Methods in the Assessment of Genotype-Phenotype Correlations in Rare Childhood Disease Through Orthogonal Multi-omics, High-throughput Sequencing Approaches

by

Szabolcs Szelinger

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2015 by the
Graduate Supervisory Committee:

David W. Craig, Co-Chair
Kenro Kusumi, Co-Chair
Michael S. Rosenberg
Matthew J. Huentelman
Vinodh Narayanan

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Rapid advancements in genomic technologies have increased our understanding of rare human disease. Generation of multiple types of biological data including genetic variation from genome or exome, expression from transcriptome, methylation patterns from epigenome, protein complexity from proteome and metabolite information from metabolome is feasible. "Omics" tools provide comprehensive view into biological mechanisms that impact disease trait and risk. In spite of available data types and ability to collect them simultaneously from patients, researchers still rely on their independent analysis. Combining information from multiple biological data can reduce missing information, increase confidence in single data findings, and provide a more complete view of genotype-phenotype correlations. Although rare disease genetics has been greatly improved by exome sequencing, a substantial portion of clinical patients remain undiagnosed. Multiple frameworks for integrative analysis of genomic and transcriptomic data are presented with focus on identifying functional genetic variations in patients with undiagnosed, rare childhood conditions. Direct quantitation of X inactivation ratio was developed from genomic and transcriptomic data using allele specific expression and segregation analysis to determine magnitude and inheritance mode of X inactivation. This approach was applied in two families revealing non-random X inactivation in female patients. Expression based analysis of X inactivation showed high correlation with standard clinical assay. These findings improved understanding of molecular mechanisms underlying X-linked disorders. In addition multivariate outlier analysis of gene and exon level data from RNA-seq using Mahalanobis distance, and its integration of distance scores with genomic data found genotype-phenotype correlations in variant prioritization process in 25 families. Mahalanobis distance scores revealed variants with large transcriptional impact in patients. In this dataset, frameshift variants were more likely result in outlier expression signatures than other types of functional variants. Integration of outlier estimates with genetic variants corroborated previously identified, presumed causal variants and highlighted new candidate in previously un-diagnosed case. Integrative genomic approaches in easily attainable tissue will facilitate the search for biomarkers that impact disease trait, uncover

pharmacogenomics targets, provide novel insight into molecular underpinnings of un-characterized conditions, and help improve analytical approaches that use large datasets.

DEDICATION

This work is dedicated to those people without whom I would not have gotten to this milestone:

My wife, Edina Pletenyik.

My daughter, Ilona Szelinger

My mother, Ildiko Herneczky.

My grandfather, Tibor Herneczky.

My grandmother, Edit Ebenhoch.

My friend, Jason Corneveaux.

ACKNOWLEDGMENTS

Funding Support: This work was supported by the Center for Rare Childhood Disorders Foundation, The Stardust Foundation, and TGen's Salt River (SRP) Graduate Student Support Program.

The author would like to extend his appreciation to the families of the Center for Rare Childhood Disorder Center without whom this work would not be possible.

In addition, the author must recognize and express gratitude to his mentor, Dr. David Craig for giving the opportunity to be part of numerous forward-looking projects over the years, for all the advice, support, and patience he provided, the numerous discussions, brainstorming sessions that resulted in new appreciation and excitement for the author's work.

The author is extremely grateful for the mentoring, career guidance and support provided by the graduate committee members, Dr. Matthew Huentelman, Dr. Kenro Kusumi, Vinodh Narayanan MD, and Dr. Michael Rosenberg.

The author must recognize his internship mentor Dr. Keith Coon at TGen, and his professors at Mesa Community College, Dr. Stan Kikkert and Dr. Lewis Obermiller for their career and educational guidance over the years.

Technical support: The author must recognize all the people who gave assistance in bioinformatics, statistical analysis and laboratory procedures, by members of Translational Genomics Research Institute. This includes Jessica Aldrich, April Allen, Alexis Christoforides, Jason Corneveaux, Dr. Brooke Hjelm, Tyler Izatt, Valerie Jones, Ahmet Kurdoglu, Christophe Legendre, Dr. Winnie Liang, James Long, Ivana Malenica, Dr. Sara Nasser, Corrie Panganiban, Keri Ramsey, Margot Redman, Rebecca Reiman, Megan Russell, Dr. Sampath Rangasamy, and Ryan Richtolt, Dr. Isabelle Schrauwen, Ashley Siniard.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

LIST OF SYMBOLS

Symbol	Gene Name
APP	amyloid beta (A4) precursor protein
AR	androgen receptor
ATG2A	autophagy related 2 homolog A
ATG2B	autophagy related 2 homolog B
CACNA1A	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
CFTR	cystic fibrosis transmembrane conductance regulator
CLNK	cytokine-dependent hematopoietic cell linker
CLNS1A	chloride channel, nucleotide-sensitive, 1A
CUBN	cubilin
DDC	dopa decarboxylase (aromatic L-amino acid decarboxylase)
DYT1	torsin family 1, member A (Torsin A)
ERCC2	excision repair cross-complementation group 2
GJA12	gap junction protein, gamma 2, 47kDa
HDHD1	haloacid dehalogenase-like hydrolase domain containing 1
HEXB	hexosaminidase B (beta polypeptide)
KCNA1	potassium voltage-gated channel, shaker-related subfamily, member 1
KRT18	keratin 18
LHX6	LIM homeobox 6
MeCP2	methyl CpG binding protein 2
MTFMT	mitochondrial methionyl-tRNA formyltransferase
MYO5C	myosin VC
OCEL1	occludin/ELL domain containing 1
PCSK1N	proprotein convertase subtilisin/kexin type 1 inhibitor
PHGR1	proline/histidine/glycine-rich 1
PLP1	proteolipid protein 1

PLXNA3	plexin A3
PNPLA4	patatin-like phospholipase domain containing 4
POLG	polymerase (DNA Directed), gamma
RNASEH2B	ribonuclease H2, subunit B
SAMHD1	SAM domain and HD domain 1
SLC25A11	*s*olute carrier family 25 member 11
SNAP29	synaptosomal-associated protein, 29kDa
STS	steroid sulfatase (microsomal), isozyme S
UBC	ubiquitin C
UBE3A	ubiquitin protein ligase E3A
UBR1	ubiquitin protein ligase E3 component N-recognin 1
UTP14A	U3 small nucleolar ribonucleoprotein, homolog A (Yeast)
VCX	variable charge, X-Linked
VCX3A	variable charge, X-Linked 3A
WDR45	WDR repeat domian protein 45

PREFACE

This dissertation is the culmination of several years of work as a group member in Dr. David Craig laboratory in the Neurogenomics Division of TGen. The author started his work as microarray based technology and whole genome association studies in common diseases including Autism Spectrum Disorder, Bipolar disorder, Alzheimer's disease, Parkinson's disease became the focus of several investigators in the division. From the start, the author found himself in a nurturing environment with projects that pushed the envelope on current technologies, to find faster, cheaper, more informative ways to study human disease. In 2007, TGen stepped into the next-generation era of high-throughput sequencing and the author and his group was responsible to develop new high-throughput, multiplexed method in next-generation sequencing that allowed the implementation of gene focused sequencing studies on large number of patients. In addition the author was responsible for the wet lab work and manuscript composition in one of the first whole genome sequencing studies at TGen that focused on the identification of causal variant in a rare childhood disorder.

The chapters in this dissertation are original works with some already published by the author. The author conceived Chapter 2 with the guidance from Drs. David Craig, Matt Huentelman and Vinodh Narayanan. The author was responsible for conceptual design, wet lab, data collection and analysis.

Chapter 3 contains a first author publication by the author in its entirety. This chapter was published as Characterization of X Chromosome Inactivation Using Integrated Analysis of Whole-Exome and mRNA Sequencing(Szelinger et al. 2014). The author was responsible for all major areas of this study including, sample preparation, wet lab, data collection, majority of data analysis, and the majority of the manuscript preparation. Drs. David Craig, Matt Huentelman, and Vinodh Narayanan helped conceive the conceptual design, Ivana Malenica and Jason Corneveaux aided with data analysis concepts and computer programming.

Chapter 4 is an original work by the author with contribution from multiple individuals. The author was responsible for study design, wet lab, data collection, data analysis, and manuscript composition. The author received guidance in conceptual design from Dr. David Craig, in data management from Dr. David Craig and Jason Corneveaux, in bioinformatics analysis from Dr. David Craig, Tyler Izatt, Gregory Frantzen, Megan Russell, and Jessica Aldrich. The author received guidance in statistical analysis from Ivana Malenica and Dr. Isabelle Schrauwen. Sample preparation and wet lab support was provided by Ashley Siniard, Rebecca Reiman, Lori Cuyugen, and Jon Adkins. Clinical phenotyping and patient related inquires was supported by Vinodh Narayanan and Keri Ramsey.

CHAPTER 1

CLINICAL DIAGNOSIS OF RARE HUMAN DISEASE IN THE ERA OF NEXT-GENERATION SEQUENCING

Introduction

There are approximately 7,000 rare diseases, and they are defined by a prevalence of less than 200,000 affected for any given rare condition in the United States alone. Although much effort has taken place, only in about half of described, rare diseases the molecular etiology of the condition has been identified (Boycott et al. 2013). A substantial portion of already described disorders are monogenic and associated with rare, pathogenic variants within a single gene (Bamshad et al. 2011). Rare variants can run in families and follow Mendelian inheritance such as autosomal recessive, autosomal dominant, or X-linked inheritance. Therefore, Mendelian inheritance models can provide a basis for the identification of causal variants in disease-associated genes in rare conditions (Bamshad et al. 2011).

Historically over the past two decades disease gene identification for a disorder of unknown genetic etiology relied on focused, candidate gene sequencing or genome mapping strategies. Candidate gene sequencing typically requires a prior knowledge of disease biology, or familiarity with suspected disease locus harboring the candidate gene. Due to a continued reliance on Sanger sequencing, candidate gene sequencing has and continues to be labor and cost intensive, which resulted in a limited number of published studies and limits its future potential for gene identification (Thomasson et al. 1991).

In some cases, mapping strategies have been possible using linkage analysis (including homozygosity mapping) to uncover disease-associated loci by tracking co-segregation of genetic variants with phenotype. Usually, mapping or linkage approaches require multiple affected patients, ideally together with family members. In linkage analysis, the family members, and their affected relatives are evaluated for regions defined by genetic markers informative towards chromosomal position that may or may not segregate with disease status. The main principle is that disease-associated markers are not necessarily causal, but they are inherited together in a region that did not recombine during sexual reproduction. Across generations, these regions become smaller and can be powerful when large pedigrees are available. Given a genetic interval, candidate gene, Sanger sequencing typically follows linkage mapping to help identify causal gene and potential pathogenic mutations. One of the first successful application of genome mapping strategy in rare disease was the identification of the gene associated with Cystic Fibrosis, cystic fibrosis transmembrane conductance regulator gene (*CFTR*), using restriction fragment length polymorphism (RFLP) in combination with Sanger sequencing (Kerem et al. 1989).

More recently homozygosity mapping utilized microarray technology to find loss-of-heterozygosity (LOH) regions for recessive traits or structural variations indicative of deletions. In mapping of LOH, this approach takes advantage of multiple affected individuals within a founder population to identify regions of homozygosity that overlap among the patients. As is the case with linkage studies, LOH analysis is followed up by Sanger sequencing across the minimal region to identify potentially pathogenic variants (Chiang et al. 2006). Similarly, the quantitative nature of these arrays can lead to identification of the genetic basis of disease by mapping deletions within individuals with a common phenotype (Craig et al. 2008).

However, candidate gene, gene mapping, and microarray studies provide information on a limited scale when compared to the complexity of the whole genome or all of the coding regions of the genome. In addition, they are usually resource intensive, low-throughput with substantial costs. In gene mapping the region of interest may be substantially large due to small number of available pedigrees in very rare conditions. Therefore, further reduction to single, disease-associated locus requires multiple mapping steps and additional families, which may be unattainable in very rare diseases and in heterogeneous conditions. Locus heterogeneity, multi-genic causes, can also result in false discovery. Thus, these studies are most informative in highly distinct diseases where a single gene with high penetrance is predicted as causal.

The emergence of high-throughput, next-generation sequencing methods, approximately 2007-2010, combined with the additional ability to capture DNA or partition DNA at defined regions opened up exhaustive single step approaches instead of the two step process of map to candidates and sequence candidates through Sanger sequencing. To date, whole genome sequencing (WGS), and whole exome sequencing (WES) have had a remarkable impact on the clinical diagnosis of rare, Mendelian disorders that have only been impeded by the ability to identify the one or two causal genetic variants from the 3 to 4 million of genetic variants differentiating any two individuals. Sequencing the entire genome of a patient by WGS allows for a global view of all genetic variations including single nucleotide polymorphisms (SNP), short insertions and deletions (indel), translocations, large chromosomal rearrangements, and copy number variations (Gilissen et al. 2014). A number of clinical WGS studies have successfully applied the comprehensive sequencing approach in the identification of causal alleles in single patient studies where disease gene identification was confounded by heterogeneous phenotype (Lupski et al. 2010; Bainbridge et al. 2011; Welch et al. 2011). Conversely, sequencing the high impact, protein-coding regions of the genome by WES helped expedite genetic diagnosis in a number of single-gene Mendelian diseases with simple inheritance patterns and well defined phenotype (S. B. Ng et al. 2009; Bamshad et al. 2011; S. B. Ng et al. 2010). WES focuses sequencing resources to approximately 1% of the genome by targeting genetic variations in coding or exonic regions for about 20,000 protein coding genes.

As a result, in a short period of time, discovery of new clinically actionable variations causing disease, or genetic diagnosis by WGS and WES shows a tremendous potential to impact the ability to diagnose, treat, and manage care for rare childhood disorders. The reduced costs, fast turnaround time, and availability of a range of exome targeting assays make WES an intriguing tool for identifying the genetic basis of disorders with unknown genetic etiology. However, WES still yields tens of thousands of variants, of which many variants remain plausible candidate causal variants, as their impact on gene function is not ascertained by WES or WGS. Thus, improvements in interpretation of genetic variants to their functional impact can further reduce the number of candidate variants and identify those that have greatest impact on gene function.

Currently, in silico predictions of a variant's functional impact are based on how the variant is predicted to alter transcription and translation of the DNA sequence. The first step in the generation of protein product from the genetic code defining a gene is the creation of intermediary

RNA copy of the gene. This messenger RNA (mRNA) goes through transcriptional modification that removes intronic sequences and leaves only exonic and un-translated regions (UTR). Genetic variants can influence how introns are spliced out, which exons are included in what combination. This can result in multiple mRNA species, or transcripts that can be translated into protein products with divergent properties. Functional predictions utilize this information to classify genetic variants into functional classes based on how they impact the mRNA transcript. The most common functional classes are loss-of-function, missense, or synonymous (McCarthy et al. 2014). Loss-of-function variants can subject a transcript to nonsense-mediated decay or loss-of-function of the translated protein. SNP and indel variants can cause a frameshift in the open reading frame of the mRNA sequence during translation and can result in altered amino acid sequence, thus categorized as frameshift variants. In addition frameshift variants can be further classified as stop-gain, stop-loss and splice donor and splice acceptor variants. Missense variants result in the change in the mRNA codon sequence, and consequently the amino acid they code for. Synonymous SNPs or indels do not change the amino acid sequence and assumed silent to function. Previously, population scale sequencing predicted that 95% of rare protein coding variants with a population frequency of <1% may have a functional impact and an individual may carry up to 500 rare functional variants (Tennessen et al. 2012). However, not all functional variants are disease causing, thus additional information is necessary to associate a variant's predicated impact to phenotype and further reduce plausible candidates.

To further narrow to the most plausible candidate variants a variety of additional filters are applied during standard WES studies that include predictions about the variants' pathogenicity, their observed frequency in the general population, presumed inheritance of the condition, and available clinical information about the patient (Gilissen et al. 2012) (Figure 1). However, even after best practice, a few hundred candidate variants may still remain as potentially causal for any disease, thus variant prioritization approaches can greatly impact genetic diagnosis (Richards et al. 2015).

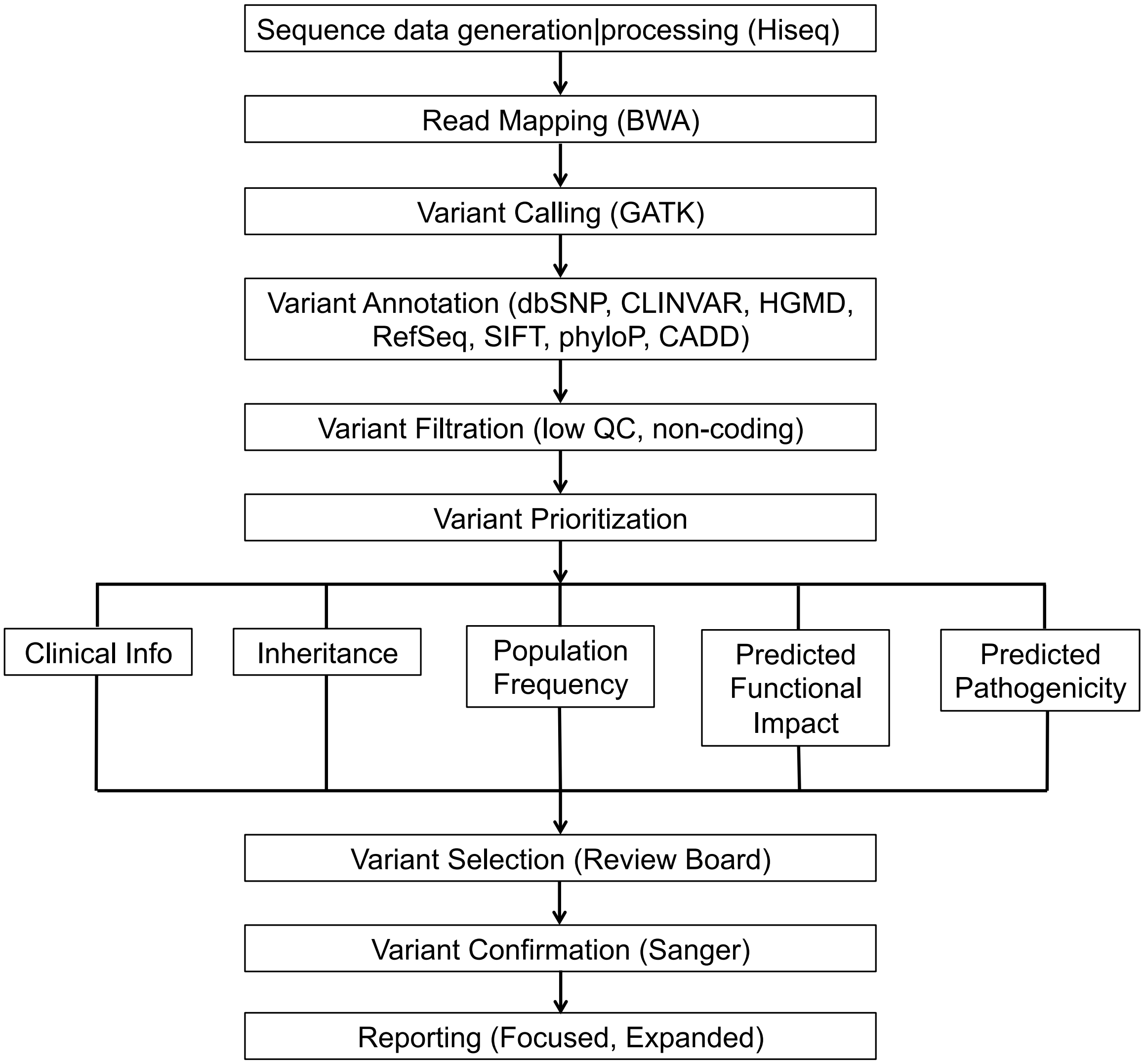
Sequence data generation|processing (Hiseq)
Read Mapping (BWA)
Variant Calling (GATK)
Variant Annotation (dbSNP, CLINVAR, HGMD, RefSeq, SIFT, phyloP, CADD)
Variant Filtration (low QC, non-coding)
Variant Prioritization
Clinical Info
Inheritance
Population Frequency
Predicted Functional Impact
Predicted Pathogenicity
Variant Selection (Review Board)
Variant Confirmation (Sanger)
Reporting (Focused, Expanded)

Figure 1. Variant Prioritization in clinical sequencing studies. This figure shows standard workflow in a clinical sequencing study. It starts by data generation and alignment to a reference genome the millions of sequenced reads. This is followed by variant identification and variant annotation. Annotation is performed to understand the variant's properties in terms of evolutionary conservation, population frequency, clinical and disease relevance, gene function, etc. On this figure under "Variant Annotation" the various acronyms refer to databases that contain information about variant properties. Post annotation, variants are filtered based on quality and based on genomic position to select the most informative variants, like coding variants. This step is followed by variant prioritization, which is a multi-tiered data reduction process utilizing accrued information about the likelihood of the variant to be pathogenic and its potential association with observed phenotype. Candidates are further evaluated by a board of clinical staff until consensus is found based on the combined evidence. The selected variant is validated and reported. Reports can be focused, or expanded, that provides information only on the most likely pathogenic variant, or provides additional variants that are incidental or with unknown significance.

As described in Figure 1, variant prioritization is a multi-tiered approach and in the following we will describe the most common information used to reduce plausible candidates and the approaches that utilize them.

One of the cornerstones of variant prioritization is the assumption made about the inheritance of the patient's disease. This information is then utilized as a filter mechanism to eliminate those genomic variants from candidate list that do not adhere to assumed inheritance model. Inheritance based reduction of variants can be used in single patient, in most Mendelian disorders, the inclusion of genotype data from multiple affected individuals or from family members can greatly expedite candidate variant reduction. For diseases that are expected to follow autosomal dominant inheritance, a number of affected patients with overlapping phenotype are usually needed; ideally with parents that are unaffected in order to identify heterozygous, and often de novo, variants (S. B. Ng et al. 2009). In disorders that are expected to follow autosomal recessive inheritance, biological parents needed to find homozygous or compound heterozygous variants (Becker et al. 2011). In cases, where causality may be due to *de novo* mutation, sequencing the family trio may be sufficient in some cases (Vissers et al. 2010). In families where consanguinity is suspected, homozygosity mapping in a single patient may be sufficient to identify a homozygous causal variant although family segregation in family trio WES data can provide additional support (Bilguvar et al. 2010).

The most widely used strategy to prioritize variants in rare disease sequencing studies is based on family inheritance by sequencing the patients and their biological parents or family trio (Farwell et al. 2014). Parents being the first order relatives of patients, nearly all of the identified genetic variants will be present in the parents, and application of Mendelian inheritance patterns, *de novo* filtering, and clinical phenotyping can reduce a substantial portion of genetic variants suspected to be associated with the patient's condition. The power of family based sequencing can be seen in 30% success rate in diseases where well characterized, phenotypically homogeneous group of patients are difficult to obtain (Farwell et al. 2014). Interestingly, recent clinical studies of hundreds of singleton patients achieved an approximately equal 25% diagnostic yield irrespective of variant prioritization or patient selection strategy. This suggests limited utility of WES alone in clinical diagnosis highlighting the need for novel approaches to improve diagnostic rate (Y. Yang et al. 2013; Y. Yang et al. 2014). As new publications emerge, revisiting cases can yield a diagnosis, suggesting that in many cases the causal variant does lie within the primary few hundred candidate variants and can be better captured by alternative filtering strategies.

Variant annotations can provide further evidence for or against a variant's causality. After sequencing and data processing, high quality variants are annotated with population frequency, in silico prediction for pathogenicity, variant type, predicted impact of variant to protein structure and function, and biochemical properties (Gargis et al. 2015). Population frequency information is usually obtained from large databases of thousands of sequenced from the 1000 Genomes Project (Consortium et al. 2012), or Exome Aggregate Consortium (Exome Aggregation Consortium). Annotations for pathogenicity, biochemical properties can be obtained from aggregate tools like dbNSFP (X. Liu, Jian, and Boerwinkle 2013). dbNSFP is a variant-level collection of predictions from a wide range of in silico tools (e.g. SIFT, MutationTaster, CADD, PolyPhen, etc.) in a single tabulated format across millions of variant loci. There are multiple variant effect predictors that annotate variants for variant type and their impact on protein function (eg. missense, non-sense, silent) including SnpEff, ANNOVAR, Variant Effect Predictor, VAAST (Cingolani et al. 2012; K. Wang, Li, and Hakonarson 2010; McLaren et al. 2010; Hu et al. 2013). Annotations are often enhanced by medical information obtained from genome-wide clinical databases that contain genotype-phenotype descriptions like ClinVar (Landrum et al. 2014), or information about known disease causing genes, or variants and their associated conditions from OMIM (OMIM), CGD (Solomon, Nguyen, and Bear 2013), and HGMD (Stenson et al. 2003).

There are two main approaches to prioritization that utilize clinical information, variant predictions based on annotation and inheritance models; a probabilistic model, and a heuristic model. Heuristic model is based on some assumptions (e.g. variant is rare and deleterious to protein function) that guide variant filtration process. In general, it starts by filtering out known variants, commonly from dbSNP, making up about 90-95% of candidates. For rare diseases, variants are further filtered by applying a population allele frequency cutoff of 1% (M. X. Li et al.

2012). This is usually followed by further reduction by filtering variants that may be present in in-house sequenced cohort. The remaining 100-500 private variants are then evaluated for pathogenicity using in silico prediction algorithms for deleteriousness to protein function (obtained from the variant annotations). Application of inheritance models and genotype segregation are also part of heuristic models, however, they can be applied at various steps of prioritization depending on researcher and assumptions made about the patient's clinical information. Probabilistic models assess sequence variants for their likelihood to be associated with disease when variants from affected patients are compared to variants obtained from control genomes (Hu et al. 2013). Filtering variants for inheritance, population frequency, or predictions for functional impact prior to comparative analysis can extend this model. In some cases heuristic and probabilistic models are combined to obtain higher confidence candidate lists (Coonrod et al. 2013).

Still, there are several factors that may impede diagnosis and yielding the approximately 30% rate for identification of the genetic basis of disease. Technological difficulties related to sequencing can impact diagnostic yield, for example. With the advent of Sanger sequencing, it became known that genomic complexity of an organism greatly influences the sequence coverage that can be achieved, and extreme GC or AT rich genomes are hard to sequence to even coverage (Ajay et al. 2011; Lam et al. 2011). This GC bias mainly manifests itself with coverage gaps. GC content bias in Mendelian disease discovery is especially problematic, because most causal variations lie in the protein coding regions of the genome which are known to have higher GC content than surrounding inter-genic regions, especially in first exons (Lander et al. 2001; Majewski 2002). Both WES and WGS are prone to coverage bias which can impact variant discovery and produce false negatives in clinical diagnosis (Ross et al. 2013). WES is also poorly powered to study copy-number variations, triplicate expansions, and large insertions deletions that may overlap exon-intron boundaries or those that are situated in non-coding, regulatory regions. There are a number of target enrichment assays available for WES each of which has its own advantage or disadvantage in terms of completeness of targeted genomic regions, capture efficiency, design strategy for exon-intron boundaries, and accuracy for SNP and indel detection (Meienberg et al. 2015; Chilamakuri et al. 2014). The non-uniform performance of these exome assays is a critical consideration prior in a WES study design.

In addition to technological challenges variant prioritization approach and available annotations can greatly impact diagnostic yield. Filtering out known variants may remove true causal mutations, as dbSNP contains rare, disease causing variants. The choice between autosomal recessive or dominant inheritance models can be influenced by the accuracy of clinical phenotyping. In addition, the annotation tool used can impact our interpretation of a variant's predicted pathogenicity. These factors can result in the exclusion of potentially deleterious variants. Alas, there are many types of annotations are available, and variant prediction approaches differ between laboratories and in many times between exome sequencing projects within the same laboratory. This may confound disease gene identification for patients with similar clinical symptoms studied at different times. In addition, continually updated annotations warrant the continued re-analysis of previously studied exomes, and may alter previously obtained results. Effort is taken by multiple groups to standardize analytical processes associated with clinical exome sequencing to provide a best-practices framework that aims to maximize the utility of current scientific knowledge to improve diagnostic yield (Gargis et al. 2015).

In addition to best-practice models there is a community effort by American College of Medical Genetics (ACMG) underway to standardize interpretation techniques for WES (Richards et al. 2015). One of the main confounding factors of variant interpretations for diagnosis is our ability to define a variant as pathogenic, which is mostly based on *in silico* predictions and available annotations as described above. As noted, there are a multitude of tools available to predict a variant's likelihood to impact gene function and thus impacting its rank on the ranked candidate variant lists. However, a systematic evaluation of prediction algorithms has shown that the concordance between prediction tools is approximately 60-65% and is dependent on the data source that the algorithm is trained on (eq. Ensembl, RefSeq) (McCarthy et al. 2014). This can result in discordant classification of a candidate variant in terms of pathogenicity and predicted functional impact. Since variant prioritization assumes that most likely pathogenic candidates will have a large functional impact on gene function, incorrect predictions can result in the enrichment of false positive variants among the top ranked candidates. Thus ACMG recommendations aim to standardize nomenclature to describe variants identified as potentially causal as “pathogenic,” “likely pathogenic,” “uncertain significance,” “likely benign,” and “benign”. It also suggests the combination of multiple annotation tools prior interpretation to improve confidence. Finally, the ACMG group acknowledges that functional studies using RNA sequencing (RNA-seq) or protein-based assays can greatly support variant predictions to gene function obtained through computational methods. Ideally, these assays would be most informative and provide stronger evidence if performed in patient derived tissue (Richards et al. 2015).

Our understanding of the genetic basis of disease-associated traits has greatly improved with the advent of high-throughput “omics” methods (Figure 2). We are now able to routinely generate comprehensive data from multiple biological systems including, genome, epigenome, transcriptome, proteome, and metabolome, with methods collectively referred to as “omics”. In addition to studying genetic variations by WGS or WES, researchers have been working on analytical methods to study the comprehensive expression profile of genes by Transcriptome sequencing, the regulation of gene expression by Chip-seq or Methylation sequencing, the diversity and structure of proteins and metabolites by Mass-spectrometry. These “omics” methods have been successfully applied to various tissues including whole blood, and recently garnered attention in single cell studies (Shapiro, Biezuner, and Linnarsson 2013). Historically, measurements from each type of omics analysis have been analyzed individually to look for associations between the biological system studied and the phenotype observed and to find predictor markers in the biological system studied to the observed phenotype. The study of individual biological systems allowed us to uncover pieces of the puzzle, but as described above in the diagnostic yield of WES studies, studying single biological systems can explain only in part the genetic etiology of human disease. Therefore additional information is necessary from other biological systems to gather evidence for causality of genetic variants in disease diagnosis.

Recently systems biology approaches have been developed that integrate multiple “omics” data types (Ritchie et al. 2015). Combination of multiple data can reduce unreliable data, improve confidence, and gather additional evidence that may reduce false positive findings from single source. In addition the complete biological background of disease-associated traits can only be deciphered by connecting the cause-and-effect relationships when considering all biological systems simultaneously.

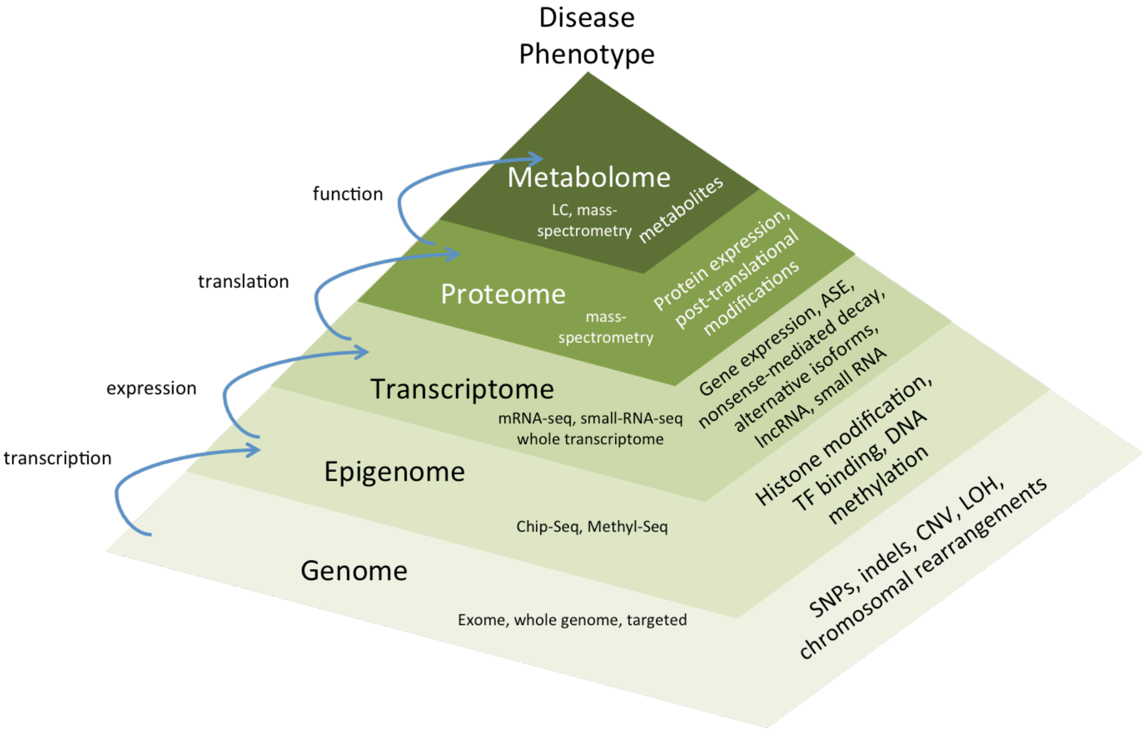
Disease Phenotype
function
translation
expression
transcription
Metabolome
LC, mass-spectrometry
metabolites
Proteome
mass-spectrometry
Protein expression, post-translational modifications
Transcriptome
mRNA-seq, small-RNA-seq whole transcriptome
Gene expression, ASE, nonsense-mediated decay, alternative isoforms, lncRNA, small RNA
Epigenome
Chip-Seq, Methyl-Seq
Histone modification, TF binding, DNA methylation
Genome
Exome, whole genome, targeted
SNPs, indels, CNV, LOH, chromosomal rearrangements

Figure 2. Multi-omics for the study of biological systems. This plot shows biological systems, genome, epigenome, transcriptome, proteome and metabolome as they related to each other and the various “omics” tools that enable us to study them (i.e. Exome-sequencing, Chip-seq, mRNA-seq, mass-spectrometry, liquid chromatography or LC). For each biological system on the right face of the pyramid, the various features are listed that the “omics” tools are able to interrogate (i.e. SNPs, histone modifications, small RNA, post-translational modification, metabolites). Arrows on the left indicate the flow of genetic information from the genome level to the final manifestation in the phenotype on top.

There are two main integration approaches: multi-staged analysis, which involves integrating information using a stepwise, hierarchical analysis approach, and meta-dimensional analysis, which refers to the concept of integrating multiple different data types into a multivariate model associated with given outcome (Ritchie et al. 2015).

Multi-stage approach divides analysis into steps with an initial association test between data types followed by association analysis of the combined data with phenotype. Many times, multi-staged integration studies use two types of data; genomic and transcriptomic (Huang et al.

2007; Lappalainen et al. 2014). Integration starts by reducing genomic data to those variants that are associated with observed phenotype based on association analysis or in silico predictions about the deleteriousness of the variants, and additional criteria about population frequency and inheritance. This reduced genomic variant list is then associated with other data types, to find those variants that are associated with transcription or epigenetic modifications, for example. This step can identify variants that are expression quantitative trait loci (eQTL), predictors for nonsense-mediated decay, alternative splicing, associated with DNA methylation metabolite, protein, miRNA levels, depending on the data type used. This step is followed by an additional step that performs either an association test of the combined data with the phenotype of interest or further filters genetic variants based on the variant's functional effect. This approach has been applied in the identification of eQTLs that are associated with drug response (Huang et al. 2007). In Autism Spectrum Disorder (ASD), integration of genomic and transcriptomic data from whole blood sequencing was able to identify potentially causal variants in monogenic forms of ASD that were missed by WES alone (Codina-Solà et al. 2015). In population scale integrative studies, utilizing heterozygous alleles from genomic data and their impact on transcript expression (i.e. allele-specific expression) have identified common genomic variations that are eQTLs (Lappalainen et al. 2014).

Meta-dimensional analysis takes an approach where multiple data-types are combined and analyzed simultaneously. This approach essentially combines data matrices from as many data types as possible into a single matrix. The combined data matrix is then used for multiple analytical approaches for association testing with outcome that include Bayesian modeling by Fridley et al. (2012), to Cox regression by Mankoo et al. (2011). This approach has been used to develop comprehensive analytical tool, ATHENA, that incorporates, copy-number variation, methylation, miRNA, and gene expression to study association with complex traits in cancer (Holzinger et al. 2014). The advantage of this approach is that it allows the study of interaction between the data types simultaneously that may be missed in a step-wise process. However, the combination of data types that was analyzed by different methods and may be at different scale can be challenging for integrations into a single matrix and may require data transformation that can reduce correlation between the data types.

Currently, most integration approaches have been trained on common disorders, like cancers and cardiovascular disease, with a relatively large number of affected and non-affected control populations. This allows for building models that can perform association analysis for expected outcome. However, in rare disease, many time only a few affected patients available, and phenotype is not well characterized, thus integrative association test would lack power.

As an alternative, in this dissertation we set out to develop an integrative approach whereby simultaneous analysis of next-generation sequencing data obtained from whole blood DNA and RNA of patients with rare childhood disorders will provide information on the functional impact of rare, coding variants, and enable us to rank them based on their functional impact. In many cases, it is easy to acquire RNA from tissues of childhood patients enrolling into a clinical sequencing study, particularly whole blood or even skin fibroblast, recognizing that many metabolic disorders are transcriptionally active in multiple tissues. Rare variants associated with dysregulation of gene expression at the RNA level is consistent, though does not prove, that the variant has a functional role, and could be prioritized differently with respect to other private variants with completely unknown functional impact. Thus integration of data from multiple biological systems in the same patient can improve standard variant prioritization procedures and unmask variants whose functional impact is not well supported by DNA sequencing alone. We hope that integration will improve our understanding of the functional impact rare variants have on cellular phenotype and disease trait that may be clinically actionable, will unveil molecular targets for pharmacogenomics, and will improve patient specific care.

CHAPTER 2

SHARED DE NOVO MUTATION IN THE WD REPEAT DOMAIN 45 PROTEIN IN A SIBLING PAIR WITH BETA-PROPELLER PROTEIN-ASSOCIATED NEURODEGENERATION

Introduction

In this chapter we set out to take advantage of the wide spectrum of genomic and functional information we can obtain from simultaneous sequencing of DNA and RNA in a sibling pair diagnosed with a rare, X-linked, neurological disorder that was previously only reported in sporadic, singleton cases. In addition this study hopes to identify the underlying molecular mechanisms leading to a more severe, lethal phenotype in the male sibling and a less severe manifestation in the female sibling. Our approach to study an X-linked disorder in the context of integrative DNA and RNA sequencing has only recently was reported and is described in Chapter 3 (Szelinger et al. 2014).

Beta-propeller protein-associated neurodegeneration (BPAN) is a newly recognized member of the neurodegeneration with brain iron accumulation (NBIA) group of disorders (MIM:300894). The disorder is also known as static encephalopathy and neurodegeneration in adulthood (SENDA). BPAN is characterized by developmental delay and intellectual disability in early childhood, followed by neurodegeneration, and characteristic MRI findings of hypointensity on T2-images in the globus pallidus and substantia nigra (Haack et al. 2012). Clinical symptoms deteriorate after adolescence with progressive loss of psychomotor skill, rigidity, and reduction or complete lack of language skills (Haack et al. 2013). Key hallmark of adult BPAN is dystonia, Parkinsonism, and dementia. A subset of patients show Rett-like symptoms, ocular defects, and gastrointestinal dysfunction (Hayflick et al. 2013). Most of the diagnosed are sporadic, singleton cases with a wide ethnic spectrum. Mutations in *WDR45,* a member of the WD40 repeat domain genes, are known to cause BPAN (Pagon et al. 1993). Patients reported to date are predominantly females, carrying *de novo,* heterozygous single nucleotide variants or small indels consistent with an X-linked *de novo* dominant model (Haack et al. 2012; Hayflick et al. 2013; Ozawa et al. 2014; Okamoto et al. 2014). There is a substantial overlap in clinical manifestation of BPAN between males and females, which suggest the role of post-zygotic mutations, and chromosomal aberrations (Haack et al. 2012). Phenotypic variability has also been suggested by existence of skewed X chromosome inactivation in the germline DNA of some cases (Saitsu et al. 2013; Haack et al. 2012). X chromosome inactivation (XCI) is a process by which one of the two X chromosomes inherited from parents are silenced by epigenetic mechanisms during early female embryonic development (Augui, Nora, and Heard 2011). Consequently, in each progeny of the cell the same X chromosome will be active leading to a mosaic pattern of X-linked gene expression; in a portion of cells the maternally inherited X is active, and in others the paternal X is active. Lyon proposed that inactivation occurs randomly in females and therefore the overall expression of X-linked genes results in an approximate equal proportion (Lyon 1961). Any deviation from this ratio results in skewedness in favor of one of the parental X and in extreme cases result in complete silencing of one of the X chromosomes. Skewed XCI can mediate phenotypic variability in X-linked diseases (J. I. Young and Zoghbi 2004).

DNA methylation status is standard assay (HUMARA) to assess the X inactivation ratio of the two chromosomes and relies on the methylation status of a polymorphic triplicate expansion in the human androgen receptor (*AR)* gene. Most females are polymorphic of this repeat and the two chromosomes can be distinguished (Amos-Landgraf et al. 2006). The methylation status of a restriction site in the proximity of the repeat is associated with X inactivation status, and enzymatic digest of the un-methylated restriction site (active X, Xa) followed by PCR amplification will lead to amplification of the methylated (inactive X, Xi) *AR* locus. On the other hand, the un-methylated, active allele will be cut by the restriction enzyme and no amplicon will be generated (Allen et al. 1992). Capillary analysis of the amplicon peaks correlates with PCR yield of the active and inactive alleles and their normalized ratios give the XCI ratio.

The contribution of X inactivation to the phenotypic heterogeneity among patients with X-linked disorders can be best studied by comprehensive genomics methods that reduce the need for multitude of molecular tests, and helps in the identification causal genes, and provides a unique view of biological processes contributing to phenotype. We performed family-based WES and RNA-seq analysis of the sibling pair and their unaffected parents. The male sibling presented with a more severe phenotype and expired by the time of this study. The female sibling presented a less severe phenotype consistent with some Rett-like symptoms. Exome sequencing identified a *de novo*, missense variant in *WDR45* that was shared by both siblings. RNA-seq analysis showed nominally significant differential regulation of *WDR45* between the siblings and parents. We used allelic expression to directly quantify X-linked, heterozygous SNP allele and combined with genotype segregation we found an extreme, non-random XCI in the female patient in favor of the maternal X. The *de novo* mutation showed allele specific expression concordant with the biased expression of the maternal X chromosome suggesting that the *WDR45* mutant allele originated on the maternal X and implicated XCI in phenotypic heterogeneity between the siblings. Comparison of RNA-seq data to methylation assay showed high correlation. This is the first study of a sibling pair sharing a *de novo* dominant, X-linked, *WDR45* mutation in BPAN.

Materials and Methods

Exome Sequencing.

Sequencing and data analysis methods are described in Chapter 4. Materials and Methods.

RNA-seq

Sequencing and data processing methods are described in Chapter 4. of Materials and Methods. Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKMs) were calculated using Cufflinks2.2.1 and plots were generated using GGplot2 (R v3.1.3) (Trapnell et al. 2013). We performed differential expression analysis between parents and patients using Cuffdiff2 in the Cufflinks package. We only retained *WDR45* and those genes for analysis that were shown to interact with *WDR45* from public interaction database BioGRID 3.3 (BioGRID).

XCI with HUMARA analysis

Females were enrolled from 29 families from the Dorrance Center for Rare Childhood Disorders including affected female patients, and if available their mothers, female siblings regardless of affected status, and female grandparents. The enrolled families are listed in Table 14, in Chapter 4. We obtained DNA and total RNA for a total of 48 participants, 5 of which had already been evaluated for XCI and described in Chapter 3. These 48 participants included 10 female patients diagnosed with Aicardi Syndrome, previously described by Schrauwen et al. (2015). Genomic

DNA from each participant was sent for HUMARA test to Greenwood Genetic Center (Greenwood, SC) and RNA-seq was performed as described in Chapter 4 Materials and Methods.

After WES and RNA-seq, the estimation of XCI ratio was performed as described in Szelinger et al. (2014) and also in Chapter 3. Briefly, in family trios, and large families, the SNP variants were phased for affected patient or sibling if paternal and maternal genotypes were available. Maternal and grandmother genotypes were not phased. Next, heterozygous genotypes were selected from the X chromosomes and pileup was created from RNA-seq data to count the number of reads mapping to each heterozygous allele. The allele counts were used to obtain allele ratio for each heterozygous locus. The allele ratio was defined as the read count of the SNP allele over the sum of counts for SNP and reference alleles. The distribution of allelic ratios across X was fitted to the beta distribution and the mean of the allele ratio distributions were used as XCI ratio. These un-scaled ratios were then scaled to 100 scale. Scaling was performed by taking the ratio of 100 and the cumulative value of the ratios of the 2 chromosomes for each patient. This difference factor was then multiplied by the un-scaled ratios of each X chromosome's allelic ratios to obtain the scaled XCI ratio. Complete skewing of XCI was defined as a ratio of <2:98/>98:2, extreme skewing was <10:90/>90:10, moderately skewed as <20:80/>80:20, and random XCI as >20:80/<80:20.

To estimate the distribution of the XCI ratios in our cohort the XCI ratios were binned based on the more dominant alleles for the RNA-seq experiment and from the inactive allele from the HUMARA assay. The bins were defined as 50-60, 60-70, 70-80, 80-90, and 90-100. XCI ratio of 59:34 is binned into the 50-60 category and a ratio of 81:18 is binned into the 80-90 category, respectively.

Calculation of statistical significance and Spearman's rank correlation was performed by the cor.test function in the R statistical package and visualized by GGplot2 (R v3.0.3).

Results

Clinical Description

We present a family with two siblings affected with BPAN enrolled into the Dorrance Center for Rare Childhood Disorders under a human research protocol approved by WIRB (Table 1). The family id for this family is 0103 (Chapter 4, Table 13).

The older sibling (Patient 0103_1) is a male who expired at age 18. He was born 4 weeks early, 4 lb. 15 oz., without complications. He rolled over front to back, and was able to sit with support; but was noted to be delayed by 6 months of age. MRI at 10 months suggested a white matter process. He developed infantile spasms and myoclonic seizures at 1 year, and was treated with vigabatrin and adrenocorticotrophic hormone (ACTH). He regressed. He has had intractable epilepsy since then, with continued daily seizures including myoclonic seizures, staring spells, and tonic seizures with apnea. At age 2, he has hypotonic, but with hyperreflexia, clonus and up going toes. By age 5, physical findings were of spastic quadriplegia, hypertonia, with cortical thumb posture. With time, he had progressive spastic quadriplegia with contractures, progressive scoliosis, cortical visual impairment and bilateral sensorineural hearing loss. Longitudal Magnetic Resonance Imaging showed cerebral and cerebellar atrophy, white matter volume loss, T2 and GRE hypointensity in the globus pallidus (GP) and substantia nigra (SN).

This female child (Patient 0103_2) is now 14 years old. Birth history was unremarkable, with a birth weight of 7 lb. 9 oz. She rolled over at 5 months, and was able to sit if propped up. Delayed development was noted early, and was seen by neurologist at 17 months. She was hypotonic; unable to sit with support, had poor visual fixation and tracking, and poor hand use. Seizures, consisting of tonic stiffening, appeared at around age 2 years, and have persisted. Seizures have been better controlled than her brothers. She made slow progress in her development but remains in a wheelchair most of the time. She is able to walk with assistance, hold a cup, throw toys, and tries to communicate. Features noted at multiple examinations include poor eye contact, bruxism, hand clasping, truncal hypotonia, peripheral hypertonia with hyperreflexia, features that are suggestive of Rett syndrome. MRI scans have suggested white matter volume loss, hypointensity in the GP and SN similarly to brother's MRI.

Table 1.

Clinical features of the patients with BPAN.

	0103_1 (P1)	0103_2 (P2)
General Characteristics		
Age (y)	18	14
Gender	Male	Female
Neuropsychiatric Symptoms		
Intellectual Disability	Severe	Moderate
Developmental	+ (6 mo)	+ (<1y)
Behavioral Problems	+	+
Cognitive	Progressive	Non-progressive
Psychopathology		Rett-like symptoms
Neurological Symptoms		
Current Status	Expired	Wheelchair/short assisted
Communication	-	Few words
Seizures present	+ (1y)	+ (2y)
Visual Impairment	+	+
Dystonia	+	+
Seizures	Epileptic, myoclonic, tonic	tonic
Muscle function	Spastic quadriplegia, contractures	Hand clasping, hypotonia
Radiology features		
MRI	hypointensity in globus pallidus, substantia nigra	hypointensity in globus pallidus, substantia nigra
Cerebral atrophy	+	n.a.
Genetic tests	Karyotype 46 XY	Karyotype 46 XX
	Fluorescence in situ hybridization	Chromosomal microarray
	FragileX, mtDNA point mutation screen, *MeCP2* sequencing, Leber hereditary optic neuropathy gene test	*MeCP2* sequencing, neuronal ceroid-lipofuscinoses, CLN3, CLN6 gene test
Molecular testing	Very long chain fatty acids	Plasma amino acids
	Lysosomal enzymatology	Urine organic acids
	Electron microscopy for leukocytes	Plasma lactate, pyruvate
	Neuronal Ceroid-Lipofuscinoses enzymatology	Acylcarnitine profile

Exome analysis.

An average of 13.9 Gb of mappable bases were sequenced for average target coverage of 103X, and greater than 85% of the target regions were covered by at least 30X. Please refer to Appendix B for QC metrics of exome sequencing.

Variant analysis identified 642,202 SNVs and short indels in the family. To identify the causal variant we annotated 16,126 missense, nonsense and short indel variants. BPAN is primarily a sporadic disease with singleton cases, and this family has two affected, thus we focused our attention on autosomal recessive and de novo variants. Due to the severe, well-characterized phenotype, we also postulated that X-linked variants may contribute to disease therefore we also evaluated variants on the X chromosome. In both affected children 140 autozygous, 27 compound heterozygous, and 17 X-linked candidate variants were identified. In addition *de novo* variants in the male child (n=57), female child (n=42), and in both children (n=25) were uncovered. Evaluation of the candidate variants for pathogenicity, led to the identification of a *de novo* missense variant in *WDR45* shared by the siblings (Table 2). The male sibling was hemizygous and the female sibling was heterozygous for the mutant allele. This variant was found in exon 10 (NM_007075.3, c.758T>C, p.Leu253Pro), and both parents carried a homozygous wild type genotype (Figure 3B). The average exome sequencing depth of the locus across the family members was 105 ± 33X above base quality cutoff of 10 (phred scaled). This mutation was not observed in the Exome Aggregate Consortium’s over sixty thousand unrelated exomes (Exome Aggregate Consortium) and had a moderate conservation score of 1.5 by phyloP, and was conserved across multiple vertebrates (Figure 3C).

Previously, Verhoeven at al reported a wheelchair bound, adult female with severe intellectual disability diagnosed with BPAN carrying an in-frame deletion at c.752-74del six bases upstream from the variant identified in the male and female siblings (Verhoeven et al. 2014). Overlap between the phenotypic manifestations between the siblings and reported case supports this variant as likely causal. *WDR45* is a repeat domain protein and amino acid changes in the repeat domains structure through missense variants can interfere with protein folding and function.

Table 2.

Candidate variants by whole-exome sequencing.

					Variant	MAF NHLBI	MAF		
chr:pos	Gene	Model	cDNA	aa	type	(EA/AA/All)	1,000	phyloP	SIFT
12:53343231	*KRT18*	*de*	c.274G>C	p.Ala92Pro	missense	NA	NA	2.43	0
X:48933095	*WDR45*	*de*	c.758T>C	p.Leu263Pro	missense	NA	NA	1.51	0
15:40648398	*PHGR1*	AR	c.215G>A	p.Gly48Asp	missense	NA	NA	2.29	0
4:10502936	*CLNK*	AR	c.1084C>	p.Arg362Cys	missense	NA	0.2	2.52	0
4:10586571			c.92C>T	p.Pro31Leu	missense	6.2/1.2/4.5	3.0	1.36	0
chr10:1697959	*CUBN*	AR	c.5924C>	p.Pro1975Le	missense	1.1/0.1/0.7	1.4	0.03	0.01
chr10:1693249			c.8635C>	p.Leu2879Ile	missense	3.7/0.5/2.7	0.9	2.59	0.27
chr15:5252795	*MYO5C*	AR	c.2878A>	p.Lys960Glu	missense	2.1/0.5/1.6	1.6	0.96	1
chr15:5254361			c.1634C>	p.Ser545Tyr	missense	1.5/0.3/1.1	0.6	2.87	0.08
4:43256191	*UBR1*	AR	c.4642A>	p.Thr1548Al	missense	7.4/1.7/5.4	3.3	0.96	1
4:43317071			c.2695A>	p.Ile899Val	missense	3.0/0.5/2.2	1.0	2.87	0.08
chrX:12905546	*UTP14*	X	c.1249G>	p.Glu417Lys	missense	0.05/0.0/0.0	NA	2.49	0.54

Model = Inheritance, AR = autosomal recessive, MAF= minor allele frequency, NHLBI= NHLBI Exome Sequencing project

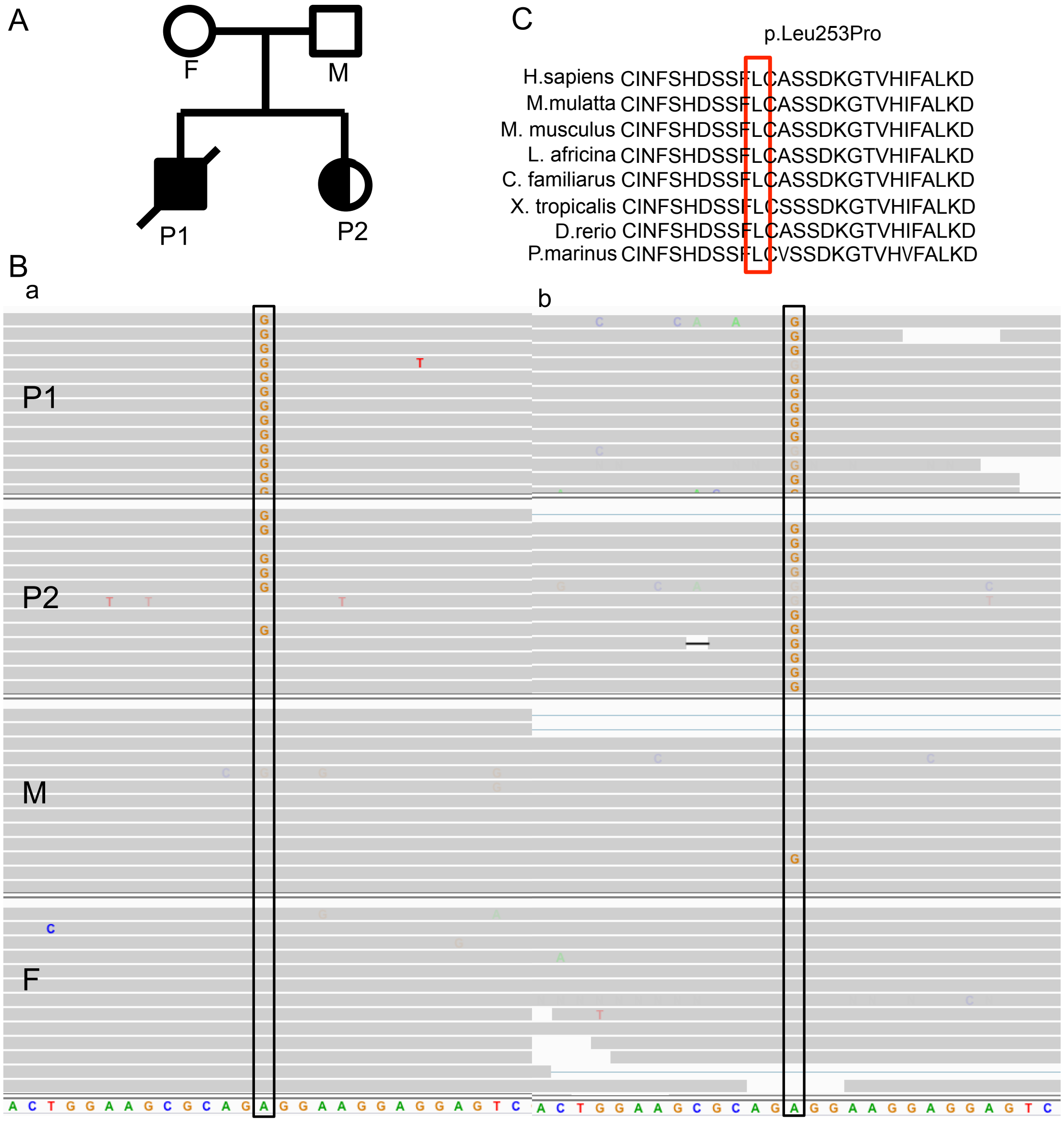
A
F
M
P1
P2
C
p.Leu253Pro
H.sapiens CINFSHDSSFLCASSDKGTVHIFALKD
M.mulatta CINFSHDSSFLCASSDKGTVHIFALKD
M. musculus CINFSHDSSFLCASSDKGTVHIFALKD
L. africina CINFSHDSSFLCASSDKGTVHIFALKD
C. familiarus CINFSHDSSFLCASSDKGTVHIFALKD
X. tropicalis CINFSHDSSFLCSSSDKGTVHIFALKD
D.rerio CINFSHDSSFLCASSDKGTVHIFALKD
P.marinus CINFSHDSSFLCVSSDKGTVHVFALKD
B
a
b
P1
P2
M
F
A C T G G A A G C G C A G A G G A A G G A G G A G T C
A C T G G A A G C G C A G A G G A A G G A G G A G T C

Figure 3. WDR45 variant allele. **A.** Family pedigree, F=father, M=mother, P1=affected male, P2=affected female. **B.** Sequencing traces obtained from Exome Sequencing (a) and mRNA sequencing (b) of c.758C>T nucleotide variant in *WDR45* for each individual. Variant allele is boxed across the traces. **C.** Amino acid conservation of *WDR45* variant allele across vertebrates.

RNA-seq

We sequenced an average of 73.4 million bases across the family members. This includes the smallest library size of 26.8 million reads for the affected male sibling (0103_1) and the largest library size of 128.1 million reads for the mother (0103_3) (Appendix C). RNA-seq analysis of female sibling's X chromosome variants revealed 117 heterozygous SNPs with dbSNP137 identifier expressed at a coverage above 20X. These were phased to 62 maternal and 55 paternal SNPs. We found 96 unphased, heterozygous SNPs in the mother. Based on the allelic ratio distributions the female child had a moderately skewed XCI ratio of 87:7 while the mother had random XCI of 56:37. Scaling the XCI ratios to 0-100 scale the XCI ratios were extreme 93:7, and random 59:41, for the patient and the mother, respectively (Figure 4). Distribution of phased SNP alleles expressed from maternal X indicated bias against the expression of paternally inherited SNP alleles in the female child and suggested that the paternal X chromosome was only active in approximately 7% of the whole blood cells. The mutant allele in *WDR45* also showed a biased expression with over 97% of reads mapping to the mutant allele and result in an allelic ratio of 0.97. The bias seen towards the maternal X expression and expression bias to similar degree toward the mutant allele suggests that the mutant allele originates on the maternal X.

WDR45 is transcribed in whole blood and the male sibling showed the lower expression of *WDR45* compared to female sibling (FPKM= 51.0 vs. 78.77) (Table 3). Using Cuffdiff2 we compared the siblings to the parents to find that *WDR45* is dysregulated at a nominal significance (p=0.045), showing a downregulation of the *WDR45* transcript. This suggests a disruptive effect of the missense variant to mRNA stability. *WDR45* mediates protein-protein interaction, so we selected 7 genes that were shown to interact with *WDR45* (Behrends et al. 2010; Oláh et al. 2011; Fischer 2008; Emanuele et al. 2011). We found that LHX6 was not expressed in whole blood and only *APP* is dysregulated at a nominal level (p=0.049). This upregulation was primarily caused by the male sibling whose expression of APP was the highest. None of the nominal significance estimates remained significant after Benjamini-Hochberg correction.

Table 3.

Expression of *WDR45* and its interacting proteins.

Gene|Ensembl ID	FPKM (0103_1)	FPKM (0103_2)	FPKM (0103_3)	FPKM (0103_4)	Log2fold	p value
UBC|ENSG00000150991	521.21	795.34	1124.93	515.55	-0.538	0.224
CLNS1A|ENSG00000074201	35.71	44.51	29.02	26.38	0.252	0.418
LHX6|ENSG00000106852	0	0	0	0	0	-
SLC25A11|ENSG00000108528	26.49	35.41	43.29	29.03	-0.475	0.413
APP|ENSG00000142192	31.44	13.40	13.28	11.82	0.615	0.049
ATG2A|ENSG00000110046	13.92	11.15	19.13	6.49	-0.197	0.506
ATG2B|ENSG00000066739	3.71	2.93	2.65	1.97	0.320	0.427
WDR45|ENSG00000196998	51.00	78.77	86.76	93.36	-0.771	0.045

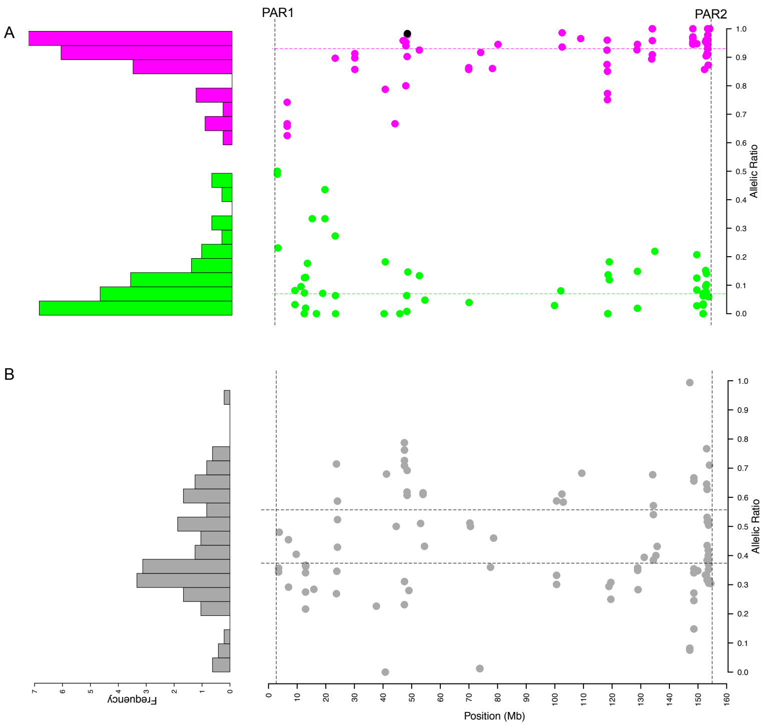

Figure 4. X inactivation by X-linked allele expression ratios. Then scatter plots show the allelic ratio of X-linked SNP variants. The corresponding reference allele ratios are not plotted. When the SNP alleles are phased an overall expression pattern of the parent-of-origin chromosome can be observed. We show the ratio of X-linked alleles between the pseudo-autosomal regions PAR1 and PAR2 on the terminal ends. **A.** The distribution of phased SNP alleles in P2 indicated a biased expression in favor of the maternally derived SNPs (magenta), over the paternal (green) X chromosome alleles. Histogram indicates a bimodal allelic ratio distribution. The black dot indicates the allelic ratio of the variant in *WDR45* suggesting that source of the mutation is the maternal X. Colored horizontal lines show the means of the paternal and maternal allelic ratio distributions scaled to 1, at 0.93 and 0.07, respectively. **B.** The distribution of allelic ratios in the mother. These X-linked variants are not phased and so colored uniformly. Histogram of the allelic ratios indicates non-normal distribution without evidence of bimodality. The two horizontal lines indicate the mean of the predicted allelic ratio distributions inferred from the data at 0.56 and 0.37, respectively.

Comparison of XCI ratio estimated by HUMARA and RNA-seq

In order to compare the utility of RNA-based estimation of XCI ratio, we determined XCI ratio from RNA-seq and from methylation assay by HUMARA for 48 total females from the Center For Rare Childhood Disorders. This cohort included 21 affected female patients (44%), 23 unaffected mothers (48%), 3 unaffected siblings (6%), and 1 grandmother (2%). We found 9 of the 48 enrolled female participants were uninformative (~19%) for the methylation assay due to homozygosity at the *AR* locus. Population scale analysis of heterozygosity in HUMARA analysis suggested ~8% of females are not polymorphic, and comparison with population data indicated that our cohort was enriched for uninformative female (Fisher's p=0.0372) (Amos-Landgraf et al. 2006). The enrichment of uninformative females was likely due to enrollment of families with multiple, related females homozygous for the *AR* allele. Methylation assay reports the XCI ratio of inactive *AR* allele over the active allele, and the signal peaks are scaled to 100. In un-scaled measurement, the ratio of the two X chromosomes did not always add up to 100 as the HUMARA assay does (Table 2). This is the result of variance in chromosome wide SNP expression across X. Un-scaled allelic ratios revealed 4 females with skewed XCI (>80:20) (~8%), when scaled, this number increased to six females (~12.5%)(Table 4). Of the 4 females predicted to have skewed XCI 2 were affected patients indicating that approximately 10% of female patients will have skewed XCI. We observed complete skewing in a single affected female patient (0118_1) based on methylation and categorized as extreme skewing by RNA-seq (Table 4). Not one participant had complete skewing by RNA-seq suggesting that X chromosome silencing is not complete across X, and that some genes may be expressed from both chromosomes. We also observed extreme skewing in 3 participants by methylation (0118_2, 0011_2, 0049_2), which were categorized as extreme (0118_2) moderately skewed (0049_2) and random XCI (0011_2) by RNA-seq. While only 2 participants (4%) had extreme skewing by RNA-seq, the methylation analysis identified 4 participants (8%). Interestingly 3 of the 4 participants with extreme skewing were mothers of affected female patients. This corresponds with previous reports that skewing can increase with age (Knudsen et al. 2007).

In 19 of the 48 participants segregation analysis could help us determine the phase of inactivation by analysis of parental variant calls. Nine participants showed biased expression of the paternal X. Additionally, nine females showed biased expression of maternally inherited X suggesting that in our small cohort the choice of maternal of paternal silencing is random. In a single case the parent-of-origin could not be determined because XCI was completely random and the phased allelic ratios were equal. In this female allelic ratio of additional phased SNPs could potentially help decipher parent-of-origin. On average 142±40 heterozygous SNP was evaluated within the X-linked regions for each female. The distribution of X inactivation ratio estimates from random 50:50 to 100:0 is right skewed towards the random XCI obtained from un-scaled RNA-seq data with 20 females (42%) of females categorized in this group (Figure 5). The distribution of XCI ratios between scaled and HUMARA data follow similar trend suggesting that scaling shift the ratios toward a normal distribution. We also observed two families (0002, 0047) with familial homozygosity at the AR locus which would normally be an uninformative test, allelic expression shows random XCI in each case increasing the available information for a more comprehensive view of molecular data. In addition in 4 of the 9 uninformative cases we were also able to determine the phase of X inactivation.

Table 4.

Results of XCI ratio estimation by HUMARA and RNA-seq.

Family	individual	status	HUMARA*	RNA-seq^{+} (un-scaled)	RNA-seq^{++} (scaled)	preferentially silenced X	# X-linked SNPs
0001	0001_1	affected	62:38	69:30	70:30	Xp	144
0001	0001_2	mother	65:35	59:39	60:40		139
0001	0001_3	sibling	60:40	46:44	51:49	Xp	212
0004	0004_2	mother	70:30	60:35	63:37		157
0011	0011_1	affected	77:23	68:30	69:31	Xm	163
0012	0012_1	affected	56:44	50:47	52:48		157
0014	0014_2	mother	65:35	61:31	66:34		191
0016	0016_2	mother	62:38	57:34	63:37		166
0018	0018_2	mother	71:29	75:22	77:23		150
0019	0019_1	affected	53:47	54:38	59:41	Xm	128
0019	0019_2	mother	55:45	64:34	65:35		100
0020	0020_1	affected	60:40	59:42	58:42		219
0025	0025_2	mother	65:35	43:54	44:56		129
0029	0029_1	affected	52:48	49:48	51:49	Xp	132
0033	0033_1	affected	63:37	57:39	59:41	Xm	122
0033	0033_2	mother	64:36	61:35	64:36		138
0034	0034_2	mother	64:36	58:34	63:37		159
0046	0046_1	affected	50:50	47:46	51:49		241
0048	0048_1	affected	70:30	56:38	60:40	Xp	92
0048	0048_2	mother	79:21	61:30	67:33		98
0049	0049_1	affected	71:29	64:28	70:30	Xm	156
0091	0091_1	sibling	68:32	61:38	62:38	Xp	116
0091	0091_3	grandmother	70:30	75:17	82:18		66
0117	0117_2	mother	55:45	47:47	50:50		114
0139	0139_2	mother	58:42	59:34	63:37		142
0140	0140_2	affected	60:40	62:42	60:40		143
0152	0152_2	mother	51:49	50:46	52:48		63
0157	0157_1	affected	55:45	46:41	53:47	Xm	97
0002	0002_2	affected	uninformative	47:47	50:50	equal	156
0002	0002_3	sibling	uninformative	63:35	65:35	Xp	188
0002	0002_3	mother	uninformative	74:25	76:24		189
0008	0008_9	affected	uninformative	47:45	51:49	Xm	123
0034	0034_1	affected	uninformative	73:25	74:26	Xp	194
0047	0047_1	affected	uninformative	49:48	51:49	Xp	171
0047	0047_2	mother	uninformative	74:25	75:25		159
0059	0059_2	mother	uninformative	65:33	66:34		144
0091	0091_2	mother	uninformative	48:48	50:50		96
0157	0157_2	mother	74:26	72:12	86:14		69
0008	0008_2	mother	88:12	73:23	76:24		96
0014	0014_1	affected	83:17	72:23	76:24	Xm	239
0023	0023_1	affected	84:16	77:22	78:42	Xm	133
0024	0024_2	mother	81:19	62:29	68:32		134
0059	0059_1	affected	82:18	67:27	71:29	Xm	134
0018	0018_1	affected	84:16	81:18	82:18	Xp	103
0011$$	0011_2	mother	93:7	78:23	77:23		148
0049	0049_2	mother	90:10	81:19	81:19		133
0118	0118_1	affected	100:0	90:6	94:6		140
0118	0118_2	mother	96:4	90:8	92:8		119

*=ratio is defined by the proportion of methylated (inactive) X over the proportion of unmethylated X (active). +=ratio is defined by the allelic ratio of higher frequency SNP alleles over SNP alleles with lower frequency. ++=XCI ratio is scaled to 0-100 from un-scaled ratios by normalizing the additive proportions of the SNP ratio distributions to 100.
$$=Light grey shaded rows indicate samples where only methylation assay predicted extreme X skewing. Dark grey shaded rows indicate cases where both HUMARA and RNA-seq predicted extreme skewing. Xp= paternal X chromosome, Xm=maternal X chromosome

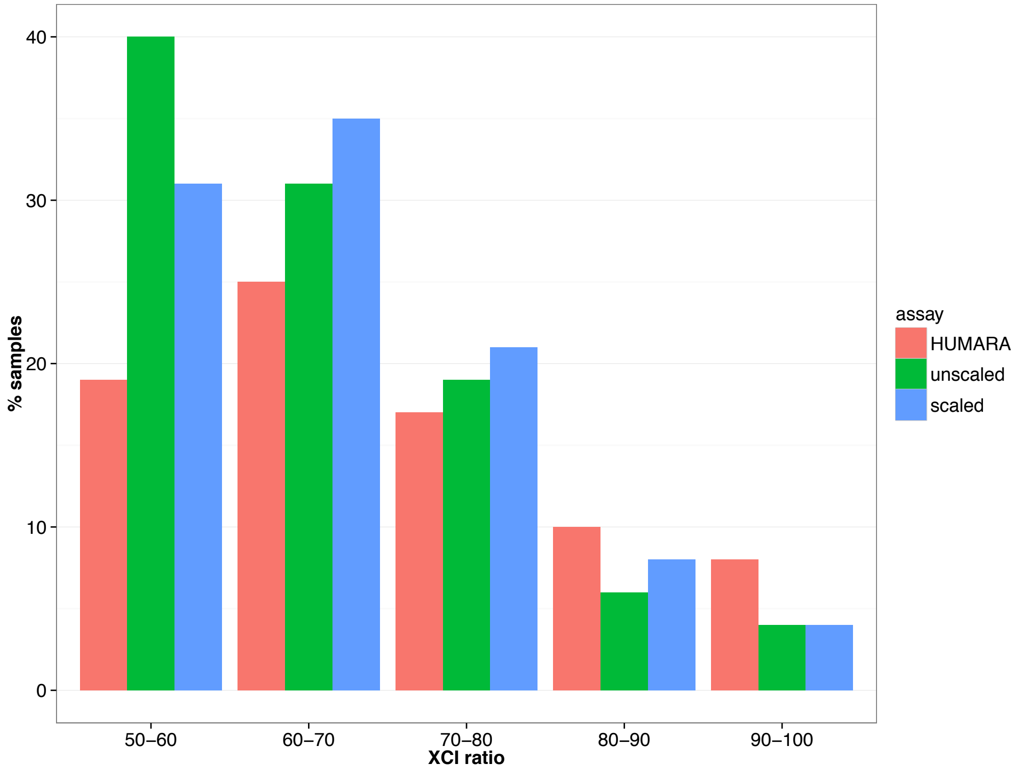

Figure 5. Distribution of X inactivation. The axes indicate 5 arbitrary bins of X inactivation ratios and the percent of total samples in each category. Each estimation method is listed in the legend. Un-scaled estimates of XCI by RNA-seq indicate an enrichment of XCI ratios at the random 50:50 level due to the difficulty of the algorithm to differentiate between overlapping allelic ratio distributions.

Previous studies of allelic expression XCI estimates have been inconclusive whether direct expression based XCI analysis correlate with DNA methylation (Amos-Landgraf et al. 2006; Swierczek et al. 2012). To that effect, we evaluated the correlation between the RNA-seq derived XCI estimates and the HUMARA method. Correlation was estimated for the 39 informative datasets with both HUMARA and RNA-seq data. Using Spearman's rank-order correlation we found statistically significant correlation between expression and methylation based estimates (Figure 6). There was strong linear correlation between HUMARA assay results and un-scaled RNA-seq estimates (Spearman: S= 1636.727, ρ = 0.834, P= 4.157e-11). Scaling the expression estimates improved the linear relationship and significance although at the more extreme XCI ratios HUMARA predicted twice as many extreme events than expression methods (Spearman: S= 1472.278, ρ = 0.850, P= 6.801e-12). This may resulted from the fact that expression estimates are based on the mean of the allelic ratio distribution and variance in allelic expression due to incomplete silencing of X, influence of imprinting, cis-, trans-regulatory elements on SNP expression may moderate extreme ratio estimates. Moreover, scaling the expression estimates preserved strong correlation with un-scaled ratios overall, although in families 091 and 0157 scaling resulted in a shift from random XCI to moderately skewed which can impact biological interpretation. Scaling the RNA-seq data did not significantly changes un-scaled estimates, as correlation of the two RNA-seq estimates was significant (Spearman: S= 410.518, ρ = 0.958, P= 2.2e-16).

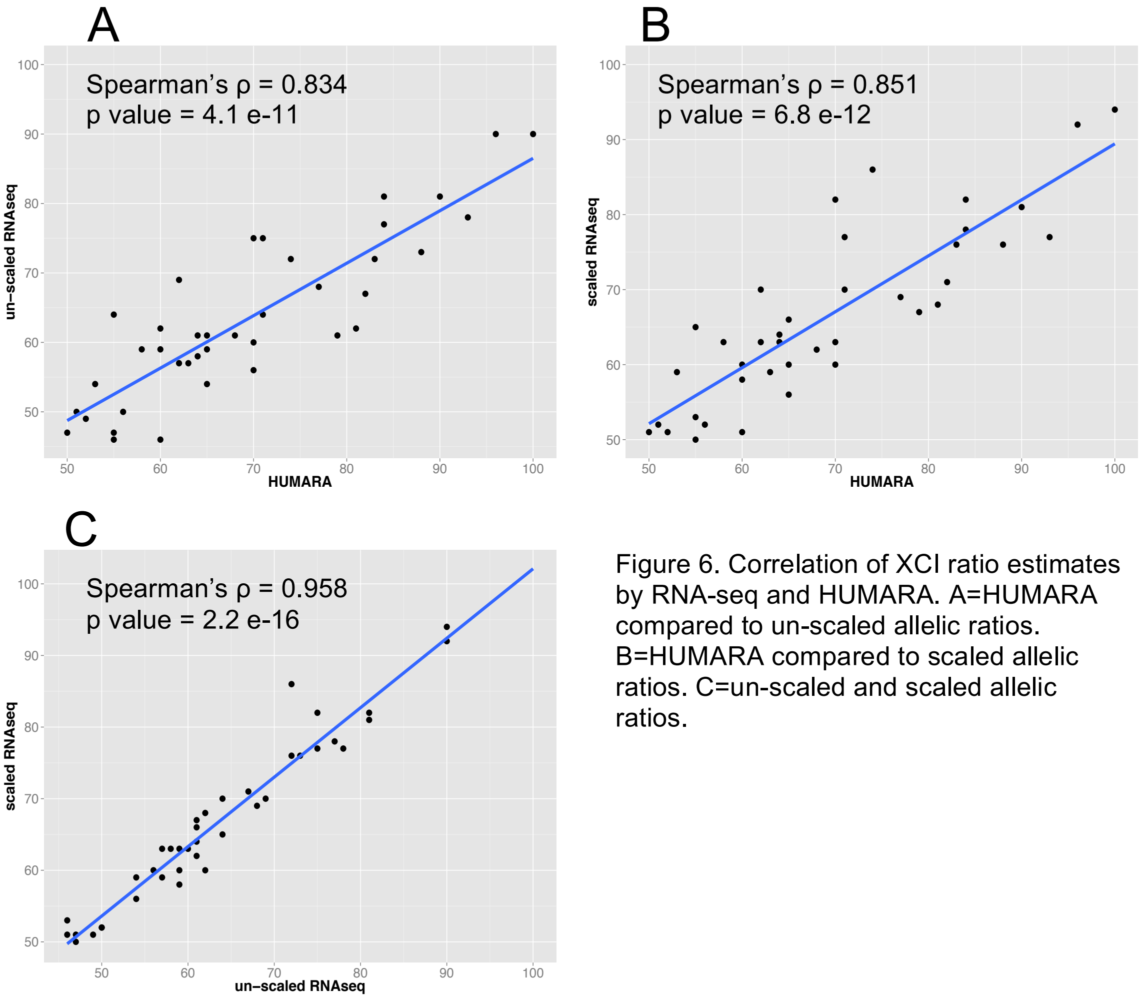

Figure 6. Correlation of XCI ratio estimates by RNA-seq and HUMARA. A=HUMARA compared to un-scaled allelic ratios. B=HUMARA compared to scaled allelic ratios. C=un-scaled and scaled allelic ratios.

Discussion

In this study, we obtained insight into the phenotypic variability in beta-propeller associated neurodegeneration by the first integrated whole-exome and RNA-seq study of a male-female sibling pair diagnosed with BPAN. Characteristic features like brain iron accumulation in the globus pallidus, and the substantia nigra, progressive neurological and psychomotor decline and seizures, Parkinsonism and dementia in adulthood are all common diagnostic of BPAN. BPAN is also known as static encephalopathy of childhood with neurodegeneration in adulthood and heterogeneity has been reported in disease manifestation (Saitsu et al. 2013; Hayflick et al. 2013). The male sibling, who has passed on, presented a more severe phenotype including epileptic, myoclonic seizures, spastic quadriplegia, and cerebral atrophy. The female sibling shows Rett-like symptoms including hyperreflexia, truncal hypotonia, and hand clasping, she is able to walk short distances, and manifests white matter volume loss.

Sequencing in the family revealed a shared, *de novo* dominant missense mutation in the X-linked *WDR45* gene. *WDR45* is a member of the WD repeat domain proteins with multiple homologs on the autosomes (Haack et al. 2012). It contains multiple, conserved 40 amino acid residues usually terminated by a tryptophan-aspartic acid repeat residues(D. Li and Roberts 2001). Their role has been implicated in signal transduction, regulation of protein complex formation, and cell-cycle control. WD repeat proteins contain a symmetrical, seven-bladed, beta-propeller motif that mediates protein-protein interaction (Haack et al. 2012). *WDR45* has been implicated in autophagy, the cell's intracellular degradation system that transports cytoplasmic molecules for degradation to the lysosomes (Lu et al. 2011). Saitsu et al. (2013) showed using autophagic flux assay that *WDR45* mutant lymphoblastoid cell lines present a blockage in the autophagic flux and affect autophagosome formation. Knockdown of rat *Wdr45* results in accumulation of autophagic structures (Lu et al. 2011). In addition the importance of autophagy in neurodevelopmental disease has been implicated as mice lacking autophagy in neurons were seen to develop psychomotor dysfunction (Hara et al. 2006).

We identified a missense mutation in exon 10 of *WDR45* that leads to the substitution of a conserved leucine to proline (Figure 3C). This amino acid residue change was in the 6th WD repeat domain of *WDR45*. Previously an in-frame deletion was reported 2 amino acid residues upstream in this exon in a middle aged female who is wheelchair bound, with severe intellectual disability, and tonic, clonic seizures that overlaps with female sibling's phenotype (Verhoeven et al. 2014). All three males reported thus far carried frameshift indels (Haack et al. 2012). It has been suggested that males with germline mutations are non-viable, and severity may depend on when the mutations occur during development. The siblings share a *de novo* mutation in germline DNA suggesting a low probability for the mutation to occur after embryogenesis. This is supported by the fact the male sibling was severely affected. However the female sibling showed milder phenotype suggesting that if the mutation occurred prior embryogenesis some other molecular mechanism may have altered her symptoms.

Previously, X inactivation was implicated as a mechanism that may explain phenotypic similarity between males and females and for the second part of our study we looked at new approach to study the role of X inactivation in the heterogeneity between the siblings. Using RNA-seq instead of traditional methylation assay, we found that the female patient had a 93:7 extremely skewed XCI in favor of the maternal X chromosome. To determine the chromosome where the mutation situated, we had to rely on allele frequency data from RNA-seq, as germline mutation occurred *de novo.* However, correlating the chromosome wide allelic ratios to the allelic expression of the wild type and mutant alleles could be used to infer the parent-of-origin of the chromosome with the mutant allele of the variant and wild type allele. Allele specific expression was observed at the mutant allele from allele ratio of 0.48 in the DNA to 0.93 in RNA. By inference, we concluded that the SNP allele likely resides on the maternal X chromosome. This was supported by almost complete loss of the wild type allele expression. The dominant allele, the SNP allele however showed very similar allele bias to the maternally inherited SNP alleles. This finding implicates maternal germline or gonadal mosaicism as the mother's blood DNA shows only wild type alleles. Previously, a familial Rett Syndrome case also implicated maternal germline mosacism as a mechanism to phenotypic heterogeneity among sibling with same X-linked mutation (Venâncio et al. 2007). In addition, Danda et al. reported two female siblings with a rare X-linked Oculo-facio-cardio-dental (OFCD) syndrome (MIM 300166) with a shared *de novo* mutation in *BCOR* an X-linked gene that was only found in the siblings (Danda et al. 2014).

RNA-seq also identified a possible mode for phenotypic variability. We found that the male sibling had most reduced expression of *WDR45* in blood and loss of function mutations in this gene lead to loss of protein product suggesting a reduced mRNA stability (Saitsu et al. 2013). However the female patient not only shows higher *WDR45* abundance in blood, she shows expression of the wild type allele suggesting that a portion of her cells express the normal protein. Saitsu et all showed that both missense and loss-of-function variants lead to protein degradation suggesting that the male sibling is likely have no protein expression in whole blood thus autophagy is severely impacted (Saitsu et al. 2013). It has been shown that even in genes that are subject to X inactivation a leaky expression can be detected in mice hybrid cells, suggesting that even at low wild type allele frequency the female patient may produce the wild type protein (F. Yang et al. 2010). BPAN phenotype is mostly brain specific, so the possibility that the female patient expresses some level of wild type protein, and that the boy only expressed the mutant could lead to the male lethality and a rescue of the more severe phenotype in the female sibling. To elucidate the role of X-inactivation in the phenotypic spectrum of this sibling pair, parent-of-origin of X inactivation of other reported cases of BPAN may be necessary as those female patients with skewed XCI have not been completely characterized (Haack et al. 2012; Saitsu et al. 2013).

RNA-seq analysis of XCI showed high correlation with DNA methylation assay. Traditionally skewing has been estimated by the methylation assay of the *AR* locus. This assay has shown good correlation with other quantitation methods based on pyrosequencing of cDNA (Mossner et al. 2013). Other expression based methods showed little correlation highlighting the problematic nature of using a small number of genes to determine XCI (Swierczek et al. 2012). Instead of selecting alleles in specific genes, integration of the genomic and functional sequencing data we were able to study allelic expression across the X-linked region of X greatly improving our ability to predict XCI status. The mean of X-linked SNP allele expression rather than one or few genes can reduces noise and improve accuracy (Cotton et al. 2013). It should be noted that while HUMARA method was uninformative in 9 cases due to homozygosity at a single locus, leveraging over hundred high quality, expressed alleles in each participant's RNA-seq method provided an XCI estimate in each participant. Variability of SNP allele expression across chromosome X resulted that that the ratio of Xi and Xa did not add up to 100 in most cases. Thus a slightly different scale confounded comparison of un-scaled XCI estimates to HUMARA assay. This inconsistency in allelic expression can be attributed to variable silencing of genes across X (Carrel and Willard 2005). Thus, pre-selection for genes that are only expressed from the active X may be able to bring XCI estimate by expression to same scale as HUMARA, but may reduce the number of SNPs to estimate XCI. Our method for scaling allele ratios and provides a basis to compare XCI ratios by HUMARA and allelic expression.

RNA-seq approach identified two female patients from the same family with extreme skewing. The female patient with extreme XCI was diagnosed with Aicardi Syndrome. Although previously, Eble et al. (2009) showed that 18% of Aicardi patients have extreme skewing, the mother in this family is un-affected suggesting a different mechanism to this patient's phenotype. Although familial skewed XCI are rare and may be by chance alone, some cases have been reported in haemophilia B and X-linked adrenoleukodsytrophy suggesting that genetic mechanisms of unidentified gene mutations may contribute to the inheritance of the mutation (Ørstavik, Orstavik, and Schwartz 1999; Z. Wang et al. 2013).

RNA-seq and segregation analysis could identify the parent-of-origin of XCI in 19 patients, aiding interpretation in the context of clinical symptoms. In half of the cases the paternal X and other half maternal X was silenced. Our cohort suggests that selection of X to be inactivated by the X inactivation process is not determined by the origin of the X chromosome but is likely determined by genetic and epigenetic factors of each chromosome. This is supported by studies showing that the choice of X inactivation can be influenced by epigenetic events that are not well understood (Gribnau et al. 2005).

In conclusion, we successfully applied integrated DNA and RNA sequencing to better understand the molecular mechanism of an X-linked disorder and its heterogeneity in a case of affected sibling pairs with shared *de novo* mutation. The two siblings share the same mutation but

XCI ratio analysis, shows that there is low expression of the wild type allele in a subset of the female sibling's cell that can lead to a less severe phenotype. We found that *WDR45* was most dysregulated in the male patient supporting the phenotypic heterogeneity between the siblings. Segregation analysis of parental genotypes and XCI analysis implicate maternal gonadal mosacism as the most likely source of the mutation and molecular mechanism. In addition we performed RNA-seq on a total of 48 females from our study, which showed high correlation with standard methylation assay. We were able to identify a familial extreme inactivation that pointed out the role X inactivation played in the female patient's phenotype. Our method improved on current assay by reporting XCI for all subjects that were uninformative for the methylation assay and added parent-of-origin information to standard quantitative analysis. Integration of next-generation sequencing methods in rare diseases will lead to a more comprehensive view of disease etiology and reduced need for individual clinical assays in patient management.

CHAPTER 3

CHARACTERIZATION OF X CHROMOSOME INACTIVATION USING INTEGRATED ANALYSIS OF WHOLE-EXOME AND MRNA SEQUENCING

Introduction

In this chapter we set out to develop a method to quantify X inactivation ratio using simultaneous sequencing of DNA and RNA. We first use simulated data to show the utility of integrated data to quantify the proportion of active and inactive X chromosomes in females. Next we apply this method to a clinical case where skewed X inactivation was identified prior this study.

Diagnosing and uncovering the genetic basis of disease has been revolutionized by WES, allowing discovery of new disease genes and improving the rate of clinical diagnosis for rare genetic conditions. Indeed, the genetic basis of childhood disorders can be identified in approximately 25% of patients, where successful molecular diagnosis frequently has a major impact on patient management and treatment (Dixon-Salazar et al. 2012; Y. Yang et al. 2013). Prioritization of candidate variants for the remaining patients remains challenging due mainly to insufficient understanding of the functional consequence of substantial fraction of candidate variants (Gilissen et al. 2012). Large scale functional characterization of genomic variation by simultaneous DNA and RNA sequencing from a patient can reveal genotype-phenotype correlation, can highlight gene expression profile that is associated with the studied genetic condition, and allows immediate evaluation of *in silico* prediction algorithms to the effect genomic variants have on gene expression, alternative splicing, exon usage, gene fusions (Z. Wang, Gerstein, and Snyder 2009). In breast and pancreatic cancer integrated analysis of DNA and RNA has been successfully utilized to obtain insight into molecular mechanisms that explain pathogenicity and uncovered potential therapeutic targets to improve patient management (Shah et al. 2012; Craig et al. 2013; Liang et al. 2012). In addition, RNA-seq has been utilized in the context of the affect epigenetic modifications have on gene expression (Babak et al. 2008; X. Wang et al. 2008). Integrative analysis of WES and RNA-seq data in X-linked disorders may also be informative both in diagnosis and gene discovery for phenotypes emerging caused by epigenetic changes such as XCI (Lyon 1961).

In the process of XCI, in females, cells undergo epigenetic inactivation of one of the inherited, parental X chromosomes resulting in consecutive daughter cells expressing one X (Muller 1932; Augui, Nora, and Heard 2011). The proportion of cells with either parental X as the active is defined by the XCI ratio that ranges from 50:50 random to 100:0 completely skewed. Epigenetic analysis of X chromosome in unaffected females indicate that XCI ratio normally distributed in the general population (Amos-Landgraf et al. 2006). Although, on the cellular level X-linked alleles are expressed in a dominant fashion, in cell populations X-linked alleles show mosaic pattern of expression, which can lead to heterogeneous phenotypes in females who are carriers for disease causing, deleterious mutations (Migeon 2006). In X-linked neurological disease, mode and magnitude of XCI can influence disease severity and outcome (Ørstavik 2009). Indeed, case-control studies demonstrate that skewed XCI is common among females who are carriers for X-linked Mental Retardation disorders (XLMR) (Plenge et al. 2002). XCI may also lead to asymptomatic carrier status by selective advantage of cells expressing the wild-type alleles(Van Esch et al. 2005). One of the difficulties diagnosing females with X-linked diseases and skewed XCI is the broad and overlapping description of clinical phenotype, the limited availability of similar patients, and lack of high-throughput, expression-based methods to estimate XCI(Ørstavik 2009). Routine, clinical method to estimate XCI ratio rely on the HUMARA differential DNA methylation assay that targets a polymorphic short tandem repeat (STR) in the human androgen receptor gene (*AR*) (Allen et al. 1992). Methylation of this repeat is associated with XCI. Although >90% of females are polymorphic at this site, it provides expression information indirectly from DNA, and, relies on a single locus (Amos-Landgraf et al. 2006). There is also conflicting evidence whether DNA methylation can reflect the quantitative expression ratio of active X (Xa) to inactive X (Xi) compared to allele-expression-based methods (Busque et al. 2009; Swierczek et al. 2012). Using next-generation sequencing of DNA and RNA simultaneously, we can scan for potential disease causing variations, and at the same time learn about the functional implications of genomic changes with the additional benefit of learning about transmission of alleles and potential imbalance in chromosome X expression. By phasing X-linked variant alleles, we can learn about the mode, or parent-of-origin of imbalance, and the magnitude can be estimated from direct measurement of relative expression of chromosome-wide heterozygous alleles.

In this chapter we present genetic and functional analysis from high-throughput sequencing of WES and RNA-seq to both (1) identify potentially pathogenic genetic mutations and (2) identify XCI ratio using phased and unphased allele-specific expression analysis. We show that high-throughput sequencing can be utilized to estimate XCI ratio on simulated data and we apply our approach to a patient with undiagnosed, heterogeneous phenotype. Using family-trio based WES with segregation analysis, we characterized a *de novo*, heterozygous deletion on Xp22.31 as potentially pathogenic, and we identified a moderately skewed XCI ratio from the RNA-seq experiment. Integration of exome and expression data revealed that the deletion occurred on the paternal X (Xp), and skewed XCI favored the expression of the cytogenetically normal, maternal X (Xm), suggesting a mechanism for the mild neurological phenotype.

Materials and Methods

In Silico Experiment

XCI results in two cell populations in females, one expressing Xm, the other expressing Xp. In theory, the degree of cellular mosaicism of X-linked allele expression can be estimated by RNA-seq using count-based approach (Figure 7A). In this approach, we obtain digital measurement of allele expression from Xm and Xp by counting sequenced reads mapping to each allele, which is directly related to the expression of the chromosome with the allele. On the X chromosome, the allele counts come from either Xa, or Xi, and the ratio of allele frequencies at a heterozygous locus correlates with the overall XCI status of the Xp and Xm chromosomes in the tissue. However, epigenetic modifications, including DNA methylation, cis-, and trans-acting elements, and chromosome strata can influence allele expression at a single locus. Therefore, chromosome-wide heterozygous allele frequency ratio can provide a better estimate of the overall expression of each parental X. In addition, when the transmission of the allele can be determined from parent to offspring by segregation analysis, and variants can be assigned a parental origin (i.e. phase), phasing the alleles can identify the parental X that is preferentially inactivated or activated. To evaluate this approach, we simulated RNA-seq reads with female, heterozygous genotypes from a pool of known, X chromosome SNPs in coding regions from the ESP6500 NHLBI Exome Sequencing Project (NHLBI Exome Sequencing Project,). The 4996 SNPs were randomly binned in two sets analogous to maternal or paternal SNPs (i.e. phased SNPs) by rand function of a perl script. In the first set, the alternative allele of the genotype was assigned as paternal (Alt-P, n=2520), and in the second set (Alt-M, n=2476), the alternative allele was assigned as the maternal allele. Using *seqtk* FASTA processing tool (seqtk) the Alt-M and Alt-P alleles were introduced into two separate chromosomes X transcriptome fasta files containing known transcripts greater than 500bp from Homo sapiens.GRCh37.62.gtf. The two modified fasta files were analogous to an X transcriptome with maternal variant alleles and one with paternal variant alleles. Next 10 million, 100bp paired reads in fastq format were generated, mapping to the two transcriptome files from above (5 million read1 and 5 million read2) using *wgsim* 0.3.1-r13 fastq simulator (wgsim). Command line options for *wgsim* included zero indel error rate, an outer distance of 150bp between the paired reads, a uniform Phred quality score of 40 for each base, and a 0.001% base error rate. The combination of these two parental, Alt-M, and Alt-P allele containing fastq files in various ratios followed by mapping them back to the chromosome X reference, and followed by estimation of allelic expression by read count provides the basis for the estimation of XCI ratio. Essentially, after the two modified fastq files with 10 million reads were generated, *seqtk* was used to subsample them randomly, and merge each set into a single fastq file analogous to the reads obtained through RNA-seq of an experimental sample. When, for example, XCI ratio of 75:25 was simulated, 7.5 million correctly paired reads were randomly sampled from Alt-M alleles containing fastq file and 2.5 million were subsampled from Alt-P fastq file and merged. In theory, after alignment and allele count, there would be a 75:25 allelic imbalance in favor of the Alt-M alleles to an overall chromosome wide 75:25 ratio since approximately 75% of reads contain alleles from Alt-M. Using this approach, RNA-seq reads were simulated for 11 expected X inactivation ratios: completely skewed X inactivation (100:0), extremely skewed X inactivation (95:5, 90:10), moderately skewed X inactivation (85:15, 80:20), and random X inactivation (75:25, 70:30, 65:35, 60:40, 55:45, 50:50).

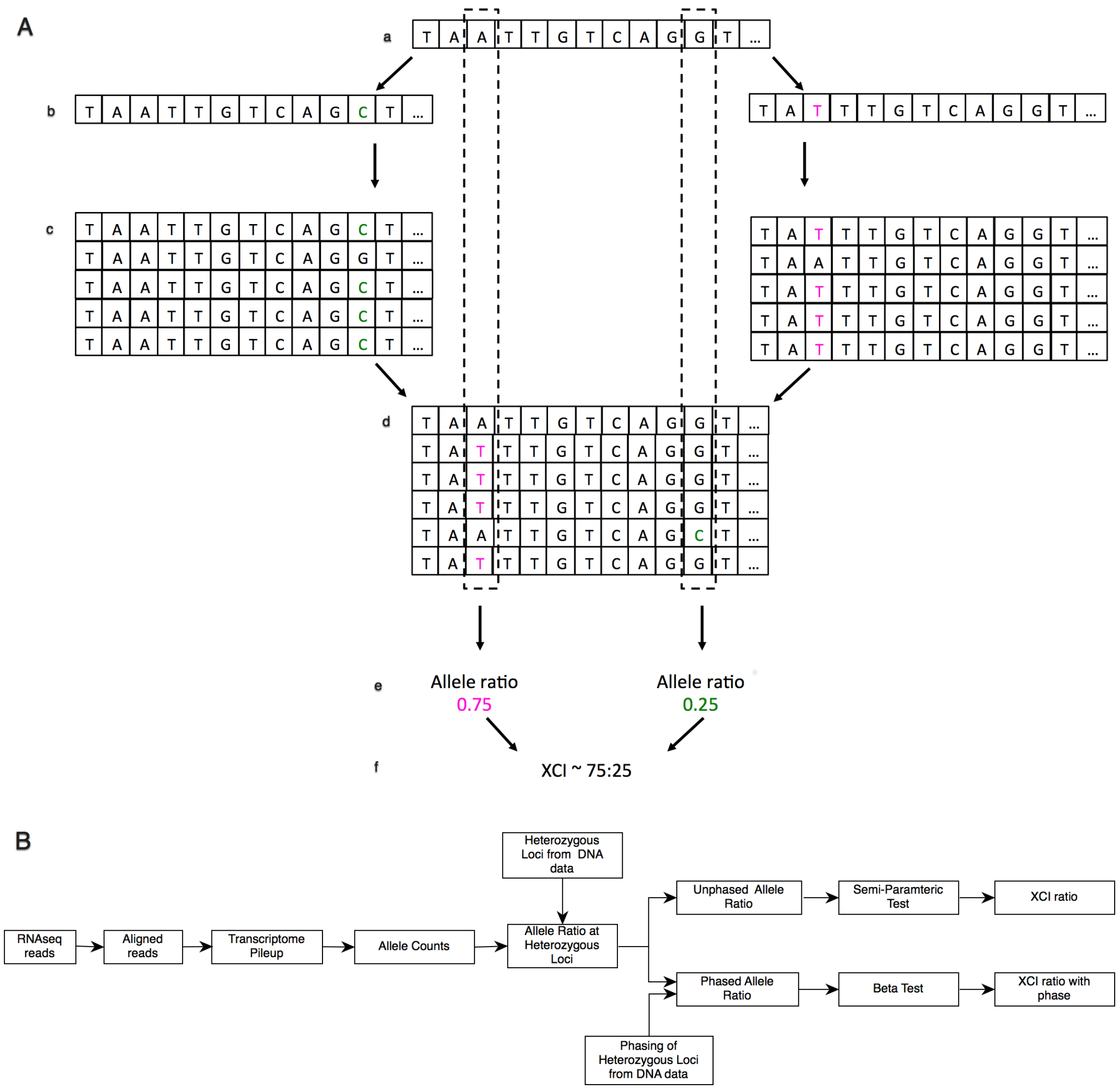
A
a
T A A T T G T C A G G T ...
b
T A A T T G T C A G C T ...
T A T T T G T C A G G T ...
c
T A A T T G T C A G C T ...
T A A T T G T C A G G T ...
T A A T T G T C A G C T ...
T A A T T G T C A G C T ...
T A A T T G T C A G C T ...
T A T T T G T C A G G T ...
T A A T T G T C A G G T ...
T A T T T G T C A G G T ...
T A T T T G T C A G G T ...
T A T T T G T C A G G T ...
d
T A A T T G T C A G G T ...
T A T T T G T C A G G T ...
T A T T T G T C A G G T ...
T A T T T G T C A G G T ...
T A A T T G T C A G C T ...
T A T T T G T C A G G T ...
e
Allele ratio
0.75
Allele ratio
0.25
f
XCI ~ 75:25
B
Heterozygous Loci from DNA data
RNAseq reads
Aligned reads
Transcriptome Pileup
Allele Counts
Allele Ratio at Heterozygous Loci
Unphased Allele Ratio
Semi-Paramteric Test
XCI ratio
Phased Allele Ratio
Beta Test
XCI ratio with phase
Phasing of Heterozygous Loci from DNA data

Figure 7. Schematic view of estimation of XCI ratio from read counts data. (A) Overview of the simulation study. From a reference transcriptome (a), two haplotypes are simulated with known variant alleles (b). Sequence read simulator generates reads with error attributes using the two haplotypes as reference (c). The reads from both read simulations are merged and aligned back to the original reference (d, dashed lines). Counting the number of reads mapping to each known allele, the allelic ratio of mapped variant alleles can be determined (e). The overall XCI ratio is determined for large number of variants by estimating the mean of the allele ratio distributions of multiple alleles (f). (B) Workflow of XCI estimation from RNA-seq experiment using phased and un-phased approaches. Essentially, RNA-seq reads are aligned followed by obtaining the transcriptome pileup at each sequenced loci. This is followed by counting the number of reads mapping to each allele across the transcriptome. Next, loci are reduced to those that contain heterozygous calls in the genomic DNA and allelic ratio is calculated at each heterozygous locus. If there is no available information on the phase of X-linked alleles at heterozygous loci, the un-phased, X-linked allelic ratios are evaluated for their distribution using semi-parametric model and XCI is reported from the parameters of the semi-parametric model. When transmission of alleles can be obtained from DNA data, the phased, X-linked allele ratios are evaluated by the beta distribution and XCI reported from the parameters of the beta model with the phase of XCI.

Estimation of XCI Ratio.

Estimation of XCI ratio followed similar steps in both the *in silico* experiment and for the patient (Figure 7B). Reads were aligned to human reference genome GRCh37.62 using TopHat2 (Kim et al. 2013). Alignment of next generation sequencing data has reference bias that may influence the allelic ratio estimate of SNP alleles. Reduction of bias can be achieved by read alignment to diploid reference incorporating parental genotype information or by reduction of mapping stringency by increasing the number of mismatches allowed in a read for alignment (Rozowsky et al. 2011; Stevenson, Coolon, and Wittkopp 2013). Therefore five and four mismatches per 100bp read length were allowed in the *in silico* and clinical experiments, respectively. Allele counts were obtained by generating a chromosome wide pileup with SAMtools mpileup command (H. Li et al. 2009). Bases with Phred quality score > 20 were counted only in the in silico and clinical experiments. Pileup was parsed by an in-house perl script. Next the allelic ratio at each heterozygous locus was calculated by dividing the number of reads mapping to the variant allele with the total number of reads mapping to the locus. After allelic ratio calculation the SNPs were further filtered for quality by following procedure: (1) SNPs within the PAR1 and PAR2 pseudo-autosomal regions were filtered out as they follow autosomal inheritance and can bias XCI ratio(Mangs and Morris 2007) (2) Filtered for high confidence variant loci from exome dataset with a genotype filter score of PASS by GATK VariantRecalibrator (McKenna et al. 2010). (3) Loci without a dbSNP identifier were filtered out (4) Variants with less than 20X coverage were filtered out.

First, phased alleles were used to estimate XCI ratio. Phasing was performed in the *in silico* experiment by assigning the heterozygous variants into their respective Alt-P and Alt-M bin, and by genotype phasing of the trio in the family study as described below. Phasing of X-linked heterozygous variants allows us to evaluate the functional profile of each inherited parental copy. By estimating the parameters (mean, variance) of each copy's allele ratio distribution we can estimate the proportion of cells with Xm or Xp as active and inactive (eg. mean allelic ratio of paternal alleles of 65% and mean allelic ratio of maternal alleles of 35 equals an estimated XCI ratio of 65:35). To control for over-dispersion of read count data from RNA-seq, phased allelic ratios were fitted to the beta distribution to estimate their mean and variance using the fitdistr module of MASS package in R (MASS) (Skelly et al. 2011; Zhou, Xia, and Wright 2011; Hardcastle and Kelly 2013; Sun 2011).

Next, XCI ratio was also estimated without phasing the alleles. When phasing information is unavailable we can lose our ability to define the activity of the parental chromosomes. In this case, the inheritance is unknown and the distribution of allele expression from the two chromosome copies may overlap suggesting similar proportion of cells with one of the parental copies active. However, alleles sampled from the two chromosome copies can have their unique distribution pattern resulting in multi-modal allele distributions. Multi-modal distributions can be understood as a mixture of two or more distributions and thus mixture models based on the expectation maximization (EM) algorithm may be used to estimate the parameters of each component or mode of the distribution. The problem with normal mixture modeling is that the number of components in the data set can greatly affect outcome and advised to account for prior modeling. The semi-parametric (SP) model, however, has no assumptions about the modality or the normality of the data and can also approximate the parameters of each component in a data distribution. In estimation of the inactivation status of the X chromosomes, the mean allelic ratios estimated by the SP model can directly correlate to the proportion of cells carrying the variant alleles. Thus, allelic expression captured in component 1 and 2 of a multi-modal allelic distribution can be thought of as indicators of the proportion of activity of parentally inherited chromosomes in the tissue. The SP method is motivated by the fact that the choice of a parametric family may not always be evident from the distribution of the data, as it is in over-dispersed and heavy-tailed distributions (Hunter, Wang, and Hettmansperger 2007). We applied Bordes et al. stochastic expectation-maximization algorithm for estimating SP model parameters for unphased data (Bordes, Chauveau, and Vandekerkhove 2007). The mean of the estimated component distributions were utilized as the expression status of each inherited chromosomes but were blind to the origin of alleles and applied to define the XCI ratio.

Family Study

The participating family of Northern European ancestry provided written consent and was enrolled into the Center For Rare Childhood Disorders Program at the Translational Genomics Research Institute (TGEN). The patient was 12 years old at the time of enrollment and verbal assent was obtained from her and documented in writing by the consenting staff person. In addition, written consent for the minor under the age of 18 years was obtained from the parents. All additional participants over 18 years of age provided written consent at the time of enrollment. The study protocol and consent procedure was approved by the Western Institutional Review Board. The primary goal of enrollment is to utilize family-trio based WES in the clinical diagnosis of previously undiagnosed, rare conditions suspected of genetic cause. The female child, now 14 years old had no clinical diagnosis at the time of enrollment, although complex neurobehavioral condition was suspected based on manifesting phenotype of emotional instability, attention deficit, and delays in development and learning. She was born at 38 weeks gestation, and required minimal respiratory assistance. There were early concerns about her development, as she didn't walk until 13-14 months of age. Behavioral problems were noted at age 2, consistent with current phenotypic description above. Treatments with medications for poor attention, impulsivity, repetitive behaviors, and learning difficulties started at age 5. She did not have convulsive seizures, but subtle events consisting of staring, loss of awareness, and tremulousness had been observed. MRIs of the brain were normal; EEG showed right posterior temporal sharp waves. The patient had an older unaffected brother, and her neurological examination was normal showing concrete ability to respond and interpret questions. Previous genetic analysis of genomic DNA from whole blood by array-based comparative genomic hybridization (aCGH) identified a heterozygous deletion between positions 6.4-8.1 Mb on chromosome X. Additionally, HUMARA DNA methylation assay at the *AR* gene identified 85:15 skewed X inactivation within peripheral blood, providing a hypothesized mechanism for the patient's moderate phenotype. To find possible causal variants that may explain her condition and to validate previous genetic and epigenetic findings whole-exome and RNA-seq sequencing was completed on genomic DNA and mRNA isolated from peripheral blood for the mother, father, and patient. Whole blood was collected into EDTA Blood tubes and PaxGene RNA tubes. Genomic DNA was isolated with DNeasy Blood & Tissue Kit (Qiagen, Germantown, MD), and total RNA was isolated from PaxGene RNA tubes using PaxGene Blood miRNA kit (Qiagen, Germantown, MD) following manufacturer's suggested protocol. Exome capture and library preparation was performed with 2μg of input genomic DNA for each participant using the TruSeq DNA sample preparation kit v2 and the TruSeq Exome Enrichment kit v2 (Illumina, San Diego, CA) following manufacturer's guidelines. The three DNA samples were sequenced as part of a pool of 6 multiplexed libraries on two lanes of a HiSeq2000 v3 flowcell using version 3 of Illumina's multiplexed paired–end sequencing chemistry for 101 bp read length (Illumina, San Diego, CA). RNA library preparation was performed for each family member from 1.5μg of total RNA using Illumina TruSeq RNA Sample Prep Kit v2 according to manufacturer's instructions (Illumina, San Diego, CA). The three RNA samples were sequenced as part of a multiplexed pool of 4 samples on a single lane of a HiSeq2000 v3 flowcell using version 3 of Illumina's multiplexed paired–end sequencing chemistry for 101 bp read length (Illumina, San Diego, CA).

Binary base calls files were generated by the Illumina HiSeq2000 RTA module during sequencing and were converted to demultiplexed fastq files using CASAVA 1.8.2 (Illumina, San Diego, CA). Quality filtered reads from exome data were aligned to reference genome with BWA 0.6.2-r126 (H. Li and Durbin 2009). Binary alignment files were converted and coordinate sorted into the standard BAM format using SAMtools 0.1.18 (H. Li et al. 2009). Aligned reads were realigned around short insertion and deletions and duplicate reads were filtered using Picard 1.79 (picard). This followed aligned base quality recalibration with GATK 2.2 (McKenna et al. 2010). Flowcell lane level sample BAMs were then merged with Picard 1.79 if samples were sequenced across multiple lanes. Variant calling was done by UnifiedGenotyper and genotype quality recalibrated using VariantRecalibrator as described in the best practice methods of GATK 2.2 (DePristo et al. 2011).

Demultiplexed fastq files obtained from the RNA-seq experiment were aligned to human reference genome using ensembl.63.genes.gtf of annotated, known transcripts with TopHat2 (Kim et al. 2013). Aligned reads were assembled into transcripts with Cufflinks 2.0.2 using known transcript annotation in ensembl.63.genes.gtf as guide and we used annotated high abundance transcript annotation of ribosomal RNA and mitochondrial genes in an ensembl.63.genes.MASK.gtf. Post transcript assembly, Cufflinks was used to calculate the relative concentration of each annotated transcript by assigning an FPKM value (Fragments Per Kilobase of transcript per Million mapped reads) to each gene and transcript (Trapnell et al. 2010).

Calculation of physical coverage.

To determine the boundaries of the interstitial deletion on X, sequence read counts were obtained across X chromosome in a 100 bp sliding window for the mother and child using previously described methods (Craig et al. 2013). This script uses the SAMtools package to parse the exome BAM file for the patient and mother (H. Li et al. 2009). The algorithm uses a sliding window across the selected chromosome in 100 bp length, and for each read mapping within the window finds its mate pair and fills in the gap between the read pairs, then counts this gapped read as one read mapping within the window. This raw read count per 100bp window is then normalized by dividing the raw read count with the total reads mapping to the sum of sliding windows. Next, the normalized coverage in each window is transformed to log2 scale in both the mother and child and log2 transformed normalized read count is deducted from each other as described in Equation 1:

$$\log 2(\frac{\#\text{ reads mapping to 100bp window for case}}{\#\text{ reads mapping in all 100bp windows for case}}) - \log 2(\frac{\#\text{reads mapping to 100bp window for control}}{\#\text{reads mapping in all 100bp windows for control}})$$

Plotting log2 differences across chromosomes allows detection of large chromosomal deletions and amplifications, where a log2 difference of -1 means a heterozygous deletion in one of the copies.

Genotype phasing.

While any given SNP or indel could be potentially causative towards a disease phenotype, SNPs could also be used as markers for segregation analysis. In this study, we were interested in the parental origin (i.e. phase) of the deletion and the X inactivation skewing. We refer to the process of phasing as determining the parent-of-origin of a molecular variant (i.e., a heterozygote SNP or mRNA transcript containing a SNP), recognizing that phasing can have broader meanings. In our analyses, we use SNPs as markers to phase a genetic interval or region, where the interval could be a deletion, gene transcript, or chromosome. For example, if the patient is “A/T” for a SNP, the mother is “A/T” and the father is “A/A”, we can determine the “T” allele is from the mother. Larger events can also be phased by examining SNP genotypes contained within the larger event (i.e., a deletion); however, this requires that one recognize that SNP genotypes should be recoded to match their ploidy. For example, males containing a single X chromosome should be understood to be “A” and not “A/A”. Likewise, SNPs within a deletion should be understood to be “T”, rather than “T/T”.

Results

Estimation of XCI Ratio from Simulated Data.

We developed a simulation study for 11 datasets to estimate XCI pattern from paired, RNA-seq reads. For each dataset, 4996 loci provided read count information to estimate XCI and on average 1600 SNPs had a minimum read depth of 20. After phasing, the allelic ratios were fitted to the beta distribution and their parameters estimated. The distributions showed increased mono-allelic expression from 50:50 random to 100:0 completely skewed XCI (Figure 8). As expected, at 50:50 XCI ratio the maternal (Alt-M alleles) and paternal (Alt-P alleles) distributions almost completely overlap with their mean ratios at around 0.5 indicating bi-allelic expression and suggesting approximately equal expression of both chromosomes (Figure 8, 50:50). At each expected XCI ratio, the experimental, mean XCI ratios obtained from the beta distributions of the phased allelic ratios showed high concordance with expected XCI (Table 5). Although we compensated for read mapping bias by allowing 5 mismatches, our results show some deviation from the expected mean XCI in each dataset. Since our reads were generated against only known transcripts of 500bp or longer, some sequence homology between transcripts and the other regions of chromosome X may have resulted in read bias affecting allelic ratio estimates. As we shift expected allelic ratios from 50:50 random toward completely skewed 100:0, we observed an increased bimodality with the two phases separating into discrete distributions.

Table 5.

Estimation of XCI Ratio of *in silico* phased SNPs by beta testing.

Expected XCI ratio (%)	**Alt-M mean ratio (%)**	**SD**	**Alt-P mean ratio (%)**	**SD**	**Observed XCI Ratio (%)**
100:0	99.64	1.81	0.06	0.31	99.64 : 0.06
95:5	95.46	11.88	3.91	10.83	95.46 : 3.91
90:10	90.63	15.12	7.96	14.63	90.63 : 7.96
85:15	84.82	14.35	13.11	14.79	84.82 : 13.11
80:20	78.91	12.91	18.01	15.40	78.91 : 18.01
75:25	74.31	12.63	22.59	11.61	74.31 : 22.59
70:30	69.76	14.02	28.87	10.94	69.76 : 28.87
65:35	63.25	11.59	34.11	10.78	63.25 : 34.11
60:40	58.76	11.47	39.15	11.58	58.76 : 39.15
55:45	54.05	11.51	42.88	14.08	54.05 : 42.88
50:50	49.25	11.79	47.84	12.00	49.25 : 47.84

XCI = X inactivation, SD = standard deviation

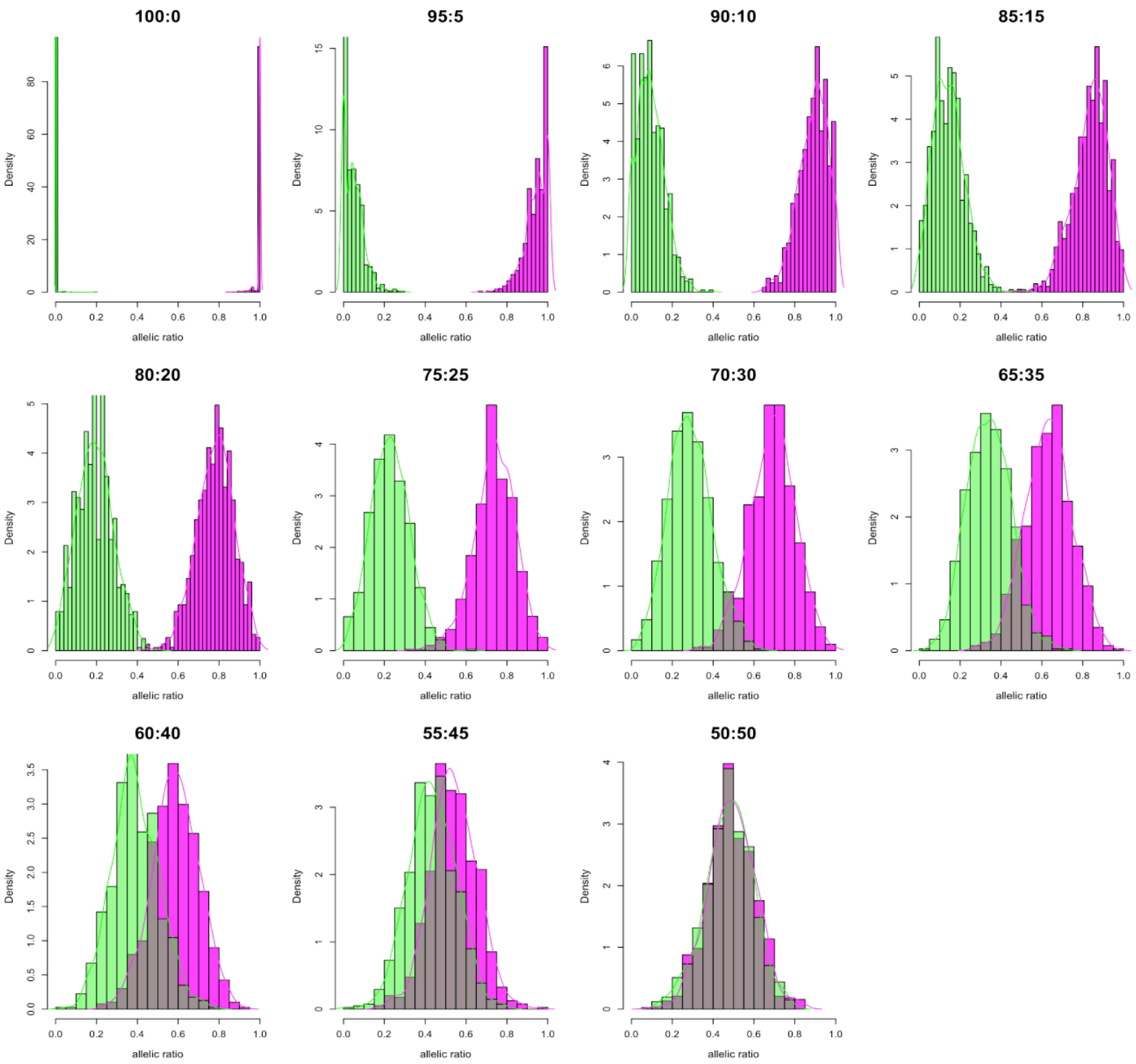

Figure 8. Phasing and distribution of *in silico* allelic ratios. Histograms of showing the allelic ratio distribution after each heterozygous SNP in the in silico data is assigned phase. Each heterozygous SNP allele was covered with at least 20 reads. Alt-M allelic ratios [magenta] and Alt-P allelic ratios [green] in bins of 20. Dark bars indicate SNP ratios that overlap between phased groups. Colored lines are the kernel density estimates of the phased allelic ratio distributions.

Coverage analysis indicated high correlation between expected and observed XCI ratios. Although Pearson's correlation was above 0.990 from coverage as low as 10X, correlation coefficient convergence with expected was achieved at > 0.999 above 20X suggesting that as low coverage RNA-seq experiments may be used for XCI ratio estimation (Figure 9). Unphased allelic ratio distribution followed a similar distribution pattern to phased dataset (Figure 10). Application of SP model to unphased allelic ratios resulted in consistent estimation of expected

XCI ratios (Table 6). The mean may be biased by the number of SNP markers available and other factors such as variants in genes that normally escape inactivation. However, our simulation shows that when relatively large number of markers is available, both beta distribution and SP model can consistently estimate the XCI ratio to the expected (Table 7).

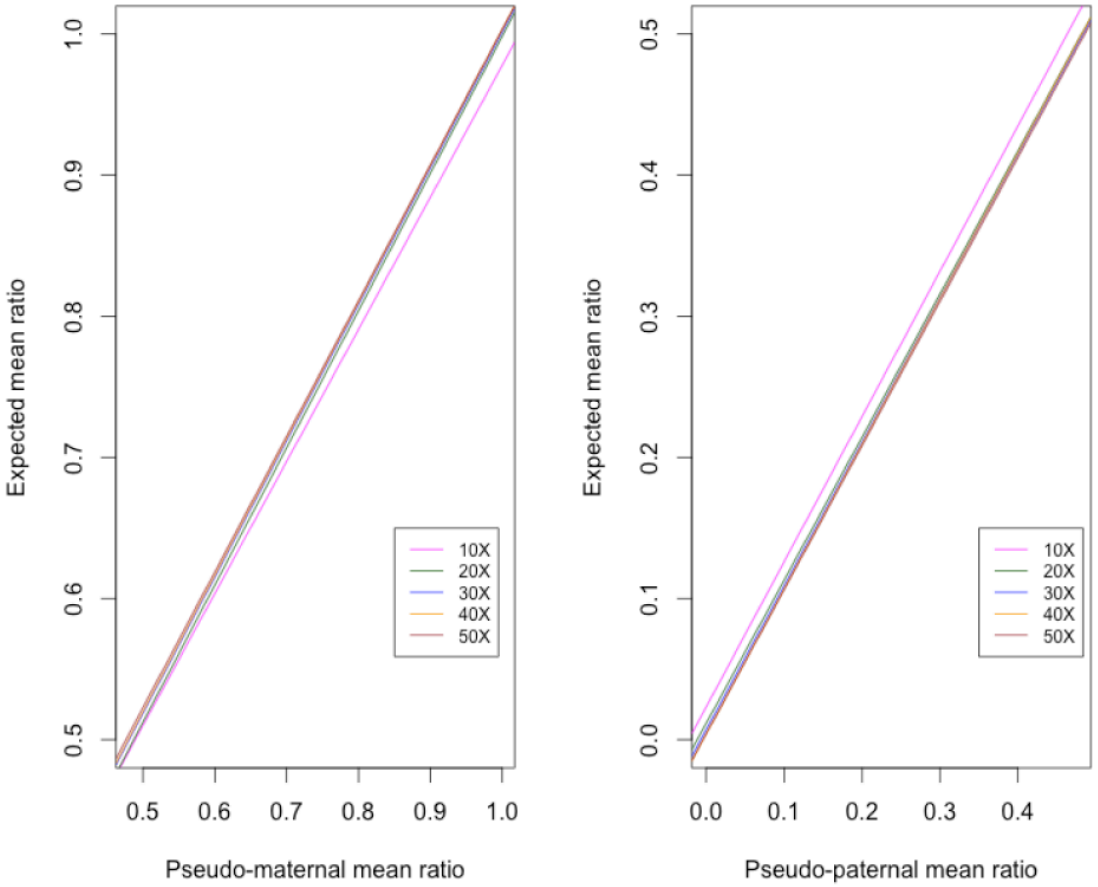

Figure 9. Correlation of expected and observed XCI ratios.
(A) The mean allelic ratio of the Alt-M alleles the *in silico* data to their corresponding expected allelic ratio. For example in 70:30 simulation, Alt maternal alleles have an observed mean allelic ratio of 69.0. (B) The mean allelic ratio of Alt-P alleles from each in silico dataset. Eg. in 70:30 simulation, Alt-P alleles have an observed allelic ratio of 27.6. Each color indicates the correlation of observed vs. expected ratios at minimum sequence coverage of 10X, 20X, 30X, 40X, and 50X. Pearson correlation coefficient was highest at r > 0.9998 above 20X read coverage.

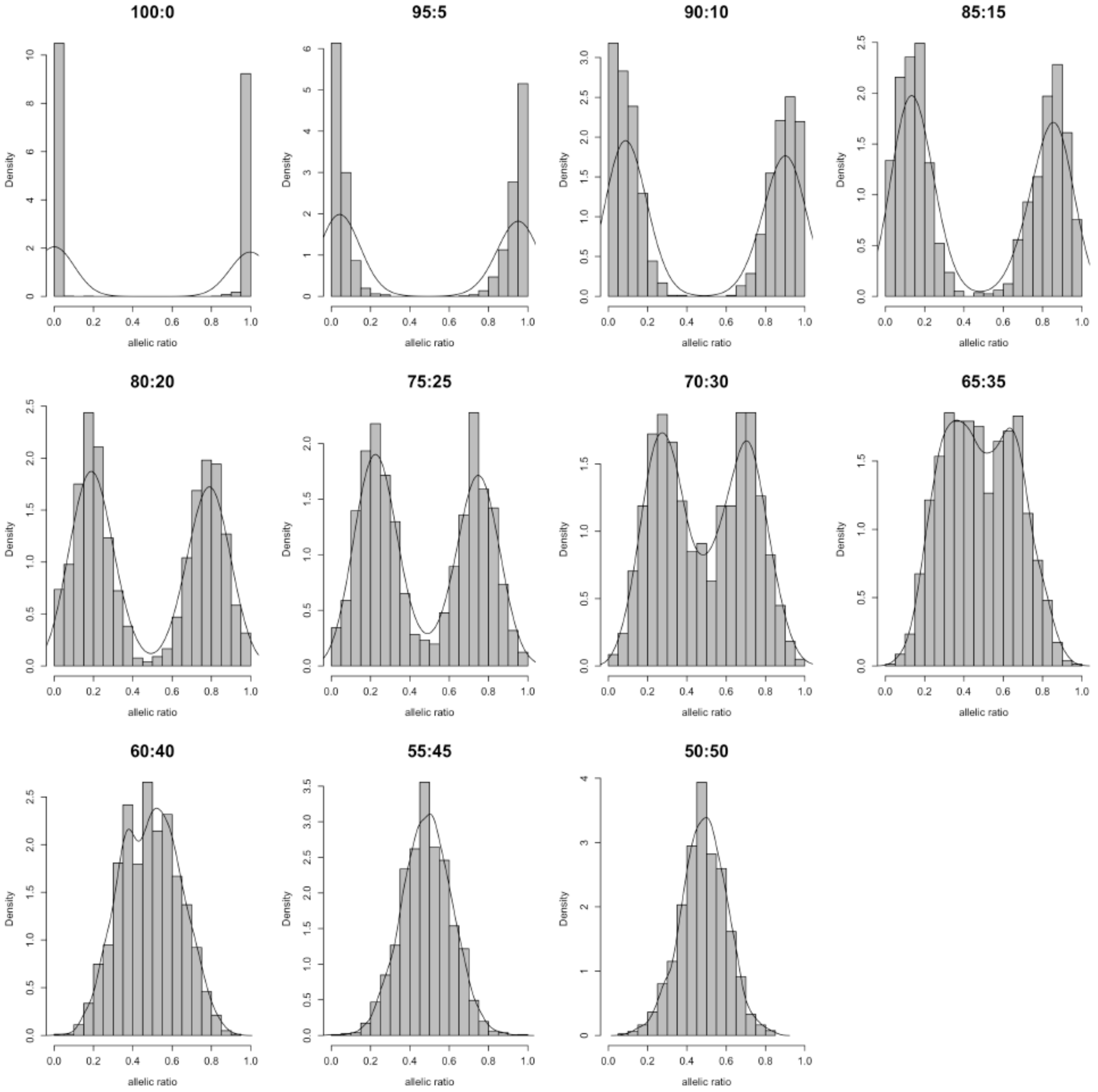

Figure 10. Un-phased allelic ratio distributions. Histograms showing the allelic ratio distribution after each heterozygous SNP in the in silico experiment when phase is not assigned. Each heterozygous SNP had to be covered with at least 20 reads. Black lines indicate the Gaussian kernel density of unphased allelic ratio distributions. Similar to phased experiments, the shift of distributions from unimodality in random XCI (50:50) toward bi-modality as XCI becomes more skewed towards 100:0 complete skewing.

Table 6.

Estimated XCI ratio of un-phased data by semi-parametric method.

Expected XCI ratio	Component 1 mean allelic ratio (%)	Standard Deviation	Component 2 mean allelic ratio (%)	Standard Deviation	Observed XCI ratio
100:0	99.6	2.2	0.0	1.1	99.6:0.0
95:5	94.8	6.2	5.2	6.2	94.8:5.2
90:10	88.9	9.1	9.7	8.5	88.9:9.7
85:15	84.7	10.2	13.7	9.4	84.7:13.7
80:20	78.9	11.5	18.7	11.0	78.9:18.7
75:25	74.3	12.0	24.2	12.0	74.3:24.2
70:30	69.0	12.8	27.6	12.4	69.0:27.6
65:35	64.7	12.5	32.7	12.5	64.7:32.7
60:40	58.0	13.9	37.0	13.6	58.0:37.0
55:45	51.0	14.8	49.7	14.9	51.0:49.7
50:50	48.6	14.0	47.8	14.0	48.6:47.8

Table 7.

Variant coverage and XCi ratio.

Expected XCI	total	≥10X	≥20X	≥30X	≥40X	≥50X
100:0	4878	3163	1606	723	288	119
95:5	4887	3203	1681	756	332	160
90:10	4882	3180	1590	694	308	136
85:15	4891	3168	1598	708	316	138
80:20	4894	3203	1591	693	323	131
75:25	4887	3176	1595	738	310	140
70:30	4875	3166	1627	712	293	137
65:35	4891	3165	1591	686	287	132
60:40	4878	3186	1623	727	313	143
55:45	4891	3239	1695	724	348	151
50:50	4879	3205	1597	703	312	126

Exome Analysis

WES resulted in an average of 139 million paired reads with average insert size of 249 base pairs [bp] corresponding to an average 14.8 gigabases (Gb) on the HiSeq2000 platform for the trio. After quality filtering, the 121 million average reads were aligned to reference with an 88% alignment rate. Approximately 97% of target regions had a mean base coverage of 10X (Table 8). Joint variant calling identified 85,708 single nucleotide variants (SNVs) and short indels in with 85.96% of calls in dbSNP135 (dbSNP). Functional evaluation of calls identified 42,192 (46%) missense, 344 non-sense (0.38%), and 48,373 (53%) silent variations. Transition/transversion ratio was 2.31 for all calls, and 2.447 for dbSNP variants. We applied various filtering approaches described elsewhere, but extensive search within Clinvar (Landrum et al. 2014), The Human Gene Mutation Database (HGMD) (Stenson et al. 2003), and OMIM (OMIM) did not identify any unambiguous genetic variants that likely caused or contributed to the child's phenotype (Gilissen et al. 2012).

Table 8.

Summary metrics of Exome sequencing.

	Mappable Paired Reads (M)	Mappable Unique Paired Reads (M)	Paired Reads Mapped (M)	Mapped Bases (Gb)	On/Near Target Mapped Bases (Gb)	Mean Coverage Captured Regions (X)	Target Regions Coverage >10X (%)	Fold Enrichment
Child	138.58	121.30	107.22	10.79	8.13	85.71	98.05	26.17
Mother	145.77	128.17	113.15	11.39	8.54	88.79	97.73	25.71
Father	133.88	115.60	101.66	10.22	7.78	84.25	97.66	26.6
Average	139.41	121.69	107.35	10.80	8.15	86.25	97.81	26.16

M = million, Gb = Gigabases, X= number of times locus was sequenced

Characterization and Phasing of Xp22.31 Deletion

Absence of candidate rare variants focused our attention to the previously identified interstitial deletion on Xp22.31. We compared log2 normalized physical coverage of the daughter's exome to the log2 normalized coverage of the mother's (see Materials and Methods), and observed those regions where the ratio fell below the threshold coverage of -1. Comparative analysis identified the deletion as heterozygous at Xp22.31 with breakpoints at 6,451,600 and 8,095,100, respectively (Figure 11). Similar comparison to the father's exome indicated that father was hemizygous for this region; therefore the deletion occurred *de novo*. The distal breakpoint is approximately 50bp upstream of *VCX3A* and the proximal breakpoint resides within the first 100bp of miR-651, a microRNA gene with no known biological function. The deletion encompasses 1,643,501bp harboring five genes and two microRNA genes (Table 9). This region was in concordance with the aCGH. The deletion was phased to Xp based on rs5933863, at X:7,270,694 G>A in the 3' un-translated UTR region of the *STS* gene (NM_000351)**.** The affected child's genotype was homozygous G/G, the mother's was heterozygous G/A, and the father's was homozygous alternative A/A. Recoding based on anticipated ploidy, the child's genotype is "G", the mother remains "G/A", and the father with a single X chromosome is recoded "A". Principles of X-linked inheritance dictate that the child must have a heterozygous genotype G/A at this position. Since she is missing the paternal allele A and has an apparent genotype of "G", there is evidence that the region containing this SNP on Xp was deleted resulting in an out-of-phase genotype (Table 10). This out-of-phase coding SNP was validated by Sanger method in the trio (Figure 12).

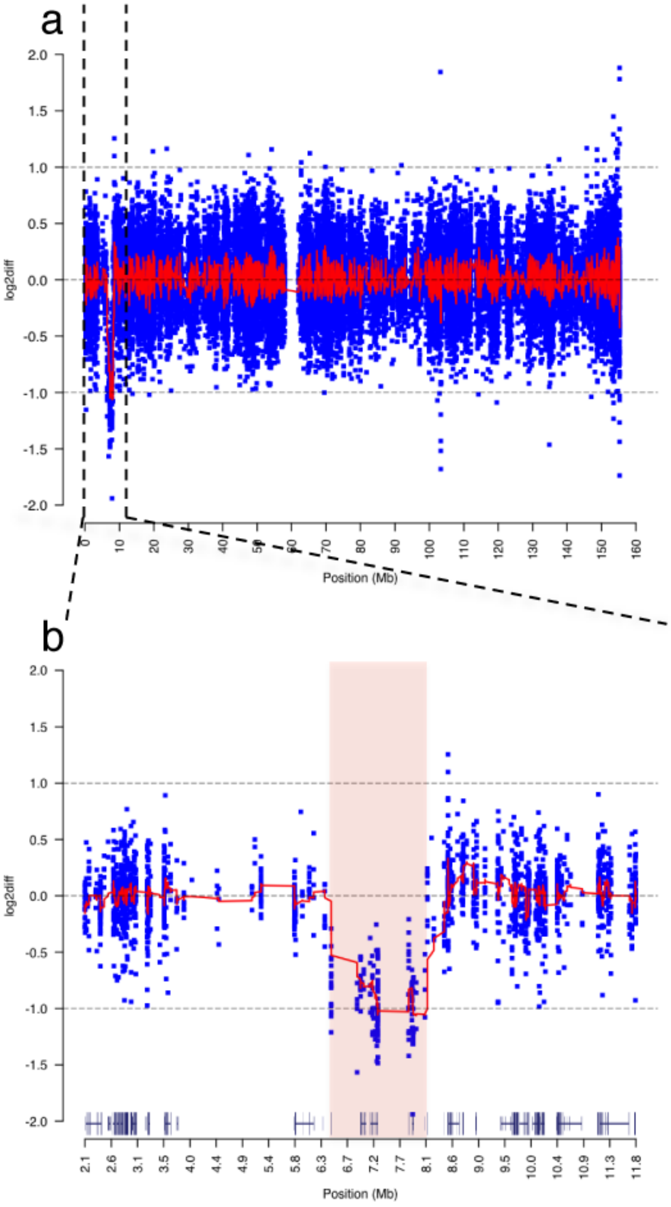

Figure 11. De Novo Deletion on Xp22.31.

(a) Chromosomal view of log2 coverage difference between affected child and mother obtained by WES. The log2 difference of normalized read coverage between affected child and mother is shown on the y axis, with each blue dot indicating log2 difference in normalized sequence coverage in a 100bp window. The red line across the chromosome is the mean log2 differences across a sliding window of 25. A large deletion on chromosome X is recognizable in the child indicated by drop in log2 difference to -1 between 0-10Mbase. (b) Zoomed in view of reduced sequence read coverage between 6.4 - 8.1Mbase of the short arm of the chromosome. The pink shaded area indicates the deletion breakpoints predicted by aCGH analysis that overlaps with deletion seen by the exome coverage analysis. Gene tracks above the x-axis was obtained from UCSC Genome Browser and contains the deleted genes *VCX3A*, *HDHD1*, *STS*, *VCX*, *PNPLA4* genes and *MI4767* microRNA genes.

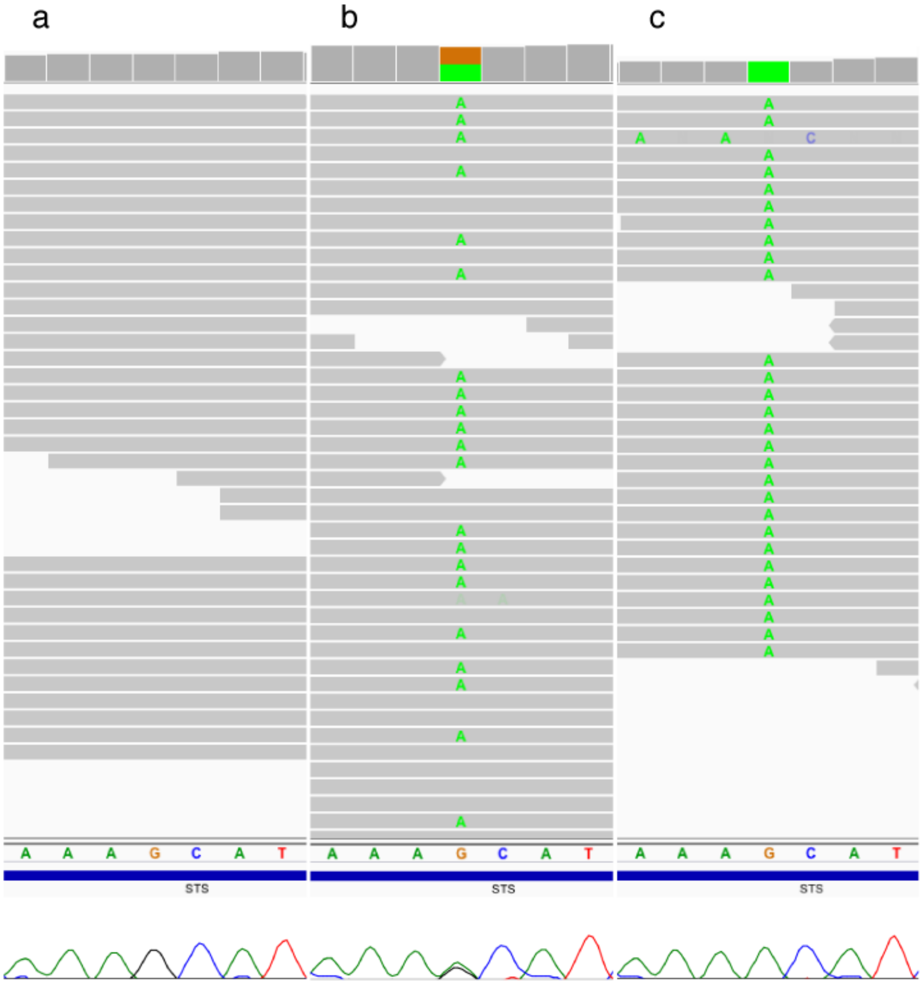

Figure 12. Determining phase of rs5933863.

Next-generation sequencing traces visualized using the Integrated Genomic Viewer (IGV) and below them the corresponding Sanger traces of rs5933863 G>A alleles in the *STS* gene that helped determine phase and origin of the 1.7Mb deletion on chromosome X (J. T. Robinson et al. 2011). Patient's IGV and Sanger traces (a) indicate that she is either homozygous G/G or hemizygous "G" genotype at this position. The mother's (b) and the father's (c) traces indicate that they are "G/A" and "A" genotype, respectively.

Table 9.

Genes within 6,4-8,1 Mb interstitial deletion.

Gene	Gene Name	Start	End	Strand	RefSeq ID	OMIM	Phenotype
VCX3A	Variably charged, X-linked 3A	6,451,659	6,453,159	-	NM_016379	300533	XLI/MR
MIR4767	microRNA 4767	7,065,901	7,065,978	+	NR_039924		
HDHD1	Haloacid dehalogenase-like hydrolase domain	6,966,961	7,066,231	-	NM_001135565	306480	
STS	Steroid sulfatase, isozyme S	7,137,472	7,272,682	+	NM_000351	300747	XLI
VCX	Variably charged, X-linked	7,810,303	7,812,184	+	NM_013452	300229	
PNPLA4	Patatin-like phospholipase domain	7,866,804	7,895,475	-	NM_004650	300102	
MIR651	microRNA 651	8,095,006	8,095,102	+	NR_030380		

XLI=X-linked ichtyosis, MR=mental retardation

Table 10.

Genotype phase of X-linked SNPs within the 6,4-8,1 Mb interstitial deletion

Chr:pos	ref	alt	rsID	Proband				Mother				Father			
				gt	ploidy	depth	depth	gt	ploi	depth	depth	gt	ploid	depth	depth
X;727069	G	A	rs5933863	G/G	1n	33	0	G/A	2n	46	43	A/A	1n	0	28
X:727099	A	A	rs1131289	G/G	1n	0	79	G/G	2n	2	176	G/G	1n	0	75
X:727222	G	A	rs13648	A/A	1n	0	1	nc	-	-	-	A/A	1n	0	1
X:786737	T	C	rs3470971	C/C	1n	0	96	C/C	2n	0	141	C/C	1n	0	56
X:786743	G	A	rs1200998	A/A	1n	0	72	A/A	2n	0	122	A/A	1n	0	57
X:786766	G	C	rs7739847	C/C	1n	1	2	G/C	2n	2	4	C/C	1n	0	1
X:786773	A	G	rs6639976	G/G	1n	0	2	G/G	2n	0	1	nc	-	-	-

chr=chromosome, pos=position, ref=reference allele, alt=alternative allele, rsID=dbSNP137 id, gt=genotype, 1n=haploid, 2n=diploid, nc=no call by lack of coverage

Estimation of XCI Ratio from RNA-seq experiment.

Sequencing the patient's mRNA resulted in an average of 116 million paired reads per sample mapping to human reference genome (Table 11). From the exome variant call set 1,729 single nucleotide variants including indels mapped to chromosome X, of which 901 were called heterozygous in the affected child. 374 calls were heterozygous SNPs within transcripts, and 325 were X-linked, outside PAR1 and PAR2 regions (Mangs and Morris 2007). 226 variants were high quality with score PASS by GATK Variant Recalibration. Next we selected variants that were previously documented in dbSNP build 135. A total of 83 SNPs were covered with at least 20 reads. 37 phased to Xm and 44 to Xp, and two Mendelian errors. The 37 Xm alleles were from 23 genes, with 19 genes with a single heterozygous expressed variant and four had more than two heterozygous expressed variants. The 44 Xp alleles were from 31 genes, and 22 of them had a single heterozygous variant expressed and 9 had more than one heterozygous variant. The allele ratio distribution indicated bimodal distribution showing lower expression of paternally inherited heterozygous SNPs (Figure 13). The XCI ratio estimated from phased alleles was 82.7:20.3 (approximately 83:20), and from the unphased allelic data was 82.2:19.2 (approximately 82:19), consistent with moderately skewed X inactivation with a ratio of 85:15 obtained by the HUMARA methylation assay. The integration of phase information had minimal affect to final estimate indicating the power of the SP model**.** In addition to the patient, we estimated XCI ratio in 4 additional female individuals from our clinical sequencing center (Figure 14). In each case XCI was estimated by our RNA-seq approach and the HUMARA assay. A single case was uninformative for the HUMARA, caused by homozygosity at the methylation sensitive repeat sequence of the *AR* locus (Figure 14, S34). In 3 out of the 5 cases (60%), the HUMARA method suggested moderately skewed XCI ratio (>80:20) (Figure 14, S14, S18, S23). However, expression analysis supported strong correlation between the three methods only in the clinical case of this report where skewed XCI was estimated by all three methods (Figure 14, S18). In three of the remaining four cases skewed XCI was not supported by the RNA-seq analysis (Figure 14, S14, S23, S34). In a single case all three methods predicted random XCI ratio (Figure 14, S11). In general there is a high concordance between the three approaches with the beta and the SP methods have the highest concordance (Pearson's r = 0.99), but these approaches have weaker correlation with HUMARA (SP Pearson's r=0.84, beta Pearson's r= 0.80). In general, we see a lower XCI ratio estimated by allele expression analysis than by HUMARA. Estimates of XCI ratio may be biased by reference bias in read mapping, insufficient coverage at heterozygous loci, and by heterogeneous gene expression driven by DNA methylation and cis-acting regulatory mechanisms.

Table 11.

Summary metrics of RNA-seq

	Quality Reads Mapped (M)	Reads Mapped in Pairs (M)	Reads Mapped in pairs (%)	Mappable Bases (Gb)	Mapped Bases (Gb)	FPKM >1.0 #genes/total annotated on X	Median Insert Size
Child	95.44	84.13	88.15	8.047	8.046	346 /2688	154
Mother	154.90	135.39	87.41	13.538	13.537	374/2688	156
Father	99.18	83.11	83.8	8.894	8.893	362/2688	154
Average	116.51	100.88	86.58	10.161	10.159	361/2688	155

M = Megabases, Gb = Gigabases, X= number of times locus was sequenced

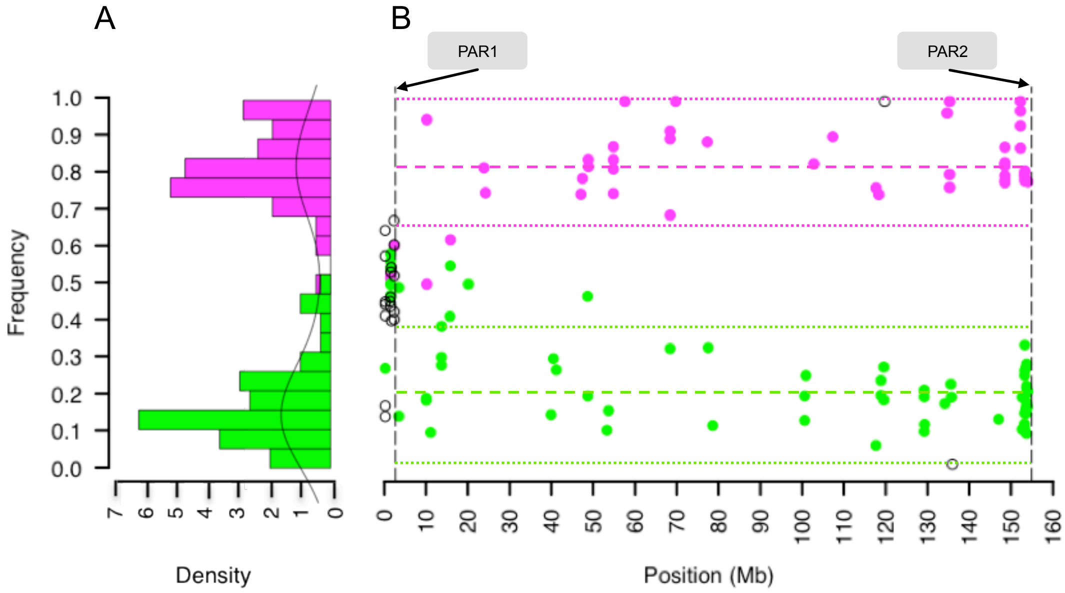

Figure 13. Phased allelic expression on chromosome X. (A) Allelic ratio of heterozygous SNPs show bimodal distribution of the expressed maternal (magenta dots, n=37) and paternal (green dots, n=44) alleles indicated biased expression of the inherited chromosomes. (B) Chromosome-wide allele frequency of the phased alleles from RNA-seq indicate that overall, maternal X has a preferential expression in the patient with mean ratio across X of 0.82.7±0.083 (dashed magenta line), compared to paternal alleles of 0.20.3±0.095 (green dashed line). Biased expression in favor of the maternally inherited alleles is preserved across the entire length of the chromosome. However, alleles within genes that potentially escape X inactivation can show bi-allelic expression as defined by an allelic ratio 2SD outside the mean of the phased allele ratios (colored, dotted lines). Essentially all high quality heterozygous SNPs with a minimum of 20X coverage could be phased based on transmission of alleles within the X-linked region. SNPs where transmission of alleles could not be determined (clear circle) lie predominantly in the pseudo-autosomal region (PAR1) except two Mendelian errors.

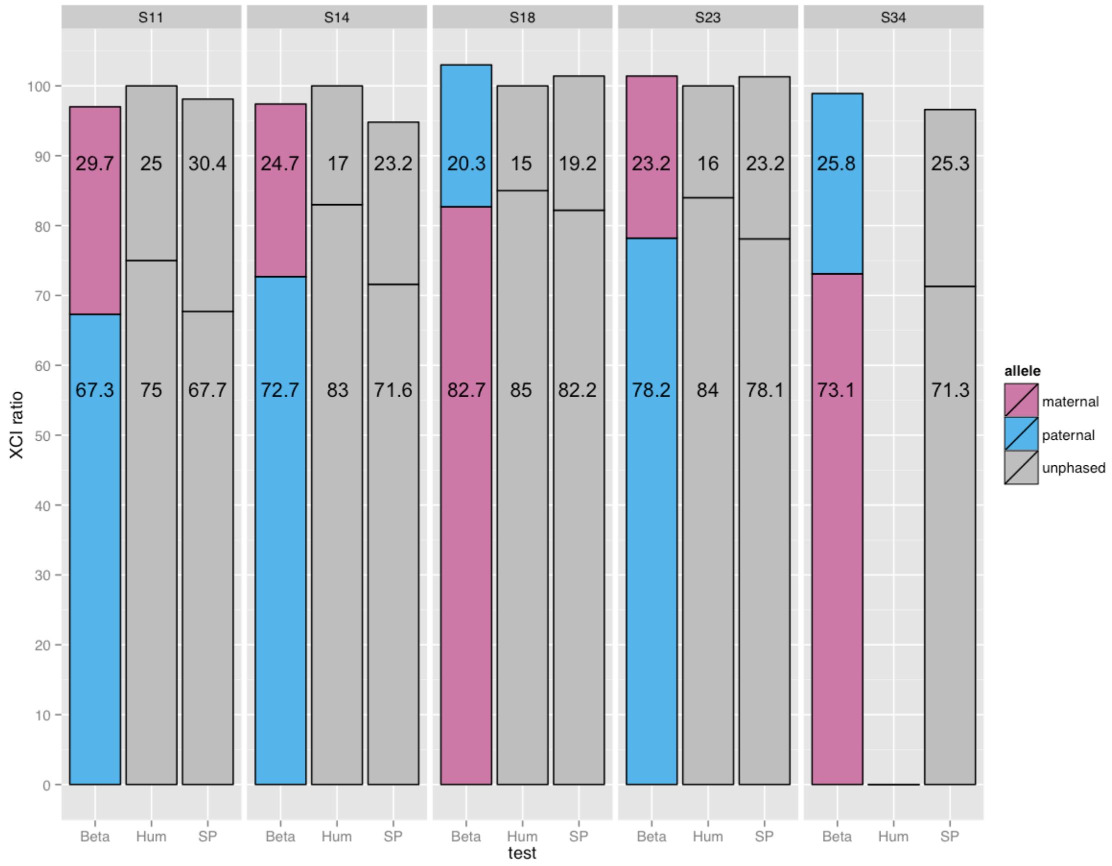

Figure 14. Estimation of XCI ratio in 5 patients. XCI estimated in five female patients. The x-axis indicates the approach (Beta= beta distribution of phased allelic expression, Hum= HUMARA DNA methylation assay, SP= semi-parametric method of unphased allelic expression). The y-axis indicates the XCI ratio (eg. S11 XCI ratio by Hum = 75:25). XCI ratio estimated by fitting allele ratios to the beta distribution can provide information about parental bias in XCI ratio as in the patient (S18) has 82.7:20.3 biased XCI that favors the expression of Xm (magenta). The ratio of allele expression from the maternal chromosome to the allele expression from the paternal chromosome (blue) gives the XCI ratio. In S18, using the beta model, we were able to determine that moderately skewed XCI ratio favored the expression of Xm compared to Xp. We had no phase information on the *AR* locus for the HUMARA assay, thus phase of XCI could not be determined. Homozygosity at the *AR* locus, in S34 shows uninformative HUMARA test, underlying the utility of RNA-seq in XCI estimation. The SP method does not consider allele phase to estimate the parameters of allele distributions, so phase of XCI could not be determined. RNA-seq estimates random XCI (<80:20) in S14 and S23 compared to moderately skewed XCI (>80:20) by HUMAR. S18 and S11 show complete concordance between the three methods. There is no clear trend that would indicate a higher likelihood of biased inactivation of either parental chromosome.

Identification of Genes that Escape X inactivation

Phased, allele-specific expression analysis highlighted a number of variants in genes that may escape inactivation. Escape of X inactivation results in bi-allelic expression of genes from Xa and Xi in the same cell and can contribute to phenotypic variability in females who are carriers of X-linked disease (Carrel and Willard 2005). Therefore a catalogue of escape genes in clinical evaluation may contribute to the better understanding of clinical symptoms and may offer treatment options. We identified escape genes in the patient by examining 325 heterozygous loci across X and the deviation of their allelic ratio from the mean allelic ratio of each phased distribution. We defined a candidate escape gene by having a heterozygous SNP with an allelic ratio two standard deviations (2SD) outside the mean allelic ratio of the chromosome-wide allelic distribution and showing bi-allelic expression. Bi-allelic expression was defined as allelic ratio between 0.1 and 0.9. Therefore if a paternally inherited variant had an allelic ratio of 0.49 and the mean allele ratio of the chromosome-wide paternal alleles was 0.203 with a standard deviation of 0.09, that variant allele ratio was greater then 2SD from the mean, thus was bi-allelic expressed. Of the 325 X-linked heterozygous alleles 15 showed bi-allelic expression in 12 genes, but 7 variants were considered false positive owing to low read coverage (<7X)(Table 12) (Y. Zhang et al. 2013). Comparison of the sufficiently covered variant loci to chromosome wide XCI screens in hybrid cell lines and fibroblast indicated that in 4 of the 6 escape genes, XCI status was consistent with previous assignments of genes as escaping from XCI using both hybrid cell line and fibroblast data. Protein Convertase 1 Inhibitor *(PCSK1N)* and Plexin A3 *(PLXNA3*) both suggest escape status in the patient, and were previously reported as subject of XCI (Carrel and Willard 2005; Y. Zhang et al. 2013). *PCSK1N* and its associated propeptide may have a role in body weight and behavior in mice, and Plexin A3 is a co-receptor of the axon guidance receptor, Neurophilin-2 *(NRP2)* but their dosage affect owing to XCI remain to be elucidated (Morgan et al. 2010). The distribution of genes that are shown to escape XCI was consistent with the regions that contain the highest density of escape genes, and were mostly located on the short arm of chromosome X (Disteche 1999).

Table 12.

Escape of XCI.

position	dbSNP	variant phase	allelic ratio	read depth	Gene ID	Carrel et al.
X:3,524,309	rs6567569	paternal	0.49	55	*PRKX*	Escape
X:10,203,342	rs41305355	maternal	0.38	8	*CLCN4*	Heterogeneous
X:10,204,267	rs4830442	maternal	0.50	12	*CLCN4*	Heterogeneous
X:15,339,588	rs148660178	maternal	0.50	2	*PIGA*	Subject
X:15,801,330	rs12841514	paternal	0.55	20	*CA5B*	Escape
X:15,801,643	rs28707735	paternal	0.56	9	*CA5B*	Escape
X:15,802,800	rs5980189	paternal	0.50	4	*CA4B*	Escape
X:20,143,370	rs13179	paternal	0.50	10	*EIF1AX*	Escape
X:41,374,523	rs5918192	paternal	0.60	5	*CASK*	Subject
X;46,358,046	rs148701104	paternal	0.50	2	*ZNF673*	-
X:48,690,749	rs11538178	paternal	0.47	15	*PCSK1N*	Subject
X:100,881,434	rs6995	paternal	0.50	4	*ARMCX3*	Subject
X:132,438,872	rs1129980	paternal	0.50	2	*GPC4*	Heterogeneous
X:153,694,334	rs5945430	paternal	0.50	8	*PLXNA3*	Subject
X:153,759,858	rs1050757	paternal	0.67	3	*G6PD*	Subject

Discussion

In this study we applied integrated WES and RNA-seq to simultaneously evaluate the functional effect of coding variations in the process of clinical diagnosis. Although previous clinical testing suggested a mechanism for the patient's disease, with the combined analysis of the trio exome and the patient's RNA expression that we are now able to hypothesize a mechanism for the observed phenotype. Variant filtration approaches after trio WES did not result in the identification of strong candidate causal variations. Although there was suggestive evidence from the aCGH that the disease pathology may be related to a heterozygous deletion on Xp22.31, it was only with incorporation of SNP phasing and comparative analysis of sequenced reads that we were able to determine that the deletion occurred *de novo*. Genes associated with neurological dysfunction including a number of variable-charge X-linked genes lie within the deletion (*VCX*, *VCX3A)* (Jiao et al. 2009). Although we were not able to detect lymphocyte expression of any of the *VCX* genes, there is suggestive evidence these genes have roles in cognitive function. *VCX3A* overexpression in rat hippocampal neurons increase neurite outgrowth that may positively influence synaptic plasticity (Jiao et al. 2009). Furthermore, some males who are hemizygous for a recurrent Xp22.31 deletion and have X-linked ichtyosis (OMIM 308100) also demonstrate mental retardation (Van Esch 2005). This region appears to be a hotspot for copy number changes, complex duplications, and triplications, suggesting that the instability of this region may contribute to disease risk (P. Liu et al. 2011). The inherent limitation of our approach is that our resolution to define the exact genomic content of the deletion is reduced by exome sequencing and can only be circumvented with whole-genome sequencing approaches.

Phased and unphased allele-specific expression in the patient was concordant with the HUMARA assay and indicated moderately skewed XCI. The contribution of skewed XCI to her condition is not clear, although the phased XCI ratio allows us to develop a hypothesis for the molecular mechanism that underlies her condition. One could hypothesize that random XCI in the patient and potential dominant negative affect of the deletion would result in a severe neurological condition. However, females who are carriers for deleterious chromosomal mutations may not present clinical symptoms owing to selective advantage and preferential expression of the normal X (Plenge et al. 2002; Desai et al. 2011). These females are usually heterozygous for an X-linked deleterious allele and have skewed XCI. The patient has skewed XCI and is heterozygous for the deletion but showing some mild neurological condition, suggesting that the preferential expression of the cytogenetically normal X may be compensating for the deleterious affect of the deletion. While insufficient cases have been reported to provide statistical significance, females who were diagnosed with Xp22.31 microduplication and preferentially silenced the X with the microduplication had normal phenotype while those who preferentially express the X with the microduplication had intellectual disability (F. Li et al. 2010). It is plausible that loss of a chromosome copy at Xp22.31 has different clinical manifestation than copy gain. Therefore the contribution of Xp22.31 rearrangements to neurological dysfunction need further study. For the patient sequenced in this study, our data are consistent with a model that the preferential expression of the cytogenetically normal, maternal X may have contributed to her mild cognitive phenotype.

Our ability to uncover molecular mechanisms by DNA and RNA-seq in patient's surrogate tissue (peripheral blood) that may correlate with phenotype in the central nervous system argues for potential benefit in clinical diagnostic cases that remain unresolved. This is supported by a number of studies that find a strong correlation in gene expression profile in blood with affected status in such diseases as Parkinson's Disease and Huntington's Disease (Scherzer et al. 2007; Borovecki et al. 2005). Previous studies evaluating the methylation status of X-linked genes and overall XCI patterns across various tissues show that XCI is concordant between tissues, including blood and brain (Bittel et al. 2008; Cotton et al. 2011). However, these studies were performed in females with no known neurological condition and showed that variable XCI status exists in about 12% of X-linked genes and variance between tissues increases with age. Studies in Rett syndrome and XCI in mice show some evidence that deleterious alleles lead to preferential silencing of the mutant X in brain tissue, but their correlation with blood has not been well characterized. (J. I. Young and Zoghbi 2004). In females with Rett Syndrome there is evidence that skewed XCI correlates with disease, however correlation between blood and brain XCI pattern was low in a small sample set. Therefore the use of whole blood to predict XCI patterns in the brain and their correlation to disease susceptibility remains to be elucidated.

Our simulation proposed an approach to estimate XCI ratio using chromosome-wide SNP expression and found that phased and unphased SNPs can equally estimate the ratio with both beta and SP model. Even if research and clinical sequencing application will be limited in sequence coverage, our method is able to predict XCI at high concordance with expected as low as 10X coverage. Our method also allowed for base error rate therefore providing a more realistic sequence data. Our approach based on read count, and relative ratio estimation of variant alleles, can be applied to other sequencing platforms and to other expressed regions of the genome that are targeted by RNA-seq. Principles of skewed expression demonstrated in this study could be relevant to imprinted portions of autosomes and therefore applicable to disorders like Prader-Willi and Angelman syndromes (Biliya and Bulla 2010). Skewed expression of autosomal heterozygous alleles can be markers for imprinted regions, and may uncover cis-regulatory elements.

Although, in our small dataset, XCI estimation from RNA-seq analysis was not fully concordant with the methylation assay, direct measurement of allele expression may provide a better estimate of the true cellular activity of each inherited chromosome copies. HUMARA assay targets a single genomic locus and relies on the methylation of a repeat sequence targeted by methylation sensitive restriction enzyme. Deletions, copy number changes, homozygosity at the *AR* locus, enzymatic and PCR inefficiency, hypo-methylation of restriction enzyme target, difficulties associated with data interpretation, and the challenges associated with the amplification of repeat regions may influence assay results (Swierczek et al. 2012).

Our approach is dependent on the accuracy and sensitivity of multiple SNP markers expressed in the X-linked region. There is heterogeneity in the regulation of X-linked gene expression by epigenetic mechanisms, therefore, sampling alleles from multiple genes with various expression levels to infer XCI ratio may be inconsistent with previous methods but excluding alleles from genes that escape XCI can provide an inaccurate picture of the X chromosome activity, and molecular characteristics of the tissue source (Carrel and Willard 2005). Therefore, we did not filter out alleles from genes that were previously reported to escape XCI. This may have contributed to an overall lower XCI ratio estimates by RNA-seq compared to the methylation assay. In addition, methylation based assessment of XCI may not be concordant with expression based methods owing to differences in assays and applied analytical methods. Challenges in RNA-seq experiments include technical and analytical variability that may affect XCI ratio therefore transcription-based validation assays may be useful to improve our approach (Carrel and Willard 2005; Moreira de Mello et al. 2010; Swierczek et al. 2008). The use of direct expression analysis of multiple SNP markers may also increase our power to accurately estimate XCI, providing a basis to improve our definition of clinically significant XCI ratio boundaries. However a more systematic screening of XCI by RNA-seq across a series of X-linked disorders in females may greatly enhance our understanding of the underlying cause of phenotypic variability.

WES identified a deleterious deletion on Xp22.31 that is in a hotspot for chromosomal rearrangements and associated with a number of neurological conditions. In addition, using allele-specific expression analysis from RNA-seq we were able to define XCI ratio in simulated and experimental data. Although the number of individuals reported, and the number of heterozygous alleles in the X-linked region may be small, both the SP and beta models could reliably estimate XCI from RNA-seq data. The benefit of the SP model is that parental sequencing and genotype phasing is not necessary to estimate XCI, it compares well to XCI based on allele phasing, and can be applied to individuals only. The combined genomic and functional data allowed formulating hypothesis for the molecular mechanism for the patient’s symptoms, which can provide a basis for further clinical studies and patient management. However, extensive functional analysis is required to assess if our hypothesis based on sequencing blood RNA can be applied to a neurological condition. Finally, our study also represents an application of high-throughput sequencing methods and their simultaneous utilization to study epigenetic mechanisms in the clinical settings and how they contribute to genetic basis of a heterogeneous disease. Rapid decrease in sequencing costs, improved analytical methods, comprehensive, integrative sequencing approaches will likely be used more in the future and may replace traditional methods that may be uninformative owing to atypical disease phenotype, low-throughput, high costs and invasiveness.

In conclusion, we showed the utility of combined analysis of genomic and functional variations on a chromosomal scale to determine XCI ratio. Application of this method showed concordance with currently available clinical test thus provides a sensible alternative in studies that apply next-generation sequencing to study complex, hard-to-diagnose phenotypes. In addition, we showed that the use of integrated approach can provide insight into the underlying molecular process potentially correlating with her symptoms.

CHAPTER 4

INTEGRATED ANALYSIS OF DNA AND RNA SEQUENCE DATA IN RARE CHILDHOOD DISORDERS BY MULTIVARIATE OUTLIER ANALYSIS OF RARE FUNCTIONAL VARIANTS

Introduction

In this chapter we set out to develop a novel framework to study the functional impact of germline DNA variants derived from patient specific tissue, to improve standard variant prioritization methods by integrated DNA and RNA sequencing. Identifying the genetic basis of disease in rare childhood disorders is often hampered by discerning which variants from a list of a few dozen to a few hundred variants are functional, and thus the focus of this chapter is to describe an approach for prioritizing those variants with a functional impact on transcription. A germline variant can have a variety of functional effects on transcription, including but not limited to exposing cryptic splicing sites causing in-frame exon skipping, causing premature truncation of transcription or altering promoter binding.

RNA-seq is a high-throughput approach that provides qualitative and quantitative information on the impact of functional variants by sequencing the transcriptome or transcribed RNA species including mRNA, long non-coding RNAs (lincRNA), small-RNAs in a tissue of interest relating to specific condition (Z. Wang, Gerstein, and Snyder 2009). This approach works essentially like DNA sequencing except the millions of sequenced mRNA fragments are mapped to a known transcript structure of the genome, or assembled without a reference transcript map to detect novel transcripts (Ozsolak and Milos 2010). The most commonly investigated properties of the transcriptome are alternative mRNA transcription and processing (de Klerk and 't Hoen 2015). Choice of promoter, exon splicing, alternative poly-adenylation directly impact the mRNA composition of the cell and can result in cellular heterogeneity affecting clinical phenotypes (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014; H. Zhang, Lee, and Tian 2005; Florea, Song, and Salzberg 2013).

Current RNA-seq analytical approaches provide multitude of information on the functional portion of the genome including but not limited to gene and isoform expression, differential gene and isoform expression, allele-specific expression, and alternative splicing and exon usage.

One common RNA-seq analysis approach is the estimation of transcriptome abundance of the sequenced sample, either alone, within groups, or by comparison to others. The number of reads mapping to transcript relates to its expression but within-sample, across-sample biological and technical variability during RNA-seq can influence estimates and only provide a relative expression level. Transcript abundance is impacted by the presence of isoforms, recognizing that we are measuring fragments that are typically much smaller than the overall transcript (in the case of Illumina next generation sequencing). A single gene may have many different types of transcripts, or here referred to as isoforms, representing possibly alternatively spliced variants of an mRNA species with different composition of exons. Due to ambiguity in read mapping, reads may map to multiple isoforms. There are numerous statistical approaches to resolve this uncertainty based on known exon structure and the quality of mapping reads (H. Jiang and Wong 2009; Trapnell et al. 2010). Currently the most common measures to quantify the expression of a transcript are transcript per million (TPM) and fragments per kilobase of transcript per million reads mapped (FPKM) (B. Li and Dewey 2011; Trapnell et al. 2010).

Finding genes harboring functional variants that are differentially expressed between two or more conditions has become a routine experimental design to study phenotypic variability in human disease. This approach is based on estimating the change in read counts in expressed transcripts between conditions followed by statistical testing if the change is greater than what would be expected just due to random variation. The final result of a differential expression study provides an estimated magnitude of change in expression (fold change) and its significance (p-value) (Rapaport et al. 2013). Differential expression analysis can provide a list of genes that are associated with given predictors (i.e. affected, unaffected) or responses (i.e. treated, untreated). There is extensive literature on experimental designs and analytical strategies for differential gene expression studies using RNA-seq which is beyond the scope of this study (Rapaport et al. 2013; Finotello and Di Camillo 2015; Oshlack, Robinson, and Young 2010).

Another analysis approach for RNA-seq is to measure allele specific expression (ASE), which was a core concept behind our prior chapter on X-inactivation. While in that chapter we were looking at a chromosome-level impact, ASE is more narrowly defined as the unequal expression of two copies of the same gene and this imbalance in expression can be important in phenotypic variability in human disease. ASE in extreme cases can result in monoallelic expression of only one copy of a gene while the other is silenced, by significantly biased expression of the two alleles. ASE may also result in allele specific transcript expression presenting biased expression a transcript with one allele over the other. This pattern of expression may be influenced by epigenetic gene regulation, XCI, or parental imprinting (Fang et al. 2012; Moreira de Mello et al. 2010). ASE studies utilize heterozygous loci across the transcriptome and quantify the relative proportion of the mRNA expression between the two alleles (Main et al. 2009). Mapping bias present in RNA-seq against alternative alleles present challenges for accurate estimation of ASE and can be tackled by the creation of reference genomes that contain both reference and alternative alleles or by adjusting the mapping algorithms for alignment stringency (Rozowsky et al. 2011; Stevenson, Coolon, and Wittkopp 2013). Genetic variants that result in allele specific gene expression are also called expression quantitative trail loci (eQTL) that have been shown to have population and tissue specificity (Lappalainen et al. 2014; Battle et al. 2014). ASE analysis provides a direct measurement of the allelic differences by counting the number reads mapping to the two alleles, and provides a probabilistic significance estimate of the difference in the from of a p-value.

Alternative splicing results in the differential inclusion of exons into mRNA. Splicing is the most prevalent regulatory mechanism with 95% of genes undergoing splicing (E. T. Wang et al. 2008). The major splicing events are exon skipping, alternative use of splice donor and acceptor sites, intron retention and mutually exclusive exons (de Klerk and 't Hoen 2015). Most common mechanisms genetic variants impact exon usage is by exon skipping which occurs in approximately 30% of human and mouse genes showing great diversity in exon usage across tissues (Sugnet et al. 2004; Florea, Song, and Salzberg 2013). Detection of alternatively spliced mRNA transcripts and their exon structure is based on counting sequenced reads in a pre-defined exon map of the transcriptome and then performing a comparative estimation of the difference among conditions (Anders, Reyes, and Huber 2012). This approach can provide information on the diversity of transcripts with various combinations of exon usage, and quantifies the difference (fold change) between the conditions allowing for hypothesis-based analysis.

Non-sense mediated decay (NMD) is an important mRNA quality control mechanism that has been associated with over 10% of all human diseases (Bidou et al. 2012). NMD is caused by mutations that lead to premature termination codon (PTC) in the mRNA sequence. PTC can lead to the degradation of mRNA transcript and to non-functional protein or truncated polypeptide. Authentic stop codon or upstream mutations resulting in the change in the open reading frame can lead to PTC, therefore variant calling and annotation has a major role in the prediction of NMD transcripts. Optimally variant detection should be in genomic context as variant callers for RNA-seq are still in their infancy (Piskol, Ramaswami, and Li 2013). The next step is to correlate the stop codon signal to abundance of transcript where PTC lies. This is challenging because genes have multiple splice variants and therefore we need to identify the transcript where the stop codon lies, or infer NMD from read counts mapping to the wild type and mutant allele. There are some methods that detect NMD sensitive transcripts (Vitting-Seerup et al. 2014), but most commonly simple allele ratio estimates are used to infer NMD (MacArthur et al. 2012).

Outlier detection is one of the major steps in many "omics" applications. High-throughput "omics" generate large amount of data and obtaining the most important information, and to perform a coherent analysis many times starts with identifying observations that deviate from the bulk of the data. Thus, outliers are data that deviate so much from other observations that are suspected to be generated by other mechanisms (Hawkins 1980). A data point that is an outlier from the other data may be indicative of low sample quality, sample stratification, technical noise, and can suggest biologically important features that correlate with clinically important traits. It is therefore important to identify them prior analysis or as the goal of the analytical process.

Outlier analysis can be grouped into two main groups, univariate and multivariate methods. Univariate statistical models often rely on assumptions made about the distribution of the data, with the expectation that that data points are independently distributed (Ben-Gal 2005).

Essentially, a univariate model would calculate the sample mean and standard deviation of a single variable and classify outliers as measurements that are 2 or 3 standard deviation away from the mean. Visually univariate outliers can be detected by using scatterplots, QQ plots or boxplots. Univariate methods have difficulty when multiple outliers exist in the sample data. This can be attributed to the fact that when multiple outliers exist in the same direction, the mean of the sample data shifts and the standard deviation estimates increase so the lesser outlier falls within the standard deviation limit and thus goes undetected. This is called the masking effect. In other cases where large outliers shift the mean and the standard deviation so much as other observations become outliers as well is called swamping effect (Ben-Gal 2005). Statistical methods like the Grubb's test (Grubbs 1969) or the Tietjen-Moore test (Tietjen and Moore 1972) exist to compensate for effects biasing outlier estimation, but require the knowledge of expected number of outliers in the data and assume normal data distributions.

In RNA-seq, where data distributions do not follow normal distribution univariate measurements are fitted to Gaussian, Poisson, or beta distributions and outliers are estimated based on the probability that point belong to the data distributions. Such approaches are utilized in differential gene, exon or transcript expression studies that use a gene-by-gene technique to test whether a single measure in a patient's condition (i.e. expression of a gene) is significantly different from the expression of the gene in a control condition/group. In essence, these tests like DESeq, or DEXSeq, use read count data to quantitate expression level, and assuming a Poisson or negative binomial distribution, model the expression levels between conditions to estimate the probability that the gene is an outlier in the patient (A. Roberts et al. 2012; Love, Huber, and Anders 2014). However analysis of the transcriptome is performed over thousands of genes, and multiple testing corrections may leave biologically relevant, outlier genes off the list of significant differential expression list.

As described above RNA-seq provides information on multiple transcriptomic features and analyzing them individually provides information about the impact of genomic variation to the specific RNA feature. However, true understanding of biological systems, like transcriptome and its diversity can be best explained by integrating measurements from multiple transcriptomic features. Essentially, the identification of genes that are significantly impacted by genetic variation can be best studied if multiple measurements from allelic expression, exon usage, transcript diversity, or gene level expression can be combined and evaluated simultaneously. Patterns in these complex datasets can provide a means of quantifying truly multivariate patterns that arise from the correlational structure of the variables. Multivariate analysis also highlights patterns that are redundant in univariate analysis, provides means to identify patterns and relationships between variables that may be missed by univariate analysis. As an example, identification of alternative exon usage obtained from a single measurement for a patient that may be considered as a measurement error can gain biological importance if gene or transcript level measurements are combined with exon measurements and multiple variables indicate an outlier pattern of the exon when applied to multivariate algorithm. Thus, multivariate analysis of multiple variables is best suited for high dimensional data sets. Detection of outliers in multivariate models is only possible by identifying interactions between the different variables within the class of data. Essentially, by adding additional dimension to univariate data, outliers detected by univariate method can be confirmed or rejected, or new outliers can be identified relying on multiple measurements. Thus taking into account the relationship of the multivariate is a critical step in multivariate analysis. Some of the more common multivariate outlier methods include statistical models, and data-mining techniques (Ben-Gal 2005). Statistical models are based on the identification of observations that lie relatively far from the center of the multivariate data distribution. Data mining techniques apply clustering of multiple variables into distinct clusters that may include multiple observations indicating relationship of observations in multi-dimensional space. Multivariate analysis is not computationally intensive and can be used as an unbiased data exploratory tool simply summarizing the variability in the data (Jombart, Pontier, and Dufour 2009).

One of the most extensively applied multivariate statistical approach for RNA-seq data is principal component analysis (PCA) (Yeung and Ruzzo 2001). PCA is primarily used to reduce multi-dimensional data into as few components as possible that explains the greatest variability in the original data. PCA based methods work best for data that is transformed to normalize data distribution and stabilize variance. The transformation results in creation of linearly uncorrelated variables from possibly correlated variables that negatively impacts clustering of variables and impacts the number of outliers detected (Yeung and Ruzzo 2001). Un-correlating the variables allows for the estimation of distances between the variables using Euclidian distance. PCA is also sensitive for the scale of the variables. In quality control procedures of RNA-seq experiments using gene expression abundance measures across multiple samples as multivariate, PCA can identify samples whose expression do not adhere to group indicating potential quality issues (Ellis et al. 2013). In cancer, PCA allows for separation of normal samples from samples with different stages of tumor progression (Veytsman et al. 2014). Using a subset of genes and their expression profile outlier PCA can also identify sub-populations of cells among hundreds of single-cell RNA-seq experiments (Buettner et al. 2015).

The most common multivariate approach that takes into account the relationship between variables in a multivariate data space is the Mahalanobis distance (MD) (Mahalanobis 1936). Given n observations from a p-dimensional dataset, the algorithm first estimates the mean of each variable, followed by estimation of covariance between each variable. This is followed by taking the square root of the quadratic multiplication of mean difference and inverse of covariance matrix. Mahalanobis measures the distance for each observation from the multidimensional mean (centroid) of the data distribution given the covariance (De Maesschalck and Jouan-Rimbaud 2000). An observation is a multivariate outlier if its probability falls under a threshold given a degrees of freedom. Since Mahalanobis scores follow a Chi-Squared distribution for normal data the degrees of freedom equals the number of variables in the dataset.

The advantage of MD is that it does not have assumptions about the scale of the variables and does not require data normalization, or transformation. In addition it allows for integration of large number of variables that are only limited by the number of observations, as MD works best when number of observations exceeds the number of variables. The utility of multivariate outlier analysis and Mahalanobis distance to quantify outliers has been demonstrated by Kothari et al. (2013) who applied continuous variables of absolute gene expression and fold change from differential gene expression in a two-dimensional data space to identify kinase expression signatures across hundreds of samples that may be targets for pharmacogenomics treatment in breast and pancreatic cancer. In addition, Schissler et al. (2015) applied log2 transformed gene expression measurements from paired normal and tumor tissue of the same breast cancer patient in a two-dimensional data space to identify dysregulated pathways. In this approach the MD is calculated for each gene within its respective pathway with the initial assumption of no difference in gene expression between the normal and tumor tissues. Thus the MD is interpreted as a signed magnitude of differential expression between tissues incorporating the variance of other genes within the pathway where the gene lies. The average of gene specific MD scores for each pathway were then used to define the pathway as potentially relevant clinically. These two approaches underscore the utility of MD in large cohorts where patients with outlier expression signatures are studied for specific genes, or in single-patient cases where all expressed genes are evaluated for outlier gene signatures that may be associated with phenotype. Both methods show the utility of MD when variables of the same or different scales are studied.

However, the application of traditional gene-level expression signatures in cross patients studies can mask distinctive signals from single patient, and may not fully explain the significance of the gene signature to disease mechanism.

To address this, we present a framework to apply multivariate outlier analysis of multiple transcriptomic signatures of gene and exon expression from RNA-seq in a group of 29 patients with rare genetic conditions (Figure 15). Our cohort includes patients with or without genetic diagnosis, but enrolled patients present clinical symptoms that are difficult to categorize and do not easily fit into any clinical disease phenotype. Thus in a sense our cohort is a collection of clinical outliers. We use gene and exon expression to search for transcriptional multivariate patterns that are rare, and outliers in the multivariate data space. Specifically, our analysis framework leverages multivariate outlier analysis by MD, recognizing that there are thousands of possible genetic variants in standard clinical exome sequencing to make diagnosis in any given pediatric disorder of unknown etiology. The goal is to provide MD score based on the expression profile for each candidate variant, such that those with the highest score are indicative of outlier expression pattern supported by gene and exon data within our cohort. In this case, if there were 500 candidate variants one would perhaps choose for in depth functional analysis the variants with highest scores before those with lower scores. Thus, we will utilize the hypothesis that variants with substantial impact on transcription by their outlier score would be more likely to be functional, and thus more relevant than a non-functional variant when considering the variant with a phenotype or disease in any child. Multivariate analysis by Mahalanobis distance of outliers provides a tested and established approach to integrate gene and exon expression values for patient specific variants so they can be simultaneously interpreted with clinical information.

In this chapter, genomic information in the form of rare variant annotation from family based DNA sequencing are integrated with MD scores for each patient. Integration of genomic data with MD scores was expected to identify rare, functional variants that have significant impact on transcription, and provide a basis to further reduce the list of potential candidate variants in rare disease diagnosis. Our results show that gene-based MD scores have association with variants predicted to have high functional impact. We also found that frameshift variants had higher outlier scores than variants in other functional classes. Using this approach we found that presumed causal variants previously identified by DNA sequencing in a subset of cases showed large functional impact corroborating the genomic findings and supporting causality. Integration of RNA-seq based outlier analysis also revealed new candidate variant in previously undiagnosed case, suggesting the utility of integrated DNA-RNA analytical approaches in the diagnosis of rare childhood diseases.

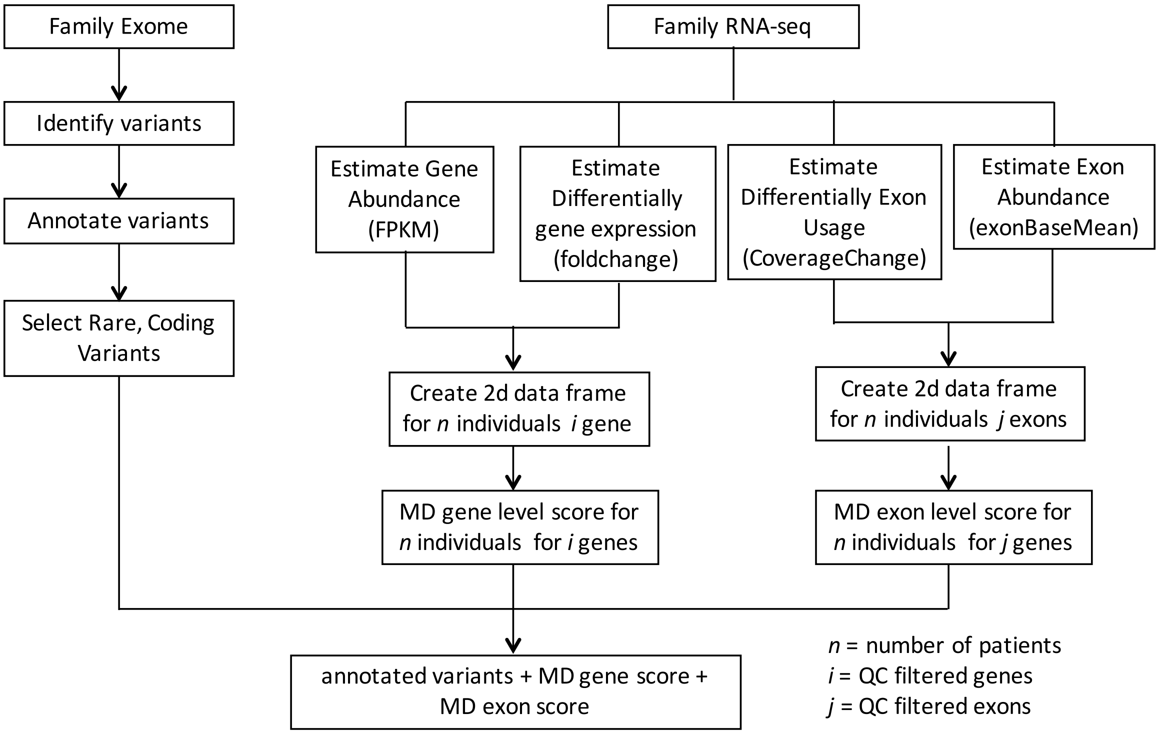
Family Exome
Identify variants
Annotate variants
Select Rare, Coding Variants
Family RNA-seq
Estimate Gene Abundance (FPKM)
Estimate Differentially gene expression (foldchange)
Estimate Differentially Exon Usage (CoverageChange)
Estimate Exon Abundance (exonBaseMean)
Create 2d data frame for n individuals i gene
Create 2d data frame for n individuals j exons
MD gene level score for n individuals for i genes
MD exon level score for n individuals for j genes
annotated variants + MD gene score + MD exon score
n = number of patients
i = QC filtered genes
j = QC filtered exons

Figure 15. Schematic overview and workflow. We prepared an integrative DNA and RNA sequencing data set by combining family-based whole exome data with family-based RNA-seq analysis results for 29 patients from the Center for Rare Childhood Disorders at TGen. Exome data was obtained from family sequencing, variants were called, annotated and filtered by in-house analytical pipeline for all 29 patients. RNA-seq was performed for the same family members followed by differential gene and exon expression analysis between each patient and their parents. Multiple measures were taken from each differential analysis and used to perform multivariate outlier analysis by Mahalanobis distance for each expressed gene and exon in the 29 patient cohort. MD scores obtained for each patient transcriptome data were integrated with variant annotations from exome sequencing to a final tabulated variant table utilized for variant prioritization.

Materials and Methods

Patients

Patients with undiagnosed genetic condition from 32 families were selected from the Dorrance Center for Childhood Disorders between 2012 and 2014. Enrollment criteria into the Center's study included, but were not limited to previously undiagnosed, possibly severe condition, an ambiguous genetic origin, and negative, or inconclusive genetic tests prior enrollment.

Standardized clinical assessment was performed by the referring physician or by the center’s clinical staff. All patients went through standard clinical evaluation prior enrollment, and remained undiagnosed. Clinical evaluation varied case-by-case and included but were not limited to karyotyping, genetic panel testing, mitochondrial DNA genotyping, magnetic resonance imaging, chromosomal microarray testing, enzymatic assays. Most patients exhibit some form of neurological phenotype and were characterized as one of the following condition: Neurologic, Multi-system, Musculoskeletal structural, Cardiac (Table 13). Written informed consent was obtained from the patients at the time of enrollment, or from parent/guardian for patients under the age of 18. The Western Institutional Review Board (WIRB) approved this study.

The goal of the center is to obtain consent and to collect biospecimen from the patient and his/her biological parents. Recognizing that not all family members could be consented and whole blood obtained, we define a family trio, with exome, and/or whole genome, and RNA sequencing was performed in the patient and their biological parents. In addition we define a singleton where whole genome or exome sequencing could be performed only in the patient. Furthermore, we define a large family where exome and\or whole genome sequencing was performed for patient, biological parents, affected or unaffected siblings, grandparents, uncles, and/or aunts. Finally, we define a parent-child duo in those families where exome or whole genome data could only be obtained from one of the biological parent and the patient. For detailed clinical description of each patient please refer to Appendix A.

Table 13.

Study patients.

Family	Patient (Gender)	Ethnicity	Family History	Age/Age of Onset	Organ System	Clinical Diagnosis
0001	1(F)	Caucasian	N	15y/<1mo	Neurologic	Neurotransmitter Disorder
0002	1(F)	Caucasian	Y	13y	Unaffected	Migraine
0002	2(F)	Caucasian	Y	21y/5y	Cardiac, Neurologic	Intellectual disability, (Wolfe-Parkinson-White syndrome)
0002	4(M)	Caucasian	Y	18y/8y	Neurologic	Hemiplegic migraine
0002	5(M)	Caucasian	Y	9y/<2y	Neurologic	Leigh Syndrome; Mitochondrial encephalopathy
0004	1(M)	Caucasian	N	17y/4mo	Neurologic	Developmental Delay | ID | Microcephaly
0005	1(M)	Caucasian	N	6y[1]/<1mo	Neurologic, Musculoskeletal structural	Nystagmus|Motor Delay|Feeding Disorder
0006	1(F)	Middle East	C|Y	12y/2-3y	Neurologic, Musculoskeletal structural	Ataxia with sensory neuropathy
0008	1(F)	Indian	N	17y/3-4mo	Musculoskeletal /structural	progressive leukoencephalopathy|spastic |global cerebral atrophy
0016	1(M)	Asian	N	10y/6y	Neurologic, Musculoskeletal structural	progressive cerebellar ataxia|dystonia
0018	1(F)	Caucasian	N	15y/2y	Neurologic	ADHD, Autism Spectrum Disorder
0019	1(F)	Caucasian	Y	11y/<1mo	Neurologic	non-progressive cerebellar ataxia, infantile dystonia
0024	1(M)	Middle East	C	17y/2 mo	Neurologic	Aicardi-Goutieres Syndrome
0025	1(M)	Caucasian	N	6y/<1mo	Musculoskeletal structural	motor delay, hypotonia|feeding disorder
0049	1(F)	Caucasian	N	11y/<1mo	Neurologic, Musculoskeletal structural	Cockayne Syndrome, COFS-2
0091	1(M)	Caucasian	Y	9y/5y	Neurologic	Schizophrenia
0103	1(M)	Hispanic	Y	19y[1]/6mo	Neurologic	NBIA| BPAN
0103	2(F)	Hispanic	Y	14y/<1y	Neurologic	NBIA| BPAN
0117	1(M)	Caucasian	N	10y/birth	Neurologic	congenital nystagmus, Pelizaeus–Merzbacher-like
0139	1(M)	Caucasian	N	19y/prenatal	Cardiac, neurologic	Situs inversus; developmental delays, chronic lung disease
0152	1(M)	Caucasian	N	3y[1]/<1y	Multi	Leigh Syndrome
0157	1(F)	Caucasian	N	5y/1y7mo	Neurologic	Developmental Delay
0011	1(F)	Caucasian	N	8y/<1mo	Neurologic	Aicardi Syndrome
0014	1(F)	Caucasian	N	14y/<1mo	Neurologic	Aicardi Syndrome
0033	1(F)	Caucasian	N	7y/3mo	Neurologic	Aicardi Syndrome
0034	1(F)	Hispanic	N	3y/<1mo	Neurologic	Aicardi Syndrome
0046	1(F)	Afr.American /Caucasian	N	4y/3mo	Neurologic	Aicardi Syndrome
0047	1(F)	Caucasian	N	14y/3mo	Neurologic	Aicardi Syndrome
0048	1(F)	Caucasian	N	8y/3mo	Neurologic	Aicardi Syndrome
0118	1(F)	Caucasian	N	18y/3mo	Neurologic	Aicardi Syndrome
0059	1(F)	Caucasian	N	9y/3mo	Neurologic	Aicardi Syndrome
0012	1(F)	Caucasian	N	12y/3y	Neurologic	Developmental delay, autism spectrum disorder
0020	1(F)	Caucasian	N	5y/birth	Multi	Neonatal progeroid disorder, failure to thrive
0023	1(F)	Hispanic	N	7y/<1y	Neurologic	Infantile choreoathetosis; dystonia, rigidity; cognition is near normal
0029	1(F)	Caucasian	Y	6y/<2y	Neurologic	Leukoencephalopathy
0140	1(F)	Caucasian	N	5y/10mo	Musculoskeletal	

Abbreviations: F=female, M=male, N = no family history, C = consanguinity or suspected consanguinity, Y= multiple affected within the family, [1] = expired.

The 32 enrolled families consisted of 18 trios (56%), 8 large families (25%), 4 singletons (13%), and 2 parent+proband duos (6%) (Figure 16). This cohort is ethnically heterogeneous, 27 patients are Caucasian (75%), 4 are Hispanic (11%), 1 of African American descent (3%), and 4 are of Asian descent (11%)In six families, there is a family history of the rare condition with multiple affected individuals (0002, 0006, 0019, 0091, 0103, 0029), and we enrolled multiple affected patients from families 0002 (n=3) and 0103 (n=2) for a total of 36 patients. One of the children (0002_1) is diagnosed as unaffected sibling but we included her due to some mild symptoms that we felt was important to decipher the phenotypic heterogeneity within the family. From the 32 families 25 families participated in the study described in this chapter (Ch.4), 5 families in the study described in Chapter 3 and 30 families in the study described in Chapter 2. There is an overlap between the studies in terms of participation and participation is described in Table 3. Among the 32 families, there were a total of 24 female patients (66.7%) and 12 (33.3%) males. The study participants included 10 females diagnosed with Aicardi Syndrome (MIM:304050). In two families (0006, 0024) the clinicians reported that there was evidence of consanguinity. In 23 patients the primary organ system that is affected is neurologic, for 5 patients a combination of neurologic and musculoskeletal symptoms were observed. Three patients show severe musculoskeletal symptoms, and two patients show extensive multi-system clinical symptoms. In ten families, the likely pathogenic, disease causing mutations using exome and genome sequencing was identified prior RNA-seq (0001, 0002, 0005, 0012, 0018, 0020, 0024, 0047, 0049, 0103).

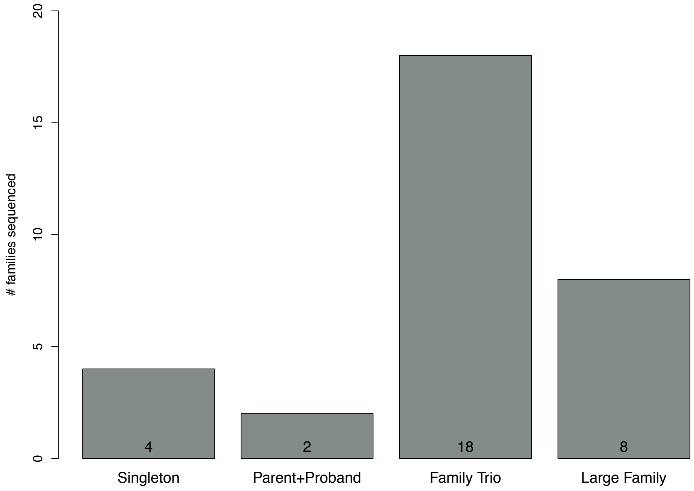

Figure 16. Family structure of enrolled patient.

Table 14.

Study Participation and sequencing

Family	WGS	Count	WES	Count	RNA-seq	Count	Study		
0001	-	-	P|M|F|S1|S2|S3	6	P|M|F|S3	4	Ch4	Ch2	
0002	-	-	P1|P2|P3|M|F|S	6	P1|P2|P3|M|F|S	6	Ch4	Ch2	
0004	P|M|F	3	-	-	P|M|F	3	Ch4	Ch2	
0005	P	1	-	-	P|M|F	3	Ch4		
0006	P	1	-	-	P|M|F	3	Ch4		
0008	-	-	P|M|F|S1	4	P|M|F	3	Ch4	Ch2	
0011	P|M|F	3	-	-	P|M|F	3	Ch4	Ch2	Ch3
0012	P	1	M|F	2	P	1		Ch2	
0014	P|M|F	3	-	-	P|M|F	3	Ch4	Ch2	Ch3
0016	-	-	P|M|F	3	P|M|F	3	Ch4	Ch2	
0018	-	-	P|M|F	3	P|M|F	3	Ch4	Ch2	Ch3
0019	-	-	P1|P2|P3|M|F|S	6	P1|M|F	3	Ch4	Ch2	
0020	P	1	-	-	P	1		Ch2	
0023	-	-	P|M|F	3	P	1		Ch2	Ch3
0024	-	-	P|M|F	3	P|M|F	3	Ch4	Ch2	
0025			P1|M|F|S1|S2	5	P1|M|F	3	Ch4	Ch2	
0029			P1|P2|M|F|S	5	P1	1		Ch2	
0033			P|M|F	3	P|M|F	3	Ch4	Ch2	
0034			P|M|F	3	P|M|F	3	Ch4	Ch2	Ch3
0046			P|M	2	P	1		Ch2	
0047			P|M|F	3	P|M|F	3	Ch4	Ch2	
0048			P|M|F	3	P|M|F	3	Ch4	Ch2	
0049			P|M|F	3	P|M|F	3	Ch4	Ch2	
0059			P|M|F	3	P|M|F	3	Ch4	Ch2	
0091			P|M|F|S1|G	5	P|M|F|S1|G	5	Ch4	Ch2	
0103			P1|P2|M|F	4	P1|P2|M|F	4	Ch4	Ch2	
0117			P|M|F	3	P|M|F	3	Ch4	Ch2	
0118			P|M	2	P|M	2		Ch2	
0139			P|M|F	3	P|M|F	3	Ch4	Ch2	
0140			P	1	P	1		Ch2	
0152			P|M|F	3	P|M|F	3	Ch4	Ch2	
0157			P|M|F	3	P|M|F	3	Ch4	Ch2	
Total		13		90		90			

P,P1,P2,P3= proband, M=mother, F=father, S,S1,S2= unaffected sibling, Ch2=Chapter 2, Ch3=Chapter 3, Ch4=Chapter 4

Biospecimens

We collected from each consented study participant whole blood in Vacutainer Blood Collection Tube (Becton, Dickinson and Company; Franklin Lakes, NJ) and in PaxGene RNA tube (Qiagen; Germantown, MD). Genomic DNA isolation was performed in multiple stages depending on the time of enrollment. Patients DNA enrolled prior January 2013 was isolated at Barrow Neurological Institute using Wizard SV Genomic Purification System (Promega, Madison, WI). From 2013 blood collections were sent out for DNA and total RNA isolation at GeneDx

(Gaithersburg, MD). From 2014, genomic DNA isolation was performed under Clinical Laboratory Improvement Amendment (CLIA) standard operating procedures. RNA isolated at TGEN followed standard manufacturer recommended protocol using PaxGene Blood miRNA kit (Qiagen; Germantown, MD).

Whole Genome Sequencing

Whole Genome Sequencing was performed for a total of 13 individuals, including 3 parents-proband trios, and 4 singletons proband. Genomic DNA from 10 of the 13 individuals were prepared and sequenced at Illumina Whole Genome Sequencing Service (Understand Your Genome, Illumina FastTrack), and one trio was prepared and sequenced at TGEN (Table 3).

TGEN library preparation was performed using Illumina suggested whole genome library preparation protocol with some modifications to achieve sequencing libraries with longer than 500bp insert size. 1μg of genomic DNA was fragmented using random shearing by sonication on the Covaris S1 system to a target insert size of approximately 1000bp. After fragmentations the sheared fragments were blunt end repaired and A base added to the 3' end of the DNA fragments. Barcoded adapters ligated to fragments by an A-to-T ligation step followed by size selection on agarose gel. DNA bands corresponding to approximately 1000bp were sliced out of the gel and purified using the Quantum Prep Freeze'N'Squeeze DNA Gel Extraction Spin Columns (Bio-Rad, Hercules, CA). Purified genomic DNA was consequently PCR amplified, and quantitated by qPCR, followed by equimolar pooling. The pooled trio was sequenced on a single HiSeq2000 flowcell using multiplexed, paired sequencing chemistry for 100 bp read length.

Whole Exome Sequencing

Coding regions were captured using TruSeq Exome Enrichment Kit v2 (Illumina, San Diego, CA) and SureSelectXT Human All Exon V5 (Agilent, Santa Clara, CA) following manufacturer recommended protocol. Sequencing was performed after prepared samples were pooled in pools of 6 for TruSeq Exome libraries, or pools of 8 for SureSelect libraries. Each pool was sequenced on two lanes of a Hiseq2000 flowcell using multiplexed paired end chemistry and 101bp read length with a goal of 100X coverage.

RNA preparation and sequencing

Total RNA was isolated from PaxGene blood tubes (Qiagen, Georgetown, MD) using manufacturer recommended protocol. The purity of the total RNA was assessed using Nanodrop ND-1000 (Thermo Scientific, Wilmington, DE), and integrity was assessed by BioAnalyzer 2100 (Agilent, Santa Clara, CA). Samples with an RNA integrity number (RIN) of at least 5 were used in this study. mRNA libraries were prepared using Illumina TruSeq stranded RNA library preparation kit and Illumina TruSeq RNA library preparation kit v2 (Illumina, San Diego, CA). The choice of kit was consistent within families. The Illumina sample preparation kits utilize oligo-dTs hybridized to magnetic beads to purify the mRNA molecules followed by thermal fragmentation. The fragmented RNA molecules were converted to first strand cDNA by random hexamer primers and reverse transcriptase enzyme. DNA Polymerase I and RNase H were used to polymerize second strand of cDNA. Double stranded cDNA molecules were end repaired to obtain blunt ends, which was followed by ligation of a single A base to each 3' end. Sequencing adaptors with unique barcodes and T overhang were ligated to A-tailed cDNA fragments creating a final sequencing library. Libraries were amplified to increase cDNA yield for sequencing and final amplified libraries are quantified by qPCR. Final libraries were evaluated for fragment size distribution using Agilent BioAnalyzer 2100. Stranded RNA library preparation includes addition of Actinomycin D to reduce DNA-dependent synthesis during first strand cDNA synthesis. Strand specificity was achieved by incorporating dUTP instead of dTTP in second strand of cDNA. Quantified libraries were equimolarly pooled based on qPCR concentrations into pools of 4, and final library pools were quantified before cluster generation for sequencing. Each pool was sequenced on a single lane of a HiSeq2000 flowcell using multiplexed, paired end sequencing chemistry for a 101bp read length.

Bioinformatics Analysis

Upon completion of sequencing runs, raw basecall files were converted to sequenced reads in FASTQ format using CASAVA 1.8.2 (Illumina, San Diego, CA). The fastqc package was used to evaluate the raw reads for overall quality (fastqc). Reads were aligned to human reference genome hs37d5.fa from the 1000 Genomes Project. The reference genome contained contigs associated with ribosomal unit, cancer causing viruses, and ERCC spike-ins. Alignment was performed by mem module of the Burrows Wheeler Aligner (BWA v0.7.8) (Li 2013), and binary alignment files were generated by SAMTOOLS v0.1.19(H. Li et al. 2009). After alignment the base quality scores were recalibrated and joint indel realignment was performed on the BAM files of each family member using Genome Analysis Toolkit (GATK v3.1-1)(McKenna et al. 2010). Duplicate read pairs were marked using PICARD v1.119 (picard). Single nucleotide polymorphisms (SNPs), short insertion and deletions were identified using HaplotypeCaller module of GATK.

RNA data

Reads were aligned to human reference genome hs37d5.fa as described above. Alignment was performed by Spliced Transcripts Alignment to a Reference (STAR_2.3.1z_r395) (Dobin et al. 2012) and binary alignment files were generated by SAMTOOLS v0.1.19(H. Li et al. 2009). Alignment was facilitated using known transcript structure of the human genome from Homo sapiens GRCh37.74.gtf (ftp://ftp.ensembl.org/pub/release-74/gtf/homo_sapiens/). Duplicate read pairs were marked using PICARD v1.119 (picard). Final bam alignments were used to estimate gene expression abundance in the form of normalized FPKM values (Fragments Per Kilobase Of Exon Per million Fragments Mapped) using the Cufflinks 2.2.1 package(Trapnell et al. 2013). Library size normalization was performed across all the families presented as follows: Post BAM file generation, individual BAM read alignment files were processed by cuffquant module of Cufflinks. Cuffquant essentially takes a transcript annotation and an alignment file from the RNA-seq experiment and pre-processes the information in the alignment with reference to the transcript coordinates by generating a binary output that is an input to cuffnorm, reducing the computational burden on the normalization step. The sample level output of cuffquant module is the input for the next module called cuffnorm. Cuffnorm takes input reference annotation and a list of output files from cuffquant, and it normalizes FPKM abundances across all the input based on geometric mean of all samples in the normalization and controlling for library size. We normalized FPKM abundances across 79 enrolled participants from 25 families including patients and their biological parents described in this chapter.

We used the cuffdiff2 module of Cufflinks to estimate differential gene expression of annotated transcripts. In each comparison the biological parents were assigned as “control”, and thus differential gene expression was based on parent compared to offspring/patient. Cuffdiff applies a geometric normalization method the number of fragments mapping to each transcript and applies an algorithm that considers cross-replicate variance and uncertainty in read mapping to different isoforms of the same gene(Trapnell et al. 2013). It models fragment counts using the beta negative distribution and reports change in expression between conditions (eq, parents, patient) on gene level with statistical significance. Cuffdiff2 calculates the log2 foldchange of gene expression between conditions. In those genes where one of the conditions had zero mapped fragments the foldchange is positive or negative infinity depending which condition has zero fragments.

Normalized FPKMs and cuffdiff2 analysis output were inserted into an in-house relational database. The Mongo database was based on dynamic, document style data structure that allowed for horizontal and vertical scaling giving flexibility to storage and access to ever-increasing omics data (mongo). We reduced the cuffdiff2 output by filtering out genes assigned “FAIL”, “LOWDATA”, “HIDATA”, or “NOTEST”.

Estimation of alternative exon usage was performed using the R 3.1.2/Bioconductor package DEXSeq v.1.12.2(Anders, Reyes, and Huber 2012). This method is based on the counts of mapped reads overlapping well annotated exons. If read overlaps exon boundaries of multiple overlapping transcripts with different boundaries for the exon, DEXSeq merges all the boundaries of the exons into a single feature and breaks it up into multiple “counting bins”. Reads were counted that overlap exon boundaries in protein coding regions using a flattened gtf annotation based on Homo sapiens.GRCh37.74.gtf. We excluded all overlapping exons of different genes to reduce the number of merged exons as DEXSeq does not count overlapping exons into their own respective bins. After read count DEXSeq uses sizefactor obtained from the geometric mean coverage of each exon across conditions to report normalized exon coverage “exonBaseMean” as the abundance estimate of exon expression. The read count dispersion due to technical and biological variability is estimated for each condition and a generalized linear model is used to estimate differential exon usage for each counting bin(Anders, Reyes, and Huber 2012). Based on DEXSeq read count we estimated a normalized coverage difference between the conditions (*E*). For each sample we calculated the total number of reads mapping to all counting bins (N_a, N_b…N_n). Next we obtained the ratio of counts per bin (*Ci*) over all the reads mapping to all counting bins for each sample across all counting bins. Finally we obtained a relative coverage difference in each exon (*Ei*) between the conditions by dividing the normalized count in the condition 1 (patient) with sum of normalized counts in condition 2 (parents) as seen in Equation 2:

$$Ei = \frac{\frac{Cia}{Na}}{\Sigma\left(\frac{Cib}{Nb}\right)}$$

The DEXSeq output and the calculated coverage difference were inserted into in-house mongo database. We used exonBaseMean and normalized coverage difference (nDiff) as measurements in multivariate outlier analysis of exons expression.

Variant Annotation Matrix

Variants identified by the HaplotypeCaller were inserted into mongo database for each family from VCF formatted variant list, followed by annotation by snpEff according to Homo sapiens GRCh37.74 annotation(Cingolani et al. 2012). SnpEff annotated variants with their predicted functional impact on amino acid change, protein structure. We included annotations for known canonical transcripts only. In addition, variants were annotated with prediction scores for functional, pathogenic affect (SIFT, CADD, MutationTaster), conservancy (phyloP, phastCons), and population frequency (dbSNP141, Exome Sequencing Project, 1000 Genomes) using the collection of annotations stored in dbNSFPv2.8 (P. C. Ng and Henikoff 2003; Kircher et al. 2014; Schwarz et al. 2014; X. Liu, Jian, and Boerwinkle 2013; Pollard et al. 2010; Siepel et al. 2005;

Consortium et al. 2010). Genotype-phenotype correlation from ClinVar was added. ClinVar contains genomic variants and their relationship to observed phenotype, health status, categorizing them based on likeliness of pathogenicity(Landrum et al. 2014). We also added disease gene annotation from Clinical Genomics Database(Solomon, Nguyen, and Bear 2013) which contains over 3000 genes that are know to be associated with genetic conditions. This database was curated into adult and pediatric disease causing genes with information on available intervention, and primary organ system to be affected. Population frequency estimates from large-scale exome sequencing project, the Exome Aggregation Consortium were also added to variant annotations (Exome Aggregation Consortium). We removed all variants that fell within 5' and 3' UTR regions, introns, upstream or downstream of genes, intergenic variants with the expectation that most rare, functional variants that may be detected by mRNA sequencing will lie within amino acid coding regions. Variants with a phred-scaled genotype confidence quality (GQ) of less than 90, and phred-scaled probability estimate (QUAL) that the SNP event exist of less than 500 as described in the VCF format guide were removed (Consortium et al. 2012). Annotated variants were further filtered for an estimated allelic frequency of less than 5%, a measurement taken from the maximum population frequency of the population frequencies reported in dbNSFP v2.9(X. Liu, Jian, and Boerwinkle 2013).

Multivariate Outlier Analysis

Central to multivariate analysis is the definition of objects and the number and nature of variables that is to be analyzed for each object. In addition interpretation of multivariate data depends on the relationships we set out to observe; whether we are looking for relationship between the objects or the variables. (Jombart, Pontier, and Dufour 2009). In this study we defined objects as patients and the variables as measurements obtained from RNA-seq analysis with the intention to study relationship between patients in terms of outlier behavior. Outlier behavior was estimated in a gene and exon level. Thus for each expressed gene and exon we built two-dimensional matrices with two vectors. In each matrix we tested a single gene or exon with n objects and p variables. In gene-based matrix the variables included two vectors of continuous values of log2 transformed, normalized gene abundance defined by the FPKM value from Cufflinks, and log2 fold change expression difference between each patient and their parents from Cuffdiff. In exon-based matrix for n objects we included two vectors with continuous values of log2 transformed, normalized exon coverage defined by exonBaseMean value in DEXSeq, and log2 normalized exon expression difference defined by the nDiff value as described above in Methods. We set following rules for estimation of the distance scores for each gene: 1) each vector had to have the same dimension, thus both FPKM and log2foldchange must be obtained for a patient, 2) genes with zero covariance were filtered out, 3) each value within the vectors had to be numeric, thus genes with a log2foldchange of negative or positive infinity were given an arbitrary value of -19 or +19, respectively. We applied this method to account for the fact that in differential expression, the lack of read fragments in one of the two condition results in a logarithmic ratio that may not capture well biological significance. Since, we were looking for extreme events, insufficient read fragment coverage can indicate biologically important events that one could pursue. 4) We defined the detection limit of a gene to FPKM ≥ 0.1, and required that >90% of objects have an expression above defined limit. For each gene that passed detection criteria we added 0.1 to the FPKM to facilitate our ability to transform the FPKM values to a logarithmic scale and apply uniformly scaled data for distance analysis.

Estimation of Mahalanobis distance for expressed exons followed similar rules as described for genes. We used exonBaseMean of ≥1 as our detection limit and required at least 90% of patients to have an expression above detection limit. We performed logarithmic transformation of exonBaseMean and nDiff prior Mahalanobis distance analysis.

We used the native Mahalanobis function of the R statistical package wrapped in a perl script to first query our database of RNA-seq expression results for each gene and exon across the 29 patients followed by loading the descriptors into the n x p vector matrix for distance analysis in R programming language(Rv3.1.2). MD score (MD) was determined for each object (n) in the n x p matrix with respect to the vector means (μ) for each vector p, and the covariance (S) of all vectors as shown in Equation 3:

$$MD^2 = (p - \mu)'S^{-1}(p - \mu)$$

Statistical Analysis.

We evaluated the gene-based and exon-based MD scores using a non-parametric Kolmogorov-Smirnov test to compare the two data distributions. This test has the advantage of no assumptions about normality of the distributions. Similar distributions would suggest that measurements are taken from the same data and exon and gene based scores would be redundant. Levene's test for equal variance test was used to test for variance in the MD scores grouped based on their functional impact. MD scores were grouped into "high" or "moderate-to-low" groups as described in the Results section. Levene's test was used to test the first assumption of the Mann-Whitney rank sum test of equal variance. Using Mann-Whitney of the MD scores, we are able to compare the variants predicted to have high or moderate-to-low functional impact. Mann-Whitney is a non-parametric rank sum test that can test for the null hypothesis of no difference in the mean ranks of the data distributions. A significant difference in mean ranks between high and moderate-to-low functional impact would indicate a more significant impact on transcription by one of the functional classes. Mann-Whitney was performed using the Wilcox test.

In addition the variants were further grouped into nine functional classes including frameshift, insertions-deletions, missense, sequence feature, splice site, splice region, start/stop, synonymous , start codon as described in the Results section. Kruskal-Wallis test was used to test the hypothesis that MD scores among the nine functional classes show different distribution of scores. This test is essentially an extension of the Mann-Whitney test for comparing more than two data sets. When multiple data distributions are compared the Kruskal-Wallis tests for the null hypothesis that the median ranks of all groups come from the same distribution. A significant finding would suggest that one of the data distributions is enriched for higher MD scores. Since the Kruskal-Wallis does not identify which functional class is enriched for outliers if the null hypothesis fails, we performed pairwise, non-parametric test of rank sums by Dunn's test (Dunn 1964). Dunn's test uses average ranking from Kruskal-Wallis rank sum test and test the null hypothesis of no difference in pairwise manner. It reports a z score as test statistics based on the difference of the average ranks and the sum of ranks between the two groups. The pairwise probability that one random value from a group is larger than a random value from another group can then be evaluated for multiple testing corrections. All tests were preformed in R programming language using ks.test, levene.test. dunn.test functions.

Results

We will present the results in three sections. In the first section the QC metrics of the DNA and RNA sequencing will be shown focusing on the obtained throughput and QC metrics that are standard procedures for large scale DNA and RNA sequencing projects. In the second section, we analyze the multivariate MD score's ability to discern transcriptionally functional variants. In the third section, we will present two families applying the approaches within the context of a genetic diagnosis.

***Quality Control Metrics*.** We first provide quality control metrics for the sequence data we generated. Between 2012 and 2014 we obtained whole genome data for 7 patients and 6 parents for 3 trios and 4 singletons for a total of 7 families with WGS data. Median genomic coverage for the trio sequenced at TGEN was 23.2 ± 0.5X and for the individuals sequenced at Illumina was 43.1 ± 6.5X (Table 15). Genomic coverage analysis indicated that in each TGEN prepared genome >90% of bases were sequenced at least 10X depth, and for the Illumina genomes >90% of bases for sequenced at least 20X reads depth (Figure 17A). We must recognize that the genomes sequenced at TGEN were prepared using non-standard methods to obtain long insert library of approximately 1000bp. Standard Illumina library preparation methods target a 350bp insert size and kits designed for clinical sequencing service have been optimized for throughput and quality. In light of this, the long insert libraries of samples 004_1, 004_2, and 004_3 have performed well. In addition we sequenced 90 whole exome samples from the remaining 19 families. The mean per base coverage for the exome target regions was 85 fold, with 93.1% of bases covered more than 10 fold and 67.1% above 50 fold (Figure 17B.) On average 4.5 million SNVs and short indels were identified in the whole genome data and 475 thousand in the exome data. The average dbSNP rate that shows the proportion of variants identified previously in human populations is 0.96 for the genomes and 0.94 for the exomes which is in line with previous findings by the 1000 Genomes project (Consortium et al. 2012). We find on average the exome data resulted in more non-synonymous variants called 3,695 in exomes compared to 2,965 in the genomes. Exome sequencing achieves a higher overall coverage in exons than genome sequencing thus increasing confidence in variant calls in exome data. Interestingly exome analysis resulted in a lower average calls in start sites (n=151) compared to the genomes (n=180). The capture of exon 1 in next-generation sequencing is a known issue caused by a higher GC content in first exons. Thus exome kits are challenged by this and the optimization of exome capture kits to leverage the efficient capture of first exons with start sites is an ongoing process. The results of variant calling and annotation can be found in Table 16.

Table 15.

Quality metrics of Whole Genome Sequencing

Individual	Gender	Total Reads(M)	Read Length (bp)	% Reads Aligned (Pairs)	QC Aligned Bases (M)	Coverage (X)	Insert Size (bp)	% Duplicates
004_1	Male	897	101	99.78	88,648	22.6	722	15.05
004_2	Female	895	101	99.85	88,636	23.7	701	11.99
004_3	Male	884	101	99.86	87,498	23.3	708	11.73
005_1	Female	1,163	100	99.51	114,343	35.7	302	2.07
006_1	Female	1,258	100	99.51	123,051	38.8	304	2.16
011_1	Female	1,296	100	99.74	127,581	38.7	295	3.81
011_2	Female	1,762	100	99.73	172,955	52.5	285	4.25
011_3	Male	1,137	100	99.69	111,390	33.8	294	3.47
012_1	Female	1,526	100	99.42	149,566	47.6	313	2.78
014_1	Female	1,565	100	99.60	154,337	48.5	317	2.70
014_2	Female	1,123	100	99.66	110,738	48.5	323	2.31
014_3	Male	1,541	100	99.66	152,066	47.5	303	2.95
020_1	Female	1,310	100	99.89	129,322	39.4	280	3.66

Abbreviations; M= million, bp=base pairs.

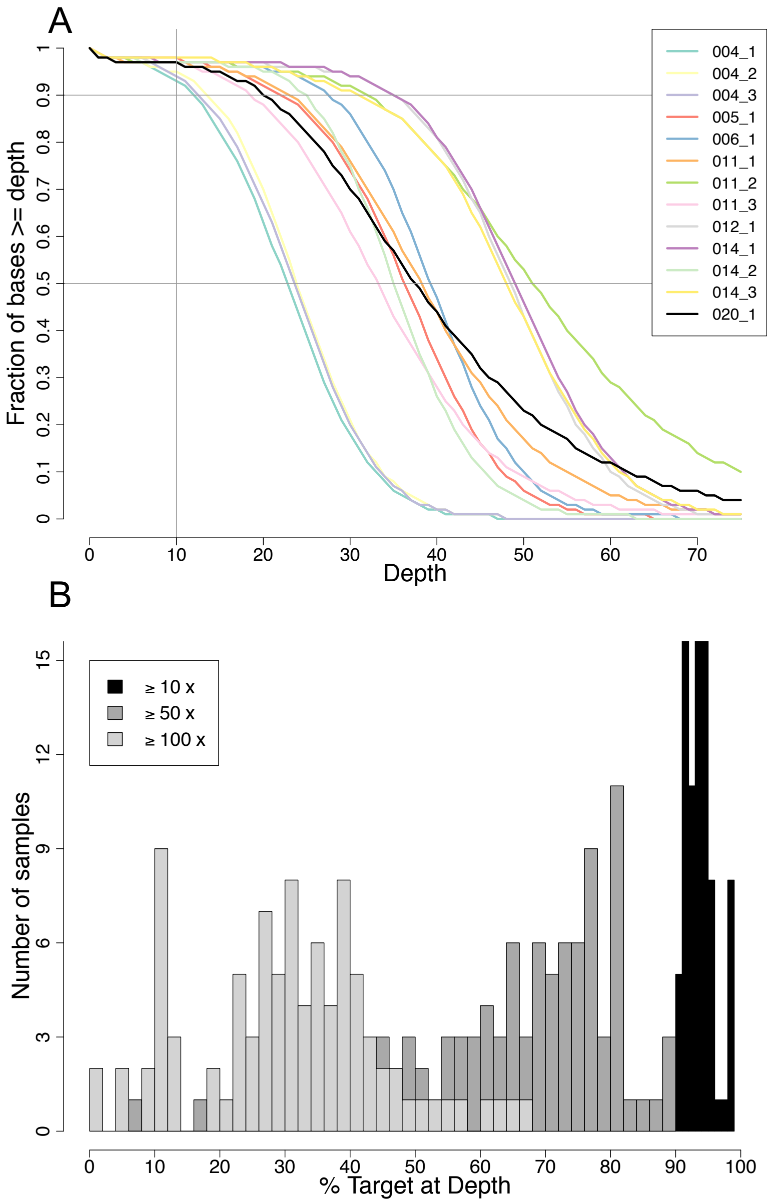

Figure 17. Coverage analysis of WGS and WES. Plot shows sequencing depth obtained for whole-genome sequencing (A) and whole-exome sequencing (B).
A) Samples 004_1, 004_2 and 004_3 are the long insert libraries with a lower overall coverage that can be seen by the three curves to the left. In these samples 50% of bases were covered by at least 20X The Illumina sequenced samples show that 50% of bases are covered at a minimum of 30X (011_3), and in some cases 50X of average depth is achieved for 50% of bases (014_2).
B) Histogram showing the percent of targets (exons) with average coverage of 10X, 50X, 100X. Target coverage for 90 exomes show that samples achieve at least 10X average target coverage in >90% of targeted regions, and between 70-80% of targets are covered at 50X for most exomes.

Table 16.

Variant call metrics for 32 families.

Assay	Family Id	Family Size	SNP count	dbSNP rate	Insertions	Deletions	het/hom ratio	Ti/Tv ratio	Frameshift	Missense	Splice Site	Start	Stop	Exon Del
WGS	0004	3	5,063,930	0.96	455,201	492,023	1.57	2.03	195	3,217	66	199	45	-
	0005	1	3,799,567	0.97	386,288	393,962	1.57	2.03	142	2,377	53	147	36	-
	0006	1	3,889,748	0.96	401,692	409,745	1.62	2.02	167	2,560	48	146	32	-
	0011	3	5,123,155	0.95	450,801	487,031	1.69	2.03	205	3,332	78	190	41	1
	0012	1	3,937,030	0.96	406,630	418,854	1.42	2.02	219	3,314	71	196	46	-
	0014	3	5,217,286	0.95	490,447	521,190	1.68	2.02	236	3,461	73	225	51	-
	0020	1	4,117,711	0.97	279,615	333,559	1.52	2.07	174	2,492	51	160	40	-
		Average	4,449,775	0.96	410,096	436,623	1.59	2.03	191	2,965	63	180	42	-
WES	0001	6	442,288	0.94	29,074	34,990	0.98	2.08	187	3,442	62	149	42	-
	0002	6	473,005	0.95	32,376	39,109	0.9	2.1	172	3,273	63	150	47	-
	0008	4	376,522	0.93	25,213	29,478	1.07	2.12	170	3,193	72	147	39	-
	0016	3	501,795	0.94	31,696	35,314	0.75	2.01	168	3,259	68	157	36	-
	0018	3	322,878	0.94	22,741	28,156	1.16	2.16	151	3,058	62	152	44	-
	0019	5	371,672	0.94	25,670	32,718	1.15	2.13	165	3,085	68	154	43	-
	0023	3	416,843	0.95	29,963	36,349	0.91	2.12	158	3,232	66	150	39	-
	0024	3	295,158	0.94	20,965	26,125	1.13	2.15	158	2,862	58	141	40	-
	0025	5	1,196,341	0.96	92,407	108,435	0.47	2.15	233	3,324	67	176	47	-
	0033	3	438,071	0.95	28,504	33,349	0.88	1.98	163	3,139	60	177	46	-
	0034	3	446,589	0.95	33,754	2,007	0.94	1.98	179	3,207	64	154	38	-
	0046	2	277,054	0.94	18,962	22,653	1.54	2.15	134	3,122	60	122	37	-
	0047	3	287,001	0.94	20,152	24,828	1.32	2.18	153	2,994	61	157	40	-
	0048	3	287,467	0.94	19,855	24,763	1.32	2.18	145	3,179	68	151	41	-
	0049	3	291,966	0.94	20,946	26,168	1.28	2.17	163	3,093	56	140	45	-
	0059	3	407,957	0.95	27,006	32,315	0.97	2.15	175	3,252	79	157	45	-
	0091	5	1,303,209	0.97	85,303	100,053	0.43	2.14	237	3,721	71	196	48	-
	0098	2	264,827	0.94	18,044	21,878	1.26	2.19	156	2,894	63	132	34	-
	0103	4	426,542	0.94	28,728	35,053	0.96	2.15	183	3,262	60	157	45	-
	0117	3	403,156	0.95	24,968	30,131	0.93	2.17	195	3,169	71	175	40	-
	0118	2	328,227	0.95	21,084	24,912	0.97	2.17	148	2,788	55	132	34	-
	0139	3	339,853	0.9	27,966	1,874	1.24	2.09	157	3,150	66	159	39	-
	0140	1	210,102	0.92	14,060	16,318	1.32	2.14	124	2,444	39	113	35	-
	0152	3	743,061	0.97	54,584	66,373	0.34	2.2	221	3,118	69	126	43	-
	0157	3	1,025,862	0.97	72,550	86,488	0.35	2.2	224	3,321	64	152	48	1
		Average	475,098	0.94	33,063	36,793	0.98	2.13	173	3,695	64	151	41	

This table shows the variant call summary of the family based whole genome and exome sequencing for all study participants presented in all chapters of this dissertation. In terms of quality metrics the most important QC metrics include the dbSNP rate and transition-transversion ratio (Ti/Tv).

For RNA-seq, on average, 96,5 million reads mapped to reference genome for a total of 90 RNA-seq libraries (Figure 18A) The smallest library size was 26.8 million reads up to the largest library size of 239.8 million reads. Refer to sample-by-sample RNA-seq metrics to Appendix C. The RNA content of the prepared libraries shows that on average 60.9% of bases were amino acid coding, 29.6% were UTR bases, 5.8% intronic, 3.7% intergenic, and 0.07% were ribosomal bases (Figure 18B). We found 14,441 protein coding genes with a median FPKM of greater than zero in all study participants. We found 190,219 exons in 18,378 protein coding genes with an exonBaseMean above zero and 133,752 exons in 14,055 genes with a median exonBaseMean of 1 (Table 17).

Table 17.

Expression estimates for protein coding genes and exons.

FPKM	0	0.001	0.01	0.1	1	5	10	50	100	500	1000
# genes	14,77	14,642	14,484	13,067	10,653	7,745	5,669	1,569	775	139	66
% genes	79.87	79.18	78.32	70.66	57.61	41.88	30.65	8.48	4.19	0.75	0.36
baseMean	0	0.001	0.01	0.1	1	5	10	50	100	500	1000
# exons	190,2	148,09	148,09	148,09	133,75	114,22	102,85	65,23	45,44	12,46	6,00
% exons	86.28	67.17	67.17	67.17	60.97	51.81	46.65	29.62	20.61	5.65	2.72
% genes	100.0	82.91	82.91	82.91	76.47	68.62	64.75	54.14	47.47	23.86	13.8

This table shows the number of protein coding genes and their percentage to the total at various expression levels estimated for genes (FPKM) and exons (baseMean)

We evaluated the gene expression correlation of protein coding genes to find outlier samples that may bias multivariate analysis (Figure 19A). Selection was done gene-by-gene those protein coding genes that showed expression >0 FPKM in at least one of the study participant (n=15,454). Correlation was also evaluated in the exon usage data. We selected exons in protein coding genes that were expressed in at least one of the patients with an exonBaseMean of >0 (n=123,945). Family 0002 patients were highly correlated as expected as exonBaseMean is calculated across conditions, and in each patient exon usage was compared to parents' exon expression (Figure 19B). Overall correlation across exons was Spearman's rho >0.9 suggesting high quality data.

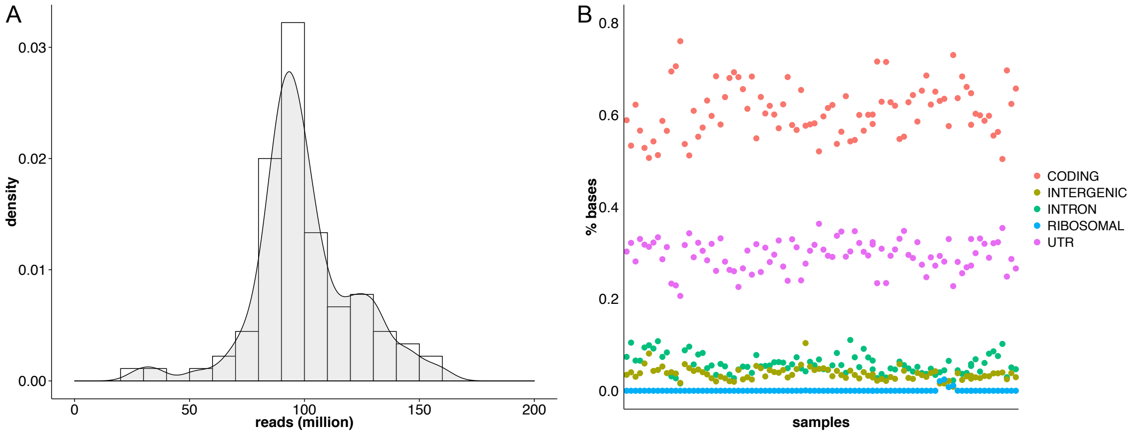

Figure 18. Mapping of RNA-seq data. A) Distribution of high quality sequenced reads across 90 individuals from RNA-seq presented in all 3 chapters. B) Proportion of bases sequenced across individuals in relation to their RNA contents. The horizontal axis shows each prepared samples and the vertical axis the % bases mapping to mRNA species.

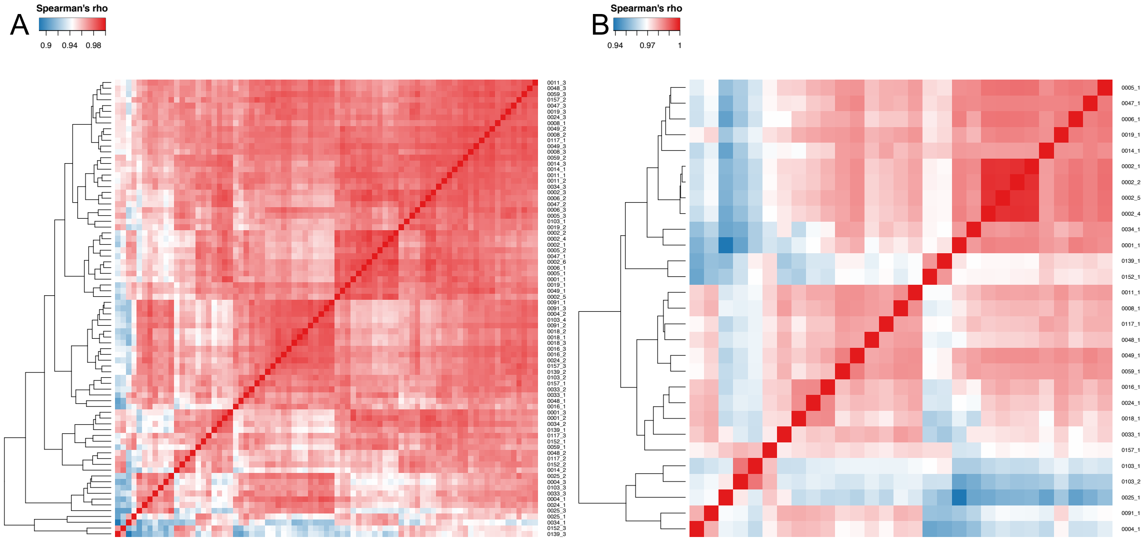

Figure 19. Sample correlations of gene and exon expression. A) Pairwise correlation of normalized FPKMs in protein coding genes across 25 families presented in this chapter for outlier analysis. The higher the correlation on a 0-1 scale the more red the cell's color. Each cell represents a comparison between two samples. B) Pairwise correlation of exonBaseMean between 25 families (29 patients). The more red the cell the higher the correlation. These plots show high correlation above 0.8 for gene based expression and >0.9 for exons, indicating that sequencing libraries were of good quality.

Overall assessment of RNA-seq found that 53.2% of protein coding genes (n=9,831) were expressed in at least 90% of samples above the detection limit of FPKM >= 0.1, and 64,1% of protein coding exons were expressed above the detection limit of exonBaseMean >= 1.0. This allows us to investigate over half of protein coding genes in whole blood for multivariate outlier testing analysis (Figure 20).

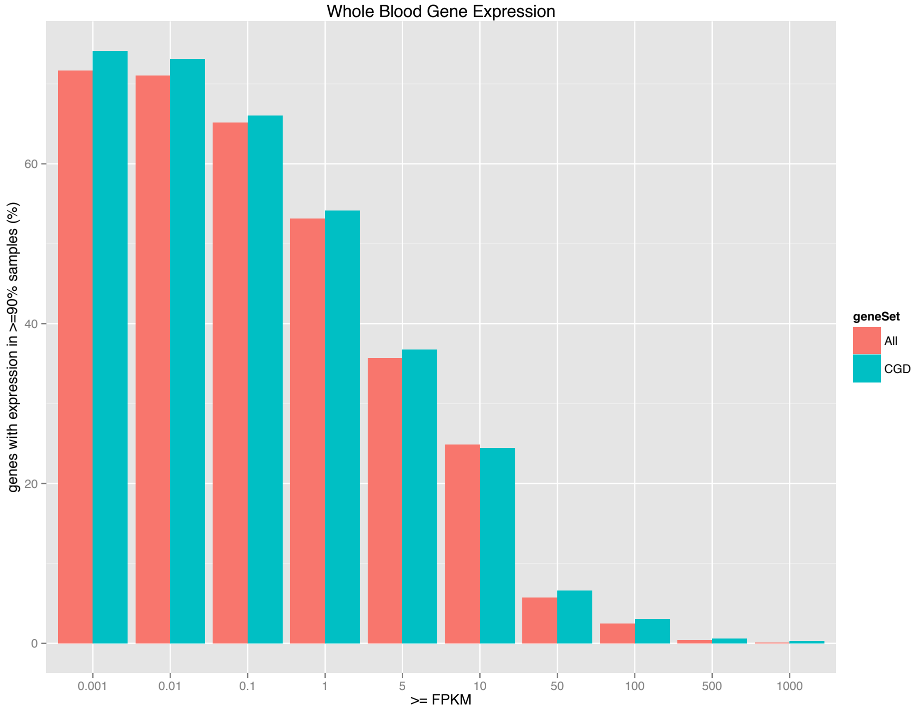

Figure 20. Transcript abundance of protein coding genes. The x-axis indicates FPKM thresholds and the vertical axis indicates the number of protein coding genes where >90% of participants had an FPKM above threshold. The orange color indicates all protein coding genes and the blue bars indicate genes in the Clinical Genomics Database. This plot shows that we could study over 60% of protein coding genes in whole blood above our detection limit of 0.1FPKM. When looking at genes that are known to cause disease we obtain a similar percentage at 0.1 FPKM.

Multivariate analysis by Mahalanobis distance. We selected RNA-seq data types of gene and exon expression as the basis for multivariate analysis because of the availability of relatively straightforward techniques to obtain their measurements, and their measurements could be used to correlate expression with in silico predictions of variant affect to gene function. Gene expression and differential gene expression are the most common methods to quantitatively and qualitatively study transcriptomic diversity and its relation to phenotype. These measurements are rapidly approaching their applications in clinical studies showing high correlation between sequencing platform and improved accuracy in their measurements (S. Li et al. 2014; Risso et al. 2014). Selection of exons during splicing can have a great impact of mRNA complexity and protein diversity in the cell. It is also suggested that over 95 percent of genes are spliced that leads to inclusion of different sets of exons in mRNA (E. T. Wang et al. 2008). In addition, exon skipping is the most common mechanism of alternative splicing occurring in over 38% of genes (de Klerk and 't Hoen 2015). DEXSeq provides exon-by-exon information on splice events without considering isoform complexity thus significant findings cannot be correlated with isoforms found in the tissue. Sulem et al. (2015) showed that 74% loss-of-function variants, including splice site, have effect on all transcripts of the gene.

Mahalanobis distance is a unitless, descriptive measure of relative distance of a data point from the centroid of the data distribution taking into account the correlation of each data point within the multivariate dataset by estimating the covariance (Mahalanobis 1936). Calculation of a covariance matrix is essentially a normalization method, thus Mahalanobis analysis does not requires input data to be normalized or scaled to a common scale. However, as a proof-of-concept we brought all multivariate to a common scale and used log transformed FPKM, exonBaseMean, and nDiff to match log transformed fold change from differential expression. Covariance estimation is a critical step in Mahalanobis analysis, because without covariance distances between multivariate would simply be the Euclidian distance that does not capture the relationship between the data points (De Maesschalck and Jouan-Rimbaud 2000). In addition, the large dynamic range of achieved RNA-seq experiments warrants data transformation to improve confidence in prognostic metrics and make data more amenable to analytical tools that assume normal distribution of input (Zwiener, Frisch, and Binder 2014; Risso et al. 2014). While our RNA-seq data does not follow normal distribution even after transformation, Kothari et al. (2013) reported successful application of Mahalanobis distance in a multivariate data space of gene expression and differential gene expression to identify outlier genes for the discovery of clinically actionable kinase gene targets in cancer therapies.

In order to integrate candidate variants we calculated outlier MD scores across 29 patients for genes and exons harboring rare variants by calculating the MD score for each candidate variant (post-filtering) and each individual. By example, a variant (GeneX Y555C) would have an associated outlier score such as 'MD score = 1.32'. In total 25,053 variants remaining after variant annotation and filtration for frequency were scored in 7,222 genes across the patient cohort. The number of variants scored can be found in Table 18. The scores ranged between 19.2558-0.0006 for genes and between 25.4-0.0009 for exons. Next we filtered our list to those genes with a single variant within patients and obtained 18,834 variants in 7,043 genes. The range of scores was 19.3-0.0001 for gene-based scores and did not change for exon-based scores (Table 18). The highest gene-based distance score was seen in 0002_4 (MD score = 19.3) and highest exon score was detected in 0103_1 (MD score = 25.4).

On average there was 863±200 variants in each patients with scores for both genes and their exons. The distribution of gene and exon distance scores indicated non-normal, right skewed distribution with very long tails for both scores suggesting that most variants have similar functional impact across patients (Figure 21). This follows our expectations that variants with functional impact that deviate from general tendencies will be rare and likely patient specific. We performed a bootstrap version of Kolgomorov-Smirnov test to find out if the gene and exon MD scores come from same distribution in the full and filtered dataset. Hypothesis testing is done with the null hypothesis that the two data sets come from the same distribution. Bootstrapping is performed by Monte Carlo simulations and allows for non-continuous data or data with many ties. Our data has many ties as many variants may have the same MD scores. Two-sample KS test indicated that gene scores and exon scores are not coming from the same distribution for full dataset and filtered dataset (full: Kolmogorov-Smirnov, D= 0.0491, P= 2.2e-16; filtered D=

0.0357, P= 7.735e-11). There are weak, linear relations between gene and exon distance measurements and overall exon scores tend to be higher than gene scores. (Spearman's Rank, S = 699004637637, P < 2.2e-16, rho= 0.37232) (Figure 22). Proportion of shared variance between the gene and exon ranked scores shows that little variance in one distance is explained by variance the other distance (R^2=0.1386) suggesting that the two distances capture different properties of the transcriptome.

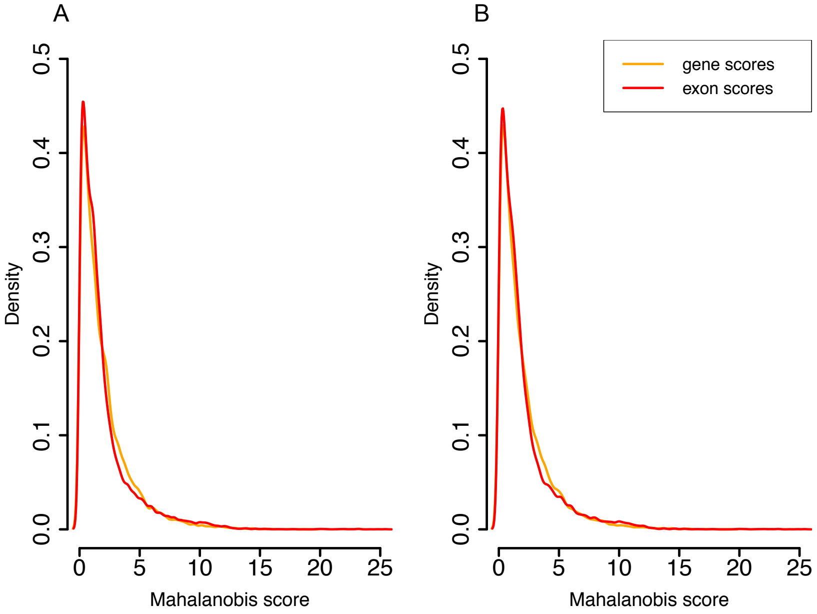

Figure 21. Distribution of MD scores. A. Scores for variants across 29 patients with both gene and exon scores n=25,053. B. Scores for variants in single hit genes n=18,834. In general we can see non-normal, very right skewed distributions. For both gene and exon scores indicating a very few MD scores with very high magnitude for the full datasets and the filtered variants as well.

Table 18.

Gene and exon measures for 29 patients.

Sample ID	# rare, coding variants	# variants with Gene Score	# variants with Exon Score	max/min Gene Score	max/min Exon Score	# of genes scored	# variants with Gene/Exon
001_1	1046	669	622	10.5/0.003	6.5/0.00009	475	594
002_5	1574	1070	928	11.5/0.0003	15.8/0.002	694	902
002_2	1574	1070	928	8.1/0.003	9.3/0.006	698	906
002_1	1574	1077	928	6.9/0.0001	12.1/0.001	699	907
002_4	1574	1077	928	19.3/0.002	5.3/0.003	698	906
004_1	524	333	314	16.0/0.007	19.8/0.05	243	302
005_1	863	537	447	6.1/0.002	12.0/0.01	368	434
006_1	1253	779	686	11.0/0.001	7.5/0.001	531	667
008_1	1979	1268	1143	13.6/0.0001	17.8/0.007	807	1085
011_1	2110	1271	1149	6.6/0.0001	9.7/0.005	792	1100
014_1	2354	1505	1469	12.9/0.05	10.3/0.004	818	1295
016_1	1994	1292	1955	11.7/0.002	11.8/0.002	566	1092
18_1	1553	1012	1512	17.3/0.03	21.5/0.006	661	845
19_1	1508	977	1468	7.2/0.0003	6.2/0.001	629	814
24_1	1499	960	1473	17.9/0.006	19.7/0.003	634	806
25_1	1426	944	1394	16.9/0.01	22.5/0.004	645	813
33_1	1758	1105	1002	10.3/0.00006	20.3/0.001	715	961
34_1	2022	1351	1153	13.5/0.005	14.4/0.0003	807	1126
47_1	1330	863	786	10.8/0.0003	5.3/0.0005	593	753
48_1	1476	899	842	10.9/0.0002	21.7/0.006	636	789
49_1	1319	871	786	8.5/0.004	7.9/0.0004	594	758
59_1	1851	1181	1057	11.3/0.0006	22.2/0.0006	780	1012
91_1	1286	818	738	10.5/0.0002	20.1/0.0002	562	696
103_1	1731	899	960	15.8/0.0003	25.4/0.3	646	831
103_2	1731	952	960	13.5/0.003	20.3/0.09	757	861
117_1	1723	1065	1000	12.6/0.0008	23.2/0.002	836	944
139_1	1610	1085	985	15.8/0.003	13.3/0.001	684	947
152_1	1545	1003	916	17.7/0.0005	14.0/0.003	662	882
157_1	1823	1173	1068	9.6/0.0004	12.4/0.005	848	1025

This table indicates the range of scores from minimum to maximum for each patient. The last column indicates the number of variants successfully evaluated for both gene-based and exon-based outlier analysis.

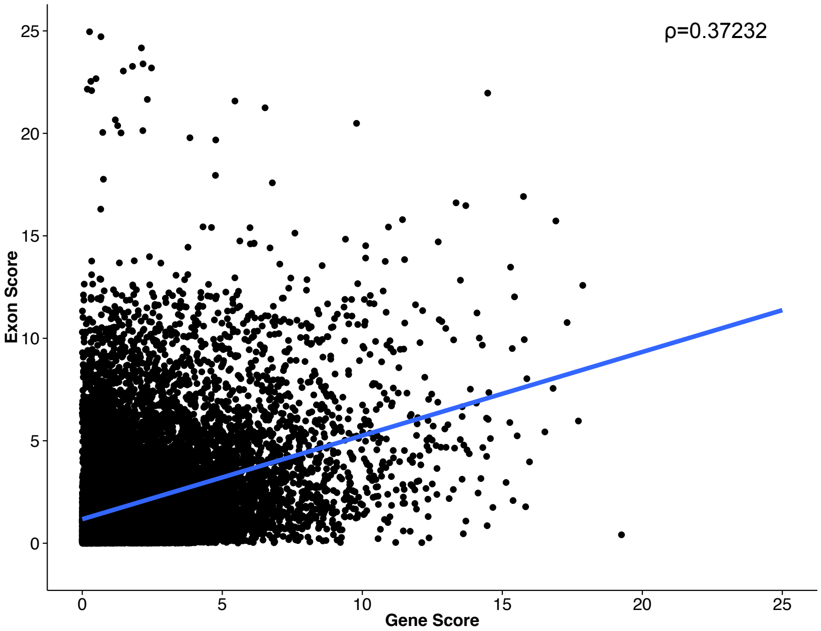

Figure 22. Gene and exon distance correlations. Black dots indicate variants (n=18,834) and their associated gene and exon scores in the 29 patients. There is a general trend of higher exon scores. Linear regression suggests a linear relationship between the two scores with a correlation coefficient of 0.37 indicating a medium effect size.

Next we evaluated the distribution of MD scores with relation to the variants' functional impact. We defined two variant classes high, and moderate-to-low. Based on available variant annotation, high functional impact variants predicted to cause frameshift, start codon loss, stop codon gained, stop codon lost, splice donor, and splice acceptor changes. The moderate-to-low functional impact variants were defined as missense, insertions-deletions, splice region, synonymous, start codon gain, and sequence feature consequences. Levene's test for equal variance within the functional classes revealed that gene scores have the same variance in both full dataset (Levene's Test, F=1.1541, P=0.2827) and filtered dataset (Levene's Test, F=0.3218, P=0.5705). We found that exon scores in the filtered dataset had unequal variances (Levene's Test, F=5.0929, P=0.02403), but in the full dataset variance between functional classes was equal (Levene's Test, F=0.0052, P=0.9424). Levene's Test tests for one of the assumptions of Mann-Whitney rank sum test, which is equal variances of the data distributions. Gene and exon scores indicate that distributions of distance scores for HIGH and MOD-LOW impact variants are similar (Figure 23). Based on the median ranks of gene-based scores we find that variants predicted in high functional impact class have higher distance scores in the full variant list (Mann–Whitney *U* ,U =7492668, P=0.006378,two-tailed) and among filtered variants as well (Mann–Whitney *U* ,U = 3375335, P= 0.001005,two-tailed) (Figure 24). Exon scores are not associated with functional class in either the full or filtered variant list (full list: Mann–Whitney *U* ,U =3186404, P=0.1741,two-tailed, filtered: Mann–Whitney *U* ,U =7176454, P=0.378,two-tailed). The median MD scores for genes were highest for variants with high functional impact (1.57) compared to 1.28 for moderate-to-low impact variants (Table 19). Results indicate that high impact functional variants have larger MD scores than variants with low moderate-to-low functional class based on overall gene scores. Lack of difference between the two functional classes among exon scores can be attributed to cryptic splice sites and alternate exon usage impacted by missense variants (Ahlborn et al. 2015).

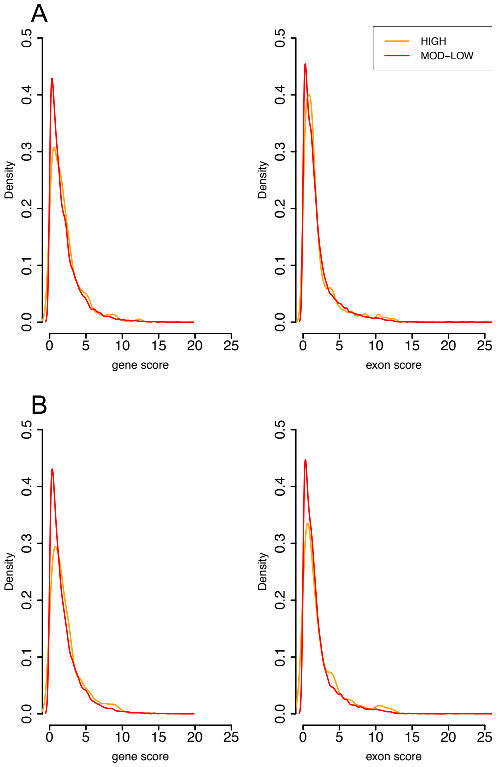

Figure 23. Distribution of MD scores within functional classes. Horizontal axis indicates the gene and exon scores and the vertical axis shows the density distribution. The orange line shows scores for variants in moderate-to-low impact functional class and yellow indicates variants in high impact functional class. The distribution of these scores also follows a non-normal, right skewed distribution.

Table 19.

MD scores for high and moderate-to-low functional variants.

		Single Hit Genes				Full Dataset			
		Gene Score		Exon Score		Gene Score		Exon Score	
		High	Mod-Low	High	Mod-Low	High	Mod-Low	High	Mod-Low
Variants (n)		330	18,504	330	18,504	574	24,479	574	24,479
Mean		2.26	1.96	2.25	1.98	2.13	1.97	2.02	1.95
Median		1.57	1.28	1.28	1.22	1.52	1.29	1.18	1.20
Std.Dev		2.17	2.11	2.58	2.35	2.11	2.37	2.10	2.32
Min		0.0018	0.0001	0.0054	0.0001	0.0018	0.0001	0.0053	0.0001
Max		12.66	19.26	12.61	25.45	12.66	19.26	12.81	25.45
Percent	25	0.73	0.51	0.53	0.51	0.62	0.58	0.53	0.51
	50	1.57	1.28	1.28	1.22	1.52	1.18	1.29	1.20
	75	2.92	2.91	2.63	2.42	2.78	2.27	2.63	2.39
	95	6.97	8.01	6.17	6.80	6.41	7.67	6.19	6.77

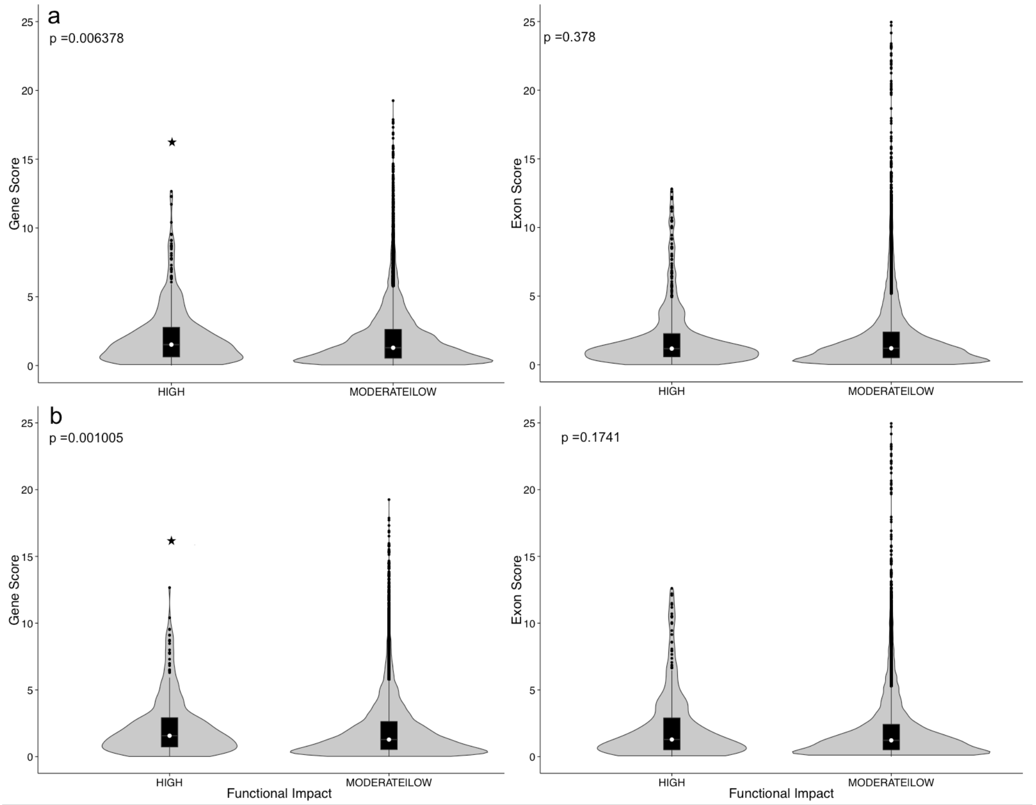

Figure 24. Variants in different functional class and their MD scores. Violin plots of the MD distance scores based on functional class (i.e HIGH, MODERATE-TO-LOW). a=includes all scored variants across the 29 patients. b=scores for single hit genes across 29 patients. Star above plot indicates statistical significance by Mann-Whitney test. The exon scores in the MOD-TO_LOW group have much longer tails indicating large variance within the dataset. Values in HIGH functional class have much shorter tails suggesting that variants with prediction of high functional impact are more likely have a more uniform behavior.

Analysis of functional effect for variants with high MD scores

After evaluating variants based on their predicted functional impact we set out to study MD scores based on their position and predicted impact on the mRNA structure. We selected genes with single coding variant for each patient. Based on available annotation, we defined 9 functional classes of variants as frameshift (SnpEff=frameshift), insertions-deletions (SnpEff=disruptive inframe insertion, disruptive inframe deletions, inframe deletion, inframe insertions) missense (SnpEff=missense), sequence feature (SnpEff=sequence feature), splice site (SnpEff=splice acceptor, splice donor), splice region (SnpEff=splice region), start|stop

(SnpEff=start lost, stop gained, stop lost), synonymous (SnpEff=synonymous, stop retained), start codon (SnpEff= start codon gain, initiator codon). Non-parametric analysis of variance of gene and exon scores across the nine functional classes shows that distribution of scores are from different distributions with different means and medians. (gene scores: Kruskal-Wallis, H=15.604, p=0.05, exon scores: Kruskal-Wallis, H=17.584,p=0.02) (Figure 25). Splice site, frameshift and nonsense variants have the highest MD scores for genes on average suggesting that they may be related with respect to their impact on transcriptional activity (Table 20). Significant dysregulation of transcripts by nonsense variants, especially variants causing nonsense-mediated decay has been demonstrated in unaffected populations (MacArthur et al. 2012).

Pairwise comparisons of MD scores by Dunn's test reveals that frameshift variants impact transcription at highest degree among the 29 patients showing statistical difference from indels (P=0.01), missense (P=0.002), sequence feature (P=0.001), splice region (P=0.009), start gained (P=0.004), and synonymous variants (P=0.003) (Table 21). Frameshift variants (n=175) are more likely to have higher Mahalanobis distances suggesting that they impact "outlierness", although this difference is not seen when frameshift variants are compared with splice site and start|stop gained or lost variants (Figure 25A). High confidence, loss of function variants resulting in the shift of the open reading frame have been implicated as most likely loss-of-function variants (MacArthur et al. 2012). Overall, in our dataset, the greatest difference to the effect genetic variants have on transcription is between frameshift and sequence feature variants (Bonferroni P=0.024).

The distribution of distance scores for exon usage is more evenly distributed indicating that alternative exon usage is not the function of a single variant type (Figure 25B). Pairwise comparison of MD scores shows that frameshift, missense, sequence feature, start|stop and synonymous variants in exons are more likely impact exon usage than inframe indels in exons (Frameshift-VS-indels P=0.019, missense-VS-indels P=0.019,sequence feature-VS-indel P=0.020,start|stop-VS-indels P=0.019, synonymous-VS-indels P=0.009). Interestingly, synonymous variants show significant difference from sequence features (P=0.003), and from splice site variants (P=0.030). Synonymous variants in exons may lie in exonic splice enhancers and they can impact alternative splicing and protein function (Rice et al. 2013; Sheikh et al. 2013). Our results also suggest that authentic splice site mutations are not necessarily accompanied by alternative exon usage and if exon usage occurs they may have similar functional impact across patients.

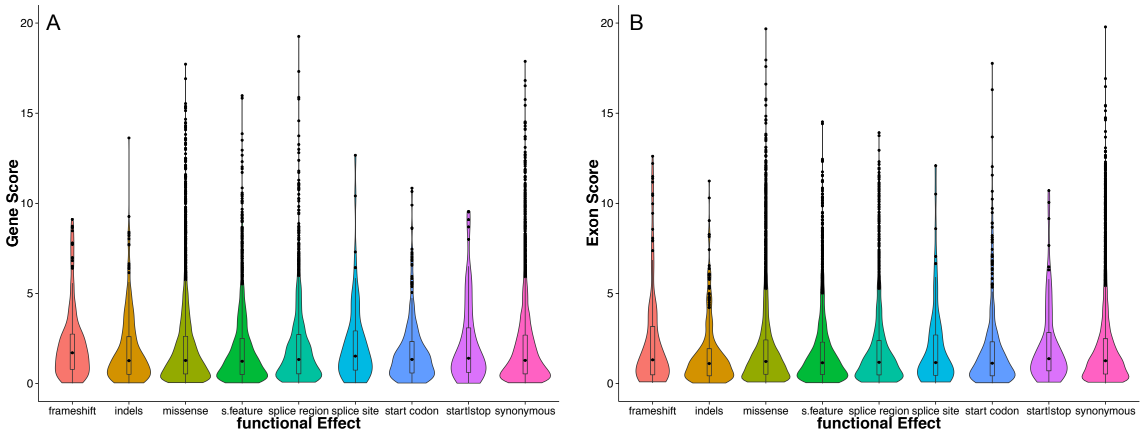

Figure 25. MD scores and functional effect. A=gene scores, B=exon scores. There are nine functional effect groups, each colored differently. Overall all groups show a non-normal right tailed distribution. Exon scores are more uniform than gene scores. Gene scores show that frameshift variants have a higher median MD scores than other groups (frameshift median = 1.703).

Table 20.

MD scores and functional class.

Gene Score

	#/class	Min	Median	Mean	Max	25%	50%	75%	95%
frameshift	175	0.0018	1.703	2.228	9.111	0.785	1.703	2.739	6.967
indels	280	0.0012	1.267	1.901	13.62	0.511	1.267	2.599	5.657
missense	6,823	0.0003	1.277	1.975	17.72	0.529	1.277	2.619	6.169
seq. feature	2,390	0.0001	1.233	1.906	15.97	0.498	1.233	2.504	6.294
splice site	70	0.0172	1.518	2.243	12.66	0.741	1.518	2.914	6.164
splice region	1,669	0.0003	1.331	2.057	19.26	0.536	1.331	2.715	6.626
start gained	344	0.0068	1.336	1.776	10.84	0.580	1.336	2.330	5.534
start|stop	84	0.0038	1.400	2.327	9.552	0.618	1.400	3.084	7.698
synonymous	6,999	0.0001	1.283	1.965	17.87	0.532	1.283	2.684	6.105

Exon Score

frameshift	175	0.034	1.312	2.351	12.61	0.481	1.312	3.164	8.825
indels	280	0.015	1.107	1.667	11.24	0.427	1.107	1.933	5.785
missense	6,823	0.000	1.227	1.993	24.17	0.512	1.227	2.421	6.884
seq. feature	2,390	0.001	1.152	1.859	25.45	0.505	1.152	2.301	6.211
splice site	70	0.003	1.176	1.963	22.67	0.474	1.176	2.389	7.067
splice region	1,669	0.005	1.161	2.074	12.08	0.454	1.161	2.688	6.875
start gained	344	0.010	1.14	2.044	24.96	0.426	1.140	2.341	7.782
start|stop	84	0.057	1.376	2.177	10.7	0.703	1.376	2.831	6.435
synonymous	6,999	0.000	1.262	2.009	25.43	0.524	1.262	2.493	6.876

25%,50%,75%,95%= the percentile cutoff value.

Table 21.

Dunn's test of pairwise comparisons of functional classes.

Gene

	frameshift	indels	missense	seq. feature	splice site	splice region	start gained	start|stop	synonymous
frameshift		0.351	0.086	0.027	1	0.329	0.147	1	0.106
indels	0.010		1	1	1	1	1	1	1
missense	0.002	0.441		1	1	1	1	1	1
seq.	0.001	0.357	0.088		1	1	1	1	1
splice site	0.318	0.119	0.108	0.068		1	1	1	1
splice	0.009	0.280	0.148	0.028	0.162		1	1	1
start	0.004	0.399	0.296	0.482	0.087	0.163		1	1
start|stop	0.284	0.114	0.099	0.059	0.479	0.157	0.080		1
synonymo	0.003	0.407	0.378	0.057	0.116	0.197	0.264	0.108	

Exon

frameshift		0.683	1	1	1	1	1	1	1
indels	0.019		0.692	1	1	1	1	0.691	0.328
missense	0.168	0.019		0.714	1	1	1	1	1
seq.	0.059	0.110	0.020		1	1	1	1	0.090
splice site	0.089	0.076	0.110	0.315		1	1	1	1
splice	0.268	0.200	0.455	0.386	0.436		1	1	1
start	0.072	0.212	0.132	0.411	0.316	0.357		1	1
start|stop	0.335	0.019	0.116	0.052	0.071	0.186	0.056		1
synonymo	0.232	0.009	0.150	0.003	0.030	0.397	0.075	0.151	

Columns and rows indicate functional class. Clear cells indicate raw p values of the test statistic and grey columns show the Bonferroni correction p –values for multiple testing from Dunn's test. Top table indicates pairwise analysis of gene scores, lower table shows pairwise comparisons of exon scores. Red values show p<0.05 significance.

Patients with known causal variants.

In this section we will show result of Mahalanobis scores from RNA-seq. integrated with genomic variants. Integration was performed for each patient and for each rare variant that remained after filtration described above in the Variant Annotation section of the Material and Methods section.

We investigated 10 patients with genetic diagnosis prior RNA-seq and multivariate analysis. Each patient had a presumed causal variant for a total of 7 genes (Table 22). In two families multiple affected siblings were diagnosed with a presumed causal variant (0002, 0103). In patient 0001_1 the presumed causal *DDC* gene had an average normalized FPKM of 0.07 under the detection limit of this study so no outlier analysis could be preformed. The average abundance for the remaining 6 causal genes was FPKM=26.78. In two patients the source of causal variation was compound heterozygous mutations (0005_1, 0049_1), three patients *de* *novo* variants contributed to disease (0047_1, 0103_1, 0103_2), one patient had a autozygous variant (0024_1) and 4 patients from families 0001 and 0002 presented causal variants that did not follow Mendelian inheritance. For each patient specific variant we calculated the percentile rank of the gene and exon distance measurement associated with the variant. In three patients the presumed causal variant ranked in the 95th percentile, and for 4 patients ranked in the 90th suggesting that presumed causal variants show elevated impact on gene regulation.

In general, distance measures for genes range from 48th percentile in family 00024 to the top ranked gene based distance score in family 0002. The casual variant ranked 48th percentile is a homozygous missense variant and the top ranked variant is a missense variant in family 0002. In family 0103 the siblings share a *de novo* variant but their ranks differ from 83rd percentile for 0103_1 compared to 96th percentile for 0103_2 indicating that difference in transcript regulation.

MD scores for exons show a greater variance from the 4h percentile in sample 0047_1 to ranking at the top in patient 0002_5. In three patients (0049_1, 0002_5, 0002_2) splice region, frameshift and cryptic splice site variations resulted in both exon and gene based distance scores in the 90th percentile. This suggests that these variants have a role in alternative splicing. In cases where gene based distance scores are not accompanied with high exon based distance scores suggests that those variants are silent to alternative exon usage, however they may negatively impact mRNA stability which is captured by a gene based distance score that is an outlier when compared to the other patients.

Functional importance of our findings were supported by *in silico* predictions of high conservancy by phyloP and moderate deleteriousness by CADD as listed in Table 22 (Siepel et al. 2005; Kircher et al. 2014).

Table 22.

Patients with known causal variants.

Patient	chr:pos	cDNA	Gene	Exon Rank	Mean FPKM	effect	gt	phyloP	phyloP Pct	phyloP R/T	CADD	CADD Pct	CADD R/T	Gene Score	Gene Pct	Gene R/T	Exon Score	Exon Pct	Exon R/T
0001_1	7:50571690	c.781+1G>C	DDC	I8/15	0.07	splice donor	0|1	5.40	0.8	90/454	17	0.79	102/454	-		-	-		-
0002_5	15:65313871	c.626C>T	MTFMT	4/9	6.6	missense	1|1	8.40	0.96	15/355	17	0.74	93/355	11.51	1.00	1/902	15.79	1.00	1/902
0002_2							1|1	8.40	0.96	15/355	17	0.74	93/355	5.23	0.99	4/902	4.06	0.99	9/902
0002_4							0|1	8.40	0.96	15/355	17	0.74	93/355	0.92	0.52	524/902	0.44	0.11	437/902
0005_1	15:89860752	c.3483-4_3497delTCAGGTGCATGTTTGCCTA	POLG	22/23	38.7	splice acceptor	0|1	-		-	-		-	1.1	0.65	153/434	0.40	0.11	389/434
	15:89866657	c.2243G>C		13/23		missense	0|1	-		-	-		-	1.1	0.65	153/434	-		-
0024_1	13:51503701	c.227A>T	RNASEH2B	3/10	24.8	missense	1|1	3.55	0.71	85/291	18	0.87	42/291	1.8	0.48	424/806	1.89	0.67	269/424
0047_1	19:17338695	c.499G>A	OCEL1	4/6	22.3	missense	0|1	5.57	0.79	66/315	15	0.72	107/317	1.1	0.62	289/753	0.04	0.04	726/753
0049_1	19:45856553	c.1703_1704delTT	ERCC2	18/23	7.4	frameshift	0|1	-	-	-	-	-	-	2.5	0.93	52/758	3.68	0.98	18/758
	19:45867589	c.719A>G		9/23		splice region	0|1	5.30	0.82	50/270	15	0.74	142/272	2.5	0.93	52/758	3.08	0.97	26/758
0103_1	X:48933095	c.758T>C	*WDR45*	10/12	60.9	missense	1	8.60	0.95	17/360	18	0.84	56/344	5.8	0.83	144/896	-	-	-
0103_2							0|1	8.60	0.95	17/343	18	0.84	57/344	4.8	0.96	40/896	-	-	-

In the following we discuss two cases where we present the utility of our approach in a case where casual mutation was known prior RNA-seq and multivariate outlier analysis, and another case where candidate variant was identified after DNA-RNA integration and outlier analysis.

The first family (0002) is a Caucasian family of six with three affected siblings and one unaffected sibling (Figure 26A). The genetic diagnosis in *MTFMT* took several years and without knowledge through extensive functional characterization of a cryptic splice-site, the functional importance of the causative variant would have been unknown. In many ways, more efficient identification of the causal variant in this case is the goal of our outlier analysis.

Within the family, patient 0002_1 was described as unaffected born in 2002, then started complaining about migraine headaches at age 9. Overall, this patient had an unremarkable phenotype.

Patient 0002_2 was suspected with Wolff-Parkinson-White Syndrome (MIM:194200) and with mitochondrial encephalomyopathy at time of enrollment. She has short stature, which was treated with growth hormone. She has learning disability, attention deficit disorder. She has cardiac conduction defect that is stable and without episodes of tachyarrhythmia. She has exercise intolerance and can walk at most 0.25 mile before getting tired. She has amblyopia and wears eye glasses. She is weak, has hyperflexible ankle joints, which was stabilized. She is stuttering that is suggestive of Tourette's Syndrome (MIM:137580). She had cardiac catheterization, ablation procedure, tonsillectomy and adenoidectomy to improve sleep. Her molecular tests showed elevated plasma and CSF lactate. She has cerebral folate deficiency, which is treated with leucovorin. She has decreased methyltetrahydrofolate level.

Patient 0002_4 is a male patient with a suspected mitochondrial disorder. He presented with headaches at 8-9 years of age. He has been having hemiplegic migraine since age 15. His molecular tests showed elevated plasma lactate. He has normal plasma amino acids, plasma lactate, CSF amino acids, CSF lactate, CSF neurotransmitters, neopterin, tetrahydrobiopterin, and methyltetrahydrofolate. He presented with normal cardiac function, and ophthalmological pulmonary evaluations were normal. His immunohistochemistry is normal. His skeletal muscle enzymology shows reduced activity of mitochondrial oxidative phosphorylation (OXPHOS) complexes I and III; abnormal high-resolution spirometry on cultured fibroblasts. Genetic testing for mtDNA deletions, *KCN1A*, *CACNA1A* gene testing are negative.

Patient 0002_5 is an affected male. Clinical diagnosis at time of enrollment was Mitochondrial encephalomyopathy with suspected Leigh Syndrome (MIM:256000). He has developmental delay and a coordination disorder. He presented with expressive language disorder with dysarthric and delayed speech. He has small stature. He is hyper with short attention span. Complex I deficiency is likely; decreased ND6 subunit was observed on skeletal muscle biopsy. His MRI showed symmetric frontal white matter (pericallosal) lesions and symmetric basal ganglia lesions. EKG and ECHO of heart showed no evidence of cardiac disease. Areas of T2 signal abnormality involving the genu of the corpus callosum extending into the bifrontal white matter with additional lesions located within the inferior left putamen and bilateral subthalamic regions. A small area of patchy enhancement involves the genu of the corpus callosum and restricted diffusion is noted around the margins of this dominant lesion centered in the genu of the CC and the left inferior putamen lesion. Ophthalmic tests showed pale optic nerves. MR Spectroscopy of the brain showed a large lactate peak over normal appearing right basal ganglia. He is normal for CSF 5' pyridoxal phosphate, CSF succinyladenosine, CSF neurotransmitters, ceruloplasmin, plasma amino acids, urine organic acids, and urine mucoplysaccharides. Lysosomal storage panel and tests for disorders of glycosylation was normal. He has abnormally low CSF 5-methyltetrahydrofolate, high CSF lactate, high CSF alanine. He has significantly increased myofiber lipid with unremarkable immunochemistry but showing complex I defect. He has normal muscle levels of Coenzyme Q10. Genetic testing for *PDHA1* gene mutations and for mtDNA point mutations and deletions tests was negative.

We sequenced the exome and mRNA of the entire family 0002. Exome sequencing achieved an average target coverage of 87.7X across targeted regions. Bioinformatics analysis identified 473,005 SNP and short indel variants with 95% of them reported in dbSNP141 and with a Ti/Tv ratio of 2.1 (Table 16). A total of 1574 protein coding variants in 1326 genes had a frequency of <5% and were evaluated for their impact on expression by Mahalanobis distance.

Variant prioritization identified a missense variant in exon 4 of *MTFMT* gene (NM_139242.3, c.626C>T, p.Ser209Leu). Patient 0002_5 and 0002_2 were homozygous for this variant and all other family members, including parents were heterozygous. (Figure 26C). The identified variant was verified by Sanger sequencing (Figure 26B). This variant is known pathogenic variant reported by Tucker et al showing that heterozygous mutation result in a frameshift and premature stop codon by skipping exon 4 during pre-mRNA processing in patients with Leigh Syndrome (Tucker et al. 2011). This finding corroborated suspicion of Leigh Syndrome in 0002_5 and resulted in the diagnosis in the affected patients although the phenotypic heterogeneity was noted across siblings. Patient 0002_5 phenotype showed similarity with reported cases and thus he was diagnosed with Leigh Syndrome. In addition Haack et al later reported that the exon 4 mutation was one of the most frequent mutations in defects of mitochondrial oxidative phosphorylation (OXPHOS) (Haack et al. 2014).

Analysis of RNA-seq reads supports exon skipping in the homozygous patients, with most reads spanning the exon 3-4 and exon 4-5 boundaries (Figure 26C). RNA-seq read data supports that heterozygous family members express a transcript in whole blood that includes exon 4 as IGV traces show reads mapping in exon 4. Differential gene expression analysis between each patient and their parents shows that *MTFMT* is more dysregulated among the homozygous patients then heterozygous patients (0002_5 p value = 0.00455, 0002_2 p value= 0.03815, 0002_1= 0.59625, 0002_4= 0.47585). Alternative exon usage analysis corroborates the prediction of exon skipping with the two homozygotes suggesting differential usage of exon 4. Taken these two RNA-seq analyses together and applying gene abundance and exon analysis by multivariate approach shows that gene-based and exon-based scores correlate with severity of phenotype and for zygosity. Patient's 0002_5 was most severely affected and *MTFMT* gene and exon score and the variant in exon 4 had the largest MD scores among all identified rare variants (Figure 27A and Table 22). Interestingly his homozygous female sibling, 0002_2 showed a similar exon-skipping pattern by RNA-seq reads, although MD scores indicated that variants in other genes and exons had greater transcriptional impact (Figure 27B). This variability may be indicators of false positives, or suggestive of the effect of other variants that lead to phenotypic heterogeneity presented among patients with the same presumed causal variant. The heterozygous siblings show similarly diminished impact of heterozygous *MTFMT* mutation to its transcriptional profile implicating the role of other genetic variants in clinical symptoms of mitochondrial condition with *MTFMT* variant (Figure 27C, D).

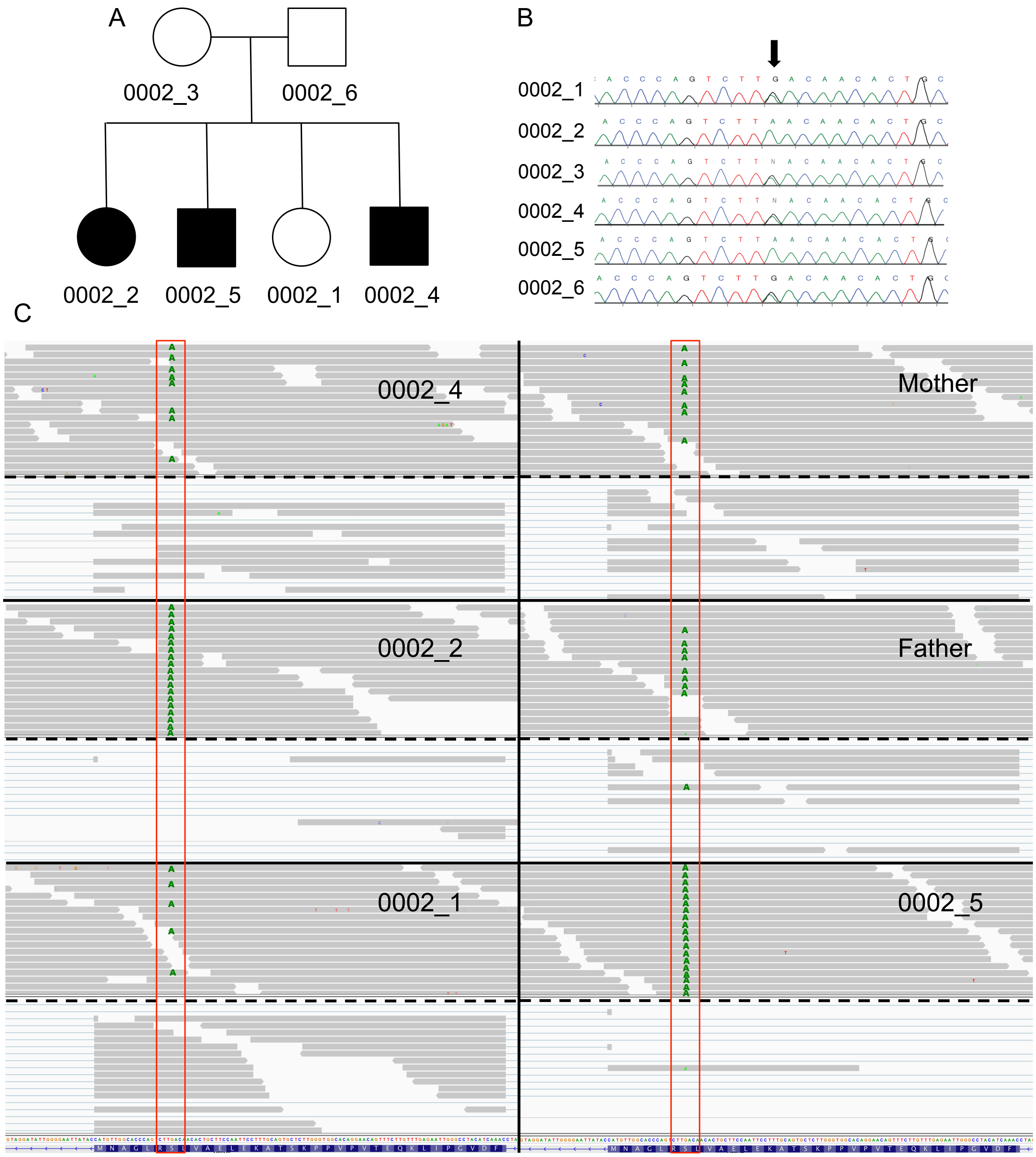

Figure 26. Family sequencing of the *MTFMT* variant. A=pedigree of family 0002. B=Sanger verification of the causal variant. C= Next-generation sequencing traces of exome and RNA-seq experiment for exon 4 including the missense variant. Solid black lines separate the traces for each family member. The red rectangle highlights the position of the causal variant with respect to exon 4. The exon can be seen by the blue horizontal bar at the bottom of the plot. For each family member the image is divided by dashed, black line. The traces above the dashed lines indicate the exome reads and the track under the dashed lines indicates the RNA-seq reads.

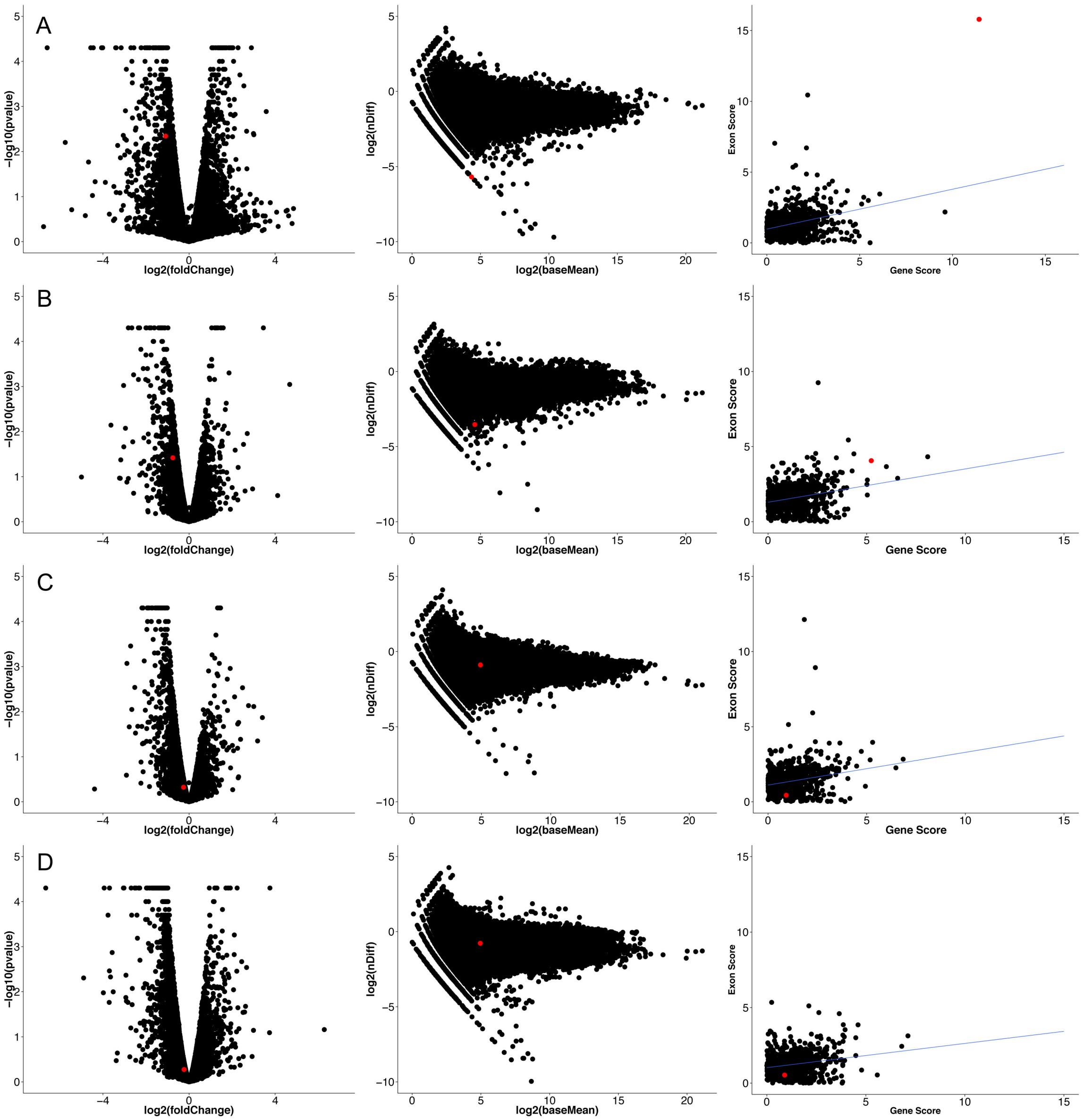

Figure 27. Expression profile of the *MTFMT* gene in family 0002. This plot shows the expression of *MTFMT* across the four siblings. A=0002_5, B=0002_2, C=0002_4, D=0002_1. The first column shows results of Cuffdiff differential expression for protein coding genes. The horizontal axis shows log2 foldchange, and vertical axis is the negative log10 of the probability that the gene is significantly dysregulated between the conditions. The second column shows the results of differential exon usage analysis. The horizontal axis is the log2 normalized exon coverage for exons with normalized coverage >1. The vertical axis shows the log2 normalized coverage difference between the patients and the parents. The third column shows a scatterplot of the gene-based and exon-based MD scores for the each rare variant. The red dot in each plot indicates the *MTFMT* gene in relation to all other genes in the analysis. Rows A and B are shows the siblings with the homozygous genotype for the *MTFMT* variant.

The second family we describe was family 117 with a single affected male. In this case, the genetic basis was not known prior to using the MD score and the candidate variant failed to be prioritized to a high enough level to warrant a genetic diagnosis. Effectively, our databases indicated this patient as undiagnosed at the time of analysis. Subsequent review of the initial genetic analysis prior to RNA-seq indicated some in the analysis group did view this as a good candidate, but the large number of other variants and other patients led to a failure to detect what on new inspection became a plausible causal variant for the genetic basis of the child's disease.

Clinical diagnosis at enrollment was suspected Pelizaeus–Merzbacher-like disease with no candidate genes identified. The patient presented with nystagmus, hypotonia, delayed development. The patient had limited speech to about 5 words, but could use signs. He was characterized with a leukodytrophy or significant dysregulation of the myelin sheet that protects nerve cells. His MRI scans showed diffuse lack of myelination of subcortical white matter, but with time some improvement, especially in the genu of corpus callosum; atrophy of splenium of corpus callosum. His urine organic acids test was negative. Genetic testing of *PLP1*, *GJA12*, *CDG* screening was negative. His CT scan was negative for calcifications.

We sequenced the exome and mRNA of the entire family 117. Exome sequencing achieved an average target coverage of 93.2X across targeted regions. Bioinformatics analysis identified 403,156 SNP and short indel variants with 95% of them reported in dbSNP141 and with a Ti/Tv ratio of 2.17 (Table 16). A total of 1723 protein coding variants in 1458 genes had a frequency of <5% and were evaluated for their impact on expression by Mahalanobis distance. Post integration of MD scores with genetic variants 944 variants from 836 genes were update with both gene-based and exon-based distance scores. Next the gene and exon based scores were ranked for all 944 variants. Gene-based ranking revealed a compound heterozygous variant in the *SNAP29* gene. The two variants rank #2 (MD Gene score= 12.29) and the two exonic scores ranked #107 of 944 scored variants (MD Exon score= 3.98), and #2 of 944 scored variants (MD Exon score= 12.81) respectively. The first variant was predicted to cause a loss of start codon in exon 1 (NM_004782, c.2T>C, p.Met1?) had a CADD score of 18, a phyloP of 5.05 and a genomic coverage of 32X. The second variant is a predicted loss-of-function, frameshift insertion in exon 2 (NM_004782, c.348_349insG, p.Gly118fs) had a genomic coverage of 76X in the patient (Figure 28). The mutations were not observed in the Exome Aggregate Consortium's over sixty thousand unrelated exomes (Exome Aggregate Consortium) *SNAP29* was classified as autosomal recessive disease causing gene in the Clinical Genomics Database. Further evaluation revealed that mutations in *SNAP2*9 are known to cause Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma, CEDNIK Syndrome (MIM:609528). CEDNIK syndrome was first described by Sprecher et al in two consanguineous families of Middle Eastern descent with homozygous frameshift deletion (Sprecher and Ishida-Yamamoto 2005). To date, patients with CEDNIK syndrome have been reported to carry homozygous frameshift insertions or deletions resulting in premature termination of the protein (Sprecher and Ishida-Yamamoto 2005; Fuchs-Telem et al. 2011). Common clinical manifestations of the disorder are roving eye movement, hypotonia, and malformation of the corpus callosum, neuropathy, microcephaly, facial dysmorphism, ichtyosis, and keratoderma. Hemizygous loss of function mutations in *SNAP29* in patients from non-consanguineous parents diagnosed with 22q.11.2 deletion syndrome show some overlap with symptoms of CEDNIK patients (McDonald-McGinn et al. 2013). The patient in this study is from a non-consanguineous family, carrying compound heterozygous variants (Figure 28C). Exon 1 mutation, on chr22:21213400:T>C was inherited from the father and is a predicted loss-of-function variant (Figure 28B). Exon 2 mutation, of 22:21224735:T>TG was inherited from the mother and is also predicted loss-of-function variant with a premature stop codon 16 amino acid resides downstream from the frameshift insertion (Figure 28A).

Sequenced reads from RNA-seq support the loss of maternally inherited transcript because the patient only express the paternal allele in exon 1 and the maternal insertion is not found in exon 2 track (Figure 28C). The mother also lacks reads that map to the insertion suggesting that the insertion is a loss-off-function variant in whole blood. This is supported by previous findings that frameshift insertions and deletion lead to truncated protein product (McDonald-McGinn et al. 2013; Fuchs-Telem et al. 2011). Multivariate outlier analysis suggested the importance of compound heterozygous mutation in *SNAP29* as gene-based MD score was ranked #2 on variant list (Figure 29). Differential gene expression does not support significant dysregulation of SNAP29 in whole blood when parents and affected patient are compared (p-value = 0.59, Cuffdiff). Exon based scores also ranked on top of the MD score list, however sequencing traces show no evidence of exon skipping. Frameshift mutation in exon 2 suggested a premature stop codon, therefore a high MD scores for exonic variants can be indicative of alternate exon usage, and in essence mRNA degradation by non-sense mediated decay.

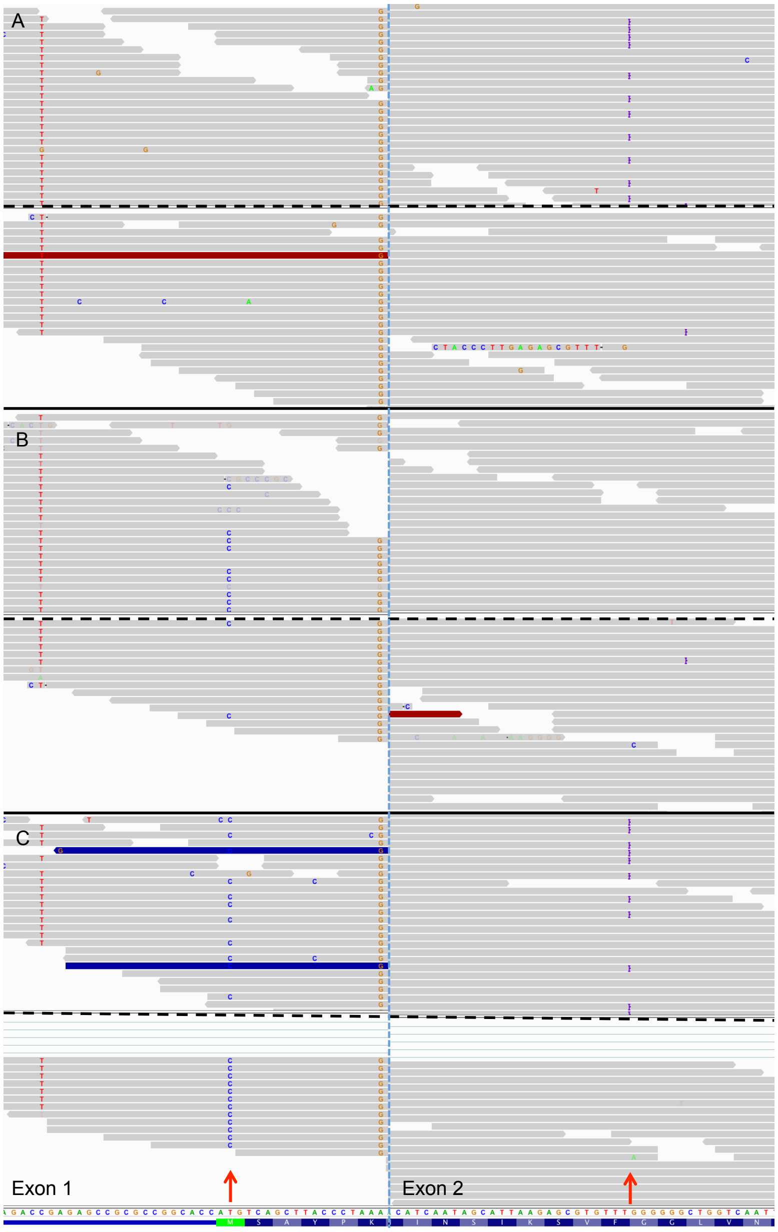
A
B
C
Exon 1
Exon 2
GACCGAGAGCCGCGCCGGCACCATGTCAGCTTACCCTAAACATCAATAGCATTAAGAGCGTGTTTGGGGGGCTGGTCAAT
M S A Y P K I N S I K S V F G G L V N

Figure 28. Exome and RNA sequencing of *SNAP29* variant.
A=Mother, B=Father, C=Patient. Blue center vertical line divides exon 1 and exon 2 tracks. Family member tracks are divided by black solid horizontal lines. Dashed horizontal lines separate the exome (upper) and RNA-seq tracks (lower). Exon 1 mutation is a chr22:21213400:T>C, and exon 2 mutation is a 22:21224735:T>TG their positions indicated by red arrows.

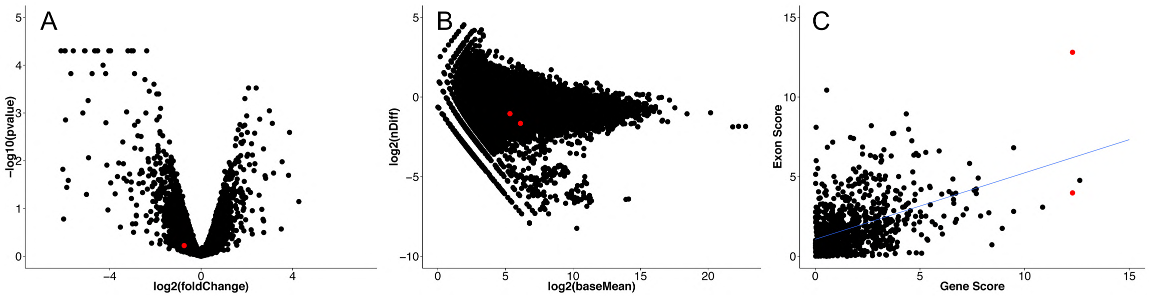

Figure 29. Expression profile of *SNAP29*. This plot shows the expression of *SNAP29* in patient 0117_1. A=results of Cuffdiff differential expression for protein coding genes. The horizontal axis shows log2 foldchange, and vertical axis is the negative log10 of the probability that the gene is significantly dysregulated. B= differential exon usage analysis. The horizontal axis is the log2 normalized exon coverage for exons with normalized coverage >1. The vertical axis shows the log2 normalized coverage difference between the patients and the parents. C=scatterplot of the gene-based and exon-based MD scores for each rare variant. The red dot in plot A indicates the SNAP gene, in B exon 1 and exon 2 results, and in C the two variants' MD scores.

Discussion

In this study, we developed a framework for integrated DNA and RNA analysis of high-throughput sequencing data in a multivariate format for 29 patients with rare childhood disorders using Mahalanobis distance for outliers to prioritize candidate variants.

The cohort represented a spectrum of rare neurological and musculoskeletal conditions with prolonged diagnostic odysseys and complex phenotype making clinical diagnosis challenging. Patients were selected for family-based DNA and RNA sequencing to utilize variant segregation with phenotype and used parental transcriptomes for comparative expression analysis. We performed outlier analysis on transcriptomic features including genes and exons and found patient specific variants that have large impact on transcription and correlate with phenotype. We obtained multiple measurements on these features including expression abundance and differential expression magnitude and applied these variables in a multivariate matrix to determine Mahalanobis distance of each patient specific feature.

After grouping variants across the 29 families based on predicted functional impact, we found that gene-based distance scores were associated with variants predicted to have high functional impact. This suggested that variants like splice acceptor, or donor, and stop codon are more likely result in an expression signature that is an outlier when gene expressions from multiple patients are compared. However, the exon-based scores did not support this finding. This may have been caused by multiple factors including the difference how gene and exon expression is estimated and also the variance in the scores. Exon score distributions had longer tails suggesting a greater variance in the data regardless of functional class (Figure 24). Exon expression in DEXSeq is determined by normalizing read counts for each exon across conditions which may introduce artifacts for those exons that are significantly differentially used between conditions (Hooper 2014). In addition, our estimation of differential exon usage is calculated based on the total number of reads sequenced per sample, and in some families the number of reads sequenced across samples varies greatly from 26 million to over 200 million reads per sample (Appendix C).

When variants were further categorized based on their predicted effect to mRNA sequence, we found that frameshift variants were associated with higher MD scores than other functional groups except splice site and start and stop codon variants. This suggests that the least frequently occurring variants tend to have the highest impact on transcription in our cohort. These three variant groups, frameshift, splice site and start|stop codon were the least frequent in our cohort of 18,834 variants (frameshift = 175/18,834, 0.9%; splice site = 70/18,834, 0.3%, start|stop codon= 84/18,834, 0.4%) (Table 20). It is important to note that variant annotation and classification into functional groups may have an impact on association analysis, and choice of annotation tool can be critical in interpretation (McCarthy et al. 2014). Therefore, future studies should evaluate those variants for association that are consistently classified between annotation tools. In addition prospective studies should increase the number of participants to increase power to detect associations between functional variants and their outlier scores.

We showed that over 50% of protein coding genes could be investigated in this study. Gene expression measured in whole blood has great implication to detect functionally active candidate variants only if candidate gene activity can be observed in blood. We recognize that many diseases manifest their phenotype in certain tissues exclusively. Thus information obtained by RNA-seq from whole blood will only be relevant if functional observations made in blood carry over to primary tissue. Previous study by Yang et al. (2013) found that exome sequencing achieved diagnostic success rate of 25% sequencing germline DNA in 250 clinical patients. Expression profile of published causal genes from Y. Yang et al. (2013) showed that 85% of causal genes were expressed in whole blood above the detection limit we set in this study and 69% were expressed above FPKM of 1 in our RNA-seq cohort. Using brain tissue data from Human BodyMap 2.0 Project analyzed simultaneously with our patient cohort, we found that 97% of previously published causal genes by Y. Yang et al., were above our detection limit and 88% were expressed above FPKM of 1.

We demonstrated that in our ten patients who had previous presumed causal variants the variants had outlier behavior and MD scores were ranked in the 90th percentile almost exclusively. In concordance with previously published data, close to half of presumed causal variants were missense (Y. Yang et al. 2013). Interestingly, we found similar proportion of presumed causal variants affecting splicing at authentic splice sites, splice regions, and exonic splice suppressor elements. This is an enrichment of splice variants compared to previously reported clinical sequencing studies (Y. Yang et al. 2013). This may be caused by our small sample size and our selection of extreme cases to be enrolled in our study.

Our study design was motivated by two factors, 1) diagnosis of rare disease can be improved upon by integrative genomics approaches, 2) rare variants have large impact on cellular phenotype that can be measured in a high-throughput manner. Unambiguity for a variant's causality can be improved by evidence from gene level signatures either from bioinformatics analyses or functional studies (MacArthur et al. 2014). Our approach obtains further evidence by integrated analysis of gene and exon based transcriptomic signatures in patient specific tissue. The correlation between predictions obtained from DNA sequencing with functional effect was previously demonstrated by MacArthur at al, who validated variant predictions of loss-of-function mutations causing nonsense-mediated decay in transcriptomic analysis of lymphoblastoid cell lines (MacArthur et al. 2012). Our study is another example of the power of integrated genomics and functional approaches have in the identification of high impact functional variants.

Our approach showed that focused, supervised data from genomic and functional sequencing can be efficiently joined and be surprisingly informative when multiple variables from

RNA-seq used for outlier analysis. Some of the confounding factors of integrative genomics approaches are the size of the data generated, the noise across data types, lack of correlation between sequencing technologies (Ritchie et al. 2015). To address these issues, we reduced our data to protein-coding variants only, which focused our attention to the most informative regions of the genome. In addition we used normalized, log transformed gene and exon abundances to reduce noise in the multivariate matrix and between samples. Normalization of gene expression across samples is an important issue as technical and biological variation can impact data interpretation. Multiple methods are suggested for normalization that are beyond the scope of this study, but we used geometric mean developed for DESeq and implemented in Cufflinks 2.2 to normalize gene expression across the 25 families (Dillies et al. 2013). We recognize that expression estimation is an important topic in RNA-seq and other approaches than FPKM have been proposed as more accurate estimators of expression abundance. However, we found that Cufflinks version 2.2 incorporated new elements addressing previous concerns of FPKM normalization, and its streamlined modular workflow was simple to implement and combined with differential gene expression analysis (Trapnell et al. 2013). In future studies of multivariate analysis accuracy of outlier estimation may be improved by use of TPM and other abundance estimators (B. Li and Dewey 2011).

Although the ultimate goal of our study was to find pathogenic variants that are supported by functional data, integration of genomic and functional data in our study only reports the magnitude of the functional impact with the MD score. Thus our findings in themselves do not prove causality. In current literature, most integrative approaches of genomic and functional variations test the integrated data for association with phenotype of interest (Schadt et al. 2005; Huang et al. 2007). A significant association is usually quantified by a p-value that can be set arbitrarily and needs adjustment for multiple testing due to the large number of variants tested. Multiple testing correction however in many cases is very conservative and leads to an inflation of false-negatives (Johnson et al. 2010). Our study is not powered to perform association analysis because we are studying extreme phenotypes with a single patient in most cases.

In two families we showed the utility of our approach by verifying the predicted affect of a presumed causal variant and by uncovering a new candidate variant. In family 0002, our method worked essentially as a validation tool. Based on previous clinical findings, *MTFMT* was the most plausible candidate variant that correlated with observed phenotype. *MTFMT*'s main role is to transfer formyl group to methionyl-trNA (met-tRNA^{Met}). met-tRNA^{Met} is essential in translation initiation and elongation in humans. The transfer of a formyl group determines the role of met-tRNA^{Met} in the translation process. Formylated met-tRNA^{Met} is associated with translation initiation in the ribosome, while un-formylated met-tRNA^{Met} is essential in translation elongation(Haack et al. 2014). Dysregulation of *MTFMT* due to mutations have been associated with altered mitochondrial oxidative phosphorylation (OXPHOS) due to inefficient translation of OXPHOS associated genes (Tucker et al. 2011). To date, mutations in *MTFMT* have been reported in two studies associated with OXPHOS dysfunction (Tucker et al. 2011; Haack et al. 2014). The c.626C>T mutation we found in our patients is the most common variant reported in 13 of 16 OXPHOS cases (Haack et al. 2012). In all but one case this variant was found in a compound heterozygous form. Exome Aggregation Consortium data of 60,706 exomes of unrelated individuals showed that this mutation had an MAF of about 0.00036%. Interestingly the two patients in our study who carry the homozygous mutation show phenotypic heterogeneity. Patient 0002_5 was diagnosed with Leigh Syndrome. One of the hallmarks of Leigh phenotype is a characteristic symmetrical brain lesion in basal ganglia and white matter loss, which was documented in the patient's MRI. In addition, Tucker et al. (2011) previously reported two cousins who had *MTFMT* mutations and cardiac dysfunction leading to a diagnosis of Wolff-Parkinson-White Syndrome (WPWS). Patient 0002_2 fits the WPWS description. Thus this family is an example of the phenotypic heterogeneity in mitochondrial disease. The integration showed the significant functional impact of the exon-skipping event that was more dominant in the homozygous patients. The full molecular characterization of the DNA-RNA predictions requires the addition of proteomic characterization of potential mechanism that lead to heterogeneity in the phenotypes of the homozygotes.

In family 0117, conventional exome sequencing approach and variant prioritization failed to identify the compound heterozygous variant in *SNAP29* as a potential candidate and only after outlier analysis of gene-based and exon-based scores, coupled with segregation analysis, this gene became a candidate. This patient does not fit the clinical description of CEDNIK syndrome, however, some phenotypic overlap, multivariate outlier analysis, and published study suggest the role of *SNAP29* in this patient. *SNAP29* is a soluble SNARE protein that has been implicated in cytoplasmic trafficking and synaptic plasticity. The importance of *SNAP29* in nerve myelination by microglia has been previously shown (Schardt et al. 2009). The overexpression of *SNAP29* and its binding partner *Rab3A* increased cell surface directed myelin proteolipid trafficking(Schardt et al. 2009). Schardt et al. (2009) also shown that in rat brain the remyelination process correlates with an increased abundance of *SNAP29* in sciatic nerves. Although no previous CEDNIK syndrome patient has been shown to have dysmyelination, or myelin related brain phenotype, the patient in this study shows diffuse lack of myelination. This patient also shows delayed development and abnormalities in the corpus callosum, which are hallmarks of patient phenotypes with loss-of-function *SNAP29* mutations and CEDNIK syndrome. Previously homozygous mutations affecting both copies of *SNAP29* showed loss of protein product by Western blot analysis (Sprecher and Ishida-Yamamoto 2005). In addition patients with heterozygous deletion encompassing *SNAP29*, and with a heterozygous mutation in the other copy of *SNAP29* has shown atypical CEDNIK phenotype (McDonald-McGinn et al. 2013). This patient has a predicted loss-of-function insertion that is predicted to result in no protein product from the maternal copy of *SNAP29*. However the patient shows expression of the paternal copy of the gene containing a loss of initiator codon mutation. This suggests a mechanism for gene translation from an alternate start site for translation machinery that may result in a modified N terminal of the nascent protein product leading to altered protein function and phenotypic presentation in the child. Investigation of the mRNA sequence of *SNAP29* shows that there are two alternate start codons downstream in exon 2, and ribosomes can initiate translation from alternative start codons through leaky scanning (Kozak 2005). In addition, RNA-seq data shows that father also expresses the mutant transcript with no phenotypic presentation. Additional molecular characterization of the mutant transcript is needed to connect DNA-RNA findings to patient’s phenotype.

In conclusion, we developed a novel framework of integrating genomic and functional information obtained from next-generation sequencing in our efforts to prioritize variants for diagnosis of complex, hard-to-diagnose childhood disorders. Our framework of combining multiple data types is mostly a proof-of-concept in our investigation of outlier expression signatures in patients who themselves are outliers. Our approach needs further development so data processing and management can be more streamlined and additional functional data can be incorporated into multivariate analysis. The promise of merging large datasets with complex information in an efficient and informative way will potentially improve clinical diagnosis, variant interpretation, speed up our search for clinically actionable biomarkers and empower novel study designs.

REFERENCES

Ahlborn, Lise B, Mette Dandanell, Ane Y Steffensen, Lars Jønson, Finn C Nielsen, and Thomas v O Hansen. 2015. “Splicing Analysis of 14 BRCA1 Missense Variants Classifies Nine Variants as Pathogenic..” *Breast Cancer Research and Treatment* 150 (2): 289–98. doi:10.1007/s10549-015-3313-7.

Ajay, S S, S C J Parker, H Ozel Abaan, K V Fuentes Fajardo, and E H Margulies. 2011. “Accurate and Comprehensive Sequencing of Personal Genomes.” *Genome Research* 21 (9): 1498–1505. doi:10.1101/gr.123638.111.

Allen, R C, H Y Zoghbi, A B Moseley, H M Rosenblatt, and J W Belmont. 1992. “Methylation of HpaII and HhaI Sites Near the Polymorphic CAG Repeat in the Human Androgen-Receptor Gene Correlates with X Chromosome Inactivation..” *American Journal of Human Genetics* 51 (6): 1229–39.

Amos-Landgraf, James M, Amy Cottle, Robert M Plenge, Mike Friez, Charles E Schwartz, John Longshore, and Huntington F Willard. 2006. “X Chromosome–Inactivation Patterns of 1,005 Phenotypically Unaffected Females.” *American Journal of Human Genetics* 79 (3). Elsevier: 493–99.

Anders, S, A Reyes, and W Huber. 2012. “Detecting Differential Usage of Exons From RNA-Seq Data.” *Genome Research* 22 (10): 2008–17. doi:10.1101/gr.133744.111.

Augui, Sandrine, Elphège P Nora, and Edith Heard. 2011. “Regulation of X-Chromosome Inactivation by the X-Inactivation Centre.” *Nature Publishing Group* 12 (6). Nature Publishing Group: 429–42. doi:10.1038/nrg2987.

Babak, Tomas, Brian DeVeale, Christopher Armour, Christopher Raymond, Michele A Cleary, Derek van der Kooy, Jason M Johnson, and Lee P Lim. 2008. “Global Survey of Genomic Imprinting by Transcriptome Sequencing.” *Current Biology* 18 (22). Elsevier Ltd: 1735–41. doi:10.1016/j.cub.2008.09.044.

Bainbridge, M N, W Wiszniewski, D R Murdock, J Friedman, C Gonzaga-Jauregui, I Newsham, J G Reid, et al. 2011. “Whole-Genome Sequencing for Optimized Patient Management.” *Science Translational Medicine* 3 (87): 87re3–87re3. doi:10.1126/scitranslmed.3002243.

Bamshad, Michael J, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. 2011. “Exome Sequencing as a Tool for Mendelian Disease Gene Discovery.” *Nature Publishing Group* 12 (11). Nature Publishing Group: 745–55. doi:10.1038/nrg3031.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, et al. 2014. “Characterizing the Genetic Basis of Transcriptome Diversity Through RNA-Sequencing of 922 Individuals..” *Genome Research* 24 (1): 14–24. doi:10.1101/gr.155192.113.

Becker, Jutta, Oliver Semler, Christian Gilissen, Yun Li, Hanno Jörn Bolz, Cecilia Giunta, Carsten Bergmann, et al. 2011. “Exome Sequencing Identifies Truncating Mutations in Human SERPINF1 in Autosomal-Recessive Osteogenesis Imperfecta.” *American Journal of Human Genetics* 88 (3). The American Society of Human Genetics: 362–71. doi:10.1016/j.ajhg.2011.01.015.

Behrends, Christian, Mathew E Sowa, Steven P Gygi, and J Wade Harper. 2010. “Network Organization of the Human Autophagy System..” *Nature* 466 (7302): 68–76. doi:10.1038/nature09204.

Ben-Gal, Irad. 2005. “Outlier Detection.” In *Data Mining and Knowledge Discovery Handbook*, edited by Oded Maimon and Lior Rokach, 131–46. New York: Springer US. doi:10.1007/0-387-25465-X_7.

Bidou, Laure, Valérie Allamand, Jean-Pierre Rousset, and Olivier Namy. 2012. “Sense From Nonsense: Therapies for Premature Stop Codon Diseases.” *Trends in Molecular Medicine* 18 (11). Elsevier Ltd: 679–88. doi:10.1016/j.molmed.2012.09.008.

Bilguvar, Kaya, Ali Kemal Öztürk, Angeliki Louvi, Kenneth Y Kwan, Murim Choi, Burak Tatlı, Dilek Yalnızoğlu, et al. 2010. “Whole-Exome Sequencing Identifies Recessive WDR62 Mutations in Severe Brain Malformations.” *Nature* 467 (7312): 207–10. doi:10.1038/nature09327.

Biliya, S, and L A Bulla. 2010. “Genomic Imprinting: the Influence of Differential Methylation in the Two Sexes.” *Experimental Biology and Medicine* 235 (2): 139–47. doi:10.1258/ebm.2009.009251.

BioGRID, Biological General Repository for Interaction Datasets. Accessed February 15, 2015. http://thebiogrid.org

Bittel, D C, M F Theodoro, N Kibiryeva, W Fischer, Z Talebizadeh, and M G Butler. 2008. “Comparison of X-Chromosome Inactivation Patterns in Multiple Tissues From Human Females.” *Journal of Medical Genetics* 45 (5): 309–13. doi:10.1136/jmg.2007.055244.

Bordes, Laurent, Didier Chauveau, and Pierre Vandekerkhove. 2007. “A Stochastic EM Algorithm for a Semiparametric Mixture Model.” *Computational Statistics & Data Analysis* 51 (11): 5429–43. doi:10.1016/j.csda.2006.08.015.

Borovecki, F, L Lovrecic, Jessica Zhou, Hyun Jeong, Florian Then, H D Rosas, S M Hersch, P Hogarth, Berengere Bouzou, and R V Jensen. 2005. “Genome-Wide Expression Profiling of Human Blood Reveals Biomarkers for Huntington's Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (31). National Acad Sciences: 11023–28.

Boycott, Kym M, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. 2013. “Rare-Disease Genetics in the Era of Next-Generation Sequencing: Discovery to Translation..” *Nature Publishing Group* 14 (10): 681–91. doi:10.1038/nrg3555.

Buettner, Florian, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. 2015. “Computational Analysis of Cell-to-Cell Heterogeneity in Single-Cell RNA-Sequencing Data Reveals Hidden Subpopulations of Cells.” *Nature Biotechnology* 33 (2): 155–60. doi:10.1038/nbt.3102.

Busque, L, Y Paquette, S Provost, D C Roy, R L Levine, L Mollica, and D Gary Gilliland. 2009. “Skewing of X-Inactivation Ratios in Blood Cells of Aging Women Is Confirmed by Independent Methodologies.” *Blood* 113 (15): 3472–74. doi:10.1182/blood-2008-12-195677.

Carrel, Laura, and Huntington F Willard. 2005. “X-Inactivation Profile Reveals Extensive Variability in X-Linked Gene Expression in Females..” *Nature* 434 (7031): 400–404. doi:10.1038/nature03479.

Chiang, Annie P, John S Beck, Hsan-Jan Yen, Marwan K Tayeh, Todd E Scheetz, Ruth E Swiderski, Darryl Y Nishimura, et al. 2006. “Homozygosity Mapping with SNP Arrays Identifies TRIM32, an E3 Ubiquitin Ligase, as a Bardet–Biedl Syndrome Gene (BBS11).” *Proceedings of the National Academy of Sciences of the United States of America* 103 (16). National Acad Sciences: 6287–92. doi:10.1073/pnas.0600158103.

Chilamakuri, Chandra Sekhar Reddy, Susanne Lorenz, Mohammed-Amin Madoui, Daniel Vodák, Jinchang Sun, Eivind Hovig, Ola Myklebost, and Leonardo A Meza-Zepeda. 2014. “Performance Comparison of Four Exome Capture Systems for Deep Sequencing..” *BMC Genomics* 15 (1). BioMed Central Ltd: 449. doi:10.1186/1471-2164-15-449.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3..” *Fly* 6 (2). Taylor & Francis: 80–92. doi:10.4161/fly.19695.

Codina-Solà, Marta, Benjamín Rodríguez-Santiago, Aïda Homs, Javier Santoyo, Maria Rigau, Gemma Aznar-Laín, Miguel del Campo, et al. 2015. “Integrated Analysis of Whole-Exome Sequencing and Transcriptome Profiling in Males with Autism Spectrum Disorders..” *Molecular Autism* 6 (1): 21. doi:10.1186/s13229-015-0017-0.

Consortium, The 1000 Genomes Project, Corresponding author, Steering committee, Production group Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, et al. 2012. “A Map of Human Genome Variation From Population-Scale Sequencing.” *Nature* 467 (7319). Nature Publishing Group: 1061–73. doi:10.1038/nature09534.

Coonrod, Emily M, Jacob D Durtschi, Rebecca L Margraf, and Karl V Voelkerding. 2013. “Developing Genome and Exome Sequencing for Candidate Gene Identification in Inherited Disorders: an Integrated Technical and Bioinformatics Approach.” *Archives of Pathology & Laboratory Medicine* 137 (3): 415–33. doi:10.5858/arpa.2012-0107-ra.

Cotton, Allison M, Lucia Lam, Joslynn G Affleck, Ian M Wilson, Maria S Peñaherrera, Deborah E McFadden, Michael S Kobor, Wan L Lam, Wendy P Robinson, and Carolyn J Brown. 2011. “Chromosome-Wide DNA Methylation Analysis Predicts Human Tissue-Specific X Inactivation.” *Human Genetics* 130 (2): 187–201. doi:10.1007/s00439-011-1007-8.

Cotton, Allison M, Bing Ge, Nicholas Light, Veronique Adoue, Tomi Pastinen, and Carolyn J Brown. 2013. “Analysis of Expressed SNPs Identifies Variable Extents of Expression From the Human Inactive X Chromosome..” *Genome Biology* 14 (11): R122. doi:10.1186/gb-2013-14-11-r122.

Craig, David W, Abraham Itty, Corrie Panganiban, Szabolcs Szelinger, Michael C Kruer, Aswin Sekar, David Reiman, Vinodh Narayanan, Dietrich A Stephan, and John F Kerrigan. 2008. “Identification of Somatic Chromosomal Abnormalities in Hypothalamic Hamartoma Tissue at the GLI3 Locus..” *American Journal of Human Genetics* 82 (2): 366–74. doi:10.1016/j.ajhg.2007.10.006.

Craig, D W, J A O'Shaughnessy, J A Kiefer, J Aldrich, S Sinari, T M Moses, S Wong, et al. 2013. “Genome and Transcriptome Sequencing in Prospective Metastatic Triple-Negative Breast Cancer Uncovers Therapeutic Vulnerabilities.” *Molecular Cancer Therapeutics* 12 (1): 104–16. doi:10.1158/1535-7163.MCT-12-0781.

Danda, Sumita, Vanessa A van Rahden, Deepa John, Padma Paul, Renu Raju, Santosh Koshy, and Kerstin Kutsche. 2014. “Evidence of Germline Mosaicism for a Novel BCOR Mutation in Two Indian Sisters with Oculo-Facio-Cardio-Dental Syndrome..” *Molecular Syndromology* 5 (5). Karger Publishers: 251–56. doi:10.1159/000365768.

dbSNP. Short Genetic Variations. Last accessed February 14, 2015. http://www.ncbi.nlm.nih.gov/SNP/

de Klerk, Eleonora, and Peter A C 't Hoen. 2015. “Alternative mRNA Transcription, Processing, and Translation: Insights From RNA Sequencing..” *Trends in Genetics* 31 (3). Elsevier: 128–39. doi:10.1016/j.tig.2015.01.001.

De Maesschalck, R, and D Jouan-Rimbaud. 2000. “The Mahalanobis Distance.” *… And Intelligent Laboratory …* 50 (1): 1–18. doi:10.1016/S0169-7439(99)00047-7.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data.” *Nature Publishing Group* 43 (5): 491–98. doi:10.1038/ng.806.

Desai, V, A Donsante, K J Swoboda, M Martensen, J Thompson, and S G Kaler. 2011. “Favorably Skewed X-Inactivation Accounts for Neurological Sparing in Female Carriers of Menkes Disease.” *Clinical Genetics* 79 (2): 176–82. doi:10.1111/j.1399-0004.2010.01451.x.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis..” *Briefings in Bioinformatics* 14 (6). Oxford University Press: 671–83. doi:10.1093/bib/bbs046.

Disteche, C M. 1999. “Escapees on the X Chromosome..” *Proceedings of the National Academy of Sciences of the United States of America* 96 (25): 14180–82.

Dixon-Salazar, T J, J L Silhavy, N Udpa, J Schroth, S Bielas, A E Schaffer, J Olvera, et al. 2012. “Exome Sequencing Can Improve Diagnosis and Alter Patient Management.” *Science Translational Medicine* 4 (138): 138ra78–138ra78. doi:10.1126/scitranslmed.3003544.

Dobin, A, C A Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and T R Gingeras. 2012. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics (Oxford, England)* 29 (1): 15–21. doi:10.1093/bioinformatics/bts635.

Dunn, Olive Jean. 1964. “Multiple Comparisons Using Rank Sums.” *Technometrics* 6 (3): 241–52. doi:10.1080/00401706.1964.10490181.

Eble, Tanya N, V Reid Sutton, Haleh Sangi-Haghpeykar, Xiaoling Wang, Weihong Jin, Richard A Lewis, Ping Fang, and Ignatia B Van den Veyver. 2009. “Non-Random X Chromosome Inactivation in Aicardi Syndrome.” *Human Genetics* 125 (2): 211–16. doi:10.1007/s00439-008-0615-4.

Ellis, Shannon E, Simone Gupta, Foram N Ashar, Joel S Bader, Andrew B West, and Dan E Arking. 2013. “RNA-Seq Optimization with eQTL Gold Standards..” *BMC Genomics* 14 (1). BioMed Central Ltd: 892. doi:10.1186/1471-2164-14-892.

Emanuele, Michael J, Andrew E H Elia, Qikai Xu, Claudio R Thoma, Lior Izhar, Yumei Leng, Ailan Guo, et al. 2011. “Global Identification of Modular Cullin-RING Ligase Substrates..” *Cell* 147 (2). Elsevier: 459–74. doi:10.1016/j.cell.2011.09.019.

Exome Aggregate Consortium. Last accessed February 14, 2015. http://exac.broadinstitute.org

Fang, Fang, Emily Hodges, Antoine Molaro, Matthew Dean, Gregory J Hannon, and Andrew D Smith. 2012. "Genomic Landscape of Human Allele-Specific DNA Methylation.." *Proceedings of the National Academy of Sciences* 109 (19): 7332–37. doi:10.1073/pnas.1201310109.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R R Forrest, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Vanja Haberle, et al. 2014. "A Promoter-Level Mammalian Expression Atlas.." *Nature* 507 (7493): 462–70. doi:10.1038/nature13182.

Farwell, Kelly D, Layla Shahmirzadi, Dima El-Khechen, Zöe Powis, Elizabeth C Chao, Brigette Tippin Davis, Ruth M Baxter, et al. 2014. "Enhanced Utility of Family-Centered Diagnostic Exome Sequencing with Inheritance Model-Based Analysis: Results From 500 Unselected Families with Undiagnosed Genetic Conditions.." *Genetics in Medicine*, November. doi:10.1038/gim.2014.154.

fastqc. A quality control tool for high throughput sequence data. Last accessed February 15, 2014. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Finotello, Francesca, and Barbara Di Camillo. 2015. "Measuring Differential Gene Expression with RNA-Seq: Challenges and Strategies for Data Analysis.." *Briefings in Functional Genomics* 14 (2). Oxford University Press: 130–42. doi:10.1093/bfgp/elu035.

Fischer, Ashwin Chari Monika M Golas Michael Klingenhäger Nils Neuenkirchen Bjoern Sander Clemens Englbrecht Albert Sickmann Holger Stark Utz. 2008. "An Assembly Chaperone Collaborates with the SMN Complex to Generate Spliceosomal SnRNPs." *Cell* 135 (3). Elsevier Inc.: 497–509. doi:10.1016/j.cell.2008.09.020.

Florea, Liliana, Li Song, and Steven L Salzberg. 2013. "Thousands of Exon Skipping Events Differentiate Among Splicing Patterns in Sixteen Human Tissues.." *F1000Research* 2: 188. doi:10.12688/f1000research.2-188.v2.

Fridley, Brooke L, Steven Lund, Gregory D Jenkins, and Liewei Wang. 2012. "A Bayesian Integrative Genomic Model for Pathway Analysis of Complex Traits.." *Genetic Epidemiology* 36 (4): 352–59. doi:10.1002/gepi.21628.

Fuchs-Telem, D, H Stewart, D Rapaport, J Nousbeck, A Gat, M Gini, Y Lugassy, et al. 2011. "CEDNIK Syndrome Results From Loss-of-Function Mutations in SNAP29." *British Journal of Dermatology* 164 (3). Blackwell Publishing Ltd: no–no. doi:10.1111/j.1365-2133.2010.10133.x.

Gargis, Amy S, Lisa Kalman, David P Bick, Cristina da Silva, David P Dimmock, Birgit H Funke, Sivakumar Gowrisankar, et al. 2015. "Good Laboratory Practice for Clinical Next-Generation Sequencing Informatics Pipelines." *Nature Publishing Group* 33 (7). Nature Publishing Group: 689–93. doi:10.1038/nbt.3237.

Gilissen, Christian, Alexander Hoischen, Han G Brunner, and Joris A Veltman. 2011. "Unlocking Mendelian Disease Using Exome Sequencing.." *Genome Biology* 12 (9). BioMed Central Ltd: 228. doi:10.1186/gb-2011-12-9-228.

Gilissen, Christian, Alexander Hoischen, Han G Brunner, and Joris A Veltman. 2012. "Disease Gene Identification Strategies for Exome Sequencing." *European Journal of Human Genetics* 20 (5). Nature Publishing Group: 490–97. doi:10.1038/ejhg.2011.258.

Gilissen, Christian, Jayne Y Hehir-Kwa, Djie Tjwan Thung, Maartje van de Vorst, Bregje W M van Bon, Marjolein H Willemsen, Michael Kwint, et al. 2014. “Genome Sequencing Identifies Major Causes of Severe Intellectual Disability.” *Nature* 511 (7509). Nature Publishing Group: 344–47. doi:10.1038/nature13394.

Gribnau, Joost, Sandra Luikenhuis, Konrad Hochedlinger, Kim Monkhorst, and Rudolf Jaenisch. 2005. “X Chromosome Choice Occurs Independently of Asynchronous Replication Timing..” *The Journal of Cell Biology* 168 (3). Rockefeller Univ Press: 365–73. doi:10.1083/jcb.200405117.

Grubbs, Frank E. 1969. “Procedures for Detecting Outlying Observations in Samples.” *Technometrics* 11 (1): 1–21. doi:10.1080/00401706.1969.10490657.

Haack, Tobias B, Matteo Gorza, Katharina Danhauser, Johannes A Mayr, Birgit Haberberger, Thomas Wieland, Laura Kremer, et al. 2014. “Phenotypic Spectrum of Eleven Patients and Five Novel MTFMT Mutations Identified by Exome Sequencing and Candidate Gene Screening.” *Molecular Genetics and Metabolism* 111 (3). Elsevier Inc.: 342–52. doi:10.1016/j.ymgme.2013.12.010.

Haack, Tobias B, Penelope Hogarth, Michael C Kruer, Allison Gregory, Thomas Wieland, Thomas Schwarzmayr, Elisabeth Graf, et al. 2012. “Exome Sequencing Reveals De Novo WDR45 Mutations Causing a Phenotypically Distinct, X-Linked Dominant Form of NBIA.” *American Journal of Human Genetics* 91 (6). Elsevier: 1144–49. doi:10.1016/j.ajhg.2012.10.019.

Haack, Tobias B, Penny Hogarth, Allison Gregory, Holger Prokisch, and Susan J Hayflick. 2013. *BPAN: the Only X-Linked Dominant NBIA Disorder*. *Metal Related Neurodegenerative Disease*. 1st ed. Vol. 110. International Review of Neurobiology. Elsevier Inc. doi:10.1016/B978-0-12-410502-7.00005-3.

Hara, Taichi, Kenji Nakamura, Makoto Matsui, Akitsugu Yamamoto, Yohko Nakahara, Rika Suzuki-Migishima, Minesuke Yokoyama, et al. 2006. “Suppression of Basal Autophagy in Neural Cells Causes Neurodegenerative Disease in Mice..” *Nature* 441 (7095): 885–89. doi:10.1038/nature04724.

Hardcastle, Thomas J, and Krystyna A Kelly. 2013. “Empirical Bayesian Analysis of Paired High-Throughput Sequencing Data with a Beta-Binomial Distribution..” *BMC Bioinformatics* 14 (1): 135. doi:10.1186/1471-2105-14-135.

Hawkins, D M. 1980. *Identification of Outliers*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-015-3994-4.

Hayflick, S J, M C Kruer, A Gregory, T B Haack, M A Kurian, H H Houlden, J Anderson, et al. 2013. “Beta-Propeller Protein-Associated Neurodegeneration: a New X-Linked Dominant Disorder with Brain Iron Accumulation.” *Brain* 136 (6): 1708–17. doi:10.1093/brain/awt095.

Holzinger, Emily R, Scott M Dudek, Alex T Frase, Sarah A Pendergrass, and Marylyn D Ritchie. 2014. “ATHENA: the Analysis Tool for Heritable and Environmental Network Associations..” *Bioinformatics (Oxford, England)* 30 (5): 698–705. doi:10.1093/bioinformatics/btt572.

Hooper, Joan E. 2014. “A Survey of Software for Genome-Wide Discovery of Differential Splicing in RNA-Seq Data..” *Human Genomics* 8 (1). BioMed Central Ltd: 3. doi:10.1186/1479-7364-8-3.

Hu, Hao, Chad D Huff, Barry Moore, Steven Flygare, Martin G Reese, and Mark Yandell. 2013. “VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix.” *Genetic Epidemiology* 37 (6): 622–34. doi:10.1002/gepi.21743.

Huang, R Stephanie, Shiwei Duan, Wasim K Bleibel, Emily O Kistner, Wei Zhang, Tyson A Clark, Tina X Chen, et al. 2007. “A Genome-Wide Approach to Identify Genetic Variants That Contribute to Etoposide-Induced Cytotoxicity..” *Proceedings of the National Academy of Sciences of the United States of America* 104 (23): 9758–63. doi:10.1073/pnas.0703736104.

Hunter, David R, Shaoli Wang, and Thomas P Hettmansperger. 2007. “Inference for Mixtures of Symmetric Distributions.” *The Annals of Statistics* 35 (1): 224–51. doi:10.1214/009053606000001118.

Jiang, H, and W H Wong. 2009. “Statistical Inferences for Isoform Expression in RNA-Seq.” *Bioinformatics (Oxford, England)* 25 (8): 1026–32. doi:10.1093/bioinformatics/btp113.

Jiao, X, H Chen, J Chen, K Herrup, B L Firestein, and M Kiledjian. 2009. “Modulation of Neuritogenesis by a Protein Implicated in X-Linked Mental Retardation.” *Journal of Neuroscience* 29 (40): 12419–27. doi:10.1523/JNEUROSCI.5954-08.2009.

Johnson, Randall C, George W Nelson, Jennifer L Troyer, James A Lautenberger, Bailey D Kessing, Cheryl A Winkler, and Stephen J O'Brien. 2010. “Accounting for Multiple Comparisons in a Genome-Wide Association Study (GWAS).” *BMC Genomics* 11 (1). BioMed Central Ltd: 724. doi:10.1186/1471-2164-11-724.

Jombart, T, D Pontier, and A-B Dufour. 2009. “Genetic Markers in the Playground of Multivariate Analysis..” *Heredity* 102 (4). Nature Publishing Group: 330–41. doi:10.1038/hdy.2008.130.

Kerem, Bat-Sheva, Johanna M Rommens, Janet A Buchanan, Danuta Markiewicz, Tara K Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee Tsui. 1989. “Identification of the Cystic Fibrosis Gene: Genetic Analysis.” *Science* 245 (4): 1073–80. doi:10.1126/science.2570460.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes Inthe Presence of Insertions, Deletions and Genefusions.” *Genome Biology* 14 (4). BioMed Central Ltd: R36. doi:10.1186/gb-2013-14-4-r36.

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. 2014. “Technical Reports.” *Nature Publishing Group* 46 (3). Nature Publishing Group: 310–15. doi:10.1038/ng.2892.

Knudsen, G P S, J Pedersen, O Klingenberg, I Lygren, and K H Ørstavik. 2007. “Increased Skewing of X Chromosome Inactivation with Age in Both Blood and Buccal Cells..” *Cytogenetic and Genome Research* 116 (1-2). Karger Publishers: 24–28. doi:10.1159/000097414.

Kothari, Vishal, Iris Wei, Sunita Shankar, Shanker Kalyana-Sundaram, Lidong Wang, Linda W Ma, Pankaj Vats, et al. 2013. “Outlier Kinase Expression by RNA Sequencing as Targets for Precision Therapy..” *Cancer Discovery* 3 (3). American Association for Cancer Research: 280–93. doi:10.1158/2159-8290.CD-12-0336.

Kozak, Marilyn. 2005. “Regulation of Translation via mRNA Structure in Prokaryotes and Eukaryotes.” *Gene* 361 (November): 13–37. doi:10.1016/j.gene.2005.06.037.

Lam, Hugo Y K, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O'Huallachain, Frederick E Dewey, et al. 2011. “Performance Comparison of Whole-Genome Sequencing Platforms.” *Nature Biotechnology* 30 (1). Nature Publishing Group: 78–82. doi:10.1038/nbt.2065.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, and W FitzHugh. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822). Nature Publishing Group: 860–921.

Landrum, M J, J M Lee, G R Riley, and W Jang. 2014. “ClinVar: Public Archive of Relationships Among Sequence Variation and Human Phenotype.” *Nucleic Acids …*. doi:10.1093/nar/gkt1113.

Lappalainen, Tuuli, Michael Sammeth, Marc R Friedländer, Peter A C t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, et al. 2014. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468). Nature Publishing Group: 506–11. doi:10.1038/nature12531.

Li, Bo, and Colin N Dewey. 2011. “RSEM: Accurate Transcript Quantification From RNA-Seq Data with or Without a Reference Genome..” *BMC Bioinformatics* 12 (1). BioMed Central Ltd: 323. doi:10.1186/1471-2105-12-323.

Li, D, and R Roberts. 2001. “WD-Repeat Proteins: Structure Characteristics, Biological Function, and Their Involvement in Human Diseases..” *Cellular and Molecular Life Sciences : CMLS* 58 (14): 2085–97.

Li, Feng, Yiping Shen, Udo Köhler, Freddie H Sharkey, Deepa Menon, Laurence Coulleaux, Valérie Malan, et al. 2010. “Interstitial Microduplication of Xp22.31: Causative of Intellectual Disability or Benign Copy Number Variant?.” *European Journal of Medical Genetics* 53 (2). Elsevier Masson SAS: 93–99. doi:10.1016/j.ejmg.2010.01.004.

Li, H, and R Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Li, H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, M X, H S Gui, J S H Kwan, S Y Bao, and P C Sham. 2012. “A Comprehensive Framework for Prioritizing Variants in Exome Sequencing Studies of Mendelian Diseases.” *Nucleic Acids Research* 40 (7): e53–e53. doi:10.1093/nar/gkr1257.

Li, Sheng, Scott W Tighe, Charles M Nicolet, Deborah Grove, Shawn Levy, William Farmerie, Agnes Viale, et al. 2014. “Multi-Platform Assessment of Transcriptome Profiling Using RNA-Seq in the ABRF Next-Generation Sequencing Study.” *Nature Biotechnology* 32 (9): 915–25. doi:10.1038/nbt.2972.

Liang, Winnie S, David W Craig, John Carpten, Mitesh J Borad, Michael J Demeure, Glen J Weiss, Tyler Izatt, et al. 2012. “Genome-Wide Characterization of Pancreatic Adenocarcinoma Patients Using Next Generation Sequencing.” Edited by Fazlul H Sarkar. *PLoS ONE* 7 (10): e43192. doi:10.1371/journal.pone.0043192.t005.

Liu, P, A Erez, S C Sreenath Nagamani, W Bi, C M B Carvalho, A D Simmons, J Wiszniewska, et al. 2011. “Copy Number Gain at Xp22.31 Includes Complex Duplication Rearrangements and Recurrent Triplications.” *Human Molecular Genetics* 20 (10): 1975–88. doi:10.1093/hmg/ddr078.

Liu, Xiaoming, Xueqiu Jian, and Eric Boerwinkle. 2013. “dbNSFP v2.0: a Database of Human Non-Synonymous SNVs and Their Functional Predictions and Annotations..” *Human Mutation* 34 (9): E2393–E2402. doi:10.1002/humu.22376.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2..” *Genome Biology* 15 (12). BioMed Central Ltd: 550–21. doi:10.1186/s13059-014-0550-8.

Lu, Qun, Peiguo Yang, Xinxin Huang, Wanqiu Hu, Bin Guo, Fan Wu, Long Lin, Attila L Kovács, Li Yu, and Hong Zhang. 2011. “The WD40 Repeat PtdIns(3)P-Binding Protein EPG-6 Regulates Progression of Omegasomes to Autophagosomes.” *Developmental Cell* 21 (2). Elsevier Inc.: 343–57. doi:10.1016/j.devcel.2011.06.024.

Lupski, James R, Jeffrey G Reid, Claudia Gonzaga-Jauregui, David Rio Deiros, David C Y Chen, Lynne Nazareth, Matthew Bainbridge, et al. 2010. “Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy.” *The New England Journal of Medicine* 362 (13): 1181–91. doi:10.1056/NEJMoa0908094.

Lyon, M F. 1961. “Gene Action in the X-Chromosome of the Mouse.” *Nature* 4773 (190): 372–73.

MacArthur, D G, T A Manolio, D P Dimmock, H L Rehm, J Shendure, G R Abecasis, D R Adams, et al. 2014. “Guidelines for Investigating Causality of Sequence Variants in Human Disease..” *Nature* 508 (7497): 469–76. doi:10.1038/nature13127.

MacArthur, Daniel G, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, et al. 2012. “A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes..” *Science* 335 (6070): 823–28. doi:10.1126/science.1215040.

Mahalanobis, P C. 1936. *On the Generalized Distance in Statistics*. Proceedings of the National Institute of Sciences (…. http://ci.nii.ac.jp/naid/10004710165/.

Main, Bradley J, Ryan D Bickel, Lauren M McIntyre, Rita M Graze, Peter P Calabrese, and Sergey V Nuzhdin. 2009. “Allele-Specific Expression Assays Using Solexa.” *BMC Genomics* 10 (1): 422. doi:10.1186/1471-2164-10-422.

Majewski, J. 2002. “Distribution and Characterization of Regulatory Elements in the Human Genome.” *Genome Research* 12 (12): 1827–36. doi:10.1101/gr.606402.

Mangs, A H, and B J Morris. 2007. “The Human Pseudoautosomal Region (PAR): Origin, Function and Future.” *Current Genomics* 8 (2). Bentham Science Publishers: 129.

Mankoo, Parminder K, Ronglai Shen, Nikolaus Schultz, Douglas A Levine, and Chris Sander. 2011. “Time to Recurrence and Survival in Serous Ovarian Tumors Predicted From Integrated Genomic Profiles.” Edited by Sumitra Deb. *PLoS ONE* 6 (11): e24709. doi:10.1371/journal.pone.0024709.s008.

McCarthy, Davis J, Peter Humburg, Alexander Kanapin, Manuel A Rivas, Kyle Gaulton, Jean-Baptiste Cazier, and Peter Donnelly. 2014. “Choice of Transcripts and Software Has a Large Effect on Variant Annotation..” *Genome Medicine* 6 (3). BioMed Central Ltd: 26. doi:10.1186/gm543.

McDonald-McGinn, Donna M, Somayyeh Fahiminiya, Timothée Revil, Beata A Nowakowska, Joshua Suhl, Alice Bailey, Elisabeth Mlynarski, et al. 2013. “Hemizygous Mutations in SNAP29 Unmask Autosomal Recessive Conditions and Contribute to Atypical Findings in Patients with 22q11.2DS..” *Journal of Medical Genetics* 50 (2). BMJ Publishing Group Ltd: 80–90. doi:10.1136/jmedgenet-2012-101320.

McKenna, A, M Hanna, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, K Garimella, et al. 2010. “The Genome Analysis Toolkit: a MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. doi:10.1101/gr.107524.110.

McLaren, W, B Pritchard, D Rios, Y Chen, P Flicek, and F Cunningham. 2010. “Deriving the Consequences of Genomic Variants with the Ensembl API and SNP Effect Predictor.” *Bioinformatics (Oxford, England)* 26 (16): 2069–70. doi:10.1093/bioinformatics/btq330.

Meienberg, Janine, Katja Zerjavic, Irene Keller, Michal Okoniewski, Andrea Patrignani, Katja Ludin, Zhenyu Xu, et al. 2015. “New Insights Into the Performance of Human Whole-Exome Capture Platforms..” *Nucleic Acids Research* 43 (11). Oxford University Press: e76–e76. doi:10.1093/nar/gkv216.

Migeon, B R. 2006. “The Role of X Inactivation and Cellular Mosaicism in Women's Health and Sex-Specific Diseases.” *JAMA: the Journal of the American Medical Association* 295 (12). Am Med Assoc: 1428–33.

mongo. Last accessed February 14, 2015. https://www.mongodb.org

Moreira de Mello, Joana Carvalho, Érica Sara Souza de Araújo, Raquel Stabellini, Ana Maria Fraga, Jorge Estefano Santana de Souza, Denilce R Sumita, Anamaria A Camargo, and Lygia V Pereira. 2010. “Random X Inactivation and Extensive Mosaicism in Human Placenta Revealed by Analysis of Allele-Specific Gene Expression Along the X Chromosome.” Edited by Edith Heard. *PLoS ONE* 5 (6): e10947. doi:10.1371/journal.pone.0010947.t001.

Morgan, Daniel J, Suwen Wei, Ivone Gomes, Traci Czyzyk, Nino Mzhavia, Hui Pan, Lakshmi A Devi, Lloyd D Fricker, and John E Pintar. 2010. “The Propeptide Precursor proSAAS Is Involved in Fetal Neuropeptide Processing and Body Weight Regulation.” *Journal of Neurochemistry*, April. doi:10.1111/j.1471-4159.2010.06706.x.

Mossner, Maximilian, Florian Nolte, Gero Hütter, Jana Reins, Marion Klaumünzer, Verena Nowak, Julia Obländer, et al. 2013. “Skewed X-Inactivation Patterns in Ageing Healthy and Myelodysplastic Haematopoiesis Determined by a Pyrosequencing Based Transcriptional Clonality Assay..” *Journal of Medical Genetics* 50 (2). BMJ Publishing Group Ltd: 108–17. doi:10.1136/jmedgenet-2012-101093.

Muller, Hermann J. 1932. “Further Studies on the Nature and Causes of Gene Mutations” 1 (21): 3–255.

Ng, Pauline C, and Steven Henikoff. 2003. “SIFT: Predicting Amino Acid Changes That Affect Protein Function..” *Nucleic Acids Research* 31 (13): 3812–14.

Ng, Sarah B, Abigail W Bigham, Kati J Buckingham, Mark C Hannibal, Margaret J McMillin, Heidi I Gildersleeve, Anita E Beck, et al. 2010. "Exome Sequencing Identifies MLL2 Mutations as a Cause of Kabuki Syndrome." *Nature Genetics* 42 (9): 790–93. doi:10.1038/ng.646.

Ng, Sarah B, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, et al. 2009. "Exome Sequencing Identifies the Cause of a Mendelian Disorder." *Nature Genetics* 42 (1). Nature Publishing Group: 30–35. doi:10.1038/ng.499.

Okamoto, Nobuhiko, Tae Ikeda, Tatsuji Hasegawa, Yuto Yamamoto, Kazumi Kawato, Tomohiro Komoto, and Issei Imoto. 2014. "Early Manifestations of BPAN in a Pediatric Patient." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 164 (12): 3095–99. doi:10.1002/ajmg.a.36779.

Oláh, Judit, Orsolya Vincze, Dezsõ Virók, Dóra Simon, Zsolt Bozsó, Natália Tõkési, István Horváth, et al. 2011. "Interactions of Pathological Hallmark Proteins: Tubulin Polymerization Promoting Protein/P25, Beta-Amyloid, and Alpha-Synuclein.." *The Journal of Biological Chemistry* 286 (39). American Society for Biochemistry and Molecular Biology: 34088–100. doi:10.1074/jbc.M111.243907.

OMIM. *Online Mendelian Inheritance in Man. Accessed Fabruary 15, 2015. http://www.ncbi.nlm.nih.gov/omim*

Oshlack, Alicia, Mark D Robinson, and Matthew D Young. 2010. "From RNA-Seq Reads to Differential Expression Results.." *Genome Biology* 11 (12): 220. doi:10.1186/gb-2010-11-12-220.

Ozawa, Tadashi, Reiji Koide, Yasuhiro Nakata, Hirotomo Saitsu, Naomichi Matsumoto, Kazushi Takahashi, Imaharu Nakano, and Satoshi Orimo. 2014. "A Novel WDR45mutation in a Patient with Static Encephalopathy of Childhood with Neurodegeneration in Adulthood (SENDA)." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 164 (9): 2388–90. doi:10.1002/ajmg.a.36635.

Ozsolak, Fatih, and Patrice M Milos. 2010. "RNA Sequencing: Advances, Challenges and Opportunities." *Nature Publishing Group* 12 (2). Nature Publishing Group: 87–98. doi:10.1038/nrg2934.

Pagon, Roberta A, Margaret P Adam, Holly H Ardinger, Stephanie E Wallace, Anne Amemiya, Lora JH Bean, Thomas D Bird, et al. 1993. "Neurodegeneration with Brain Iron Accumulation Disorders Overview." Seattle (WA): University of Washington, Seattle.

picard. A set of tools (in Java) for working with next generation sequencing data in the BAM format. Last accessed February 14, 2015. http://broadinstitute.github.io/picard/

Piskol, Robert, Gokul Ramaswami, and Jin Billy Li. 2013. "AR TICLEReliable Identification of Genomic Variants From RNA-Seq Data." *American Journal of Human Genetics* 93 (4). The American Society of Human Genetics: 1–11. doi:10.1016/j.ajhg.2013.08.008.

Plenge, Robert M, Roger A Stevenson, Herbert A Lubs, Charles E Schwartz, and Huntington F Willard. 2002. "Skewed X-Chromosome Inactivation Is a Common Feature of X-Linked Mental Retardation Disorders.." *American Journal of Human Genetics* 71 (1): 168–73. doi:10.1086/341123.

Pollard, Katherine S, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. 2010. "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies.." *Genome Research* 20 (1). Cold Spring Harbor Lab: 110–21. doi:10.1101/gr.097857.109.

R. The R Project for Statistical Computing. Last accessed February 14, 2015. http://www.r-project.org

Rapaport, Franck, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. 2013. “Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data.” *Genome Biology* 14 (9). BioMed Central Ltd: R95. doi:10.1186/gb-2013-14-9-r95.

Rice, Gillian I, Martin A M Reijns, Stephanie R Coffin, Gabriella M A Forte, Beverley H Anderson, Marcin Szynkiewicz, Hannah Gornall, et al. 2013. “Synonymous Mutations in RNASEH2ACreate Cryptic Splice Sites Impairing RNase H2 Enzyme Function in Aicardi-Goutières Syndrome.” *Human Mutation* 34 (8): 1066–70. doi:10.1002/humu.22336.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, et al. 2015. “Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology..” *Genetics in Medicine* 17 (5): 405–23. doi:10.1038/gim.2015.30.

Risso, Davide, John Ngai, Terence P Speed, and Sandrine Dudoit. 2014. “Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples.” *Nature Biotechnology* 32 (9): 896–902. doi:10.1038/nbt.2931.

Ritchie, Marylyn D, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. 2015. “Methods of Integrating Data to Uncover Genotype–Phenotype Interactions.” *Nature Publishing Group* 16 (2). Nature Publishing Group: 85–97. doi:10.1038/nrg3868.

Roberts, Adam, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, Lior Pachter, and Cole Trapnell. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols* 7 (3). Nature Publishing Group: 562–78. doi:10.1038/nprot.2012.016.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1). Nature Publishing Group: 24–26. doi:10.1038/nbt0111-24.

Ross, Michael G, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. 2013. “Characterizing and Measuring Bias in Sequence Data.” *Genome Biology* 14 (5). BioMed Central Ltd: R51. doi:10.1186/gb-2013-14-5-r51.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, et al. 2011. “AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework.” *Molecular Systems Biology* 7 (1). Nature Publishing Group: 1–15. doi:10.1038/msb.2011.54.

Saitsu, Hirotomo, Taki Nishimura, Kazuhiro Muramatsu, Hirofumi Kodera, Satoko Kumada, Kenji Sugai, Emi Kasai-Yoshida, et al. 2013. “De Novo Mutations in the Autophagy Gene WDR45 Cause Static Encephalopathy of Childhood with Neurodegeneration in Adulthood.” *Nature Publishing Group* 45 (4). Nature Publishing Group: 445–49. doi:10.1038/ng.2562.

Schadt, Eric E, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, et al. 2005. “An Integrative Genomics Approach to Infer Causal Associations Between Gene Expression and Disease.” *Nature Genetics* 37 (7): 710–17. doi:10.1038/ng1589.

Schardt, Anke, Bastian G Brinkmann, Miso Mitkovski, Michael W Sereda, Hauke B Werner, and Klaus-Armin Nave. 2009. “The SNARE Protein SNAP-29 Interacts with the GTPase Rab3A: Implications for Membrane Trafficking in Myelinating Glia.” Edited by Rashmi Bansal, Wendy B Macklin, and Jean de Vellis. *Journal of Neuroscience Research* 87 (15): 3465–79. doi:10.1002/jnr.22005.

Scherzer, Clemens R, Aron C Eklund, Lee J Morse, Zhixiang Liao, Joseph J Locascio, Daniel Fefer, Michael A Schwarzschild, et al. 2007. “Molecular Markers of Early Parkinson's Disease Based on Gene Expression in Blood..” *Proceedings of the National Academy of Sciences of the United States of America* 104 (3): 955–60. doi:10.1073/pnas.0610204104.

Schissler, A Grant, Vincent Gardeux, Qike Li, Ikbel Achour, Haiquan Li, Walter W Piegorsch, and Yves A Lussier. 2015. “Dynamic Changes of RNA-Sequencing Expression for Precision Medicine: N-of-1-Pathways Mahalanobis Distance Within Pathways of Single Subjects Predicts Breast Cancer Survival..” *Bioinformatics (Oxford, England)* 31 (12). Oxford University Press: i293–i302. doi:10.1093/bioinformatics/btv253.

Schrauwen, Isabelle, Szabolcs Szelinger, Ashley L Siniard, Jason J Corneveaux, Ahmet Kurdoglu, Ryan Richholt, Matt De Both, et al. 2015. “A De Novo Mutation in TEAD1 Causes Non-X-Linked Aicardi Syndrome..” *Investigative Ophthalmology & Visual Science* 56 (6). The Association for Research in Vision and Ophthalmology: 3896–3904. doi:10.1167/iovs.14-16261.

Schwarz, Jana Marie, David N Cooper, Markus Schuelke, and Dominik Seelow. 2014. “MutationTaster2: Mutation Prediction for the Deep-Sequencing Age..” *Nature Methods* 11 (4): 361–62. doi:10.1038/nmeth.2890.

seqtk. Toolkit for processing sequences in FASTA/Q formats. Last accessed February 14, 2015. https://github.com/lh3/seqtk

Shah, Sohrab P, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, et al. 2012. “The Clonal and Mutational Evolution Spectrum of Primary Triple-Negative Breast Cancers.” *Nature*, April. doi:10.1038/nature10933.

Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. “Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science.” *Nature Publishing Group* 14 (9). Nature Publishing Group: 618–30. doi:10.1038/nrg3542.

Sheikh, Taimoor I, Kirti Mittal, Mary J Willis, and John B Vincent. 2013. “A Synonymous Change, P.Gly16Gly in MECP2 Exon 1, Causes a Cryptic Splice Event in a Rett Syndrome Patient..” *Orphanet Journal of Rare Diseases* 8 (1). BioMed Central Ltd: 108. doi:10.1186/1750-1172-8-108.

Siepel, Adam, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, et al. 2005. “Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes..” *Genome Research* 15 (8). Cold Spring Harbor Lab: 1034–50. doi:10.1101/gr.3715005.

Skelly, D A, M Johansson, J Madeoy, J Wakefield, and J M Akey. 2011. “A Powerful and Flexible Statistical Framework for Testing Hypotheses of Allele-Specific Gene Expression From RNA-Seq Data.” *Genome Research* 21 (10): 1728–37. doi:10.1101/gr.119784.110.

Solomon, B D, A D Nguyen, and K A Bear. 2013. “Clinical Genomic Database.” In. doi:10.1073/pnas.1302575110/-/DCSupplemental.

Sprecher, E, and A Ishida-Yamamoto. 2005. “A Mutation in SNAP29, Coding for a SNARE Protein Involved in Intracellular Trafficking, Causes a Novel Neurocutaneous Syndrome Characterized by Cerebral ….” *The American Journal of Human Genetics* 77 (2): 242–51. doi:10.1086/432556.

Stenson, Peter D, Edward V Ball, Matthew Mort, Andrew D Phillips, Jacqueline A Shiel, Nick S T Thomas, Shaun Abeysinghe, Michael Krawczak, and David N Cooper. 2003. “Human Gene Mutation Database (HGMD): 2003 Update.” *Human Mutation* 21 (6): 577–81. doi:10.1002/humu.10212.

Stevenson, Kraig R, Joseph D Coolon, and Patricia J Wittkopp. 2013. “Sources of Bias in Measures of Allele-Specific Expression Derived From RNA-Seq Data Aligned to a Single Reference Genome.” *BMC Genomics* 14 (1). BMC Genomics: 1–1. doi:10.1186/1471-2164-14-536.

Sugnet, C W, W J Kent, M Ares, and D Haussler. 2004. “Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice..” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 66–77.

Sulem, Patrick, Hannes Helgason, Asmundur Oddson, Hreinn Stefansson, Sigurjon A Gudjonsson, Florian Zink, Eirikur Hjartarson, et al. 2015. “Identification of a Large Set of Rare Complete Human Knockouts.” *Nature Publishing Group* 47 (5). Nature Publishing Group: 448–52. doi:10.1038/ng.3243.

Sun, Wei. 2011. “A Statistical Framework for eQTL Mapping Using RNA-Seq Data.” *Biometrics* 68 (1): 1–11. doi:10.1111/j.1541-0420.2011.01654.x.

Swierczek, S I, L Piterkova, J Jelinek, N Agarwal, S Hammoud, A Wilson, K Hickman, C J Parker, B R Cairns, and J T Prchal. 2012. “Methylation of AR Locus Does Not Always Reflect X Chromosome Inactivation State.” *Blood* 119 (13): e100–e109. doi:10.1182/blood-2011-11-390351.

Swierczek, S I, N Agarwal, R H Nussenzveig, G Rothstein, A Wilson, A Artz, and J T Prchal. 2008. “Hematopoiesis Is Not Clonal in Healthy Elderly Women.” *Blood* 112 (8): 3186–93. doi:10.1182/blood-2008-03-143925.

Szelinger, Szabolcs, Ivana Malenica, Jason J Corneveaux, Ashley L Siniard, Ahmet A Kurdoglu, Keri M Ramsey, Isabelle Schrauwen, et al. 2014. “Characterization of X Chromosome Inactivation Using Integrated Analysis of Whole-Exome and mRNA Sequencing..” Edited by Osman El-Maarri. *PLoS ONE* 9 (12): e113036. doi:10.1371/journal.pone.0113036.

Tennessen, J A, A W Bigham, T D O'Connor, W Fu, E E Kenny, S Gravel, S McGee, et al. 2012. “Evolution and Functional Impact of Rare Coding Variation From Deep Sequencing of Human Exomes.” *Science* 337 (6090): 64–69. doi:10.1126/science.1219240.

Thomasson, H R, H J Edenberg, D W Crabb, X L Mai, R E Jerome, T K Li, S P Wang, Y T Lin, R B Lu, and S J Yin. 1991. “Alcohol and Aldehyde Dehydrogenase Genotypes and Alcoholism in Chinese Men..” Edited by Francesc Palau. *American Journal of Human Genetics* 48 (4). Elsevier: 677–81. doi:10.1371/journal.pone.0018931.

Tietjen, Gary L, and Roger H Moore. 1972. “Some Grubbs-Type Statistics for the Detection of Several Outliers.” *Technometrics* 14 (3): 583–97. doi:10.1080/00401706.1972.10488948.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. doi:10.1038/nbt.1621.

Trapnell, Cole, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. 2013. “Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq..” *Nature Biotechnology* 31 (1): 46–53. doi:10.1038/nbt.2450.

Tucker, Elena J, Steven G Hershman, Caroline Köhrer, Casey A Belcher-Timme, Jinal Patel, Olga A Goldberger, John Christodoulou, et al. 2011. “Mutations in MTFMT Underlie a Human Disorder of Formylation Causing Impaired Mitochondrial Translation.” *Cell Metabolism* 14 (3): 428–34. doi:10.1016/j.cmet.2011.07.010.

Van Esch, H. 2005. “Deletion of VCX-a Due to NAHR Plays a Major Role in the Occurrence of Mental Retardation in Patients with X-Linked Ichthyosis.” *Human Molecular Genetics* 14 (13): 1795–1803. doi:10.1093/hmg/ddi186.

Van Esch, Hilde, Marijke Bauters, Jaakko Ignatius, Mieke Jansen, Martine Raynaud, Karen Hollanders, Dorien Lugtenberg, et al. 2005. “Duplication of the MECP2 Region Is a Frequent Cause of Severe Mental Retardation and Progressive Neurological Symptoms in Males..” *American Journal of Human Genetics* 77 (3): 442–53. doi:10.1086/444549.

Venâncio, Margarida, Mónica Santos, Susana Aires Pereira, Patrícia Maciel, and Jorge M Saraiva. 2007. “An Explanation for Another Familial Case of Rett Syndrome: Maternal Germline Mosaicism.” *European Journal of Human Genetics* 15 (8): 902–4. doi:10.1038/sj.ejhg.5201835.

Verhoeven, Willem M A, Jos I M Egger, David A Koolen, Helger Yntema, Simone Olgiati, Guido J Breedveld, Vincenzo Bonifati, Bart P.C. van de Warrenburg. 2014. “Beta-Propeller Protein-Associated Neurodegeneration (BPAN), a Rare Form of NBIA: Novel Mutations and Neuropsychiatric Phenotype in Three Adult Patients.” *Parkinsonism and Related Disorders* 20 (3). Elsevier Ltd: 332–36. doi:10.1016/j.parkreldis.2013.11.019.

Veytsman, Boris, Lei Wang, Tiange Cui, Sergey Bruskin, and Ancha Baranova. 2014. “Distance-Based Classifiers as Potential Diagnostic and Prediction Tools for Human Diseases..” *BMC Genomics* 15 Suppl 12: S10. doi:10.1186/1471-2164-15-S12-S10.

Vissers, Lisenka E L M, Joep de Ligt, Christian Gilissen, Irene Janssen, Marloes Steehouwer, Petra de Vries, Bart van Lier, et al. 2010. “A De Novo Paradigm for Mental Retardation.” *Nature Genetics* 42 (12). Nature Publishing Group: 1–5. doi:10.1038/ng.712.

Vitting-Seerup, Kristoffer, Bo Torben Porse, Albin Sandelin, and Johannes Waage. 2014. “spliceR: an R Package for Classification of Alternative Splicing and Prediction of Coding Potential From RNA-Seq Data.” *BMC Bioinformatics* 15 (1). BMC Bioinformatics: 1–7. doi:10.1186/1471-2105-15-81.

Wang, Eric T, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” *Nature* 456 (7221): 470–76. doi:10.1038/nature07509.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants From High-Throughput Sequencing Data..” *Nucleic Acids Research* 38 (16). Oxford University Press: e164–64. doi:10.1093/nar/gkq603.

Wang, Xu, Qi Sun, Sean D McGrath, Elaine R Mardis, Paul D Soloway, and Andrew G Clark. 2008. “Transcriptome-Wide Identification of Novel Imprinted Genes in Neonatal Mouse Brain.” Edited by Anne C Ferguson-Smith. *PLoS ONE* 3 (12): e3839. doi:10.1371/journal.pone.0003839.t001.

Wang, Z, M Gerstein, and M Snyder. 2009. “RNA-Seq: a Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10 (1). Nature Publishing Group: 57–63.

Wang, Zhihong, Aizhen Yan, Yuxiang Lin, Haihua Xie, Chunyan Zhou, and Fenghua Lan. 2013. “Familial Skewed X Chromosome Inactivation in Adrenoleukodystrophy Manifesting Heterozygotes From a Chinese Pedigree.” Edited by Bart Dermaut. *PLoS ONE* 8 (3): e57977. doi:10.1371/journal.pone.0057977.t003.

Welch, J S, P Westervelt, L Ding, D E Larson, J M Klco, S Kulkarni, J Wallis, K Chen, J E Payton, and R S Fulton. 2011. “Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene.” *JAMA: the Journal of the American Medical Association* 305 (15). Am Med Assoc: 1577–84.

Wu, Jiaxin, Yanda Li, and Rui Jiang. 2014. “Integrating Multiple Genomic Data to Predict Disease-Causing Nonsynonymous Single Nucleotide Variants in Exome Sequencing Studies.” Edited by Greg Gibson. *PLoS Genetics* 10 (3): e1004237. doi:10.1371/journal.pgen.1004237.s004.

Yang, F, T Babak, J Shendure, and C M Disteche. 2010. “Global Survey of Escape From X Inactivation by RNA-Sequencing in Mouse.” *Genome Research* 20 (5): 614–22. doi:10.1101/gr.103200.109.

Yang, Yaping, Donna M Muzny, Fan Xia, Zhiyv Niu, Richard Person, Yan Ding, Patricia Ward, et al. 2014. “Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing.” *JAMA: the Journal of the American Medical Association* 312 (18). American Medical Association: 1870–10. doi:10.1001/jama.2014.14601.

Yang, Yaping, Donna M Muzny, Jeffrey G Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, et al. 2013. “Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders.” *The New England Journal of Medicine*, October, 131002140031007. doi:10.1056/NEJMoa1306555.

Yeung, Ka Yee, and Walter L Ruzzo. 2001. “Principal Component Analysis for Clustering Gene Expression Data..” *Bioinformatics (Oxford, England)* 17 (9). Oxford University Press: 763–74. doi:10.1093/bioinformatics/17.9.763.

Young, Juan I, and Huda Y Zoghbi. 2004. “X-Chromosome Inactivation Patterns Are Unbalanced and Affect the Phenotypic Outcome in a Mouse Model of Rett Syndrome..” *American Journal of Human Genetics* 74 (3): 511–20. doi:10.1086/382228.

Zhang, Haibo, Ju Youn Lee, and Bin Tian. 2005. “Biased Alternative Polyadenylation in Human Tissues..” *Genome Biology* 6 (12). BioMed Central Ltd: R100. doi:10.1186/gb-2005-6-12-r100.

Zhang, Y, A Castillo-Morales, M Jiang, Y Zhu, L Hu, A O Urrutia, X Kong, and L D Hurst. 2013. “Genes That Escape X-Inactivation in Humans Have High Intraspecific Variability in Expression, Are Associated with Mental Impairment but Are Not Slow Evolving.” *Molecular Biology and Evolution*, September. doi:10.1093/molbev/mst148.

Zhou, Y H, K Xia, and F A Wright. 2011. “A Powerful and Flexible Approach to the Analysis of RNA Sequence Count Data.” *Bioinformatics (Oxford, England)* 27 (19): 2672–78. doi:10.1093/bioinformatics/btr449.

Zwiener, Isabella, Barbara Frisch, and Harald Binder. 2014. “Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures..” *PLoS ONE* 9 (1): e85150. doi:10.1371/journal.pone.0085150.

Ørstavik, K H, R E Orstavik, and M Schwartz. 1999. “Skewed X Chromosome Inactivation in a Female with Haemophilia B and in Her Non-Carrier Daughter: a Genetic Influence on X Chromosome Inactivation?.” *Journal of Medical Genetics* 36 (11): 865–66.

Ørstavik, Karen Helene. 2009. “X Chromosome Inactivation in Clinical Practice.” *Human Genetics* 126 (3): 363–73. doi:10.1007/s00439-009-0670-5.

APPENDIX A

CLINICAL DESCRIPTION OF PATIENTS

Family 0001, Patient 0001_1

Caucasian family of 6 with single affected female. This is a 14 year-old girl with a history of hypotonia, weakness and motor delay. Her parents and three sisters are unaffected.

Birth history: Second pregnancy for this mother, complicated by mild hypertension during the third trimester. Fetal movement may have been reduced compared to her first child. Delivery was normal and without complication, with normal Apgar scores.

By 3-4 months, she was noted to be floppy, with poor head control. Development was notably delayed by 8 months (she was babbling, rolling over, able to commando crawl, but not able to sit). Neurological evaluations and testing was started by 11 months of age. She was visually attentive and had no facial weakness. She couldn’t maintain sitting, she reached for objects without tremor, and there was a decrease in axial and appendicular tone. Tendon reflexes were present. EEG, blood count, metabolic profile, creatine kinase, lactate, ammonia, Acylcarnitine profile, plasma amino acids, very long chain fatty acids, and urine organic acids were all normal. At age 2, muscle biopsy did not lead to a specific diagnosis. Enzyme testing for mitochondrial disease was negative. Over the years, consistent findings on exam have been small stature, hypotonia, weakness, and decreased endurance. Her speech was dysarthric and difficult to understand, but fluent. She was able to walk for short periods, especially after a rest, but she fatigued quickly. She had a mild gaze apraxia, poor control of her neck muscles, and hypotonia especially at the shoulders and neck. She had a strong grip and briskly active tendon reflexes. Her heel cords were tight. Nerve conduction velocities were normal; EMG studies were suggestive of a mild myopathic process, without myotonia, or decremental response on repetitive stimulation. Edrophonium challenge test did not produce any change in her EMG or improve her strength. Cognitive development has always been normal. She never had trouble controlling bladder or bowel function.

At age 7, she had her first spinal fluid examination for neurotransmitter metabolites and pterins; this was reported as normal, but re-examination suggested mildly low homovanillic acid (HVA) and slightly elevated 3-ortho-methylDOPA. Therapeutic trial of L-DOPA/Carbidopa was not effective; trial of pyridostigmine was also not effective. She and her family were enrolled in the research study at the Neurogenetics Center at St. Joseph's Hospital and TGen.

At age 8 she was seen at the Mayo Clinic, and EMG (including single fiber EMG) was not consistent with a congenital myasthenia syndrome. A second muscle biopsy was not diagnostic. Features on examination again included hypotonia, neck muscle weakness, dystonic posturing of the feet and episodes of ocular dystonia. A second spinal fluid examination was done and again showed slightly low HVA and elevated 3-ortho-methyl-DOPA. A second trial of L-DOPA/Carbidopa was not effective. She continued to have spells of dystonia in her legs, and ocular dystonic attacks (oculogyric crises). She was getting weaker – by age 10, she was using a motorized wheelchair, was having trouble chewing, and was losing weight; the idea of placing a feeding gastrostomy tube was being contemplated.

At age 10.5 years, she was started empirically on a combination of bromocriptine and selegiline, based on the hypothesis that she might have a variant of AADC (aromatic amino-acid decarboxylase) deficiency. She had a dramatic response to this treatment – within 6 months, she was completely out of the wheelchair, and was able to walk to school and around school all day; she did not have any more episodes of falling.

At this time, she is active, can walk and run and dance; she does fatigue after a full day of school. Her primary problems now remain difficulty with speech, neck and lower back posture, and short stature.

She is maintained now only pramipexole, a dopamine receptor agonist, and has been taken off selegiline.

MRI scan: MRI at 3-4 months of age was normal. At age 2, a second MRI of the brain and spine was normal. A third brain MRI done at age 5 was normal. MRI of the C-spine done at age 7 was normal.

Molecular tests: see case description above.

Genetic tests: at age 7 *GCH1* gene testing was normal (for GTP cyclohydrolase deficiency). At age 8, gene testing did not detect mutation in the *TH* (tyrosine hydroxylase) gene.

Family 0002

Clinical Description in Chapter 4.

Family 0004, Patient 0004_1

Caucasian family of five with single affected male. Clinical diagnosis at time of enrollment was leukoencephalopathy, developmental delay, microcephaly, and intellectual disability. Patient could start to walk at age 6. Patient presents hypotonia, developmental delay, autism spectrum symptoms, feeding disorder, scoliosis, and microcephaly.

MRI scan: Normal brain MRI, EEG, and ERG.

Molecular tests: Enzyme tests were normal for lactate. Tests for plasma amino acids and urine organic acids were negative. Muscle biopsy showed normal histology with slightly increased cytochrome C oxidase level.

Genetic tests: Normal Fragile X, *MeCP2* sequence, FISH for Angelman, *UBE3A* sequence

Family 0005, Patient 0005_1

Caucasian family of 5 with single affected male. Clinical diagnosis at time of enrollment without a suspected causal gene is Pelizaeus-Merzbacher-like disease with nystagmus and motor delay. He presented motor delay and nystagmus at infancy with feeding disorder.

MRI scan: Her MRI initially thought to show abnormal myelination. Follow-up MRI showed T2 hyperintensity in dentate nucleus of cerebellum and bilateral thalamic signal abnormalities. ***Molecular tests:*** Lysosomal enzyme and very long chain fatty acids test was normal.

Genetic tests: *PLP*, *GJA12*, *GJC2* gene and duplication test was negative.

This patient was enrolled in the study presented in Chapter 4 and his mother was enrolled in the RNA-seq-HUMARA study presented in Chapter 2.

Family 0006, Patient 0006_1

This is a Middle Eastern family of four with X affected. Here clinical diagnosis at time of enrollment without a suspected gene is ataxia with sensory neuropathy, similar to Friderich's Ataxia. Parents are first cousins so parental consanguinity is suspected. Sister is also affected with NF1 disease and radius dysplasia.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

Family 0008, Patient 0008_1

Caucasian family of four with single affected female. Here clinical diagnosis at enrollment without a suspected causal gene is progressive leukoencephalopathy, spastic quadriparesis, global cerebral atrophy and neurodegenerative disorder. She presented feeding problems as an infant including colicky behavior and vomiting. She showed failure to thrive. She was diagnosed with cerebral palsy at age 2. At time of enrollment she presents contractures, ocular bobbing, myoclonic jerks but reflexes are present. Brain CT scan indicated progressive global atrophy without calcifications.

MRI scan: Abnormal. MRI at 8 month with white matter volume loss.

Molecular tests: not available

Genetic tests: She had negative Rett syndrome genetic test, and BAC array indicated no large structural variant in here genome.

This patient was enrolled in the study presented in Chapter 4. She and her mother were also presented in the RNA-seq-HUMARA study in Chapter 2.

Family 0011, Patient 0011_1

Caucasian family of three with single affected female. Her clinical diagnosis at enrollment without a suspected causal gene is Aicardi Syndrome. Prenatal ultrasound showed brain cysts and prenatal MRI was suggestive of agenesis of corpus callosum. Congenital “hydrocephalus”; s/p fenestration of cerebral cysts; subsequent third ventriculostomy at 3 months. Ophthalmology exam at 2 weeks showed choreoretinal lacunae a hallmark of Aicardi syndrome. She presented infantile spasms at 3 months of age and has severe developmental delay.

MRI scan: Post-natal MRI showed agenesis of corpus callosum.

Molecular tests: not available

Genetic tests: not available

This patient and here mother were enrolled in the study presented in Chapters 1 and 2.

Family 0012, Patient 0012_1

Caucasian family of four with single affected female and unaffected younger brother. Clinical diagnosis at enrollment was developmental delay, autism spectrum disorder. Genetic test identified *de novo* interstitial deletion 2q23.1 – q24.2. She had early feeding problems and failure to thrive. Presented delayed milestones. At age 3 she was diagnosed with Autism Spectrum Disorder. Clinical test showed normal EEG.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient participated in RNA-seq and HUMARA comparison of XCI ratio study described in Chapter 2.

Family 0016, Patient 0016_1

Asian family of four with single affected male and unaffected sister. Clinical diagnosis at enrollment was progressive cerebellar ataxia, dystonia. Patient walked at 15 months of age, and started talking at 2 years of age. Patient presents dysarthria, motor delay, progressive ataxia, and dystonia, tight heel cords. Spine X-ray is normal.

MRI scan: Normal.

Molecular tests: Lysosomal enzyme test, plasma amino acid, urine organic acid, acylcarnitines, creatine, guanidinoacetate all negative. CPK, alphafetoprotein, B12, ceruloplasmin test are normal. Muscle biopsy is normal.

Genetic tests: Fragile X, *MPS7*, mtDNA, Ataxia (recessive) panel, spinocerebellar ataxia gene panel was all negative. Array CGH found 68kb heterozygous deletion at 1p36.11.

This patient was enrolled in the DNA-RNA study in Chapter 4 and his mother was enrolled into the study presented in Chapter 2.

Family 0018, Patient 0018_1

Clinical description can be found in Chapter 2. This family was enrolled in she study presented in Chapter 2, and 3. The patient and her mother also participated in the RNA-seq-HUMARA study described in Chapter 2.

Family 0019, Patient 0019_1

Caucasian family of six with 3 affected and one unaffected children. Clinical diagnosis of female patient at time of enrollment was autosomal recessive non-progressive cerebellar ataxia infantile dystonia. Patient was delivered by C-section for failure to progress; immediately after delivery, was arching her back, eyes were rolled up. Arching and rigidity with involuntary eye movements continued as a neonate, continued until age 4 yrs.

Patient was delayed in motor milestones - rolling over at 1 year; sat up at 2 years; pulled to stand 3 years; walking at 4 years with gait trainer. Clinical examination at age 1.5-2 years showed hypotonia, preserved reflexes, tongue thrusting. Patient has poor writing skills, dysarthric speech, better cognition, and possible myopathy. Patient has normal nerve conduction and EMG.

Clinical test at 7 years of age, showed that she could walk with crutches, had broad based ataxic gait, able to climb, and movements are slow. She had low muscle tone, head lag, and action tremor; but presented no spasticity.

MRI scan: MRI shows mild cerebellar vermis atrophy.

Molecular tests: lactate, pyruvate, and lysosomal enzyme levels are normal.

Genetic tests: MeCP2, Spinal muscular atrophy (SMA) test negative.

This patient was enrolled in the study presented in Chapter 4 and her mother was enrolled into the study presented in Chapter 2.

Family 0020, Patient 0020_1

Caucasian family of 5 with single affected female and two unaffected brothers. Clinical diagnosis at time of enrollment was Neonatal progeroid disorder, failure to thrive. She has a feeding disorder, lipodystrophy, and cutis marmorata.

MRI scan: Normal.

Molecular tests: Plasma amino acids, urine organic acids, cholesterol, triglycerides, Acyl carnitine profile were all normal.

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 2.

Family 0023, Patient 0023_1

Caucasian family of three with single affected female. Clinical diagnosis at time of enrollment was infantile choreoathetosis, dystonia. She was born premature at 26-27 weeks, but she had a relatively normal NICU course. She presented apnea at 8 months of age and development regressed. She had developmental delay with rigidity, fisting, head lag, and hyperreflexia. She has near normal cognition.

MRI scan: She had two MRIs both of which were normal and EEG was normal as well. CT scan for calcification was negative.

Molecular tests: Urine amino acids, organic acids normal. Total and free plasma carnitine tests were normal. Copper and ceruloplasmin were normal. CSF neurotransmitter metabolites, tetrahydropbiopterin/neopterin profile, methyltetrahydrofolate, amino acids were all normal. Lysosomal enzymes were also normal.

Genetic tests: *MeCP2* point mutation, deletion and duplication gene test was negative. Congenital Disorder of Glycosylation was negative.

This patient was enrolled in the study presented in Chapter 2.

Family 0024, Patient 0024_1

Middle Eastern family of seven with single affected male and four unaffected siblings. Clinical diagnosis at enrollment was Aicardi-Goutieres Syndrome. Suspected parental consanguinity.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4 and his mother was enrolled into the study presented in Chapter 2.

Family 0025, Patient 0025_1

Caucasian family of four with single affected male and unaffected female sibling. Clinical diagnosis at enrollment was suspected feeding disorder, choreoathetosis. He had neonatal feeding difficulty and showed no reaction to pain. As a neonate he had megacystis, and urinary retention. He presented hypotonia and delayed motor development with poor head control. He has bladder-emptying problem.

MRI scan: Brain MRI showed cavum septum pellucidi. MRI of spine was normal. EMG normal

Molecular tests: Urine acyl glycines, urine organic acids, and TORCH titers all normal. Normal serum CPK, lactate, ammonia all negative. Plasma short chain fatty acids showed mild ketonemia. CSF neurotransmitter metabolites, tetrahydrobiopterin, neopterin, methyltetrahydrofolate were all normal. Muscle biopsy showed normal ETC complex enzyme activity.

Genetic tests: Array CGH was negative for copy number changes.

Family 0029, Patient 0029_1

Caucasian family of seven with two affected siblings (male, and female) and three unaffected brothers. Clinical diagnosis at time of enrollment was Aicardi-Goutieres Syndrome.

The female patient had normal development until age 17 months, when developmental regression and spastic quadriparesis developed.

MRI scan: MRI showed delayed myelination. CT scan showed no calcifications.

Molecular tests: Lysosomal enzymes test was normal. CSF neopterin was slightly elevated.

Genetic tests: FISH for Pelizaeus-Merzbacher disease was negative, spastic paraparesis panel was also negative. Gene test in SMAHD1 found a heterozygous variant in exon 12 at I448T.

This patient was enrolled in the study presented in Chapter 2.

Family 0033, Patient 0033_1

Caucasian family of 5 with single affected female and unaffected brother and sister. Clinical diagnosis at time of enrollment was Aicardi syndrome. She started to present seizures at 10 weeks of age. She has choreoretinal lacunae.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and she with her mother was presented in Chapter 2.

Family 0034, Patient 0034_1

Hispanic family of 5 with single affected female and a brother and an unaffected sister. Clinical diagnosis at time of enrollment was Aicardi Syndrome.

Prenatal ultrasound was suggestive of ventriculomegaly, and possible agenesis of corpus callosum. At birth diagnosed with asymmetric ventriculomegaly, agenesis of the corpus callosum, and L microphthalmia with optic nerve dysplasia. She smiled at 2 months of age. Myoclonic seizures and infantile spasms in clusters presented at 3.5 months. Seizures were controlled with vigabatrin and valproate. Ophthalmology exam showed microphthalmia, bilateral optic nerve colombomas with variable size, choreoretinal lacunae surrounding optic nerves in both eyes, sparing fovea. Combination of infantile spasms with agenesis of corpus callosum and optic nerve coloboma/choreoretinal lacunae led to diagnosis of Aicardi Syndrome.

MRI scan: MRI showed dilation of posterior portion of ventricles, left more than right; third ventricle was elevated. CT scan showed features of agenesis of corpus callosum, and colpocephaly. EEGs showed slowing and bursts of epileptiform activity, primarily from left hemisphere.

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and she with her mother was presented in Chapter 2.

Family 0046, Patient 0046_1

African American family of 5 with single affected female and two half-sisters unaffected. Clinical diagnosis at time of enrollment was Aicardi Syndrome. Pregnancy was normal, prenatal ultrasound suggested agenesis of corpus callosum. Prenatal MRI showed partial agenesis of corpus callosum MRI at birth partial agenesis of corpus callosum. Seizures were noted at 6 weeks of age; ophtho exam at 3 months of age showed retinal lacunae and consequently was diagnosed with Aicardi syndrome. She has intractable epilepsy and spams like seizures.

MRI scan: Brain MRI at 3 months of age also showed cortical dysplasia in left frontal lobe and partial agenesis of corpus callosum with preserved genu and anterior body of corpus callosum.

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 2.

Family 0047, Patient 0047_1

Caucasian family of five with single affected female. Clinical diagnosis at time of enrollment was Aicardi Syndrome. She is a high functioning Aicardi patient. She walked at 2.5 years and uses a few single words; finger feeds, and is able to use the toilet. Normal birth. She showed infantile spasms and retinal lesions at 3 months of age. She has seizures.

MRI scan: MRI showed only a small remnant of the splenium of corpus callosum; posterior fossa arachnoid cyst requiring shunting;

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and she with her mother was presented in Chapter 2.

Family 0048, Patient 0048_1

Caucasian/Filipino family with a single affected female. Clinical diagnosis at time of enrollment was Aicardi Syndrome. She had infantile spasms and seizures starting at 10 weeks of age.

She has global developmental delay, failure to thrive, trunkal hypotonia, scoliosis with trunkal curvature. She has the Aicardi characteristic of lacunae

MRI scan: MRI showed absent corpus callosum, heterotopic gray matter in frontal lobes, intracranial cysts, ventricular dilatation ex vacuo, and abnormal sulcation patter.

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and she with her mother was presented in Chapter 2.

Family 0049, Patient 0049_1

Caucasian family of five with single affected female. Clinical diagnosis at enrollment was Cockayne syndrome or Cerebro-Oculo-Facio-Skeletal Syndrome (COFS type 2).

She presents severe delay in motor and cognitive development.

She has intrauterine growth retardation, congenital cataracts, congenital nystagmus, specifically continuous rotary and horizontal nystagmus. She has microcephaly, developmental delay, hypotonia and dystonia. She has Dysphagia, failure to thrive, and scoliosis.

MRI scan: MRI shows diffuse T2 hyperintensities in the entire white matter indicative of leukoencephalopathy.

Molecular tests: Plasma amino acids, urine organic acids, extended newborn screen all negative.

Lysosomal enzymes, very long chain fatty acids are negative.

Genetic tests: Congenital disease of glycosylation screen is negative. 3. FISH for Prader-Willi syndrome was negative.

This patient was enrolled in the study presented in Chapter 4, and she with her mother were presented in Chapter 2.

Family 0059, Patient 0059_1

Caucasian family with single affected female. Her clinical diagnosis at enrollment without a suspected causal gene is Aicardi Syndrome. She presented seizures at 3 months of age. She has a cyst in the brain and is getting smaller. She has preserved, almost entire corpus callosum She has lacunae in one eye and her vision is improving.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and she with her mother was presented in Chapter 2.

Family 0091, Patient 0091_1

Caucasian family of four with single affected male and unaffected female sibling. Clinical diagnosis at enrollment was Schizophrenia, which was diagnosed at age 7. He has language difficulties and did not start to speak until age 4. He is also presenting symptoms characteristics of Bipolar disorder. He has a propensity for violence and aggression. He has commanding auditory and visual hallucinations. EEG was normal.

Maternal grandmother and her sister as well as maternal great-grandmother diagnosed with schizophrenia. Asperger's runs on father's side of family.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and his mother and grandmother presented in Chapter 2.

Family 0103, Patient 0103_1 and 0103_2

Clinical description can be found in Chapter 2. These patients were enrolled in the study presented in Chapter 2 and 3.

Family 0117, Patient 0117

Caucasian family of three with single affected male. Clinical diagnosis at enrollment was Pelizaeus–Merzbacher-like disease (leukodystrophy) with no candidate genes identified.

Patient present nystagmus, hypotonia, delayed development. Limited speech, can only say about 5 words, but can use signs.

MRI shows diffuse lack of myelination of subcortical white matter, but with time some improvement, especially in the genu of corpus callosum; atrophy of splenium of corpus callosum

Molecular tests: urine organic acids negative.

Genetic tests: PLP1, GJA12, CDG screening is negative.

CT scan negative for calcifications.

This patient was enrolled in the study presented in Chapter 4, and his mother presented in Chapter 2.

Family 0118, Patient 0018_1

Caucasian family with single affected female. Clinical diagnosis at enrollment was Aicardi Syndrome.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient and her mother were enrolled in the study presented in Chapter 2.

Family 0139, Patient 0139_1

Caucasian family of three with single affected male. Clinical diagnosis at enrollment was not available. Self reported case. Patient has very small stature and features. Total situs inversus, wide set eyes, small low set ears, developmental and speech delays.

Patient is unable to feed through mouth, VP shunt and g-tube and fundo.

He has chronic lung disease, immotile cilia syndrome.

MRI scan: not available

Molecular tests: not available

Genetic tests: not available

This patient was enrolled in the study presented in Chapter 4, and his mother presented in Chapter 2.

Family 0140, Patient 0140_1

This is an adopted, Caucasian female patient. Clinical diagnosis at enrollment was suspected Dystonia.

She has abnormal gait and posture that fluctuates without weakness; left leg with choreoathetotic or dystonic posture; can walk one minute and then is crawling the next because she cannot walk

She was evaluated for Torticollis. She has low set ears flat nasal bridge.

MRI scan: MRI showed stable very small 2.2mm syrinx at T12-L1. EMG was normal.

Molecular tests: CSF was normal.

Genetic tests: *DYT1* mutation was negative. Array cGH for insertions deletions was negative.

This patient was enrolled in the study presented in Chapter 2.

Family 0152, Patient 0152_1

Caucasian family with single affected male. Clinical diagnosis at enrollment was Leigh's Syndrome with suspected causal mechanism. This is and old Amish family. Patient can crawl, sit without support, pull to stand, and cruise. He can say about 12 words and continues to expand.

MRI scan: MRI showed bilateral signal intensity alterations involving the anterior aspect of the subthalamic regions and the substantia nigra and pars reticularis of the mid brain. CT scan was normal

Molecular tests: spectroscopy suggested subtle alterations in the lactate profile.

Genetic tests: indicated that the child had a regions of homozygosity (ROH) across 9 chromosomes. Genes associated with Mitochondrial complex I deficiency/Leigh's syndrome are in these regions (*NDUFAF2* and *NDUFS3*).

This patient was enrolled in the study presented in Chapter 4 and her mother was enrolled in a study presented in Chapter 2.

Family 0157, Patient 0157_1

Caucasian family with affected female. Clinical diagnosis at enrollment was not available. Patient presents delayed development, Autism Spectrum behavior. She has normal skin.

MRI scan: MRI showed non-specific symmetric prominence of T2-weighted high signal within the eperitrigonal white matter; delay in myelination or dysmyelination.

Molecular tests: patient has borderline low levels of vitamin A level and undetectable DHEAS level.

Genetic tests: FISH study showed a de novo 1.6 Mb deletion at Xp 22.31 (6,456,510-8,077,333)

APPENDIX B

WHOLE EXOME SEQUNECING METRICS

							PCT target bases covered			
ID	Study Chapter	Gender	Capture assay	HQ Read (M)	PCT Aligned reads	Mean target coverage	2X	10X	20X	30 X
0001_1	Ch2|Ch4	F	TruSeq	39.22	0.840	19.4	0.8	0.59	0.39	0.23
0001_3	Ch2|Ch4	M	TruSeq	127.96	0.842	54.8	0.9	0.91	0.82	0.71
0001_2	Ch2|Ch4	F	TruSeq	128.11	0.835	55.4	0.9	0.91	0.82	0.71
0001_4	Ch2	F	TruSeq	129.77	0.895	91.8	0.9	0.95	0.92	0.88
0001_5	Ch4	F	TruSeq	131.60	0.899	94.9	0.9	0.95	0.92	0.89
0001_6	Ch4	F	TruSeq	147.49	0.899	104.5	0.9	0.95	0.93	0.90
0002_5	Ch2|Ch4	M	TruSeq	193.32	0.872	124.4	0.9	0.95	0.92	0.89
0002_2	Ch2|Ch4	F	TruSeq	135.25	0.877	87.3	0.9	0.93	0.89	0.84
0002_4	Ch2|Ch4	M	TruSeq	125.71	0.872	80.3	0.9	0.93	0.88	0.82
0002_1	Ch2|Ch4	F	TruSeq	138.80	0.884	85.2	0.9	0.92	0.87	0.82
0002_6	Ch4	M	TruSeq	114.83	0.884	71.6	0.9	0.91	0.84	0.77
0002_3	Ch2|Ch4	F	TruSeq	123.56	0.884	77.4	0.9	0.92	0.86	0.79
0008_1	Ch2|Ch4	F	TruSeq	152.24	0.880	100.0	0.9	0.94	0.90	0.86
0008_3	Ch4	M	TruSeq	141.72	0.876	94.8	0.9	0.94	0.90	0.85
0008_2	Ch2|Ch4	F	TruSeq	129.64	0.883	84.9	0.9	0.93	0.89	0.83
0008_4	Ch4	F	TruSeq	133.06	0.883	87.7	0.9	0.93	0.89	0.84
0012_3	Ch4	M	TruSeq	92.04	0.878	54.3	0.9	0.90	0.82	0.72
0012_2	Ch4	F	TruSeq	94.63	0.876	56.5	0.9	0.91	0.83	0.73
0016_1	Ch4	M	TruSeq	62.47	0.868	30.0	0.9	0.82	0.64	0.45
0016_3	Ch4	M	TruSeq	142.58	0.895	101.2	0.9	0.95	0.93	0.89
0016_2	Ch2|Ch4	F	TruSeq	146.68	0.898	101.6	0.9	0.95	0.92	0.90
0018_1	Ch2|Ch3|Ch	F	TruSeq	138.58	0.883	81.7	0.9	0.91	0.86	0.79
0018_3	Ch2|Ch3|Ch	M	TruSeq	133.88	0.878	77.8	0.9	0.91	0.85	0.77
0018_2	Ch2|Ch3|Ch	F	TruSeq	145.77	0.882	84.9	0.9	0.92	0.87	0.80
0019_1	Ch2|Ch4	F	TruSeq	90.53	0.873	54.6	0.9	0.91	0.83	0.72
0019_4	Ch4	F	TruSeq	105.24	0.875	68.2	2.2	0.96	0.92	0.85
0019_5	Ch4	M	TruSeq	134.00	0.872	83.4	0.9	0.93	0.90	0.84
0019_3	Ch4	M	TruSeq	91.15	0.873	53.3	0.9	0.91	0.83	0.71
0019_2	Ch2|Ch4	F	TruSeq	94.00	0.871	57.4	0.9	0.91	0.84	0.74
0019_6	Ch4	F	TruSeq	99.06	0.871	60.3	0.9	0.91	0.85	0.76
0023_1	Ch2|Ch3	F	TruSeq	134.41	0.884	74.9	0.9	0.91	0.85	0.78
0023_3	Ch2|Ch3	F	TruSeq	164.02	0.882	94.1	0.9	0.92	0.87	0.82
0023_2	Ch2|Ch3	M	TruSeq	152.04	0.876	82.4	0.9	0.91	0.86	0.80
0024_1	Ch2|Ch4	F	TruSeq	120.12	0.881	76.7	0.9	0.91	0.86	0.79
0024_3	Ch4	M	TruSeq	112.98	0.885	71.5	0.9	0.91	0.84	0.75
0024_2	Ch2|Ch4	F	TruSeq	130.30	0.891	83.2	0.9	0.92	0.86	0.80
0025_1	Ch2|Ch4	F	TruSeq	156.44	0.882	92.3	0.9	0.92	0.87	0.82
0025_3	Ch4	M	TruSeq	147.60	0.878	86.8	0.9	0.92	0.86	0.80
0025_2	Ch2|Ch4	F	TruSeq	176.04	0.869	102.9	0.9	0.93	0.89	0.84

0025_4	Ch4	F	Agilent	129.23	0.978	140.1	0.9	0.98	0.97	0.95
0025_5	Ch4	M	Agilent	134.79	0.970	147.1	0.9	0.98	0.97	0.95
0029_1	Ch2	M	TruSeq	115.83	0.888	82.9	0.9	0.91	0.87	0.82
0029_2	Ch2	F	TruSeq	104.61	0.881	73.1	0.9	0.89	0.82	0.75
0029_3	Ch2	M	TruSeq	144.50	0.871	102.5	0.9	0.93	0.90	0.86
0029_4	Ch2	F	TruSeq	124.04	0.877	87.9	0.9	0.92	0.89	0.85
0029_5	Ch2	M	TruSeq	151.58	0.840	43.5	0.9	0.89	0.76	0.60
0033_1	Ch2|Ch4	F	TruSeq	139.65	0.900	100.2	0.9	0.95	0.92	0.89
0033_3	Ch2|Ch4	M	TruSeq	145.74	0.894	102.5	0.9	0.95	0.93	0.89
0033_2	Ch2|Ch4	F	TruSeq	146.21	0.896	104.4	0.9	0.95	0.93	0.90
0034_1	Ch2|Ch3|Ch	F	TruSeq	134.89	0.899	97.8	0.9	0.95	0.92	0.89
0034_3	Ch2|Ch3|Ch	M	TruSeq	121.93	0.897	87.7	0.9	0.94	0.92	0.88
0034_2	Ch2|Ch3|Ch	F	TruSeq	121.11	0.900	81.4	0.9	0.94	0.91	0.87
0046_1	Ch2	F	TruSeq	112.76	0.874	45.4	0.9	0.88	0.76	0.63
0046_2	Ch2	F	TruSeq	149.68	0.886	79.5	0.9	0.92	0.86	0.79
0047_1	Ch2|Ch4	F	TruSeq	96.12	0.875	58.4	0.9	0.91	0.84	0.74
0047_3	Ch2|Ch4	M	TruSeq	98.56	0.872	61.4	0.9	0.91	0.85	0.76
0047_2	Ch2|Ch4	F	TruSeq	119.45	0.877	73.4	0.9	0.92	0.88	0.81
0048_1	Ch2|Ch4	F	TruSeq	96.96	0.871	58.3	0.9	0.91	0.84	0.75
0048_3	Ch2|Ch4	M	TruSeq	93.18	0.873	55.6	0.9	0.91	0.83	0.73
0048_2	Ch2|Ch4	F	TruSeq	106.08	0.869	61.3	0.9	0.91	0.85	0.77
0049_1	Ch2|Ch4	F	TruSeq	140.43	0.877	71.4	0.9	0.91	0.85	0.77
0049_3	Ch4	M	TruSeq	134.70	0.873	68.5	0.9	0.91	0.84	0.76
0049_2	Ch2|Ch4	F	TruSeq	125.76	0.878	65.0	0.9	0.90	0.83	0.74
0059_1	Ch2|Ch4	F	TruSeq	143.69	0.879	103.1	0.9	0.94	0.92	0.89
0059_3	Ch4	M	TruSeq	152.70	0.875	110.6	0.9	0.94	0.92	0.89
0059_2	Ch2|Ch4	F	TruSeq	151.30	0.883	108.4	0.9	0.94	0.92	0.89
0091_1	Ch2|Ch4	F	TruSeq	125.67	0.876	78.5	0.9	0.93	0.89	0.81
0091_3	Ch4	M	TruSeq	170.94	0.872	105.6	0.9	0.94	0.92	0.87
0091_5	Ch2	F	Agilent	143.42	0.978	154.3	0.9	0.98	0.97	0.95
0091_2	Ch2|Ch4	F	TruSeq	130.68	0.882	83.3	0.9	0.94	0.90	0.83
0091_4	Ch2	F	TruSeq	158.58	0.873	100.2	0.9	0.94	0.91	0.86
0103_1	Ch2|Ch4	M	TruSeq	144.44	0.891	109.0	0.9	0.93	0.91	0.87
0103_2	Ch2|Ch4	F	TruSeq	123.90	0.891	91.5	0.9	0.93	0.89	0.85
0103_4	Ch2|Ch4	M	TruSeq	143.73	0.889	106.4	0.9	0.93	0.90	0.87
0103_3	Ch2|Ch4	F	TruSeq	140.95	0.885	104.7	0.9	0.93	0.90	0.87
0117_1	Ch4	M	TruSeq	135.07	0.884	95.9	0.9	0.94	0.91	0.86
0117_3	Ch4	M	TruSeq	126.66	0.882	92.0	0.9	0.93	0.90	0.86
0117_2	Ch2|Ch4	F	TruSeq	127.41	0.889	91.7	0.9	0.93	0.90	0.86
0118_1	Ch2	F	TruSeq	135.49	0.887	98.1	0.9	0.93	0.90	0.87
0118_2	Ch2	F	TruSeq	142.71	0.886	102.9	0.9	0.94	0.91	0.87
0139_1	Ch4	M	TruSeq	161.13	0.873	86.0	0.9	0.94	0.91	0.87
0139_3	Ch4	M	TruSeq	103.65	0.880	58.1	0.9	0.92	0.87	0.79
0139_2	Ch2|Ch4	F	TruSeq	136.84	0.877	73.3	0.9	0.93	0.90	0.85

0140_1	Ch2	F	TruSeq	151.16	0.877	80.8	0.9	0.94	0.90	0.86
0152_1	Ch2|Ch4	M	Agilent	84.85	0.961	96.8	0.9	0.98	0.95	0.90
0152_3	Ch4	M	Agilent	66.82	0.978	78.0	0.9	0.97	0.93	0.86
0152_2	Ch2|Ch4	F	Agilent	93.09	0.979	109.0	0.9	0.98	0.96	0.92
0157_1	Ch2|Ch4	F	Agilent	110.45	0.978	122.7	0.9	0.98	0.96	0.94
0157_3	Ch4	M	Agilent	112.65	0.980	128.9	0.9	0.98	0.97	0.94
0157_2	Ch2|Ch4	F	Agilent	103.48	0.979	116.9	0.9	0.98	0.96	0.93

APPENDIX C

RNA-SEQ SEQUENCING METRICS

Id	Study Chapter	Gender	RIN	Reads HQ	HQ Base	Bases mapp	Median CV	5’-3’ Bias
0001_4	Ch2	F	7.8	154.7	13.1	12.8	0.39	0.65
0001_1	Ch2|Ch4	F	8.6	119.6	10.0	9.6	0.37	0.74
0001_3	Ch2|Ch4	M	8.6	117.1	9.3	9.0	0.37	0.76
0001_2	Ch2|Ch4	F	8.8	82.2	7.4	7.1	0.38	0.77
0002_1	Ch2|Ch4	F	8.9	108.3	9.1	8.8	0.38	0.75
0002_2	Ch2|Ch4	F	8.9	152.5	12.7	12.3	0.38	0.75
0002_3	Ch2|Ch4	F	8.9	125.9	10.4	10.0	0.37	0.72
0002_4	Ch2|Ch4	M	9.4	106.0	9.2	8.9	0.37	0.74
0002_5	Ch2|Ch4	M	8.9	127.3	9.6	9.3	0.38	0.81
0002_6	Ch4	M	9	162.1	12.8	12.4	0.38	0.71
0004_1	Ch4	M	8.2	128.3	8.2	8.0	0.41	0.67
0004_2	Ch2|Ch4	F	7.5	230.6	14.5	14.0	0.42	0.67
0004_3	Ch4	M	4.8	184.8	10.3	10.0	0.50	0.48
0005_1	Ch4	M	9.1	117.4	9.8	9.5	0.38	0.74
0005_3	Ch4	M	8.9	164.8	12.0	11.6	0.39	0.71
0005_2	Ch4	F	8.4	126.3	11.4	11.0	0.39	0.73
0006_1	Ch4	F	8.5	111.5	9.4	9.1	0.37	0.73
0006_3	Ch4	M	8.7	98.8	7.1	6.9	0.38	0.76
0006_2	Ch4	F	8.8	109.9	8.9	8.6	0.37	0.77
0008_1	Ch2|Ch4	F	7.7	116.7	9.0	8.7	0.45	0.57
0008_3	Ch4	M	7.3	134.2	10.5	10.2	0.42	0.63
0008_2	Ch2|Ch4	F	7.3	166.3	11.0	10.6	0.44	0.58
0011_2	Ch2|Ch3|Ch4	F	7.1	128.3	8.7	8.4	0.41	0.69
0011_3	Ch3|Ch4	M	5.9	142.1	9.4	9.2	0.45	0.59
0011_1	Ch2|Ch3|Ch4	F	8.8	126.9	9.1	9.1	0.38	0.74
0012_1	Ch2	F	8.1	135.5	10.1	9.7	0.39	0.72
0014_2	Ch2|Ch3|Ch4	F	8.6	132.9	10.5	10.1	0.41	0.71
0014_3	Ch3|Ch4	M	8.3	141.2	10.0	9.6	0.40	0.69
0014_1	Ch2|Ch3|Ch4	F	8.5	135.5	9.6	9.5	0.39	0.70
0016_1	Ch4	M	8.5	120.6	10.0	9.7	0.38	0.71
0016_2	Ch2|Ch4	F	8.5	159.1	11.1	10.7	0.38	0.70
0016_3	Ch4	M	8.7	102.4	8.5	8.2	0.40	0.74
0018_1	Ch2|Ch3|Ch4	F	7.5	109.3	8.6	8.4	0.39	0.74
0018_2	Ch2|Ch3|Ch4	F	8	174.8	14.5	14.1	0.39	0.74
0018_3	Ch2|Ch3|Ch4	M	7.9	111.0	9.5	9.2	0.39	0.73
0019_1	Ch2|Ch4	F	8.1	177.1	12.6	12.2	0.40	0.72
0019_2	Ch2|Ch4	F	8.8	152.2	9.9	9.5	0.38	0.73
0019_3	Ch4	M	7.3	124.3	9.8	9.5	0.43	0.65
0020_1	Ch2	F	8.5	122.0	10.1	9.7	0.37	0.72
0023_1	Ch2|Ch3	F	8	151.7	10.3	10.0	0.39	0.72
0024_1	Ch2|Ch4	F	7.4	145.9	11.3	10.9	0.39	0.74

0024_2	Ch2|Ch4	F	7.5	116.6	9.5	9.2	0.39	0.73
0024_3	Ch4	M	8.1	104.8	8.9	8.7	0.40	0.69
0025_1	Ch2|Ch4	F	5.1	239.8	14.7	14.0	0.49	0.56
0025_3	Ch4	M	5.9	130.8	8.5	8.2	0.46	0.65
0025_2	Ch2|Ch4	F	6	198.7	12.0	11.4	0.46	0.62
0033_2	Ch2|Ch4	F	6.3	123.1	10.0	9.9	0.43	0.72
0033_3	Ch2|Ch4	M	6.4	111.2	8.0	7.9	0.45	0.71
0033_1	Ch2|Ch4	F	8.7	110.1	9.3	9.2	0.43	0.68
0034_2	Ch2|Ch3|Ch4	F	8.6	104.9	9.0	8.9	0.41	0.80
0034_3	Ch2|Ch3|Ch4	M	7.9	118.8	9.4	9.3	0.42	0.77
0034_1	Ch2|Ch3|Ch4	F	9	112.7	9.9	9.9	0.39	0.75
0046_1	Ch2		8.8	102.3	8.7	8.3	0.37	0.76
0047_3	Ch2|Ch4	M	7.8	108.8	9.0	8.9	0.38	0.75
0047_2	Ch2|Ch4	F	7.5	114.2	9.2	9.1	0.41	0.65
0047_1	Ch2|Ch4	F	8.8	120.5	8.7	8.6	0.38	0.76
0048_2	Ch2|Ch4	F	6	109.8	8.2	8.1	0.45	0.60
0048_3	Ch2|Ch4	M	6.1	165.9	10.4	10.3	0.44	0.62
0048_1	Ch2|Ch4	F	6.9	135.0	8.8	8.7	0.43	0.65
0049_1	Ch2|Ch4	F	7.5	149.7	10.8	10.7	0.40	0.70
0049_2	Ch2|Ch4	F	7	124.3	9.2	9.1	0.41	0.61
0049_3	Ch4	M	6.8	110.8	9.0	8.9	0.42	0.69
0059_1	Ch2|Ch4	F	7.3	64.6	5.3	5.3	0.43	0.63
0059_2	Ch2|Ch4	F	7	148.4	10.8	10.7	0.41	0.69
0059_3	Ch4	M	7.7	137.0	9.8	9.7	0.40	0.71
0091_1	Ch2|Ch4	F	7.3	116.5	9.4	9.4	0.43	0.66
0091_2	Ch2|Ch4	F	7.1	127.8	9.5	9.4	0.43	0.63
0091_3	Ch4	M	7.6	132.0	9.1	9.0	0.43	0.66
0091_4	Ch2	F	7.1	102.0	7.9	7.8	0.43	0.72
0091_5	Ch2	F	5.6	194.5	12.1	11.4	0.39	0.68
0103_4	Ch2|Ch4	M	6.9	128.1	10.9	7.4	0.46	0.63
0103_3	Ch2|Ch4	F	8.9	45.2	3.6	2.4	0.47	0.67
0103_2	Ch2|Ch4	F	7	76.0	9.6	6.4	0.45	0.66
0103_1	Ch2|Ch4	M	6.9	26.8	2.8	2.0	0.44	0.70
0103_2	Ch2|Ch4	F	7	76.0	9.6	6.4	0.45	0.66
0103_3	Ch2|Ch4	F	8.9	45.2	3.6	2.4	0.47	0.67
0103_4	Ch2|Ch4	M	6.9	128.1	10.9	7.4	0.46	0.63
0117_1	Ch4	M	8.4	101.8	7.3	7.2	0.40	0.69
0117_2	Ch2|Ch4	F	7.9	185.3	13.2	13.1	0.41	0.69
0117_3	Ch4	M	8.5	88.1	6.4	6.4	0.44	0.64
0118_1	Ch2	F	8.1	137.8	9.8	9.7	0.39	0.76
0118_2	Ch2	F	8.4	116.6	9.3	9.2	0.39	0.71
0139_1	Ch4	M	9.6	202.8	13.4	12.8	0.37	0.79
0139_3	Ch4	M	5.2	200.6	13.5	13.0	0.40	0.71
0139_2	Ch2|Ch4	F	7.7	183.7	11.9	11.4	0.53	0.59

0140_1	Ch2	F	7.7	235.2	16.1	15.5	0.39	0.79
0152_1	Ch4	M	8.1	163.3	13.4	12.2	0.40	0.73
0152_3	Ch4	M	7.4	231.4	15.5	14.7	0.40	0.73
0152_2	Ch2|Ch4	F	8.1	99.6	8.3	7.4	0.40	0.71
0157_1	Ch2|Ch4	F	7.1	199.1	13.1	12.2	0.43	0.69
0157_3	Ch4	M	7	181.2	12.1	11.3	0.41	0.71
0157_2	Ch2|Ch4	F	7.6	174.6	12.4	11.5	0.42	0.71