico* prediction tools over the last 20 years to help prioritise potentially pathogenic variants (including MaxEntScan [35], GeneSplicer [36] , SPANR [37] and VEP [38]), many are still overlooked. Equally, due to an underappreciation and the undereducation of the impact of splicing variants, particularly synonymous variants and variants lying outside of the coding regions, prediction tools are reported to overlook splicing events and provide incorrect predictions up to 76.1% of the time [39]. Lee et al. [27] support this suggestion, reporting only 4.3% of predicted damaging splice variants actually resulting in an impact after RNA-Seq validation. This is an area in need of improvement in order for splicing prediction to become accurate and routine using WES or WGS and suggests an underestimation of their potentially pivotal role in identification providing genetic diagnosis. RNA-Seq, however, aids the annotation of these synonymous variants or VUSs in order to validate disease involvement. Additionally, the recent development of the FRASER algorithm holds promise for effective identification of these variants [40].

The advantages of RNA-Seq were noted by Kremer et al. [4] who demonstrated that although candidate splicing variants were often sequenced in WGS and WES, few were being validated due to lack of appropriate assays and biomaterials. This therefore emphasises the importance of combining traditional NGS techniques with RNA-Seq in both diagnostic and research settings to tackle the increasing problem of VUS interpretation and identifying true damaging variants. With five aberrant splice sites identified on average in their cohort of neuromuscular and inborn errors of metabolism (IBM) patients, exon skipping and new exon formation were identified as the most common types of splicing mutations. The importance of quantitative proteomics and validation of any identified variants via western blotting was also emphasised. Cummings et al. [22] created a splicing prediction algorithm in 50 WES/WGS-negative muscle disorder patients, detecting abnormal exon junctions and abnormal splicing events in up to 190 genes per individual, compared to GTEx control data. This included synonymous variants in both *RYR1* (n = 1) and *POMGNT1* (n = 1) that were discovered to influence aberrant splicing in patients. Furthermore, they identified deep intronic

variants in *DMD* that led to pseudoexon construction and early stop codon creation in three patients. For both types of variants their contribution would have been typically overlooked amongst the mass of WGS variants prioritized and even undetected using current frontline WES diagnostic techniques.

Kernohan et al. [21] tackled the largest difficulty that arises with application of using RNA-Seq—tissue availability. In contrast to Cummings et al. [22], the availability and non-invasive nature of collecting RNA from whole blood are widely appreciated. Kernohan et al. [21] demonstrate the capability of using RNA-Seq on blood leukocytes to identify the underlying pathogenic variant in a single patient. Here, they detected a previously unidentified splice mutation in *ASAH1*, resulting in an autosomal recessive spinal muscular atrophy diagnosis. This indicates the value of transcriptomics even in smaller cohorts or even individual patients.

Frésard et al. [24] and Lee et al. [28] expand this work further by conducting RNA splicing studies on larger heterogenous disease cohorts (n = 94 and n = 100, respectively), containing adult and paediatric patients, on whole blood as well as other tissues, including muscle and fibroblast cells. Both studies reported increased diagnostic rates (7.5 and 18%, respectively) across a range of disorders, with Frésard et al. [24] stressing the feasibility of using blood in patients where biopsy is not routine. Variants identified were both intronic and exonic mutations that triggered a wide range of splicing events, despite previously being completely missed or overlooked VUSs during WES and WGS. Of note again, neuromuscular and musculoskeletal patients were amongst the highest proportion of patients in these studies, suggesting a bias of success for these types of rare disease patients.

Strict filtering criteria (HPO relevant genes with a deleterious variant (CADD ≥ 10) within 20 bp of a splice junction) enabled Frésard et al. [24] to achieve a reduction of candidate variants to 0.14%, with 32% of individuals still remaining with at least one gene candidate. It was commented that such stringent filtering will still lead to a loss of some potentially pathogenic variants; thus expert analysis is still vital to conduct RNA-Seq analysis effectively. Throughout the studies, GTEx data was commonly used as control data. Frésard et al. [24] also emphasise the importance of accounting for batch effects in large control cohorts, such as GTEx, as well as matching controls for tissue and sex, in order to show the largest difference in RNA data and identify variants. Overall, the *in silico* prediction tools available have been seen to be inconsistent in the prediction of splicing variants. RNA-Seq offers an opportunity to correctly identify both intronic and exonic variants leading to varied splicing events across a range of diseases and tissues.

Allele-Specific Expression or Mono-Allelic Expression (MAE)

Mono-allelic expression (MAE) refers to the incidence where imbalanced expression of alleles occurs, or only one of a patients two alleles is being expressed. The silencing of one allele may occur due to deletions or variants that cause genetic or epigenetic changes. This is a phenomenon that can occur normally with well-known examples including immunoglobulin gene exclusion, X-inactivation, and genomic imprinting [41]. However, abnormal allele expression or allelic imbalance can result in a clear disease phenotype, such as Prader-Willi syndrome [42]. In these scenarios, the investigation of unbalanced allelic expression is an area of interest in genetic diagnosis that RNA-Seq can shed light on, by aiding variant annotation.

During the filtering and prioritisation of compound heterozygous variants following WES or WGS, a recessive form of inheritance is assumed; thus single heterozygous rare coding variants are often dismissed. However, Kremer et al. [4] suggest how these variants when solely expressed due to allelic silencing mimic homozygous variants with similar and identifiable effects at RNA level. They prioritised rare SNVs (minor allele frequency (MAF) < 0.001) and high coverage (> 10)) following RNA-Seq that were mono-allelically expressed in more than 80% of these reads (p < 0.05). This resulted in an average of 6 MAE events per sample, a number noted to be small enough for manual inspection, in contrast to WGS. Here, aberrant expression and MAE reprioritized an *ALDH18A1* variant in a WES-negative patient, after it was identified in a compound heterozygous state with a previously annotated VUS nonsense variant. Proteomic validation of the variant supported pathogenicity.

Frésard et al. [24] proposed a higher rate of MAE per sample (n = 94) vs. GTEX controls. Following filtering of rare and deleterious variants in their heterogenous cohort (n = 94), 111 variants demonstrated allelic imbalance towards the mutated allele, including 96 splice variants and 15 stop-gain mutations. In particular, a *EFDH2* variant, involved in B-cell apoptosis and inflammatory response, led to the genetic diagnosis of a previously undiagnosed individual with idiopathic cardiomyopathy. This example highlights the potential of using RNA-Seq in combination with other NGS techniques to identify variants or even reprioritize variants previously discarded due to lack of annotation available.

Limitations of RNA-Seq

Tissue Specificity

Despite recent publications succeeding in providing an increased diagnostic rate of 7.5–36%, limitations to RNA-Seq

still remain. The most prominent challenge is finding the appropriate balance of tissue specificity of gene expression vs. accessibility and invasiveness of tissue collection. Some larger studies have focused on mitochondrial and neuromuscular diseases, where muscle biopsies and fibroblast collection are routine as part of disease management. The benefit of using fibroblasts in mitochondrial disorders was supported by Kremer et al. [4] highlighting that 68% of disease genes reported in OMIM are expressed in fibroblasts and with many recent papers still arguing the ineffectiveness of blood for being representative in many conditions, despite it being easily accessible and feasible [23, 27, 28]. In this regard, Murdock et al. [28] reported that half of the pathogenic splice variants identified did not have an observable functional consequence following RNA-Seq in blood, whereas all had a functional consequence in fibroblasts. Furthermore, principal component analysis (PCA) studies reveal that fibroblasts have genes with comparatively better expression levels (Transcript Per Million (TPM) > 10) vs. blood, most prominently in aorta cardiomyopathy genes (80% vs. 24%) and with the only anomaly being immunodeficiency genes (45% vs. 58%). In contrast, Frésard et al. [24] reports 70.6% of OMIM genes to be expressed in blood, as well as 76% of neurological disease panel genes (n = 284) being well expressed in blood. Furthermore, they go on to suggest the appropriateness of blood, as causative loss of function (LOF) and missense mutations were seen at higher levels in genes that are expressed in multiple tissues with 66% of genes that are most intolerant to LOF mutations (pLI < 0.9) being expressed in blood. Interestingly, Maddirevula et al. [25] and Kernohan et al. [21] support the use of blood, with Kernohan et al. [21] showing that 71% of motor neuron panel genes were seen expressed in the blood of at least half of their cohort. Musculoskeletal and neuromuscular diseases are still frequently those seen to be successfully aided from RNA-Seq experiments, with appropriate tissues available. Rentas et al. [26] highlight the use of LCL over blood, demonstrating 90% in sensitivity using their pipeline which utilises splicing and expression data, with a reported 1.8 fold higher expression of neurodevelopmental genes in LCL than in blood, in their cohort (n = 5). Nevertheless, many studies have pinpointed neurologically pathogenic variants in blood and other tissues, despite their non-haematological phenotypes [24]. The recently developed MAJIQ-CAT resource offers the opportunity to identify clinically accessible tissues offering the best representation of splicing events in the gene or tissue of interest [43]. Overall, the discussion about the best tissue to use for RNA-Seq is ongoing, with the general consensus being that if the affected tissue is easily accessible, then it should be prioritized for use in RNA-Seq. However, when not available, whole blood or other GTEx identified "proxy" tissues have still been seen to effectively diagnose individuals with a variety of pathologies.

Potential emerging techniques could begin to tackle the tissue specificity problem where disease-relevant tissue is unattainable. One option is the non-trivial procedure of reprogramming patient cells of different tissue types into "disease" cells. Yamanaka's factors can induce specific forced gene expression from easily accessible tissues from the patient, resulting in induced pluripotent stem cell (iPSC) formation that can then be differentiated into the desired and patient-specific disease tissue type [44]. Alternatively, a more recent advancement using transdifferentiation techniques may be used to produce patient-specific cells in a faster and more efficient manner [23]. Another aspect to consider is that RNA-Seq typically relies on the incorrect assumption that all cells within the same tissue are relatively homogeneous, an exciting potential solution here being single-cell RNA-Sequencing (scRNA-Seq).

Database Biases

The variability of gene expression throughout a person's lifespan is another consideration for RNA-Seq. Despite the vast resources already available, there is still a need for more comprehensive gene expression datasets encompassing a range of ages. GTEx donor information describes that the majority are aged between 50 and 70 years (64.7%), causing potential difficulties in paediatric cases [20]. Further database biases are noted including gender and race whereby 67.1% of data is provided from male donors, and the racial input of data is predominantly white (84.6%) [20]. Increased variation in database recruitment, alongside control matching could further increase diagnostic yields.

Validation of Transcriptomic Data

Finally, RNA-Seq will provide information on gene expression, allele-specific expression, and splicing events; however, gene expression is not always a good proxy for protein abundance. Following transcription, regulation at or prior to the translational level is also present, e.g. nonsense-mediated decay (NMD). Therefore, validating the impact of RNA splicing variants at the level of the protein is good practice.

Conclusion

In conclusion, despite the exponential growth and development of NGS techniques, at least 50% of individuals investigated with WES or WGS for a genetic diagnosis remain undiagnosed, thereby hindering disease management at many levels. Reasons include sequencing bias of WES, alongside poor *in silico* prediction tools and filtering, and the vast number of VUSs produced in WGS. Recent publications highlight

the success of employing transcriptomic data provided by RNA-Seq to improve diagnostic yields in a wide range of rare diseases, using both fibroblasts and muscle tissue, with muscular and neurological conditions reporting the most success. Publications demonstrate the significance of conducting RNA sequencing in both individual undiagnosed cases and larger cohorts, particularly in highlighting previously underappreciated non-coding and synonymous variants. Tissue specificity remains an issue, though even whole blood has been seen to successfully increase diagnostic rates. This suggests that efficient protocols for the use of RNA-Seq in both diagnostic and research settings are established. It should however be acknowledged that variants will still be overlooked, even when applying RNA-Seq, namely, if they are not detectable by short-read sequencing such as RNA-Seq, filters applied to the transcriptomic data are too stringent, or if the causative variant does not affect expression. Overall, reported studies emphasise the versatility and practicality of using RNA-Seq as part of a multi-omics approach in genetically undiagnosed cases.

Declarations

Conflicts of Interest

The authors declare no competing interest.

References

Papers of particular interest, published recently, have been highlighted as:

- • Of importance
- •• Of major importance

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(D1):D789–98.
2. Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. Am J Hum Genet. 1988;42(5):677–93.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
4. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8:1–11.
5. Kremer LS, Wortmann SB, Prokisch H. "Transcriptomics": molecular diagnosis of inborn errors of metabolism via RNA-sequencing. J Inherit Metab Dis. 2018;41(3):525–32.
6. Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, et al. Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. JAMA Pediatr. 2017;171(9):855–62.
7. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090–6.
8. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12(10):683–91.
9. Montes M, Sanford BL, Comiskey DF, Chandler DS. RNA Splicing and disease: animal models to therapies. Trends Genet. 2019;35(1):68–87.
10. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 2018;59(3):253–68.
11. Gallo RC. Molecular biology: reverse transcriptase, the DNA polymerase of oncogenic RNA viruses. Nature. 1971;234(5326):194–8.
12. Shaffer AL, Wojnar W, Nelson W. Amplification, detection, and automated sequencing of gibbon interleukin-2 mRNA by Thermus aquaticus DNA polymerase reverse transcription and polymerase chain reaction. Anal Biochem. 1990;190(2):292–6.
13. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
14. Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun. 2019;10(1):1–12.
15. Reeser JW, Martin D, Miya J, Kautto EA, Lyon E, Zhu E, et al. Validation of a targeted RNA sequencing assay for kinase fusion detection in solid tumors. J Mol Diagnost. 2017;19(5):682–96.
16. Liu Y, Hu J, Liu D, Zhou S, Liao J, Liao G, et al. Single-cell analysis reveals immune landscape in kidneys of patients with chronic transplant rejection. Theranostics. 2020;10(19):8851–62.
17. Depledge DP, Mohr I, Wilson AC. Going the distance: optimizing RNA-Seq strategies for transcriptomic analysis of complex viral genomes. J Virol. 2018;93(1):1–9.
18. Wang K, Kim C, Bradfield J, Guo Y, Toskala E, Otieno FG, et al. Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement. Genome Med. 2013;5(7):67.
19. Besser J, Carleton HA, Gerner-smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect. 2018;24(4):335–41.
20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E. The genotype-tissue expression (GTEx) project : nature genetics : Nature Publishing Group. Nat Genet. 2013;45:580–5.
21. Kernohan KD, Zappala Z, Hartley T, Kevin S, Wagner J, Xu H, et al. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. Hum Mutat. 2017;38(6):611–4.
22. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209.
23. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. Am J Hum Genet. 2019;104(3):466–83.

24. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25(6):911–9.
25. Maddirevula S, Kuwahara H, Ewida N, Shamseldin HE, Patel N, Alzahrani F, et al. Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. Genome Biol. 2020;21(1):1–21.
26. Rentas S, Rathi KS, Kaur M, Raman P, Krantz ID, Sarmady M, et al. Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. Genet Med. 2020;22(5):927–36.
27. Lee H, Huang AY, Wang LK, Yoon AJ, Renteria G, Eskin A, et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genet Med. 2020;22(3):490–9.
28. Murdock DR, Dai H, Burrage LC, Rosenfeld JA, Ketkar S, Müller MF, et al. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J Clin Invest. 2020;131:e101500.
29. De Vooght KMK, Van Wijk R, Van Solinge WW. Management of gene promoter mutations in molecular diagnostics. Clin Chem. 2009;55:698–708.
30. Heimer G, Kerätär JM, Riley LG, Balasubramaniam S, Eyal E, Pietikäinen LP, et al. MECR mutations cause childhood-onset dystonia and optic atrophy, a mitochondrial fatty acid synthesis disorder. Am J Hum Genet. 2016;99(6):1229–44.
31. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. Am J Hum Genet. 2018;103(6):907–17.
32. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
33. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet. 1992;90(1–2):41–54.
34. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 2005;579(9):1900–3.
35. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2–3):377–94.
36. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 2001;29(5):1185–90.
37. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, RKC Y, Hua Y, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806.
38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):1–14.
39. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–44.
40. Mertes C, Scheller IF, Yépez VA, Çelik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. Nat Commun. 2021:1–13.
41. Fahrner JA, Bjornsson HT. Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin States. Annu Rev Genomics Hum Genet. 2014, 15:269–93.
42. Peters J. The role of genomic imprinting in biology and disease: an expanding view. Nat Rev Genet. 2014;15(8):517–30.
43. Aicher JK, Jewell P, Vaquero-Garcia J, Barash Y, Bhoj EJ. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. Genet Med. 2020;22(7):1181–90.
44. Warren L, Lin C. mRNA-based genetic reprogramming. Mol Ther. 2019;27(4):729–34.
45. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.
46. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
47. Kim D, Langmead B. Salzberg1 SL. HISAT: a fast spliced aligner with low memory requirements Daehwan HHS Public Access. Nat Methods. 2015;12(4):357–60.
48. Picard Tools - By broad institute [Internet]. Available from: http://broadinstitute.github.io/picard
49. Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28(11):1530–2.
50. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139–40.