EXOME SEQUENCING TO IDENTIFY THE GENETIC BASES FOR LYSOSOMAL STORAGE DISEASES OF UNKNOWN ETIOLOGY

by

NAN WANG

A thesis submitted to the
Graduate School-New Brunswick
Rutgers, the State University of New Jersey
And
The Graduate School of Biomedical Sciences
In partial fulfillment of the requirements
For the degree of
Master of Science
Graduate Program in Microbiology and Molecular Genetics
Written under the direction
Dr. Jinchuan Xing
And approved by

____

____

____

New Brunswick, New Jersey
October, 2014

ABSTRACT OF THE THESIS

Exome Sequencing to Identify the Genetic Bases for Lysosomal Storage Disease of Unknown Etiology

By NAN WANG

Thesis Director:

Dr. Jinchuan Xing

Lysosomes are membrane-bound, acidic eukaryotic cellular organelles. As an enzyme container, they play important roles in the degradation of macromolecules. Monogenic mutations resulting in the loss of enzyme activities in the lysosome may lead to severe health problems, such as neurodegeneration and early death. These conditions are categorized as Lysosomal Storage Diseases (LSDs). The diagnosis of LSDs is typically straightforward. However, in some cases the underlying genetic defects remain unknown. Here, we performed whole exome sequencing on 14 suspected LSD cases, with the goal of finding the causal mutations. From the raw sequence data, we first identified DNA variants in each individual using three variant discovery pipelines: the Genome Analysis Toolkit, LifeScope™ Genomic Analysis Software and CLC Genomics Workbench. For each variant calling dataset, we then used the Variant Annotation Analysis Search Tool (VAAST) to prioritize disease-causing mutations in 848 candidate LSD genes. As a probabilistic disease gene finder, VAAST integrates allele frequency, amino acid substitution severity and conservation information into a composite likelihood framework. Different from hard filtering methods, VAAST preserves all the candidates by listing them according to their disease-causing potential. To obtain the detailed information of each mutation and add one more layer of mutational prediction, we performed SIFT analysis for each dataset. Afterward, tier study was conducted to accommodate the discrepancies between different pipelines and further reprioritize the candidate variants. Finally, based on the mutational validation and functional analysis, we identified nine mutations in six genes to be candidate LSDs causal variants in five individuals, including both known and novel mutations. In summary, our project utilized various bioinformatics analyses tools to decode the extensive exome sequencing data and identify candidate variants for downstream functional studies. The study results provide valuable insights into the genetic basis of LSDs.

Acknowledgements

I would like to thank my advisor, Dr. Jinchuan Xing. His guidance and encouragement made my transition from veterinary medicine to human genetics and bioinformatics an exciting and productive one. My curiosity for science has never stopped when I worked with him. Besides his great impact on my academic life, his support and kindness to my personal life influenced me in a way beyond expression. I may never get rid of the miserable memory of what I've suffered from my atypical allergic asthma in America, but I will never forget the tremendous help and spiritual support he offered me during the unique and special period. I give my deepest appreciation to him. It's my honor to work with him as his first graduate student.

I would like to thank Dr. Peter Lobel and Dr. David Sleat for their support and direction of this research. I wouldn't be able to finish the thesis without many valuable discussions with them.

I would like to thank Dr. Kumar Dibyendu, Dr. Robert Donnelly, Erika Gedvilaite and Jui Wan Loh for their devotion and hard work for this project.

I would like to thank the postdoc Dr. Hong-Seok Ha in our lab. I learned so much from many talks and discussions with him. He never hesitates to answer my questions. His passion in science makes me believe in the beauty of being a real scientist.

I would also like to thank all the lab members in our lab and Dr. Kevin Chen's lab. Talking with them always make me happy and enjoy thinking.

I am very thankful to Dr. Tara Matise and Dr. Peter Lobel for being supportive. They are all great scientists I admire. It's my great honor to have them as my committee.

My stay in Rutgers was amazing and productive. The knowledge I've leant and experience I've been through is far more precious to me. Thank you, Rutgers!

Dedication

To my beloved parents for their infinite love, for the sacrifice and devotion they made to grant me all the virtues that make me who I am and what I can be

&

To my beloved grandparents, who live and will live in my heart forever

&

To my dear fiancé for his love and support during my darkest and happiest time

Table of Contents

Additional files:

Supplementary Material I. Lysosome Storage Disease candidate gene list

Supplementary Material II. All candidate variants prioritized after Tier Study and SIFT Analysis

List of Tables

List of Figures

1. Introduction

Lysosome Storage Diseases (LSDs) are inherited metabolic multisystemic disorders caused by specific mutations in genes encoding lysosomal enzymes, resulting in reduced lysosomal trafficking, substrate accumulation and cellular dysfunction [1]. Multiple tissues and organs can be affected, especially the ones with high turnover-rate lysosomal enzymatic substrate. Clinical manifestations include bone deformities, decline in vision and hearing, organomegaly (especially in spleen and liver), cardiac disease and other symptoms. The most severe manifestations involve the central nervous system, leading to mental retardation. Many LSDs share similar clinical presentation and progressively deteriorate with consequence of premature death [2].

LSDs comprise a group of more than 50 known rare monogenic disorders [3]. Although individually LSDs are rare with incidences ranging from 1:57,000 (Gaucher disease) to 1:4,200,000 (Sialidosis) [4], collectively they have an overall prevalence about 1 in 5,000-7,700 live births [5], which is likely underestimated due to a significant number of undiagnosed or misdiagnosed cases. Many of these cases may represent atypical clinical LSDs manifestations with a mild or delayed onset due to partial loss-of-function mutations [6]. Two possible reasons may account for LSDs with unknown etiology: defects in lysosomal proteins that are not currently associated with human diseases or in unidentified lysosomal proteins [3].

Biochemical and genetic analyses are the classic strategies to study the etiology of LSDs [7]. More recently, highly sensitive proteomic approaches have been applied and largely expanded the knowledge of lysosomal proteins [2]. However, these molecular methods are subjected to certain limitations and drawbacks such as low specificity, low throughput, contamination, complexity [3]. Moreover, currently most clinical and molecular analyses focus on specific type of common LSD, such as Gaucher disease [8]. Therefore, these strategies are limited to exploring the genetic bases of unknown etiology LSDs in large scale.

In this study, we adopted whole exome sequencing (WES) and multiple bioinformatics techniques to identify known and novel LSDs mutations from 14 patients with potential LSDs. With the decrease of sequencing cost, WES has been widely used in human genetic studies since the first successful application of WES in identifying the genetic cause of Miller Syndrome [9]. As a cost-effective method for identifying disease-causal mutations, WES covers nearly all protein-coding regions but only requires ~5% of the sequencing throughput of a whole human genome.

With enormous data from WES, bioinformatics tools are required to manipulate and analyze the data efficiently. To eliminate the bias of a certain tool, we adopted three different variant-calling pipelines to discover the mutations in all individuals. Then we used Variant Annotation Analysis Selection Tool (VAAST) [10] to prioritize all the mutations and narrow down our candidate gene list. As a probabilistic disease gene finder, VAAST integrates allele frequency, amino acid substitution severity and conservation information into a composite likelihood framework. Different from hard filtering disease finding tools [11], VAAST prioritizes genes according to their disease-causing potential, ranking candidate genes without excluding any candidates, largely decreasing the false negative rate. It has been successfully applied in finding the causal mutation of an X-linked disorder with very limited sample size [12]. VAAST has been shown to outperform other mutation effect prediction software, including SIFT, Polyphen-2, CASM and Mutation Taster [13]. It's a very flexible and powerful tool with the properties that allows both dominant and recessive mode of inheritance, and allows the scoring of splice sites mutation and indels (insertions and deletions), in addition to single-nucleotide coding variants. In this study we applied WES and multiple bioinformatics tools to study the genetic bases of LSDs with unknown etiology in an exome wide scale.

2. Methods/Experimental Procedures

2.1 Sample information

Cell lines from 14 genetically unrelated patients were obtained by Dr. David Sleat (Center for Advanced Biotechnology and Medicine and Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey, United States of America) and Dr. Peter Lobel (Center for Advanced Biotechnology and Medicine and Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey, United States of America) from several sources (Table 2.1).

Table 2.1. Source or phenotype information of 14 samples.

Sample ID	Clinical Information
00RD098	LSD cases from Netherlands - no confirmed diagnosis
01RD492	
02RD297	
82RD265	
95RD414	
99RD299	
B1278	pycnodystosis-like, Cathepsin K positive
CABMHF11	thromocytopenia and splenomegaly, tests for Niemann-Pick negative
CABMHF210	untrastructure suggestive of NPC but tests negative
CABMHF311	neuronal storage, sphingolipidoses excluded by enzyme assays
CABMHF412	neurodegeneration
CABMHF52	severe neurodegeneration
HL508Pa	adult neuronal ceroid lipofuscinosis
TC98307	metaphyseal acroscyphodysplasia

The patients from whom these cell lines were derived displayed a spectrum of typical LSD phenotypes, such as different levels of neurodegeneration. Both histopathological and clinical evidence support that they were suffering LSD. However, none of the cases could be clinically diagnosed of being subjected to the defect of known LSD causal mutations – they were negative for known LSD tests. Research protocols involving human subjects were approved by the Institutional Review Board of the University of Medicine and Dentistry of New Jersey.

2.2 Exome Sequencing and Reads Mapping

The Whole Exome sequencing was performed on SOLiD platform in 50x25 bp format by Sequencing and Non-Coding RNA Program (The University of Texas MD Anderson Cancer Center, Houston, Texas). The exomes were enriched with Agilent SureSelect DNA - Human All Exon 50Mb Kit. Raw sequences were aligned to the human reference genome (version hg 19) using LifeScope ™ Genomic Analysis Software (http://www.lifetechnologies.com/lifescope), a tool kit developed by Life Technologies for SOLiD sequencing data.

2.3 Variant Calling

Variant discovery by the Genome Analysis Tool Kit (GATK) (version 2.8-1) genotyping pipeline roughly followed the GATK Best Practices recommendations [14]. Parameters are tuned based on the SOLiD data. Briefly, the raw sequence alignments (in binary alignment map (BAM) format) were reordered based on chromosomal coordinates with Picard tool (version 1.80) (http://picard.sourceforge.net), and then sorted with Samtools-0.1.19 [15] to adjust the

SOLiD sequencing format to be suitable for downstream GATK variant calling pipeline. Picard was used to index the BAM files, followed by a series of GATK alignment-processing procedures: indel realignment, remove duplication, base recalibration, which were all applied to individual BAM files. Then, a multi-individual genotype calling were performed on all individuals with GATK UnifiedGenotyper to generate the raw genotype call in a single variant calling format (VCF) file. Single nucleotide variants (SNVs) and indels (insertions and deletions) located outside of the targeted exome regions were removed based on the target-region definition provided by SureSelect DNA - Human All Exon 50Mb Kit. Lastly, the quality scores of both SNVs and indels were recalibrated with VariantRecalibrator according to the GATK recommended parameters. Detailed commands can be found in Appendix A.

Variant discovery by LifeScope™ Genomic Analysis Software and CLC Genomic Workbench were performed under near-default parameters following the manufacture's recommendation. CLC Variants with Depth of Coverage coverage (DP) less than 10 were removed. Variant calling with LifeScope™ Genomic Analysis Software was performed by Dr. Kumar Dibyendu (Waksman Genomics Core Facility, Waksman Institute, Rutgers University, Piscataway, New Jersey, United States of America). Variant calling with CLC Genomic Workbench was performed by Dr. Robert Donnelly (Department of Pathology & Laboratory Medicine, New Jersey Medical School, Rutgers University, Newark, New Jersey, United States of America).

2.4 Coverage Calculation

Variant coverage was calculated by a GATK tool DepthofCoverage with the following command: java -Djava.io.tmpdir=/lab01/tmp -Xmx168g -jar /usr/local/gatk-

```
2.5-2/GenomeAnalysisTK.jar --omitDepthOutputAtEachBase -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -T DepthOfCoverage -o
Lyso_coverage_L.txt -I bamlist.list -pt readgroup -L
padded_exome_region.interval_list > Lyso_L_ct15.log
```

2.5 Variant Calling Pipeline Comparisons

We calculated the number of all types of variants from the individual VCF files with a bash code. Since GATK called variants in population, it retained the no-call variants; while the LifeScope and CLC called variants individually, no-call variants have been excluded from the individual variant file. Therefore, we filtered the GATK no-call variants first, and then calculated the variants number of different categories in different individuals. The variant categories included transcript variant, gene variant, intron variant, exon variant and different types of coding variants: amino acid substitution, frameshift variant, splice donor variant, splice acceptor variant, stop gained, stop lost, stop retained and inframe variant. We calculated the average number of variants in each variant category of all the individuals for each variant calling pipeline, and compared the three different pipelines to determine the difference among them. Lastly we calculated the number of variants that shared by at most four individuals in each variant category of different pipelines and compared them. In our experience, four individuals are the optimal number to balance the sensitivity and specificity of the downstream analysis.

2.6 VAAST Analysis

The variant data were analyzed using the Variant Annotation, Analysis and Search Tool (VAAST) package. The VAAST analyses were conducted in parallel for the three datasets derived from different variant-calling pipelines. The variants in VCF were converted to Genome Variation Format (GVF) with vaast_converter. Genotypes with quality score (GQ) less than 30 were converted to no call. The variants were annotated based on their functional impact using VAT (Variant Annotation Tool). Gender information was utilized for the individuals with known gender to increase the accuracy of analysis on the sex chromosomes. Then, annotated variants shared by less than four individuals were combined into one condensed file using VST (Variant Selection Tool). Lastly, the VAAST analysis was conducted within the candidate genetic regions under both a dominant and recessive mode of inheritance. Both splice sites and indels were analyzed. The analysis was performed under the assumption of allowing 1% prevalence of a variant in the background population, which means we estimated the expected allele frequency of a disease-causing allele within the background population to be 1% or lower. Variants in each gene were scored as a group. Candidate genes were ranked based on their disease-causing probability and the p-value for each gene was determined by a permutation test. VAAST analysis was performed within a list of candidate genetic regions defined by Dr. Peter Lobel and Dr. David Sleat. Detailed commands can be found in Appendix B. The candidate gene list can be found in Supplementary Material I.

2.7 SIFT Analysis

The chromosome coordinates, reference allele, alternative allele, patient identifier information of all variants from VAAST output file were exported with a python code to make the SIFT input file. For example: “18,21140411,T,C,CABMHF11” stands for a mutation at chromosome 18, position 21140411 that is T to C in individual

CABMHF11. SIFT analysis was conducted at http://sift.jcvi.org/ with SIFT/PROVEAN Human SNPs database.

2.8 Tier Study

A bash code was scripted to perform the Tier Study. By intersecting the three data sets, we obtained different tier groups. Tier One contained variants that shared by all three data sets, while Tier Two included variants that shared by two data sets. The variants that were unique to one data set were categorized to Tier Three. Variants were ranked according to the order of Tier One, Tier Two and Tier Three. Under each tier group, homozygous variants were ranked with higher priority than heterozygous variants. Results can be found in Supplementary Material II.

2.9 Candidate Selection

After reprioritizing all the variants by the tier study, we selected the initial candidate variants based on the following criteria:

1. Homozygous mutation in each tier group, because homozygous variants have no compensation effect and are likely to contribute to the disease.
2. Compound heterozygous, two heterozygous variants that are in the same gene in one individual but at different chromosomal positions may disrupt the gene function.
3. Severe mutations based on VAAST and SIFT analysis results, i.e., the variant with higher VAAST score and the “Damaging” mutation marked by SIFT.
4. All of the three standards mentioned above should follow the tier group priorities, where Tier One ranks higher than Tier Two, and Tier Two ranks higher than Tier Three.

After the initial selection, we narrowed down our candidate variants list to 10 interesting mutations on 6 genes in 5 individuals as potential disease causing candidates. Detailed information for the 10 mutations is shown in Table 2.10.

2.10 Sanger Sequencing Validation

After candidate variant selection, we validated candidate variants by Sanger sequencing. Primers (Table 2.10) were designed with Primer3 [16] for PCR amplification of the genetic region covering the variants sites. The primers and PCR procedure used for the gender determination were as described previously [17]. Agarose gel electrophoresis was performed to validate the size of the amplicons. The gel with the right size of DNA fragment was cut and further purified with the Wizard® SV Gel and PCR Clean-Up System. Genes with unclear band was cloned with ZERO BLUNT TOPO kit (Life Technologies). In addition to single-mutation validation, larger fragments containing two close-by mutations were further amplified and cloned from *NPC1* and *SLC31A1*. Individual clones of the large fragments were sequenced to determine if the two mutations are on the same or different chromosomal copies. Purified DNA products or molecular cloning products were sequenced by ABI 3730 DNA Sequencer (GenScript). Then genotypes were examined with BioEdit Sequence Alignment Editor.

Table 2.10 Primers used for gene validation. Primers used for large fragment amplification were marked in red.

Individual ID	Gene Name	Coordinate	Reference Allele	Alternative Allele	Amplicon Primer Forward	Amplicon Primer Reverse	Length	Large Fragment Length
00RD098	GLB1	chr3:33099692	G	A	TTCCCTGCTCTTTTTTCACTCACAG	CTGCAATTTCTGTTACTACAAACACC	235	
	GLB1	chr3:33055721	A	G	TCCTTCCCTCCCCAGCTCACTGTG	GAATTCAAACCCTTCCCATGAAGAC	328	
82RD265	SLC31A1	chr9:116021039	C	T	CAAGCAGTCTGACCAAAAGGT	CAGGCATGGAATTGTAGCGAA	385	2096
	SLC31A1	chr9:116022721	G	T	AAGTACCCATGAGTTGCCAGA	CTTCAACAACTTCCCACTGCA	382	
95RD414	GLA	chrX:100653420	C	A	ATGGCTGCTCCTTTATTCATGT	AAACCAAGAAAGTGTGGTTGCT	397	
	SMPD1	chr11:6413175	C	A	TGACTGTGCAGACCCACTGT	TGCTTTCATGGTTACCCACA	308	
CABMHF11	NPC1	chr18:21136367	C	T	TGATTCCTGCCATGAGATAGCAACT	CCCATCTAGCAGTAGTCAACATGTA	556	4415
	NPC1	chr18:21140411	T	C	GTATTTCAGTGGGCTTTTCTTTGAGT	CATGGAGGTATTTGTTTCTTGTCCTA	457	
	SMPD1	chr11:6415259	G	A	CACCATCCCTGTTGTCCCATGGAGT	CACAGGGCTCCGAGGGTGGGT	713	
CABMHF311	AP3B1	chr5:77396837	T	.	TTGAGACAAATGTTGATTCAGGA	TTGGGACATGTAAATGAAAGGT	325	

3. Results

3.1 Analysis Overview

A schematic describing the entire analysis procedure is shown in Figure 3.1. In brief, WES was performed on 14 individuals. Three variant calling pipelines -- GATK, CLC and Lifescope -- were implemented to identify variants on all the samples independently. For each variant dataset, we applied VAAST analysis to prioritize candidate genes. Because of the disconcordance of variant calling methods, Tier Study was performed to rerank candidates. Then, SIFT analysis was conducted to add one more layer of genetic information and evaluation of candidate mutations. Lastly, disease causing candidate mutations were selected based on our criteria and in select cases, mutational or functional validations were or will be performed to test the genetic bases of disease.

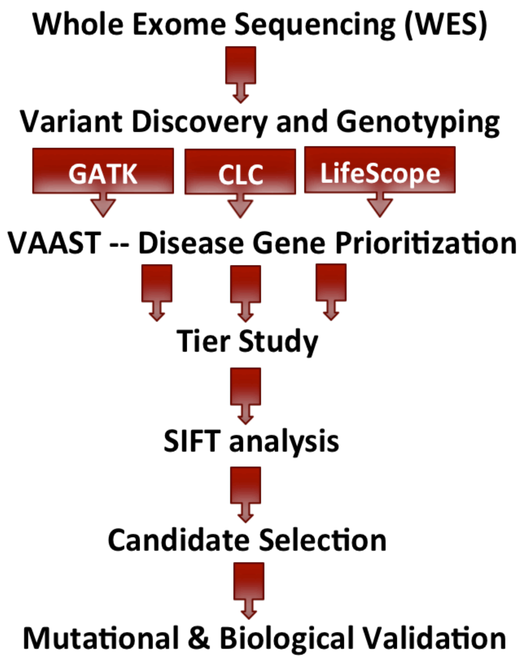

Whole Exome Sequencing (WES)
Variant Discovery and Genotyping
GATK
CLC
LifeScope
VAAST -- Disease Gene Prioritization
Tier Study
SIFT analysis
Candidate Selection
Mutational & Biological Validation

Figure 3.1 Analysis procedure overview.

3.2 Exome Sequencing and Variant Discovery Pipelines

We performed WES on 14 genetically unrelated patients suspected to have LSDs. We calculated the sequencing coverage. The average exome-wide mean coverage for all the samples was 22.63-fold (Table 3.2).

Table 3.2 Individual coverage calculation results.

Individual ID	Total Bases	Mean Coverage
00RD098	2037480379	22.5
01RD492	2179163363	24.07
02RD297	2023856437	22.35
82RD265	2421322830	26.74
95RD414	1584030773	17.5
99RD299	1941261757	21.44
B1278	1891984968	20.9
CABMHF11	1996568992	22.05
CABMHF210	1652477376	18.25
CABMHF311	1835520789	20.27
CABMHF412	2304968208	25.46
CABMHF52	1306118808	14.43
HL508Pa	2108762505	23.29
TC983077	2007500640	22.17
Average	1949358416	21.53

We implemented three different variant calling pipelines to genotype all the samples. They are Genomic Analysis Tool Kit (GATK) [14], LifeScope™ Genomic Analysis Software and CLC Genomic Workbench (For clarity, GATK, CLC and LifeScope, respectively, were used as abbreviations in the rest of this thesis).

3.3 Variant Calling Pipeline Comparisons

To compare the overall performance of the three variant calling pipelines, we counted the number of different variants in each individual.

First, we compared different types of variants averaged per individual among the three calling pipelines (Table 3.3.1). Overall, GATK identified a similar mutation load distribution with LifeScope, but CLC found a much smaller number of variants. For functional variants, such as frameshift_variant, stop_retained variant, LifeScope called more than GATK and CLC.

Table 3.3.1 Overall mutation load comparison. The mutation loads of different variant categories was listed under each variant calling pipeline.

Variant Type	GATK	LifeScope	CLC
transcript_variant	171581	152331	45507
gene_variant	173767	157751	48699
intron_variant	115128	97290	17148
exon_variant	5804	6157	2873
amino_acid_substitution	41000	29550	14186
frameshift_variant	206	596	180
splice_donor_variant	139	196	31
splice_acceptor_variant	257	635	27
stop_gained	427	747	163
stop_lost	81	73	23
stop_retained	29	26	10
inframe_variant	189	303	79

Because we are mainly interested in rare disease-causing variants, next we compared the number of total variants after we filtered the variants according to allele frequency (at most 0.285 within the 14-individual dataset or present in no more than three individuals) (Table 3.3.2). After filtering, GATK called a similar number of mutations to CLC output. However, the number of variants from LifeScope is larger than the other two call sets. The functional variants showed the same pattern.

Table 3.3.2 Total variants comparison after filtering by allele frequency. Only variants shared by no more than three individuals were presented.

Variant Type	GATK	LifeScope	CLC
transcript_variant	35267	99421	30478
gene_variant	35602	102764	32574
intron_variant	19541	64156	12530
exon_variant	1338	3973	1848
amino_acid_substitution	9857	19856	9027
frameshift_variant	59	461	121
splice_donor_variant	37	163	21
splice_acceptor_variant	32	535	16
stop_gained	137	576	119
stop_lost	9	44	10
stop_retained	8	19	9
inframe_variant	45	208	49

From the comparison, we can see that the variant calling pattern is different among the pipelines. Therefore, we performed downstream analyses on all three data sets to reduce pipeline bias and make full use of all variant calling information.

3.4 VAAST Analysis

With the aim of finding the disease causing mutations, we conducted the VAAST analysis using variants from each variant calling pipeline. The VAAST package contains Variant Annotation Tool (VAT), Variant Selection Tool (VST) and Variant Analysis Tool (VAAST). All variants were annotated with VAT according to their features.

For those cases where we have the gender information, we specified the gender option for VAT analysis to increase the accuracy of analysis of variants on the sex chromosomes. After successfully annotating all of the variants, we ran the VST, a tool to select variants of interest based on our hypothesis. In this study, all the 14 patients are genetically unrelated thus it is unlikely that they shared the exact same disease-causing mutation. Therefore, we only targeted variants that are shared by less than 4 individuals, which means we condensed the entire variants that only appeared in at most 3 individuals into a single file in the process of VST. By doing this we could largely exclude common false-positive variants generated during the variant calling process and thus narrow down our candidate variant pool.

The last step is VAAST, prioritizing the candidate genes according to their disease-causing potential. A VAAST score is assigned to each variant and the higher the score, the more likely it is a disease-causing defect. VAAST analysis was performed under both dominant and recessive modes of inheritance respectively, while allowing splice site and indel variant finding. To narrow down the candidate genes, we obtained a list of 848 LSD candidate genes, including both known disease genes and suspected genes from Dr. Peter Lobel and Dr. David Sleat. Based on the candidate gene list, we conducted VAAST analysis only within the genomic regions containing candidate genes. This way, the efficiency and effectiveness of our analysis was enhanced. The different pipelines produced different numbers of candidates scored by the VAAST. CLC workbench gave 123 variants, GAKT gave 153 variants and LifeScope scored 338 variants. Removing genes that overlapped across the three methods, 408 candidate variants were identified.

3.5 Comparison of VAAST Results among Different Variant-calling Data Sets

We compared the variants scored by VAAST of the three data sets, and found that different pipelines resulted in different amounts of variants. They shared some variants, but all three sets had their own unique variants. Moreover, the variant genotype wasn't always concordant among different pipelines. Some of them were consistent among all the pipelines; some of them are consistent within two pipelines, while others are unique to certain pipeline. Low concordance of multiple variant-calling pipelines is known because of the difference of variant-calling algorithms and their parameterization efficacy. However, it has been shown that the concordance rates of novel and unique-to-pipeline SNVs increase for variants called by an increasing number of pipelines [18]. Therefore, to make full use of all three bioinformatics pipelines, we conducted a tier study, reprioritizing the variants based on their concordance.

By intersecting different pairs of datasets in terms of both variants and their genotypes, we obtained different tier group. As showed in Figure 3.5, the Tier One group contained the 53 variants that shared by all three datasets; the Tier Two group included variants shared by two groups – 1 variants shared by CLC and GATK, 14 variants shared by CLC and LifeScope, and 94 variants shared by LifeScope and GATK; lastly, the Tier Three group possessed variants unique to one dataset – 55 variants unique to CLC, 177 variants unique to LifeScope and 5 variants unique to GATK.

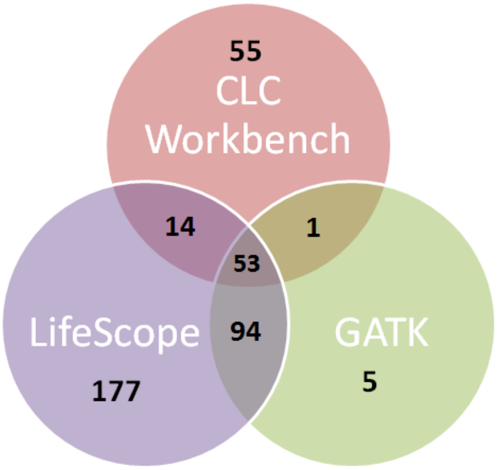


Figure 3.5 Venn diagram of pipeline comparisons. The number of variants was listed in corresponding tier group.

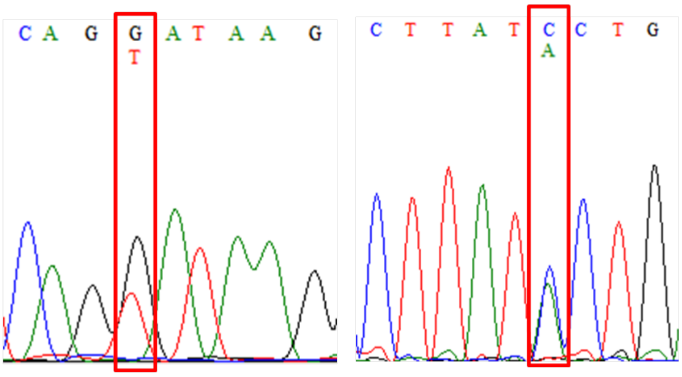
3.6 SIFT Analysis

To acquire more information and evaluation of the variants, we conducted SIFT (Sorting Tolerant From Intolerant) analysis [19]. SIFT predicts the functional effect of amino acid substitution mutations based on the degree of conservation of amino acid residues in closely related sequence alignments. It provides a user-friendly outcome for each mutation, scoring the mutation based on their predicted functional disruption severity and categorizing them into two groups: Damaging and Tolerated. Among 408 identified variants, SIFT predicted 167 damaging variants. Besides grouping variants, SIFT also provided the detailed information of codon change and residue substitution, which were used for data mining and mutation validation. SIFT results can be found in Supplementary Material II.

3.7 Mutation Validation

Sanger sequencing was performed to verify the 10 candidate variants, including 9 SNPs and one 1bp deletion in corresponding patient genomes. For all 9 SNPs we validated the alternative allele: 3 mutations were validated as homozygous and the others were heterozygous. Two examples were shown in Figure 3.7.1. For the 1bp deletion, the Sanger sequencing result showed a homozygous 3bp deletion at the position (Figure 3.8.3).

(a) GLA_Forward GLA_Reverse



(b) SMPD1_Forward SMPD1_Reverse

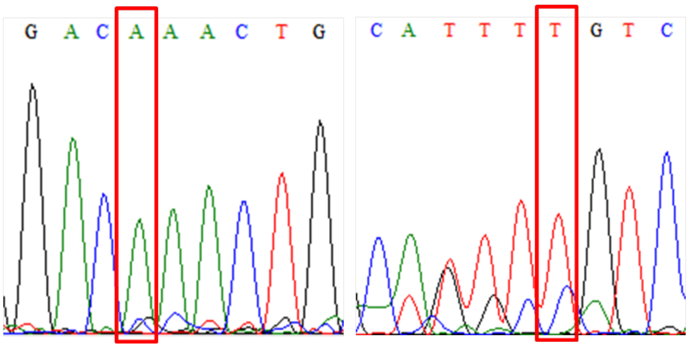


Figure 3.7.1 Sequence electropherogram examples of gene *GLA* and *SMPD1*. (a) Variant "X,100653420,C,A,95RD414" on gene *GLA* was heterozygous with signals of both reference allele (G) and alternative allele (T) on the forward strand. Sequencing of the reverse strand confirmed the heterozygous signals of both reference allele (C) and alternative allele (A). (b) Variant "11,6413175,C,A,95RD414" on the gene *SMPD1* was homozygous with the predominant signal of mutant allele (A) on the forward strand and predominant signal of mutant allele (T) on the reverse strand.

Because the candidate mutation in individual 95RD414 is on the X chromosome, the gender of this case was confirmed by PCR using the methods developed by Hedges et.al[17]. We used both *AluSTXa* and *AluSTYa*, the two monomorphic *Alu* insertions fixed on X chromosome and Y chromosome respectively, to validate the gender of the individual. The agarose gel electrophoresis result showed a single band for both of the two insertions on individual 95RD414 – 878bp for *AluSTXa* and 199bp for *AluSTYa* (Figure 3.7.2), which means that 95RD414 is a female. Therefore, this variant is heterozygous, consistent with the Sanger validation result (Figure 3.7.1(a)).

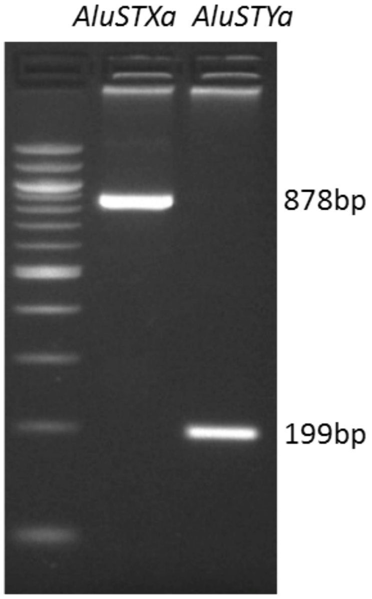


Figure 3.7.2 Mobile element-based gender determination. An agarose gel chromatograph from the analysis of individual 95RD414 using the genetic systems *AluSTXa* and *AluSTYa* is shown. Females are distinguished by the presence of one DNA fragments, while males have two amplicons. The size of DNA fragment is listed on the right side.

3.8 Candidate Disease-causing Mutations

Based on our mutational validation results and data mining analysis, we found 9 candidate disease-causing mutations on six genes in five individuals. To check if the validated mutations are known pathogenic variants, we searched the Human Gene Mutation Database (HGMD) [20] and identified 4 known pathogenic variants in four genes; 2 novel variants in known pathogenic genes; and, most excitingly, 3 novel mutations in two genes related with lysosomal function (Table 3.8.1). These are discussed on a case by case basis below.

Table 3.8.1 Candidate Disease-causing Mutations. Ten mutations on six genes in five individuals were listed in the table. Both known and novel variants are included.

VARIANT TYPE	KNOWN (PATHOGENIC) VARIANTS					NOVEL VARIANTS IN KNOWN PATHOGENIC GENES		NOVEL PATHOGENIC VARIANTS		
INDIVIDUAL ID	**00RD098**		**95RD414**		**CABMHF11**			**82RD265**		**CABMHF311**
GENE SYMBOL	**GLB1**		**SMPD1**	**GLA**	**NPC1**		**SMPD1**	**SLC31A1**		**AP3B1**
GENE DESCRIPTION	Galactosidase beta		Sphingomyelin phosphodiesterase 1	Galactosidase alpha	Niemann-Pick disease, type C	Niemann-Pick disease, type C1	Sphingomyelin phosphodiesterase 1	Solute carrier family 31 (copper transporters), member 1	Solute carrier family 31 (copper transporters), member 2	Adaptor-related protein complex 3, beta 1 subunit
COORDINATE	chr3:33099692	chr3:33055721	chr11:6413175	chrX:100653420	chr18:21140411	chr18:21136367	chr11:6415259	chr9:116021039	chr9:116022721	chr5:77396838
LENGTH	677	677	631	429	1278	1278	631	190	190	1045
STRAND	-1	-1	1	-1	-1	-1	1	1	1	-1
CODON_CHANGE	TTT [C/T]GC CAC	GTG [T/C]GC AGC	GAC [C/A]AA CTG	CAG [G/T]AT AAG	AAC A[A/G]T GCC	GCT C[G/A]C CTG	ATC [G/A]GC CTT	GCC [C/T]GA GAG	GCA [G/T]TG GTA	GAA [AAG/-] AAA
POSITION	208	521	294	313	222	389	492	90	181	754
REFRENECE RESIDUE	R	C	Q	D	N	R	G	R	V	K
ALTERNATIVE RESIDUE	C	R	K	Y	S	H	S	*	L	.
TYPE	Single AA Change	Single AA Change	Single AA Change	Single AA Change	Single AA Change	Single AA Change	Single AA Change	Nonsense	Single AA Change	Deletion
SIFT PREDICTION	Damaging	Tolerated	Damaging	Damaging	Tolerated	Damaging	Tolerated	NA	Damaging	Deleterious
TIER GROUP	Tier2		Tier2 homo	Tier1	Tier3	Tier3	Tier3 homo	Tier 2	Tier 2	Tier3 homo
SANGER VALIDATION	Het	Homo, C (ref is T)	Homo, A (ref is C)	Het	Het	Het	Het	Het	Het	Deletion
PHENOTYPE	Gangliosidosis GM1		Niemann-Pick disease	Fabry disease	Niemann-Pick C disease	Niemann Pick C type 1 disease	Niemann Pick A disease			
REFRENCE	Boustany (1993) Am J Hum Genet 53, 881	Caciotti (2005) Hum Mutat 25, 285	Pavlu (1997) J Inherit Metab Dis 20, 615	Eng (1993) Am J Hum Genet 53, 1186	Park (2003) Hum Mutat 22, 313					

1) Case 00RD098

We found two previously reported mutations on the protein coded by Galactosidase beta (*GLB1*): R208C [21] and C521R [22]. β-galactosidase (GLB1) enzyme deficiency will cause GM1 Gangliosidosis. R208C was validated as a heterozygous variant in case 00RD098. It is a known pathogenic allele according to Silva et al [23]. The other mutation, C521R, was validated as a homozygous variant with alternative allele G (reference is A). C521R is reported to be a netural polymorphism by Silva et al [23]. However, Caciotti et al [24] expressed GLB1 with C521R mutation and found a 75% decrease in the GLB1 activity. We performed a beta-galactosidase enzyme assay [25] for 00RD098, and no significant diminish was observed. In the 1000 Genome Project, the alternative allele G is the major allele, with the allele frequency of 76% in African and 100% in European (http://browser.1000genomes.org/Homo_sapiens/Variation/Population?db=core;r=3:33055721-33055721;source=dbSNP;v=rs4302331;vdb=variation;vf=3279748). Therefore, this mutant is most likely to be a neutral polymorphism rather than pathogenic in individual 00RD098.

2) Case 95RD414

Sphingomyelin phosphodiesterase (SMPD1) is a lysosomal acid sphingomyelinase that converts sphingomyelin to ceramide. Defects of *SMPD1* gene result in Niemann-Pick A and B diseases. A known pathogenic mutant Q294K [26] on SMPD1 was found in this case. Note that this allele is sometimes referred to as Q292K [26]. Comparing the sequence containing the Q292K mutation from the original paper [27] with SMPD1 sequence (http://www.ncbi.nlm.nih.gov/protein/NP_000534.3), we can see that all of the reported alleles are -2 positions compared to the reference sequence. Homo-allelic tranversion C to A has been confirmed to be disease-causing [26]. Our Sanger sequencing validation confirmed that the Q294k mutant in 95RD414 is homozygous (Figure 3.7.1(b)). Therefore, it's likely that this established pathogenic mutation produced an LSD phenotype in case 95RD414. Moreover, X chromosomal mutation D313Y [28] in Galactosidase alpha (*GLA*) is validated to be heterozygous (Figure 3.7.1(a)) and could also have functional impact. Downstream validation is required for further confirmation.

3) Case CABMHF11

We found a known pathogenic mutation N222S [29] on Niemann-Pick disease, type C3 gene (*NPC1*). *NPC1* encodes a large protein that resides in the membrane of endosomes and lysosomes, mediating intracellular cholesterol trafficking. Low-density lipoproteins (LDL) carry cholesterol in the plasma. Circulatory LDL is endocytosed by cells and is delivered to late endosomal/lysosomal compartments where the cholesterol esters are hydrolyzed. Normally, the free cholesterol is transported out of lysosomes. Impairment in lysosomal cholesterol transport arises when either of two proteins NPC1 or NPC2 are defective, causing NPC disease. . [30]. Besides the known pathogenic NPC1 variant, we also found a novel variant on NPC1 in the same individual: R389H. There are known pathogenic mutations at the position 389 with different amino acid substitutions: R389L [31] and R389C [29], which both lead to Niemann-Pick disease C (Table 3.8.2), making this mutation a very promising candidate. Moreover, since two mutations are present on *NPC1* in case CABMHF11, the two mutations can collectively lead to disease through compound heterogeneity. We amplified and cloned a large *NPC1* DNA fragment containing the two variant loci and sequenced different clones to determine if the two mutations are present on different chromosomes. The sequencing results supported our hypothesis (Figure

3.8.1). The mutant alleles were separately located on the two copies of chromosome 18: mutant allele C located on one copy of chromosome 18 at position 21136367, while mutant allele G located on another copy of chromosome 18 at position chr18: 21140411. Protein coded by either chromatid would be disrupted by a mutant allele. Fillipin staining, which detects the accumulation of cholesterol in lysosomes, will be performed to validate the functional defects of *NPC1* in our case.

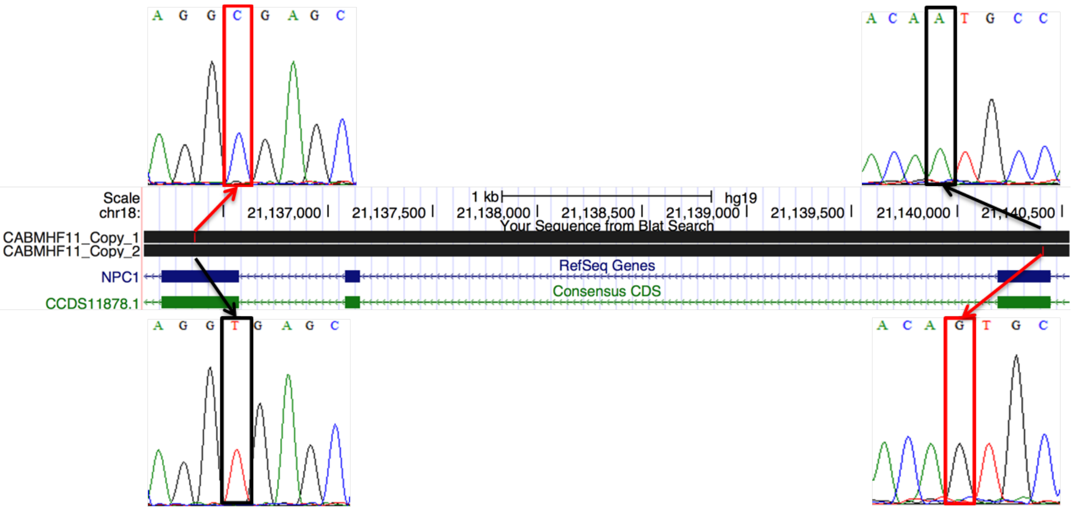


Figure 3.8.1 *NPC1* Compound Heterogeneity Illustration. Two copies of gene *NPC1* in individual CABMHF11 are shown as thick solid black track. Two mutants on different copies are marked in red line. Red arrows lead to the electropherogram of the mutant allele (circled by red box) and its flanking sequence, while black arrows lead to the mutant-corresponding reference allele (circled by black box) and its flanking sequence. The reference is shown in blue. Consensus coding sequence is shown in green. The sequence plot is generated by UCSC genome browser Custom Tracks tool. For gene annotation, exonic regions are shown as solid boxes, while non-exonic regions are shown as thin lines, with arrows indicating the direction of the gene.

Additionally, in individual CABMHF11, we also found another novel variant in the known pathogenic gene *SMPD1*: G492S. Although no pathogenic mutations are known at position 492, there are known pathogenic mutations very close to it: T488A [32], and Y490N [33] (Table 3.8.2).

Table 3.8.2 Related variants of novel mutations on known pathogenic genes.

INDIVIDUAL ID		**CABMHF11**		**CABMHF11**	
GENE SYMBOL		NPC1		SMPD1	
GENE DESCRIPTION		Niemann-Pick disease, type C1		Sphingomyelin phosphodiesterase	
POSITION		389		492	
REFERENCE RESIDUE		R		G	
ALTERNATIVE RESIDUE		H		S	
TYPE		Single AA Change		Single AA Change	
PHENOTYPE		Niemann Pick C type 1 disease		Niemann Pick A disease	
CHARACTERISTICS		There was mutation at the same position, but not the same amino acid substitution		There was mutation around this position	
HGMD INFORMATION	**HGMD Accession Number**	CM096652	CM032619	CM093885	CM023159
	Codon Change	CGC-CTC	tCGC-TGC	aACT-GCT	cTAC-AAC
	Amino Acid	Arg-Leu	Arg-Cys	Thr-Ala	Tyr-Asn
	Codon Number	389	389	488	490
	Phenotype	Niemann-Pick disease C	Niemann-Pick disease C	Niemann-Pick disease	Niemann-Pick disease
	Reference	Fancello (2009) Neurogenetics 10, 229	Park (2003) Hum Mutat 22, 313	Rodríguez-Pascau (2009) Hum Mutat 30, 1117	Simonaro (2002) Am J Hum Genet 71, 1413

4) Case 82RD265

We found 2 novel mutations in a lysosome-functional related gene in individual 82RD265: a nonsense mutation R90* and a non-synonymous V181L mutation in the Solute carrier family 31 (copper transporters), member 1 (*SLC31A1*). The V181L was predicted as damaging by SIFT. *SLC31A1* was proposed to be a high-affinity copper uptake gene. It’s responsible for the uptake of at least 80% of copper and other metals into cells [34]. The dysfunction of copper metabolism can leads to human disease such as Wilson disease, which is characterized by dramatic build-up of intracellular hepatic copper with subsequent hepatic and neurologic abnormalities [35]. Although SLC31A1 has been reported to localize on the plasma membrane[36], another SLC31 family member *SLC31A2* has been shown to localize in lysosomes and facilitates cellular copper uptake [37]. All the indirect evidence implied that our mutations are highly likely to be involved in the LSDs. Moreover, similar to the case CABMHF11, two heterozygous mutations of *SLC31A1* in the same individual implied the possibility of compound heterogeneity loss-of-function of these two mutations. We did similar large fragment amplification, cloning and sequencing analyses. Tests results supported our hypothesis (Figure 3.9.2). The two mutations are located on different chromosomes, so each copy of the *SLC31A1* has a heterozygous mutation in individual 82RD265. Further functional analysis to test the cellular copper uptake would be necessary to demonstrate the effect of these mutations.

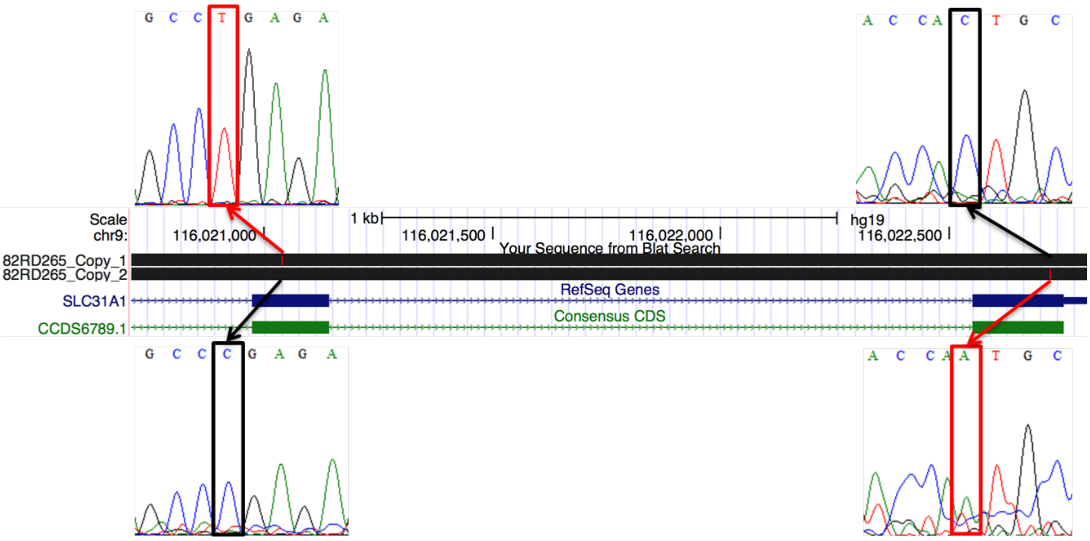


Figure 3.8.2 *SLC31A1* Compound Heterogeneity Illustration. Two copies of gene *SLC31A1* in individual 82RD265 are shown as thick solid black track. Two mutants on different copies are marked in red line. Red arrows lead to the electropherogram of the mutant allele (circled by red box) and its flanking sequence, while black arrows lead to the mutant-corresponding reference allele (circled by black box) and its flanking sequence. The reference is shown in blue. Consensus coding sequence is shown in green. The sequence plot is generated by UCSC genome browser Custom Tracks tool. For gene annotation, exonic regions are shown as solid boxes, while non-exonic regions are shown as thin lines, with arrows indicating the direction of the gene.

5) Case CABMHF311

We found another novel mutation in individual CABMHF311. It is an inframe deletion of Lysine at the postion754 on Adaptor-related protein complex 3, beta 1 subunit coded by gene *AP3B1*. This protein is involved in protein trafficking to lysosomes or specialized endosomal-lysosomal organelles such as pigment granules, melanosomes, and platelet dense granules [38]. Adaptor protein complex 3 is a ubiquitous cytoplasmic complex that shuttles cargo proteins from the trans-Golgi and a tubular-endosomal compartment to endosome-lysosome-related organelles. Lack of the beta-3A subunit of this complex causes Hermansky-Pudlak syndrome type 2 [39]. A study on a dog disease -- canine cyclic neutropenia -- indicated *AP3B*1 as their candidate genes because a lysine deletion led to a polyA track and in turn results in transcriptional slippage in *AP3B1* mRNA at the equivalent position to the human lysine deletion [40]. In human the locus appears to be a mix of simple repeats. The repetitive feature makes the locus susceptible for replication slippage and creating small indels both during replication and transcription, resulting in frameshifting mutations. The *AP3B1* sequence can be found in Appendix C. Based on our Sanger sequencing validation, we see a clean 3bp deletion signal (Figure 3.8.3), which implied that this mutation is homozygous. Therefore, we believe that deleterious mutations on these lysosomal related proteins are very likely to be disease causing. Of course, functional validation must be pursued for confirmation of our hypothesis. To test if the lysine deletion results in a polyA tract that in turn leads to transcriptional infidelity, we will conduct realtime-PCR to examine the *AP3B1* RNA expression level of patient CABMHF311. Significant decrease or complete absence of expression level will confirm the loss of function of *AP3B1* due to the deletion.

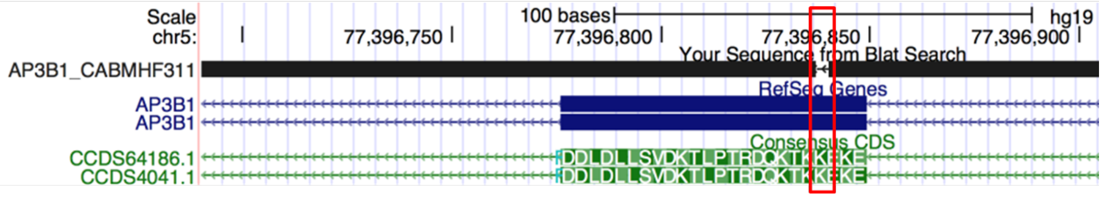


Figure 3.8.3 ***AP3B1*** **3bp inframe deletion signal.** The sequence of gene *AP3B1* in individual CABMHF311 is shown as thick solid black track. The reference is shown in blue. The consensus coding amino acid sequence is shown in green. 3bp inframe deletion signal is circled in red box. It is a lysine deletion according to the amino acid sequence. The sequence plot is generated by UCSC genome browser Custom Tracks tool. For gene annotation, exonic regions are shown as solid boxes, while non-exonic regions are shown as thin lines, with arrows indicating the direction of the gene.

4. Discussion

In this study, we applied multiple bioinformatics tools to decode and analyze the WES data from 14 suspected LSD patients. Nine mutations on 6 genes in 5 individuals have been selected as candidate disease-causing mutations. Four of nine mutations are known pathogenic mutations. This result supports the accuracy of our selection and implies the applicability of using WES or whole genome sequencing (WGS) as a genetic diagnostic method for LSDs with the decrease of sequencing cost. We also found novel mutations in known pathogenic genes, expanding the category of known pathogenic mutations in LSDs and also improving the sensitivity of genetic testing. Most excitingly, we found novel mutations in genes that are related to lysosomal function but have not previously been associated with human disease, e.g., SLC31A1, which is important for downstream functional research to unveil the genetic bases of LSDs with unknown etiology.

In addition to proteomics study, bioinformatics methods provide a novel approach for identifying the etiology of unknown LSD cases. We were able to search disease-causing mutations on an exome-wide scale, and multiple variant calling pipelines were adopted to reduce technical bias. Pipeline discordance was expected because of their difference in post-alignment data processing (e.g., different quality filters), analysis parameters, and the underlying models utilized by each algorithm [18]. However, by combing VAAST, SIFT and a Tier Study, we were able to prioritize candidate genes in an informative way and we identified nine disease causing mutations.

However, we did not find disease-causing mutations for all 14 patients. Several reasons may account for the limitation. First of all, there are limitations in our hypotheses. Our study was based on two hypotheses: 1) fewer than 4 individuals would share the disease causing mutations because the cases are thought to be genetically unrelated; 2) the diseases were caused by monogenic variant. If the suspected LSDs were caused by a common mutation shared by more than 4 individuals within our samples, the sensitivity of our selection would be largely decrease because they have been excluded by VAAST analysis at the VST step. If the diseases were caused by the epistatic effect of multiple less-deleterious mutations, the specificity of our selection would also decline. Second, each bioinformatics method has its own false-positive/false-negative rate, and the combined false-positive/false-negative rate limits the power of our selection. For example, our tier study biases toward the variants that are shared by all three datasets, and biases against true variants exist in each pipeline-unique dataset. VAAST, a probability disease gene finder, may rank false positive mutations high if it is present in multiple individuals. Furthermore, SIFT prediction on mutation effect is not always 100% accurate. Although we tried to reduce the false negative rate by combing different methods together, we could still miss certain causal mutations in our data set. Third, we limited our search region within the 848 candidate genes. Any mutations outside of these regions will not be detected, such as mutations in genes encoding novel lysosomal proteins. Fourth, WES focuses on exonic variants in the genome. If a disease is caused by the disruption of a regulatory element, the disease-causing mutations will not be discovered in our datasets.

For the future directions, we can expand the search scale to whole exome beyond the 848 candidate genes. If there are still undiagnosed patients, we can expand to whole genome, increasing the chances of finding disease-causing mutations in non-coding regions. Modifying the VST parameters and increasing our candidate mutation pool can be applied to evaluate mutations shared by more than 4 individuals. Thorough downstream functional analysis of candidate novel mutations is important to unveil novel molecular mechanisms of LSDs.

5. Conclusion

By using three variant-calling pipelines and three prioritization bioinformatics tools, we identified both known pathological mutations and novel LSDs causing mutations on 6 genes in 5 individuals from the WES data of 14 patients suspected to have LSDs. With the application of high-throughput sequencing and bioinformatics analysis, we were able to efficiently identify a list of candidate genes, offering feasible amounts of candidates for downstream biochemical and molecular research. We provide a new approach for genetic testing-based diagnosis of LSDs and shed light on the genetic bases of LSDs with unknown etiology.

Appendix A. GATK variant calling commands

The GATK variant calling commands of individual 00RD098 are listed below as examples.

Picard: ReorderSam

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gkno/gkno_launcher/tools/picard/dist/ReorderSam.jar
I=00RD098.bam O=00RD098.reordered.bam
R=/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta
VALIDATION_STRINGENCY=LENIENT
```

Samtools: sort

```
samtools sort -m 16G -@ 32 00RD098.reordered.bam 00RD098.reordered.sorted
```

Picard: BuildBamIndex

```
java -Djava.io.tmpdir=/lab01/tmp -jar
/usr/local/gkno/gkno_launcher/tools/picard/dist/BuildBamIndex.jar I=00RD098.reordered.sorted.bam
```

GATK:RealignerTargetCreator

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gatk-2.5-2/GenomeAnalysisTK.jar -T
RealignerTargetCreator -I 00RD098.reordered.sorted.bam -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -o 00RD098.reordered.sorted.intervals -nt
32
```

GATK: IndelRealigner

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gatk-2.5-2/GenomeAnalysisTK.jar -T IndelRealigner -
I 00RD098.reordered.sorted.bam -R /lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -
compress 0 -targetIntervals 00RD098.reordered.sorted.intervals -o
00RD098.reordered.sorted.realigned.bam
```

Samtools: rmdup

```
samtools rmdup 00RD098.reordered.sorted.realigned.bam
00RD098.reordered.sorted.realigned.marked.bam
```

Picard: BuildBamIndex

```
java -Djava.io.tmpdir=/lab01/tmp -jar
/usr/local/gkno/gkno_launcher/tools/picard/dist/BuildBamIndex.jar
I=00RD098.reordered.sorted.realigned.marked.bam
```

GATK: BaseRecalibrator

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gatk-2.5-2/GenomeAnalysisTK.jar -T
BaseRecalibrator -I 00RD098.reordered.sorted.realigned.marked.bam -o
00RD098.reordered.sorted.realigned.marked.bam.perged.grp -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -knownSites
/lab01/DataSets/hg19/GATK_bundle/hg19/dbsnp_137.hg19.vcf -nct 32 --solid_nocall_strategy
PURGE_READ
```

GATK: PrintReads

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gatk-2.5-2/GenomeAnalysisTK.jar -T PrintReads  -I
00RD098.reordered.sorted.realigned.marked.bam -BQSR
00RD098.reordered.sorted.realigned.marked.bam.nocallperged.grp -
o 00RD098.reordered.sorted.realigned.marked.nocallperged.recalibrated.bam -
R /lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -nct 32
```

Once the alignment processing for each individual were finished, UnifiedGenotyper was applied to genotype all the variants and merge them into a single VCF file.

GATK: UnifiedGenotyper

```
java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/gatk-2.5-2/GenomeAnalysisTK.jar -T
UnifiedGenotyper -I all_bam_files  -o Lyso_Nan.vcf -
R /lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta --
dbsnp /lab01/DataSets/hg19/GATK_bundle/hg19/dbsnp_137.hg19.vcf -glm BOTH
```

GATK: SelectVariants

```
/usr/local/jdk1.7.0_40/jre/bin/java -Djava.io.tmpdir=/lab01/tmp -Xmx64g -jar
/usr/local/GenomeAnalysisTK-2.7-1/GenomeAnalysisTK.jar -T SelectVariants -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta --variant
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan.vcf -o
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_Exome_Selected.vcf
–L
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/ExomeEnrichment/S02972011/S02972011_Pa
dded_noheader.bed
```

GATK: VariantRecalibrator (SNP)

```
/usr/local/jdk1.7.0_40/jre/bin/java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/GenomeAnalysisTK-2.7-
1/GenomeAnalysisTK.jar -T VariantRecalibrator -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -input
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_Exome_Selected.vcf
-mode SNP -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an
InbreedingCoeff -recalFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_SNP_target.recal -
tranchesFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_SNP_target.tranches
-rscriptFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_SNP_target.plots.R
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
/lab01/DataSets/hg19/GATK_bundle/hg19/hapmap_3.3.hg19.vcf -
resource:omni,known=false,training=true,truth=false,prior=12.0
/lab01/DataSets/hg19/GATK_bundle/hg19/1000G_omni2.5.hg19.vcf -
resource:dbsnp,known=true,training=false,truth=false,prior=6.0
/lab01/DataSets/hg19/GATK_bundle/hg19/dbsnp_137.hg19.vcf
```

GATK: VariantRecalibrator (Indel)

```
/usr/local/jdk1.7.0_40/jre/bin/java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/GenomeAnalysisTK-2.7-
1/GenomeAnalysisTK.jar -T VariantRecalibrator -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -input
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_Exome_Selected.vcf
-mode INDEL -resource:mills,known=false,training=true,truth=true,prior=12.0
/lab01/DataSets/hg19/GATK_bundle/hg19/Mills_and_1000G_gold_standard.indels.hg19.vcf -
resource:dbsnp,known=true,training=false,truth=false,prior=2.0
/lab01/DataSets/hg19/GATK_bundle/hg19/dbsnp_137.hg19.vcf -an DP -an FS -an ReadPosRankSum -
an MQRankSum -numBad 1000 --maxGaussians 4 -recalFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_INDEL_target.recal
-tranchesFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_INDEL_target.tranc
```

```
hes -rscriptFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_INDEL_target.plots.
R;
```

GATK: ApplyRecalibration (SNP)

```
/usr/local/jdk1.7.0_40/jre/bin/java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/GenomeAnalysisTK-2.7-
1/GenomeAnalysisTK.jar -T ApplyRecalibration -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -input
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_Exome_Selected.vcf
--ts_filter_level 99.0 -tranchesFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_SNP_target.tranches
-recalFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_SNP_target.recal -
mode SNP -o
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_recalibrated.SNP.vc
f -nt 32;
```

GATK: ApplyRecalibration (Indel)

```
/usr/local/jdk1.7.0_40/jre/bin/java -Djava.io.tmpdir=/lab01/tmp -jar /usr/local/GenomeAnalysisTK-2.7-
1/GenomeAnalysisTK.jar -T ApplyRecalibration -R
/lab01/DataSets/hg19/GATK_bundle/hg19/ucsc.hg19.fasta -input
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_recalibrated.SNP.vc
f --ts_filter_level 99.0 -tranchesFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_INDEL_target.tranc
hes -recalFile
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_INDEL_target.recal
-mode INDEL -o
/lab01/Projects/Nan_Projects/Lysosome/variant_calling/Nan_vcf_files/Lyso_Nan_recalibrated.SNP.IN
DEL.vcf -nt 32;
```

Appendix B. VAAST analysis commands

Convert VCF to GVF:

```
nice /usr/local/VAAST/bin/vaast_tools/vaast_converter --build hg19 -i 00RD098  00RD098.vcf
```

Filter GVF:

```
nice awk --posix '($1 !~ /^$/) && ($1 !="chrM") && ($3 !="gap") {print $0}' 00RD098.gvf > 00RD098.filter.gvf
```

Sort GVF:

```
nice /usr/local/VAAST_hao_dev/bin/vaast_tools/vaast_sort_gff -i -n 00RD098.filter.gvf
```

Variant Annotation:

```
nice /usr/local/VAAST_hao_dev/bin/VAT -f /lab01/Projects/VAAST_Projects/Data/Features/genes_only_ref_GRCh37.p10_top_level_chr_only_uniq.gff -a /lab01/Projects/VAAST_Projects/Data/Fasta/vaast_hsap_chrs_hg19.fa --sex male -v quiet /lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/Filter_Sort/CABMHF52.filter.sorted.gvf > /lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/VAT.GVF_GENDER/CABMHF52.vat.gvf 2>CABMHF52_log&
```

Variant Selection:

```
nice /usr/local/VAAST_hao_dev/bin/VST -o 'S("<4",0..13)' -b hg19 /lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/VAT.GVF/*.gvf> no_g_w_gender.cdr 2>no_g_w_gender.log
```

VAAST (Recessive) analysis:

```
nice /usr/local/VAAST_hao_dev/bin/VAAST -iht r -lh y -splice_site --indel -d 1e4 -r 0.01 -m lrt -mp1 8 --less_ram -fast_gp -regions /lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/Second_Round_VAAST/Target_region/Nov15th_uniq_gene_region.bed -l /lab01/Projects/VAAST_Projects/Data/Phastcons/phastcons-hg19-vertebrate.txt -o /lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/Second_Round_VAAST/Nov20th_VAAST/Results/vaast_no_g_w_gender/Lyso_sec_Nov20_nog_wgender_recessive /lab01/Projects/VAAST_Projects/Data/Features/genes_only_ref_GRCh37.p10_top_level_chr_only_uni
```

```
q.gff /lab01/Projects/VAAST_Projects/Data/Background_CDR/1304-doped.cdr
/lab01/Projects/Nan_Projects/Lysosome/waksman_vcf_files/VAAST_Lysosome_Analysis/Second_Round_VAAST/Nov20th_VAAST/CDR_no_g_w_gender/no_g_w_gender.cdr
2>Lyso_sec_Nov20_nog_wgender_recessive_log&
```

Appendix C. The sequence of gene *AP3B1*

Lysine 803 was marked in yellow

Amino Acid Sequence (1095aa):

MSSNSFPYNEQSGGGEATELGQEATSTISPSGAFGLFSSDLKKNEDLKQMLESNKDSAKLDAMKRIVGMI

AKGKNASELFPAVVKNVASKNIEIKKLVYVYLVRYAEEQQDLALLSISTFQRALKDPNQLIRASALRVLS

SIRVPIIVPIMMLAIKEASADLSPYVRKNAAHAIQKLYSLDPEQKEMLIEVIEKLLKDKSTLVAGSVVMA

FEEVCPDRIDLIHKNYRKLCNLLVDVEEWGQVVIIHMLTRYARTQFVSPWKEGDELEDNGKNFYESDDDQ

KEKTDKKKKPYTMDPDHRLLIRNTKPLLQSRNAAVVMAVAQLYWHISPKSEAGIISKSLVRLLRSNREVQ

YIVLQNIATMSIQRKGMFEPYLKSFYVRSTDPTMIKTLKLEILTNLANEANISTLLREFQTYVKSQDKQF

AAATIQTIGRCATNILEVTDTCLNGLVCLLSNRDEIVVAESVVVIKKLLQMQPAQHGEIIKHMAKLLDSI

TVPVARASILWLIGENCERVPKIAPDVLRKMAKSFTSEDDLVKLQILNLGAKLYLTNSKQTKLLTQYILN

LGKYDQNYDIRDRTRFIRQLIVPNVKSGALSKYAKKIFLAQKPAPLLESPFKDRDHFQLGTLSHTLNIKA

TGYLELSNWPEVAPDPSVRNVEVIELAKEWTPAGKAKQENSAKKFYSESEEEEDSSDSSSDSESESGSES

GEQGESGEEGDSNEDSSEDSSSEQDSESGRESGLENKRTAKRNSKAKGKSDSEDGEKENEKSKTSDSSND

ESSSIEDSSSDSESESEPESESESRRVTKEKEKKTKQDRTPLTKDVSLLDLDDFNPVSTPVALPTPALSP

SLMADLEGLHLSTSSSVISVSTPAFVPTKTHVLLHRMSGKGLAAHYFFPRQPCIFGDKMVSIQITLNNTT

DRKIENIHIGEKKLPIGMKMHVFNPIDSLEPEGSITVSMGIDFCDSTQTASFQLCTKDDCFNVNIQPPVG

ELLLPVAMSEKDFKKEQGVLTGMNETSAVIIAAPQNFTPSVIFQKVVNVANVGAVPSGQDNIHRFAAKTV

HSGSLMLVTVELKEGSTAQLIINTEKTVIGSVLLRELKPVLSQG

Nucleotide Sequence (3285 nt):

ATGTCCAGCAATAGTTTTCCTTACAATGAGCAGTCCGGAGGAGGGGAGGCGACGGAGCTGGGTCAGGAGG

CGACCTCAACCATTTCCCCCTCGGGGGCCTTCGGCCTCTTTAGCAGCGATTTGAAGAAGAATGAAGATCT

AAAGCAAATGTTAGAGAGCAACAAAGATTCTGCTAAACTGGATGCTATGAAGCGGATTGTTGGGATGATT

GCAAAAGGGAAAAATGCATCTGAACTGTTTCCTGCTGTTGTGAAGAATGTGGCCAGTAAAAATATTGAGA

TCAAGAAGTTGGTATATGTTTACCTGGTTCGATATGCTGAAGAACAGCAGGATCTTGCACTCCTGTCCAT

AAGCACTTTTCAGCGAGCTCTGAAGGACCCAAACCAACTAATTCGTGCAAGCGCTTTGAGAGTTCTGTCA

AGTATTAGAGTGCCAATTATTGTACCTATCATGATGCTTGCTATTAAGGAAGCTTCTGCTGACTTATCAC

CATATGTTAGGAAGAATGCAGCCCATGCAATACAAAAATTATACAGCCTTGATCCAGAGCAGAAGGAAAT

GTTAATTGAAGTAATTGAAAAACTTCTGAAAGATAAAAGCACATTGGTAGCTGGCAGTGTTGTGATGGCT

TTTGAAGAAGTATGCCCGGACAGAATAGATCTGATTCATAAAAATTACCGCAAGCTATGTAACTTACTAG

TGGATGTTGAAGAGTGGGGGCAGGTTGTCATAATCCACATGCTAACTCGATATGCTCGGACACAGTTTGT

CAGCCCTTGGAAAGAGGGTGATGAATTAGAAGACAATGGAAAGAATTTCTACGAATCTGATGATGATCAG

AAGGAAAAGACTGACAAAAAGAAGAAGCCGTATACTATGGATCCAGATCATAGACTCTTAATTAGAAATA

CAAAGCCTTTGCTTCAGAGCAGGAATGCTGCGGTGGTTATGGCAGTTGCTCAGCTGTATTGGCACATATC

ACCAAAATCTGAAGCTGGCATAATTTCTAAATCACTAGTGCGTTTACTTCGTAGCAATAGGGAGGTGCAG

TATATTGTCCTACAAAATATAGCAACTATGTCAATTCAAAGAAAGGGGATGTTTGAACCTTATCTGAAGA

GTTTCTATGTTAGGTCAACTGATCCAACTATGATCAAGACACTGAAGCTTGAAATTTTGACAAACTTGGC

AAATGAAGCCAACATATCAACTCTTCTTCGAGAATTTCAGACCTATGTGAAAAGCCAGGATAAACAATTT

GCAGCAGCCACTATTCAGACTATAGGCAGATGTGCAACCAACATCTTGGAAGTCACTGACACGTGCCTCA

ATGGCTTGGTCTGTCTGCTGTCCAACAGGGATGAAATAGTTGTTGCTGAAAGTGTGGTTGTTATAAAGAA

ATTACTGCAAATGCAACCTGCACAACATGGTGAAATTATTAAACATATGGCCAAACTCCTGGACAGTATC

ACTGTTCCTGTTGCTAGAGCAAGTATTCTTTGGCTAATTGGAGAAAACTGTGAACGAGTTCCTAAAATTG

CCCCTGATGTTTTGAGGAAGATGGCTAAAAGCTTCACTAGTGAAGATGATCTGGTAAAACTGCAGATATT

AAATCTGGGAGCAAAATTGTATTTAACCAACTCCAAACAGACAAAATTGCTTACCCAGTACATATTAAAT

CTCGGCAAGTATGATCAAAACTACGACATCAGAGACCGTACAAGATTTATTAGGCAGCTTATTGTTCCGA

ATGTAAAGAGTGGAGCTTTAAGTAAATATGCCAAAAAAATATTCCTAGCACAAAAGCCTGCACCACTGCT

TGAGTCTCCTTTTAAAGATAGAGATCATTTCCAGCTTGGCACCTTATCTCATACTCTCAACATTAAAGCT

ACTGGGTACCTGGAATTATCTAATTGGCCAGAGGTGGCGCCCGACCCATCAGTTCGAAATGTAGAAGTAA

TAGAGTTGGCAAAAGAATGGACCCCAGCAGGAAAAGCAAAGCAAGAGAATTCTGCTAAGAAGTTTTATTC

TGAATCTGAGGAAGAGGAGGACTCTTCTGATAGTAGCAGTGACAGTGAGAGTGAATCTGGAAGTGAAAGT

GGAGAACAAGGCGAAAGTGGGGAGGAAGGAGACAGCAATGAGGACAGCAGTGAGGACTCCTCCAGTGAGC

AGGACAGTGAGAGTGGACGGGAGTCAGGCCTAGAAAACAAAAGAACAGCCAAGAGGAACTCAAAAGCCAA

AGGAAAAAGTGATTCTGAAGATGGGGAGAAGGAAAATGAAAAATCTAAAACTTCAGATTCTTCAAATGAC

GAATCTAGTTCAATAGAAGACAGTTCTTCCGATTCTGAATCAGAGTCAGAACCTGAAAGTGAATCTGAAT

CCAGAAGAGTCACTAAGGAGAAAGAAAAGAAAACAAAGCAAGATAGAACTCCTCTTACCAAAGATGTTTC

ACTTCTAGATCTGGATGATTTTAACCCAGTATCCACTCCAGTTGCACTTCCCACACCAGCTCTTTCTCCA

AGTTTGATGGCTGATCTTGAAGGTTTACACTTGTCAACTTCCTCTTCAGTCATCAGTGTCAGTACTCCTG

CATTTGTACCAACGAAAACTCACGTGCTGCTTCATCGAATGAGTGGAAAAGGACTAGCTGCCCATTATTT

CTTTCCAAGACAGCCTTGCATTTTTGGTGATAAGATGGTCTCTATACAAATAACACTGAATAACACTACT

GATCGAAAGATAGAAAATATCCACATAGGGGAAAAAAAACTTCCTATAGGCATGAAAATGCATGTTTTTA

ATCCAATAGACTCTCTTGAGCCTGAGGGATCCATTACAGTTTCAATGGGTATTGACTTTTGTGATTCTAC

TCAGACTGCCAGTTTCCAGTTGTGTACCAAGGATGATTGCTTCAATGTTAATATTCAGCCACCTGTTGGA

GAACTGCTTTTACCTGTGGCCATGTCAGAGAAAGATTTTAAGAAAGAGCAAGGAGTGCTAACAGGAATGA

ATGAAACTTCTGCTGTAATCATTGCTGCACCACAGAATTTCACTCCCTCTGTGATCTTTCAGAAGGTTGT

AAATGTAGCCAATGTAGGTGCAGTCCCTTCTGGCCAGGATAATATACACAGGTTTGCAGCTAAAACTGTG

CACAGTGGGTCATTGATGCTAGTCACAGTGGAACTGAAGGAAGGCTCTACAGCCCAGCTTATCATAAACA

CTGAGAAAACTGTGATTGGCTCTGTTCTGCTGCGGGAACTGAAGCCTGTCCTGTCTCAGGGGTAA

Bibliography

1. Segatori, L., *Impairment of homeostasis in lysosomal storage disorders.* IUBMB Life, 2014. **66**(7): p. 472-7.

2. Sleat, D.E., M. Jadot, and P. Lobel, *Lysosomal proteomics and disease.* Proteomics Clin Appl, 2007. **1**(9): p. 1134-46.

3. Lubke, T., P. Lobel, and D.E. Sleat, *Proteomics of the lysosome.* Biochim Biophys Acta, 2009. **1793**(4): p. 625-35.

4. Meikle, P.J., et al., *Prevalence of lysosomal storage disorders.* JAMA, 1999. **281**(3): p. 249-54.

5. Sanderson, S., et al., *The incidence of inherited metabolic disorders in the West Midlands, UK.* Arch Dis Child, 2006. **91**(11): p. 896-9.

6. Maire, I., *Is genotype determination useful in predicting the clinical phenotype in lysosomal storage diseases?* J Inherit Metab Dis, 2001. **24 Suppl 2**: p. 57-61; discussion 45-6.

7. Charles R. Scriver, W.S.S., Barton Childs and Arthur L.Beaudet, *The Metabolic and Molecular Bases of Inherited Diseases, 4 volumn set.* Dec 15, 2000.

8. Beutler, E., *Gaucher disease: multiple lessons from a single gene disorder.* Acta Paediatr Suppl, 2006. **95**(451): p. 103-9.

9. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.

10. Yandell, M., et al., *A probabilistic disease-gene finder for personal genomes.* Genome Res, 2011. **21**(9): p. 1529-42.

11. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.

12. Rope, A.F., et al., *Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency.* Am J Hum Genet, 2011. **89**(1): p. 28-43.

13. Hu, H., et al., *VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix.* Genet Epidemiol, 2013. **37**(6): p. 622-34.

14. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

15. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

16. Steve Rozen, H.J.S., *Primer3.* 1998.

17. Hedges, D.J., et al., *Mobile element-based assay for human gender determination.* Anal Biochem, 2003. **312**(1): p. 77-9.

18. O'Rawe, J., et al., *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.* Genome Med, 2013. **5**(3): p. 28.

19. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

20. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

21. Boustany, R.M., W.H. Qian, and K. Suzuki, *Mutations in acid beta-galactosidase cause GM1-gangliosidosis in American patients.* Am J Hum Genet, 1993. **53**(4): p. 881-8.

22. Caciotti, A., et al., *Role of beta-galactosidase and elastin binding protein in lysosomal and nonlysosomal complexes of patients with GM1-gangliosidosis.* Hum Mutat, 2005. **25**(3): p. 285-92.

23. Silva, C.M., et al., *Six novel beta-galactosidase gene mutations in Brazilian patients with GM1-gangliosidosis.* Hum Mutat, 1999. **13**(5): p. 401-9.

24. Caciotti, A., et al., *Primary and secondary elastin-binding protein defect leads to impaired elastogenesis in fibroblasts from GM1-gangliosidosis patients.* Am J Pathol, 2005. **167**(6): p. 1689-98.

25. Sleat, D.E., et al., *Rat brain contains high levels of mannose-6-phosphorylated glycoproteins including lysosomal enzymes and palmitoyl-protein thioesterase, an enzyme implicated in infantile neuronal lipofuscinosis.* J Biol Chem, 1996. **271**(32): p. 19191-8.

26. Pavlu, H. and M. Elleder, *Two novel mutations in patients with atypical phenotypes of acid sphingomyelinase deficiency.* J Inherit Metab Dis, 1997. **20**(4): p. 615-6.

27. Schuchman, E.H., et al., *Human acid sphingomyelinase. Isolation, nucleotide sequence and expression of the full-length and alternatively spliced cDNAs.* J Biol Chem, 1991. **266**(13): p. 8531-9.

28. Eng, C.M., et al., *Nature and frequency of mutations in the alpha-galactosidase A gene that cause Fabry disease.* Am J Hum Genet, 1993. **53**(6): p. 1186-97.

29. Park, W.D., et al., *Identification of 58 novel mutations in Niemann-Pick disease type C: correlation with biochemical phenotype and importance of PTC1-like domains in NPC1.* Hum Mutat, 2003. **22**(4): p. 313-25.

30. Carstea, E.D., et al., *Niemann-Pick C1 disease gene: homology to mediators of cholesterol homeostasis.* Science, 1997. **277**(5323): p. 228-31.

31. Fancello, T., et al., *Molecular analysis of NPC1 and NPC2 gene in 34 Niemann-Pick C Italian patients: identification and structural modeling of novel mutations.* Neurogenetics, 2009. **10**(3): p. 229-39.

32. Rodriguez-Pascau, L., et al., *Identification and characterization of SMPD1 mutations causing Niemann-Pick types A and B in Spanish patients.* Hum Mutat, 2009. **30**(7): p. 1117-22.

33. Simonaro, C.M., et al., *The demographics and distribution of type B Niemann-Pick disease: novel mutations lead to new genotype/phenotype correlations.* Am J Hum Genet, 2002. **71**(6): p. 1413-9.

34. Wee, N.K., et al., *The mammalian copper transporters CTR1 and CTR2 and their roles in development and disease.* Int J Biochem Cell Biol, 2013. **45**(5): p. 960-3.

35. de Bie, P., et al., *Molecular pathogenesis of Wilson and Menkes disease: correlation of mutations with molecular defects and disease phenotypes.* J Med Genet, 2007. **44**(11): p. 673-88.

36. Zhou, B. and J. Gitschier, *hCTR1: a human gene for copper uptake identified by complementation in yeast.* Proc Natl Acad Sci U S A, 1997. **94**(14): p. 7481-6.

37. van den Berghe, P.V., et al., *Human copper transporter 2 is localized in late endosomes and lysosomes and facilitates cellular copper uptake.* Biochem J, 2007. **407**(1): p. 49-59.

38. Dell'Angelica, E.C., et al., *Association of the AP-3 adaptor complex with clathrin.* Science, 1998. **280**(5362): p. 431-4.

39. Fontana, S., et al., *Innate immunity defects in Hermansky-Pudlak type 2 syndrome.* Blood, 2006. **107**(12): p. 4857-64.

40. Benson, K.F., et al., *Paradoxical homozygous expression from heterozygotes and heterozygous expression from homozygotes as a consequence of transcriptional infidelity through a polyadenine tract in the AP3B1 gene responsible for canine cyclic neutropenia.* Nucleic Acids Res, 2004. **32**(21): p. 6327-33.