Winter 12-2019

Finding and Analyzing de novo Mutations in the Exomes of Parent-Offspring Trios of Spontaneous Chiari Malformation Type 1 Patients

Brian Leon Ricardo
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Finding and Analyzing *de novo* Mutations in the Exomes of Parent-Offspring Trios of Spontaneous Chiari Malformation Type 1 Patients

by

Brian X. León-Ricardo

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

December 2019

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank DBBS and the CSB program for the support during the research done for this thesis. Dr. Christina Gurnett, Dr. Gabriel Haller and Dr. Brooke Sadler who and the rest of the Gurnett & Dobbs lab. Dr. Gurnett gave me a chance during a time of transition between labs and has never failed to support me through the problems that have occurred during my time in her lab. The support, encouragement and conversations about science with her and the rest of the lab have enriched my time at Washington University.

Dr. Skeath and Dr. Shadding who advised me throughout my whole journey and welcomed to Washington University through IMSD. I would not have made it to this point without their calm heads and good advise.

Celine St. Piere, Kayla Nygaard, Caryn Carson, and Juan Macias have been invaluable peers who have proofread, edited, brainstormed and made me laugh when most needed. They have become a family away from home.

Finally, I'd like to thank my family, there is no amount of space where I could fully describe and account their support, but I hope this simple thanks and acknowledgment gets me a bit closer.

Brian X. Léon Ricardo

Washington University in St. Louis

December 2019

Dedicated to my mother,
Ondina Ricardo Orozco
Making it this far would not be possible without your sacrifice, support, and love.

ABSTRACT OF THE THESIS

Finding and Analyzing *de novo* Mutations in the Exomes of Parent-Offspring Trios of Spontaneous Chiari Malformation Type 1 Patients

by

Brian X. León-Ricardo

Masters of Arts Degree in Biological Sciences

Computational and Systems Biology

Washington University in St. Louis, 2019

Professor Christina Gurnett

Chiari Malformation Type 1 (CM1) is a neurodevelopmental disorder that occurs when one of the cerebellar tonsils herniates past the foramen magnum causing headaches, motor or sensory deficits, sleep apnea, and difficulty swallowing. This disorder is estimated to affect 1% of the population but due to the need of neuroimaging for diagnosis and the presence of asymptomatic patients there is still uncertainty about the exact proportion of the population affected. CM1 often presents itself with other neurodevelopmental disorders such as syringomyelia, scoliosis, and known genetic syndromes such as Klippel-Feil and Marfan syndromes. Twin, family, and familial clustering studies have established a genetic component to CM1 etiology, but have failed to ascertain any specific causal gene. The difficulty ascertaining causal genes, its comorbidity with multiple different syndromes, and the complex symptomatology of its patients indicate genetic heterogeneity. Other neurodevelopmental disorders with genetic heterogeneity such as Autism Spectrum Disorder and Epileptic Encephalopathies have had success finding genes of interests by looking for *de novo* mutants

(DNMs) from spontaneous patient trios. With this in mind, we sequenced the exomes of a cohort of 29 offspring-parent trios affected with CM1 in search of candidate causative DNMs. Using previously established methods that predict which variants in the exome are DNMs, we found 44 variants that passed multiple filtering steps for quality, likelihood of being real DNMs, and potential to be causative. Three of these variants were classified as stopgain which made them likelier to be detrimental. These three were validated and analyzed for their potential role in CM1 risk. From thousands of possible variants, we successfully obtained a shortlist of genes to further study in future studies.

Chapter 1: Chiari Malformation Type 1

1.1 Description and Symptoms

Chiari Malformation Type 1(CM1) is a developmental disorder that occurs as a result of structural defects where a cerebellar tonsil herniates into the spinal cord from the base of the skull. This herniation can be acquired or be congenital and is highly comorbid with syringomyelia, a fluid-filled cyst within the spinal cord, and scoliosis, a sideways curvature of the spine.[1–5] Normally, the cerebellar tonsils are located inside the skull in an area known as the posterior cranial fossa. In CM1 patients, however, the cerebellar tonsils protrude more than 5 mm into the spinal canal.[1,3,6,7] Many individuals with CM1 are asymptomatic, but symptomatic patients report a multitude of complaints.[6] The majority experience headaches, affecting 66% to 87% of patients. Other symptoms include motor or sensory deficits in extremities, oropharyngeal dysfunctions, sleep apnea, nausea, and ataxia.[3,7]

The exact cause of CM1 has not been established but a common theory is that due to variation in the size of the posterior fossa it can be be too small for the cerebellum causing what is called cramping.[6,8] Not all cases feature this cerebellar cramping which contributes to the theory of CM1 being multifactorial.[6,8,9] Although not always predictive, it has also been observed that more severe herniation of the cerebellar tonsils increases the likelihood of patients showing or developing symptoms.[10] Cerebellar tonsils position in the skull can change slightly during growth from childhood to young adulthood, which has been suggested as a reason for why some CM1 cases resolve with age.[9] In cases where symptoms are severe and early onset, patients can be treated by decompression surgery which increases space in the skull. CM1 primarily diagnosed with neuroimaging (i.e. CT scans or MRI) and it's detection is often

incidental. Otherwise, it is difficult to diagnose due to the non-specific and common nature of its symptoms and the high variance of symptoms between patients.[7]

1.2 Prevalence of Chiari Malformation Type 1

Since neuroimaging is the only certain method for diagnosis, the exact prevalence of CM1 in the population is difficult to calculate. The prevalence of CM1 is also higher in pediatric versus adult patients, therefore estimates will vary based on the composition of the sampled population.[10] Despite these challenges, multiple studies have attempted to ascertain the proportion of the population affected by CM1 with current estimates ranging from 0.77%[6] to 3.6%.[10] The study with the largest sample size is a retroactive image analysis by Aitken et al. (2009), which evaluated 5,248 pediatric scans and found that 1% had CM1 on neuroimaging. However, this approach is biased since it only sampled children that were symptomatic or were scanned for a separate medical reason, and the study may then over-represent the fraction of children with CM1. The Rotterdam Project, which investigated many neurological abnormalities, used a less biased approach that analyzed the MRI images of 2000 adult subjects with a mean age of 63.3 years without selecting for specific symptoms or phenotype. This study found CM1 in 0.9% of the population.[11] Although CM1 prevalence estimates vary, approximately 1 in 100 people are affected, though many of those are asymptomatic.[6] CM1 is one of the most common neurodevelopmental disorders, and the most severe cases can be the source of complications and pose a near lethal threat in cases that severe and early onset.[12]

1.3 Comorbidities with CM1: Syringomyelia, Scoliosis and Genetic Syndromes

CM1 often presents alongside other neurodevelopmental disorders, the most common being syringomyelia and scoliosis. Syringomyelia occurs when cerebrospinal fluid (CSF) collects in the spinal cord or brain stem creating a cyst. The observed comorbidity rates for CM1 with syringomyelia have ranged from 23%[13] to 85%[10]. This large range is likely due to patient heterogeneity, the decreased severity of CM1 with age, and incomplete patient data. Nevertheless, publications reporting on the surgical intervention of CM1 have established that many syringomyelia cases resolve along with CM1, providing evidence that both diseases occur from a similar structural defect.[14]

Scoliosis, another neurodevelopmental disorder characterized by a curvature of the spine, is also often associated with CM1. The frequency of CM1 patients with comorbid scoliosis ranges broadly from 18% to 50% in cohorts comorbid with connective tissue disorders.[15] One publication of individuals younger than 5 years and presenting with Scoliosis and CM1 reported that decompression surgery had successfully treated both even though the surgery was not performed to alleviate scoliosis.[16] Similarly to syringomyelia, this adds to the evidence suggesting a relationship between the structural causes for scoliosis and CM1.

In addition to syringomyelia and scoliosis, 20 other syndromes and disorders have been found to be comorbid with CM1, such as the two connective tissue disorders Marfan Syndrome and Klippel-Feil.[17] The large number of other diseases comorbid with CM1 suggests that a given number of the cases that are not spontaneous might be the result of multifactorial inheritance and indicates a genetic contribution to its etiology. Additionally, family-based and twin studies have been found to also suggest a genetic contribution to CM1, even though few causative genes or mutations have been identified.

1.4 Evidence From Twin and Family Studies

In a recent study, Abbot *et al.* looked for reported CM1 cases in two healthcare providers and matched them to their pedigrees from the Utah Population Database. They calculated relative risk (RR) which is commonly used to find evidence supporting genetic contribution to a given phenotype. The RR works as a measurement of likelihood of finding an individual with CM1 in a cohort given the presence of an affected relative. This is calculated by counting CM1 cases to obtain a cohort-specific population disease rate. In turn, this rate is used to predict the number of CM1 cases expected given the cohort size and compared to those observed in the form of an unbiased ratio.[18]

When looking at first- (parents and siblings), second- (aunts and grandparents) and third- (cousins) degree relatives, the authors calculated a RR of 4.54, 1.20 and 1.36 respectively. The RR scores for first- and third-degree relatives were significantly elevated with *p*-values smaller than 0.001, indicating familial clustering of CM1. They also theorized that the reason why second-degree relatives were not significantly elevated was due to reduced ascertainment in older generations from lack of neuroimaging.[18]

A case study by Nagy *et al.* (2016) focused on a family with five confirmed CM1 cases and eight additional individuals with symptoms, the highest recorded number of CM1 cases in a single family to date. This family also exhibited a high incidence of a rare inherited connective tissue disorder, Ehlers-Danlos syndrome that affects skin and joints. The cases of Ehlers-Danlos were correlated with CM1, suggesting a possible link between these disorders.[19]

Finally, Speer *et al.* (2003) reported six cases of twins where one or both had CM1. Of these, three cases were monozygotic twins with concordant CM1 (two of which also had syringomyelia), one was a set of dizygotic twins with concordant CM1, and the other two were dizygotic twins where only one twin was affected.[20] Together these family studies provide more

support for a genetic contribution for CM1 without the need to identify a specific candidate gene or genes responsible for the malformation.

1.5 Genetic Association and Linkage of Chiari Type 1

Family-based studies, twin studies, and comorbidity with syndromes of known genetic origin support a genetic contribution to CM1, but they do not identify possible causative genes or mutations. To identify genes linked to CM1, Boyles *et al.* (2006) performed a linkage analysis of 23 families with 71 affected CM1 individuals and were able to find two regions on chromosomes 9 and 15 associated with CM1. The chromosome 15 locus included a gene, *FBN1*, that encodes an extracellular matrix (ECM) protein previously associated with the connective tissue disorder Marfan Syndrome, which is also associated with CM1.[17,21] In another whole genome linkage study on CM1, Markunas *et al.* (2013) increased their detection power by stratifying their cohort of 367 individuals into patients that did or did not show signs of connective tissue disorders. Using this approach they were able to pinpoint two regions on chromosome 8 associated with CM1.[22] These regions contained the growth factors *GDF6* and *GDF3*, which have previously been implicated in Klippel-Feil syndrome, another connective tissue disorder associated with CM1.[17] Using the same dataset, but stratified based on cranial base morphometrics, they also found significant association between CM1 and regions on chromosomes 1 and 22.[23]

A case-control association study for CM1 with 451 patients and 524 controls identified 18 SNPs and 14 genes as possible candidates. While many of the SNPs and genes did not overcome multiple correction testing, four SNPs within *CDX1*, *FLT1*, and *ALDH1A2* were marginally significant.[24] The same group also conducted a joint eQTL analysis using expression

data from the blood, cranial tissue and dura mater of 43 individuals with CM1 and identified 239 genes with a highly significant correlated expression in both tissues. This study identified three genes (*IPO8*, *XYLT1*, and *PRKAR1A*) as potential candidates of CM1 etiology due to their function in osteoblast differentiation, alterations in which could contribute to a smaller posterior fossa.[25] Finally, two other studies independently investigating copy number variants in locus 16p11.2 found that one of it's three patients had CM1 along with their other neurodevelopmental disorders and the second uncovered CM1 as one of the most common associated abnormalities out of a cohort of 246 individuals.[26,27]

More recent studies have focused on finding rare variants with strong effects. Two of these studies used exome sequencing to detect variants in CM1 patients. The first study in 2016 evaluated the child-parent trio exome of a boy with CM1 and Dent disease Type 2 (DD2). By sequencing his exome and those of his parents, they found a de novo deletion in *OCRL1*, the gene causal for DD2, as well as a rare inherited mutation in the *INPP5B* protein family.[28] *OCRL1* and *INPP5B* are paralogs that play a role in ciliogenesis, suggesting that disrupting ciliogenesis can lead to CM1. Another CM1 study looked at the exomes of two parent-offspring trios and the exome of 65 sporadic CM1 cases. In the two families, they found three heterozygous missense variants in *DKK1*, *LRP4*, and *BMP1*. These are part of the *WNT* and bone morphometric protein (*BMP*) pathways functionally responsible for the normal development of the posterior fossa, suggesting a link between *WNT* and *BMP* signaling with CM1. They also looked for genetic variants that were located on the three genes of interest in the exomes of the 65 sporadic cases and found variants in two of them.[29]

1.6 Genetic Heterogeneity in Chiari Malformation Type 1

The association and exome studies performed so far are just beginning to shine a light on the genes and pathways involved in CM1 etiology. These studies have found genes involved in different biological functions, along with genes that are causal to other syndromes, e.g. Marfan syndrome. Nevertheless, the association, linkage, and exome studies failed to establish strong certainty to any of the genes and pathways they found in their data. Chiari Malformation Type 1 research is not unique to these challenges and similar trends have been observed by groups researching other neurodevelopmental disorders such as Autism Spectrum Disorder (ASD) and Epileptic Encephalopathies (EE).[30] The possibility of many genetic changes being able to produce similar disease phenotypes fits as a reason for the difficulty finding overlap between the genes implicated from the previously mentioned papers. In which case predictability for diagnosis of the disease for genetic counseling and understanding its basic biology will continue to be a difficulty. As CM1 research continues an individual approach to a collective of CM1 cases could benefit the search for the etiology of the disease.[31] In contrast to ASD and EE, CM1 benefits from higher sensitivity when identifying patients since a herniation of the cerebellum is more quantitative than behavioral traits or diseases that appear intermittently. In that way even though so far the genetic heterogeneity has been the major challenge, bigger studies and the compiling of the data of each will hold the key to understanding CM1.

Chapter 2: Finding *de novo* Mutants in Patient-Parent Exome Trios

2.1 Abstract

Chiari Malformation Type 1 (CM1) is a neurodevelopmental disorder that occurs when one of the cerebellar tonsils herniates past the foramen magnum causing headaches, motor or sensory deficits, sleep apnea, and bowel and bladder incontinence. CM1 is classified as a neurodevelopmental disorder with a likely heterogeneous genetic etiology due to the complex symptomatology of affected patients. Other disorders with similar complex genetic relationships, such as Epileptic Encephalopathy (EE) and Autism Spectrum Disorder (ASD), have used trio exome studies to find an enrichment of *de novo* mutations (DNMs) in specific genes and pathways associated with these disorders, suggesting specific mutations, genes, and pathways involved in the pathogenesis of EE and ASD. We analyzed the exome of 29 parent-offspring trios of early onset and severe CM1 cases and called the genetic variants for each of them. All variants were annotated by finding their corresponding gene, pathogenicity and allele frequency data from previously published online databases. Using two different DNM prediction algorithms, TrioDeNovo and DeNovoGear, we calculated a score to predict real DNMs from false positives. The variants with the highest likelihood of being real DNMs were further analyzed for functional roles and any interaction with genes previously implicated as candidate. The top variants with high confidence of being real and with a good potential of being causal were Sanger validated.

2.2 Introduction

Our understanding of the genetics of CM1 is still in its infancy. With only a handful of genetic loci and genes implicated to CM1 etiology, and even fewer direct connections to the biological processes, we have much to learn. Future studies on CM1 should be shaped and informed by successful approaches observed in other similar neurodevelopmental diseases. Previous groups have gleaned important insight into the genetic basis of other neurodevelopmental diseases by looking for spontaneous cases of diseases and using whole exome sequencing (WES) of parent-offspring trios to identify potential causal *de novo mutations* (DNM) in coding regions.[32] This approach has successfully identified putative causative mutations in a number of neurodevelopmental disorders, such as Epileptic Encephalopathies (EE) and Autism Spectrum Disorder (ASD), regardless of the genetic heterogeneity characteristic of them .[32–35]

The challenge of disorders with genetic heterogeneity

These two disorders are defined by a group of phenotypic traits or symptoms, and patients diagnosed with them can exhibit some or all of the traits. Some of these defining traits or symptoms can be phenotypically very similar but caused by different genotypes. One way in which this happens is when two changes in gene function result in different molecular consequences affecting an overall network that control a single phenotype or function. This variance of ASD and EE genotypes causing similar phenotypes is the described genetic heterogeneity.[30] As an example, four publications have found loss of function *de novo* variants in different exons of the *SCN2A* gene of ASD patients. This gene encodes a subunit of the sodium voltage-gated ion channels which play a crucial role in neuronal activity.[34,36–38] Genetic heterogeneity poses a challenge to association and linkage studies by reducing their power of detection for relevant variants. For these two methods to be successful they rely on the

population being analyzed being relatively homogeneous or a method for good stratification, all of which are difficult to accomplish for disorders of heterogeneous backgrounds and reduces the number for disorders that are hard to ascertain. To compound the problem even more, older methods of doing these analysis will only common variants to be causative. These common variants will only explain a percentage of the sampled affected population which increases the difficulty even more. In the case of ASD, it's been observed that common variants explain about 40% of cases in Autism and more recently DNMs are being found to also play a causal role.[30,32,39] All these difficulties would be part of any endeavor to find CM1 causative genes using this methods, since both CM1 and ASD have an estimated 1% prevalence and have shown genetic heterogeneity.

Two solutions for the loss statistical power when detecting causal genes in a cohort are increasing the number being studied to thousands or increasing the homogeneity of the cohort being analyzed. Of these two options, ASD research has been successful doing both by increasing its cohort sizes and creating more stringent cohorts, e.g. the Simmon Simplex Collection. In some cases attempting to increase specificity by removing ambiguous cases and creating distinct phenotypic subsets is much more difficult and impractical solution. Unlike autism CM1 is ultimately defined by an empirical measurement of herniation of the cerebellar tonsils and although ascertainment bias still exist having less false positive cases increases the specificity overall.

Success finding candidate causal genes in EE and ASD trios

Epileptic encephalopathies refer to a broad group of conditions characterized by recurring epileptic seizures along with developmental and learning disabilities. Multiple approaches have been used to find the underlying genetic causes of EE, with six GWAS studies resulting in the identification of eight candidate loci.[40] In parallel, the Epi4K consortium and the Epilepsy

Phenome/Genome Project performed WES analysis of a cohort of 264 unaffected parents and affected offspring trios and found 329 confirmed *de novo* variants that potentially contribute to EE.[35] An excess of DNMs were uncovered in genetic regions intolerant to genetic changes and some were found in genes that had been previously associated with EE, such as *STXBP1* and *SCN1A*, providing confidence in their approach.[41] In contrast, a study using only ten trios of individuals with epilepsy of unknown etiology and unaffected parents found and confirmed 15 DNMs in nine of the ten patients. Variants were located in genes such as *SCN1A, CDKL5* and *EEFIA2* which were previously associated with early onset EE.[42] In this study the confirmed DNMs were often found in patients exhibiting the earliest and most severe symptoms, suggesting the most severe early onset cases are likelier to be caused by DNMs.

Autism Spectrum Disorder is another neurodevelopmental disorder with a complex genetic origin. Like EE and CM1, there is evidence of genetic contributions to its etiology from twin and family studies, but more progress has been made in finding associated and causal genes in ASD development. In early experiments, GWAS studies calculated the role of common variants in autism heritability, finding only modest associations in individual variants but collectively explaining 15% to 50% of ASD cases.[43] Later whole exome and whole genome sequencing of autism patient trios led to the discovery of enrichment for *de novo* mutations in the genomes of autism patients.[44] As of 2019, lists of genes associated with ASD have reached 253, with additional rare CNVs and SNVs being reported regularly.[45] This presents us with the next big challenge for these studies: If hundreds of genes and variants are potentially associated with ASD, how do you prioritize them? Some studies have focused on aggregating the data from multiple projects, and finding over-representation of not just single nucleotide DNMs but of copy-number variants and functional enrichment from protein-protein interaction and co-

expression networks.[45] These studies used online databases of projects like ClinVar and ExAC, tissue type expression analysis (TSEA), and network analysis like HumanBase and the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) to identify the biological processes relevant to ASD etiology.[46,47] Most of the identified mutations are missense mutations likely to be disruptive to the gene, and they are present in pathways such as synaptic functioning, chromatin remodeling, WNT signaling, transcriptional regulation, interactions with FMR1 and, more broadly, MAPK signaling.[34,45,48]

Success has also been found for exome trio projects searching for DNMs. Sanders *et al.* (2012) used the Simon Simplex Collection (SSC) of trios to study 238 families, where exomes of unaffected parents, affected children, and an unaffected sibling were sequenced in 200 of the families, found 125 DNMs in the affected probands. They focused on 13 that were nonsense or in splice sites, of which only one gene had a DNM in two unrelated probands. They modeled the probability of finding two of these mutations in brain-expressed genes by chance and found that, with 150,000 iterations and $p = 0.008$, it was significantly unlikely.[36] Another successful project by Al-Mubarak *et al.* (2017) featured a much smaller sample size with sequenced exomes of 19 trios from Saudi Arabia. Similar to the SSC group, these cases were all sporadic. They found DNMs in 17 of the 19 trios, of which 3 were missense DNMs confirmed by Sanger sequencing and previously associated with ASD in the literature.[49]

Increasing the likelihood of finding causal DNMs by specific cohort selection

The studies reviewed illustrate how pairing trio studies with WES can be used to find DNMs. The two key parameters to the success of a WES approach are read depth and sample size. Since the average human genome is expected to have 1 to 2 DNMs in coding sequences, finding these variants requires a high enough depth such that artifact of sequencing and variants calling can be overcome.[50] There also need to be sufficient trios that meet diagnostic criteria in

order to reduce the effect of genetic heterogeneity, and when focusing on severe cases it is expected that missense and potential gene disruptive DNMs are most likely responsible.[32,39,51] When severe neurodevelopmental diseases appear spontaneously in a family, normal Mendelian inheritance is expected to be disrupted. One reason for the expectations it the lack of family history for the disease. Additionally, in cases of severe and early onset neurodevelopmental disorders fitness is reduced, making mutations likely to be lost due to purifying selection. From the biological reasoning and trends observed in ASD and EE a proportion of causative alleles for neurodevelopmental disorders will very likely be de novo, especially so for the most severe, early onset cases.

2.3 Methods

Patient Recruitment and Sequencing

Trios were composed of Chiari Malformation Type 1 patients and their parents recruited from the St. Louis Children's Hospital Neurology and Neurosurgery Clinics. Patients were selected for participation based on criteria that increased the chances of finding *de novo* variants involved in the etiology of the Chiari Malformation Type 1 patients. All the patient's came from families with no known or reported history of CM1, and one of two criteria. A cerebellar tonsil herniation greater than greater than 10 mm, considered severe, or showed symptoms from before the age of 10, considered early-onset. DNA was extracted from cheek swabs from all the members of the trios. To create an exome sequencing library we used the the 65-Mb Illumina Tru-Seq Exome enrichment kit to select the coding DNA sequences for selective amplification. The libraries were paired-end sequenced on an Illumina HiSeq 2000 to at minimum average of 30x coverage per trio.

Sequence Alignment, Quality Control and Variant calling

The reads were mapped to the human genome version GRCh37/hg19 using the Burrows-Wheels Aligner.[52] Quality control and variant calling for the mapped reads was done using the Genome Analysis Toolkit (GATK) version following the steps outline in it's use manual online and published in the Depristo *et al.* 2011 article.[53] All 102 trio members were processed by the GATK pipeline alongside 5,829 other exomes. These exomes were made into libraries and sequenced with similar protocols, and processed for quality control and mapped using the same methods. The reason to include the CM1 trios in a joint variant calling method of so many other samples was to increase the certainty and quality of the genetic variants detected. By calculating

population wide statistics the software does a better job to distinguish between the real biological signals and those from library or sequencing specific artifacts.

In brief, the aligned reads go through four major steps, base quality score recalibration (BQSR), haplotype calling with the gVCF method, and variant quality score recalibration (VQSR). In the BQSR step the individual quality scores of each nucleotide reads are recalculated after obtaining an error rate specific to three sources of covariation: the reported quality score, the cycle of the sequencing when the base was read, and the dinucleotide context of each base.[53] The SNP calling step is done by HaplotypeCaller which is described in their original protocol. This method calculates the likelihood of each particular genotype in all the locations of the genome using the quality of the mapping as variables to measure certainty.[54] It then merges all the individual samples into a single joint called project using its "GenotypeGVCFs" algorithm.[55] This method generates De Bruijn Graphs and uses a Hidden Markov Model algorithm to decide the identity of the base in that position for each sample taking into consideration the identity of the reference and all other samples in that position.[55] Lastly, the VQSR step decides which variants are real by calculating a Gaussian Mixed Model, comprised of two distributions, one for real SNPs and another for calling errors. Using an expectation maximization algorithm it decides which are real variant calls and which are probable computational artifacts. After the model is created, it calculates the likelihood of each called SNP belonging to the the real distribution versus the false positive one.[53] Once all these steps were were finished, an output list of the variant genotypes for each of the samples was created in Variant Calling Format (VCF). Once the full GATK pipeline was completed the exomes for the 34 trios of CM1 patients were extracted. The output also contained multiple statistics describing the number of reads used, and measuring the quality and certainty of the

genotypes called in the variant positions. The software also calculated scores measuring the bias between reference and alternate allele reads on the quality of their mapping, the position of the variant on each sequencing read and the frequency and quality of the strand sequenced supporting each of the alleles. Some of these statistics were used before processing to reduce false positives by selecting for variants that had a minimum depth (DP) of 8 for each member of a trio, an allele balance (AB) between 0.20 and 0.80 for the heterozygous calls, and a genotype quality score higher or equal to 20. All other calculated statistics were used in later steps to discern the confidence of the genotypes for the variants of interest.

Finding *de novo* single nucleotide variants with TrioDeNovo and DeNovoGear

To find *de novo* mutations (DNMs) from the called SNVs I used three different methods. The first approach to find DNMs relied on simple filtering by finding a genotype on the patients that were incompatible with Mendelian inheritance, i.e. variants were the affected patient in the trio was heterozygous while the parents were homozygous. Other than that, I also chose only those likely to be causal by keeping only non synonymous mutations. The other two approaches were two different algorithms, called TrioDeNovo (TDN) and DeNovoGear (DNG), that calculate certainty of variants being DNMs given their quality statistics.[56,57] The mathematical frameworks used to create these algorithms were tested using variants obtained from older and more basic software. As the quality of variant calling increases, many of the false positives are remove before getting to these DNM predicting steps. Nothing guarantees that these software will outperform basic filtering of variants that have been heavily processed by GATK. If successful the combination of GATK and these software would reduce the computational time and the number of false positives. To implement this we used the filtered variants as input for TDN and DNG.

TrioDeNovo gave each variant a *de novo* quality (DQ) score using the default parameters. This score is the log odds of the reads from a potential DNM belonging to a mutation model that reflects a *de novo* event over a model that supports reads from a polymorphism:

$$DQ = \log_{10}\left(\frac{P(Reads|Model_1)}{P(Reads|Model_0)}\right)$$

The TDN algorithm uses the Genotype Likelihood (GL) calculated by GATK for each candidate DNM, the alternative allele frequencies from the data of the 1000 Genome Project, and derived a simplification for the absolute allelic mutation rate and transition rate of each mutation model. The full derivation for the simplifications of the mutation rates of each model are outside the scope of this thesis but can be found in the algorithm's publication Wei *et al.* 2015.

DeNovoGear shares the same goal of TDN but takes a likelihood approach to obtain a probability for each found potential DNM. It used the GATK outputted GL for each member of the trios, the calculated probabilities of transmission from the parents to the child given their read depth information and mapping quality, and calculated the prior probabilities of observing a polymorphism or a DNM at any site in the genome from user provided mutation frequencies. We used both their default prior probability values of 1.0×10^{-3} and 1.0×10^{-8} for the polymorphism rate and the DNM rate respectively. Their polymorphism prior was calculated from fitting a beta-binomial likelihood model to the 1000 Genome Project and the mutation rate was calculated from validated DNMs in this same dataset.[50,57]

Annotation of the candidate DNMs

Along with the output from *de novo* variant prediction algorithms, the candidate DNMs were annotated with data from multiple online databases and the GATK calling software. The information added for each of the variants can be classified into four broader categories: mutation quality statistics, mutation type annotations, population frequency from large genome

projects, and potential pathogenicity. All except the quality statistics from GATK were compiled and put together using the Annovar software package.[58] Random variants were selected and viewed in the UCSC Genome Browser to double check their annotations. The quality statistics that were extracted from the GATK using GATK's VariantToTable function. The output selected included mapping quality, base quality, depth of sequencing for the position, allele specific depth and allele balance for the heterozygous calls, and multiple different bias tests. The bias test assigned each variant a significance score by comparing the difference between mapping quality and base calling quality of the reads used to call an alternate allele versus those that matched the reference genome. An additional bias test checked if the alternate allele supporting reads were clustered at the 5' end of them. This helps correct for systematic sequencing errors that become more likely after a certain number of cycles in the pyrosequencing.

Mutation type annotations included the genes were the variants resided, the functional annotation, e.g. exonic or intronic, and the specific exonic functional change, e.g. synonymous, non-synonymous, or stop gain. The gnomAD and ExAC databases were downloaded from Annovar's repository and were searched by the software for matches to any of the variants selected as candidate DNMs.[gnomad and exac publications] For pathogenicity annotation CADD version 1.3 was used to obtain scores for all given positions.[cadd ref] The other pathogenicity scores were obtained from the Database of Non-synonymous Functional Prediction (dbNSFP) which compiles scores for non-synonymous mutations from SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, fitCons, PhyloP and SiPhy.[59]

Merging of the full dataset

The GATK, DNG, TDN, and Annovar output were merged using a custom made algorithm coded in the Python3.5 language. Although tools like Bedtools are well established and can perform the merging of multiple genomic regions in short periods of time, the streamlining of these steps is crucial for a manageable process. A custom script allowed to correctly manage multiple alleles in the same location from different trios, account for variants that did not have output from some of the annotations, and perform verification that each unique trio and variant position pair was correctly kept. The python output generates a dataframe that is compliant with the R language's data reading functions and ready to be visualized in any data array or calculation sheet processor regardless of operating system. All the software code and steps are available on my online repository[1].

Dataset filtering in R and curation with online databases

To filter and generate a final list of candidate DNMs of interest, we used the programming language R version 3.4.4 for statistical analysis and RStudio version 1.2 for visualization. Other than the base functions included with R we used the svglite, ggforce, ggplot2, ggpubr, and wesanderson libraries to complement figure generation and the reshape2, limma, tidiverse, export, mclust, and mixtools for additional statistical tools. The annotated source code used for all the processing is on my repository[1] . The filters created with the the code classified each candidate into certain groups. The groups were those that passed or failed the VQSR, those that were processed by TDN or DNG, the subset the previous group with high confidence scores for each algorithm, those that passed the manual filtering schema, and annotation groups by functional role and exonic effect for those that fell in gene exons. For the group of high certainty sets, the cutoff was selected by finding the lowest value before the highest scoring population group within the distribution of scores, i.e. the last local minima

1 https://github.com/bxleon/dnmScripts.git

before the peak for the highest score in the distribution density plots. The resulting variants of high quality and high confidence were then sorted by trio and for each trio we searched for variants concordant between algorithms, high pathogenicity prediction scores or with a non-synonymous functional annotation. Those likely pathogenic DNMs were individually processed by searching in the HumanBase and UCSC genome browser databases. The search focused on finding significance in tissue specific expression, protein-protein, co-expression, and theoritical interactions, membership to groups of know molecular function, and any phenotypic associations.

2.4 Results

Cohort phenotype descriptions and total number of variants

Out of the 34 parent-offspring exomes of CM1 patients collected only 30 fit the criteria of having no family history and one or both severity indicators. The severity indicators chosen were an age of diagnosis younger than 10 to classify as early onset and herniation of at least twice the diagnostic cutoff of 5mm for severe herniation. The 30 CM1 patients had an mean age of 5.9 years and a mean herniation of 14 mm. From the cohort, 20 (66.7%) of them were severe enough to require surgical intervention, 15 (50%) had a spinal syrinx, and 4 (13%) also had scoliosis. We had symptom incidence information for 26 of the 30 CM1 patients, of which 20 (54%) were affected with headaches and 8 (31%) had some degree of difficulty swallowing. The data can be found byin **Table 2.1**.

After joint calling the exomes from the 30 trios, there were over 200,000 variant sites. This was reduced to 879 total with a mean of 30 .3 variants per trio after filtering for quality and keeping variants inconsistent with the parental genotypes, also known as mendelian violations. We called these variants candidate DNMs since anywhere from 0 to 3 of them are expected to be real and were the input for both *de novo* mutation predicting software.[44,50] The count of each candidate DNM per trio is visualized in **Figure 2.1** by the purple bars.

Table 2.1 Symptoms and comorbidity of the CM1 cohort [a]Age of diagnosis, the ones with asterisks are estimated based on time of appointment. Green values are below our threshold for early onset of younger than 10 [b] Millimeter of cerebellar tonsil herniation. Green pass our threshold of twice the 5 mm cutoff [c] 1 for presence of a syrinx, 0 for absence. [d] 1 for presence of scoliosis, 0 for absence. [e] 1 for cases severe enough to require surgery, 0 for no surgery. [f] 1 for reported headaches, 0 for not reported [g] 1 for reported problems swallowing or gagging, 0 for not reported.

Family ID	Age [a]	Herniation[b]	Syrinx[c]	Scoliosis[d]	Surgeries[e]	Headache[f]	Swallowing[g]
1008	14	28	0	0	1	1	0
1013	15	12	1	0	1	1	0
1040	8	22	0	0	1	1	1
1076	3	9	1	0	0	0	0
1091	4	10	1	0	0	0	0
1095	13	25	1	0	1	1	1
1096	14	11	1	0	1	1	0
1098	3	9	1	0	1	0	0
1112	3	23	0	0	1	1	1
1113	15	10	1	0	1	1	0
1132	4	9	0	0	0	0	0
1136	7	17	1	0	1	1	0
1142	< 9*	20	1	0	1	—	—
1143	1.1	6	0	0	0	0	0
1150	13	16	1	0	1	1	1
1155	1.9	7	0	0	0	0	0
1177	3	—	1	1	0	0	0
1180	8	13	1	1	1	0	0
1199	< 10*	—	1	0	1	—	—
1203	1	12	0	0	0	—	—
1207	2	14	0	0	0	0	1
1210	4	17	0	0	0	—	—
1238	4	15	1	1	1	0	0
1261	2	9	0	0	1	0	1
1266	2	14	0	0	1	1	1
1272	2	18	0	0	1	1	0
1280	6	7	1	1	1	0	0
1309	7	13	0	0	0	1	0
1332	2	16	0	0	1	1	0
1338	2	10	0	0	1	1	1

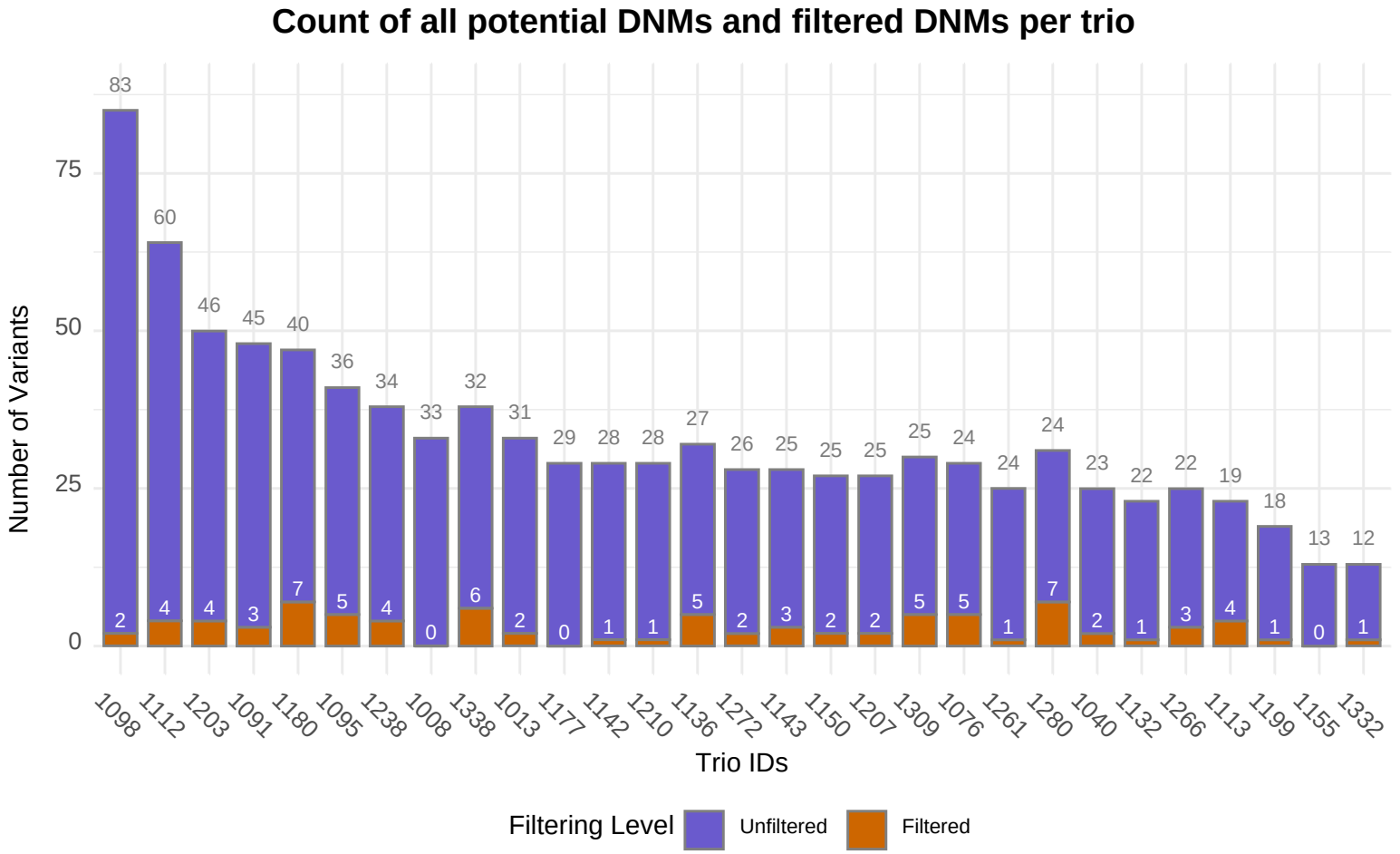

Figure 2.1 Bar plot of candidate and high quality variants per Trio
The purple bars indicate the total number of called variants in the offspring that was incompatible with the parental genotypes. The total count is displayed on top of each. The orange bars inside the purple bars are the subset after filtering for variants with good quality, and above the threshold of 0.80 for DNG scores and 11.5 for TDN scores.

A threshold of 11.5 was selected for the high certainty TrioDeNovo values

As established in the methods, TrioDeNovo's calculates a statistic for certainty of a DNM being a true positive and calls it the DQ score. These DQ scores are the log of the likelihood of the variant being a DNM over the likelihood of the null, where the variant is an artifact. Since the scores have no upper limit, the decision of what the cutoff to separate high likelihood candidate DNMs from low was obtained by looking at the distribution of all scores. From the 879 candidate DNMs the TrioDeNovo algorithm calculated DQ scores for 683 (73%), and dismissed the rest because of low confidence or nonsensical result. The mean of the scores is 7.21, but the cutoff used was that of 11.5. The value was decided after observing **Figure 2.2** and selecting the peak density with the highest DQ value. From the plotted density of the distribution of scores we can see two clear peaks, one around the DQ score of 5.9 and the second around the DQ score of 12.5. We expect the real DNMs to be those in the second group with DQ scores distributed around the second peak. For this reasons we selected a cutoff of 11.5, dashed orange line in the figure, and classified anything above it as a member of the high quality TrioDeNovo set, (hqTDN). The hqTDN set has 39 variants (6% of the total) from 20 of the 29 trios.

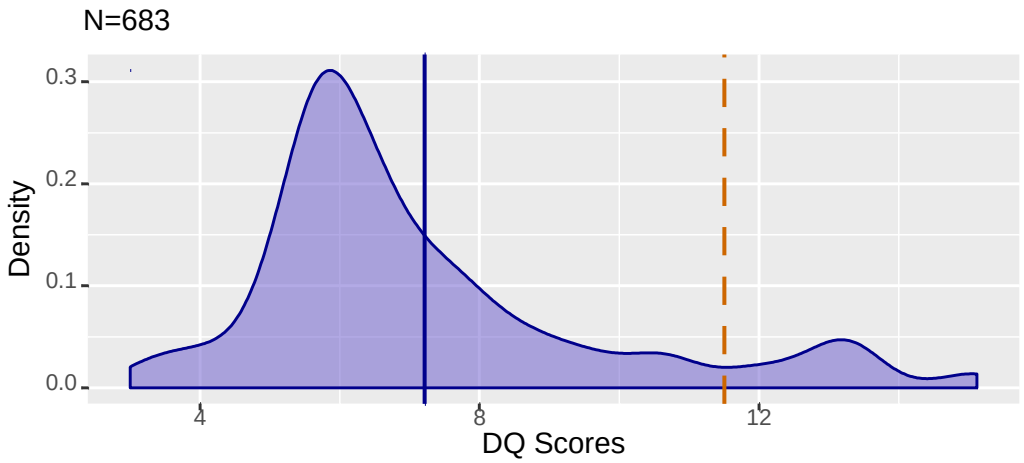

Figure 2.2 Density plot of TrioDeNovo's DQ Scores. Distribution of the Denovo Quality scores from TrioDeNovo for 683 candidate DNMs. The solid blue line is at the mean value of 7.21 and the dashed orange line is at the selected cutoff of 11.5.

A threshold of 0.75 was selected as the high certainty DeNovoGear values

DNG calculated a posterior probability for the candidates DNMs that were provided as input. Of the 879 input variants it generated a score for 482 variants (55% of total) who met the software's minimum quality thresholds and did not generate a nonsense result. Since these are posterior probability (PP) values they will range from 0 to 1 and I refer to them as the DNG scores. The resulting DNG scores had a mean of 0.414 and when these scores are plotted it shows a bimodal distribution (**Figure 2.3**). The first group around the score value of 0.046 are the predicted false positives and the second group around the score value of 0.98 are the candidate DNMs predicted to be true positives. Instead of using a the mean value as a cutoff we decided to use the score value of 0.75 as the threshold, dashed orange line in the figure. This value in the distribution density plots is approximately where the density starts to increase for the second peak. Of the 879 candidate DNMs, 85 (9.7%) variants have a score above the threshold and are part of the high quality DeNovoGear set (hqDNG).

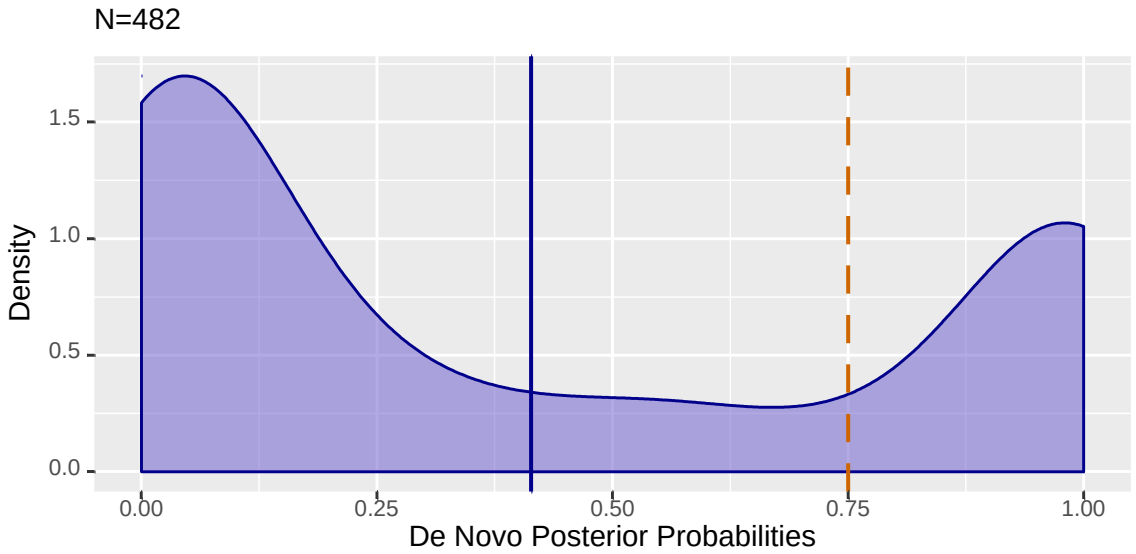

Figure 2.3 Density plot of DNG's Posterior Probabilities. Distribution of the *de novo* posterior probabilities that DNG calculated for 482 candidate DNMs. The solid blue line is at the mean value of 0.414 and the dashed orange line is at the selected cutoff of 0.75.

Depth of sequencing is significantly different between variants analyzed by the algorithms and those that were not

We compared the overlap of variants that TDN and DNG were able to process as a heuristic way to test how flexible these algorithms are with their input. Unlike the data they were created and tested with, our very large joint calling method involves multiple computations and recalculations of parameters that they will use to score each variant. Variables related to the format of the input can easily be changed or manipulated through programming, but the underlying distribution of quality scores, which reads are trusted or discarded, and what genotype to decide in ambiguous cases are all calculations that cannot be easily reverted once processed and ready to analyze. We adjusted TDN's minimum score to report from 5.00 to 0.00 and the minimum depth per member of trio from 5 to 8. For DNG we reduced the minimum depth per individual from 10 to 8. In **Figure 2.4** we see that from the 879 candidate DNM variants, DNG and TDN respectively calculated scores for 482 (53%) and 683 (79%) of the total. The overlap between them was 479 (54%) variants intersected and all DNG having no unique variants processed makes the union 686 (79%) all together.

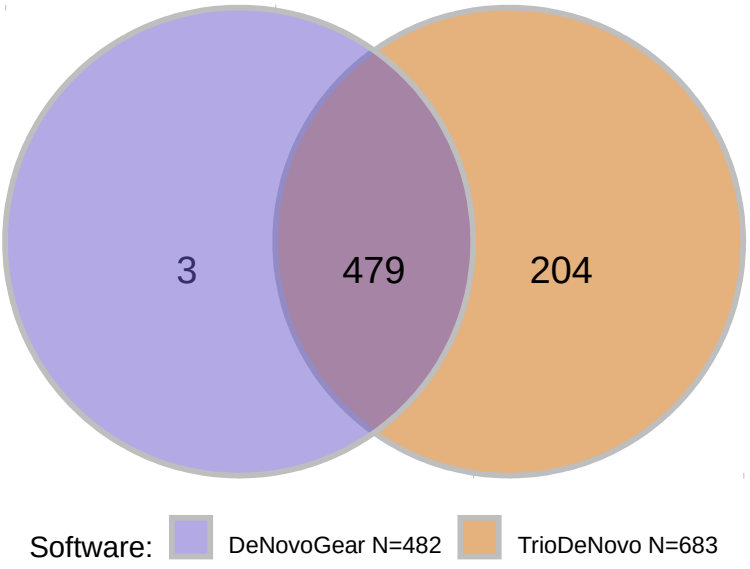

Figure 2.4 Venn Diagram of candidate DNMs processed by DNG and TDN. Number of candidate DNMs that each algorithm was able to process.

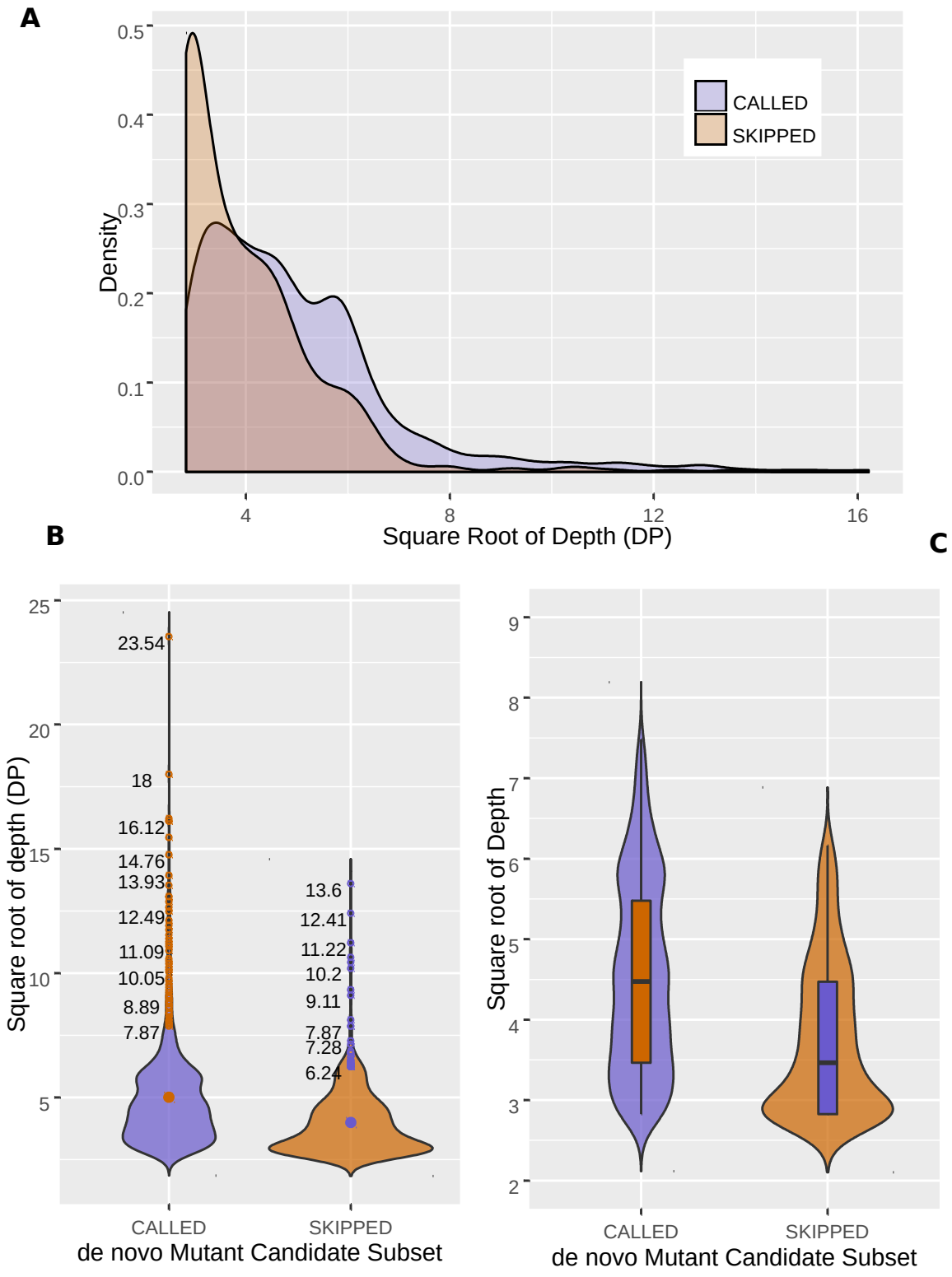

Figure 2.5 Violin plots of the distribution of depth between called and discarded variants. (A) Show the density shift from the group of called variants and those skipped for quality of complexity issues. **(B)** Plot of unfiltered read data with outliers marked by small circles and the mean of distribution by a filled on. **(C)** Both the distribution and quartiles of the read depth without outliers.

The density plot of variant depth in **Figure 2.5 A** is both the called and skipped sets, with 193 (21% of total) skipped and the 689 processed by the algorithms. It show that the majority of the density for both lies under 50 reads and it has noticeable skew to the left indicating possible outliers. The processed variants have a mean depth of 29.3 with a standard deviation (sd) of 31.8 and the skipped variants a mean of 17.8 and an sd of 16.1. The evidence of outliers is clear when comparing the top 11 variants whose depths range from 207 to 554 reads and make up less than 5% of the total of number of variants. The violin plots in **Figure 2.5B** better illustrate the effect the outliers have on the distribution and the need to remove them so that any statistical assessment comparing the two groups is accurate to the real population. This sort of distribution and outliers are not unexpected since read depth data has been previously described as a mixture model of a both a Poisson and a Negative Binomial distribution with over-dispersion. This makes the use of standard statistical outlier exclusion methods inappropriate for this data set since they expect a normal distribution and if used it would likely remove real variants.

For the reasons discussed, I removed values higher than the mean read depth plus it's square root multiplied by 5 which was a method proposed in a 2015 review of quality control and artifact finding in trio exomes.[60] With a mean depth of 29.3 and 17.8 in the called and skipped groups respectively, the calculated cutoffs are of 56.37 and 43.11. After removal this leaves us with 91.7% of called and 94.4% of skipped variants. Both with and without the removed data points the difference in depth between the groups is significant with a *p*-value below 2.2×10^{-6} when calculated using Mann-Whitney U statistical test. This test was selected over others for it's robustness with data that is not normally distributed. This statistically significant difference between the depth of called and skipped variants is expected. Nevertheless there is concern was concern that the DNM software isn't correctly compensating for it.

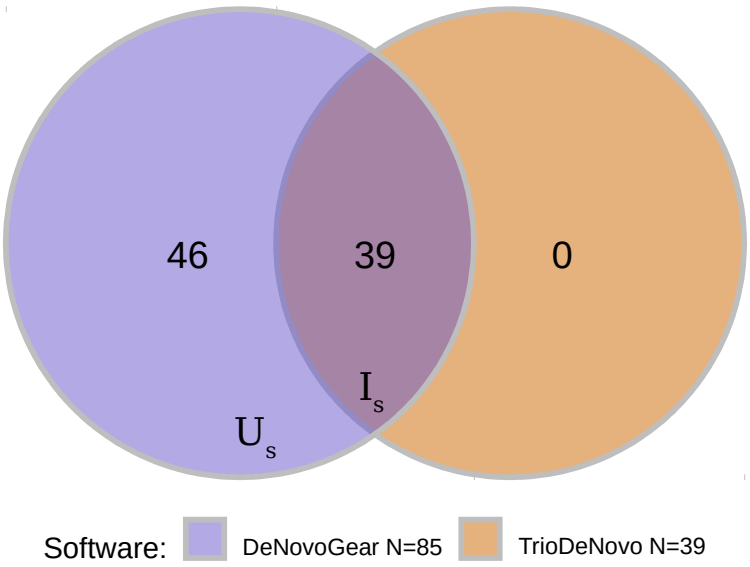

Figure 2.6 Venn Diagram of candidate DNMs processed by DNG and TDN. Number of candidate DNMs that passed the quality threshold for high confidence candidates. The letters Is represent the Intersect set of variants that are shared between TDN and DNG. The Us represents the unique set from DNG.

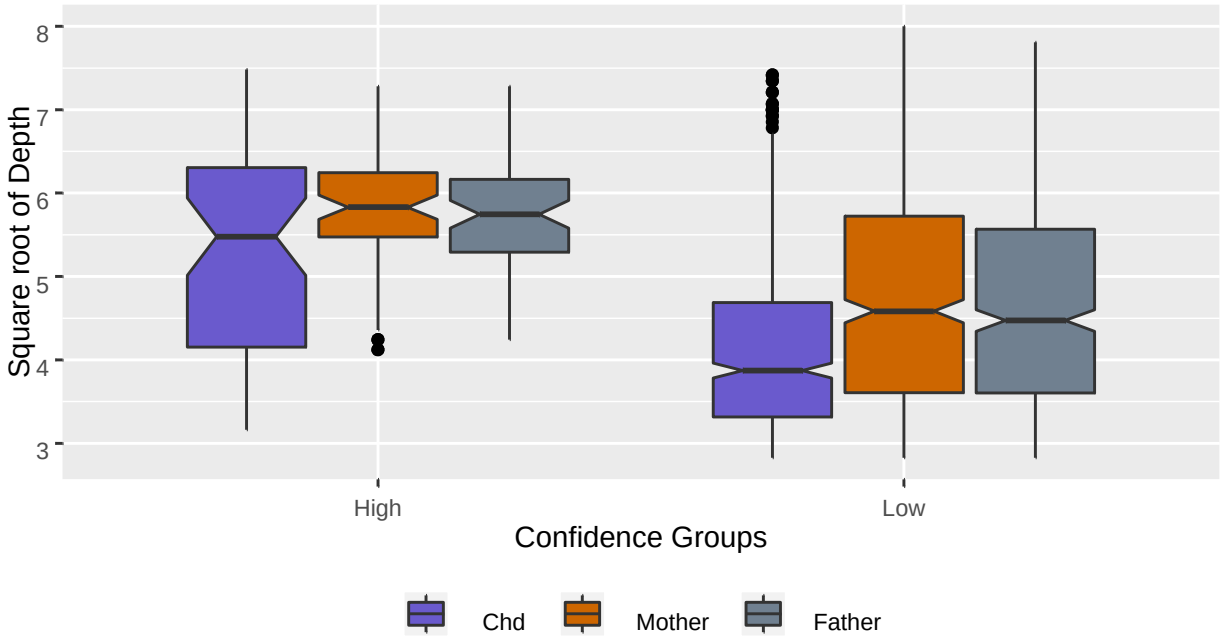

Figure 2.7 Boxplot comparing the distribution of depth between confidence sets The group of the left is the high confidence variants and on low confidence on the right. Each group is divided into each member of the trio.The variant depth values were transformed by using their root square values and removing outliers as described for Figure 2.5

The DNG algorithm generated a larger group of high confidence variants

The real value of these programs will depend on their ability to reduce groups of hundreds of candidates to a group of small size with high confidence. After applying the minimum quality cutoffs of 0.75 and 11.5 to the group of candidate DNMs, the number of total candidate variants decreased from 482 to 85 and from 683 to 39 for TDN and DNG respectively. These smaller groups were considered as candidate DNMs of high and low confidence. **Figure 2.6** shows the overlap for the variants that TDN and DNG processed and considered as high confidence variants. Because the high quality variants from the TDN algorithm were a subset of those from DNG, I chose to use the DNG scores as the primary predictive score. With this more manageable size of high confidence variants, I tested for multiple parameters for significant differences using the Pairwise Wilcoxon Rank Sum Test with the Benjamini-Hochberg Procedures to correct for multiple comparisons. Of the metrics the most significant differences was comparing between groups sorted by the role of the members of the trios (child, mother and father groups) and those variants that pass the DNG score threshold for confidence (high and low groups) for a total of six groups. The depth values were transformed by applying the square root to each depth value which helped to better illustrate the overdispersed distribution for the six groups.

Figure 2.7 is the resulting boxplot of the square root of depth in the variant positions between the six groups. The high and low confidence groups have significantly different sequencing depth value with a p-value $< 2x10^{-16}$. When the groups are analyzed without dividing for high or low confidence and looking between the familial roles, the parents don't have a significant difference in depth (p-value = 0.35) but there is a significant difference between the depth of the children in the cohort and the parental members. Those comparison have a p-value $< 2x10^{-16}$ and p-value = $4.9x10^{-16}$ for the child-mother and child-father pairs. The significant

differences in sequencing depth between high and low certainty groups, along with the significant difference between variants processed and skipped, indicate just how important sequencing depth is for the calculation of scores for both algorithms. It also indicates a bias where depth artificially increases the confidence scores calculated by the DNM predicting algorithms. A potential fix would be to scale or normalize scores based on total depth of the trio to increase true DNMs prediction. The best way to successfully normalize for depth would be by comparing and modeling the depth of groups of variants that are pass or fail sequencing validation. Unfortunately this would require a lot of sequence validation and is outside of the scope of this project.

Depth and Quality Scores were not significantly different between the members of each trio; but they did significantly differ between the candidate DNMs in the intersect of both algorithms and those from the union set which can be seen represented in **Figure 2.6**. This could mean TDN is less likely to assign high scores based on depth. Although as a counter observation to this, the threshold chosen to differentiate between the two sets is an *ad hoc* value decided from visualizing their distributions and selecting for a high scoring subset. It might just be there is a better method to decide a cutoff for TDN. Ultimately it may result that most of the high confidence variants unique to DNG are artifacts of one type or another, but from Sanger validation, we know that TDN missed true stop-gain variant in the CRIM1 gene which is in the set of variants unique to TDN.

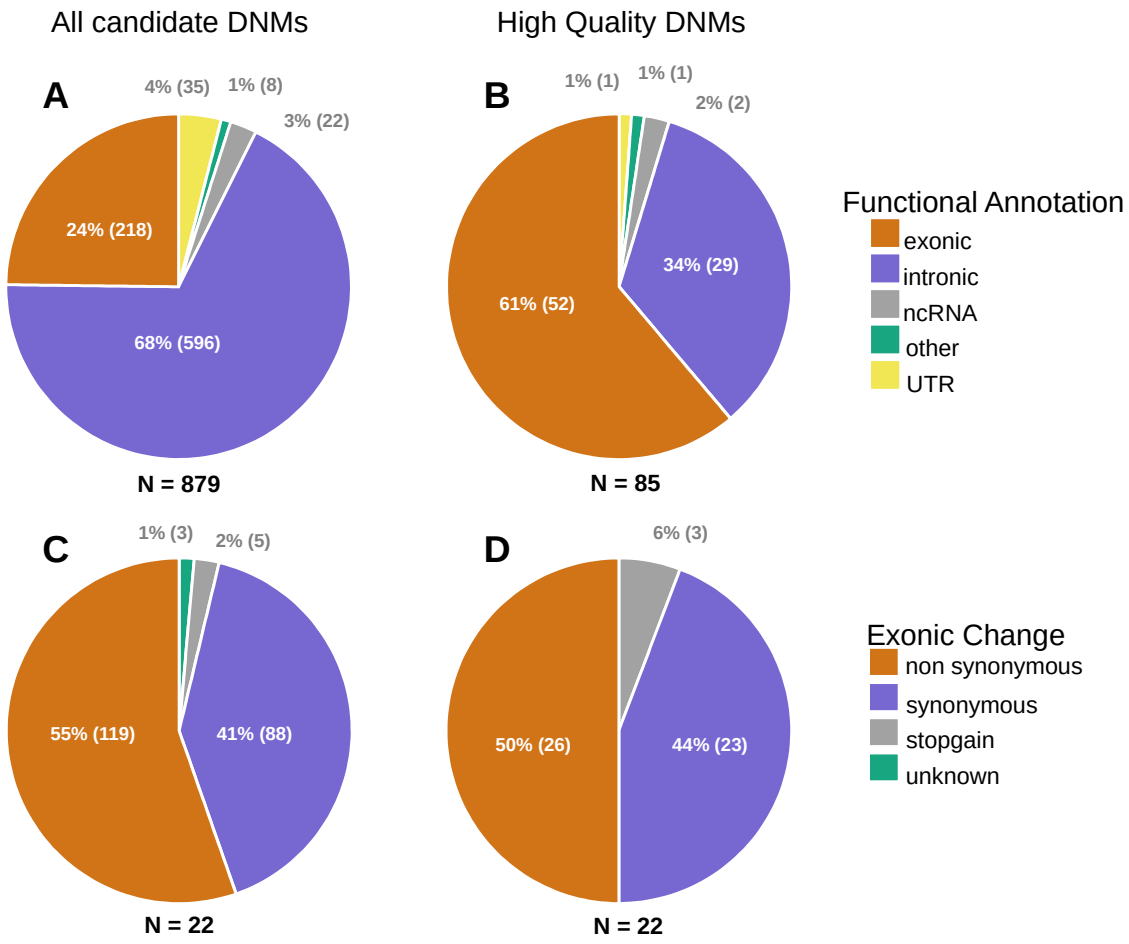

Figure 2.8 Distribution of Functional and Exonic Annotations for all candidate DNMs and the high confidence subset. (A) functional annotation for all candidate DNMs, **(B)** functional annotation of the high confidence subset shown in A **(C)** Exonic functional change type from the exons in pie A **(D)** Exonic functional change type from the exons in pie B

The functional annotation for the complete and high confidence DNM candidate sets were consistent with expectations

Looking at the functional annotation of each variant is an effective way to decide likelihood of pathogenicity and add context to the potential molecular changes . **(Figure 2.8A)** For the overall population of variants the majority are annotated as intronic sequences, which is a probably caused by lower quality score at the boundaries of exons selected for during library preparation. This is further supported when the certainty cutoff is applied and a large number of the candidate DNMs that fall on intronic regions are eliminated.**(Figure 2.8B)** It is to be expected that when selecting for variants that are inconsistent with parental genotype you will have a population biased to contain real *de novo* events and artifacts caused by low sequence quality at the extreme of sequencing reads. From the exonic candidate DNMs the distribution between synonymous and non synonymous variants doesn't change much between the groups, **(Figure 2.8C & D)** likely because the population of variants in exonic regions fall in more reliable regions in between the probes during exon enrichment of the sequencing library.

The high quality variants feature three validated stop-gain variants

After filtering for those variants with a high confidence score, the other annotation fields were used to remove variants that might be real DNMs but would be unlike to be causative. The first criteria was find which of these variants were previously detected by the ExAC and Gnomad projects. If they were found, anything with a frequency in any of the populations higher than 0.001 was removed, which left us with 44 variants with high confidence score and either rare or not found by population sequencing studies. As a way to compare how effective this was to a manual approach we compared the number of variants we would have selected if we had selected for variants from every GATK called SNV with two criteria: nonsynonymous annotation which

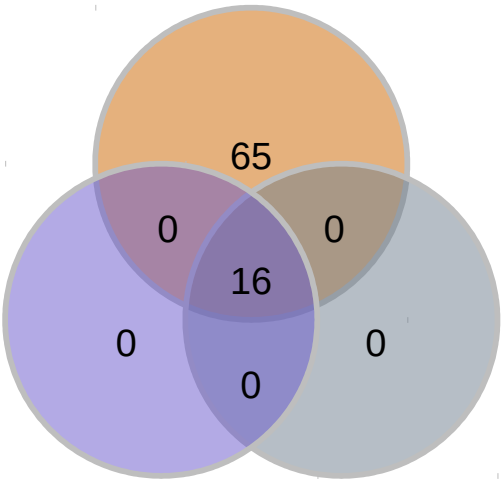

Figure 2.9 Venn diagram of non-synonymous mutation overlap between selecting methods TDN stand for TrioDeNov, DNG is DeNovoGear and manual is the set of nonsynonymous variants that don't appear in databases.

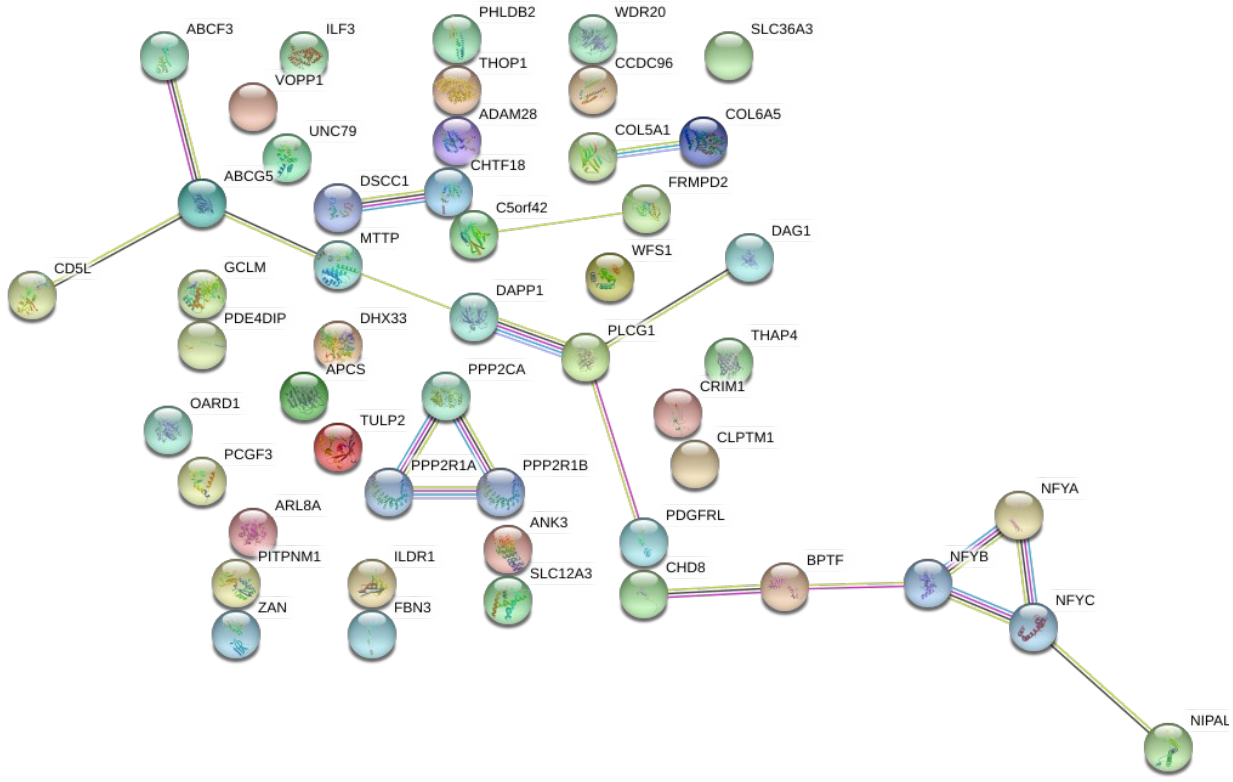

Figure 2.10 Graph from string analysis. String analysis result from the 44 variants.
Known interactions: cyan lines for curated databases and for experimentally determined interactions
Other methods: yellow lines are connected from textmining publications, the black lines are co-expression, and the pale blue line is protein homology. Some genes are not in the list of 44 variants but were added by STRING to display connections with intermediaries

Chrom[a]	Coord[b]	Trio[c]	Change[d]	Functional[e] Change	Exonic[f] Change	Gene[g]	TDN:[h] DQ	DNG:[I] Post Prob
5	37221474	1272	T/C	exonic	nonsynonymous SNV	C5orf42	10.62	0.998483
10	49457193	1266	C/A	exonic	nonsynonymous SNV	FRMPD2	12.55	0.999995
3	111671543	1261	G/A	exonic	nonsynonymous SNV	PHLDB2	13.07	0.999995
19	45495655	1238	G/A	exonic	nonsynonymous SNV	CLPTM1	13.11	0.999995
19	10799237	1013	A/G	exonic	nonsynonymous SNV	ILF3	13.32	0.999997
1	94360245	1113	A/G	exonic	nonsynonymous SNV	GCLM	10.62	0.998473
20	39793951	1238	G/A	exonic	nonsynonymous SNV	PLCG1	13.58	0.999998
1	157803081	1076	G/A	exonic	nonsynonymous SNV	CD5L	13.6	0.999998
10	61833498	1095	C/T	exonic	nonsynonymous SNV	ANK3	12.1	0.999949
8	24199262	1095	G/A	exonic	nonsynonymous SNV	ADAM28	13.75	0.999999
1	159557979	1210	C/A	exonic	nonsynonymous SNV	APCS	8.82	0.976516
14	102675564	1332	T/A	exonic	nonsynonymous SNV	WDR20	12.72	0.999997
11	67265662	1280	G/A	exonic	nonsynonymous SNV	PITPNM1	13.32	0.999997
1	24775998	1076	G/A	exonic	nonsynonymous SNV	NIPAL3	13.23	0.999996
3	130187839	1338	A/T	exonic	nonsynonymous SNV	COL6A5	12.55	0.999995
4	7044040	1076	T/C	exonic	nonsynonymous SNV	CCDC96	15.12	1
14	21868728	1136	G/A	exonic	stopgain	CHD8	13.23	0.999996
2	36771526	1095	T/A	exonic	stopgain	CRIM1	11.77	0.999973
1	202107112	1338	G/A	exonic	stopgain	ARL8A	13.02	0.999994
16	845149	1136	C/T	exonic	synonymous SNV	CHTF18	13.45	0.999998
2	44052119	1180	C/T	exonic	synonymous SNV	ABCG5	13.15	0.999995
3	49548223	1040	T/C	exonic	synonymous SNV	DAG1	15.12	1
14	93963550	1338	A/G	exonic	synonymous SNV	UNC79	13.15	0.999995
17	65955689	1203	A/G	exonic	synonymous SNV	BPTF	13.32	0.999997
4	100571409	1180	T/C	intergenic	na	MTTP;DAPP1	13.42	0.999998
6	41062258	1112	A/T	intronic	na	NFYA;OARD1	10.02	0.998473
3	183907605	1143	G/A	intronic	na	ABCF3	13.32	0.999997
3	121707276	1136	C/A	intronic	na	ILDR1	9.12	0.987901
5	150660805	1112	C/A	intronic	na	SLC36A3	8.65	0.965148
9	137690226	1199	A/T	intronic	na	COL5A1	12.88	0.999998
7	100367450	1098	A/G	intronic	na	ZAN	9.55	0.981967
4	738497	1309	G/A	intronic	na	PCGF3	9.1	0.950693
17	5359553	1203	G/A	intronic	na	DHX33	8.94	0.930295
7	55565253	1280	G/C	intronic	na	VOPP1	8.04	0.871307
19	49391545	1272	G/A	intronic	na	TULP2	11.22	0.999618
16	56913993	1280	G/C	intronic	na	SLC12A3	12.42	0.999994
5	133536263	1266	C/G	intronic	na	PPP2CA	9.12	0.987997
4	6293802	1180	T/C	intronic	na	WFS1	9.52	0.98125
19	8190680	1136	A/G	intronic	na	FBN3	9.7	0.987524
8	17446968	1280	G/A	intronic	na	PDGFRL	8.52	0.839481
19	2807857	1076	C/T	intronic	na	THOP1	15.12	1
19	9732290	1098	C/T	ncRna_exonic	na	ZNF561-AS1	8.42	0.798211
1	144951959	1150	T/C	ncRna_intronic	na	LOC100996724	8.83	0.912628
2	242576691	1040	C/G	UTR5	na	THAP4	13.5	1

Table 2.2 Candidate de novo Mutants

(a)chrom and (b)pos are the chromosome and position for each variants
(c)trio is the ID for the family
(d)change is the mutation with the reference and then the variant
(e) functional and (f)exonic change are the variant type annotation
(g)gene is the gene name
(h)TDN:DQ is the *de novo* quality score assigned by TrioDeNovo
(I)DNG is the posterior probability assigned by DeNovoGear

are the likeliest pathogenic, no presence in public databases and allele not present in the parents' genotypes. **Figure 2.9** shows that we would not have obtained variants unique from those using DNM predictive algorithms, but we did obtain a much smaller set with 16 nonsynonymous in common with the manual method's group of 81. We also took into scores computed to predict pathogenicity by CADD, Polyphen, Fathmm, and Sift. Out of 16 non synonymous mutations that Polyphen evaluated, it assigned 7 as possibly or probably damaging. Sift scored 9 out the 16 nonsynonnymous and the only integenic variant as damaging. Fathmm assigned scores as damaging for 10 of 16 nonsynonymous, 3 of 3 stopgain and the 1 of 1 intergenic variants. Of all the variants, 3 of the 3 stopgain variants have been sequenced and confirmed so far.

We searched for interactions in the STRING database **(Figure 2.10)** and 3 networks with more than 3 nodes were found . The first and largest including genes *ABCF3, CD5L, ABCG5*, *MTTP*, *DAPP1*, *PLCG1*, *DAG1* and *PDGFRL* with an assigned PPI enrichment p-value of 3.07×10^{-9}. The 6 network gene with CHD8, BPTF, NIPAL2, NFYC, NFYA and NFYB has a PPI enrichment *p*-value of 1.31×10^{-7}. These score measure the significance of finding these interactions in comparison to the size and number of proteins chosen at random. With a false discovery rate of 0.044 the CHD8 and BPTF in the second network have the brain development biological process descriptor.

Chrom[a]	Coord[b]	Trio[c]	Change[d]	CADD[e]	ClinVar[f]	Polyphen2[g]	Polyphen[h] Pred	Fathmm[i]	Fathmm[j] Pred	Sift[k]	Sift[l] Pred
5	37221474	1272	T/C	12.73	0	0.067	B	0.716	D	0.205	T
10	49457193	1266	C/A	12.68	0	0.091	B	0.819	D	0.03	D
3	111671543	1261	G/A	7.397	0	0.899	P	0.975	D	0.034	D
19	45495655	1238	G/A	6.124	0	1	D	0.996	D	0	D
19	10799237	1013	A/G	4.484	0	0	B	0.606	D	0.844	T
1	94360245	1113	A/G	3.837	0	1	D	0.988	D	0	D
20	39793951	1238	G/A	2.74	0	0.755	P	0.996	D	0.054	T
1	157803081	1076	G/A	1.994	0	0.809	P	0.118	T	0.022	D
10	61833498	1095	C/T	1.77	0	0.026	B	0.961	D	0.004	D
8	24199262	1095	G/A	1.63	0	0	B	0.006	T	0.197	T
1	159557979	1210	C/A	0.418	0	0.109	B	0.076	T	0.078	T
14	102675564	1332	T/A	0.098	0	0.827	P	0.97	D	0.073	T
11	67265662	1280	G/A	0.051	0	0.535	P	0.894	D	0.008	D
1	24775998	1076	G/A	0.031	0	0.935	D	0.971	D	0.01	D
3	130187839	1338	A/T	0.019	0	0.029	B	0.022	T	0.761	T
4	7044040	1076	T/C	0.002	0	0.007	B	0.053	T	0.005	D
14	21868728	1136	G/A	12.91	1	na	na	0.967	D	na	na
2	36771526	1095	T/A	11.26	0	na	na	0.922	D	na	na
1	202107112	1338	G/A	4.158	0	na	na	0.978	D	na	na
16	845149	1136	C/T	14.25	0	na	na	na	na	na	na
2	44052119	1180	C/T	6.159	0	na	na	na	na	na	na
3	49548223	1040	T/C	6.06	0	na	na	na	na	na	na
14	93963550	1338	A/G	2.548	0	na	na	na	na	na	na
17	65955689	1203	A/G	1.166	0	na	na	na	na	na	na
4	100571409	1180	T/C	2.664	0	na	na	0.98	D	0.003	D
6	41062258	1112	A/T	24.4	0	na	na	na	na	na	na
3	183907605	1143	G/A	23.9	0	na	na	na	na	na	na
3	121707276	1136	C/A	12.06	0	na	na	na	na	na	na
5	150660805	1112	C/A	7.797	0	na	na	na	na	na	na
9	137690226	1199	A/T	6.936	0	na	na	na	na	na	na
7	100367450	1098	A/G	5.708	0	na	na	na	na	na	na
4	738497	1309	G/A	4.468	0	na	na	na	na	na	na
17	5359553	1203	G/A	4.385	0	na	na	na	na	na	na
7	55565253	1280	G/C	4.073	0	na	na	na	na	na	na
19	49391545	1272	G/A	3.137	0	na	na	na	na	na	na
16	56913993	1280	G/C	2.721	0	na	na	na	na	na	na
5	133536263	1266	C/G	2.461	0	na	na	na	na	na	na
4	6293802	1180	T/C	1.344	0	na	na	na	na	na	na
19	8190680	1136	A/G	1.199	0	na	na	na	na	na	na
8	17446968	1280	G/A	0.81	0	na	na	na	na	na	na
19	2807857	1076	C/T	0.591	0	na	na	na	na	na	na
19	9732290	1098	C/T	8.135	0	na	na	na	na	na	na
1	144951959	1150	T/C	8.01	0	na	na	na	na	na	na
2	242576691	1040	C/G	7.624	0	na	na	na	na	na	na

Table 2.3 Pathogenic Annotation of the high confidence candidate DNMs

(a)chrom and (b)pos are the chromosome and position for each variants

(c)trio is the ID for the family

(d)change is the mutation with the reference and then the variant

(e) Is the assigned phred adjusted CADD score, values higher than 5 can be pathogenic for non synonymous and10 for others

(f) Presence of absence in the clinvar database, 1 for yes, 0 for no

(g) Polyphen score and (h) the prediction of its effect. B for benign, P for probably pathogenic and D for likely damaging

(i) Fathmm score and (j) the prediction of its effect. D for damaging and T for tolerated

(k) Sift score and (l) the prediction of its effect. D for damaging and T for tolerated

2.5 Discussion

Gene Networks can play a crucial role in understanding the genetics of lesser studied disorders

As the scientific community keeps increasing the accuracy of their predictive tools by creating more advanced methods to analyze and integrate previously published data, research groups interested in other disorders and questions get to use those tools and knowledge to their benefit. CM1 research features a modest number of previous experiments focusing on understanding the genetic contribution to its cause, but this makes it a prime candidate to reap the benefits from the research of other neurodevelopmental disorders. This thesis aimed to expand our understanding of CM1 analyzing the trios exome of severe or early onset spontaneous cases of CM1. The golden standard for success for this experimental design is finding highly deleterious mutations in the same gene that with further functional analysis are proved to have a strong enough effect to be at the least partially causative. Although that is undeniably a good scenario, with a group of 29 trios this is highly unlikely to happen. We did achieve to narrow down from over 800 potential variants to less than 60 variants with positive annotations and quality metrics of being potentially related to CM1 etymology. In comparison with current large studies that have thousands of exomes in their analysis, our project has just under 30 and still detected 3 stopgain mutations that were validated. These variants have to be further analyzed with experiments beyond the scope of this thesis, but the annotation and data from predictive algorithms from methods like the network analysis in HumanBase and STRING helped to narrow down which of these variants would be of a higher likelihood to be important for understanding CM1.

DeNovoGear gets more high-quality results, TrioDeNovo can processes more variants and both need to be supplemented with other statistics to increase certainty

Both algorithms performed comparably even though their mathematical frameworks used different formulas and assumptions to generated a predictive score. TrioDeNovo required the least amount of preparation, less user provided inference of mutation rates, and no need to adjust parameters. It also didn't provide an user friendly method to change or adjust those internal parameters and assumptions. Nevertheless, it narrowed down to 39 variants assigned with high confidence of being real DNMs. DeNovoGear required more attention and input from the user, but with suitable enough default parameters and fast running speeds, it offered greater control. Overall this resulted in TDN making it nearly impossible to adjust to compensate when false positives were missed or to understand what was the metric that caused the true positive to be classified as a false positive. In contrast, the ability to adjust all the filtering steps and initial statistical values for the DNG software allowed for excellent calibration which could take into account previously confirmed DNM. TDN and DNG demonstrated a significant bias towards high depth of sequencing. This bias is not entirely unexpected since both incorporate Phred likelihood scores obtained from variant calling software, like GATK and Bedtools, to perform their calculations and therefore rely on these calling algorithms to normalize for read depth within and between trios. This feature opens the possibility of newer methods of mutation calling biasing and inflating signals that influence the underlying mathematical assumptions.

Their primary benefit over blindly looking for variants that are missing in the parental variant sites is using the known parental relationship between the three sets of variants in a trio. A possible reason for the inflation or bias in prediction of these two software could be that they were designed and tested using two types of data. The first type of data used is *in silico*

generated variants from modeling an artificial trio, this gives complete certainty of know true positive DNMs and control over different statistics of the exomes and variants such as sequencing depth and quality. The second sets come from *in vivo* samples from large projects. These sets include population studies such as the 1000 genome projects or datasets of specific phenotypes like the Simon Simplex Collection for ASD. In contrast to the reality of smaller research project, an average dataset of human sequencing is likely to lack uniformity in depth and is often doesn't feature sequencing depth much higher than 30x. The sequencing of patients in a hospital setting will happen through a larger span of time and often features sequencing using different versions on protocols that have been updated as time passes. All this variability within a cohort's sample preparation, sequencing platform used, individual idiosyncrasies between the people creating the libraries and many other variables become a sources of error that is often not taken into account during the benchmark tests in the published with these sets.

These considerations make it crucial as the end-user and researcher using these methods to be aware of the underlying biases and ready to detect them. It is a given fact that increasing sequencing depth can benefit the finding and identification of novel mutations and that large numbers of trios in a cohort will increases the power of the study. In contrast, these same features create a need to for the filtering and normalization of the data. Since normalization isn't always possible or viable, we use the other quality statistic generated by the variant calling step of the experiments and took it a step further by combining and filtering based on their relationships.

We found three genes with stopgain mutations the rationale for their role on CM1 etyology

With a cohort of 29 trios, we did not find recurrent genes but we have already validated 3 of the 3 stop-gain variants (CHD8, CRIM1 and ARL8A) Further validation will be done for the other high CADD scores in the list of 44 variants. CRIM1 has previously been associated with

ASD, which hints that we are heading in the right direction. Ultimately diseases with high heterogeneity are expected to be difficult to associate to a small group of genes, but the presence of 3 stopgain variants in different trios on a small trio cohort indicates the strength of selecting good candidate trios based on the medical history and severity of the phenotype being studied. Adding three stop gain variants which are validated in severe cases is already a step forward in the understanding of CM1's biology. The network analysis of these variants included the CHD8 and CHTF18 interacting in the result of the STRING analysis, both being members of the SH2 domain of protein families. All these connections and genes are themselves nodes in larger networks that we still don't have a clear picture of and by adding genes, keeping true positives, and removing false positive we increase the power and have a better picture of which genetic interactions are disrupted in CM1. There are also many other experiments happening with datasets of CM1 patients of which very few have the parental genetic information and some are not spontaneous like the trio cohort selected for this project. This group of genes can now become part of the analysis being done with those exomes and increase the odds of elucidating pathways and gene networks to focus on.

Conclusion

Annotating and identifying de *novo* mutations is quickly becoming a popular a reliable source for the discovery of genetic risk of neurodevelopmental disorders. Most of these studies benefit from previous studies that have found genes of interest using other methods. This project adds to the body of work that previous CM1 research have done to understand its etiology. From 29 trios with thousands of genetic variants, we found and confirmed three stopgain mutations. We also identified 22 variants in genes that had not been previously identified associated or linked with CM1. From the multiple research projects finding genes related to the origin of neurodevelopmental disorders, Autism spectrum disorder research has found the most success in numbers. They have benefited from finding co-ocurring genes of interests from multiple experimental sources, and it is my hope that the list of genes obtained from this thesis will aid future research projects by increasing the certainty of genes of interests that may play a role in CM1 etyology.

From a technical perspective, using GATK's joint calling method with TrioDeNovo and DeNovoGear led to finding some critical biases that can be corrected for in future projects. Depth of sequencing significantly influenced the predictive scores calculated for DNMs, along with the difference in the sequencing depth of the members of the trio. Taking these metrics into consideration will be crucial for the removal of outliers in projects using algorithms designed for single trios on joint called trios. Additionally, a method to scale the confidence scores by depth can increase the success in future experiments when discerning between false and true positives. The three stop gain mutations in the list of potential true positive DNMs that was created through this analysis have already been confirmed, more of these candidates are to be validated in future

experiments. Having a validated set of true and false positives will be beneficial for the detection of other features and metrics of the candidate DNMs that successfully distinguish between both groups. Additionally, the true DNMs will increase the confidence of the variants and genes of interest found in other exomes of CM1 patients from the Gurnett's lab internal database.

References

1. Burina, A. & Ibrahimagić, O. Ć. ARNOLD – CHIARI MALFORMATION AND SYRINGOMYELIA. **38**, 44–46 (2009).
2. Godzik, J. *et al.* Comparison of spinal deformity in children with Chiari I malformation with and without syringomyelia: matched cohort study. *Eur. Spine J.* **25**, 619–626 (2016).
3. Pindrik, J. & Johnston, J. M. Clinical Presentation of Chiari I Malformation and Syringomyelia in Children. *Neurosurg. Clin. N. Am.* **26**, 509–514 (2015).
4. Strahle, J. *et al.* The association between Chiari malformation Type I, spinal syrinx, and scoliosis. *J. Neurosurg. Pediatr.* **15**, 607–611 (2015).
5. Strahle, J. *et al.* Syrinx location and size according to etiology: identification of Chiari-associated syrinx. *J. Neurosurg. Pediatr.* **16**, 21–29 (2015).
6. Meadows, J., Kraut, M., Guarnieri, M., Haroun, R. I. & Carson, B. S. Asymptomatic Chiari Type I malformations identified on magnetic resonance imaging. *J. Neurosurg.* **92**, 920–926 (2000).
7. Langridge, B., Phillips, E. & Choi, D. Chiari Malformation Type 1: A Systematic Review of Natural History and Conservative Management. *World Neurosurg.* **104**, 2–5 (2017).
8. Aitken, L. A. *et al.* Chiari Type I Malformation in a Pediatric Population. *Pediatr. Neurol.* **40**, 449–454 (2009).
9. Smith, B. W. *et al.* Distribution of cerebellar tonsil position: Implications for understanding Chiari malformation. *J. Neurosurg.* **119**, 812–819 (2013).
10. Kahn, E. N., Muraszko, K. M. & Maher, C. O. Prevalence of Chiari I Malformation and Syringomyelia. *Neurosurg. Clin. N. Am.* **26**, 501–507 (2015).
11. Vernooij, M. W. *et al.* Incidental findings on brain MRI in the general population. *N. Engl. J. Med.* **357**, 1821–1828 (2007).
12. Greenberg, J. K. *et al.* Chiari malformation Type I surgery in pediatric patients. Part 2: complications and the influence of comorbid disease in California, Florida, and New York. *J. Neurosurg. Pediatr.* **17**, 525–532 (2016).
13. Strahle, J. *et al.* Chiari malformation Type I and syrinx in children undergoing magnetic resonance imaging: Clinical article. *J. Neurosurg. Pediatr.* **8**, 205–213 (2011).
14. Tubbs, R. S. *et al.* Institutional experience with 500 cases of surgically treated pediatric Chiari malformation Type I. *J. Neurosurg. Pediatr.* **7**, 248–256 (2011).
15. Milhorat, T. H., Bolognese, P. A., Misao Nishikawa, M., McDonnell, N. B. & Francomano, C. A. Syndrome of occipitoatlantoaxial hypermobility, cranial settling, and Chiari malformation Type I in patients with hereditary disorders of connective tissue. *J. Neurosurg. Spine* **7**, 601–609 (2007).

16. Greenlee, J. D. W., Donovan, K. A., Hasan, D. M. & Menezes, A. H. Chiari I Malformation in the Very Young Child: The Spectrum of Presentations and Experience in 31 Children Under Age 6 Years. *Pediatrics* **110**, 1212–1219 (2002).
17. Loukas, M. *et al.* Associated disorders of Chiari Type I malformations: A review. *Neurosurg. Focus* **31**, 1–6 (2011).
18. Abbott, D., Brockmeyer, D., Neklason, D. W., Terrlink, C. & Cannon-Albright, L. A. Population-based description of familial clustering of Chiari malformation Type I. *J. Neurogsurgery* **128**, 460–465 (2018).
19. Nagy, L., Mobley, J. & Ray, C. Familial aggregation of chiari malformation: Presentation, pedigree, and review of the literature. *Turk. Neurosurg.* **26**, 315–320 (2016).
20. Speer, M. C. *et al.* Chiari type I Malformation with or without syringomielia: Prevalence an Genetics. *Jorunal Genet. Couns.* **12** , 297–311 (2003).
21. Boyles, A. L. *et al.* Phenotypic definition of Chiari type I malformation coupled with high-density SNP genome screen shows significant evidence for linkage to regions on chromosomes 9 and 15. *Am. J. Med. Genet. Part A* **140A**, 2776–2785 (2006).
22. Markunas, C. A. *et al.* Stratified Whole Genome Linkage Analysis of Chiari Type I Malformation Implicates Known Klippel-Feil Syndrome Genes as Putative Disease Candidates. *PLoS One* **8**, (2013).
23. Markunas, C. A. *et al.* Genetic evaluation and application of posterior cranial fossa traits as endophenotypes for chiari type I malformation. *Ann. Hum. Genet.* **78**, 1–12 (2014).
24. Urbizu, A. *et al.* Chiari Malformation Type I: A Case-Control Association Study of 58 Developmental Genes. *PLoS One* **8**, 1–10 (2013).
25. Markunas, C. A. *et al.* Identification of Chiari Type i Malformation subtypes using whole genome expression profiles and cranial base morphometrics. *BMC Med. Genomics* **7**, 1–15 (2014).
26. Schaaf, C. P. *et al.* Expanding the clinical spectrum of the 16p11.2 chromosomal rearrangements: Three patients with syringomyelia. *Eur. J. Hum. Genet.* **19**, 152–156 (2011).
27. Steinman, K. J. *et al.* 16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort. *Am. J. Med. Genet. Part A* **170**, 2943–2955 (2016).
28. Duran, D. *et al.* Digenic mutations of human OCRL paralogs in Dent's disease type 2 associated with Chiari I malformation. *Hum. Genome Var.* **3**, 16042 (2016).
29. Merello, E. *et al.* Exome sequencing of two Italian pedigrees with non-isolated Chiari malformation type i reveals candidate genes for cranio-facial development. *Eur. J. Hum. Genet.* **25**, 952–959 (2017).
30. An, J. Y. & Claudianos, C. Genetic heterogeneity in autism: From single gene to a pathway perspective. *Neurosci. Biobehav. Rev.* **68**, 442–453 (2016).

31. McGinniss, M. J. & Kaback, M. M. *Heterozygote Testing and Carrier Screening*. *Emery and Rimoin's Principles and Practice of Medical Genetics* (Elsevier, 2013). doi:10.1016/B978-0-12-383834-6.00031-8
32. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
33. De Rubeis, S. & Buxbaum, J. D. Genetics and genomics of autism spectrum disorder: Embracing complexity. *Hum. Mol. Genet.* **24**, R24–R31 (2015).
34. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285–299 (2012).
35. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
36. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
37. An, J. Y. *et al.* Towards a molecular characterization of autism spectrum disorders: an exome sequencing and systems approach. *Transl. Psychiatry* **4**, e394–e394 (2014).
38. Jiang, Y. *et al.* Utilizing population controls in rare-variant case-parent association tests. *Am. J. Hum. Genet.* **94**, 845–853 (2014).
39. Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P. & Eichler, E. E. Recurrent de novo mutations in neurodevelopmental disorders: Properties and clinical implications. *Genome Med.* **9**, 1–16 (2017).
40. Leu, C., Coppola, A. & Sisodiya, S. M. Progress from genome-wide association studies and copy number variant studies in epilepsy. *Curr. Opin. Neurol.* **29**, 158–167 (2016).
41. Zhu, X. *et al.* A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLoS Genet.* **13**, 1–12 (2017).
42. Veeramah, K. R. *et al.* Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* **54**, 1270–1281 (2013).
43. Vorstman, J. A. S. *et al.* Autism genetics: Opportunities and challenges for clinical translation. *Nat. Rev. Genet.* **18**, 362–376 (2017).
44. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
45. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
46. Szklarczyk, D. *et al.* The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
47. Landrum, M. J. *et al.* ClinVar : public archive of relationships among sequence variation and human phenotype. **42**, 980–985 (2014).
48. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).

49. Al-Mubarak, B. *et al.* Whole exome sequencing reveals inherited and de novo variants in autism spectrum disorder: A trio study from Saudi families. *Sci. Rep.* **7**, 1–14 (2017).
50. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
51. Jin, Z. B. *et al.* Identification of de novo germline mutations and causal genes for sporadic diseases using trio-based whole-exome/genome sequencing. *Biol. Rev.* **93**, 1014–1031 (2018).
52. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
53. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).
54. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
55. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
56. Wei, Q. *et al.* A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* **31**, 1375–1381 (2015).
57. Ramu, A. *et al.* DeNovoGear: De novo indel and point mutation discovery and phasing. *Nat. Methods* **10**, 985–987 (2013).
58. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
59. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
60. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).