# Genetic diagnosis of Mendelian disorders via RNA sequencing

Laura S Kremer[1,2,21], Daniel M Bader[3,4,21], Christian Mertes[3], Robert Kopajtich[1,2], Garwin Pichler[5], Arcangela Iuso[1,2], Tobias B Haack[1,2], Elisabeth Graf[1,2], Thomas Schwarzmayr[1,2], Caterina Terrile[1], Eliška Koňaříková[1,2], Birgit Repp[1,2], Gabi Kastenmüller[6], Jerzy Adamski[7], Peter Lichtner[1], Christoph Leonhardt[8], Benoit Funalot[9], Alice Donati[10], Valeria Tiranti[11], Anne Lombes[12,13,14], Claude Jardel[12,15], Dieter Gläser[16], Robert W. Taylor[17], Daniele Ghezzi[11], Johannes A Mayr[18], Agnes Rötig[8], Peter Freisinger[19], Felix Distelmaier[20], Tim M Strom[1,2], Thomas Meitinger[1,2], Julien Gagneur[3,4,*], Holger Prokisch[1,2,*].

1. Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany
2. Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, 81675 München, Germany
3. Department of Informatics, Technische Universität München, 85748 Garching, Germany
4. Quantitative Biosciences Munich, Gene Center, Department of Biochemistry, Ludwig Maximilian Universität München, 81377 Munich, Germany
5. Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, 82152 Martinsried, Germany
6. Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
7. Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany
8. Neuropädiatrie, Neonatologie, 78050 Villingen-Schwenningen, Germany
9. INSERM U1163, Université Paris Descartes - Sorbonne Paris Cité, Institut Imagine, 75015 Paris, France
10. Metabolic Unit, A. Meyer Children's Hospital, Florence, Italy
11. Unit of Molecular Neurogenetics, Foundation IRCCS (Istituto di Ricovero e Cura a Carettere Scientifico) Neurological Institute “Carlo Besta”, 20126 Milan, Italy
12. Inserm UMR 1016, Institut Cochin, 75014 Paris, France
13. CNRS UMR 8104, Institut Cochin, 75014 Paris, France
14. Université Paris V René Descartes, Institut Cochin, 75014 Paris, France
15. AP/HP, GHU Pitié-Salpêtrière, Service de Biochimie Métabolique, 75013, Paris, France
16. Genetikum, Genetic Counseling and Diagnostics, 89231 Neu-Ulm, Germany
17. Wellcome Trust Centre for Mitochondrial Research, Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK
18. Department of Pediatrics, Paracelsus Medical University, A-5020 Salzburg, Austria
19. Department of Pediatrics, Klinikum Reutlingen, 72764 Reutlingen, Germany
20. Department of General Pediatrics, Neonatology and Pediatric Cardiology, University Children's Hospital, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany
21. These authors contributed equally to this work.

* Correspondence should be addressed to Holger Prokisch (prokisch@helmholtz-muenchen.de) or Julien Gagneur (gagneur@in.tum.de).

## *Abstract*

**Across a large variety of Mendelian disorders, ~50-75% of patients do not receive a genetic diagnosis by whole exome sequencing indicative of underlying disease-causing variants in non-coding regions. In contrast, whole genome sequencing facilitates the discovery of all genetic variants, but their sizeable number, coupled with a poor understanding of the non-coding genome, makes their prioritization challenging. Here, we demonstrate the power of transcriptome sequencing to provide a confirmed genetic diagnosis for 10% (5 of 48) of undiagnosed mitochondrial disease patients and identify strong candidate genes for patients remaining without diagnosis. We found a median of 1 aberrantly expressed gene, 5 aberrant splicing events, and 6 mono-allelically expressed rare variants in patient-derived fibroblasts and established disease-causing roles for each kind. Private exons often arose from sites that are weakly spliced in other individuals, providing an important clue for future variant prioritization. One such intronic exon-creating variant was found in three unrelated families in the complex I assembly factor TIMMDC1, which we consequently established as a novel disease-associated gene. In conclusion, our study expands the diagnostic tools for detecting non-exonic variants of Mendelian disorders and provides examples of intronic loss-of-function variants with pathological relevance.**

Despite the revolutionizing impact of whole exome sequencing (WES) on the molecular genetics of Mendelian disorders, ~50-75% of the patients do not receive a genetic diagnosis after WES[1–6]. The disease-causing variants might be detected by WES but remain as variants of unknown significance (VUS, Methods) or they are missed due to the inability to prioritize them. Many of these VUS are synonymous or non-coding variants that may affect RNA abundance or isoform but cannot be prioritized due to the poor understanding of regulatory sequence to date compared to coding sequence. Furthermore, WES covers only the 2% exonic regions of the genome. Accordingly, it is mostly blind to regulatory variants in non-coding regions that could affect RNA sequence and abundance. While the limitation of genome coverage is overcome by whole genome sequencing (WGS), prioritization and interpretation of variants identified by WGS is in turn limited by their amount[7–9].

With RNA sequencing (RNA-seq), limitations of the sole genetic information can be complemented by directly probing variations in RNA abundance and in RNA sequence, including allele-specific expression and splice isoforms. At least three extreme situations can be directly interpreted to prioritize candidate disease-causing genes for a rare disorder. First, the expression level of a gene can lie outside its physiological range. Genes with expression outside their physical range can be identified as expression outliers, often using a stringent cutoff on expression variations, for instance using the Z-score[10] or statistics at the level of whole gene sets[11,12]. The genetic causes of such aberrant expression includes rare variants in the promoter[13] and enhancer but also in coding or intronic regions[10]. Second, RNA-seq can reveal extreme cases of allele-specific expression (mono-allelic expression), whereby one allele is silenced, leaving only the other allele expressed. When assuming a recessive mode of inheritance, genes with a single heterozygous rare coding variant identified by WES or WGS analysis are not prioritized. However, mono-allelic expression of such variants fits the recessive mode of inheritance assumption. Detection of mono-allelic expression can thus help re-prioritizing heterozygous rare variants. Reasons for mono-allelic expression can be genetic. A pilot study validated compound heterozygous variants within one gene as cause of TAR syndrome, where one allele is deleted and the other harbors a non-coding variant that reduces expression[14]. Mono-allelic expression can also have epigenetic causes such as X-chromosome inactivation or imprinting on autosomal genes, possibly by random choice[15,16]. Third, splicing of a gene can be affected. Aberrant splicing has long been recognized as a major cause of Mendelian disorders (reviewed in ref. [17–19]). However, the prediction of splicing defects from genetic sequence is difficult because splicing involves a complex set of cis-regulatory elements that are not yet fully understood. Some of them can be deeply located in intronic sequences[20] and are thus not covered by WES. Hence, direct probing of splice isoforms by RNA-seq is important, and has led to the discovery of multiple splicing defects based on single gene studies: skipping of multiple exons (exon 45-55)[21] and creation of a new exon by a deep intronic variant in *DMD*[22], intron retention in *LMNA* caused by a 5' splice site variant[23], and skipping of exon 7 in *SMN1* caused by a variant in a splicing factor binding site[24]. Altogether, RNA-seq promises to be an important complementary tool to facilitate molecular diagnosis of rare genetic disorders. However, no systematic study to date has been conducted to assess its power.

Here, we established an analysis pipeline to systematically detect instances of i) aberrant expression, ii) aberrant splicing, and iii) mono-allelic expression of the alternative allele to complement whole exome sequencing based genetic diagnosis. We considered applying our approach on patients diagnosed with a mitochondrial disorder for three reasons. First, mitochondrial diseases collectively represent one of the most frequent inborn errors of metabolism affecting 2 in 10,000 individuals[25]. Second, the broad range of unspecific clinical symptoms and the genetic diversity in mitochondrial diseases makes molecular diagnosis difficult and WES often results in variants of unknown significance. As a consequence of the bi-genomic control of the energy-generating oxidative phosphorylation (OXPHOS) system, mitochondrial diseases may result from pathogenic mutations of the mitochondrial DNA (mtDNA) or nuclear genome. More than 1,500 different nuclear genes encode mitochondrial proteins[26] and causal defects have been identified in approximately 300 genes and presumably more additional disease-associated genes still awaiting identification[27]. Third, since the diagnosis often relies on biochemical testing of a tissue sample, fibroblast cell lines are usually available from those patients. Moreover, for many patients, the disease mechanisms can be assayed in epidermal fibroblast cell lines even though the disease may manifest in different tissues[28]. This allows rapid demonstration of the necessary and sufficient role of candidate variants by perturbation and complementation assays[29]. This also indicates that disease-causing expression defects, if any, should be detectable in these cell lines.

## Results

We performed RNA-seq on 105 fibroblast cell lines from patients with a suspected mitochondrial disease including 48 patients for which whole exome sequencing based variant prioritization did not yield a genetic diagnosis (Fig. 1, Methods). After discarding lowly expressed genes, RNA-seq identified 12,680 transcribed genes (at least 10 reads in 5% of all samples, Methods, Supplementary Data 1). We systematically prioritized genes with the following three strategies: i) genes with aberrant expression level[11–13], ii) genes with aberrant splicing[22,30], and iii) mono-allelic expression of rare variants[14] (Fig. 1) to estimate their disease association. All strategies are based on the comparison of one patient against the rest. We assumed the causal defects to differ between patients, which is reasonable for mitochondrial disorders with a diversity of ~300 known disease-causing genes (Supplementary Data 2). Therefore, the patients serve as good controls for each other.

Once normalized for technical biases, sex, and biopsy site (Supplementary information and Supplementary Fig. 1), the samples typically presented few aberrantly expressed genes (median of 1, Fig. 2a, Supplementary Table 1) with a large effect ($|\text{Z-score}| > 3$) and significant differential expression (Hochberg adjusted $P$-value $< 0.05$). Among the most aberrantly expressed genes across all samples, we found 2 genes encoding mitochondrial proteins, *MGST1* (one case) and *TIMMDC1* (two cases) to be significantly down-regulated (Fig. 2b-d and Supplementary Fig. 2). For both genes, WES did not identify any variants in the respective patients, no variant is reported to be disease-associated, and no case of potential bi-allelic rare variant is listed in our in-house database comprising more than 1,200 whole-exomes from mitochondrial patients and 15,000 WES dataset available to us from different ongoing research projects. To evaluate the consequences of diminished RNA expression at the protein level, we performed quantitative proteomics in a total of 31 fibroblast cell lines (including these three patients, and further 17 undiagnosed and 11 diagnosed patients, Methods, Supplementary Table 2, Supplementary Data 3) from a second aliquot of cells taken at the same time as the RNA-seq aliquot. Normalized RNA and protein expression levels showed a median rank correlation of 0.59, comparable to what has been previously reported[31,32] (Supplementary Fig. 3). Patient #73804 showed ~2% of control MGST1 level whilst the lack of detection of TIMMDC1 in both patients (#35791 and #66744) confirmed an even stronger effect on protein expression, indicating loss of function (Fig. 2e and Supplementary Fig. 4). MGST1, a microsomal glutathione S-transferase, is involved in the oxidative stress defense[33]. Consequently, the loss of expression of MGST1 is not only a likely cause of the disease of this patient, who suffers from an infantile-onset neurodegenerative disorder similar to a recently published case with another defect in the reactive oxygen species (ROS) defense system (Supplementary Information)[34], but also suggests a treatment with antioxidants. Both TIMMDC1 patients presented with muscular hypotonia, developmental delay, and neurological deterioration, which led to death in the first 3 years of life (Supplementary Information). Consistent with the described function of TIMMDC1 as a respiratory chain complex I assembly factor[35,36], we found isolated complex I deficiency in muscle (Supplementary Fig. 2), and globally decreased levels of complex I subunits in fibroblasts by quantitative proteomics (Fig. 2e and Supplementary Fig. 2) and western blot (Fig. 2f). Re-expression of *TIMMDC1* in these cells increased complex I subunit levels (Fig. 2f). These results not only validate TIMMDC1-deficiency as disease causing but also provide compelling evidence for an important function of TIMMDC1 in complex I assembly.

We identified aberrant splicing events by testing for differential splicing in each patient against the others, using an annotation-free algorithm able to detect also novel splice sites (Methods, median of 5 abnormal events per sample, Fig. 3a). Among the 175 aberrant spliced genes detected in the undiagnosed patients, the most abundant events were, apart from differential expression of isoforms, exon skipping followed by the creation of new exons (Fig. 3b). Two genes encoding mitochondrial proteins, *TIMMDC1* and *CLPP,* which were among the 20 most significant genes, caught our attention (Supplementary Table 3). Out of 136 exon-junction reads overlapping the acceptor site of *CLPP* exon 6 for patient #58955, 82 (percent spliced in[37] $\Psi = 60\%$) skipped exon 5, and 14 ($\Psi = 10\%$) showed a 3'-truncated exon 5, in striking contrast to other samples (Fig. 3c). The likely genetic cause of these two splice defects is a rare homozygous variant in exon 5 of *CLPP* affecting the last nucleotide of exon 5 (c.661G>A, p.Glu221Lys $1.2\times10^{-5}$frequency in the ExAC database[38], Supplementary Fig. 5). Both detected splice defects result in truncated CLPP and western blots corroborated the complete loss of full-length CLPP (Supplementary Fig. 5). Our WES variant filtering reported this variant as a VUS and classified *CLPP* as one among 30 other potentially bi-allelic affected candidate genes (Supplementary Table 1). Since the variant was of unknown significance, the patient remained without genetic diagnosis. The loss of function found by RNA-seq and confirmed by Western blotting now highlights clinical relevance of the variant within *CLPP*. *CLPP* encodes a mitochondrial ATP-dependent endopeptidase[39] and CLPP-deficiency causes Perrault syndrome[40,41] (OMIM #601119) which is overlapping with the clinical presentation of the patient investigated here including microcephaly, deafness, and severe psychomotor retardation (Supplementary Information). Moreover, a study recently showed that Clpp-/- mice are deficient for complex IV expression[42], in line with complex IV deficiency of this patient (Supplementary Fig. 5).

Split read distribution indicated that both TIMMDC1-deficient patients expressed almost exclusively a *TIMMDC1* isoform with a new exon in intron 5 (Fig. 3d). This new exon introduces a frameshift yielding a premature stop codon (p.Gly199_Thr200ins5*, Fig. 3e). Moreover, this new exon contained a rare variant (c.596+2146A>G) not listed in the 1,000 Genomes Project[7,8]. Whole genome sequencing demonstrated that this variant is homozygous in both patients (Fig. 3e), the only rare variant in this intron and close to the splice site (+6 of the new exon). We could not identify any rare variant in the promoter region or in any intron-exon boundary of *TIMMDC1*. Additionally, when testing six prediction tools for splicing events, this deep intronic rare variant is predicted by SpliceAid2[43] to create multiple binding sites for splice enhancers. Together with the correctly predicted new acceptor and donor sites by SplicePort[44] (Feature generation algorithm score 0.112 and 1.308, respectively) this emphasizes the influence of this variant in the creation of the new exon. Besides, the four other tools predicted no significant change in splicing[45–48]. We further discovered an additional family in our in-house WGS database (consisting of 36 patients with a suspected mitochondrial disorder and 232 further patients with unrelated diseases) carrying the same homozygous intronic variant. In this family three affected siblings presented with similar clinical symptoms although without a diagnosis of a mitochondrial disorder (Fig. 3e, Supplementary Fig. 2). Two siblings died before the age of 10. A younger brother (#96687), now 6 years of age, presented with muscle hypotonia, failure to thrive and neurological impairment (Supplementary Information), similar to the patients described above. Western blot analysis confirmed TIMMDC1-deficiency (Fig. 2f) and impaired complex I assembly, which was restored after re-expression of *TIMMDC1* (Fig. 2g). The discovery of the same intronic *TIMMDC1* variant in three unrelated families from three different ethnicities provides convincing evidence on the causality of this variant for the TIMMDC1 loss-of-function.

In almost all non-TIMMDC1-deficiency samples, we noticed a few split reads supporting inclusion of the new exon (Fig. 3d), consistent with an earlier report that many cryptic splice sites are not entirely repressed but active at low levels[49]. We set out to quantify this phenomenon and to interrogate the frequency of private exons originating from weakly spliced exons, independent of their possible association with disease. Consequently, we modeled the distribution of $\Psi$ for the 1,603,042 splicing events detected genome-wide in 105 samples as a mixture of three components. The model classified splicing frequencies per splice site as strong (20%, with $\Psi > 5.3\%$), weak (16%, with $0.16\% < \Psi < 5.3\%$), or background (64%, with $\Psi < 0.16\%$, Methods, Fig. 3f and Supplementary Fig. 6). Strikingly, the majority (70%, 4.4-fold more than by chance) of the 17 discovered private exons originated from weak splice sites (Fig. 3f bottom). These data confirm that weakly spliced cryptic exons are loci more susceptible to turn into strongly spliced sites than other intronic regions. These weak splicing events are usually dismissed as ’noise’ since they are only supported by few reads in a given sample. Our analysis shows that they can be detected as accumulation points across multiple individuals. Hence, these results suggest that the prioritization of deep intronic variants of unknown significances gained through WGS could be improved by annotating weak splice sites and their resulting cryptic exons.

As a third approach, we searched for mono-allelic expression (MAE) of rare variants. In median per sample, 35,521 heterozygous SNVs were detected by WES, of which 7,622 were sufficiently covered by RNA-seq to call MAE (more than 10 reads), 20 showed MAE (Hochberg adjusted $P$-value < 0.05, allele frequency ≥ 0.8), of which 6 were rare variants (minor allele frequency < 0.001, Methods, Fig. 4a). Amongst the 18 rare mono-allelic expressed variants in patient #80256 was a VUS in *ALDH18A1* (c.1864C>T, p.Arg622Trp, Fig. 4b), encoding an enzyme involved in mitochondrial proline metabolism[50]. This VUS had been seen in WES compound heterozygous with a nonsense variant (c.1988C>A, p.Ser663*, Fig. 4b and Supplementary Fig. 7). Variants in *ALDH18A1* had been reported to be associated with cutis laxa III (OMIM #138250)[51,52], yet the patient did not present cutis laxa. Because of this inconsistent phenotype and the unknown significance of the non-synonymous variant, the variants in *ALDH18A1* were not regarded as disease causing. However, RNA-seq-based aberrant expression (Supplementary Fig. 7) and mono-allelic expression analysis prioritized *ALDH18A1* again. Our systematically performed validation by quantitative proteomics revealed severe reduction down to ~2% ALDH18A1 (Fig. 4c), indicating that the rare MAE variant affects translation or protein stability. Metabolomics profile of blood plasma was in accordance with a defect in proline metabolism (Fig. 4d) and the following changes in urea cycle. Patient fibroblasts showed a growth defect that was rescued by supplementation of proline, validating impaired proline metabolism as the underlying molecular cause (Fig. 4e). Our experimental evidence strongly suggests that the two observed variants are causal. Moreover, a recent report[53] on *ALDH18A1* patients extended the phenotypic spectrum to spastic paraplegia (OMIM #138250), which resembles the symptoms of our patient (Supplementary information).

In another patient (#62346) we found borderline non-significant low expression of *MCOLN1* with 10 of 11 reads expressing an intronic VUS (c.681-19A>C, Fig. 4f). This intronic variant was detected as part of a retained intron, which introduced a nonsense codon (p. Lys227_Leu228ins16*, Fig. 4f and Supplementary Fig. 8). When looking at the WES data we could additionally identify a heterozygous nonsense variant (c.832C>T, p.Gln278*). The allele with the exonic nonsense mutation was not expressed, most likely due to nonsense-mediated decay. Mutations in *MCOLN1* are associated with mucolipidosis (OMIM #605248). The symptoms of the patient were initially suggestive for mucolipidosis, but none of the enzymatic tests available for mucolipidosis type 1, 2, and 3 revealed an enzyme deficiency in blood leukocytes (Supplementary information). Moreover, *MCOLN1* was missed by our WES variant filter since the intronic variant was not prioritized. Hence, the WES data could not be conclusive about *MCOLN1*. In contrast, the RNA-seq data demonstrated two loss-of-function alleles in *MCOLN1* and therefore established the genetic diagnosis.

## Discussion

Altogether, our study demonstrates the utility of RNA sequencing in combination with bioinformatics filtering criteria for genetic diagnosis by i) discovering a new disease-associated gene, ii) providing a diagnosis for 10% (5 of 48) of undiagnosed cases, and iii) identifying a limited number of strong candidates. We established a pipeline for the detection of aberrant expression, aberrant splicing and mono-allelic expression of rare variants, that is able to detect significant outliers, i.e. a median of 1, 5, and 6, respectively. Overall, for 36 patients our pipeline provides a strong candidate gene, i.e. a known disease-causing or mitochondrial protein-encoding gene, like *MGST1* (Fig 5a, Supplementary Table 1). This manageable amount, similar to the median number of 16 genes with rare potentially bi-allelic variants detected by WES, allows manual inspection and validation by disease experts. While filtering by frequency is highly efficient when focusing on the coding region, frequency filtering is not as effective for intronic or intergenic variants identified by WGS. The loss-of-function character observed on RNA level thus improved interpretation of VUS identified by genotyping.

We focused our analysis on one sample preparation pipeline, which has several advantages. Based on our experience, expression outliers can only reliably be detected after extensive normalization process. This needs information about all technical details starting from the biopsy, growth of the cells, to the RNA extraction and library preparation. Usually not all this information is available in published data sets. For detecting aberrant splicing such as new exons, we would recommend not to mix different tissues because splicing can be tissue-specific. Mono-allelic expression is the most robust of all criteria in this respect because it only relies on read counts within a sample. Overall, we recommend not relying on a single sample being compared to public RNA-seq datasets. Instead, RNA-seq should be included in the pipeline of diagnostic centers in order to generate matching controls over time. The situation is similar for whole exome and whole genome sequencing, where the control for platform-specific biases is important.

Here, we included genetically diagnosed patients in our RNA-seq analysis pipeline to increase the power for the detection of aberrant expression and aberrant splicing in fibroblast cell lines. However, when applied to the 40 diagnosed cases with WES and RNA-seq available, aberrant splicing detected 6 out of 8 cases with a causal splicing variant, mono-allelic expression recovered 3 out of 6 patients with heterozygous missense variants compound with a stop or frameshift variant, and aberrant expression recovered 3 out of 9 stop variants. Counterintuitively, only one of the 9 frame-shift variants did lead to a detectable RNA defect, i.e. mono-allelic expression of a near splice site intronic variant within a retained intron. The partial recovery of stop and frameshift variants may reflect incomplete non-sense mediated decay. For none of the 14 genes where missense variants were disease causing, a RNA defect could be detected with our pipeline. This is expected, since missense variants more likely affect protein function rather than RNA expression (Supplementary table 4).

To our surprise, many newly diagnosed cases were caused by a defective splicing event, which caused loss of function (Fig 5b), confirming the increasing role of splicing defects in both Mendelian[54,55] and common disorders[30]. In the case of *TIMMDC1*, the causal variant was intronic, and thus not covered by WES. Even when detected by WGS, such deep intronic variants are difficult to prioritize from the sequence information alone. Here, we showed that RNA-seq of large cohorts can provide important information about intronic positions that are particularly susceptible to affect splicing when mutated. We showed that private exons often arise from loci with weak splicing of about 1%. This suggests that rare variants affecting such cryptic splice sites are more likely to affect splicing and that these can be detected as positions with low yet consistent splicing. We reason that analysis of a RNA-seq compendium of healthy donors across multiple tissues such as GTEx[56] could provide tissue-specific maps of cryptic splice sites useful for prioritizing intronic variants.

Genetic disorders typically show specificity to some tissues, some of which might not be easily accessible for RNA-sequencing. It is therefore natural to question whether transcriptome sequencing of an unaffected tissue can help diagnosis. Here, we performed RNA-seq on patient derived dermal fibroblast cell lines. The fibroblast cell lines are the byproducts of muscle biopsies routinely undertaken in the clinic to biochemically diagnose mitochondrial disorders with enzymatic assays. By using fibroblast cell lines we overcome the limited accessibility of the affected tissues, which in the case of mitochondrial disorders are often high energy demanding tissues like brain, heart, skeletal muscle or liver. It turns out that many genes with a mitochondrial function are expressed in most tissues[57], including fibroblasts. Hence, extreme regulatory defects such as loss of expression or aberrant splicing of genes encoding mitochondrial proteins can be detected in fibroblasts, even though the physiological consequence on fibroblasts might be negligible. This property might be true for other diseases: the tissue-specific physiological consequence of a variant does not necessarily stem from tissue-specific expression of the gene harboring the variant. In many cases, tissue-specificity might be due to environmental or cellular context, or from tissue-specific expression of further genes. Hence, tissue-specificity does not preclude RNA-seq of unaffected tissues from revealing the causative defect for a large number of patients. Moreover, non-affected tissues have the advantage that the regulatory consequences on other genes are limited and therefore the causal defects are more likely to stand out as outliers[58].

Parallel to our effort, another study systematically investigated the usage of RNA-seq for molecular diagnosis with a similar sample size, using muscle biopsies from rare neuromuscular disease patients[55]. Analogously to our approach, not only exome sequencing-based VUS candidates were validated, but also new disease-causing mechanisms identified using RNA-seq data. Despite a few differences in the approach (expression outliers were not looked for, only samples of the affected tissues were considered and using samples of healthy donors as controls), the results are in line with ours whereby aberrant splicing also turns out to be a frequent disease-causing event. Moreover, the success rate was even higher (35%) confirming the relevance of using RNA-seq for diagnosis of Mendelian disorders.

In conclusion, we predict that RNA sequencing will become an essential companion of genome sequencing to address undiagnosed cases of genetic disease.

# *Methods*

## *Exome sequencing*

Exome sequencing was essentially performed as described previously[59]. In brief, exonic regions were enriched using the SureSelect Human All Exon kit from Agilent (Supplemental Data 4) followed by sequencing as 100 bp paired-end runs on an Illumina HiSeq2000 and Illumina HiSeq2500 (AG_50MB_v4 and AG_50MB_v5 exome kit samples) or as 76 bp paired-end runs on the Illumina GAIIx (AG_38MB_v1 and AG_50MB_v3 exome kit samples).

## *Exome alignment and variant prioritization*

Read alignment to the human reference genome (UCSC Genome Browser build hg19) was done using Burrows-Wheeler Aligner[60] (v.0.7.5a). Single-nucleotide variants and small insertions and deletions (indels) were detected with SAMtools[61,62](version 0.1.19). Variants with a quality score below 90, a genotype quality below 90, a mapping quality below 30, and a read coverage below 10 were discarded. The reported variants and small indels were annotated with the most severe entry by the Variant Effector Predictor[63] based on The Sequence Ontology term ranking[64]. The candidate variants for one patient are filtered to be rare, affect the protein sequence and potentially both alleles.

Variants are rare with a minor allele frequency < 0.001 within the ExAC database[38] and a frequency < 0.05 among our samples. Variants affect the protein, if they are a coding structural variant or their mutation type is one of *ablation*, *deletion*, *frame-shift*, *incomplete*, *start lost*, *insertion*, *missense*, *splice*, *stop gain*, *stop retain*, *unstart*, *unstop*. A potential biallelic effect can be caused by either a homozygous or at least two heterozygous variants in the same gene, whereas in latter case we assume that the heterozygous variants are on different alleles (Supplementary Fig. 9). This filter is designed for a recessive type disease model and does not account for a single heterozygous variant that could be disease-causing in a dominant way.

## *Variant of unknown significance*

"A variation in a genetic sequence whose association with disease risk is unknown. Also called unclassified variant, variant of uncertain significance, and VUS." (see https://www.cancer.gov/publications/dictionaries/genetics-dictionary?cdrid=556493 )

## *Cell culture*

Primary patient fibroblast cell lines, normal human dermal fibroblasts (NHDF) from neonatal tissue (Lonza), and 293FT cells (Thermo Fisher Scientific) were cultured in high glucose DMEM (Life Technologies) supplemented with 10% FBS, 1% penicillin/streptomycin, and 200 µM uridine at 37 °C and 5% $CO_2$. All fibroblast cell lines have been tested negative for mycoplasma contamination.

### *RNA sequencing*

Non-strand specific, polyA-enriched RNA sequencing was performed as described earlier[28]. Briefly, RNA was isolated from whole-cell lysates using the AllPrep RNA Kit (Qiagen) and RNA integrity number (RIN) was determined with the Agilent 2100 BioAnalyzer (RNA 6000 Nano Kit, Agilent). For library preparation, 1 µg of RNA was poly(A) selected, fragmented, and reverse transcribed with the Elute, Prime, Fragment Mix (Illumina). End repair, A-tailing, adaptor ligation, and library enrichment were performed as described in the Low Throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina). RNA libraries were assessed for quality and quantity with the Agilent 2100 BioAnalyzer and the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). RNA libraries were sequenced as 100 bp paired-end runs on an Illumina HiSeq2500 platform.

### *Processing of RNA sequencing files*

RNA-seq reads were demultiplexed and mapped with STAR[65] (version 2.4.2a) to the hg19 genome assembly (UCSC Genome Browser build). In addition to the default parameters we detected gene fusions and increased sensitivity for novel splice junctions (chimSegmentMin=20, twopassMode=”Basic”). Analysis was restricted to the 27,682 UCSC Known Genes[66] (genome annotation version hg19) of chromosomes 1 to 22, M, X, or Y. Per gene, reads that are paired with mates from opposite strands and that overlapped completely within the gene on either strand orientation were counted using the *summarizeOverlaps* function of the R/Bioconductor GenomicAlignments[67] package (parameters: mode=intersectionStrict, singleEnd=FALSE, ignore.strand=TRUE, fragments=FALSE). If the 95th percentile of the coverage across all samples was below 10 reads the gene was considered “not expressed” and discarded from later analysis.

### *Computing RNA fold changes and differential expression*

Before testing for differential expression between one patient of interest and all others, we controlled for technical *batch effect*, *sex*, and biopsy site as inferred from the expression of *hox* genes (Supplementary information, Supplementary Data 1). We modeled the RNA-seq read counts $K_{i,j}$ of gene $i$ in sample $j$ with a generalized linear model using the R/Bioconductor DESeq2 package[68,69]:

$$K_{i,j} \sim NB\left(s_j \times q_{i,j}, \alpha_i\right)$$

$$\log_2(q_{i,j}) = \beta_i^0 + \beta_i^{condition}\mathbf{x}_{i,j}^{condition} + \beta_i^{batch}\mathbf{x}_{i,j}^{batch} + \beta_i^{sex}\mathbf{x}_{i,j}^{sex} + \beta_i^{hox}\mathbf{x}_{i,j}^{hox}$$

Where NB is the negative binomial distribution; $\alpha_i$is a gene specific dispersion parameter; $s_j$ is the size factor of sample $j$; $\beta_i^0$ is the intercept parameter for gene $i$. The value of $\mathbf{x}_{i,j}^{condition}$ is 1 for all RNA samples $j$ of the patient of interest, thereby allowing for biological replicates, and 0 otherwise. The resulting vector $\beta_i^{condition}$ represents the $\log_2$-fold changes for one patient against all others. Z-scores were computed by dividing the fold changes by the standard deviation of the normalized expression level of the respective gene. The *P*-values corresponding to the

$\beta_i^{condition}$ were corrected for multiple testing using the Hochberg family-wise error rate method[70].

### Detection of aberrant splicing

The LeafCutter[71] software was utilized to detect aberrant splicing. Each patient was tested against all others. To adjust LeafCutter to the rare disease setting, we modified the parameters to detect rare clusters, capture local gene fusion events and to detect junctions unique to a patient (minclureads=30; maxintronlen=500,000; mincluratio=1e-5, Supplementary Data 5). Furthermore, one sample was tested against all other samples (min_samples_per_group=1; min_samples_per_intron=1). The resulting *P*-values were corrected for multiple testing using a family-wise error rate approach[70].

The significant splice events (Hochberg adjusted *P*-value < 0.05) detected in the undiagnosed patients were visually classified as exon skipping, exon truncation, exon elongation, new exon, complex splicing (any other splicing event or a combination of the aforementioned ones) and false positives (n=73, Fig 3b). However, due to LeafCutter's restriction to split reads it is difficult to detect intron retention events, since in a perfect intron retention scenario no split-reads are present.

For further analysis, only reads spanning a splice junction, so called split reads, were extracted with a mapping quality of greater than 10 to reduce the false positive rate due to mapping issues. Each splice site was annotated as belonging to the start or end of a known exon or to be entirely new. For the reference exon annotation the GENCODE release 24 based on GRCh37 was used[72]. The percent spliced in (Ψ) values for the 3' and 5' sites were calculated as described earlier[37]:

$$\psi_5(D,A) = \frac{n(D,A)}{\sum_{A'} n(D,A')} \quad \text{and} \quad \psi_3(D,A) = \frac{n(D,A)}{\sum_{D'} n(D',A)}$$

Where *D* is a donor site and *A* is an acceptor site. *n(D,A)* denotes the number of reads spanning the given junction. *D'* and *A'* represent all possible donor and acceptor sites, respectively.

Classification of splice sites into background, weak and strong was done by modeling the distribution of the $\psi_5$ and $\psi_3$-values with three components. Identifiability of the three components was facilitated by considering three groups of junctions depending on previous annotation of splice sites: 'no side is annotated', 'one side is annotated' and 'both sides are annotated'. Specifically, the number of split reads *n(D,A)* of a junction conditioned on the total number of reads $N(D,A) = \sum_{A'} n(D,A')$, for $\psi_5$, and $N(D,A) = \sum_{D'} n(D',A)$, for $\psi_3$, was modeled as:

$$P\big(n(D,A)\big|N(D,A)\big) = \sum_{c\in\{bg,wk,st)} \sum_{s=0,1,2} \pi_{s,c}\, BetaBin(n(D,A)|N(D,A),\alpha_c,\beta_c)$$

where *c* is the component index, *s* the number of annotated sites and BetaBin the beta-binomial distribution. Hence, the components were modeled to have the same parameters $\alpha_c, \beta_c$

in all three groups but their mixing proportions $\pi_{s,c}$ to be group-specific. Fitting was performed using the expectation-maximization algorithm. For the initial step, the data points were classified as background ($\psi < 0.001$), weak spliced ($\psi < 0.1$) and canonical ($\psi >= 0.1$). After convergence of the clustering the obtained parameters were used to estimate the probability for each junction side to belong to a given class.

## *Detection of mono-allelic expression*

For mono-allelic expression analysis only heterozygous single nucleotide variants with only one alternative allele detected from exome sequencing data were used. The same quality filters were used as mentioned in the exome sequencing part, but no frequency filter was applied. To get allele counts from RNA sequencing for the remaining variants the function *pileLettersAt* from the R/Bioconductor package *GenomicAlignments*[67] was used. The data was further filtered for variants with coverage of at least 10 reads on the transcriptome.

The DESeq2 package[68,69] was applied on the final variant set to estimate the significance of the allele-specific expression. Allele-specific expression was estimated on each heterozygous variant independently of others (i.e. without phasing the variants). For each sample, a generalized linear model was fitted with the contrast of the coverage of one allele against the coverage of the other alleles (*condition*). Specifically, we modeled $K_{i,j}$ the number of reads of variant *i* in sample *j* as:

$$K_{i,j} \sim NB\left(s_j \times \ q_{i,j}, \alpha\right)$$

$$\log_2\left(q_{i,j}\right) = \beta_i^0 \ + \beta_i^{condition} \mathbf{x}_{i,j}^{condition}$$

Where NB is the negative binomial distribution; the dispersion parameter $\alpha$ was fixed for all variants to $\alpha = 0.05$, which is approximately the average dispersion over all samples based on the gene-wise analysis; $s_j$ is the size factor of each condition; $\beta_i^0$ is the intercept parameter for variant *i*. The value of $\mathbf{x}_{i,j}^{condition}$ is 1 for the alternative alleles and 0 otherwise. The resulting $\beta_i^{condition}$ represents the $\log_2$-fold changes for the alternative allele against the reference allele. The independent filtering by DESeq2 was disabled (*independentFiltering=FALSE*) to keep the coverage outliers among the results. To classify a variant as mono-allelically expressed a cutoff of $\left|\beta_i^{condition}\right| \geq 2$ was used, which corresponds to an allele frequency ≥ 0.8, and we filtered Hochberg adjusted *P*-values to be smaller than 0.05.

## *Mass spectrometric sample preparation*

We performed quantitative proteomics from a second aliquot of cells taken at the same time as the RNA-seq aliquot. Mass spectrometric sample preparation was done as described earlier[73]. Briefly, cells were lysed in SDC lysis buffer (1% sodium deoxycholate, 10 mM TCEP, 40 mM CAA, 100 mM Tris pH 8.5), boiled for 10 min at 95ºC, sonicated and diluted 1:1 with water for LysC and trypsin digestion. The dilution buffer contained appropriate amounts of proteolytic enzyme to ensure a ratio of 1:50 (µg enzyme / µg protein). Digestion was performed at 37°C overnight. Peptides were acidified, loaded on SDB-RPS (poly-styrenedivinylbenzene) material and eluted. Eluted peptides were collected in autosampler vials and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus, 5305 000.304). Peptides were resuspended in buffer A* (2% ACN, 0.1% TFA) and were sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

## *Mass spectrometric data acquisition*

2 µg of peptides per sample were loaded for 100 min gradients separated on a 50 cm column with 75 µm inner diameter in-house packed with 1.9 µm C18 beads (Dr. Maisch GmbH). Reverse phase chromatography was performed at 50ºC with an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific) coupled to the Q Exactive HF[74] mass spectrometer (Thermo Fisher Scientific) via a nanoelectrospray source (Thermo Fisher Scientific). Peptides were loaded in buffer A (0.1% volume/volume formic acid) and eluted with a nonlinear gradient. Operational parameters were real-time monitored by the SprayQC software[75]. Raw files were analysed by the software MaxQuant[76] (version 1.5.3.2) and peak lists were searched against the Homo sapiens Uniprot FASTA database (Version 2014/4) and a common contaminants database (247 entries) by the Andromeda search engine[77]. Label-free quantification was done using the MaxLFQ algorithm[78] (for detailed parameters see Supplementary Table 5) integrated into MaxQuant.

## *Processing of proteome intensities*

The LFQ intensities and gene names were extracted for 6,566 protein groups from the MaxQuant output file *proteinGroups.txt*. For protein groups with more than one member, the first member was chosen to represent the group as single protein with a distinct gene name (similar to earlier studies[79]). MaxLFQ intensities of 0 actually represent non-quantified peaks and were therefore replaced with missing values (NA). The 10 samples that had a frequency of missing values higher than 50% were considered bad quality and were discarded. Furthermore, proteins were discarded because they had no gene name assigned (n=198), were not the most abundant among their duplicates (n=295), were not expressed in any sample (n=93), because their 95th percentile was not detected (n=549), which was also considered as not expressed, analogously to RNA filtering. Finally, 5,431 proteins and 31 samples were considered for further analysis (Supplementary Data 3).

## *Computing protein fold changes and differential expression*

Since the mass spectrometric measurements of all samples were done in a single run, no technical artifacts could be found with a hierarchical clustering. Protein differential expression for each patient compared to the others was tested using moderated T-test approach as implemented in the R/Bioconductor limma package[80]. The transcriptome covariates for sex and HOX effects were used in the linear model for normalization.

## *Transduction and Transfection*

Overexpression of *TIMMDC1* in fibroblast cell lines was performed by lentivirus-mediated expression of the full-length *TIMMDC1* cDNA (DNASU Plasmid Repository) using the ViraPower HiPerform Lentiviral TOPO Expression Kit (Thermo Fisher Scientific)[81]. *TIMMDC1* cDNA was cloned into the pLenti6.3/V5-TOPO expression vector and cotransfected into 293FT cells with the packaging plasmid mix using Lipofectamine 2000. After 24 h, the transfection mix was replaced with high glucose DMEM supplemented with 10% FBS. After further 72 h, the viral particle containing supernatant was collected and used for transduction of the fibroblast cell lines. Selection of stably expressing cells was performed using 5 µg/mL Blasticidin (Thermo Fisher Scientific) for 2 weeks.

### *Immunoblotting*

Total fibroblast cell lysates were subjected to whole protein quantification, separated on 4-12% precast gels (Lonza) by SDS-PAGE electrophoresis and semi-dry transferred to PVDF membranes (GE Healthcare Life Sciences).  The membranes were blocked in 5% non-fat milk (Bio Rad) in TBS-T for 1 h and immunoblotted using primary antibodies against CLPP (Abcam, ab56455), MCOLN1 (Abcam, ab28508), NDUFA13 (Abcam, ab110240), NDUFB3 (Abcam, ab55526), NDUFB8 (Abcam, ab110242), TIMMDC1 (Abcam, ab171978), and UQCRC2 (Abcam, ab14745) for 1 h at RT or ON at 4°C. Signals were detected by incubation with HRP-conjugated goat anti-rabbit and goat anti-mouse secondary antibodies (Jackson Immuno Research Laboratories) for 1 h and visualized using ECL (GE Healthcare Life Sciences).

### *Blue native PAGE (BN-PAGE)*

Fresh fibroblast cell pellets were resuspended in PBS supplemented with 0.25 mM PMSF and 10 U/mL DNAse I and solubilized using 2 mg digitonin/mg protein. The mixture was incubated on ice for 15 min followed by addition of 1 mL PBS and subsequent centrifugation for 10 min at 10000 g and 4°C. The pellet was resuspended in 1x MB (750 mM ε-aminocaproic acid, 50 mM bis-Tris, 0.5 mM EDTA, pH 7.0) and subjected to whole protein quantification. Membrane proteins were solubilized at a protein concentration of 2 µg/µL using 0.5% (v/v) *n*-dodecyl-β-d-maltoside (DDM) for 1 h on ice and centrifuge for 30 min at 10000 g at 4°C.  The supernatant was recovered and whole protein amount was quantified. Serva Blue G (SBG) was added to a final concentration of 0.25% (v/v) and 60 µg protein were loaded on NativePAGE 4-16% Bis-Tris gels (Thermo Fisher Scientific). Anode buffer contained 50 mM Bis-Tris, pH 7.0, blue cathode buffer contained 15 mM Bis-Tris, 50 mM Tricine, pH 7.0, 0.02% SBG. Electrophoresis was started at 40 V for 30 min and continued at 130 V until the front line proceeded 2/3 of the gel. Subsequently, blue cathode buffer was replaced by clear cathode buffer not containing SBG (15 mM Bis-Tris, 50 mM Tricine, pH 7.0). Proteins were wet transferred to PVDF membranes and immunoblotted using primary antibodies against NDUFB8 to visualize complex I and UQCRC2 to visualize complex III.

### *Proline supplementation growth assay*

We modified a method established earlier[51]. For the comparative growth assay, equal number of cells (n=250) from patient and control were seeded in 96-well plates and cultured in DMEM containing 10% of either normal or dialyzed FBS. Medium with normal FBS contains small molecules, whereas medium with dialyzed FBS is free of molecules with a molecular weight smaller than 10,000 mw (Proline-free medium). To confirm the effect of Proline deprivation, DMEM containing dialyzed FBS was supplemented with 100 µM L-Proline to rescue the growth defect. After paraformaldehyde fixation, nuclei were stained with 4',6-diamidino-2-phenylindole

(DAPI) and cell number was determined using a Cytation3 automated plate reader (BioTek, USA).

### *Cellular ROS production*

Intensity of hydroethidine (HEt) oxidation products as a measure of cellular ROS production was quantified in living skin fibroblasts using epifluorescence microscopy as described previously[82].

## *References*

1. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).

2. O'Donnell-Luria, A. H. & Miller, D. T. A Clinician's perspective on clinical exome sequencing. *Human Genetics* **135,** 643–654 (2016).

3. Shashi, V. *et al.* The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet. Med.* **16,** 176–182 (2014).

4. Ankala, A. *et al.* A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. *Ann. Neurol.* **77,** 206–214 (2015).

5. Taylor, R. W. *et al.* Use of Whole-Exome Sequencing to Determine the Genetic Basis of Multiple Mitochondrial Respiratory Chain Complex Deficiencies. *Jama* **312,** 68 (2014).

6. Lieber, D. S. *et al.* Targeted exome sequencing of suspected mitochondrial disorders. *Neurology* **80,** 1762–1770 (2013).

7. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

8. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

9. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47,** 717–726 (2015).

10. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *bioRxiv* 1–22 (2016). doi:10.1101/074443

11. Zeng, Y. *et al.* Aberrant Gene Expression in Humans. *PLoS Genet.* **11,** 1–20 (2015).

12. Guan, J. *et al.* Exploiting aberrant mRNA expression in autism for gene discovery and diagnosis. *Hum. Genet.* **135,** 1–15 (2016).

13. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *Am. J. Hum. Genet.* **98,** 299–309 (2016).

14. Albers, C. a *et al.* Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat. Genet.* **44,** 435–9, S1-2 (2012).

15. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* **16,** 653–664 (2015).

16. Eckersley-Maslin, M. A. & Spector, D. L. Random monoallelic expression: Regulating gene expression one allele at a time. *Trends Genet.* **30,** 237–244 (2014).

17. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease* **1792,** 14–26 (2009).

18. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17,** 19–32 (2015).

19. Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.* **18,** 472–482 (2012).

20. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-. ).* **347,** 1254806–1254806 (2015).

21. Muntoni, F., Torelli, S. & Ferlini, A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. *Lancet. Neurol.* **2,** 731–40 (2003).

22. Gonorazky, H. *et al.* RNAseq analysis for the diagnosis of muscular dystrophy. *Ann. Clin. Transl. Neurol.* **3,** 55–60 (2016).

23. Morel, C. F. *et al.* A LMNA Splicing Mutation in Two Sisters with Severe Dunnigan-Type Familial Partial Lipodystrophy Type 2. *J. Clin. Endocrinol. Metab.* **91,** 2689–2695 (2006).

24. Qu, Y. *et al.* A rare variant (c.863G&gt;T) in exon 7 of SMN1 disrupts mRNA splicing and is responsible for spinal muscular atrophy. *Eur. J. Hum. Genet.* **24,** 864–870 (2016).

25. Gorman, G. S. *et al.* Mitochondrial diseases. *Nat. Rev. Dis. Prim.* **2,** 16080 (2016).

26. Elstner, M., Andreoli, C. & Ahting, U. MitoP2: an integrative tool for the analysis of the mitochondrial proteome. *Mol. Biotechnol.* **40,** 306–315 (2008).

27. Mayr, J. A. *et al.* Spectrum of combined respiratory chain defects. *J. Inherit. Metab. Dis.* **38,** 629–640 (2015).

28. Haack, T. B. *et al.* ELAC2 mutations cause a mitochondrial RNA processing defect associated with hypertrophic cardiomyopathy. *Am. J. Hum. Genet.* **93,** 211–223 (2013).

29. Haack, T. B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat. Genet.* **42,** 1131–4 (2010).

30. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352,** 600–4 (2016).

31. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387 (2014).

32. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (2016).

33. Lee, K. K., Shimoji, M., Hossain, Q. S., Sunakawa, H. & Aniya, Y. Novel function of glutathione transferase in rat liver mitochondrial membrane: role for cytochrome c release from mitochondria. *Toxicol. Appl. Pharmacol.* **232,** 109–18 (2008).

34. Holzerova, E. *et al.* Human thioredoxin 2 deficiency impairs mitochondrial redox homeostasis and causes early-onset neurodegeneration. *Brain* **139,** 346–54 (2016).

35. Guarani, V. *et al.* TIMMDC1/C3orf1 Functions as a Membrane-Embedded Mitochondrial Complex I Assembly Factor through Association with the MCIA Complex. *Mol. Cell. Biol.* **34,** 847–861 (2014).

36. Andrews, B., Carroll, J., Ding, S., Fearnley, I. M. & Walker, J. E. Assembly factors for the membrane arm of human complex I. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18934–9 (2013).

37. Pervouchine, D. D., Knowles, D. G. & Guig, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29,** 273–274 (2013).

38. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–91 (2016).

39. Halperin, T., Zheng, B., Itzhaki, H., Clarke, A. K. & Adam, Z. Plant mitochondria contain proteolytic and regulatory subunits of the ATP-dependent Clp protease. *Plant Mol. Biol.* **45,** 461–468 (2001).

40. Jenkinson, E. M. *et al.* Perrault Syndrome Is Caused by Recessive Mutations in CLPP, Encoding a Mitochondrial ATP-Dependent Chambered Protease. *Am. J. Hum. Genet.* **92,** 605–613 (2013).

41. Jenkinson, E. M. *et al.* Perrault syndrome: further evidence for genetic heterogeneity. *J. Neurol.* **259,** 974–976 (2012).

42. Szczepanowska, K. *et al.* CLPP coordinates mitoribosomal assembly through the regulation of ERAL1 levels. *EMBO J.* **35,** 2566–2583 (2016).

43. Piva, F., Giulietti, M., Burini, A. B. & Principato, G. SpliceAid 2: a database of human

splicing factors expression data and RNA target motifs. *Hum. Mutat.* **33,** 81–5 (2012).

44. Dogan, R. I., Getoor, L., Wilbur, W. J. & Mount, S. M. SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Res.* **35,** W285-91 (2007).

45. Timmermans, M. J. T. N. *et al.* Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res.* **38,** e197–e197 (2010).

46. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37,** e67 (2009).

47. Yeo, G., Hoon, S., Venkatesh, B. & Burge, C. B. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 15700–5 (2004).

48. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

49. Kapustin, Y. *et al.* Cryptic splice sites and split genes. *Nucleic Acids Res.* **39,** 5837–5844 (2011).

50. Adams, E. & Frank, L. Metabolism of Proline and the Hydroxyprolines. *Annu. Rev. Biochem.* **49,** 1005–1061 (1980).

51. Baumgartner, M. R. *et al.* Hyperammonemia with reduced ornithine, citrulline, arginine and proline: a new inborn error caused by a mutation in the gene encoding delta(1)-pyrroline-5-carboxylate synthase. *Hum. Mol. Genet.* **9,** 2853–8 (2000).

52. Fischer-Zirnsak, B. *et al.* Recurrent De Novo Mutations Affecting Residue Arg138 of Pyrroline-5-Carboxylate Synthase Cause a Progeroid Form of Autosomal-Dominant Cutis Laxa. *Am. J. Hum. Genet.* **97,** 483–92 (2015).

53. Coutelier, M. *et al.* Alteration of ornithine metabolism leads to dominant and recessive hereditary spastic paraplegia. *Brain* **138,** 2191–205 (2015).

54. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17,** 407–21 (2016).

55. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *bioRxiv* (2016). doi:10.1101/074153

56. Gibson, G. Human genetics. GTEx detects genetic effects. *Science* **348,** 640–1 (2015).

57. Vafai, S. B. & Mootha, V. K. Mitochondrial disorders as windows into an ancient organelle. *Nature* **491,** 374–83 (2012).

58. Gagneur, J. *et al.* Genotype-Environment Interactions Reveal Causal Pathways That

Mediate Genetic Effects on Phenotype. *PLoS Genet.* **9,** e1003803 (2013).

59. Mayr, J. A. *et al.* Lack of the mitochondrial protein acylglycerol kinase causes sengers syndrome. *Am. J. Hum. Genet.* **90,** 314–320 (2012).

60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

61. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

63. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

64. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6,** R44 (2005).

65. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

66. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22,** 1036–1046 (2006).

67. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).

68. Anders, S. & Huber, W. DESeq: Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

70. Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802 (1988).

71. Li, Y. I., Knowles, D. A. & Pritchard, J. K. LeafCutter: Annotation-free quantification of RNA splicing. *bioRxiv* 44107 (2016). doi:10.1101/044107

72. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

73. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324 (2014).

74. Scheltema, R. A. *et al.* The Q Exactive HF, a Benchtop mass spectrometer with a pre-

filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **13,** 3698–708 (2014).

75. Scheltema, R. A. & Mann, M. SprayQc: A Real-Time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11,** 3458–3466 (2012).

76. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–72 (2008).

77. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).

78. Cox, J., Hein, M. Y., Luber, C. a & Paron, I. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. ...* **13,** 2513–2526 (2014).

79. Cheng, Z. *et al.* Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol. Syst. Biol.* **12,** 855–855 (2016).

80. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

81. Van Haute, L. *et al.* Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in NSUN3. *Nat. Commun.* **7,** 12039 (2016).

82. Forkink, M., Smeitink, J. a M., Brock, R., Willems, P. H. G. M. & Koopman, W. J. H. Detection and manipulation of mitochondrial reactive oxygen species in mammalian cells. *Biochim. Biophys. Acta* **1797,** 1034–44 (2010).

## *Acknowledgements*

We are grateful to the participating families. Further we thank Dr. Lina Florentin for providing DNA samples of the parents of #35791. This study was supported by the German Bundesministerium für Bildung und Forschung (BMBF) through the E-Rare project GENOMIT (01GM1603 and 01GM1207, HP and TM, FWF I 920-B13 for JAM, and J41J11000420001 for DG), through the Juniorverbund in der Systemmedizin "mitOmics" (FKZ 01ZX1405A, JG, and FKZ 01ZX1405C, TBH), and the DZHK (German Centre for Cardiovascular Research, LSK, TM). The study was furthermore supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) within the framework of the Munich Cluster for Systems Neurology (EXC 1010 SyNergyDFG, TM) and a Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM) supports DMB. HP, VT, and JAM are supported by EU FP7 Mitochondrial European Educational Training Project (317433). CM, JG and HP are supported by EU Horizon2020 Collaborative Research Project SOUND (633974). RWT is supported by a Wellcome Trust Strategic Award (096919/Z/11/Z) and the MRC Centre for Neuromuscular Diseases (G0601943). DG is supported by Telethon-Italy (GGP15041). We thank the Pierfranco and Luisa Mariani Foundation and the Cell lines and DNA Bank of Paediatric Movement Disorders and Mitochondrial Diseases of the Telethon Genetic Biobank Network (GTB09003).

## *Author contributions*

Project planning: T.M., J.G., H.P. Experimental design: H.P. Review of phenotypes, sample collection and biochemical analysis: C.L., B.F., A.D., V.T., A.L., D.G., R.T., D.G., J.A.M., A.R., P.F., F.D., and T.M. Investigation L.S.K., D.M.B., and C.M. Data curation and analysis: L.S.K., D.M.B., C.M., T.M.S., and H.P. Cell biology experiments: L.S.K., R.K., A.I., C.T., E.K., and B.R. Exome, genome, and RNA sequencing; L.S.K., R.K., E.G., T.S., P.L. and T.M.S. Exome analysis: L.S.K., R.K.,T.B.H., and H.P. Quantitative proteomics: L.S.K. and G.P. Metabolomic studies: L.S.K., G.K., and J.A.  Manuscript writing: L.S.K., D.M.B., C.M., J.G., and H.P. Visualization L.S.K., D.M.B., and C.M. Critical revision of the manuscript: all authors.

## *Competing financial interests*

The authors declare that they have no competing interests.

## *Materials & Correspondence*

Correspondence and material requests should be addressed to: Holger Prokisch (prokisch@helmholtz-muenchen.de) or Julien Gagneur (gagneur@in.tum.de).

Supplementary tables and R code to reproduce paper figures are available online at our webserver (https://i12g-gagneurweb.in.tum.de/public/paper/mitoMultiOmics).

## *Figure Legends*

### *Figure 1: Strategy for genetic diagnosis using RNA-seq*

The approach we followed started with RNA sequencing of fibroblasts from unsolved WES patients. Three strategies to facilitate diagnosis were pursued: Detection of aberrant expression (e.g. depletion), aberrant splicing (e.g. exon creation) and mono-allelic expression of the alternative allele (i.e. A as alternative allele). Candidates were validated by proteomic measurements, lentiviral transduction of the wildtype (wt) allele or, in particular cases, by specific metabolic supplementation.

### *Figure 2: RNA expression outlier detection and validation*

(a) Aberrantly expressed genes (Hochberg corrected *P*-value < 0.05 and |Z-score| > 3) for each patient fibroblasts.

(b) Gene-wise RNA expression volcano plot of nominal *P*-values (- $\log_{10}$ *P*-value) against Z-scores of the patient #35791 compared against all other fibroblasts. Absolute Z-scores greater than 5 are plotted at ±5, respectively.

(c) Same as (b) for patient #73804.

(d) Sample-wise RNA expression is ranked for the genes *TIMMDC1* (top) and *MGST1* (bottom). Samples with aberrant expression for the corresponding gene are highlighted in red (#73804, #35791, and #66744).

(e) Gene-wise comparison of RNA and protein fold changes of patient #35791 against all other fibroblast cell lines. Subunits of the mitochondrial respiratory chain complex I are highlighted (red squares). Reliably detected proteins that were not detected in this sample are shown separately with their corresponding RNA fold changes (points below solid horizontal line).

(f) Western blot of TIMMDC1, NDUFA13, NDUFB3, and NDUFB8 protein in three fibroblast cell lines without (#62346, #91324, #NHDF) and three with a variant in *TIMMDC1* (#35791, #66744, #96687), and fibroblasts re-expressing *TIMMDC1* ("-T") (#35791-T, #66744-T, #96687-T). UQCRC2 was used as loading control. MW, molecular weight; CI, complex I subunit; CIII, complex III subunit.

(g) Blue native PAGE blot of the control fibroblasts re-expressing *TIMMDC1* (NHDF-T), the control fibroblasts (NHDF), patient fibroblasts (#96687), and patient fibroblast re-expressing *TIMMDC1* (#96687-T). Immunodecoration for complex I and complex III was performed using NDUFB8 and UQCRC2 antibodies, respectively. CI, complex I subunit; CIII, complex III subunit.

## *Figure 3: Aberrant splicing detection and quantification*

(a) Aberrant splicing events (Hochberg corrected *P*-value < 0.05) for all fibroblasts.

(b) Aberrant splicing events (n=175) grouped by their splicing category in undiagnosed patients (n=48) after manual inspection.

(c) *CLPP* Sashimi plot of exon skipping and truncation events in affected and unaffected fibroblasts (red and orange, respectively). The RNA coverage is given as the $\log_{10}$ RPKM-value and the number of split reads spanning the given intron is indicated on the exon-connecting lines. At the bottom the gene model of the RefSeq annotation is depicted. The aberrantly spliced exon is colored in red.

(d) Same as in (c) for *TIMMDC1*. At the bottom the newly created exon is depicted in red within the RefSeq annotation track.

(e) Coverage tracks (light red) for patients #35791, #66744, and #91324 based on RNA and whole genome sequencing. For patient #91324 only WGS is available. The homozygous SNV c.596+2146A>G is present in all coverage tracks (vertical orange bar). The top tracks show the genomic annotation: genomic position on chromosome 3, DNA sequence, amino acid translation (grey, stop codon in red), the RefSeq gene model (blue line), the predominant additional exon of *TIMMDC1* (blue rectangle), and the SNV annotation of the 1000 Genomes Project (each black bar represents one variant).

(f) Percent spliced in (Ψ) distribution for different splicing classes and genes. Top: Histogram giving the genome-wide distribution of the 3' and 5' Ψ-values based on all reads over all samples. Middle: The shaded horizontal bars represent the densities (black for high density) of the background, weak and strong splicing class, respectively (Methods). Bottom: Ψ-values of the predominant donor and acceptor splice sites of genes with private splice sites (i.e found dominant in at most two samples) computed over all other samples.

## *Figure 4: Detection and validation of mono-allelic expression of rare variants*

(a) Distribution of heterozygous single nucleotide variants (SNVs) across samples for different consecutive filtering steps. Heterozygous SNVs detected by exome sequencing (black), SNVs with RNA-seq coverage of at least 10 reads (gray), SNVs where the alternative allele is mono-allelically expressed (alternative allele frequency > 0.8 and Benjamini-Hochberg corrected *P*-value < 0.05, blue), and the rare subset of those (ExAC minor allele frequency < 0.001, red).

(b) Fold change between alternative (ALT+1) and reference (REF+1) allele read counts for the patient #80256 compared to total read counts per SNV within the sample. Points are colored according to the groups defined in (a).

(c) Gene-wise comparison of RNA and protein fold changes of the patient #80256 against all other patients' fibroblasts. The position of the gene *ALDH18A1* is highlighted. Reliably detected proteins that were not detected in this sample are shown separately with their corresponding RNA fold changes (points below solid horizontal line).

(d) Relative intensity for metabolites of the proline biosynthesis pathway (inlet) for the patient #80256 and 16 healthy controls of matching age. Equi-tailed 95% interval (whiskers), 25th, 75th percentile (boxes) and median (bold horizontal line) are indicated. Data points belonging to the patient are highlighted (red circles and triangles, if Student's *t*-test *P*-value < 0.05).

(e) Cell counts under different growth conditions for the normal human dermal fibroblast (NHDF) and patient #80256. Both fibroblasts were grown in fetal bovine serum (FBS), dialyzed FBS (without proline) and dialyzed FBS with proline added. Boxplot as in (d). *P*-values are based on a two-sided Wilcoxon test.

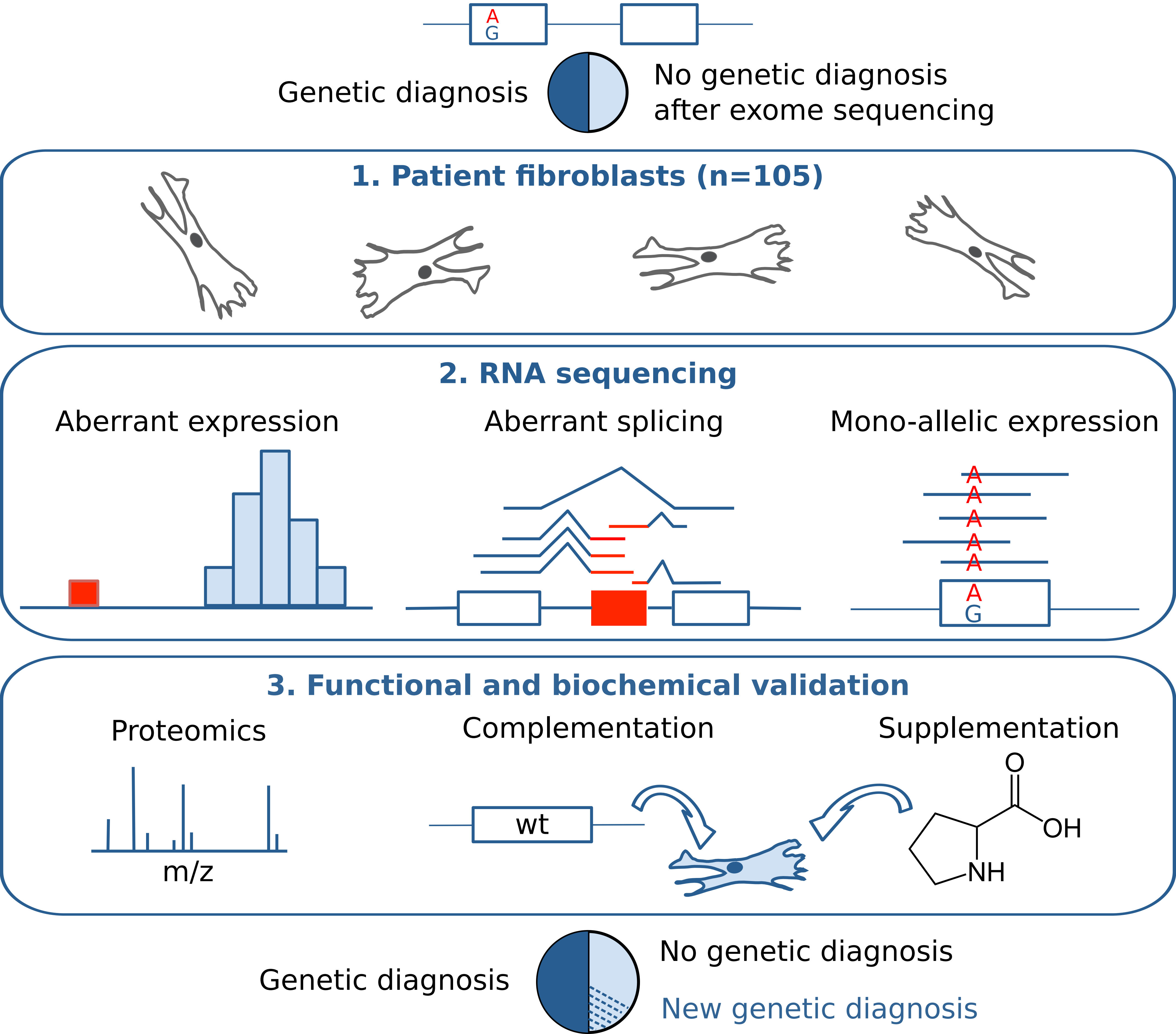
(f) Intron retention for *MCOLN1* in patient #62346. Tracks from top to bottom: genomic position on chromosome 19, amino acid translation (red for stop codons), RefSeq gene model, coverage of whole exome sequencing of patient #62346, RNA-seq based coverage for patients

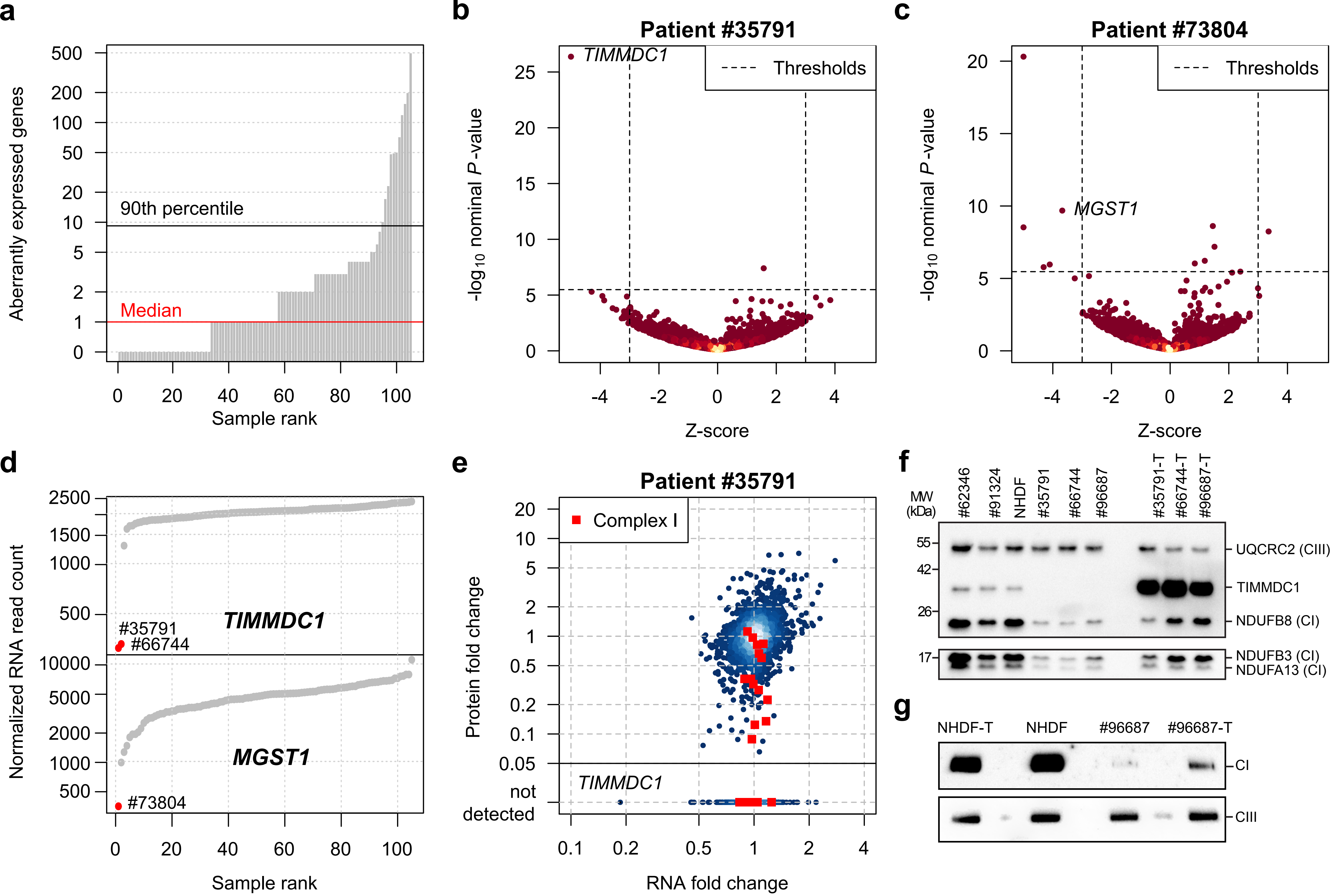
#62346 and #85153 (red and orange shading, respectively). SNVs are indicated by non-reference colored bars with respect to the corresponding reference and alternative nucleotide.

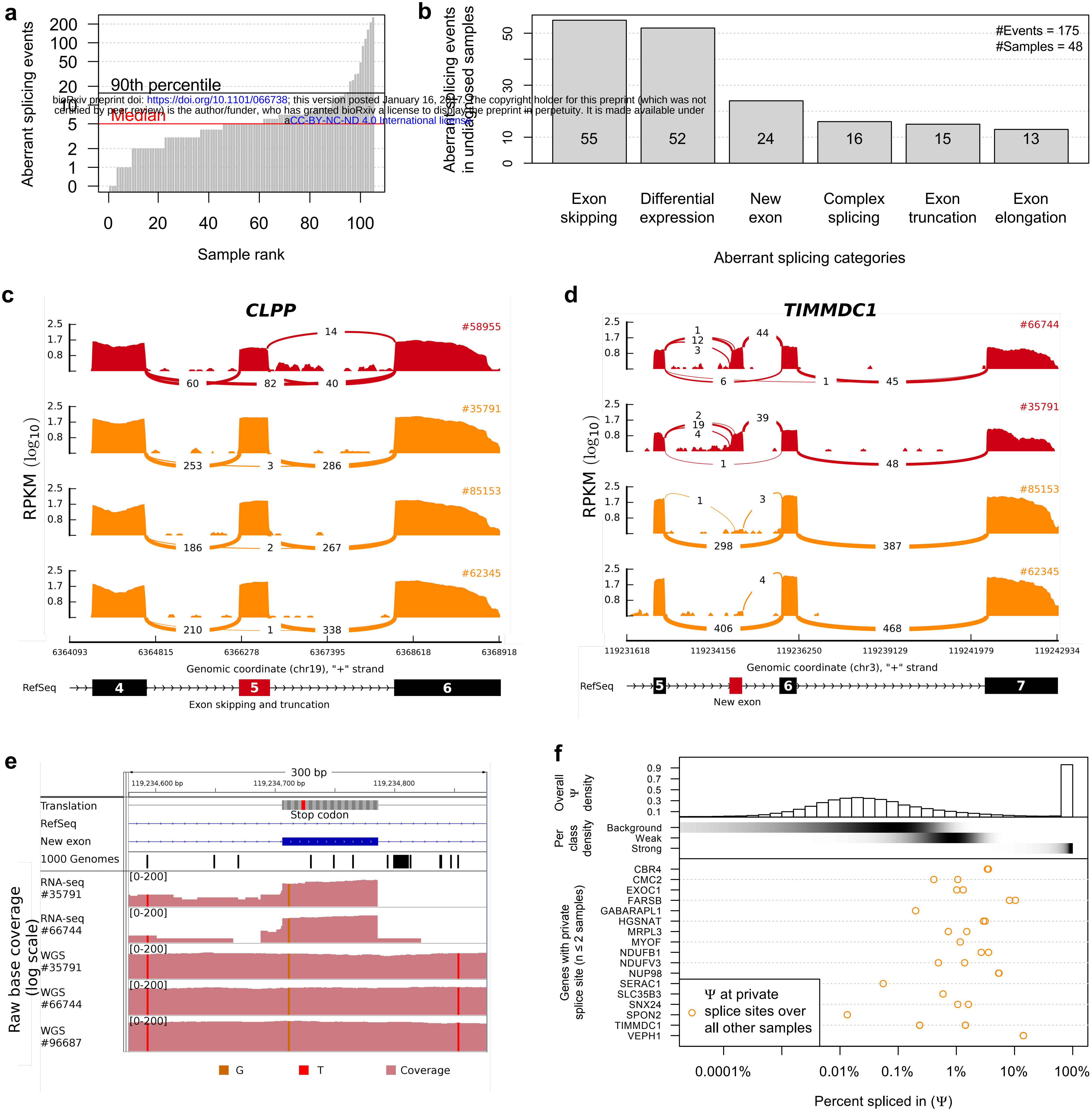
### *Figure 5: Validation summary*

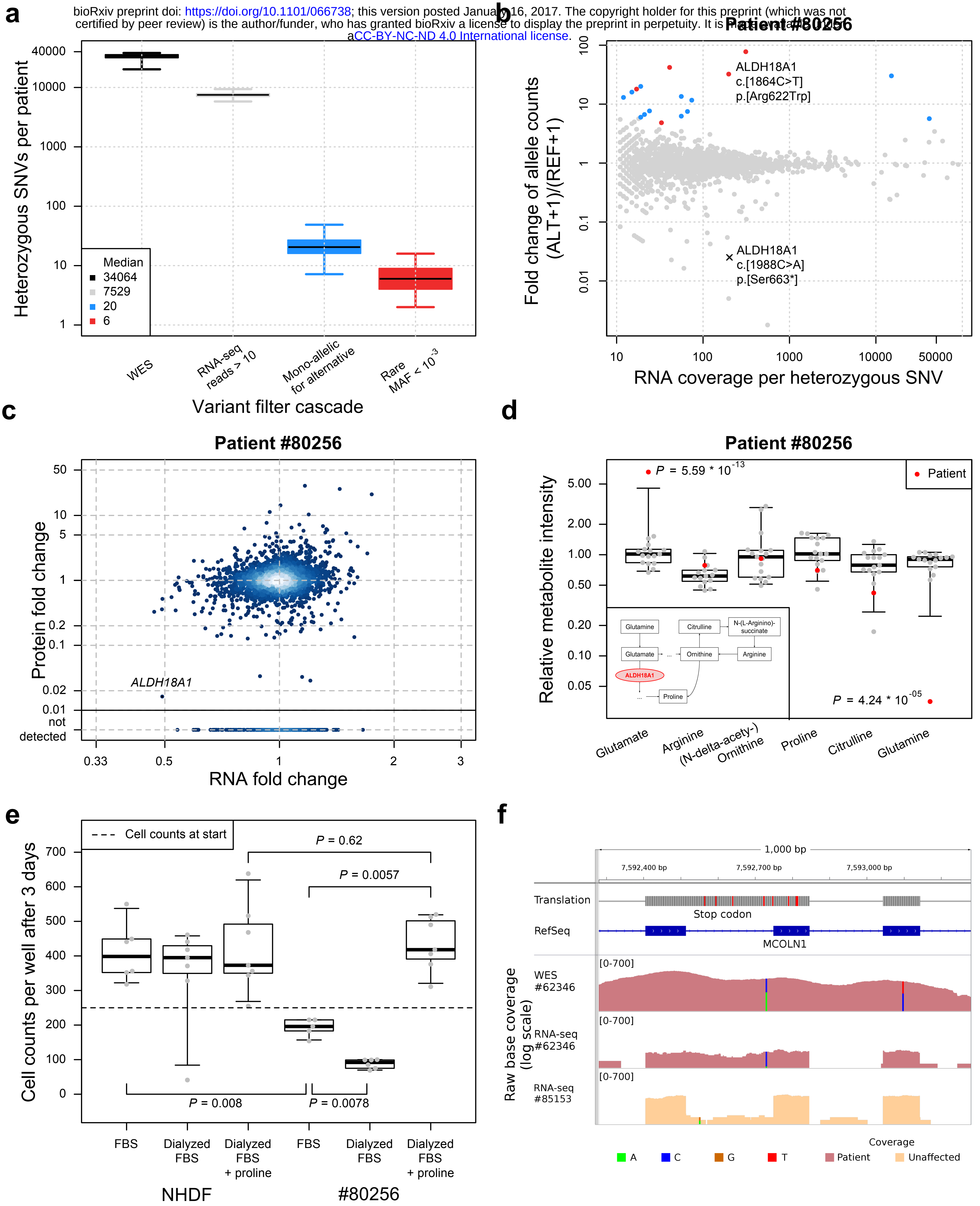
(a) Discovery and validation of genes with RNA defects in newly diagnosed patients, i.e. *TIMMDC1* (n=2 patients), *CLPP*, *ALDH18A1*, and *MCOLN1*, and patients with strong candidates, i.e. *MGST1*. The median number (± median absolute deviation) of candidate genes is given per detection strategies. Dotted check, manual inspection not statistically significant.

(b) Schematic representation of variant causing splicing defects for TIMMDC1 (top, new exon red box), CLPP (middle, exon skipping and truncation), and MCOLN1 (bottom, intron retention). Variants are depicted by a red star.


A
G
Genetic diagnosis
No genetic diagnosis
after exome sequencing
1. Patient fibroblasts (n=105)
2. RNA sequencing
Aberrant expression
Aberrant splicing
Mono-allelic expression
A
A
A
A
A
A
G
3. Functional and biochemical validation
Proteomics
m/z
Complementation
wt
Supplementation
O
OH
NH
Genetic diagnosis
No genetic diagnosis
New genetic diagnosis


a
Aberrantly expressed genes
500
200
100
50
20
10
5
2
1
0
90th percentile
Median
0
20
40
60
80
100
Sample rank
b
Patient #35791
TIMMDC1
Thresholds
-log10 nominal P-value
25
20
15
10
5
0
-4
-2
0
2
4
Z-score
c
Patient #73804
Thresholds
MGST1
-log10 nominal P-value
20
15
10
5
0
-4
-2
0
2
4
Z-score
d
Normalized RNA read count
2500
1500
1000
500
#35791
#66744
TIMMDC1
10000
5000
2000
1000
500
MGST1
#73804
0
20
40
60
80
100
Sample rank
e
Patient #35791
Complex I
Protein fold change
20
10
5
2
1
0.5
0.2
0.1
0.05
not detected
TIMMDC1
0.1
0.2
0.5
1
2
4
RNA fold change
f
MW (kDa)
#62346
#91324
NHDF
#35791
#66744
#96687
#35791-T
#66744-T
#96687-T
55
42
26
17
UQCRC2 (CIII)
TIMMDC1
NDUFB8 (CI)
NDUFB3 (CI)
NDUFA13 (CI)
g
NHDF-T
NHDF
#96687
#96687-T
CI
CIII


a
Aberrant splicing events
200
100
50
20
10
5
2
1
0
90th percentile
Median
0
20
40
60
80
100
Sample rank
b
Aberrant splicing events in undiagnosed samples
50
30
10
0
#Events = 175
#Samples = 48
55
52
24
16
15
13
Exon skipping
Differential expression
New exon
Complex splicing
Exon truncation
Exon elongation
Aberrant splicing categories
c
CLPP
RPKM (log10)
2.5
1.7
0.8
#58955
14
60
82
40
#35791
253
3
286
#85153
186
2
267
#62345
210
1
338
6364093
6364815
6366278
6367395
6368618
6368918
Genomic coordinate (chr19), "+" strand
RefSeq
4
5
6
Exon skipping and truncation
d
TIMMDC1
RPKM (log10)
2.5
1.7
0.8
#66744
1
12
3
44
6
1
45
#35791
2
19
4
39
1
48
#85153
1
3
298
387
#62345
4
406
468
119231618
119234156
119236250
119239129
119241979
119242934
Genomic coordinate (chr3), "+" strand
RefSeq
5
6
7
New exon
e
300 bp
119,234,600 bp
119,234,700 bp
119,234,800
Translation
Stop codon
RefSeq
New exon
1000 Genomes
Raw base coverage (log scale)
RNA-seq #35791
[0-200]
RNA-seq #66744
[0-200]
WGS #35791
[0-200]
WGS #66744
[0-200]
WGS #96687
[0-200]
G
T
Coverage
f
Overall Ψ density
0.9
0.7
0.5
0.3
0.1
Per class density
Background
Weak
Strong
Genes with private splice site (n ≤ 2 samples)
CBR4
CMC2
EXOC1
FARSB
GABARAPL1
HGSNAT
MRPL3
MYOF
NDUFB1
NDUFV3
NUP98
SERAC1
SLC35B3
SNX24
SPON2
TIMMDC1
VEPH1
Ψ at private splice sites over all other samples
0.0001%
0.01%
0.1%
1%
10%
100%
Percent spliced in (Ψ)


a
Heterozygous SNVs per patient
40000
10000
1000
100
10
1
Median
■ 34064
■ 7529
■ 20
■ 6
WES
RNA-seq reads > 10
Mono-allelic for alternative
Rare MAF < $10^{-3}$
Variant filter cascade
b
Patient #80256
Fold change of allele counts (ALT+1)/(REF+1)
100
10
1
0.1
0.01
ALDH18A1 c.[1864C>T] p.[Arg622Trp]
ALDH18A1 c.[1988C>A] p.[Ser663*]
10
100
1000
10000
50000
RNA coverage per heterozygous SNV
c
Patient #80256
Protein fold change
50
10
5
1
0.2
0.1
0.02
0.01
not detected
ALDH18A1
0.33
0.5
1
2
3
RNA fold change
d
Patient #80256
Relative metabolite intensity
5.00
2.00
1.00
0.50
0.20
0.10
0.05
$P = 5.59 * 10^{-13}$
● Patient
$P = 4.24 * 10^{-05}$
Glutamine
Citrulline
N-(L-Arginino)-succinate
Glutamate
...
Ornithine
Arginine
ALDH18A1
...
Proline
Glutamate
Arginine
(N-delta-acety-) Ornithine
Proline
Citrulline
Glutamine
e
Cell counts per well after 3 days
700
600
500
400
300
200
100
0
Cell counts at start
P = 0.62
P = 0.0057
P = 0.008
P = 0.0078
FBS
Dialyzed FBS
Dialyzed FBS + proline
FBS
Dialyzed FBS
Dialyzed FBS + proline
NHDF
#80256
f
1,000 bp
7,592,400 bp
7,592,700 bp
7,593,000 bp
Translation
Stop codon
RefSeq
MCOLN1
Raw base coverage (log scale)
[0-700]
WES #62346
[0-700]
RNA-seq #62346
[0-700]
RNA-seq #85153
Coverage
A
C
G
T
Patient
Unaffected