Journal of Inherited Metabolic Disease - Review:

Title:

The diagnosis of inborn errors of metabolism by an integrative "multi-omics" approach: a perspective encompassing genomics, transcriptomics and proteomics

Author names:

Sarah Louise Stenton[1,2]
Laura Sophie Kremer[1,2]
Robert Kopajtich[1,2]
Christina Ludwig[3]
Holger Prokisch[1,2]

Author affiliations:

1. Institute of Human Genetics, Technische Universität München, München 48559, Germany
2. Institute of Human Genetics, Helmholtz Zentrum München, München 85764, Germany
3. Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), Technische Universität München, München, Germany

Corresponding author:

Holger Prokisch (prokisch@helmholtz-muenchen.de)

Word count (excluding summary, acknowledgments, references and figure legends):

3864

Number of figures and tables:

7 figures

0 tables

Abstract:

Given the rapidly decreasing cost and increasing speed and accessibility of massively parallel technologies, the integration of comprehensive genomic, transcriptomic, and proteomic data into a "multi-omics" diagnostic pipeline is within reach. Even though genomic analysis has the capability to reveal all possible perturbations in our genetic code, analysis typically reaches a diagnosis in just 35% of cases, with a diagnostic gap arising due to limitations in prioritization and interpretation of detected variants. Here we review the utility of complementing genetic data with transcriptomic data and give a perspective for the introduction of proteomics into the diagnostic pipeline. Together these methodologies enable comprehensive capture of the functional consequence of variants, unobtainable by the analysis of each methodology in isolation. This facilitates functional annotation and reprioritization of candidate genes and variants - a promising approach to shed light on the underlying molecular cause of a patient's disease, increasing diagnostic rate, and allowing actionability in clinical practice.

1 sentence take-home message:

The parallel analysis of "multi-omics" methodologies increases the capability to decipher the molecular genetic cause of a patient's disease.

Details of the contributions of individual authors:

This review was written by Sarah Louise Stenton under the guidance of Holger Prokisch. Laura Sophie Kremer and Robert Kopajtich conceptualized figures and revised the manuscript. Christina Ludwig critically read the manuscript. Holger Prokisch is the acting guarantor.

The name of the corresponding author:

Holger Prokisch (prokisch@helmholtz-muenchen.de)

A competing interest statement:

We declare no competing interests associated with the manuscript.

Details of funding:

1. mitoNet German Network for Mitochondrial Diseases [grant number 01GM1113C (to H.P.)]
2. The GENOMIT European Network for Mitochondrial Disease [grant number 01GM1207 (to H.P. and S.L.S.)]
3. The SOUND EU Horizon2020 Collaborative Research Project [grant number 633974 (to H.P.)]
4. The PhD Program Medical Life Science and Technology scholarship, TUM School of Medicine (to S.L.S.)

Compliance with Ethics Guidelines:

Conflict of Interest:

Sarah Louise Stenton, Laura Sophie Kremer, Robert Kopajtich, Christina Ludwig and Holger Prokisch declare that they have no conflict of interest.

Details of ethics approval:

No ethical approval required.

Informed Consent:

No informed consent required.

A patient consent statement:

No patient consent required.

Animal Rights:

This article does not contain any studies with human or animal subjects performed by the any of the authors.

Documentation of approval from the Institutional Committee for Care and Use of Laboratory Animals (or comparable committee):

Not applicable.

A list of approximately six keywords:

Multi-omics, Genomics, Transcriptomics, Proteomics, Diagnostics

Introduction:

Diagnostics in Mendelian disease has been transformed by genomic sequencing. While in the past it has been a long, frustrating and often losing battle to elucidate the genetic cause of a patient's disease, genomic sequencing has made this - at least conceptually - possible for every patient. Despite the revolutionary impact of coding-sequence interrogation by whole-exome sequencing (WES) just 35% (24-68%) of patients receive a genetic diagnosis (Clark MM et al 2018). Diagnostic yield can increase to over 65% in carefully selected cohorts where novel genetic diagnoses are sought after (Tarailo-Graovac M et al 2016). The diagnostic shortfall indicates detection evasion of causative variants or detection of variants of uncertain significance (VUS) due to still existing gaps in knowledge. An overview of next generation sequencing (NGS) can be seen in Figure 1.

Inborn errors of metabolism (IEM) are a heterogeneous group of more than 1000 diseases (Ferreira CR et al 2019). Though individually rare, they have a significant cumulatively burden, with a global birth prevalence of 50 per 100,000 (Waters D et al 2018). IEM manifest with phenotypes ranging from non-specific delay in attaining developmental milestones, to acute decompensation with severe metabolic acidosis and sudden premature death. The result of such heterogeneity is difficultly in diagnosis. Metabolic biomarkers aid diagnosis and the introduction of comprehensive metabolomic approaches has further streamlined diagnostic workup, placating the need for multiple targeted biochemical assays guided by clinical phenotype (Coene KL et al 2018). However, metabolic biomarkers often lack specificity and so, as IEMs are linked to an underlying genetic defect, they are most effectively characterized by identification of the causative gene and its downstream products.

Extensive phenotypic overlap exists across the many distinct genetic etiologies of IEM (Lake NJ et al 2016) and conversely, the phenomenon of pleiotropy is commonplace. Typically, IEM are seen to be inherited in a “one-gene one-disease” manner. However, in about 5% of sequenced cases, multiple independent genetic diagnoses are made within one individual, leading to superimposed traits, blending of phenotypes, and reduced phenotypic similarity between cases (Posey JE et al 2017, Tarailo-Graovac M et al 2016, Yang Y et al 2014, Yang Y et al 2013). Moreover, in rare cases, synergistic heterogeneity, where disease results from multiple partial defects within one or more metabolic pathways (Vockley J et al 2000), and genetic modifiers (Wiesinger C et al 2015, Yildiz Y et al 2013) can play a role.

Given that only 35% of Mendelian diseases are solved by WES, the majority of undiagnosed cases are likely subject to limitations in variant-calling and prioritization, and the inability to detect intronic and regulatory pathogenic variants. Whole genome sequencing (WGS) enables complete coverage of the genome and has higher sensitivity than WES for certain coding variants, indels, chromosomal rearrangements and copy number variants (Ouwehand WH et al 2019). WGS may also detect short tandem repeat (STR) expansions in IEM, such as the recently described expansion of a GCA-repeat in

GLS, resulting in glutaminase deficiency (van Kuilenburg et al 2019). However, interpretation is often hindered by difficulty in prioritization of the vast numbers of variants detected and our incomplete understanding of the non-coding sequence, and so the diagnostic yield is only modestly increased to just over 40% by WGS (Clark MM et al 2018). Furthermore, a number of indels, somatic mutations, structural variants such as balanced translocations and inversions, and repeat expansions, may elude detection by current sequencing technologies and algorithms (Dolzhenko et al 2017). It is unclear to what extent these variants contribute to the diagnostic gap. Uncovering the true disease-causing variant(s) by WGS is therefore rendered no less difficult, with the primary challenge shifting from the capacity to detect variants, to the ability to distinguish true pathogenicity from the plethora of benign variants present between individuals, and to distinguish clinically relevant impact (MacArthur D et al 2014, Goldstein DB et al 2013).

This challenge can be addressed with a complementary approach for the parallel detection and functional annotation of variants utilizing genetic and advanced "multi-omics" methods (Wanders RJ et al 2018, Stenton and Prokisch 2018), such as transcriptomics, proteomics, metabolomics, glycomics, and lipidomics. Thereby promoting improved diagnostic yield in IEM and characterization of disease pathophysiology, such as in deciphering the role of the TIMMDC1 protein as a mitochondrial complex I assembly factor, by detection of reduced levels of complex I subunits on proteomics (Kremer LS et al 2017).

Here we summarize the current state of integrating transcriptomics and the novelty of integrating proteomics into the diagnostic pipeline - providing a perspective for the future. The role of metabolomics, which is already well established in the clinic, glycomics, and lipidomics in IEM diagnosis are reviewed comprehensively elsewhere (Graham E et al 2018, Bakar NA et al 2018, Lydic TA et al 2018). Though this review will focus mostly on studies of mitochondrial and neuromuscular disease (NMD) cohorts, the utility of a "multi-omics" diagnostic pipeline is not limited to one disease subset. Indeed, in the Kremer et al 2017 study, analysis was performed genome-wide, showing the far-reaching impact this approach can achieve in the diagnosis of the rare disease community for diseases of hitherto unclear etiology.

Genomics:

WES interrogates the exonic coding regions, constituting 2% of the human genome, and is a cost-effective comprehensive strategy to identify coding variation. An inescapable shortfall of WES is the inability to capture variants in the non-coding regions of the genome. This can be overcome by WGS, which allows detection of non-exonic variants in addition to structural variants (Ouwehand WH et al 2019, Taliun et al 2019, Lionel AC et al 2018, Merker JD et al 2018).

The development of WES and WGS is reflected by a synergistic improvement of bioinformatic pipelines for sequence alignment and variant calling (Supernat A et al 2018, Sandmann S et al 2017). WES yields on average 20,000 coding variants per individual (Van Hout CV et al 2019, Dewey FE et al 2016). Focusing on predicted amino acid-altering variants, around 9,000 variants remain. This challenge escalates considerably when interpreting WGS, which yields over 3 million variants per individual, 30,000 of which are rare (Ouwehand WH et al 2019, Taliun et al 2019). Moreover, despite the vast numbers of variants identified, numerous elude calling algorithms. Value is therefore found in analyzing genomic data with multiple bioinformatic pipelines (Sandmann S et al 2017).

The short variants called, SNVs (single nucleotide variants) and indels, are comprised mostly of common polymorphisms. A stringent filter of 0.1-1.0% MAF (minor allele frequency) to detect rare variants is therefore applied, yielding on average 100-400 variants per sample. This is inclusive of 20 rare loss-of-function variants (Van Hout CV et al 2019, Dewey FE et al 2016). To reveal the causative variant(s), population data (ExAC and gnomAD) (Lek M et al 2016), computational and predictive data (CADD, VEP, PolyPhen-2, SIFT, PON-P2, and MutationTaster2, among others) (Rentzsch P et al 2018, McLaren W et al 2016, Adzhubei et al 2013, Vaser R et al 2016, Niroula A et al 2015, Schwarz JM et al 2014), functional data, segregation data including the use of trio sequencing, and allelic data must be considered in aggregate (Figure 2). The American College of Medical Genetics (ACMG) proposes variant curation guidelines for the classification of variants using evidence derived from these factors (Richards S et al 2015).

Diseases from the IEM group most often show an autosomal recessive inheritance pattern. Discovery of known pathogenic variant(s) listed in disease-variant databases (Clinvar and HGMD (Human Gene Mutation Database)) (Landrum MJ et al 2017, Stenson PD et al 2017) or protein truncating loss-of-function variant(s) allows for a definitive diagnosis. However, in many cases, this rigorous step-wise approach to pathogenic variant discovery is inconclusive or results in a variant that is more difficult to interpret, such as rare missense, intronic, and synonymous variants. Strategies complementary to WES and WGS therefore need to be explored. Transcriptomics and quantitative proteomics have started to be used for the automated, high-throughput, and cost-effective functional validation of variants, replacing quantitative RT-PCR and cDNA sequencing, and western blot analysis to investigate the impact of a VUS on transcript and protein stability, respectively. Here we summarize their promising and emerging role in elucidating the missing heritability of Mendelian disease by providing invaluable data to localize causative variants when analyzed in parallel with genomic sequencing.

Transcriptomics:

The diagnostic power of combining DNA with RNA sequencing to tackle inconclusive WES cases is evident, achieving an additional gain in diagnostic yield of 10 to 35% in mitochondrial disease and NMD cohorts, respectively (Kremer LS et al 2017, Cummings BB et al 2017). RNA sequencing (RNA-seq) is a methodology for sequencing the entire transcriptome. It enables direct probing and quantification of the impact of coding as well as non-coding variants on RNA abundance and sequence. Furthermore, evidence is accumulating that pathogenic variants frequently occur deep within intronic regions, over 100 base pairs away from exon-intron junctions (Vaz-Drago R et al 2017), highlighting the important of direct interrogation of RNA-seq data as a complementary tool to identify variants eluding WES.

Three situations arise that can be interpreted to prioritize candidate disease-causing genes for a rare disease: aberrant expression, aberrant splicing and mono-allelic expression (MAE) (Figure 3). Aberrant expression, resulting in expression outliers, implies impaired gene expression and can result from both coding variants and non-coding variants in regulatory regions such as promoters, enhancers, and suppressors, in addition to RNA degradation through nonsense-mediated decay (NMD). Statistical testing is performed to compute these expression outliers as statistically significant deviations from normal physiological range, considering the population distribution (Kremer LS et al 2017).

Aberrant splicing is caused by variation affecting splice sites or splice motifs in coding and noncoding regions. These variants result in aberrant splice isoforms, which can be quantified, and regularly cause the generation of a premature stop codon, provoking degradation of the RNA by NMD. Deleterious variants can lead to a plethora of splicing abnormalities ranging from exon creation, skipping, extension and truncation to intron inclusion, due to activation of non-canonical splice sites or changes in splicing regulatory elements. Aberrant splicing is not a rare phenomenon amongst pathogenic variants, with 10-30% of pathogenic variants and VUS predicted to alter the native splice pattern (Lee M et al 2017, Soemedi R et al 2017), and accounting for 9% of variants reported in HGMD (HGMD database accessed on April 2019). Aberrant splicing is therefore well established as a significant cause of Mendelian disease.

MAE, or “allelic expression imbalance”, whereby one allele is silenced leaving only the other allele expressed, plays an important role in reprioritizing heterozygous rare SNVs called by WES and WGS (Gonorazky HD et al 2019, Kremer LS et al 2017, Cummings BB et al 2017, Falkenberg KD et al 2017). In the genomic analysis pipeline, these variants are filtered out due to the lack of a partner variant in biallelic autosomal recessive inheritance (Eilbeck K et al 2017). MAE promotes overrepresentation of such heterozygous variants, mimicking homozygosity, and fits the recessive mode of inheritance assumption.

When applying stringent filtering for rare events with strong effect size, appropriate for analysis in the rare and recessive disease setting, a systematic genome-wide analysis reveals a median of one aberrant expression, six MAE, and five aberrant splicing events per sample (Kremer LS et al 2017). The manageable number of detected events readily enables manual inspection of the data to identify the responsible pathogenic coding and non-coding variants, which were not originally detected or prioritized in the WES and WGS analysis pipeline.

There is growing interest in the bioinformatic field to further optimize the RNA-seq analysis pipeline for disentangling noise inherent to the sample from disease-relevant variation and in the development of appropriate statistical models for outlier detection. Challenges arise in the assessment of statistical significance, choice of arbitrary cut-offs, and the reliance on manual correction for confounders. For the detection of aberrant gene expression, the statistical model OUTRIDER has recently been developed to overcome these challenges (Brechtmann F et al 2018), including the adequate correction of hidden batch effect without compromise to the relevant sample information. Further method development is nevertheless required to handle aberrant splicing events. With ever increasing resolution and cohort sizes in the “omics” approaches, computational power may also become a limitation.

RNA-seq is not without limitation. The integrity of RNA molecules extracted from a tissue is of paramount importance for subsequent analysis. The RNA integrity number (RIN) is a reliable measure of RNA quality. Samples with a RIN>7 are well suited to transcriptomics. Furthermore, since RNA is highly sensitive to pre-analytical artefact, it provides a quality measure for the tissue itself, and quality control for further “omics” studies. Though any single tissue can be used for genomic sequencing, tissue selection requires careful consideration as the genes involved in a specific disease are not ubiquitously expressed across all tissues (Li X et al 2017). However, if a given gene is expressed, there is high correlation of gene regulation between different tissues (Qi T et al 2018). The 10,000 transcribed protein-coding genes reliably detected across tissue types typically represent around 65% of all genes and cover 65-78% of disease genes recognized by OMIM (Amberger JS et al 2014) (Figure 4). In IEM and in mitochondrial disease, a major subset of IEM, 68-85% and 75-88% of known disease genes (Ferreira CR et al 2019, Stenton and Prokisch 2018) are reliably detected, respectively. The usage of up-to-date annotations such as GTEx (Genotype-Tissue Expression Consortium 2015) or GENCODE (Harrow J 2012), strand-specific sequencing, and the inclusion of non-coding transcripts, is further increasing the number of genes which can be analyzed. Tissue selection and tissue-specific gene expression however remain caveats in RNA-seq as obtaining a complete analysis of the transcriptome by interrogating one tissue is not biologically feasible. Furthermore, the top candidate tissue for testing - the primary affected tissue - is frequently inaccessible for sampling, necessitating choice of a surrogate tissue. Selection of the optimal tissue for study can be guided by gene expression databases, such as GTEx comprising approximately 11,500 RNA-seq samples across more than 50 tissue sites (https://gtexportal.org/home/) and the recently developed web-based tool PAGE (Panel Analysis of Gene Expression) (Gonorazky HD et al 2019).

RNA-seq is fruitless if the underlying causative variant does not influence RNA abundance and sequence, as is frequently the case for missense mutations, accounting for a large proportion of the variant burden in Mendelian disease. To discover the impact of such variants, in addition to providing independent validation of RNA-seq findings, proteomics comes into play.

Proteomics:

Proteomics gives insight into the composition, structure and function of the proteome (Aebersold R and Mann M 2016) and to protein activity, interaction, and localization. Moreover, interrogation of proteomic data is a powerful approach to reveal impaired protein synthesis, folding, stability, and degradation, due to the detection of reduced or diminished protein level and in the detection of specific proteomic signatures (Kremer LS et al 2017).

Proteomics has been introduced for functional validation of variants in multiple studies, reliably detecting 4,500-5,500 proteins (Kremer LS et al 2017). Given advances in technology since, both the number of proteins detected and the sample throughput have seen substantial improvement. Over 8,000 proteins, out of 10,000 proteins expressed in a given cell type (Frejno M et al 2017, Lapek Jr JD et al 2017, Roumeliotis TI et al 2017, Lawrence RT et al 2015), have been shown to be reliably quantified with around 6,000 proteins found in common across batches (McAlister GC et al 2014). High-throughput is enabled by multiplexed quantification using isobaric chemical tags, for example Tandem Mass Tags (TMT), to analyze up to 10 samples in parallel (Thompson A et al 2003) (Figure 5). TMT-labeling in combination with extensive fractionation and state-of-the-art mass spectrometry (MS3-level quantification) has proven to significantly improve quantitative precision and accuracy, to increase the dynamic quantification range, and to provide a highly robust and reproducible detection method for the vast majority of the expressed proteome in a given cellular system (McAlister GC et al 2014).

Alike to RNA-seq, outliers can be detected in the proteomic data. Strongly reduced protein level in one sample compared to other samples or to controls, provides robust evidence for pathogenicity of underlying variants in autosomal recessive traits. Therefore, a proteome-wide screen for expression outliers provides valuable information - by inferring the protein encoding gene(s) in which pathogenic variants may harbor - and paves the way for proteomics, above and beyond functional validation, as a routine test in the diagnostic pipeline (Kremer LS et al 2017). Our own preliminary data of ~100 proteomes indicate a median of less than five protein expression outliers out of 7,500 quantified proteins per sample, a feasible number for further exploration (Prokisch H 2018).

Proteins frequently function in large, dynamic and labile multiprotein complexes, the function and stability of which depends on the availability of all constituents. One can therefore look not only for aberrant expression of an isolated protein, but for multiple constituents of a protein complex adding evidence for the pathological characterization of underlying VUS (Lake NJ et al 2017, Kremer LS et al 2017, Stroud DA et al 2016). This is exemplified by a powerful study coupling gene-editing technology with quantitative proteomics in dissecting the essentiality of the 45 subunits of mitochondrial respiratory chain complex I for its full assembly and function (Stroud DA et al 2016). Furthermore, complexome analysis, specifically developed to yield insight into the size, composition, and stability of protein complexes, is a powerful method to study the impact of VUS on protein complex assembly. Complexome analysis separates native intact proteins and protein complexes, indicating the composition of functional complexes by the identification of co-migrating proteins (Heide H and Wittig I 2013). In our hands, over 4,500 proteins are detected in total, around 50% of which appear within protein complexes (Prokisch H 2018). Complexome analysis has proven fruitful in novel disease gene validation (Heide H et al 2012).

The main limitations of proteomics are tissue selection, as with RNA-seq, in addition to measurement sensitivity which is limited due to the large dynamic range of protein concentrations across sample tissues.

Integration of transcriptomics and proteomics for functional annotation:

No single methodology can capture the complexity of the all molecular events resulting in a genetic disease (Karczewski KJ and Snyder MP 2018). Therefore, for the discovery of novel disease-genes and pathogenic variants, analysis of inconclusive WES samples in parallel with transcriptomics and proteomics in a "multi-omics pipeline" is advocated. All three "omics" methodologies are analyzed on the gene-level. This allows them to be used in a truly complementary fashion - tracking the impact of a variant along the pathway of RNA abundance and form, to protein level, and vice versa, tracing aberrant expression and splicing pattern back to the responsible protein-encoding or regulatory variants - with integration of results to localise causative variants on the gene-level. Both the transcriptome and proteome pipelines result in less than 20 outliers per sample. In a diagnostic setting, this allows manual inspection and validation of significant observations using additional visualization tools, such as IGV (Integrative Genome Viewer), Sashimi, and volcano plots (Figure 6). Counting of normal and aberrant splice isoforms, expressed alleles, and protein levels provides a quantitative measure of the functional consequence of the genetic variation, alike to an enzyme activity measurement. All require reference ranges and interpretation. In some cases it may prove difficult to pinpoint the causative variant, necessitating further follow-up studies to reach a firm diagnosis. The integration of additional "omics" is more challenging as they cannot easily be mapped on the gene-level.

Examples of the value of multi-omics analysis are: a case of TIMMDC1 defect due to an aberrant splicing event, further supported by aberrantly low RNA and protein expression in addition to instability of other complex I subunits (Figure 6) and, a case of ALDH18A1 defect due to MAE of a variant affecting translation or protein stability, resulting in aberrantly low protein expression, and further validated by integration of the metabolic profile in accordance with a defect in proline metabolism (Kremer LS et al 2017).

For the establishment of a definitive diagnosis, transcriptomics and proteomics act as independent validation for one another. When the variants under interrogation harbor within a known disease gene and in a patient with a clinical phenotype in-keeping with the known genotype-phenotype association, a diagnosis is directly facilitated by these holistic methodologies - negating the need for further sophisticated, time-consuming and individually tailored functional validation experiments of frequently found VUS. Conversely, when the variants harbors within a candidate gene, or when the clinical phenotype is incongruent with the known genotype-phenotype association, further validation such as complementation (Figure 7) and supplementation in patient derived cell-lines or replication of the variant(s) or other variants in the gene of question in multiple cases with similar phenotypes, is required to establish solved status. Collaboration via platforms such as GeneMatcher (Bruel AL et al 2018), Matchmaker Exchange (Sobreira NL et al 2017), and GENOMIT (GENOMIT.eu), among others, supports the collection of cases.

Moreover, as the results of RNA-seq and proteomics are quantitative, the level of residual mRNA or protein of the affected protein and potential modifying factors can be explored in the context of providing an explanation for the phenotypic spectrum observed in IEM (Wiesinger C et al 2015, Yildiz Y et al 2013).

Essential to the interpretation of these large “omics” data sets and to the increase in diagnostic yield, is the inclusion of phenotype data (Deelen P et al 2018). HPO “Human Phenotype Ontology” terms describe 13,000 clinical abnormalities with 156,000 annotations to hereditary diseases - mapped into an ontological, multi-tiered structure (Köhler S et al 2016). HPO terms enable standardization of data and promote the inclusion of clinical information into computational method development, such in Exomiser and PhenIX - phenotype-driven approaches to exome analysis (Pengelly RJ et al 2017) - among others, thereby maximizing interpretive capacity. When a genotype-phenotype correlation is well established, deep phenotyping allows prioritization of VUS in a strong candidate gene, which may routinely be overlooked, and so initiates targeted or “omics” functional follow-up studies (Lord J et al 2019).

Future directions

Moving forward, long read RNA-seq can be employed to capture complex isoforms by sequencing the full-length cDNA copies of RNA molecules, not possible with short-read technologies due to the fragmentation of transcripts for sequencing. These techniques can also be applied at the single-cell level, to remove transcriptome diversity across bulk heterogeneous cell populations, and to illuminate the transcriptional profile and aberrant transcriptional events in a single-cell type (Byrne A et al 2017).

Through the identification and validation of novel variants, "omics" technologies allow the discovery of novel points of therapeutic intervention, such as in novel splice variants amenable to treatment with antisense oligonucleotides (AONs). AONs are short synthetic DNA or RNA molecules that bind to complementary RNA sequence to modulate pre-mRNA splicing outcome (Anna A et al 2018), and are now approved in spinal muscular atrophy (SMA) and Duchenne's muscular dystrophy (DMD) (Bergsma AJ et al 2018). Here, RNA-seq may also play a future role in monitoring treatment effect by the quantification of generation of the properly spliced transcripts.

Conclusion

The innovative "multi-omics" approach, uniting genomics, transcriptomics and proteomics, is invaluable in order to pinpoint the molecular cause of Mendelian disease. It is a rapidly expanding field facilitated by systematic DNA, RNA, and protein profiling, and the continual refinement of methodology, statistical modelling, and bioinformatic analyses, combined with expert curation.

Figure legends:

(Figure 1) Next generation sequencing (NGS) overview. The NGS library is prepared by fragmenting genomic DNA (in WES and WGS) or cDNA (in RNA-seq). To produce cDNA for RNA-seq, mRNA is selected from total RNA by poly-A selection, followed by synthesis of the first and second strand of cDNA. Once the DNA or cDNA is fragmented, adapters and barcodes are ligated to allow multiplexing of samples. The prepared library is amplified and sequenced prior to reference genome alignment. Once reads are aligned to the genome, data analysis such as calling identified differences between the reference and the sequenced reads as variants can commence.

(Figure 2) The conventional genomic approach to the diagnosis of Mendelian disease. Variants called from whole exome sequencing (WES) and whole genome sequencing (WGS) are filtered and prioritized based on clinical and genetic data. This includes utilization of variant and genotype-phenotype association databases such as ClinVar (Landrum MJ et al 2017), HGMD (Stenson PD et al 2017), and OMIM (Amberger JS et al 2014). Typically, this rigorous stepwise approach does not reach a definitive diagnosis in over 50% of cases - these cases remain inconclusive and often harbor variants of uncertain significance (VUS).

(Figure 3) The “multi-omics” approach to the diagnosis of Mendelian disease. Genomic data from inconclusive WES and WGS cases is analyzed in parallel with transcriptomic and proteomic data. Transcriptomics enables detection of aberrant transcript expression, aberrant splicing, and mono-allelic expression (MAE) events, while proteomics enables detection of aberrant protein expression - providing evidence to localize causative variants on the gene-level. Integration of transcriptomics into the analysis pipeline achieves diagnostic gains of >10% in unsolved cases, this number is expected to increase with the integration of proteomics.

(Figure 4) Tissue selection for transcriptomic investigation. Here we give examples of investigated tissue types using transcriptomics. The number of detected genes varies between 10,000-15,000. The number of detected IEM and mitochondrial disease genes is shown by tissue. Almost two thirds of OMIM disease genes (Online Mendelian Inheritance in Man - a comprehensive compendium of human genes and genetic phenotypes) are detected.

(Figure 5) Schematic representation of a multiplexed proteomic experiment. Example of a Tandem Mass Tag (TMT) labelling proteomic experiment, which allows us to quantify almost 8,000 proteins in a human fibroblast cell line. The multiplexing of samples reduces missing values and improves quantitative precision and statistical analysis.

(Figure 6) An example of the value of “multi-omics” analysis. Analysis of an unsolved mitochondrial disease case revealed *TIMMDC1,* a complex I assembly factor, as the underlying diagnosis of this patient’s disease. This is due to an aberrant splicing event (red-affected in comparison orange-control) resulting in exon inclusion, a premature stop codon, and nonsense mediated decay of the transcript (a). In the Sashimi plot, 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. The gene model of the RefSeq annotation is depicted and the aberrantly spliced exon is coloured in red beneath. This is supported by aberrantly low RNA expression visualized by a volcano plot (b) and aberrantly low protein expression of TIMMDC1 but also instability of other complex I subunits (red) visualized by a scatter plot of protein versus RNA log-fold change (c). The volcano plot shows p-values and Z-scores for all transcripts detected in the sample. The scatter plot summarizes protein and RNA changes in a given sample against the median of all other samples investigated. The underlying variant was found to be a homozygous deep intronic variant of unknown significance (VUS) which was subsequently validated by complementation of the wild-type (WT) gene into the patient-derived fibroblast cell line and cellular phenotypic rescue (d).

(Figure 7) Lentiviral-mediated complementation experiment in patient derived fibroblast cell lines. Wild-type (WT) cDNA of the gene of interest is cloned into the pLenti vector using appropriate primers. The pLenti expression construct containing the gene of interest plus the lentiviral packaging mix is co-transfected into 293T/293FT cells using the Lipofectamine reagent. Following an incubation period to allow transcription and synthesis of the viral RNA and viral proteins and viral budding, the supernatant containing the lentivirus is collected and mixed with cell culture medium. The lentivirus is applied to patient and control cell lines which subsequently undergo blasticidin selection to select successfully transduced cells. The complete method can be found in Kremer LS and Prokisch H 2017.

Acknowledgements

The proteomic data has been generated at BayBioMS, TU Munich in collaboration with Christina Ludwig.

References:

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*. 2013;76(1):7.20. 1-7.20. 41.

Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016 Sep;537(7620):347.

Amberger JS, Bocchini CA, Schiettecatte F, et al. OMIM. org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2014;43(D1):D789-D798.

Anna A, Monika G. Splicing mutations in human genetic disorders: Examples, detection, and confirmation. *J Appl Genet*. 2018;59(3):253-268.

Bakar NA, Lefeber DJ, van Scherpenzeel M. Clinical glycomics for the diagnosis of congenital disorders of glycosylation. *J Inherit Metab Dis*. 2018;41(3):499-513.

Bergsma AJ, in't Groen SL, van den Dorpel, et al. A genetic modifier of symptom onset in pompe disease. *EBioMedicine*. 2019.

Bergsma AJ, van der Wal E, Broeders M, et al. Alternative splicing in genetic diseases: Improved diagnosis and novel treatment options. In: *International review of cell and molecular biology.* Vol 335. Elsevier; 2018:85-141.

Brechtmann F, Mertes C, Matusevičiūtė A, et al. OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data. *The American Journal of Human Genetics*. 2018;103(6):907-917.

Bruel AL, Vitobello A, Mau-Them FT, et al. 2.5 years' experience of GeneMatcher data-sharing: A powerful tool for identifying new genes responsible for rare diseases. *Genet Med*. 2018. doi: 10.1038/s41436-018-0383-z [doi].

Byrne A, Beaudin AE, Olsen HE, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature communications*. 2017;8:16027.

Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ genomic medicine. 2018;3.

Coene KL, Kluijtmans LA, van der Heeft E, et al. Next-generation metabolic screening: Targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *J Inherit Metab Dis*. 2018;41(3):337-353.

Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9(386):10.1126/scitranslmed.aal5209. doi: eaal5209 [pii].

Deelen P, van Dam S, Herkert JC, et al. Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis. *bioRxiv*. 2018:375766.

Dewey FE, Murray MF, Overton JD, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016 Dec 23;354(6319):aaf6814.

Dolzhenko E, van Vugt JJFA, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27(11):1895-1903. doi: 10.1101/gr.225672.117 [doi].

Eilbeck K, Quinlan A, Yandell M. Settling the score: Variant prioritization and mendelian disease. *Nature Reviews Genetics*. 2017;18(10):599.

Falkenberg KD, Braverman NE, Moser AB, et al. Allelic expression imbalance promoting a mutant PEX6 allele causes zellweger spectrum disorder. *The American Journal of Human Genetics*. 2017;101(6):965-976.

Ferreira CR, van Karnebeek CD, Vockley J, et al. A proposed nosology of inborn errors of metabolism. *Genetics in Medicine*. 2019;21(1):102.

Frejno M, Chiozzi RZ, Wilhelm M, et al. Pharmacoproteomic characterisation of human colon and rectal cancer. Molecular systems biology. 2017 Nov 1;13(11):951.

Goldstein DB, Allen A, Keebler J, et al. Sequencing studies in human genetics: Design and interpretation. *Nature Reviews Genetics*. 2013;14(7):460.

Graham E, Lee J, Price M, et al. Integration of genomics and metabolomics for prioritization of rare disease variants: A 2018 literature review. *J Inherit Metab Dis*. 2018;41(3):435-445.

GTEx Consortium. Human genomics. the genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660. doi: 10.1126/science.1262110 [doi].

Gonorazky HD, Naumenko S, Ramani AK, et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare mendelian disease. *The American Journal of Human Genetics*. 2019;104(3):466-483.

Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research. 2012 Sep 1;22(9):1760-74.

Heide H, Bleier L, Steger M, et al. Complexome profiling identifies TMEM126B as a component of the mitochondrial complex I assembly complex. Cell metabolism. 2012 Oct 3;16(4):538-49.

Heide H, Wittig I. Methods to analyse composition and dynamics of macromolecular complexes. Biochemical Society Transactions. 2013 Sept 23;41(5): 1235-1241

Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018 May;19(5):299.

Köhler S, Vasilevsky NA, Engelstad M, et al. The human phenotype ontology in 2017. *Nucleic Acids Res*. 2016;45(D1):D865-D876.

Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of mendelian disorders via RNA sequencing. *Nature communications*. 2017;8:15824.

Kremer LS, Prokisch H. Identification of disease-causing mutations by functional complementation of patient-derived fibroblast cell lines. InMitochondria 2017 (pp. 391-406). Humana Press, New York, NY.

Lake NJ, Compton AG, Rahman S, et al. Leigh syndrome: One disorder, more than 75 monogenic causes. *Ann Neurol*. 2016;79(2):190-203.

Lake NJ, Webb BD, Stroud DA, et al. Biallelic mutations in MRPS34 lead to instability of the small mitoribosomal subunit and leigh syndrome. *The American Journal of Human Genetics*. 2017;101(2):239-254.

Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2017 Nov 20;46(D1):D1062-7.

Lapek Jr JD, Greninger P, Morris R, et al. Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. Nature biotechnology. 2017 Oct;35(10):983.

Lawrence RT, Perez EM, Hernández D, et al. The proteomic landscape of triple-negative breast cancer. Cell reports. 2015 Apr 28;11(4):630-44.

Lee M, Roos P, Sharma N, et al. Systematic computational identification of variants that activate exonic and intronic cryptic splice sites. *The American Journal of Human Genetics*. 2017;100(5):751-765.

Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285.

Li X, Kim Y, Tsang EK, et al. The impact of rare variation on gene expression across tissues. Nature. 2017 Oct;550(7675):239.

Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genetics in Medicine. 2018 Apr;20(4):435.

Lord J, Gallone G, Short PJ, et al. Pathogenicity and selective constraint on variation near splice sites. *Genome Res*. 2019;29(2):159-170. doi: 10.1101/gr.238444.118 [doi].

Lydic TA, Goo Y. Lipidomics unveils the complexity of the lipidome in metabolic diseases. *Clinical and translational medicine*. 2018;7(1):4.

McAlister GC, Nusinow DP, Jedrychowski MP, et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Analytical chemistry. 2014 Jul 3;86(14):7150-8.

MacArthur D, Manolio T, Dimmock D, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469.

McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122.

Merker JD, Wenger AM, Sneddon T, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine. 2018 Jan;20(1):159.

Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction method for fast and reliable identification of harmful variants. *PloS one*. 2015;10(2):e0117380.

Ouwehand WH, et al. Whole-genome sequencing of rare disease patients in a national healthcare system. *bioRxiv*. 2019:507244.

Pengelly RJ, Alom T, Zhang Z, et al. Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Scientific reports*. 2017;7(1):13509.

Posey JE, Harel T, Liu P, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N Engl J Med*. 2017;376(1):21-31.

Prokisch H. 2018. Transcriptomics in Rare Disease Diagnosis. SSEIM, 4-7 Sept 2018, Athens Greece.

Qi T, Wu Y, Zeng J, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nature communications. 2018;9.

Rentzsch P, Witten D, Cooper GM, et al. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2018;47(D1):D886-D894.

Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*. 2015;17(5):405.

Roumeliotis TI, Williams SP, Gonçalves E, et al. Genomic determinants of protein abundance variation in colorectal cancer cells. Cell reports. 2017 Aug 29;20(9):2201-14.

Sandmann S, De Graaf AO, Karimi M, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific reports*. 2017;7:43169.

Schwarz JM, Cooper DN, Schuelke M, et al. MutationTaster2: Mutation prediction for the deep-sequencing age. *Nature methods*. 2014;11(4):361.

Sobreira NL, Arachchi H, Buske OJ, et al. Matchmaker exchange. *Current protocols in human genetics*. 2017;95(1):9.31. 1-9.31. 15.

Soemedi R, Cygan KJ, Rhine CL, et al. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. 2017;49(6):848.

Stenson PD, Mort M, Ball EV, et al. The human gene mutation database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. 2017;136(6):665-677.

Stenton SL, Prokisch H. Advancing genomic approaches to the molecular diagnosis of mitochondrial disease. *Essays Biochem*. 2018;62(3):399-408. doi: 10.1042/EBC20170110 [doi].

Stroud DA, Surgenor EE, Formosa LE, et al. Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature*. 2016;538(7623):123.

Supernat A, Vidarsson OV, Steen VM, et al. Comparison of three variant callers for human whole genome sequencing. *Scientific reports*. 2018;8(1):17851.

Tarailo-Graovac M, Shyr C, Ross CJ, et al. Exome sequencing and the management of neurometabolic disorders. *N Engl J Med*. 2016;374(23):2246-2255.

Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *BioRxiv*. 2019:563866.

Thompson A, Schäfer J, Kuhn K, et al. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem*. 2003;75(8):1895-1904.

Van Hout CV, Tachmazidou I, Backman JD, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK biobank. *bioRxiv*. 2019:572347.

van Kuilenburg AB, Tarailo-Graovac M, Richmond PA, et al. Glutaminase deficiency caused by short tandem repeat expansion in GLS. *N Engl J Med*. 2019;380(15):1433-1441

Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. *Hum Genet*. 2017;136(9):1093-1111.

Vaser R, Adusumalli S, Leng SN, et al. SIFT missense predictions for genomes. *Nature protocols*. 2016;11(1):1.

Vockley J, Rinaldo P, Bennett MJ, et al. Synergistic heterozygosity: Disease resulting from multiple partial defects in one or more metabolic pathways. *Mol Genet Metab*. 2000;71(1-2):10-18.

Wanders RJ, Vaz FM, Ferdinandusse S, et al. Translational metabolism: A multidisciplinary approach towards precision diagnosis of inborn errors of metabolism in the omics era. *J Inherit Metab Dis*. 2018.

Waters D, Adeloye D, Woolham D, et al. Global birth prevalence and mortality from inborn errors of metabolism: A systematic analysis of the evidence. *Journal of global health*. 2018;8(2).

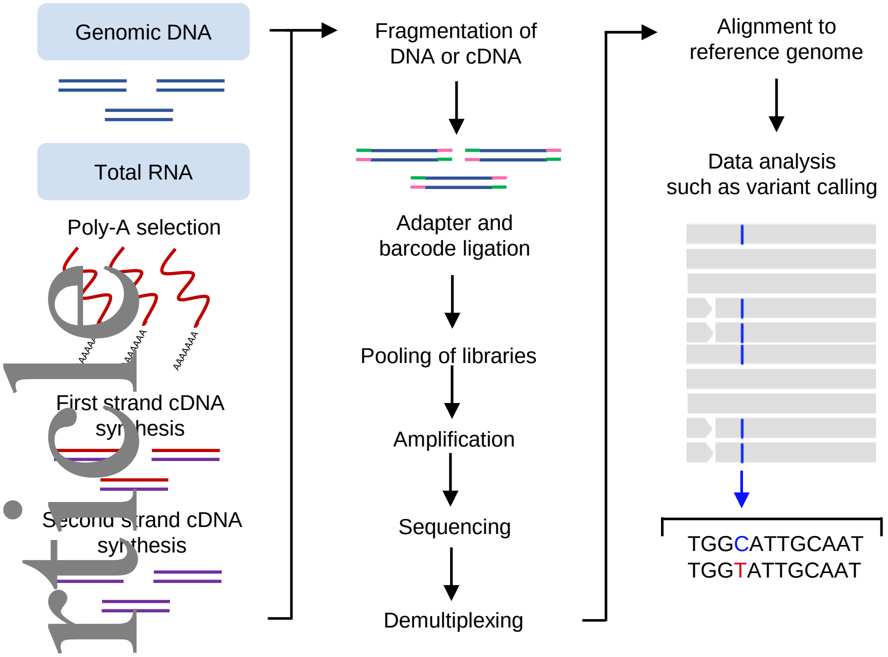
Wiesinger C, Eichler FS, Berger J. The genetic landscape of X-linked adrenoleukodystrophy: Inheritance, mutations, modifier genes, and diagnosis. *Appl Clin Genet*. 2015;8:109-121. doi: 10.2147/TACG.S49590 [doi].

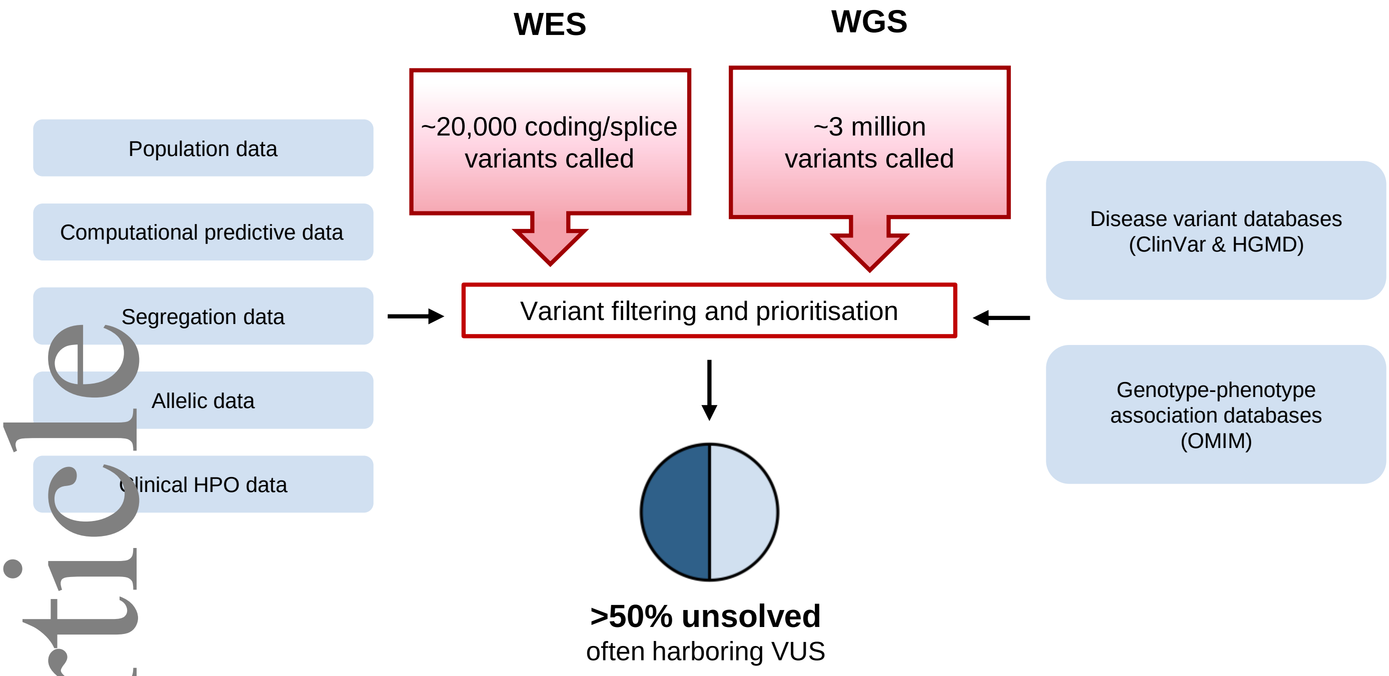
Wortmann SB, Mayr JA, Nuoffer JM, et al. A guideline for the diagnosis of pediatric mitochondrial disease: The value of muscle and skin biopsies in the genetics era. *Neuropediatrics*. 2017;48(04):309-314.

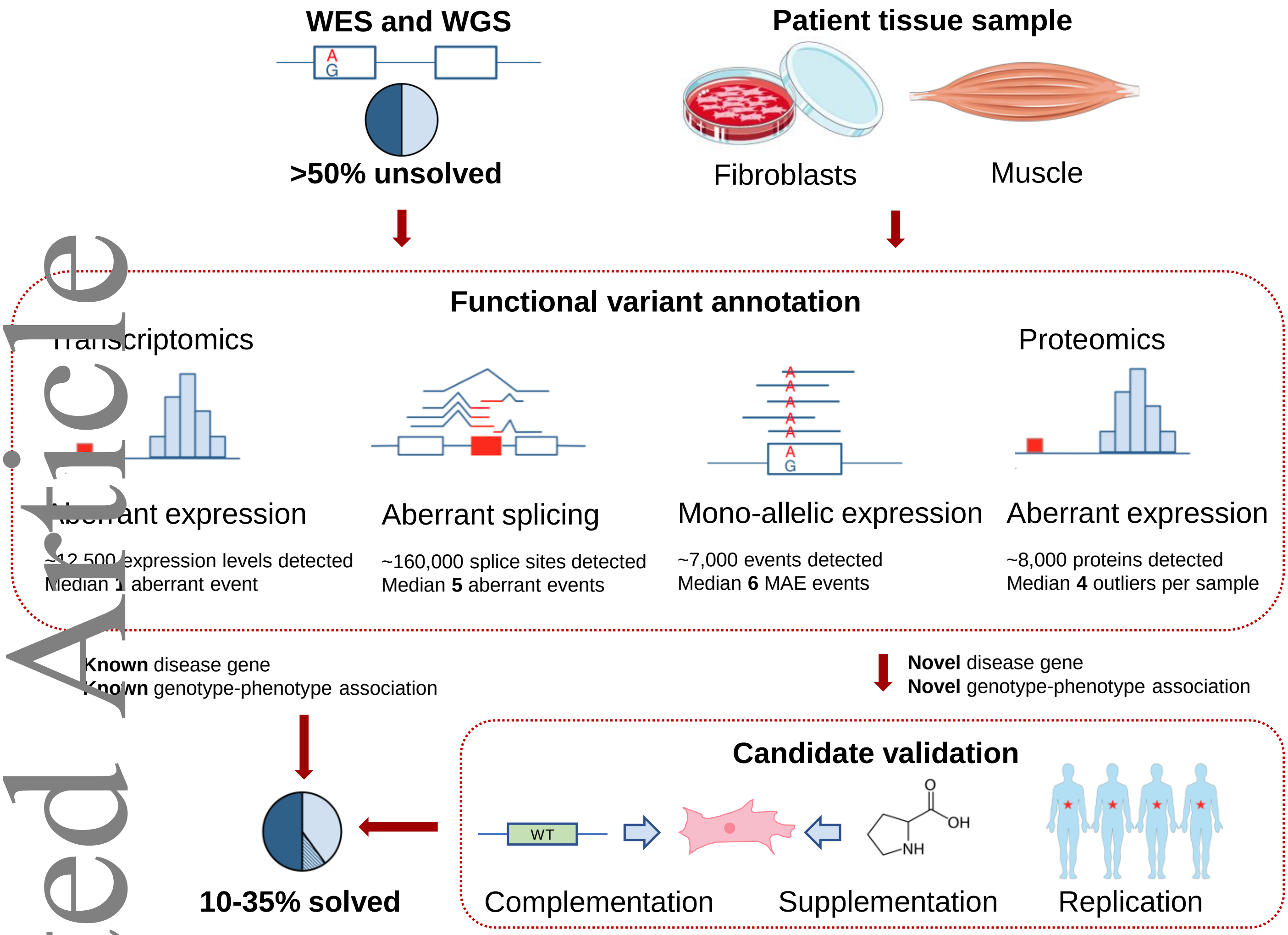
Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*. 2013;369(16):1502-1511.

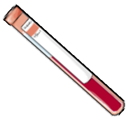
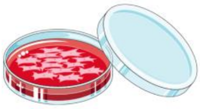
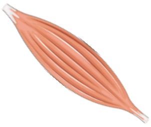
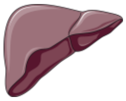
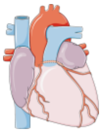
Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*. 2014;312(18):1870-1879.

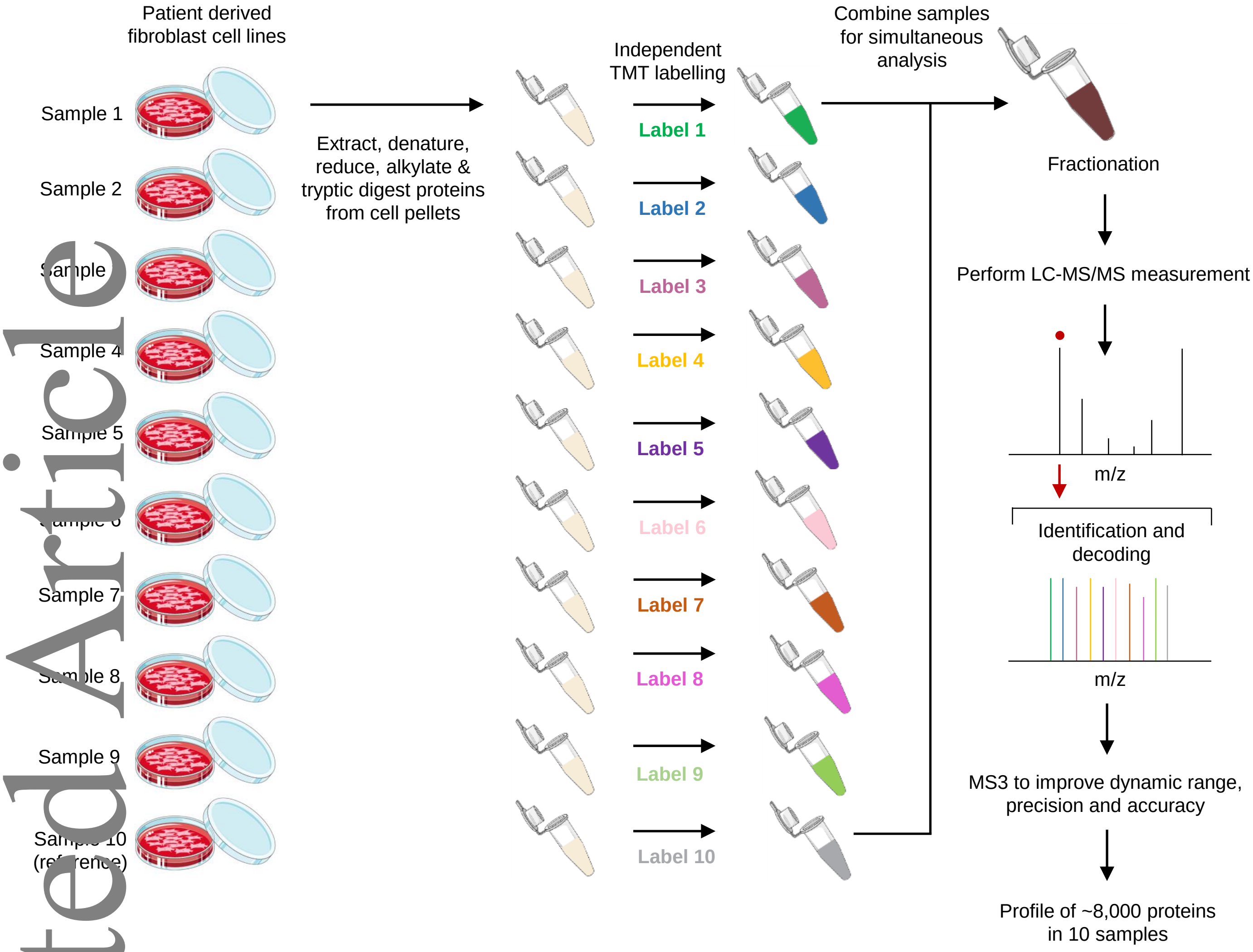
Yildiz Y, Hoffmann P, Vom Dahl S, et al. Functional and genetic characterization of the non-lysosomal glucosylceramidase 2 as a modifier for gaucher disease. *Orphanet journal of rare diseases*. 2013;8(1):151.

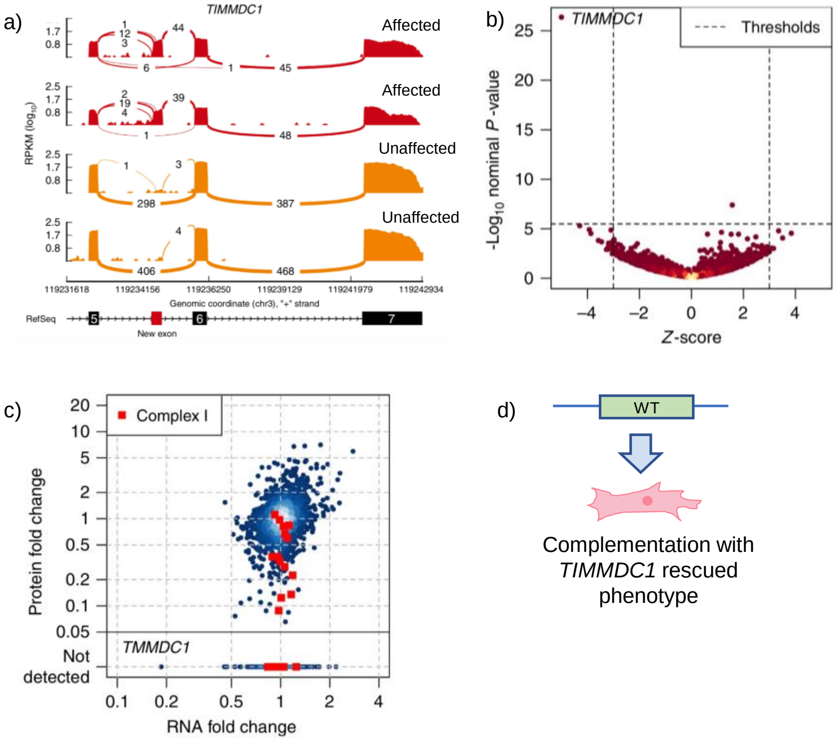

Genomic DNA
Total RNA
Poly-A selection
AAAAAA
AAAAAAA
AAAAAAA
First strand cDNA synthesis
Second strand cDNA synthesis
Fragmentation of DNA or cDNA
Adapter and barcode ligation
Pooling of libraries
Amplification
Sequencing
Demultiplexing
Alignment to reference genome
Data analysis such as variant calling
TGGCATTGCAAT
TGGTATTGCAAT

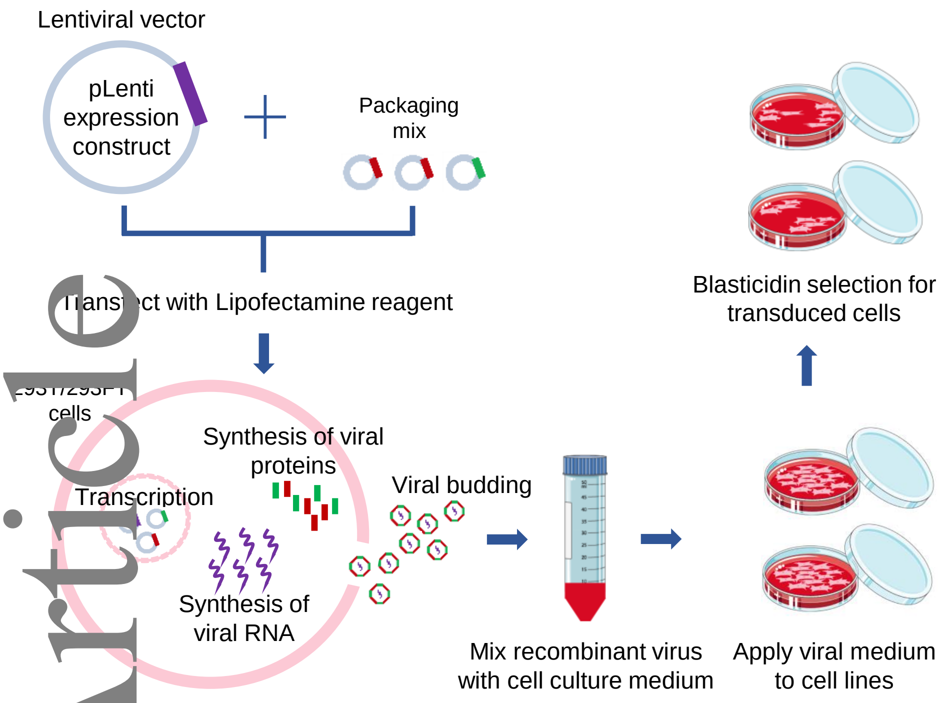

WES
WGS
~20,000 coding/splice variants called
~3 million variants called
Population data
Computational predictive data
Segregation data
Allelic data
Clinical HPO data
Variant filtering and prioritisation
Disease variant databases (ClinVar & HGMD)
Genotype-phenotype association databases (OMIM)
>50% unsolved
often harboring VUS


WES and WGS
A
G
>50% unsolved
Patient tissue sample
Fibroblasts
Muscle
Functional variant annotation
Transcriptomics
Proteomics
Aberrant expression
~12,500 expression levels detected
Median **1** aberrant event
Aberrant splicing
~160,000 splice sites detected
Median **5** aberrant events
A
A
A
A
A
A
G
Mono-allelic expression
~7,000 events detected
Median **6** MAE events
Aberrant expression
~8,000 proteins detected
Median **4** outliers per sample
Known disease gene
Known genotype-phenotype association
Novel disease gene
Novel genotype-phenotype association
Candidate validation
WT
O
OH
NH
Complementation
Supplementation
Replication
10-35% solved

Detected Transcripts	Blood	Fibroblasts	Muscle	Kidney	Liver	Heart
						
RefSeq genes	10,807	11,051	11,687	14,300	13,855	12,543
OMIM disease genes (n=3,768)	2,432 (65%)	2,574 (68%)	2,853 (76%)	2,952 (78%)	2,929 (78%)	2,914 (77%)
IEM disease genes (n=1,022)	694 (68%)	737 (72%)	767 (75%)	867 (85%)	873 (85%)	788 (77%)
Mitochondrial disease genes (n=325)	245 (75%)	266 (82%)	281 (86%)	286 (88%)	284 (87%)	278 (86%)


Patient derived fibroblast cell lines
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10 (reference)
Extract, denature, reduce, alkylate & tryptic digest proteins from cell pellets
Independent TMT labelling
Label 1
Label 2
Label 3
Label 4
Label 5
Label 6
Label 7
Label 8
Label 9
Label 10
Combine samples for simultaneous analysis
Fractionation
Perform LC-MS/MS measurement
m/z
Identification and decoding
m/z
MS3 to improve dynamic range, precision and accuracy
Profile of ~8,000 proteins in 10 samples


a)
TIMMDC1
Affected
Affected
Unaffected
Unaffected
RPKM ($\log_{10}$)
2.5
1.7
0.8
1
12
3
44
6
1
45
2
19
4
39
1
48
1
3
298
387
4
406
468
119231618
119234156
119236250
119239129
119241979
119242934
Genomic coordinate (chr3), "+" strand
RefSeq
5
6
7
New exon
b)
TIMMDC1
Thresholds
$-\text{Log}_{10}$ nominal P-value
25
20
15
10
5
0
−4
−2
0
2
4
Z-score
c)
Complex I
Protein fold change
20
10
5
2
1
0.5
0.2
0.1
0.05
Not detected
TMMDC1
0.1
0.2
0.5
1
2
4
RNA fold change
d)
WT
Complementation with *TIMMDC1* rescued phenotype


Lentiviral vector
pLenti expression construct
Packaging mix
Transfect with Lipofectamine reagent
293T/293FT cells
Synthesis of viral proteins
Transcription
Synthesis of viral RNA
Viral budding
Mix recombinant virus with cell culture medium
Apply viral medium to cell lines
Blasticidin selection for transduced cells