frontiers in Molecular Biosciences

REVIEW
published: 01 June 2021
doi: 10.3389/fmolb.2021.647277

How Machine Learning and Statistical Models Advance Molecular Diagnostics of Rare Disorders Via Analysis of RNA Sequencing Data

Lea D. Schlieben[1,2], *Holger Prokisch*[1,2]* *and Vicente A. Yépez*[1,3]

[1]School of Medicine, Institute of Human Genetics, Technical University of Munich, Munich, Germany, [2]Institute of Neurogenomics, Helmholtz Zentrum München, Neuherberg, Germany, [3]Department of Informatics, Technical University of Munich, Munich, Germany

Rare diseases, although individually rare, collectively affect approximately 350 million people worldwide. Currently, nearly 6,000 distinct rare disorders with a known molecular basis have been described, yet establishing a specific diagnosis based on the clinical phenotype is challenging. Increasing integration of whole exome sequencing into routine diagnostics of rare diseases is improving diagnostic rates. Nevertheless, about half of the patients do not receive a genetic diagnosis due to the challenges of variant detection and interpretation. During the last years, RNA sequencing is increasingly used as a complementary diagnostic tool providing functional data. Initially, arbitrary thresholds have been applied to call aberrant expression, aberrant splicing, and mono-allelic expression. With the application of RNA sequencing to search for the molecular diagnosis, the implementation of robust statistical models on normalized read counts allowed for the detection of significant outliers corrected for multiple testing. More recently, machine learning methods have been developed to improve the normalization of RNA sequencing read count data by taking confounders into account. Together the methods have increased the power and sensitivity of detection and interpretation of pathogenic variants, leading to diagnostic rates of 10–35% in rare diseases. In this review, we provide an overview of the methods used for RNA sequencing and illustrate how these can improve the diagnostic yield of rare diseases.

Keywords: rare disorders, RNA sequencing, machine learning, statistical models, aberrant expression, aberrant splicing, mono-allelic expression

OPEN ACCESS

Edited by:
Silvia Bottini,
Université Côte d'Azur, France

Reviewed by:
Kai Wang,
University of Pennsylvania, United States
Monkol Lek,
Yale University, United States

***Correspondence:**
Holger Prokisch
prokisch@helmholtz-muenchen.de

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *29 December 2020*
Accepted: *10 May 2021*
Published: *01 June 2021*

Citation:
Schlieben LD, Prokisch H and Yépez VA (2021) How Machine Learning and Statistical Models Advance Molecular Diagnostics of Rare Disorders Via Analysis of RNA Sequencing Data. Front. Mol. Biosci. 8:647277. doi: 10.3389/fmolb.2021.647277

Abbreviations: ACMG/AMP, American College of Medical Genetics and Genomics and the Association for Molecular Pathology; ANEVA-DOT, analysis of expression variation–dosage outlier test; CADD, Combined Annotation Dependent Depletion; DNA, deoxyribonucleic acid; FDR, false discovery rate; FRASER, Find Rare Splicing Events in RNAseq; GTEx, Genotype-Tissue Expression; iPSCs, induced pluripotent stem cells; LeafCutterMD, LeafCutter for Mendelian disease; LIMS, laboratory information management system; MAE, mono-allelic expression; MMsplice, modular modeling of splicing; NMD, nonsense-mediated decay; OUTRIDER, OUTlier in RNAseq fInDER; PCA, principal component analysis; PEER, probabilistic estimation of expression residuals; PTM, posttranslational modification; RNA, ribonucleic acid; RNAseq, RNA sequencing; RPKM, reads per kilobase million; scRNAseq, single-cell RNA sequencing; SIFT, sorting intolerant from tolerant; SNV, single nucleotide variant; SPOT, SPlicing Outlier deTection; SVA, surrogate variable analysis; SV, surrogate variable; TPM, transcripts per million; V^G, genetic variation; WES, whole exome sequencing; WGS, whole genome sequencing.

INTRODUCTION

Rare diseases are defined as life-threatening or chronically debilitating diseases with a low prevalence (<5 in 10,000) (Moliner and Waligora, 2017). Between 263 and 446 million individuals are currently affected worldwide (Nguengang Wakap et al., 2020). The majority, ~80%, of rare disorders are of genetic origin. Rare genetic disorders are predominantly caused by rare variants in a single gene (Pogue et al., 2018). Identification of causal variants confirms the clinical diagnosis in patients with a suspected disorder, further allowing for suitable treatment options, early interventions, and genetic counseling of family members. The advent of next-generation sequencing technologies about a decade ago has transformed the diagnostic workflow by streamlining thousands of diagnostic assays into just a few. Rapidly decreasing costs, automation of high-throughput sequencing technologies, and advances in bioinformatic approaches facilitated the implementation of genome sequencing into routine diagnostics and a diagnosis can—in principle—now be made for nearly every patient with a genetic disorder (Sawyer et al., 2016; Stenton et al., 2020). Nevertheless, this is still not reached and the causative variant or associated gene cannot be determined in many patients by DNA sequencing due to limited knowledge about genotype–phenotype associations of rare variants.

By focusing on the ~2% coding regions of the human genome, the molecular diagnostic rate of whole exome sequencing (WES) for the detection of causal pathogenic variants in children with suspected genetic disease is about 35% (Clark et al., 2018). Extending the genetic analyses to the noncoding regions increases the diagnostic yield to 41% by whole genome sequencing (WGS) (Clark et al., 2018). As a result, the majority of patients with rare genetic disorders do not receive a genetic diagnosis and remain unsolved.

Assuming a high penetrance of genetic variants in Mendelian disorders, rare genetic disorders have to be caused by rare variants. With 40,000–200,000 rare variants (minor allele frequency <0.005) present in a typical human genome (The 1000 Genomes Project Consortium, 2015), they are largely abundant. However, many are private, and only for a few, the functional consequences are known, restricting prioritization of potentially pathogenic variants (Uricchio et al., 2016). Guidelines for the interpretation of sequence variants have been developed by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology [ACMG/AMP; (Richards et al., 2015)], integrating population, computational, functional, and segregation data. A variety of computational tools to predict the deleteriousness of a variant [CADD, MetaLR; (Dong et al., 2014; Kircher et al., 2014)], and the impact of a variant on protein function [PolyPhen-2, SIFT, and MutationAssessor; (Ng and Henikoff, 2001; Adzhubei et al., 2010; Reva et al., 2011)] and on splicing [spliceAI and MMsplice; (Cheng et al., 2019; Jaganathan et al., 2019)] has been developed. Functional validation can further confirm or disprove these predictions. Specifically, profiling the transcriptional level of a tissue at a defined time using RNA sequencing (RNAseq) can help to identify and prioritize pathogenic variants in three situations: 1) altered expression levels, 2) abnormal splicing events, and 3) detection of mono-allelic expression.

In 2016, the first studies systematically using RNAseq from muscle biopsies and fibroblasts cell lines to increase the diagnostic rate of rare disorders were released. They used arbitrary thresholds (Cummings et al., 2016, 2017) and statistical methods (Kremer et al., 2016, 2017) to call transcript aberrations. Since then, the bioinformatic approaches applied to RNAseq data have been advanced with machine learning methods such as OUTRIDER or FRASER, translating into more precise outlier calling (Brechtmann et al., 2018; Mertes et al., 2021). Machine learning methods are known to provide more robust predictions than statistical models (Bzdok et al., 2018). The diagnostic potential of complementing DNA sequencing with RNAseq for solving previously inconclusive WES cases is unequivocal. RNAseq-based diagnostics has led to a diagnostic yield of 7.5–18% in rare disease cohorts with no prior patient restrictions (Kremer et al., 2017; Frésard et al., 2019; Murdock et al., 2021; Yépez et al., 2021b) and of 35% in a cohort of patients including cases with predicted splice defects (Cummings et al., 2017; Gonorazky et al., 2019). In these studies, RNAseq led to the validation of novel pathogenic variants in known disease genes [e.g., chr21:47,409,881 C>T in the COL6A1 gene in Cummings et al. (2017)], but also to the discovery and validation of a new disease gene, TIMMDC1, where a deep intronic variant caused the activation of a cryptic splice site in two unrelated families (Kremer et al., 2017). RNAseq has been performed in the clinically accessible tissues whole blood, skeletal muscle, and skin-derived fibroblasts.

The implementation of multi-omics data in rare genetic disorders is calling for new methods of machine learning and statistical algorithms to remove sample covariation and detect expression or splicing outliers (**Figure 1**). In this review, we provide an overview of the methods used for the detection of aberrant expression, aberrant splicing, and mono-allelic expression in the context of rare disease diagnostics. In addition, we illustrate how these can improve the diagnostic yield of rare diseases.

ABERRANT EXPRESSION

Previous studies demonstrated the ability of genetic variation to influence gene expression (Montgomery et al., 2010; Hulse and Cai, 2013; Li et al., 2014; Zhao et al., 2016). Gene expression levels outside the physiological range, the so-called gene expression outliers, are associated with Mendelian and common disorders. In common disorders, there is the concept that the combination of many variants contributes to the disease risk. Those genetic risk factors also include variants causing aberrant expression and can be summarized in polygenic risk scores. On the other hand, Mendelian disorders are monogenic. Despite most Mendelian disorders being caused by variants in the protein-coding region of the DNA, aberrant expression events are, however, frequently caused by noncoding variants in regions such as enhancers, promoters, and suppressors, as well as by RNA degradation

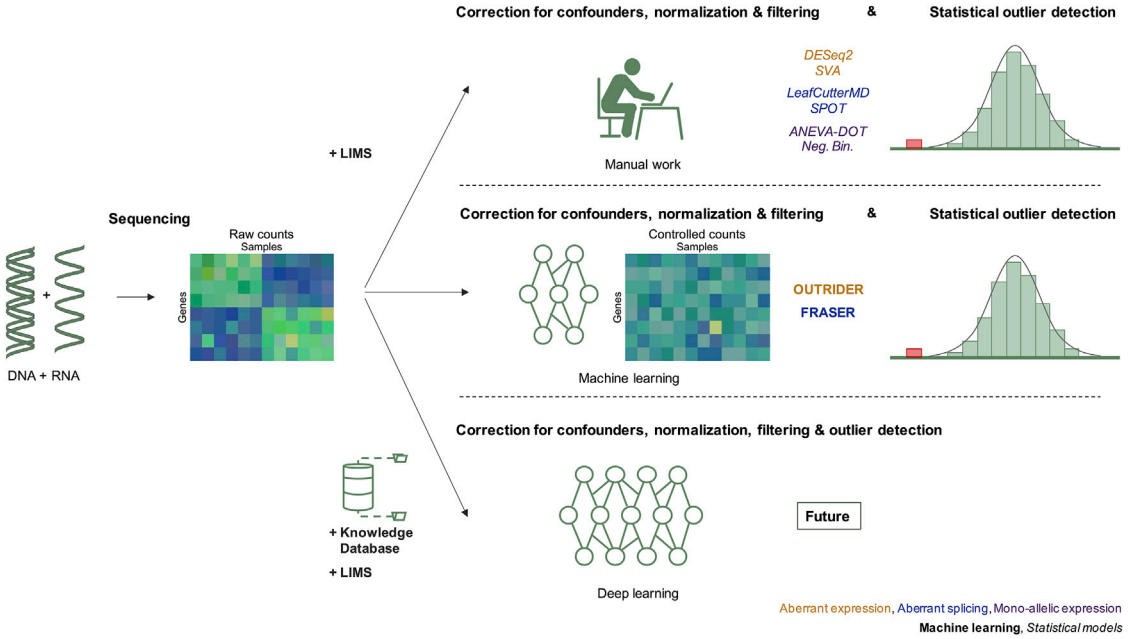

FIGURE 1 | Evolution of RNAseq data modeling in the diagnosis of rare genetic disorders. The integration of RNAseq in molecular diagnostics includes three steps, data generation by DNA and RNA sequencing resulting in raw counts, correction for confounders, and detection of outliers after statistical modeling. Before 2018, the correction for confounders was based on manual work considering known technical and biological variables recorded in the LIMS. Since 2018, methods applying machine learning have been introduced. We predict that in the future, deep learning approaches will integrate knowledge databases, sequencing data and combine normalization, and statistical outlier detection. Next to each diagram, the different methods used to detect RNAseq outliers of the corresponding cases are given. LIMS, laboratory information management system.

via nonsense-mediated decay (NMD) (Zeng et al., 2015; Li et al., 2017). Although gene expression altering variants can be detected by DNA sequencing, the functional consequences are difficult to predict. The quantification of RNA abundance and primary structure by RNAseq allow them to directly measure the consequences of genetic variants on gene expression.

Counts of reads aligned to genes are used as the basis for quantifying gene expression and detecting expression outliers (**Figure 2A**). These read counts should be normalized for sample sequencing depth by dividing the total depth or using size factors (Anders et al., 2013). Read counts can additionally be controlled for gene length resulting in the metrics reads per kilobase million (RPKM) (Mortazavi et al., 2008) or transcripts per million (TPM) (Wagner et al., 2012). Gene expression profiles are known to covariate due to both physiological regulation and technical artifacts. Therefore, several statistical methods, including principal component analysis (PCA) and surrogate variable analysis (SVA), have been adapted or specifically developed to correct for technical artifacts keeping biological-relevant signals. PCA has been used to cluster and reduce the dimension of gene expression matrices (Yeung and Ruzzo, 2001). Further, it has been shown that the top principal components can effectively explain the variation of gene expression (Ma and Dai, 2011; Todorov et al., 2018) and has been thus used to remove technical covariation. The SVA algorithm was introduced to capture gene expression heterogeneity thereby increasing biological accuracy and reproducibility of analyses in genome expression studies (Leek and Storey, 2007). Finally, using probabilistic estimation of expression residuals (PEER) on gene expression data outputs hidden factors explaining the expression variability (Stegle et al., 2012). All three methods are applied to remove unwanted covariation and improve gene expression data analysis.

The systematic implementation of gene expression analysis to detect potentially disease-associated genes causing aberrant expression in affected individuals has been successfully used as a complementary method in four studies by detecting expression outliers using distinct approaches (Kremer et al., 2017; Brechtmann et al., 2018; Frésard et al., 2019; Gonorazky et al., 2019). All approaches have been used in a gene-specific analysis, even if different isoforms of a gene are presented. One established way used to define an outlier data point is via its Z-score. Z-scores are defined as the difference between the observed value and the mean of the population, divided by the standard deviation of the population. Outliers are then commonly defined as observations with a |Z-score| greater than a cutoff, depending on the research field.

Gonorazky et al. (2019) applied this Z-score approach in a two-step procedure. First, genes with an RPKM |Z-score| $\geq$ 1.5 were defined as candidate outliers. Second, these candidates were compared to a control group and designated as outlier genes if their expression had at least a two-fold change with respect to the mean of the control group. Multiple testing was performed using Bonferroni's method, yielding a median of 17 aberrantly expressed neuromuscular-related genes per sample. This allowed the detection of six causal genes out of 25 samples in a cohort of rare muscular disorders. While this approach led to

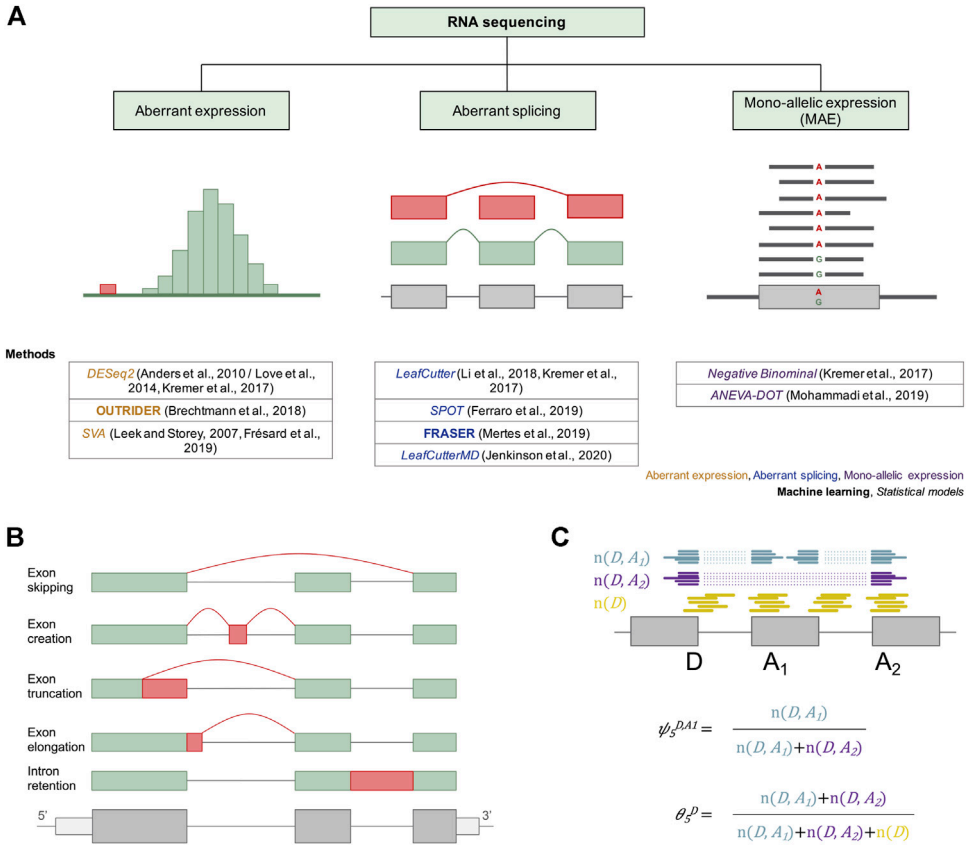

FIGURE 2 | Current approaches for the molecular diagnosis of rare genetic disorders *via* RNAseq. **(A)** Different machine learning and statistical approaches have been adapted or developed and applied to facilitate diagnosis *via* aberrant expression, aberrant splicing, and mono-allelic expression detection. For aberrant expression, the distribution of read counts is depicted, with outliers in red. For detection of aberrant splicing, an exon skipping event is depicted in red. For mono-allelic expression, a SNV with MAE of the alternative allele A is shown. Below each diagram the different methods used to detect RNAseq outliers of the corresponding cases are provided. **(B)** Different types of alternative splicing scenarios are depicted. Different mRNA isoforms can be generated from a single pre-mRNA through exon skipping, exon creation, exon truncation, exon elongation, and intron retention. Exons are represented as green and red boxes. Introns are illustrated as solid lines between the boxes. **(C)** A subset of a gene model is shown with split reads going from splice donor D to splice acceptors A_1 (blue) and A_2 (purple), and non-split reads spanning the exon–intron boundary at the splice donor D (yellow). These reads can then be converted into the percent spliced-in (ψ) and the splicing efficiency (θ) metrics using the formulae below the gene model.

some diagnoses, it lacks normalization for confounders and always requires the comparison with a control group from the same tissue as the affected samples. The Z-score approach has also been used in expression quantitative trait loci studies of common disorders (Smail et al., 2020).

Another way of defining outliers is by performing a statistical fit in the whole population and testing the distribution of the residuals (i.e., the difference between the real and the predicted value). This approach was used by Frésard et al. (2019) by applying a regression model on corrected TPMs using surrogate variables (SVs) and regression splines. In their cohort, the top two SVs significantly correlated with sequencing batch and sequencing facility. Adding the top two SVs and regression splines decreased the variation of the residuals, decreased the number of outlier genes per sample, and increased the coefficient of determination (R^2). Z-scores were generated from the residuals of this model, and values with a |Z-score| ≥ 2 were classified as outliers (**Figure 3A**), resulting in a median of 343 genes per sample. This very high number of outliers can be explained by the approach lacking multiple testing. Nevertheless, it allowed them to detect causal variants in four out of 80 samples in a cohort of various genetic disorders.

Kremer et al. (2017) adapted the method for differential expression DESeq2 (Love et al., 2014) and used it in a one vs. rest fashion. This approach models the counts using a negative binomial distribution parameterized with a mean and dispersion (**Figure 3B**). The mean was estimated taking into account size factors, batch, sex, and biopsy site of each sample, while the dispersion was specific for each gene. Gene-sample combinations with an adjusted *p*-value, using Hochberg's method, lower than 0.05 and |Z-score| > 3 were classified as outliers, deeming a median of one outlier per sample. Four out of 48 cases were diagnosed using this approach. In this case, the correction for sample covariation was performed using known factors; therefore, latent confounders were not taken into account.

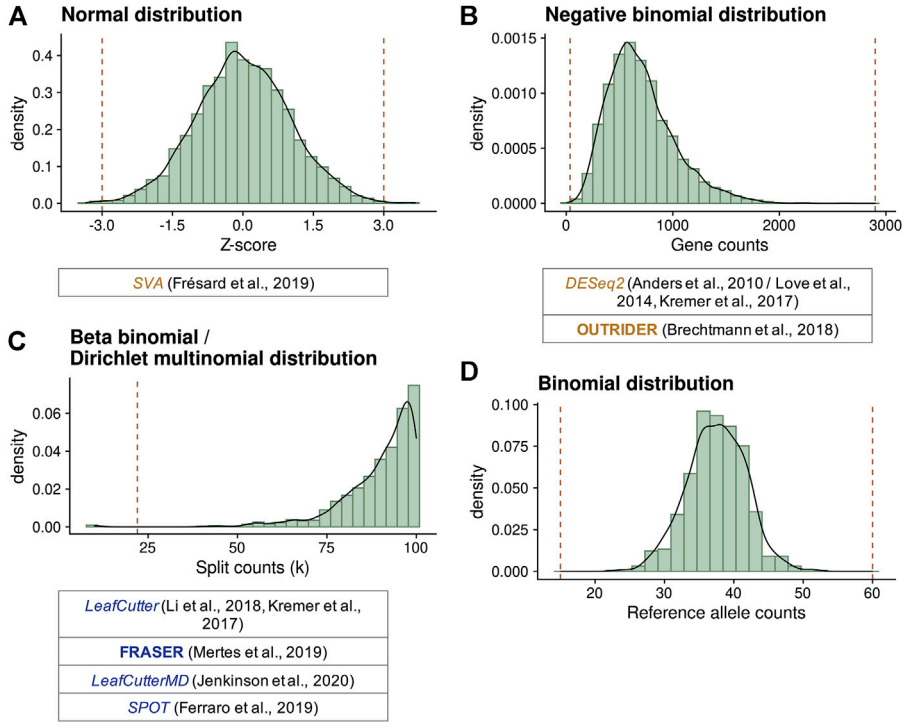

FIGURE 3 | Statistical distributions used to model gene expression data and compute outliers. **(A)** Standard normal distribution simulating standardized residual or Z-scores. Here, cutoffs defining outliers at ±3 (vertical lines) were selected. **(B)** Negative binomial distribution simulating gene counts with a mean of 700 and dispersion of 5. Values lower than 36 or greater than 2,899 (vertical lines) have a probability lower than 10^{-5} to occur, reflecting expression outliers. **(C)** Beta-binomial distribution simulating the distribution of split counts (n (D, A)) given a total of 100, with a ψ expectation of 0.9 and a correlation of 0.1. Values lower than 22 (vertical line) have a probability lower than 10^{-5} to occur, reflecting splicing outliers. Dirichlet-multinomial distributions are multidimensional extensions of beta-binomial distributions. **(D)** Binomial distribution simulating the counts of each allele on a selected variant with a total of 75 counts and a probability of 0.5. Biological-relevant cutoffs at 20 and 80% of allelic ratio (15 and 60 counts) indicate MAE. Extensions of this distribution have been proposed to account for the genetic variation by ANEVA-DOT and the MAE test of Kremer et al. (2017). Below each diagram, different methods using these distributions are given.

OUTRIDER was the first method applying a machine learning model to detect gene expression outliers (Brechtmann et al., 2018), instead of adaptations of methods used for the detection of differential expression. By using a denoising autoencoder, OUTRIDER controls for common covariation observed in gene expression. Autoencoders are machine learning models introduced to find low-dimensional representations of high-dimensional data (Hinton and Zemel, 1993). They achieve this by learning certain features from the data distribution by encoding it into a black-box representation and decoding it to denoised data (Hinton and Zemel, 1993). A subclass of autoencoders, called denoising autoencoders, is specialized to reconstruct corrupted high-dimensional data by exploiting correlations in the data (Vincent et al., 2008). OUTRIDER is taking advantage of this unsupervised learning method and applies it on gene-level counts (Brechtmann et al., 2018). Log-centered, size factor normalized gene counts are used as input for the autoencoder. The output is the parameters of the negative binomial distribution that the gene counts are assumed to follow, which are the expected counts and dispersions for each gene (**Figure 3B**). Multiple testing is performed on two-sided *p*-values using the Benjamini–Yekutieli false discovery rate (FDR) method, which holds under positive dependence because of gene co-expression. Expression outliers are defined as the gene-sample combinations with an FDR ≤0.05, resulting in a handful of outliers per sample, depending on the number of samples. In a subset of GEUVADI's cohort (Lappalainen et al., 2013) comprising 100 control samples from different sequencing centers (CNAG CRG, N = 31; ICMB, N = 28; and UNIGE, N = 41) and ancestries (British, N = 12; Finnish, N = 26; Tuscan, N = 22; Utah, N = 16; and Yoruba, N = 24), OUTRIDER's denoising autoencoder was able to control for covariation (Yépez et al., 2021b). The sample size of each subgroup was relatively similar; therefore, the autoencoder is yet to be tested in cases where a subgroup is substantially underrepresented. Murdock et al. (2021) applied OUTRIDER in a cohort of 78 DNA-unsolved patients with diverse disorders and diagnosed five of them with aberrantly expressed genes. OUTRIDER was also applied to a cohort of 303 rare disease patients sequenced in the same center but from different ancestries (mostly European and Asian), which led to the identification of 26 aberrantly expressed disease causal genes (Yépez et al., 2021a).

Currently, OUTRIDER is the only available automated method to compute expression outliers from gene expression matrices. It outperformed methods that use Z-scores on counts normalized using PEER and PCA in three different benchmarks.

First, it achieved a higher precision in recovering injected outliers in simulations. Second, OUTRIDER had a higher enrichment of rare moderate and high impact variants among outliers, as shown in GTEx samples that there is a strong association between rare variants and expression outliers (Li et al., 2017). Third, it was able to recover all the five pathogenic events from the Kremer et al. (2017) dataset, while PCA missed two and PEER missed one of the five pathogenic events. Larger datasets with more confirmed pathogenic variants would further improve benchmarking.

After obtaining an expression outlier in a disease-associated gene, it is mandatory for a molecular diagnosis to identify causative rare genetic variants. Filters including allele frequency, computational predictions, segregation, and incurrence in unaffected controls are applied. The variants can be located in regulatory regions like enhancers, promoters, or suppressors of the corresponding gene, but also in the coding or intronic regions affecting splicing or creating a nonsense codon causing NMD. In Yépez et al. (2021a), using WES, the cause of aberrant expression remained elusive in 65% of outliers and without the identification of a causative variant, the case remains undiagnosed. This fraction can be further reduced by using WGS, allowing the discovery of structural variants, which explain up to 25% of expression outliers (Ferraro et al., 2020). A significant fraction of expression outliers can be explained by NMD affecting a single allele only, showing the sensitivity of RNAseq studies (Li et al., 2017; Ferraro et al., 2020; Yépez et al., 2021a). In the majority of cases solved *via* aberrant expression, the causal variant affects splicing. Some of the variants have been prioritized after WES; however, they usually remain as variants of uncertain significance. A large fraction of variants are outside the splice region and frequently deep intronic. Without detection of aberrant splicing, they would not have been prioritized. In addition, in some cases, aberrant expression was caused by deletions in the 5'UTR (e.g., NM_004544.3 c.-99_-75del causing 50% depletion in *NDUFA10*) or promoter regions (e.g., NM_016617.2 c.-273_-271del causing 40% depletion in *UFM1*) (Yépez et al., 2021a).

SPLICING OUTLIERS

The concept of alternative splicing was initially introduced in 1978 based on the discrepancy between human protein-coding genes (~25,000) and human proteins (>90,000) (Gilbert, 1978). More than 95% of human genes undergo alternative splicing, acting to a certain extent in a tissue, or development-specific, or signal transduction–dependent manner (Pan et al., 2008; Nilsen and Graveley, 2010). Alternative splicing gives rise to different isoforms of the mature mRNA of a gene (Baralle and Giudice, 2017). Various forms of alternative splicing are known, including exon skipping, generation of new exons, exon truncation, exon elongation, and intron retention (**Figure 2B**). Alternative splicing is strictly regulated, and aberrant splicing is an underlying cause of genetic diseases (Wang et al., 2015). The splicing mechanism is complex and variable. Even without genetic variation, it is difficult to quantitatively predict splicing. Functional consequences of rare variants within splice regions are challenging to predict, especially in the case of deep intronic variants. Moreover, splice defects are quantitative, often resulting in multiple isoforms with different frequencies. Analysis of RNAseq using patients' samples enables the evaluation of splice variant consequences and detection of aberrant splice events, *de novo* and not predicted by DNA sequence.

To quantify splicing, reads spanning from the donor site of an exon to the acceptor site of another exon (split reads, n (D, A)) and reads overlapping an exon–intron boundary (non-split reads, n (D) for donor and n (A) for acceptor site) are counted and aggregated per junction (**Figure 2C**). These can then be converted into the intron-centric metrics percent-spliced-in (ψ) and splicing efficiency (θ) (Pervouchine et al., 2013). The ψ index is computed as the ratio between reads mapping to the given intron (n (D, A)) and all split reads sharing the same donor (ψ_5) or acceptor site (ψ_3), respectively. For the detection of partial or full intron retention, the splicing efficiency metric, defined as the ratio of all split reads and the full read coverage at a given splice site, is used (**Figure 2C**).

To detect aberrant splicing in patient samples suffering from rare genetic disorders, distinct, already available methodologies were applied to RNAseq data. One approach, used by Cummings et al. (2017) and Gonorazky et al. (2019), to detect aberrant splicing consists of comparing normalized split reads of affected individuals against those of controls and other affected samples. In both studies, normalized split reads were obtained by dividing them with the maximum number of split reads of a shared exon–intron junction. In Cummings et al. (2017), aberrant junctions were those whose normalized value was the highest in the sample of interest and twice or higher than the next highest. This approach resulted in a median of 190 aberrantly spliced genes per sample and allowed them to diagnose 10 out of 50 individuals with muscular disorders. In Gonorazky et al. (2019) for a junction to be aberrant, either the donor or the acceptor site cannot be annotated in GENCODE, must not be present more than five times in control samples (from GTEx), and it has to be unique among the affected cohort. On median, five aberrant junctions per sample were identified and led to the diagnosis of eight out of 25 samples with neuromuscular disorders. In a second approach for the detection of aberrant splicing, Kremer et al. (2017) adapted the method LeafCutter (Li et al., 2018), originally developed to test for differential usage in intron clusters between two groups, to work in a one vs. rest manner. LeafCutter's Dirichlet-multinomial's approach returns a *p*-value per intron per sample, later corrected for multiple testing using Hochberg's method. Outliers were defined as those with an adjusted *p*-value < 0.05, yielding a median of five outliers per sample. Although the method contributed to the diagnosis of three out of 48 cases, it does not control for sample covariation. However, Frésard et al. (2019) showed the existence of sample covariation on ψ values. Therefore, they applied a similar approach as they did for the aberrant expression analysis but regressed out principal components accounting for 95% of the variation instead of SVs and splines. The first three principal components are correlated with RIN number, batch, and sequencing facility. Splicing outliers were defined as those with a |Z-score| ≥2, yielding on average 540 outliers per sample. This

approach allowed them to diagnose two out of 80 patients. Frésard's methodology nevertheless is limited by not offering control for multiple testing and low power to detect aberrant splicing in splice sites with low reads.

Bearing in mind the limitations of the approaches described, three specialized methods to systematically detect aberrant splicing were developed: FRASER, LeafCutterMD, and SPOT. FRASER (Find Rare Splicing Events in RNAseq) is an approach combining machine learning and statistical models to detect aberrant splicing from RNAseq data (Mertes et al., 2021). Using the same rationale as OUTRIDER for detecting aberrant expression, FRASER uses a denoising autoencoder automatically controlling for latent confounders. Further, FRASER fits for each intron a beta-binomial distribution on the intron-centric metrics ψ_5, ψ_3, and θ, independently. The distribution is parameterized with a sample intron-specific proportion expectation and an intron-specific correlation (**Figure 3C**). *p*-values are computed and two multiple testing steps are performed, the first at the junction level using Holm's method (Holm, 1979), and the second at the gene level using Benjamini–Yekutieli's method (Benjamini and Yekutieli, 2001). Splicing outliers are defined as the intron-sample combinations with an FDR <0.10. A $|\Delta\ \psi| > 0.3$ is recommended as an additional filter for the identification of pathological-relevant variation, where it corresponds to the difference between the observed ψ and the expected ψ. Application of FRASER in the rare disease cohort from Kremer et al. (2017) identified all three previously detected pathogenic splicing aberrations, plus an intron-retention event missed by LeafCutter, and a synonymous variant causing a splice defect missed by Kremer et al. (Mertes et al., 2021). FRASER was also applied to the GEUVADIS multicenter and multi-ancestry cohort and was able to remove sample covariation for all metrics (Yépez et al., 2021b). In addition, FRASER has been used by Murdock et al. (2021) leading to the diagnosis of four out of 78 subjects and by Yépez et al. (2021a) leading to the diagnosis of 19 (12 in combination with aberrant expression) subjects from various ancestries.

LeafCutter for Mendelian disease (LeafCutterMD) was introduced in 2020 as an adaptation of LeafCutter (Li et al., 2018) to detect outlier splicing events (Jenkinson et al., 2020). Like its predecessor, LeafCutterMD uses an intron-based clustering approach, in which all split counts belonging to the same cluster are modeled together. Thus, it uses a Dirichlet-multinomial distribution, which is a generalized higher-dimensional version of the beta-binomial distribution (**Figure 3C**). Both distribution parameters allow it to account for biological variability and uncertainties due to statistical sampling. In simulations, the power of LeafCutterMD (1—the probability of a Type II error) is up to 50% higher than the power of LeafCutter (Jenkinson et al., 2020). Two-sided *p*-values are estimated for each intron of each cluster, later being corrected for multiple testing. In the same study, LeafCutterMD was applied to a cohort of 128 individuals with an undiagnosed genetic disease, out of which three were found to have splicing aberrations by manual inspection after variants were detected on them. LeafCutter failed to identify all three of them as outliers, while LeafCutterMD did identify them, demonstrating its improvement and application to rare disease cohorts. However, it does not account for latent confounders.

SPOT (Splicing Outlier deTection), the third novel approach developed to detect aberrant splicing, fits a Dirichlet-multinomial distribution on each of the intron clusters generated by LeafCutter (Ferraro et al., 2020), thus obtaining estimates of the distribution parameters for each cluster. Using these parameters, it generates 1 million random values and computes the Mahalanobis distance (Mahalanobis, 1930) of each of these values to the Dirichlet-multinomial distribution. This distance takes into account the covariance of the split reads. Empirical *p*-values are computed for each sample-intron cluster by comparing its Mahalanobis distance against the 1 million simulated ones. This method is yet to be tested in a rare disease cohort.

By inserting outliers in samples from skin and brain from GTEx, FRASER obtained higher precision in detecting the simulated outliers at all recall levels than SPOT and LeafCutterMD (Mertes et al., 2019). Following the principle that rare variants in the splice regions can disrupt splicing (Rivas et al., 2015), Mertes et al. (2021) also benchmarked by performing an enrichment of rare variants in the splice region among splicing outliers called by the three specialized methods. FRASER obtained a higher enrichment of rare splice-region variants and variants predicted to affect splicing [using MMsplice (Cheng et al., 2019)] than SPOT and LeafCutterMD across all tissues from the GTEx dataset (GTEx Consortium, 2017).

The variability of splicing is much higher than of gene expression and can often not be linked to a genetic variant in cis (Ferraro et al., 2020; Yépez et al., 2021a). Published pathogenic splice-disrupting variants discovered *via* RNAseq are usually within the annotated splice region and, as such, likely to cause aberrant splicing, or within the coding region or deep intronic regions activating novel splice sites. Only a minor fraction of these variants was functionally validated. Methods like MMSplice (Cheng et al., 2019) or SpliceAI (Jaganathan et al., 2019) can be further used to pinpoint which variant is most likely causing the splicing aberration.

MONO-ALLELIC EXPRESSION

Besides aberrant expression and splicing, RNAseq contains information about allelic expression, the expression of the maternal and paternal haplotype of an individual. In the case of allele-specific expression (mono-allelic expression, MAE) one allele is silenced and only the other allele is expressed. MAE is an extreme form of allelic imbalance. The reasons for MAE can be diverse and may be driven by loss-of-function genetic variants or epigenetic effects, such as imprinting of autosomal genes (Santoni et al., 2017) or inactivation of the X chromosome (Knight, 2004; Tukiainen et al., 2017). Assuming a recessive mode of inheritance, heterozygous variants are not considered to be disease causing if present alone (Albert and Kruglyak, 2015). The analysis of MAE can prioritize such rare heterozygous variants identified by WES (Kremer et al., 2017).

Detection of mono-allelically expressed genes relies on counting the expressed alleles at genomic positions of single-nucleotide heterozygous variants (SNVs, **Figure 2A**). Thereafter, it is tested whether these counts are evenly distributed among both alleles, or whether there is a pronounced skew toward one of the alleles (**Figure 3D**). These counts can be transformed into ratios by dividing the counts aligning to each allele over the total counts. Five methods have been developed to detect MAE in the context of rare genetic disorders.

Cummings et al. (2017) computed a 95% confidence interval of the mean allele balance for each gene using GTEx samples and compared the balance of the affected samples against it. Frésard et al. (2019) scaled reference allele ratios across samples thus converting them into Z-scores, ranked them, and further used them to support the findings from aberrant expression. Gonorazky et al. (2019) scaled alternative allele ratios into Z-scores and compared them against the values from GTEx samples which directed to the cause of disease in three cases.

Kremer et al. (2017) proposed a negative binomial test with a fixed dispersion for all genes. The negative binomial distribution accounts for a dispersion parameter not present in the binomial distribution. The test outputs a *p*-value for each SNV-sample combination which is later corrected for multiple testing. To distinguish between allelic imbalance and MAE, alleles with an alternative allele ratio greater than 0.8 or lower than 0.2 and Benjamini–Hochberg adjusted *p*-values < 0.05 were considered to be mono-allelically expressed.

Mohammadi et al. (2019) introduced ANEVA, a generative model to quantify genetic variation (V^G) in gene dosage (i.e., expression) within a population. Using V^G, they further developed ANEVA-DOT which implements a binomial-logit-normal test on each SNV to detect MAE. The rationale is that the variance of each gene is different and should be taken into account when testing for MAE. Moreover, it takes into account reference allele alignment bias and the probability of the allelic count to be incorrect (i.e., assignment to the reference allele when it corresponds to the alternative or vice versa). MAE variants are those with an FDR <0.05. ANEVA-DOT was applied to the cohort of 70 rare Mendelian muscle dystrophy and myopathy patients described in Cummings et al. (2017). In that cohort, 16 out of 70 patients had MAE pathogenic variants. ANEVA-DOT was able to recover all of them, plus it outperformed other tests that use binomial and beta-binomial distributions by obtaining the highest recall of the 16 true causal genes and the lowest number of reported outlier genes. Moreover, ANEVA-DOT detected a novel MAE in one proband from this cohort leading to a new diagnosis (Mohammadi et al., 2019). In order to estimate V^G, a large cohort with DNA and RNA sequence information is needed. Even using the large GTEx data collection (with a median of more than 200 samples per tissue), ANEVA estimates for V^G could only be computed for 4,962 genes (in median per tissue).

No formal benchmark has been done between the negative binomial method and ANEVA-DOT, and the application of ANEVA-DOT is currently limited by not providing genome-wide calling. Unlike for aberrant expression or splicing, one limitation of MAE is that it requires a detected variant to be able to compute the allelic counts. Obtaining MAE in genes constrained to variation, shorter genes, or ethnicities similar to the reference genome is, therefore, limited. So far, the added value of MAE has been lower with respect to that of aberrant expression or splicing and often in combination with aberrant expression (Kremer et al., 2017; Gonorazky et al., 2019; Yépez et al., 2021b).

LIMITATIONS AND FUTURE PERSPECTIVES OF RNASEQ IN THE MOLECULAR DIAGNOSIS OF RARE DISORDERS

The application of technologically and bioinformatically advanced next-generation sequencing methodologies has increased the number of patients with rare genetic disorders getting a clear molecular diagnosis. Nevertheless, a large fraction of those patients remains unsolved by DNA sequencing. Integration of RNAseq approaches appears promising to improve the diagnostic yield, yet several challenges remain to be addressed in the future.

Gene fusion, which consists of genetic material from different genes being merged and transcribed together, can be detected from RNAseq data and has proven to be successful in cancer diagnostics (Mertens et al., 2015; Dai et al., 2018). A variety of methods including Manta (Chen et al., 2016), ChimeraScan (Iyer et al., 2011), or STAR-Fusion (Haas et al., 2019) have been developed to call gene fusions. Statistical improvements, such as the inclusion of factors like allele frequency or repetitive matching to better distinguish between "bona fide gene fusions" from artifacts in CICERO will further enhance the quality of the fusion calls (Tian et al., 2020). Early studies have shown that calling gene fusions can also lead to diagnose rare diseases (Oliver et al., 2019), but its systematic application is yet to be explored.

Although all patients' cells share an almost identical genome, each cell type and subtype displays different levels of gene expression and gene isoforms. Moreover, it is assumed that not only different cell types exhibit various transcriptomes but also that the transcriptome of cells within a tissue varies (Shalek et al., 2014). Given the tissue-specific expression of genes and mRNA isoforms, analysis of disease-relevant tissues is for many diseases of great value for the interpretation of genetic variants (Wang et al., 2008; Melé et al., 2015; Cummings et al., 2017). However, obtaining biopsies of disease-relevant tissue is unfeasible for many rare genetic diseases. Blood, for instance, is the most easily accessible tissue, yet the blood transcriptome is not well suited for the analysis of a number of rare diseases (The GTEx Consortium, 2015; Cummings et al., 2017; Gonorazky et al., 2019). Fibroblast or myoblast cell lines based on skin or muscle biopsies are often used as surrogates. Patient cell lines are usually only available to a limited extent, and indefinite proliferation is not always possible. Consequently, transformed, immortalized cell lines, or animal models have been applied as disease models. Yet, neither model provides the possibility to fully replicate the physiology of patient cell

lines nor do they consider aspects of distinct gene expression and splicing in various subgroups of tissues (Anderson and Francis, 2018). By reprogramming mature patient cell lines into patient-specific induced pluripotent stem cells (iPSCs), the patient's genotype is retained. Patient-specific iPSCs, where as many as 27,046 protein-coding and nonprotein-coding genes are expressed, provide a suitable model to study the consequences of the disease genotype on RNA level (Hamazaki et al., 2017; Bonder et al., 2019). Due to the differentiation and self-renewal properties of iPSCs, their differentiation into disease-relevant tissues is an approach to overcome the limited accessibility of tissues. Generation and differentiation of iPSCs carrying disease-relevant mutations has already shown the ability to replicate the patients' phenotype. Simulating rare genetic diseases with the help of patient-specific iPSCs now further offers the possibility to analyze the transcriptome of the tissue of interest and thus to investigate new pathomechanisms (Sterneckert et al., 2014).

Some genetic variants affect the expression of a gene in a cell type–specific manner. Transcriptome analysis of an entire cell population, the so-called bulk RNAseq, is typically based on RNA extracted from tissues that may contain several cell types. These bulk RNAseq analyses provide a bulk average and are not cell type–specific. When analyzing a blood sample, the sensitivity for the detection of outliers depends on the contribution of the affected cell types in the blood sample, and the composition of different cell types may change over time. Whole transcriptome sequencing of single cells (scRNAseq) has the potential to identify and characterize rare cell populations (Hwang et al., 2018; Lederer and La Manno, 2020). ScRNAseq has generated comprehensive compendiums of cell types per tissue and identified cell types that play a critical role in genetic diseases, which can guide cell type–specific investigations (The Tabula Muris Consortium et al., 2018; Aizarani et al., 2019). In particular, scRNAseq has already been used as a powerful tool in fields such as immunology (Mizoguchi et al., 2018; Zhang, 2019), cancer (Cheng et al., 2021; Velten et al., 2021), and neurological diseases (Jäkel et al., 2019; Mathys et al., 2019). ScRNAseq of diseased tissues or patient-specific iPSCs could further advance the precision medicine approach in the field of rare genetic disorders (Choi et al., 2020).

Within the process of alternative splicing of human premessenger RNAs, multiple mRNA isoforms are generated from a single gene leading to differences in protein isoforms, structure, and function (Park et al., 2018). mRNA isoform-specific defects can result in different diseases. Standard and commonly used RNAseq platforms generate short reads of 150–300 bases, generally spanning at most two exon junctions per read. Short read RNAseq methodologies are constrained by their need for computational reconstruction of single short reads into entire transcripts (Steijger et al., 2013). Most of the human mRNAs, however, are longer than 3 kb (Piovesan et al., 2019) and with more than 100 kb Titin represents the longest disease-relevant human transcript (Bang et al., 2001). Hence, short read sequencing methods are not the best method of choice for detection and quantification of mRNA isoform expression and do not phase alternative exons. Recently developed long read sequencing technologies, producing reads of up to 100 kb, enable the accurate identification and quantification of mRNA transcript isoforms (Oikonomopoulos et al., 2020; Uapinyoying et al., 2020). Various bulk or single-cell long read RNAseq approaches revealed a significant number of missing mRNA isoforms in the current transcript annotations, suggesting a more complex mRNA isoform scene than previously assumed (Sharon et al., 2013; Tilgner et al., 2013; Anvar et al., 2018; Gupta et al., 2018; Glinos et al., 2021). Moreover, long reads improve sequence alignment, reduce the number of multi-mapped reads, allow phasing of variants, and increase the number of split-reads thereby improving the calling of aberrant splicing (Mantere et al., 2019; Amarasinghe et al., 2020; Mitsuhashi and Matsumoto, 2020). The power of long read transcriptome data in disease diagnostics has already been successfully demonstrated in diseases, such as Alzheimer's disease (De Roeck et al., 2017) or X-linked dystonia parkinsonism (Aneichyk et al., 2018). Methods to model and normalize gene expression might need to adapt as quantification of longer reads might statistically differ from that of short reads.

Transcriptome sequencing is used to assess the effects of variants on gene expression. Thresholds for pathological gene expression are not yet established, and aberrant RNA expression could be compensated on the protein level and not necessarily cause disease. Buffering mechanisms exist and transcript levels cannot be directly translated to protein levels (Battle et al., 2015). Moreover, low RNA expression levels not called as a statistical outlier may be pathologically relevant. Therefore, for the clinical interpretation according to the ACMG/AMP criteria, additional functional studies, such as proteomics and functional assays, are often needed for the validation of aberrant RNA expression (Richards et al., 2015). With an increasing number of patients with aberrant expression, thresholds of aberrant gene expression can be established. Proteomic approaches have advanced, now allowing protein quantification, study of protein–protein interactions, and identification of posttranslational modifications by high-throughput proteomics (Wang et al., 2014; Stenton and Prokisch, 2020). For example, integrated genome, transcriptome, and proteome analyses allowed the validation of a rare variant causing aberrant gene expression of genes such as TIMMDC1 (Kremer et al., 2017), *PTCD3* (Borna et al., 2019), or *MRPS34* (Lake et al., 2018). More recently, Kopajtich et al. (2021) demonstrated the effectiveness of integrating genomics, transcriptomics, and proteomics in a systematic diagnostic application to discover the genetic cause of 20% of unsolved patients with suspected mitochondrial disorders. These recent advancements in bioinformatic approaches help to combine these multi-omics data and enable their holistic analysis.

Due to the expanding availability and automation of DNA and RNA sequencing, they are increasingly applied in diagnostics. While the demand is decreasing on the wet lab side, the increasing throughput of such methods and generation of large data sets requires larger computational infrastructure, adaptation, and automation of bioinformatic algorithms. Depending on the algorithms used, memory consumption and computation time may increase exponentially, logarithmically, or otherwise

depending on the size of the dataset to be analyzed (Baichoo and Ouzounis, 2017). The machine learning approaches OUTRIDER and FRASER led to a shorter computational time with respect to DESeq and LeafCutter and therefore speed up analysis. Moreover, both tools have been integrated in a computational workflow, DROP, which also automates the preprocessing and quality control steps (Yépez et al., 2021b). Besides seeking to obtain high precision, newly developed methods must also pursue yielding results in a low computational time and being able to model hundreds or thousands of samples together.

Another challenge remaining in the analysis of RNAseq data is the replication of outliers. There are no studies about replication of outliers using multiple biopsies from the same tissue from the same patient. However, Ferraro et al. (2020) and Mertes et al. (2021) performed replication analyses using different tissues from the same individuals from the GTEx cohort. Ferraro et al. (2020) found that a median of 5.1% expression outliers, 8.7% of splicing outliers (using SPOT), and 10.7% of mono-allelically expressed genes (using ANEVA-DOT) that were detected as aberrant in one tissue were replicated in another. This was confirmed by Mertes et al. (2021) who found that more than 80% of splicing outliers are found in only one tissue, for SPOT, LeafCutterMD, and FRASER. These low replication numbers likely reflect the differences between the tissues but could also indicate variation within cells of the same tissue. Pooled CRISPR screens combined with scRNAseq have emerged as powerful tools for profiling the functional effects of genetic variants at the single-cell level (Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016; Datlinger et al., 2017). Using different cell types in parallel, those approaches have the potential to study the tissue specificity of splice variants and replication across tissues.

To properly normalize and model read counts and split reads, a minimum number of samples are needed. OUTRIDER and FRASER recommend at least 60 and 30 samples, respectively. Integrating affected with external control samples can help reach this minimum, as long as they were originated from the same tissue and sequenced using a similar protocol (Frésard et al., 2019; Yépez et al., 2021b). When integrating samples, it is recommended to first inspect the plots of the normalized counts (e.g., *via* heatmaps or principal components), especially if the affected samples are too diverse with respect to the controls (e.g., ancestry, age, or disease). Gene expression could depend on developmental status. Therefore, it is recommended to consider adequate control samples. For example, for pediatric cases, the CZI pediatric cell atlas or developmental GTEx can serve as appropriate control datasets (Taylor et al., 2019). Gene coverage also plays a role in outlier detection. Yépez et al. (2021b) showed that less expression and splicing outliers were detected in samples with lower sequencing depth (~30 vs. ~85 million reads). A systematic study to find the minimum coverage a gene and a junction should have in order to be detected as an outlier, and whether genes that are very highly expressed tend to be more prone to be called as expression or splicing outliers, is pending. Likewise, the sequencing protocol might influence the detected outliers. The total RNAseq protocol contains more immature splice transcripts in comparison to poly-A enriched. Those immature splice transcripts increase the noise and can be misinterpreted as aberrant splicing. On the other hand, using a poly-A enriched protocol might have an impact on detecting aberrant splicing on the 5' end of larger genes. In a poly-A enriched cohort, Yépez et al. (2021a) showed that genes with many exons tend to have more splicing outliers than genes with fewer exons, even after correcting for multiple testing inside each gene, but there was no information about the position of the aberrant junctions inside each gene. An analysis of a dataset with samples from the same donors sequenced using both protocols [such as the one in Chen et al. (2020)] will be valuable.

Finally, the approaches introduced here used different thresholds and cutoffs to define outliers and to filter out genes or junctions with low expression. Fine-tuning thresholds and cutoffs to obtain the desired precision-recall balance is of utmost importance and need to be adopted to the question. Thresholds can be defined by biological and statistical significance. A reduction to less than half of gene expression with respect to the median could reflect a pathological situation specifically in genes prioritized by the finding of rare DNA variants. If the data are explored to discover aberrant expression, considering the high number of tests (in the 10,000 s for gene expression, 100,000 s for aberrant splicing, and in between for MAE), multiple testing correction is necessary. These tests are not independent due to gene co-expression, split counts on a same cluster being coupled, and allelic expression being the same among all SNVs of a gene. Many multiple testing methods have been implemented, yet no agreement on a common guideline has been reached.

CONCLUSION

In diagnostics of rare disorders, it is important to be able to evaluate the functional consequences of genetic variants. A correct molecular diagnosis enables to study the natural history and pathomechanisms of the disease which may lead to targeted therapy. The current diagnostic gap in DNA sequencing can in most cases be traced back to difficulties in variant prioritization. RNAseq has now been shown to be effective in increasing the diagnostic yield. Numerous diagnostic laboratories have already implemented DNA sequencing technologies and the establishment of RNAseq protocols in these laboratories would be easy and straightforward. Machine learning approaches have automated normalization and denoising of confounded RNAseq data, providing gene expression data in a high-throughput manner ready for analysis. Statistical methods have been adopted to the analysis of aberrant expression on the quantitative and qualitative level. However, pathological variants and statistical significance may need different thresholds. Statistical models optimized to control for false positive hits may be too stringent and miss pathological

events. The limited number of positive controls for pathological aberrant expression results in an unsecure validation of established methods for diagnostics. With increasing datasets, this shortcoming has to be addressed in the near future and will force the establishment of guidelines. The application of machine learning is only in its beginning, and we foresee that deep learning methods will further improve the diagnostics of rare disorders.

AUTHOR CONTRIBUTIONS

This article was written by LS and VY under the guidance of HP.

FUNDING

This work was supported by the BMBF (German Federal Ministry of Education and Research) through mitoNET German Network for Mitochondrial Diseases (grant number 01GM1906B (to HP and LS)), PerMiM Personalized Mitochondrial Medicine (grant number 01KU2016A (to HP)), and the Medical Informatics Initiative CORD-MI (Collaboration on Rare Diseases (to VY.)). The Bavarian State Ministry of Health and Care funded this work within its framework of DigiMed Bayern (grant number DMB-1805-0002 (to LS and HP)).

REFERENCES

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuãez, J. K., Chen, Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* 167, 1867–1882.e21. doi:10.1016/j.cell.2016.11.048

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

Aizarani, N., Saviano, A., SagarMailly, L., Durand, S., Herman, J. S., et al. (2019). A Human Liver Cell Atlas Reveals Heterogeneity and Epithelial Progenitors. *Nature* 572, 199–204. doi:10.1038/s41586-019-1373-2

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biol.* 21, 30. doi:10.1038/nmeth.2714

Albert, F. W., and Kruglyak, L. (2015). The Role of Regulatory Variation in Complex Traits and Disease. *Nat. Rev. Genet.* 16, 197–212. doi:10.1038/nrg3891

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., et al. (2013). Count-based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor. *Nat. Protoc.* 8, 1765–1786. doi:10.1038/nprot.2013.099

Anderson, R. H., and Francis, K. R. (2018). Modeling Rare Diseases with Induced Pluripotent Stem Cell Technology. *Mol. Cel. Probes* 40, 52–59. doi:10.1016/j.mcp.2018.01.001

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., et al. (2018). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* 172, 897–909.e21. doi:10.1016/j.cell.2018.02.011

Anvar, S. Y., Allard, G., Tseng, E., Sheynkman, G. M., de Klerk, E., Vermaat, M., et al. (2018). Full-length mRNA Sequencing Uncovers a Widespread Coupling between Transcription Initiation and mRNA Processing. *Genome Biol.* 19, 46. doi:10.1186/s13059-018-1418-0

Baichoo, S., and Ouzounis, C. A. (2017). Computational Complexity of Algorithms for Sequence Comparison, Short-Read Assembly and Genome Alignment. *Biosystems* 156–157, 72–85. doi:10.1016/j.biosystems.2017.03.003

Bang, M.-L., Centner, T., Fornoff, F., Geach, A. J., Gotthardt, M., McNabb, M., et al. (2001). The Complete Gene Sequence of Titin, Expression of an Unusual ≈700-kDa Titin Isoform, and its Interaction with Obscurin Identify a Novel Z-Line to I-Band Linking System. *Circ. Res.* 89, 1065–1072. doi:10.1161/hh2301.100981

Baralle, F. E., and Giudice, J. (2017). Alternative Splicing as a Regulator of Development and Tissue Identity. *Nat. Rev. Mol. Cel Biol.* 18, 437–451. doi:10.1038/nrm.2017.27

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., et al. (2015). Impact of Regulatory Variation from RNA to Protein. *Science* 6, 664–667. doi:10.1126/science.1260793

Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* 29, 24. doi:10.1214/aos/1013699998

Bonder, M. J., Smail, C., Gloudemans, M. J., Frésard, L., Jakubosky, D., D'Antonio, M., et al. (2019). Systematic Assessment of Regulatory Effects of Human Disease Variants in Pluripotent Cells. *bioRxiv*, 784967. doi:10.1101/784967

Borna, N. N., Kishita, Y., Kohda, M., Lim, S. C., Shimura, M., Wu, Y., et al. (2019). Mitochondrial Ribosomal Protein PTCD3 Mutations Cause Oxidative Phosphorylation Defects with Leigh Syndrome. *Neurogenetics* 20, 17. doi:10.1007/s10048-018-0561-9

Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., et al. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* 103, 907–917. doi:10.1016/j.ajhg.2018.10.025

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus Machine Learning. *Nat. Methods* 15, 233–234. doi:10.1038/nmeth.4642

PCAWG Transcriptome Core Group Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., et al. (2020). Genomic Basis for RNA Alterations in Cancer. *Nature* 578, 129–136. doi:10.1038/s41586-020-1970-0

Chen, L., Yang, R., Kwan, T., Tang, C., Watt, S., Zhang, Y., et al. (2020). Paired rRNA-Depleted and polyA-Selected RNA Sequencing Data and Supporting Multi-Omics Data from Human T Cells. *Sci. Data* 7, 376. doi:10.1038/s41597-020-00719-4

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications. *Bioinformatics* 32, 1220–1222. doi:10.1093/bioinformatics/btv710

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., et al. (2019). MMSplice: Modular Modeling Improves the Predictions of Genetic Variant Effects on Splicing. *Genome Biol.* 20, 48. doi:10.1186/s13059-019-1653-z

Cheng, S., Li, Z., Gao, R., Xing, B., Gao, Y., Yang, Y., et al. (2021). A Pan-Cancer Single-Cell Transcriptional Atlas of Tumor Infiltrating Myeloid Cells. *Cell* 184, 792–809. doi:10.1016/j.cell.2021.01.010

Choi, J. R., Yong, K. W., Choi, J. Y., and Cowie, A. C. (2020). Single-Cell RNA Sequencing and its Combination with Protein and DNA Analyses. *Cells* 9, 1130. doi:10.3390/cells9051130

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., et al. (2018). Meta-analysis of the Diagnostic and Clinical Utility of Genome and Exome Sequencing and Chromosomal Microarray in Children with Suspected Genetic Diseases. *Npj Genomic Med.* 3, 16. doi:10.1038/s41525-018-0053-8

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., et al. (2017). Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing. *Sci. Transl. Med.* 9, eaal5209. doi:10.1126/scitranslmed.aal5209

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, R., et al. (2016). Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing. *bioRxiv*, 074153. doi:10.1101/074153

Dai, X., Theobard, R., Cheng, H., Xing, M., and Zhang, J. (2018). Fusion Genes: A Promising Tool Combating against Cancer. *Biochim. Biophys. Acta BBA - Rev. Cancer* 1869, 149–160. doi:10.1016/j.bbcan.2017.12.003

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., et al. (2017). Pooled CRISPR Screening with Single-Cell Transcriptome Readout. *Nat. Methods* 14, 297-301. doi:10.1038/nmeth.4177

De Roeck, A., Van den Bossche, T., van der Zee, J., Verheijen, J., De Coster, W., Van Dongen, J., et al. (2017). Deleterious ABCA7 Mutations and Transcript rescue Mechanisms in Early Onset Alzheimer's Disease. *Acta Neuropathol. (Berl.)* 134, 475–487. doi:10.1007/s00401-017-1714-x

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853-1866.e17. doi:10.1016/j.cell.2016.11.038

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2014). Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Hum. Mol. Genet.* 24, 2125–2137. doi:10.1093/hmg/ddu733

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., et al. (2020). Transcriptomic Signatures across Human Tissues Identify Functional Rare Genetic Variation. *Science* 369, eaaz5900. doi:10.1126/science.aaz5900

Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., et al. (2019). Identification of Rare-Disease Genes Using Blood Transcriptome Sequencing and Large Control Cohorts. *Nat. Med.* 25, 911–919. doi:10.1038/s41591-019-0457-8

Gilbert, W. (1978). Why Genes in Pieces? *Nature* 271, 501. doi:10.1038/271501a0

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., et al. (2021). Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing. *bioRxiv*, 427687. doi:10.1101/2021.01.22.427687

Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* 104, 466–483. doi:10.1016/j.ajhg.2019.01.012

GTEx Consortium (2017). Genetic Effects on Gene Expression across Human Tissues. *Nature* 550, 204–213. doi:10.1038/nature24277

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., et al. (2018). Single-cell Isoform RNA Sequencing Characterizes Isoforms in Thousands of Cerebellar Cells. *Nat. Biotechnol.* 36, 1197–1202. doi:10.1038/nbt.4259

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy Assessment of Fusion Transcript Detection via Read-Mapping and De Novo Fusion Transcript Assembly-Based Methods. *Genome Biol.* 20, 213. doi:10.1186/s13059-019-1842-9

Hamazaki, T., El Rouby, N., Fredette, N. C., Santostefano, K. E., and Terada, N. (2017). Concise Review: Induced Pluripotent Stem Cell Research in the Era of Precision Medicine. *Stem Cell Dayt. Ohio* 35, 545–550. doi:10.1002/stem.2570

Hinton, G. E., and Zemel, R. S. (1993). Autoencoders, Minimum Description Length and Helmholtz Free Energy. *Proc. 6th Int. Conf. Neural Inf. Processing*, 3–10. doi:10.5555/2987189.2987190

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* 6, 65–70.

Hulse, A. M., and Cai, J. J. (2013). Genetic Variants Contribute to Gene Expression Variability in Humans. *Genetics* 193, 95–108. doi:10.1534/genetics.112.146779

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA Sequencing Technologies and Bioinformatics Pipelines. *Exp. Mol. Med.* 50, 96. doi:10.1038/s12276-018-0071-8

Iyer, M. K., Chinnaiyan, A. M., and Maher, C. A. (2011). ChimeraScan: a Tool for Identifying Chimeric Transcription in Sequencing Data. *Bioinformatics* 27, 2903–2904. doi:10.1093/bioinformatics/btr467

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535–548. doi:10.1016/j.cell.2018.12.015

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., et al.(2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167, 1883–1896.e15. doi:10.1016/j.cell.2016.11.039

Jäkel, S., Agirre, E., Mendanha Falcão, A., van Bruggen, D., Lee, K. W., Knuesel, I., et al. (2019). Altered Human Oligodendrocyte Heterogeneity in Multiple Sclerosis. *Nature* 566, 543–547. doi:10.1038/s41586-019-0903-2

Jenkinson, G., Li, Y. I., Basu, S., Cousin, M. A., Oliver, G. R., and Klee, E. W. (2020). LeafCutterMD: an Algorithm for Outlier Splicing Detection in Rare Diseases. *Bioinformatics* 36, 4605–4615. doi:10.1093/bioinformatics/btaa259

Karollus, A., Avsec, Ž., and Gagneur, J. (2020). Predicting Mean Ribosome Load for 5'UTR of any length using deep learning. *bioRxiv*. doi:10.1101/2020.06.15.152728

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 35, 125–131. doi:10.1093/nar/gkl924

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* 46, 310–315. doi:10.1038/ng.2892

Knight, J. C. (2004). Allele-specific Gene Expression Uncovered. *Trends Genet.* 20, 113–116. doi:10.1016/j.tig.2004.01.001

Kopajtich, R., Smirnov, D., Stenton, S. L., Loipfinger, S., Meng, C., Scheller, I. F., et al. (2021). Integration of Proteomics with Genomics and Transcriptomics Increases the Diagnostic Rate of Mendelian Disorders. *medRxiv*, 21253187. doi:10.1101/2021.03.09.21253187

Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., et al. (2016). Genetic Diagnosis of Mendelian Disorders via RNA Sequencing. *bioRxiv*, 066738. doi:10.1101/066738

Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., et al. (2017). Genetic Diagnosis of Mendelian Disorders via RNA Sequencing. *Nat. Commun.* 8, 15824. doi:10.1038/ncomms15824

Lake, N. J., Webb, B. D., Stroud, D. A., Richman, T. R., Ruzzenente, B., Compton, A. G., et al. (2018). Biallelic Mutations in MRPS34 Lead to Instability of the Small Mitoribosomal Subunit and Leigh Syndrome. *Am. J. Hum. Genet.* 102, 713. doi:10.1016/j.ajhg.2018.03.015

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans. *Nature* 501, 506–511. doi:10.1038/nature12531

Lederer, A. R., and La Manno, G. (2020). The Emergence and Promise of Single-Cell Temporal-Omics Approaches. *Curr. Opin. Biotechnol.* 63, 70–78. doi:10.1016/j.copbio.2019.12.005

Leek, J. T., and Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *Plos Genet.* 3, 12. doi:10.1371/journal.pgen.0030161

Li, X., Battle, A., Karczewski, K. J., Zappala, Z., Knowles, D. A., Smith, K. S., et al. (2014). Transcriptome Sequencing of a Large Human Family Identifies the Impact of Rare Noncoding Variants. *Am. J. Hum. Genet.* 95, 245–256. doi:10.1016/j.ajhg.2014.08.004

Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., et al. (2017). The Impact of Rare Variation on Gene Expression across Tissues. *Nature* 550, 239–243. doi:10.1038/nature24267

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., et al. (2018). Annotation-free Quantification of RNA Splicing Using LeafCutter. *Nat. Genet.* 50, 151–158. doi:10.1038/s41588-017-0004-9

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Ma, S., and Dai, Y. (2011). Principal Component Analysis Based Methods in Bioinformatics Studies. *Brief. Bioinform.* 12, 714–722. doi:10.1093/bib/bbq090

Mahalanobis, P. C. (1930). On Tests and Measures of Group Divergence. *J. Asiat. Soc. Bengal* 26, 541–588.

Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* 10, 426. doi:10.3389/fgene.2019.00426

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., et al. (2019). Single-cell Transcriptomic Analysis of Alzheimer's Disease. *Nature* 570, 332–337. doi:10.1038/s41586-019-1195-2

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human Genomics. The Human Transcriptome across Tissues and Individuals. *Science* 348, 660–665. doi:10.1126/science.aaa0355

Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The Emerging Complexity of Gene Fusions in Cancer. *Nat. Rev. Cancer* 15, 371–381. doi:10.1038/nrc3947

Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., et al. (2021). Detection of Aberrant Splicing Events in RNA-Seq Data Using FRASER. *Nat. Commun.* 12, 529. doi:10.1038/s41467-020-20573-7

Mertes, C., Scheller, I., Yepez, V. A., Celik, M. H., Liang, Y., Kremer, L. S., et al. (2019). Detection of Aberrant Splicing Events in RNA-Seq Data with FRASER. *bioRxiv*, 866830. doi:10.1101/2019.12.18.866830

Mitsuhashi, S., and Matsumoto, N. (2020). Long-Read Sequencing for Rare Human Genetic Diseases. *J. Hum. Genet.* 65 (1), 11-19. doi:10.1038/s10038-019-0671-8

Mizoguchi, F., Slowikowski, K., Wei, K., Marshall, J. L., Rao, D. A., Chang, S. K., et al. (2018). Functionally Distinct Disease-Associated Fibroblast Subsets in Rheumatoid Arthritis. *Nat. Commun.* 9, 789. doi:10.1038/s41467-018-02892-y

Mohammadi, P., Castel, S. E., Cummings, B. B., Einson, J., Sousa, C., Hoffman, P., et al. (2019). Genetic Regulatory Variation in Populations Informs Transcriptome Analysis in Rare Disease. *Science* 366, 351–356. doi:10.1126/science.aay0256

Moliner, A. M., and Waligora, J. (2017). The European Union Policy in the Field of Rare Diseases. *Adv. Exp. Med. Biol.* 1031, 561–587. doi:10.1007/978-3-319-67144-4_30

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbettf, J., et al. (2010). Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population. *Nature* 464, 773–777. doi:10.1038/nature08903

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi:10.1038/nmeth.1226

Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., et al. (2021). Transcriptome-directed Analysis for Mendelian Disease Diagnosis Overcomes Limitations of Conventional Genomic Testing. *J. Clin. Invest.* 131, e141500. doi:10.1172/JCI141500

Ng, P. C., and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Res.* 11, 863–874. doi:10.1101/gr.176601

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., et al. (2020). Estimating Cumulative point Prevalence of Rare Diseases: Analysis of the Orphanet Database. *Eur. J. Hum. Genet. EJHG* 28, 165–173. doi:10.1038/s41431-019-0508-0

Nilsen, T. W., and Graveley, B. R. (2010). Expansion of the Eukaryotic Proteome by Alternative Splicing. *Nature* 463, 457–463. doi:10.1038/nature08909

Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P., and Ragoussis, J. (2020). Methodologies for Transcript Profiling Using Long-Read Technologies. *Front. Genet.* 11, 606. doi:10.3389/fgene.2020.00606

Oliver, G. R., Tang, X., Schultz-Rogers, L. E., Vidal-Folch, N., Jenkinson, W. G., Schwab, T. L., et al. (2019). A Tailored Approach to Fusion Transcript Identification Increases Diagnosis of Rare Inherited Disease. *PLOS ONE* 14, e0223337. doi:10.1371/journal.pone.0223337

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing. *Nat. Genet.* 40, 1413–1415. doi:10.1038/ng.259

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* 102, 11–26. doi:10.1016/j.ajhg.2017.11.002

Pawlikowska, I., Wu, G., Edmonson, M., Liu, Z., Gruber, T., Zhang, J., et al. (2014). The Most Informative Spacing Test Effectively Discovers Biologically Relevant Outliers or Multiple Modes in Expression. *Bioinformatics* 30, 1400–1408. doi:10.1093/bioinformatics/btu039

Pervouchine, D. D., Knowles, D. G., and Guigo, R. (2013). Intron-centric Estimation of Alternative Splicing from RNA-Seq Data. *Bioinformatics* 29, 273–274. doi:10.1093/bioinformatics/bts678

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., and Caracausi, M. (2019). Human Protein-Coding Genes and Gene Feature Statistics in 2019. *BMC Res. Notes* 12, 315. doi:10.1186/s13104-019-4343-8

Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V., Aguiar, L. R., Carvalho, J. L. de., et al. (2018). Rare Genetic Diseases: Update on Diagnosis, Treatment and Online Resources. *Drug Discov. Today* 23, 187–195. doi:10.1016/j.drudis.2017.11.002

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-Indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. doi:10.1038/nbt.4235

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res.* 39, e118. doi:10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–423. doi:10.1038/gim.2015.30

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., et al. (2015). Effect of Predicted Protein-Truncating Genetic Variants on the Human Transcriptome. *Science* 348, 666–669. doi:10.1126/science.1261877

Santoni, F. A., Stamoulis, G., Garieri, M., Falconnet, E., Ribaux, P., Borel, C., et al. (2017). Detection of Imprinted Genes by Single-Cell Allele-specific Gene Expression. *Am. J. Hum. Genet.* 100, 444–453. doi:10.1016/j.ajhg.2017.01.028

Sawyer, S. L., Hartley, T., Dyment, D. A., Beaulieu, C. L., Schwartzentruber, J., Smith, A., et al. (2016). Utility of Whole-exome Sequencing for Those Near the End of the Diagnostic Odyssey: Time to Address Gaps in Care. *Clin. Genet.* 89, 275–284. doi:10.1111/cge.12654

Scotti, M. M., and Swanson, M. S. (2016). RNA Mis-Splicing in Disease. *Nat. Rev. Genet.* 17, 19–32. doi:10.1038/nrg.2015.3

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., et al. (2014). Single-cell RNA-Seq Reveals Dynamic Paracrine Control of Cellular Variation. *Nature* 510, 363–369. doi:10.1038/nature13437

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A Single-Molecule Long-Read Survey of the Human Transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi:10.1038/nbt.2705

Singh, R. K., and Cooper, T. A. (2012). Pre-mRNA Splicing in Disease and Therapeutics. *Trends Mol. Med.* 18, 472–482. doi:10.1016/j.molmed.2012.06.006

Smail, C., Ferraro, N. M., Durrant, M. G., Rao, A. S., Aguirre, M., Li, X., et al. (2020). Integration of rare large-effect expression variants improves polygenic risk prediction. *medRxiv*, 2020.12.02.20242990. doi:10.1101/2020.12.02.20242990

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nat. Protoc.* 7, 500–507. doi:10.1038/nprot.2011.457

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., et al. (2013). Assessment of Transcript Reconstruction Methods for RNA-Seq. *Nat. Methods* 10, 1177-1184. doi:10.1038/nmeth.2714

Stenton, S. L., Kremer, L. S., Kopajtich, R., Ludwig, C., and Prokisch, H. (2020). The Diagnosis of Inborn Errors of Metabolism by an Integrative "Multi-omics" Approach: A Perspective Encompassing Genomics, Transcriptomics, and Proteomics. *J. Inherit. Metab. Dis.* 43, 25–35. doi:10.1002/jimd.12130

Stenton, S. L., and Prokisch, H. (2020). Genetics of Mitochondrial Diseases: Identifying Mutations to Help Diagnosis. *EBioMedicine* 56, 102784. doi:10.1016/j.ebiom.2020.102784

Sterneckert, J. L., Reinhardt, P., and Schöler, H. R. (2014). Investigating Human Disease Using Stem Cell Models. *Nat. Rev. Genet.* 15, 625–639. doi:10.1038/nrg3764

Taylor, D. M., Aronow, B. J., Tan, K., Bernt, K., Salomonis, N., Greene, C. S., et al. (2019). The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution. *Dev. Cel* 49, 10–29. doi:10.1016/j.devcel.2019.03.001

Tazi, J., Bakkour, N., and Stamm, S. (2009). Alternative Splicing and Disease. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1792, 14–26. doi:10.1016/j.bbadis.2008.09.017

The 1000 Genomes Project Consortium (2015). A Global Reference for Human Genetic Variation. *Nature* 526, 68–74. doi:10.1038/nature15393

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* 348, 648–660. doi:10.1126/science.1262110

The Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis, et al. (2018). Single-cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris. *Nature* 562, 367–372. doi:10.1038/s41586-018-0590-4

Tian, L., Li, Y., Edmonson, M. N., Zhou, X., Newman, S., McLeod, C., et al. (2020). CICERO: a Versatile Method for Detecting Complex and Diverse Driver Fusions Using Cancer RNA Sequencing Data. *Genome Biol.* 21, 126. doi:10.1186/s13059-020-02043-x

Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M., and Snyder, M. (2013). Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing. *G3 (Bethesda).* 3, 387–397. doi:10.1534/g3.112.004812

Timp, W., and Timp, G. (2020). Beyond Mass Spectrometry, the Next Step in Proteomics. *Sci. Adv.* 6, eaax8978. doi:10.1126/sciadv.aax8978

Todorov, H., Fournier, D., and Gerber, S. (2018). Principal Components Analysis: Theory and Application to Gene Expression Data Analysis. *Genomics Comput. Biol.* 4, 100041. doi:10.18547/gcb.2018.vol4.iss2.e100041

Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., et al. (2017). Landscape of X Chromosome Inactivation across Human Tissues. *Nature* 550, 244–248. doi:10.1038/nature24265

Uapinyoying, P., Goecks, J., Knoblach, S. M., Panchapakesan, K., Bonnemann, C. G., Partridge, T. A., et al. (2020). A New Long-Read RNA-Seq Analysis Approach Identifies and Quantifies Novel Transcripts of Very Large Genes. *bioRxiv*, 898627. doi:10.1101/2020.01.08.898627

Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S., and Hernandez, R. D. (2016). Selection and Explosive Growth Alter Genetic Architecture and Hamper the Detection of Causal Rare Variants. *Genome Res.* 26, 863–873. doi:10.1101/gr.202440.115

Velten, L., Story, B. A., Hernández-Malmierca, P., Raffel, S., Leonce, D. R., Milbank, J., et al. (2021). Identification of Leukemic and Pre-leukemic Stem Cells by Clonal Tracking from Single-Cell Transcriptomics. *Nat. Commun.* 12, 1366. doi:10.1038/s41467-021-21650-1

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). "Extracting and Composing Robust Features with Denoising Autoencoders," in Proceedings of the 25th international conference on Machine learning - ICML '08, Helsinki, Finland (New York: . ACM Press), 1096–1103. doi:10.1145/1390156.1390294

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples. *Theor. Biosci.* 131, 281–285. doi:10.1007/s12064-012-0162-3

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* 456, 470–476. doi:10.1038/nature07509

Wang, Y.-C., Peterson, S. E., and Loring, J. F. (2014). Protein post-translational Modifications and Regulation of Pluripotency in Human Stem Cells. *Cell Res* 24, 143–160. doi:10.1038/cr.2013.151

Wang, Y., Liu, J., Huang, B. O., Xu, Y.-M., Li, J., Huang, L.-F., et al. (2015). Mechanism of Alternative Splicing and its Regulation. *Biomed. Rep.* 3, 152–158. doi:10.3892/br.2014.407

Yépez, V. A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N. H., Alston, C. L., et al. (2021a). Clinical Implementation of RNA Sequencing for Mendelian Disease Diagnostics. *medRxiv*, 21254633. doi:10.1101/2021.04.01.21254633

Yépez, V. A., Mertes, C., Müller, M. F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., et al. (2021b). Detection of Aberrant Gene Expression Events in RNA Sequencing Data. *Nat. Protoc.* 16, 1276–1296. doi:10.1038/s41596-020-00462-5

Yeung, K. Y., and Ruzzo, W. L. (2001). Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics* 17, 763–774. doi:10.1093/bioinformatics/17.9.763

Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C. L., and Cai, J. J. (2015). Aberrant Gene Expression in Humans. *PLOS Genet.* 11, 1–20. doi:10.1371/journal.pgen.1004942

Zhang, F. (2019). Defining Inflammatory Cell States in Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-Cell Transcriptomics and Mass Cytometry. *Nat. Immunol.* 20, 928–942. doi:10.1038/s41590-019-0378-1

Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T. J., Lee, C. M., Banskota, S., et al. (2016). A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *Am. J. Hum. Genet.* 98, 11. doi:10.1016/j.ajhg.2015.12.023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Schlieben, Prokisch and Yépez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.