Bioinformatics, 2024, 09–15
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: September 15, 2024
Manuscript Category

Bioinformatics

Analysis of outlier expression and splicing tools in optimal transcriptomics-based pipeline to improve diagnostics of monogenic disorders

Lonneke van Brussel[1], Ellen van de Geer-de Jong[1], Mia L. Pras-Raves[1], Hanneke W.M. van Deutekom[1], Richard H. van Jaarsveld[1], Mireia Olivella García[2]

[1]Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands, [2]Faculty of Science and Technology, Universitat de Vic-Universitat Central de Catalunya, Catalonia, Barcelona, Spain.

Abstract

For rare genetic disorders, correct and in time diagnosis improves disease management and treatment. Although routine DNA based testing partly covers this need, it leaves the majority of cases unresolved. RNA-seq is an upcoming tool in genome diagnostics, found to improve the diagnostic yield. Here, we set out to install an optimal Nextflow nf-core RNA-seq pipeline, in order to provide additional diagnostic value for patients with monogenic disorders. We analyzed aberrant expression and splicing results with the tools OUTRIDER and FRASER for 28 individuals, including 15 undiagnosed patients. To enable the detection of nonsense-mediated decay, samples under normal culture conditions were compared to samples treated with cycloheximide. In addition, we compared results between these tools for aberrant splicing events. Analysis was facilitated by using a Jupyter Notebook web application, adjusted to our requirements. We were able to confirm all positive control samples for aberrant gene expression with OUTRIDER and for aberrant splicing with the FRASER analysis. Candidate genes for 9 out of 15 patients are found with aberrant gene expression analysis and no candidate genes with aberrant splicing events. FRASER analysis demonstrated at least a three-fold increase in positive predictive value for the top 20 aberrant splicing events within our settings and sample set, compared to OUTRIDER aberrant exon expression analyses. Our results demonstrate additional value of OUTRIDER and FRASER in a Nextflow nf-core pipeline to improve diagnostics of monogenic disorders.

Contact: l.vanbrussel@umcutrecht.nl
Supplementary information: Supplementary data are available at https://github.com/lonnekevanbrussel/Final-Master-Project/tree/main

1 Introduction

About 300 million people worldwide are affected by a rare disorder, out of which eighty percent have a genetic cause. At the time of writing, over 6000 rare genetic disorders have been identified, and seventy percent of those start in childhood.[1] A genetic diagnosis is instrumental in clinical care, providing subsequent treatment and prognostic information. However, establishing a definitive genetic diagnosis can take many years, precluding optimal disease management.[2]

In recent years, Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) have greatly improved the diagnostic yield for rare genetic disorders, providing a resolution for 25-35% of cases[3-5]. Unfortunately, this leaves the majority of cases unresolved. Limitations of WES are the frequently missed repeat expansions and non-coding variants, impacting gene expression and splicing. Up to 30% of pathogenic variants are suggested to fall within non-coding regions[6]. Even though WGS does cover these regions, challenges remain due to the vast number of over 3 million Single Nucleotide Variants (SNVs) per sample, coupled with the limited knowledge and predictive power of algorithms regarding the functional and clinical impact of these variants.

RNA sequencing (RNA-seq) is a method that provides direct insight into gene expression and splicing caused by genetic changes and therefore

improves interpretation. For diagnosing rare genetic disorders, RNA-seq is an upcoming though not yet standard tool. A growing number of reports demonstrate additional diagnostic value for diverse cohorts of rare disorders.

By comparing the RNA-seq profiles of different tissue types, skin fibroblasts are found to provide the most broad and useful information for diagnostic purposes[9,10]. In addition to relative over- or under-expression, aberrant splicing and skewed or monoallelic expression (MAE) can be identified. The diagnostic yield is found to be 7.5-36%, which suggests RNA-seq can become an effective complementary tool for standard genetic diagnostics [7-11].

One-third of all inherited human diseases are caused by nonsense or frameshift mutations, introducing a premature termination codon in a transcript[12]. These corrupt mRNA transcripts are prone to be detected and broken down by the Nonsense Mediated Decay (NMD) pathway, a fundamental mechanism of cellular mRNA degradation. This process prevents incorrect proteins from arising. However, NMD not only eliminates corrupt transcripts, but it also modulates the levels of a variety of naturally occurring transcripts. This variability can preclude the identification of corrupt transcripts with RNA-seq.[13,14]

In case the NMD pathway of the cell is blocked, occurrence of corrupt transcripts will increase more strongly with respect to normal transcripts. Blocking of the NMD pathway can be obtained by using cycloheximide (CHX), a natural occurring fungicide produced by the bacterium Streptomyces griseus.[15] By comparing the difference in gene expression between normal cell conditions and CHX conditions, the ability to identify pathogenic variants is enhanced.

The diagnostic value of RNA-seq for rare genetic disorders highly depends on the ability to identify outlier transcriptomic events within a single patient. Rare disease diagnostics therefore requires a different approach compared to more broadly used RNA-seq analysis, for cancer research or other more common diseases. In the latter case, small, predefined groups of several replicate samples are compared, generally between treatment and control, to detect a subtle fold change between controlled populations. These approaches borrow information across genes to estimate the within-group variability, aimed to increase the robustness of the differential-expression analysis[16]. In contrast, for rare diseases, one sample of interest is compared to a reference group.

One tool to find aberrant expressed genes in RNA-seq samples of rare genetic diseases is OUTRIDER[16] (OUTlier in RNA-seq fInDER). It works by fitting a negative binomial model to RNA-seq read counts, correcting for variations in sequencing depth and apparent co-variations across samples. Read counts that significantly deviate from the distribution are detected as outliers. OUTRIDER makes use of an autoencoder, to control automatically for covariation patterns among genes, not known a priori. It uses an adapted Principal Component Analysis (PCA) method. By introducing corrupted counts in an iterative procedure, an optimal encoding dimension is obtained to control for confounders. Additionally, OUTRIDER can be used to detect outliers at exon (or intron) level[7], indicating alternative splicing events in an indirect manner, since aberrant expression can be a result of aberrant splicing. OUTRIDER is however developed to identify outliers in gene expression, exon-level usage is not mentioned in the original article[16].

An approach employing a parallel tool to detect aberrant splicing events directly can be beneficial, because a direct approach focuses more specific on the location and sort of event. FRASER[17] (Find RAre Splicing Events in RNA-seq) is used to detect aberrant splicing events directly by counting splice sites and junctions. FRASER can identify any type of aberrant splicing event from exon skipping and alternative donor usage to intron retention. Like OUTRIDER, FRASER also uses the autoencoder approach and models the read count ratios in the psi values (percent spliced in) and considers both split and non-split reads within a single metric for all splicing events - the Intron Jaccard Index. It works by fitting a beta binomial model - where the probability of success in a fixed number of trials is unknown or random - to the psi values obtained from RNA-seq read counts and correcting for apparent co-variations across samples. Values that significantly deviate from the distribution are identified as outliers.

To enable standardized diagnostic analysis, Nextflow[18] is a powerful, flexible and well-documented workflow language. Recently, several standard 'nf-core'[19] bioinformatics pipelines were developed. Rnaseq and rnasplice are existing nf-core pipelines that can be used to analyze RNA sequencing data, providing more common methods as DESeq2[20], DEXSeq[22] and EdgeR[21] for differential gene expression and splicing. These methods are however not suitable to detect outliers for diagnosing rare genetic disorders, as mentioned before. To obtain an optimal pipeline for rare genetic disease diagnostics, appropriate tools need to be added to the standard options.

Our primary goal is to implement a transcriptomics analysis for patients with monogenic disorders. In our setting, we combine a Nextflow nf-core rnaseq pipeline and built in OUTRIDER and FRASER tools to detect outliers in gene expression and splicing. Additionally, we compared both methods for detecting aberrant splicing events and developed a technique to enhance the identification of outliers while reducing false positive findings in OUTRIDER aberrant exon expression analysis.

To facilitate the analysis, gene associated clinically relevant phenotypes from OMIM (Online Mendelian Inheritance in Man)[23] and other online catalogs were linked to the expression and splicing results. In addition, we used and adjusted an existing web accessible Jupyter Notebook[24] application[7] to our requirements in order to visualize and interpret the results.

Our RNA-Rare pipeline will be of additional diagnostic value for patients with monogenic disorders.

2 Methods

2.1 Patient samples and ethical consideration

We selected 15 individuals who were suspected to have a variety of monogenic disorders. All patients previously underwent routine genome diagnostic testing, which did not result in a diagnosis. In addition, we collected positive control samples, parents and a non-diseases individual (**Table 1**). Samples are anonymized and specific genetic variants are not mentioned in this report for privacy reasons. This study is performed within the ethical framework at the UMC Utrecht.

2.2 Nextflow nf-core rnaseq pipeline

A Nextflow nf-core rnaseq bioinformatics pipeline is set up on a Linux based high-performance computing (HPC) cluster environment using GRCh38[25] as a reference genome and GENCODE version 44[26] for gene annotation. The pipeline takes collected unstranded, paired-end FASTQ files as input, which are processed consecutively by TRIMGALORE version 0.6.7 for adapter trimming[27], STAR ALIGN v2.7.10a for alignment[28] and SAMTOOLS v.1.16.1 to merge and index[29], creating BAM and BAM-index files. Read counting was performed with SUBREAD FEATURECOUNTS v2.0.1[30], separately for gene and exon level with settings for multimapping, largest overlap and extra attributes gene_id and gene_name, extracted from the

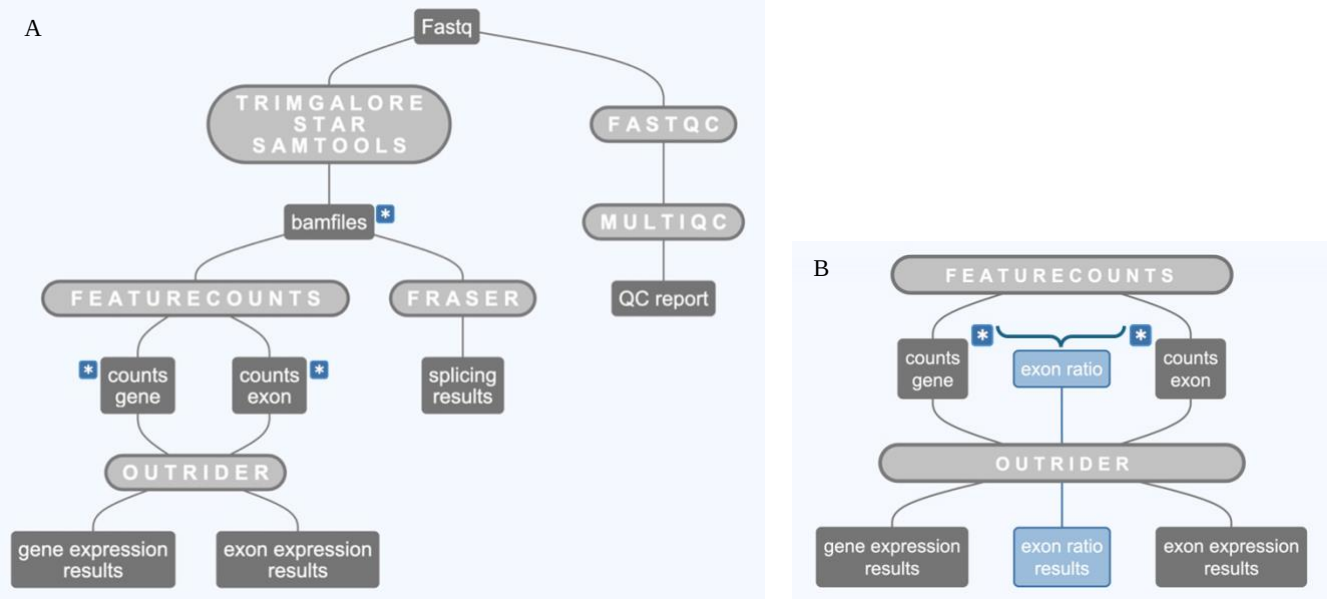

Figure 1. Schematic overview workflow RNA-Rare pipeline
(**A**) Workflow 'RNA-Rare' pipeline in diagnostic setting. Besides running the workflow as a whole, sub workflows are created to run for fastq-to-bam, quality control, featureCounts, OUTRIDER and FRASER seperately. These sub workflows are used to collect sets of BAM and featureCount files corresponding the sample set, for testing, and can be used when errors occur in diagnostic setting.
(**B**) Sub workflow OUTRIDER with added exon ratio procedure.
* At these points in the workflow, sets of BAM and featureCount files are collected for all samples (untreated and CHX) by using the sub workflows. Collected sets are compiled as input reference set, alongside a 'sample of interest' to the OUTRIDER and FRASER modules.

annotation file. Quality Control is performed alongside by FastQC[31] and MultiQC version 1.17[32]. In addition to the existing Nextflow nf-core rnaseq pipeline, we created new submodules in the workflow for aberrant expression with OUTRIDER v1.18.1[16] and splicing with FRASER v1.99.4[17]. We added sub workflows to start the pipeline at several entry points with corresponding input files. See **Figure 1A** for a schematic overview of our diagnostic 'RNA-Rare' pipeline.

2.3 OUTRIDER settings

For the analysis, and prior to running OUTRIDER, the raw counts from featureCounts were manually filtered to retain only protein coding genes and exons. Within the OUTRIDER module, the *filterExpression* function filters out lowly expressed genes. This method uses a cut-off value of Fragments Per Kilobase per Million mapped reads (FPKM) > 1 for the 95th percentile over all samples per gene or exon. Filtering steps, amounts and percentages are listed in **Table S1**.

Because technical and biological confounders can adversely affect the detection of aberrant features, the functions *sizeFactors* and *controlForConfounders* are executed. The latter uses an autoencoder that automatically controls for technical and biological confounders present in the data. Therefore, an encoding dimension q needs to be set. The optimal value of q can be determined using the *findEncodingDim* function. Default number of iterations is set to 15 for this function. The data is fitted within this procedure.

Generated output values are normalized counts per feature (gene or exon) and mean corrected counts over all samples – mean normalized counts over all samples per gene or exon. In addition, z-scores, p-values and transcriptome-wide False Discovery Rate (FDR) adjusted p-values are calculated.

2.4 Exon-ratio method

Because we were not able to confirm all control samples with the OUTRIDER exon-level analysis, a method was developed to determine the expression differences between all exons within the corresponding gene. To this extent, the ratios of exon FPKM over the corresponding gene FPKM values were calculated, multiplied by 100 and rounded to the nearest integer value - to simulate counts. OUTRIDER expects integer values and will otherwise raise an error. From here on, this method will be referred to as 'exon ratio'.

The obtained exon-ratio dataset was filtered prior to running the OUTRIDER module, with identical resulting exon ids as obtained with the exon-level *filterExpression* step, mentioned in Section 2.3. The resulting, filtered 'exon-ratio' files serve as input to the OUTRIDER module. Within the OUTRIDER module, the exon-ratio files are filtered one last time, at this point only to eliminate exons with zero counts in all samples, to prevent the OUTRIDER algorithm from raising an error (**Table S1**). **Figure 1B** shows an overview of the exon-ratio workflow added to the RNA-Rare pipeline.

2.5 FRASER settings

FRASER uses BAM files as input format. After counting junctions and splice sites, delta psi values are computed. Delta psi is the difference between the observed and expected psi value (Jaccard metric). Subsequently, filtering is performed, based on settings suggested by FRASER vignette[48] and 'Gagneurlab' DROP[34] pipeline[49]. At least one sample has 20 (or more) reads, at least 5% of the samples per junction need to have at least 10 counts. The minimal delta psi variation for an intron to pass the filter is set to 0.05. Filtering results are listed in **Table S1**.

Table 1. All individuals with annotation regarding the sample set, their presenting clinical phenotype, expected gene or exon with expression or splice aberration, priorly known type of variant and effect.

Individual	Annotation	Clinical phenotype	Expected gene	Expected exon	Type variant and effect	Included in sample set
01	Positive control gene-level parent of 26	Developmental and epileptic encephalopathy	CAD		Deletion 1 base on one allele resulting in frameshift and NMD	Yes
02	Positive control gene-level		USP11		Balanced translocation chromosome X	Yes
03	Positive control gene-level & exon-level + replicate sample	Nephronophtisis	NPHP1		Deletion on one allele, splice site variant on other	Yes + No
04	Positive control exon-level	Familial aortic aneurysm	LOX	ENSE00003647722.1	Deletion 3 nucleotides, no frameshift - exonskip	Yes
05	Positive control exon-level	Increased risk of (endometrial, colorectal) cancer	MSH6	ENSE00003670482.1	Exonskip	Yes
06	Positive control exon-level	Spastic paraplegia	SPG7	ENSE00003537923.1	Intronic variant, alternative splice site	Yes
07	Positive control exon-level	Renal cell carcinoma	PBRM1	Multiple	Distal insertion of exons 7-30	Yes
08	Patient – sibling of 28	Lactic acidosis				Yes
09	Patient	Early onset epileptical encephalopathy, mitochondriopathy	COX20, TRAPPC12, GFM1, CARS2			Yes
10	Patient	Symptoms like Smith-Kingsmore syndrome: ID, facial features, small thorax, short limbs	MTOR			Yes
11	Patient	Short stature and develop-mental delay, expected mitochondriopathy	HAVCR1			Yes
12	Patient	ID, abnormal facial shape, abnormality of the kidney	HAVCR1, GMEB2			Yes
13	Patient	Primary microcephaly				Yes
14	Patient	Symptoms like Myhre Syndrome - Microcephaly, ID facial features	SMAD4			Yes
15	Patient	Female, 2 years, microcephaly, skeletal defects, hearing loss, Ventricular Septal Defect, gastrointestinal, renal defects	SALL4			Yes
16	Patient	Intrauterine growth retardation, CNS malformation, motor delay	DNAH6			Yes
17	Patient	Spastic paraplegia	SPG7		pathogenic variant in one allele, patient has dominant phenotype	Yes
18	Patient	Cardiomyopathy	DSP			Yes
19	Patient	Dwarfism				Yes
20	Patient		CRK			Yes
21	Patient				Translocation 46,XY,t(2;6)(q35;q23.3) or t(2;6)(q37.1;q25.1)	Yes
22	Parent of 27					Yes
23	Parent of 27					Yes
24	Parent of 26					Yes
25	Non-diseased individual					Yes
26	Positive control – child of 01&24	Developmental and epileptic encephalopathy	CAD		Deletion 1 base on one allele resulting in frameshift and NMD	No
27	Patient – child of 22&23	Cutis laxa, spastic paraplegia microcephaly, short stature, ID, hypotonia/weak limbs	ALDH18A1			No
28	Patient - sibling of 08	Lactic acidosis				No

ID: Intellectual disability

Hyperparameter optimization is used to estimate the dimension q of the latent space of the data. It works by artificially injecting outliers into the data and subsequently comparing the AUC by recalling these outliers for different values of q. The best q is determined with the *bestQ* function - to control for technical and biological confounders. Subsequently, the data is fitted. In both steps, Principal Component Analysis (PCA) is used as implementation method for optimization and fitting. The results are annotated using biomaRt v2.59.0[33] with GRCh38 as a reference. P-values are determined and FDR correction is performed transcriptome-wide.

2.6 Reference set in OUTRIDER and FRASER

The OUTRIDER and FRASER modules in our workflow are set up to compare one sample to a reference set. For the analyses of this project, one 'sample of interest' is therefore excluded from the sample set and functions as input sample, to be compared to the other samples – the reference set. OUTRIDER returns outliers in all samples independently. In diagnostic settings, outlier results for samples of the reference set are excluded from the results. For the analyses of this project, we retained results for all samples, to prevent unnecessary repetitions.

2.7 Visualization, clinical interpretation and code availability

To facilitate determination of the potential diagnostic value of an aberrant expressed feature or splicing event, gene associated clinical phenotypes were added to the OUTRIDER and FRASER results. The Ensemble API[35] is used for this purpose. If an OMIM entry exists for a consulted gene, the associated phenotype is attached to the result as a first choice, otherwise a different source is chosen.

A web-based Jupyter Notebook Voilà[50] app, created by researchers at the Erasmus Medical Center[7], was adjusted to our data and specific requirements to facilitate analyses of OUTRIDER results. Several plots and tables were made available with supported filter options for combinations of all samples, conditions and features (**Figure S12**).

Finally, BAM file coverage and sashimi plots were studied in Integrative Genomics Viewer (IGV) v2.17.4[36] with GRCh38 as a reference.

The code for the RNA-Rare pipeline, OUTRIDER and FRASER modules, Jupyter Notebook app and additional scripts used for the analyses can be found in **Section 6**.

3 Results

3.1 Study design

We set out to install an optimal pipeline for RNA-seq diagnostics for monogenic disorders. To that end, we collected dermal fibroblasts for 15 individuals who were suspected to have a variety of monogenic disorders. These individuals previously underwent routine genome diagnostic testing, which did not result in a diagnosis. In addition to these patients, we collected fibroblasts from four parents, one sibling, seven positive controls and one non-diseased control sample (**Table 1**).

To facilitate automated and quantitative analysis, we created a sample set of 25 samples. This set includes 14 patients (individual 8 to 21), 4 parents (individual 1, and 22 to 24), 3 positive controls for gene-level analysis (individual 1 to 3; including 1 parent – individual 1), 5 positive controls for exon-level analysis (individual 3 to 7 - individual 3 is used for gene and exon-level analyses) and 1 non-diseased control sample (individual 25). Three samples were excluded from the sample set; two children (individual 26 and 27) and one sibling (individual 28) were excluded, so the other sibling and both sets of parents remained in the sample set to prevent bias. This sample set will be the foundation of our diagnostic 'RNA-Rare' pipeline, enabling the identification and validation of pathogenic variants in the RNA-seq data. The excluded children and sibling, as well as one replicate of a positive control sample (for individual 3), were separately analyzed with the sample set excluding the corresponding parents or sibling.

We generated RNA sequencing data as FASTQ files from all fibroblasts under normal culture conditions (untreated) and from cultures treated with cycloheximide (CHX) to enable the detection of transcripts subject to nonsense-mediated decay. FASTQ files are the input format of our 'RNA-Rare' pipeline, generating BAM files and featureCounts files for the sample set. Exon-ratio files are obtained from featureCount gene and exon-level files. **Figure 1** shows the diagnostic workflow setup, sub workflows are used to collect the sets of different files. Additions to the workflow for the exon-ratio method outside the pipeline, are elaborated in **Figure 1**.B and Section **2.4**.

In summary, we created sample sets of 25 individuals for untreated and CHX treated conditions. For both conditions, we subsequently collected BAM files and featureCounts files, the latter at both gene and exon-level, and exon-ratio files. The sets of featureCounts files are input to the OUTRIDER algorithm, while FRASER uses BAM files as input format.

3.2 Filtering and Normalization

For optimal expression outlier detection, we filtered the featureCounts data by selecting only protein coding genes and exons, as coding genes are of interest for diagnostic purposes. This filtering is performed manually and prior to the OUTRIDER module. In addition, we filtered the data to exclude low expressed genes or exons respectively, as it increases the number and confidence of outliers[19]. The latter step executes within the OUTRIDER module (Sections **2.3-2.4**). In the FRASER module, filtering is performed on junction read counts and delta psi values (Section **2.5**). Cut-off settings, resulting amounts and percentages of these steps are listed in **Table S1** and visualized in **Figures S1 and S3**.

FeatureCounts data at gene-level, filtered for protein coding genes, were analysed with Principal Component Analysis (PCA) for untreated and CHX treated samples, showing two distinct groups by treatment type (**Figure 2**). This method revealed sample swaps and one outlier during the course of this project, which were resolved or removed accordingly, demonstrating the necessity of this procedure.

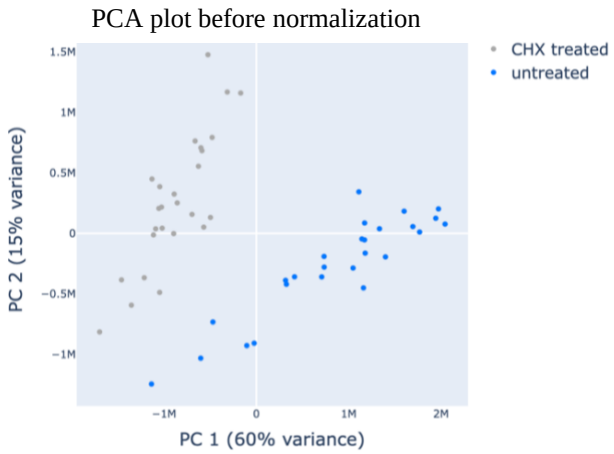

Figure 2. PCA plot of featureCounts data at gene-level, colored by treatment condition, showing similarities in count data after dimension reduction. Two distinct clusters can be distinguished by treatment type.

Consecutively, the filtered count and psi value data is normalized respectively in the OUTRIDER and FRASER modules with the autoencoder procedures. These procedures control for technical and biological confounders, since they can adversely affect the detection of aberrant features. Heatmaps prior and after normalization demonstrate less batches are present in the data and correlations are closer to zero after normalization (**Figures S2 and S4**). This indicates that the normalization succeeded.

3.3 Gene expression analysis

For OUTRIDER gene expression analysis, we first ranked the results by lowest p-value (**Table S3.A**). No significant results with transcriptome-wide FDR adjusted p-value < 0.05 were found. We analyzed the top 20 results in the Jupyter Notebook app and in IGV (top 4 results are depicted in **Figure S5**). The coverage in IGV is compared with a minimum of two other samples. 17 out of 20 top outlier candidates presented with at least 40% down or upregulation (normalized counts versus mean corrected, **Table S3A**). We determined cut-off values of |z-score| > 2.5 and p-value < 0.01, yielding a total of 2017 remaining aberrant gene expression results (0.67%; **Figure 3**, **Table S2**).

All three positive control samples selected for gene-level analysis (individual 1, 2 and 3) showed aberrant gene expression (**Table 2**, **Figure 4**). In addition, we were able to confirm NMD in the gene of interest of individual 1. This individual has a deletion on one allele of the CAD gene that causes a frameshift. The resulting corrupt transcript is subject to NMD. Besides downregulation of CAD in the untreated sample, we found normal gene expression for the CHX treated sample of this individual. This confirms the transcript is subject to NMD as expected (**Figure 4**.A and B).

Candidate genes were found for 9 out of 15 patients (**Table 5**). Besides meeting the determined cut-off values, the associated phenotype of the candidate genes corresponds to the patients' clinical phenotypes. All candidate genes were studied in the Jupyter Notebook app and in IGV. We detected no causative coding variants for any of the candidate genes by studying the corresponding BAM files in IGV. Five patients demonstrated downregulated gene expression for five (groups of) genes in untreated samples and CHX treated samples as well. This suggests regulatory causes. For four individuals eight candidate genes were found downregulated in untreated samples whereas normal in CHX-treated samples. This suggests a possible subjection to NMD, although no causative coding variants were detected in IGV. Lastly, one upregulated gene was detected.

Additional follow-up tests are currently being performed to find evidence of causative variants, by (re-)analysis of WES or WGS data to detect mutations in intronic or regulatory regions and epigenetic tests regarding other regulatory elements.

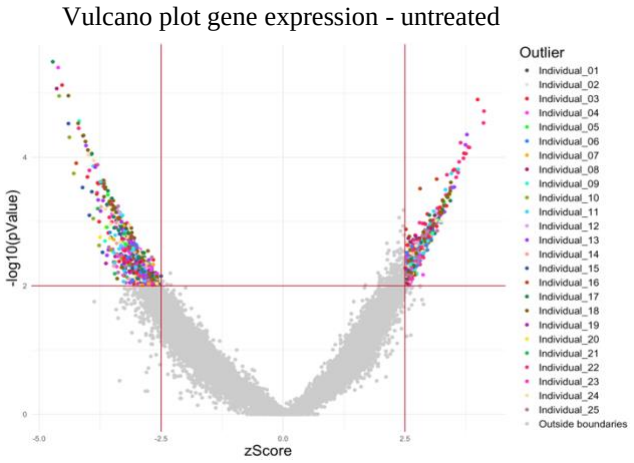

Figure 3. Vulcano plot for aberrant gene expression including all untreated samples of the sample set. Red lines indicate cut-off values |z-score| > 2.5 and p-value < 0.01. 2017 remaining outlier genes are colored by individual.

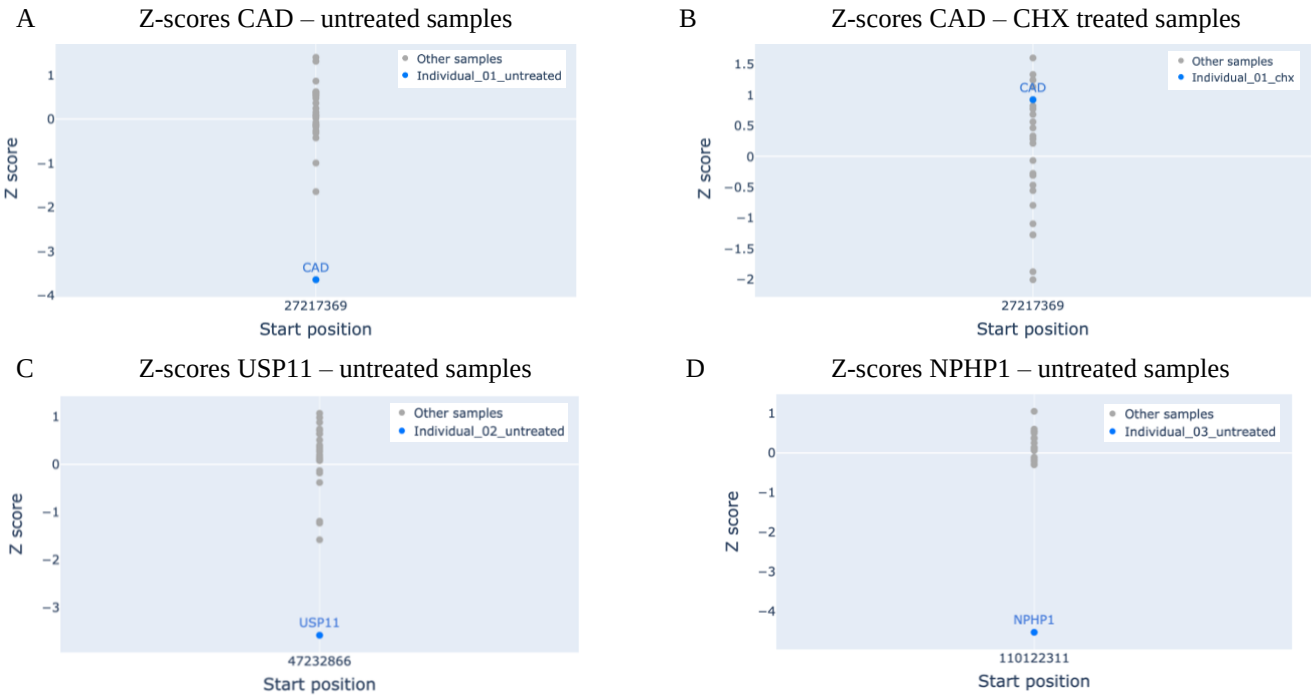

Figure 4. Z-score plots for all three positive control samples for gene-level analysis, indicating that OUTRIDER detects the positive controls as outliers. (**A**) shows downregulation of the CAD gene in the untreated sample for individual 1 compared to the other untreated samples of the sample set, whereas normal gene expression for the CHX treated sample of individual 1 in (**B**), confirming the CAD transcript is subject to NMD for this individual. (**C**) and (**D**) confirm downregulation of USP11 and NPHP1 respectively for individual 2 and 3 compared to the other untreated samples of the sample set.

Table 2. OUTRIDER gene-level results for positive control samples with priorly known gene-level mutations and expected gene expression aberrations

Individual	Gene	z-score	p-value	Rank1[1]	Rank2[2]	normalized counts[3]	mean corrected[4]	z-score CHX
01	CAD	-3,65	0.000911	67	3	3607	5664	0.92
26	CAD	-3,33	0.000911		5	3833	5564	-0.16
02	USP11	-3,58	0.000504	1	1	1400	3156	-4,13
03	NPHP1	-4,53	0.000008	3	1	50	286	not in data
03 replicate[5]	NPHP1	-4,08	0.000092	19	1	93	290	not in data

[1]Rank 1 is based on results for all samples
[2]Rank 2 is based on results per sample/individual
[3]Normalized counts, counts after normalization in OUTRIDER
[4]Mean corrected: mean normalized expression counts for all samples of this gene
[5]For individual 3 two replicate samples are used in separate runs, comparable results are obtained

3.4 Exon expression and splicing analysis

To detect aberrant splicing events, we studied both FRASER splicing and OUTRIDER exon expression results. In addition, we compared the outcome of these analyses.

The results per analysis are ranked by lowest p-value to determine the top findings. FRASER detected 61 significant aberrant splicing events with transcriptome-wide FDR adjusted p-value < 0.05, whereas OUTRIDER yielded no significant findings for multiple testing correction. The top 20 results for both methods (**Table S3.B** and **S3.E**) were analyzed in the Jupyter Notebook app and in IGV sashimi plots to confirm aberrant splicing events. For FRASER, we were able to confirm 16 out of 20 top results as unique aberrant splicing events, a positive predictive value (PPV) of 0.8, whereas among OUTRIDER exon-level results only 2 out of 20 were detected (**Table 4**, PPV 0.1).

Moreover, the five positive control samples selected for exon-level analysis were all detected by FRASER with a p-value < 10^{-6}, whereas only one positive control sample showed outlier results with a |z-score| > 2.5 and p-value < 0.01 for OUTRIDER exon-level analysis (**Table 3** and **Table 5**). For FRASER results, we determined cut-off values for p-value < 0.001 and |deltaPsi| > 0.05 (**Figure 5**).

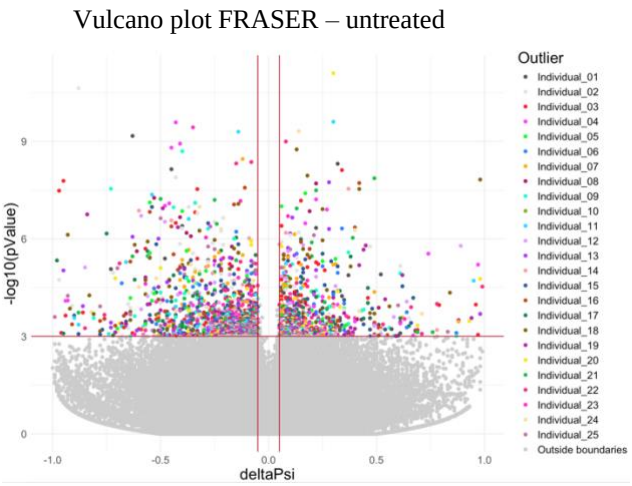

Figure 5. Vulcanoplot FRASER splicing results for untreated samples, showing cut-off values p-value < 0.001 and |deltaPsi| < 0.05 for aberrant splicing events. 2163 results remain and are colored by individual.

Because four out of five positive control samples selected for splicing analysis were not detected with OUTRIDER exon-level analysis, we developed the exon-ratio method. This method calculates a ratio of the expression differences between all exons within the corresponding gene, instead of comparing singly exon expression or all exons transcriptome-wide, elaborated in Section **2.4**. With the OUTRIDER exon-ratio analysis, we found aberrant exon expression for four out of five positive control samples (**Table 3**). The differences in exonic z-scores between the exon-level and exon-ratio results for untreated positive control samples are depicted in **Figure 6** and **Figure S11** (IGV sashimi plots in **Figure S10)**. These results demonstrate improved detection of splicing events in our positive control samples with the exon-ratio method.

Additionally, we ranked OUTRIDER exon-ratio results by lowest p-value (**Table S3.C**). We did not detect any significant results with transcriptome-wide FDR adjusted p-value and we could only confirm an aberrant splicing event for one out of the 20 top results. Hence, the majority of results in the ranked top 20 are still false positive findings.

To minimize the number of false positives, we determined factors to filter the exon-ratio results. We detected a large number of results of multiple exons of the same gene with low z-scores, whereas no splicing events could be confirmed in an IGV sashimi plot. In addition, splicing events were difficult to interpret and confirm in sashimi plots for samples with low (mean) expression. We focused our filtering steps based on one downregulated exon per gene. The four steps to achieve this filtering are summarized and visualized in **Figure 7** and **Table S2**. The final results after these filtering steps are ranked by p-value and studied in the Jupyter Notebook app and IGV (**Table S3.D**). After this filtering, we confirmed 5 out of 20 top results as unique aberrant splicing events and therefore true positive findings (positive predictive value 0.25).

In summary, FRASER detected a 3.2-fold improvement in PPV for the top 20 results compared to the filtered exon-ratio results, and 8-fold compared to the (unfiltered) exon-level OUTRIDER results (**Table 4**). All five positive controls with an aberrant splicing event were confirmed as outliers by FRASER, whereas OUTRIDER detected a maximum of four out of five positive control samples with a rather lower level of certainty. In addition, all true positive findings in the OUTRIDER exon-level and exon-ratio analysis top 20 results were also detected as outliers by FRASER, while only 4 out of the 16 true positives detected by FRASER, were also found in the exon-ratio results (with |z-score|>2.5). We did not detect any (new) candidate genes regarding aberrant splicing in our patient cohort.

Table 3. OUTRIDER Expression and FRASER splicing results for positive control samples, selected for exon-level analyses, with priorly known splice aberrations.

Indiv	Gene	OUTRIDER Gene level z-score untreated	z-score CHX	Exon level z-score	p-value	Rank 1[1]	Exon ratio z-score	p-value	Rank 1[1]	Rank 2[1]	FRASER Splicing p-value	p-adjusted	delta-psi	Rank 1[1]
03	NPHP1	-4.53	-	-4.05	0.000788	2	-4,62	1.190e-05	524	-[2]	3.301e-08	0.00855	-0.97	35
04	LOX	-0.54	-0.82	-1,56	0.127993	1000+	-4,35	1.147e-05	510	2	1.173e-09	0.00054	-0.41	10
05	MSH6	0.67	-1.53	-2,46	0.015561	1000+	-3,15	0.002072	1000+	360	2.350e-06	0.47128	-0.38	148
06	SPG7	-2.7	1.09	-2.73[3]	0.009413	1000+	0.38	0.744146	1000+	-[2]	6.952e-06	1	0.25	229
07	PBRM1	0.46	-	-1.82	0.084906	1000+	-2.52	0.019608	1000+	-[2]	2.166e-06	0.50452	-0.37	146

[1]Rank 1 results ordered by lowest p-value, Rank 2 results ordered by lowest p-value, after filtering steps (**Figure 7**, **Table S2**) are applied,
Ranks are based on results for all samples
[2]Rank 2 results for Individual 3-NPHP1, Individual 6-SPG7 and Individual 7-PBRM1 are missing, since they did not meet the filtering steps criteria.
[3]Individual 6 has an alternative splice site and partial intron retention in SPG7 (see sashimi plot, **Figure S10D**). Downregulated exon expression is not expected and is therefore not a confirmation of this splicing event.

Table 4. Number of false and true positives in top results 20 lists for all exon expression and splicing analyses of untreated samples (based on **Table S3A-E**). Findings are determined true positive when an aberrant splicing event is detected in the IGV sashimi plot, while the event is not present in at least two randomly selected other samples.

	OUTRIDER Exon level	Exon ratio before filtering	Exon ratio after filtering	FRASER Splicing
True Positives	2	1	5	16
False Positives	17	18	15	0
Double/unclear	1	1	0	4

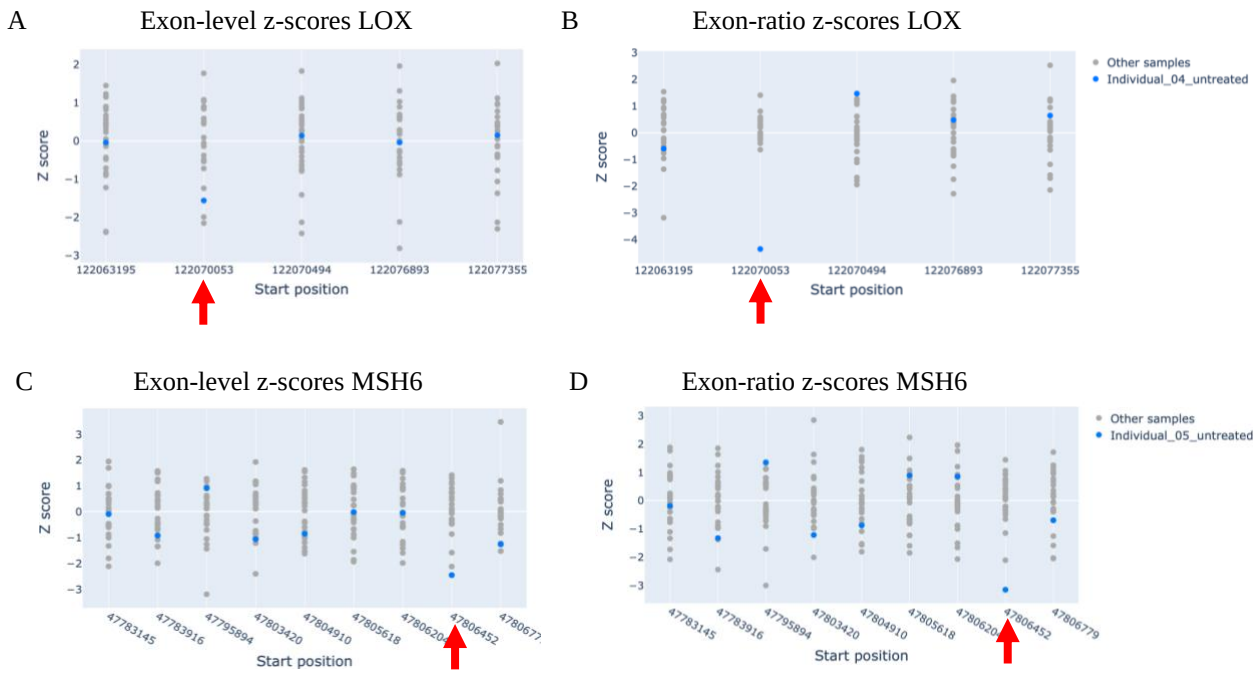

Figure 6. Exonic z-score differences between OUTRIDER exon-level (**A, C**) and exon-ratio **(B, D**) results for two of the positive control samples, demonstrating improved detection of exon skips with the exon-ratio method. All figures depict z-scores for untreated samples. The z-scores of the LOX and MSH6 gene for individual 4 and 5 respectively are depicted in blue, whereas all other samples of the sample set are depicted in grey. OUTRIDER exon-level analysis did not detect exonic outliers (|z-score| < 2.5) (**A** and **C**), whereas the exon-ratio results did show downregulated exon expression at the expected positions and individuals (**B** and **D**).

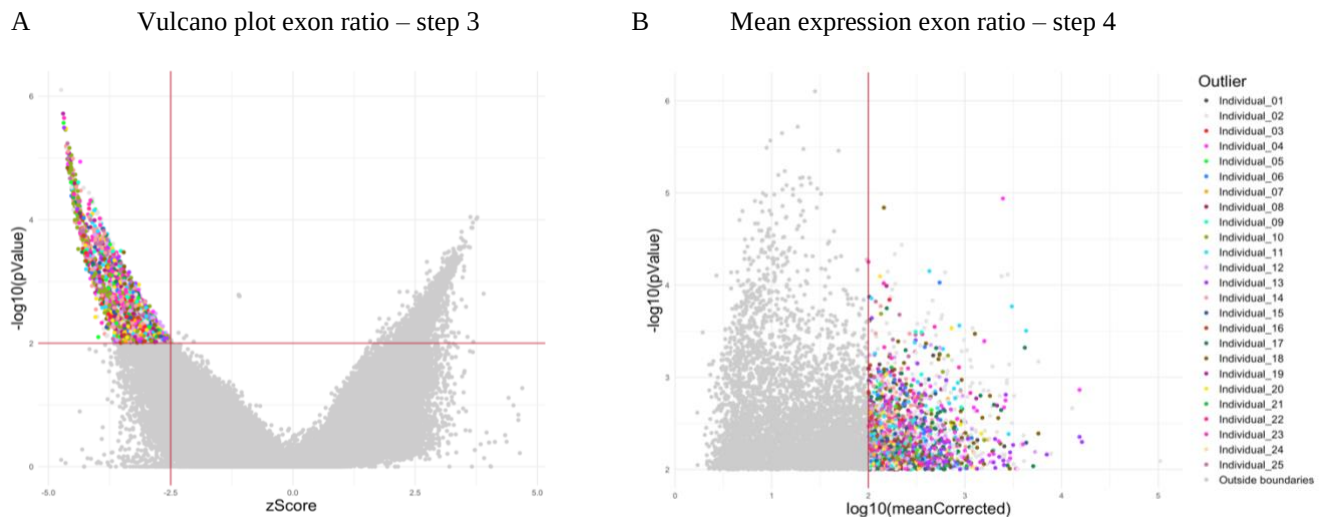

Figure 7. OUTRIDER exon ratio results plotted for filtering steps 3 and 4. We focused our filtering on one downregulated exon per gene.
Step 1 filters out exons of *gene*-sample combinations with *gene*-level normalized counts < 200.
Step 2 In addition, exons of genes with multiple aberrant expressed exons: z-score < -2 are filtered out.
Step 3 applies cut-off values of z-score < -2.5 and p-value < 0.01 on the remaining *exon ratio* results. (**A**) shows the results of step 2 and cut-off values of step 3.
Step 4 applies a cut-off value of > 100 mean corrected counts. (**B**) shows the results of step 3 for log10(meanCorrected) values and p-values with cut-off values of step 4. This cut-off is chosen, since aberrant splicing events are difficult to confirm in IGV sashimi plots with low mean exon expression over all samples.

3.5 Candidate genes

Individual 8 and 28

Two brothers, individual 8 and 28, presented with symptoms of lactic acidosis. Gene expression analysis showed downregulation of SLC38A5 to almost zero normalized counts (**Table 5**, **Figure S6**) for the untreated and CHX treated samples, compared to the other samples of the sample set. WES analysis did not reveal any variant. Follow-up WGS analysis will screen for enhancer or promoter mutations. Since SLC38A5 is located at chromosome X, males show mono-allelic expression, explaining the nearly zero gene expression. SLC38A5 is associated with hepatic encephalopathy and Mahvash disease according to GeneCards[37].

Individual 10

Individual 10 presented with a clinical phenotype corresponding to Smiths-Kingsmore syndrome. This syndrome is a neurodevelopmental disorder, including symptoms such as macrocephaly, intellectual disability, and a small thorax. OUTRIDER gene expression analysis detected downregulation of five HOXB genes (**Table 5**). **Figure S7** depicts the corresponding results in Jupyter Notebook app plots and IGV.

HOXB genes are not recognized as clinically relevant in the OMIM catalog, however HOX genes are a well-known group of evolutionarily conserved genes that encode for a family of transcription factors that regulate early developmental morphogenetic processes and continue to be expressed into adulthood. Ten HOX genes are found to cause human disorders, such as intellectual disability, facial and dysmorphic features. Moreover, mouse models show several cervical and sternal malformations in HOXB gene loss of function phenotypes[38].

We did not find evidence with WES analysis regarding any coding mutations or repeat expansions within the HOXB genes. Additional epigenetic follow-up will search for causative evidence in corresponding regulatory elements.

Individual 12

Five aberrant expressed candidate genes are found for individual 12, all showing downregulation for the untreated sample compared to the sample set, whereas normal or upregulated gene expression for the CHX treated sample, suggesting the transcripts can be subject to NMD (**Figure S8**). Individual 12 presented with symptoms such as intellectual disability, abnormal facial shape and abnormality of the kidneys. All candidate genes phenotypes resemble at least two out of three symptoms of the patient, as found in the OMIM online catalog and are autosomal dominant disorders. We did not find any coding variants in IGV or additional results at exon-level, exon-ratio or aberrant splicing indicating a cause of NMD for any of the candidate genes in the untreated and CHX treated samples.

No follow up is yet performed. The first step would be to consult the physician to discuss the findings. Replicate samples can be re-examined for RNA-seq and WES to study reproducibility.

Individual 13

Individual 13 presented with symptoms of primary microcephaly. ZNF462 is found at rank 7 for gene expression outliers regarding this individual (**Table 5**, **Figure S9**). ZNF462 is downregulated for the untreated sample and normal for the CHX treated sample. This finding suggests possible subjection to NMD.

A heterozygous mutation in ZNF462 is associated with Weiss-Kruszka syndrome, causing abnormal head shape in some patients[23]. IGV analysis shows very low coverage for the untreated sample of individual 13 compared to control samples, however no coding mutation supporting NMD is yet found. Replicate samples can be examined after consulting the physician.

Table 5. Candidate genes obtained with OUTRIDER gene expression analysis. Rank is determined for individual results, ordered by p-value

Individual	Candidate Gene	Associated clinal phenotype Source OMIM or otherwise mentioned	Inheri-tance[1]	z-score untreated	p-value	Rank[2]	normalized counts	mean corrected	z-score CHX
08	SLC38A5	Hepatic encephalopathy, Mahvash disease (GeneCards[37])	X	-4.64	0.000009	1	4	2052	-3.77
28[3]	SLC38A5	Hepatic encephalopathy, Mahvash disease	X	-4.63	0.000009	1	5	2032	-3,79
10	HOX2B	Anterior transformation of C2. C1 and sternum		-4,59	0.000011	1	7	1386	-4,47
	HOX4B	malformations. Facial paralysis with abnormal		-4,38	0.000049	2	0	547	-
	HOX3B	cranial nerve VII (in mouse)		-4,29	0.000178	3	0	1542	-3,58
	HOXB7	Encode transcription factors with function in		-3,89	0.000901	4	0	553	-3,62
	HOXB6	developmental processes[38]		-3,77	0.002357			654	
11	PRDX6	Peroxiredoxin: Catalyzes reduction H_2O_2 to H_2O and alcohols (uniProt)		+3,45	0.000179	3	32053	11864	+2,21
	CELF2	Developmental and epileptic encephalopathy[1]	AD	-3,4	0.001634	17	651	3454	-2.31
	PEX19	Peroxisome biogenesis disorder: short stature, developmental delay	AR	-2,95	0.007865	71	1950	2226	-0,38
12	CDC42BPB	Abnormal head shape, ID	AD	-3,39	0.000671	1	9554	11750	-0.85
	SRCAP	ID, dysmorphic facial features, abnormal kidneys	AD	-3,28	0.001688	3	360	545	-1,18
	CNOT1	ID, microcephaly	AD	-3,11	0.001876	8	6375	7799	-1,19
	SETD2	ID, dysmorphic facial features	AD	-2,62	0.007974	14	3899	4595	-0.56
	MTOR	ID, abnormal facial shape renal asymmetry	AD	-2,58	0.010968		4123	4629	-0,75
13	ZNF462	Abnormal head shape and facial features	AD	-3,69	0.000472	7	236	635	-0,59
14	SALL1	Microcephaly, ID,	AD	-3,26[3]	0.004538	5	62	627	-3,21
		facial features		-1,66[3]	0.005436	1000+	70	555	
	SRRM2	ID		-2,98	0.003447		17308	23446	-0.25
15	SALL1	Microcephaly hearing loss, Ventricular septal defect, gastrointestinal, renal defects	AD	-3,97	0.000799	3	0	555	-3,9
	SAMD9L	CNS malformation, motor delay	AD	-3,82	0.000374	3	222	1403	-2,79
27[4]	FLNB	short stature, skeletal features	AD	-3.35	0.001352	9	10250	21500	-2,42

[1]AD: Autosomal Dominant inheritance, AR: Autosomal Recessive inheritance, X: gene located at chromosome X
[2]Ranks are based on results per individual/sample
[3]Z-score -3.26 is found in a run where individual 12 is left out of the sample set.
[4]Results obtained with the sample set excluding corresponding parents or sibling

4 Discussion

We here show that our RNA-Rare pipeline demonstrates promising results for a cohort of 15 patients with monogenic disorders and yet inconclusive results in standard diagnostic testing. We performed RNA-seq in our cohort on dermal fibroblasts, untreated and treated with CHX, followed by OUTRIDER outlier analysis of gene and exon expression and FRASER outlier analysis of splicing events.

We were able to confirm aberrant gene expression with OUTRIDER outlier analysis for all three positive controls selected for gene-level analysis and provided evidence for NMD for one positive control. All five positive control samples with priorly known splicing events were detected as outlier with FRASER analysis.

Not all positive controls with splicing events were detected with OUTRIDER. Only one out of these five positive controls were detected as outlier with OUTRIDER exon-level analysis and four out of five when we applied a self-developed method to increase detection of aberrant exon expression, by determining the exon-ratio for the corresponding gene. In the top 20 results for OUTRIDER exon-level analysis, only two unique true positive aberrant splicing events were found and five were detected with the exon-ratio analysis, the latter after subsequent filtering. With FRASER analysis, we detected 16 out of 20 top results to be unique true positive aberrant splicing events, which is a 3.2-fold improvement in positive predictive value.

We obtained candidate genes for 9 out of 15 patients with aberrant gene expression, and none with exon expression or splicing aberrations. In addition to these transcriptomic aberrations, further evidence needs to be provided. Therefore, follow-up tests are performed, by (re-)analysis of WES, WGS or epigenetic analysis to find evidence in coding, non-coding or regulatory regions. This follow up is currently carried out for several patients.

We did not detect any significant transcriptome-wide FDR adjusted findings within the OUTRIDER results. This could be due to the size of our sample set. Literature on OUTRIDER suggests a minimum sample size of 50-60[16], whereas our sample set consists of 25 samples. Despite the smaller sample size, we were able to identify outliers and confirm positive controls for gene-level analysis. Sensitivity is expected to rise when we increase our sample set. We did not find a minimal suggested sample size in literature on FRASER.

Moreover, since aberrant expression is a result of splicing, OUTRIDER aberrant exon expression is an indirect indicator of aberrant splicing, compared to the more direct method of counting junctions and splice sites as FRASER uses. This can be an explanation for the larger number of false positives among the OUTRIDER exon-level results. We found that aberrant exon expression does not imply an aberrant splicing event per se. Exon expression fluctuates throughout the gene and read counting depend on the featureCounts settings, with more variables to set, compared to FRASER, where counting is part of the module.

We demonstrated an improvement in detecting exonic outliers in the positive control samples with the self-developed exon-ratio method. The exon-ratio calculations are based on gene and exon-level read counts and not on transcript-level. Especially for larger genes with many exons this can give rise to false positive findings, although it will be limited because we used one type of tissue. About two thirds of alternative splicing events have been shown to be tissue specific[41]. Using transcripts instead of genes for the exon-ratio method could reduce the number of false positive findings.

Furthermore, we performed filtering on lowly expressed exons manually within the OUTRIDER module for the exon-ratio method and additionally afterwards to remove false positive results. The applied extra filtering steps on the OUTRIDER exon-ratio results, focuses on only one downregulated

exon per gene. However, this filtering will even so reduce the number of true positives, because it excludes genes with true positive splicing events in multiple and upregulated exons.

Another suggestion regarding the filtering, is to set the parameters of the OUTRIDER *filterExpression* function stricter. E.g. to use a lower percentile and, or higher FPKM cut-off value. Reducing the lowly expressed genes, can increase detection of aberrantly expressed genes[39]. We did some tests with alternative *filterExpression* settings, but subsequent filtering steps were still necessary to encounter only slight increases in results. Optimizing multiple filtering steps can be considered when the sample size increases, and a high percentage of false positive results remain.

In contrast to the improved results that we found for FRASER over OUTRIDER exon-level analysis, other research found higher sensitivity with the OUTRIDER exon-level analysis over FRASER[7]. This study however used a cohort of 96 individuals, suggesting that sample size can indeed be a drawback in our analyses.

Other general factors can be of influence on the results. Our sample set does not distinguish between type of disorder (e.g. neurological, mitochondrial, specific), sex or age. Besides the autoencoder and hyperparameter optimization methods that control for confounders, sensitivity could increase by using a more homogenous sample set, since a computational model will presumably not surpass true biological differences. In the literature it is mentioned that some biological confounding effects, such as co-regulation, cannot be entirely excluded even after controlling by the autoencoder.[16]

The MultiQC report demonstrated a good mean and per sequence quality score. We did encounter a high level of duplicated sequences. This reflects the fact that we left out the advised UMItools[40] deduplication in our 'RNA-Rare' pipeline setting. We did so because of incorrect settings and a lack of time. This method removes technical duplicates with corresponding UMIs. Integration of this method in our pipeline would probably improve the sensitivity as well.

Finally, for optimal robustness of the results, three replicates for each individual need to confirm the findings. We did use one replicate for individual 3, with similar results. Also, the positive control with a deletion in the CAD gene was confirmed for both individual 1 and 26 (child and one parent) as well as results for the two siblings (individual 8 and 28) with similar symptoms. This can be extended to three replicates for all samples.

In the coming years, we expect to expand our sample set to about 100 samples and UMI deduplication settings will be corrected, so the tool can be included in our pipeline. As previous reseach findings suggest, OUTRIDER analysis at exon-level will presumably yield better results and therefore increased diagnostic value when a larger sample size is used. Our findings however suggest that OUTRIDER analysis to detect aberrant exon expression will become at most additional, perhaps even unnecessary, compared to FRASER, since the latter produces a much higher positive predictive value and sensitivity. Although further testing needs to be performed to confirm this hypothesis when our sample size increases, and our pipeline will be improved with mentioned suggestions.

5 Conclusion

Our findings demonstrate promising results achieved for patients with monogenic disorders and yet inconclusive results from standard diagnostic testing with our 'RNA-Rare' pipeline. We were able to confirm aberrant gene expression and NMD with OUTRIDER outlier analysis and aberrant splicing events with FRASER analysis for all our positive controls. At least a three-fold improvement in positive predictive value was found for FRASER over OUTRIDER analysis in detecting aberrant splicing events. Furthermore, we found candidate genes for 9 out of 15 patients. Moreover, the OUTRIDER and FRASER modules can provide improvements to the existing 'rnaseq' and 'rnasplice' pipelines, Nextflow nf-core has to offer. We therefore propose 'rnarare' to be considered as a new nf-core option, with OUTRIDER and FRASER added to detect aberrant gene expression and splicing events, with a minimum sample size of 50-60. We suggest UMItools deduplication to be included in the workflow, as well as a more homogeneous sample set on gender, age and type of disorder.

6 Data and code availability

RNA-Rare pipeline

- UMC github DxNextflowRNA https://github.com/UMCUGenetics/DxNextflowRNA/tree/feature/add_outrider
- Nextflowmodules: OUTRIDER https://github.com/UMCUGenetics/NextflowModules/tree/3e59c78eb06c564b11292a66255c4bbc488edd51/Outrider/1.20.0
- Nextflowmodules: FRASER https://github.com/UMCUGenetics/NextflowModules/tree/feature/add_outrider/Fraser/1.99.3

UMC github Jupyter Notebook Voilà web-based app

- https://github.com/UMCUGenetics/rnaseq-voila/tree/feature/custom_settings_umcu

Scripts used for actions and analyses outside of the RNA-Rare pipeline and Jupyter Notebook app

- https://github.com/UMCUGenetics/rnaseq-voila/tree/feature/custom_settings_umcu/scripts

7 Supplementary Information

Supplementary Information can be found online at https://github.com/lonnekevanbrussel/Final-Master-Project/blob/main/

8 References

1. EURORDIS. Rare Diseases: Understanding this Public Health Priority. Rare Dis. 2005;1–14. https://www.eurordis.org/wp-content/uploads/2009/12/princeps_document-EN.pdf
2. Nguengang Wakap, S., Lambert, D.M., Olry, A. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 28, 165–173 (2020). https://doi.org/10.1038/s41431-019-0508-0
3. Alfares A, Aloraini T, Subaie LA, Alissa A, Qudsi AA, Alahmad A, Mutairi FA, Alswaid A, Alothaim A, Eyaid W, Albalwi M, Alturki S, Alfadhel M. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genet Med. 2018 Nov;20(11):1328-1333. doi: 10.1038/gim.2018.41. Epub 2018 Mar 22. PMID: 29565419.
4. Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, Vertino-Bell A, Smaoui N, Neidich J, Monaghan KG, McKnight D, Bai R, Suchy S, Friedman B, Tahiliani J, Pineda-Alvarez D, Richard G, Brandt T, Haverfield E, Chung WK, Bale S. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016 Jul;18(7):696-704. doi: 10.1038/gim.2015.148. Epub 2015 Dec 3. PMID: 26633542.
5. Boycott KM, Hartley T, Biesecker LG, Gibbs RA, Innes AM, Riess O, Belmont J, Dunwoodie SL, Jojic N, Lassmann T, Mackay D, Temple IK, Visel A, Baynam G. A Diagnosis for All Rare Genetic

Diseases: The Horizon and the Next Frontiers. Cell. 2019 Mar 21;177(1):32-37. doi: 10.1016/j.cell.2019.02.040. PMID: 30901545.

6. Ma M, Ru Y, Chuang LS, Hsu NY, Shi LS, Hakenberg J, Cheng WY, Uzilov A, Ding W, Glicksberg BS, Chen R. Disease-associated variants in different categories of disease located in distinct regulatory elements. BMC Genomics. 2015;16 Suppl 8(Suppl 8):S3. doi: 10.1186/1471-2164-16-S8-S3. Epub 2015 Jun 18. PMID: 26110593; PMCID: PMC4480828.
7. Dekker J, Schot R, Bongaerts M, de Valk WG, van Veghel-Plandsoen MM, Monfils K, Douben H, Elfferich P, Kasteleijn E, van Unen LMA, Geeven G, Saris JJ, van Ierland Y, Verheijen FW, van der Sterre MLT, Sadeghi Niaraki F, Smits DJ, Huidekoper HH, Williams M, Wilke M, Verhoeven VJM, Joosten M, Kievit AJA, van de Laar IMBH, Hoefsloot LH, Hoogeveen-Westerveld M, Nellist M, Mancini GMS, van Ham TJ.
Web-accessible application for identifying pathogenic transcripts with RNA-seq: Increased sensitivity in diagnosis of neurodevelopmental disorders. Am J Hum Genet. 2023 Feb 2;110(2):251-272. doi: 10.1016/j.ajhg.2022.12.015. Epub 2023 Jan 19. PMID: 36669495; PMCID: PMC9943747.
8. Rentas S, Rathi KS, Kaur M, Raman P, Krantz ID, Sarmady M, Tayoun AA. Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. Genet Med. 2020 May;22(5):927-936. doi: 10.1038/s41436-019-0741-5. Epub 2020 Jan 8. PMID: 31911672.
9. Murdock DR, Dai H, Burrage LC, Rosenfeld JA, Ketkar S, Müller MF, Yépez VA, Gagneur J, Liu P, Chen S, Jain M, Zapata G, Bacino CA, Chao HT, Moretti P, Craigen WJ, Hanchard NA; Undiagnosed Diseases Network; Lee B. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J Clin Invest. 2021 Jan 4;131(1):e141500. doi: 10.1172/JCI141500. PMID: 33001864; PMCID: PMC7773386.
10. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, Kao D, Ohri K, Viththiyapaskaran S, Tarnopolsky MA, Mathews KD, Moore SA, Osorio AN, Villanova D, Kemaladewi DU, Cohn RD, Brudno M, Dowling JJ. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am J Hum Genet. 2019 Mar 7;104(3):466-483. doi: 10.1016/j.ajhg.2019.01.012. Epub 2019 Feb 28. Erratum in: Am J Hum Genet. 2019 May 2;104(5):1007. doi: 10.1016/j.ajhg.2019.04.004. PMID: 30827497; PMCID: PMC6407525.
11. Yépez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, Ban R, Beblo S, Berutti R, Blessing H, Ciara E, Distelmaier F, Freisinger P, Häberle J, Hayflick SJ, Hempel M, Itkis YS, Kishita Y, Klopstock T, Krylova TD, Lamperti C, Lenz D, Makowski C, Mosegaard S, Müller MF, Muñoz-Pujol G, Nadel A, Ohtake A, Okazaki Y, Procopio E, Schwarzmayr T, Smet J, Staufner C, Stenton SL, Strom TM, Terrile C, Tort F, Van Coster R, Vanlander A, Wagner M, Xu M, Fang F, Ghezzi D, Mayr JA, Piekutowska-Abramczuk D, Ribes A, Rötig A, Taylor RW, Wortmann SB, Murayama K, Meitinger T, Gagneur J, Prokisch H. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. 2022 Apr 5;14(1):38. doi: 10.1186/s13073-022-01019-9. PMID: 35379322; PMCID: PMC8981716.
12. Dyle MC, Kolakada D, Cortazar MA, Jagannathan S. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated RNA decay. Wiley Interdiscip Rev RNA. 2020 Jan;11(1):e1560. doi: 10.1002/wrna.1560. Epub 2019 Jul 29. PMID: 31359616; PMCID: PMC10685860.
13. Baker, K. E.; Parker, R. (2004). "Nonsense-mediated mRNA decay: Terminating erroneous gene expression". Current Opinion in Cell Biology. 16 (3): 293–299. doi:10.1016/j.ceb.2004.03.003. PMID 15145354
14. Liana F. Lareau, Angela N. Brooks, David A.W. Soergel, Qi Meng and Steven E. Brenner The Coupling of Alternative Splicing and Nonsense-Mediated mRNA Decay http://compbio.berkeley.edu/people/brenner/pubs/lareau-2007-landes-nmd.pdf
15. Müller, Franz; Ackermann, Peter; Margot, Paul (2012). "Fungicides, Agricultural, 2. Individual Fungicides". Ullmann's Encyclopedia of Industrial Chemistry. Weinheim: Wiley-VCH. doi:10.1002/14356007.o12_o06. ISBN 978-3527306732.
16. Brechtmann F*, Mertes C*, Matuseviciute A*, Yepez V, Avsec Z, Herzog M, Bader D M, Prokisch H, Gagneur J; OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data; AJHG; 2018; DOI: https://doi.org/10.1016/j.ajhg.2018.10.025
17. Scheller I, Lutz K, Mertes C, et al. Improved detection of aberrant splicing with FRASER 2.0 using the Intron Jaccard Index, medrXiv, 2023, https:// doi.org/ 10.1101/ 2023.03.31.23287997
18. P. Di Tommaso, et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017) doi:10.1038/nbt.3820
19. The nf-core framework for community-curated bioinformatics pipelines. Philip Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen.Nat Biotechnol. 2020 Feb 13. doi: 10.1038/s41587-020-0439-x.
20. Ignatiadis, N., Klaus, B., Zaugg, J.B., Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nature Methods, 13:7. 10.1038/nmeth.3885
21. Chen Y, Chen L, Lun ATL, Baldoni P, Smyth GK (2024). “edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets.” bioRxiv. doi:10.1101/2024.01.21.576131.
22. Anders S, Reyes A, Huber W (2012). “Detecting differential usage of exons from RNA-seq data.” Genome Research, 22, 4025. doi:10.1101/gr.133744.111.
23. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), August 2024. World Wide Web URL: https://omim.org/
24. Kluyver, T. et al., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90.
25. Schneider VA, Graves-Linsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017May;**27**(5):849–64. doi: https://doi.org/10.1101/072116
26. Frankish A, Carbonell-Sala S, Diekhans M, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res*. 2023;51(D1):D942-D949. doi:10.1093/nar/gkac1071
27. Center for Quantitative Life Sciences, Oregon State University. https://software.cqls.oregonstate.edu/updates/trim_galore-0.6.7/#trim_galore-067

28. van de Geijn, B., McVicker, G., Gilad, Y. et al. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods 12, 1061–1063 (2015). https://doi.org/10.1038/nmeth.3582
29. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. doi: 10.1093/gigascience/giab008. PMID: 33590861; PMCID: PMC7931819.
30. Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general pur- pose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30. http://www.ncbi.nlm.nih.gov/pubmed/24227677
31. Andrews S (2010). FastQC: a quality control for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
32. Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354
33. Durinck S, Spellman P, Birney E, Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols, 4, 1184–1191
34. Yépez, V.A., Mertes, C., Müller, M.F. et al. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc 16, 1276–1296 (2021). https://doi.org/10.1038/s41596-020-00462-5
35. Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R. S. Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, Paul Flicek. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics 2014 31: 143-14. doi: 10.1093/bioinformatics/btu613
36. Robinson, J., Thorvaldsdóttir, H., Winckler, W. et al. Integrative genomics viewer. Nat Biotechnol 29, 24–26 (2011). https://doi.org/10.1038/nbt.1754
37. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses (PMID: 27322403; Citations: 3,306) Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5
38. Shane C. Quinonez, Jeffrey W. Innis,Human HOX gene disorders, Molecular Genetics and Metabolism, Volume 111, Issue 1, 2014, Pages 4-15, ISSN 1096-7192, https://doi.org/10.1016/j.ymgme.2013.10.012.
39. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. Annu Int Conf IEEE Eng Med Biol Soc. 2015; 2015: 6461-4. doi: 10.1109/EMBC.2015.7319872. PMID: 26737772; PMCID: PMC4983442.
40. Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. March 2017 27: 491-499; Published in Advance January 18, 2017, doi:10.1101/gr.209601.116
41. Rodriguez JM, Pozo F, di Domenico T, Vazquez J, Tress ML. An analysis of tissue-specific alternative splicing at the protein level. PLoS Comput Biol. 2020;16(10):e1008287. Published 2020 Oct 5. doi: 10.1371/journal.pcbi.1008287

9 Web resources

42. Nextflow documentation, https://www.nextflow.io/docs/latest/index.html
43. Nextflow training, https://training.nextflow.io/basic_training/
44. Nf-core/rnaseq, https://nf-co.re/rnaseq/3.14.0
45. FeatureCounts https://manpages.debian.org/testing/subread/featureCounts.1.en.html
46. OUTRIDER vignette, https://www.bioconductor.org/packages/release/bioc/vignettes/OUTRIDER/inst/doc/OUTRIDER.pdf,
47. OUTRIDER gagneurlab pipeline https://github.com/gagneurlab/drop/tree/cc2ee6b9975335b7104565ac108ac1ff8456038f/drop/modules/aberrant-expression-pipeline
48. FRASER vignette, https://bioconductor.org/packages/release/bioc/vignettes/FRASER/inst/doc/FRASER.pdf,
49. FRASER gagneurlab pipeline https://github.com/gagneurlab/drop/tree/cc2ee6b9975335b7104565ac108ac1ff8456038f/drop/modules/aberrant-splicing-pipeline
50. Voilà app, https://github.com/voila-dashboards/voila