Exome sequence analysis of siblings affected with Niemann-Pick type C disease

Shaun Carswell

A thesis submitted to Victoria University of Wellington in part fulfillment of the requirement for the degree of Masters of Cell and Molecular Bioscience

2017

Abstract

Mutations in either the Niemann-Pick type C1 or C2 (NPC1/NPC2) gene result in a fatal lysosomal storage disorder, Niemann-Pick type C (NP-C) disease, for which there is no effective cure. The disease is characterized by systemic and neurodegenerative symptoms arising from toxic accumulation of unesterified cholesterol within the late endosome and lysosome, with a common cause of death for patients being respiratory failure or recurrent infection of pulmonary tissue. Interestingly, the disease symptoms are heterogeneous, with age of onset and severity varied, even among siblings with the same mutations in the NPC1 or NPC2 gene causing this monogenic disease. To date there is no clear explanation for disease severity in siblings with the same mutation. As siblings are raised in the same environment, the major hypothesis of this thesis is that there are genetic modifiers that explain variation in disease severity within siblings. To determine if there are genetic variants associated with disease severity, exomes were sequenced from five sibling pairs exhibiting divergent onset and progression of NPC disease. Out of 23,105 genes, 26 variants were identified that were predicted to have functional consequences in NP-C patients, of which homozygous MUC5B and MARCH8 variants segregated across siblings that exhibited increased and decreased severity of disease, respectively. A cluster of variants was discovered on chromosome 11 belonging to the matrix metalloproteinase (MMP) family. Further investigation of one of these variants, a frameshift insertion in MMP-12, confirmed that this locus regulates the accumulation of unesterified cholesterol in primary neurons derived from a murine model of NPC disease. However, this region on chromosome 11 did not have any statistically significant copy number alteration detectable through a depth of coverage analysis. Overall, these results provide groundwork into the sequence variants mediating disease severity, which with further investigations, may be novel pharmacological targets to treat NPC disease.

Acknowledgements

Andrew Munkacsi – For being the bravest researcher I know of, who places the science first and foremost, before everything else – including grants. It takes courage to see through the unimportant systems and cut right to the chase. Thank you for taking a chance on me, and letting me get lost in the thickets of research, trusting that I'd find my way back.

Stephen Sturley – Without leadership of this project and the generous contribution of these sequences, I would never have been able to start this thesis.

Paul Teesdale Spittle – The support throughout my masters and his insights into protein structure and guidance to ask the appropriate questions to consider with *in silico* modelling are greatly appreciated.

The physicians and families associated with this study – These tremendous individuals who experience NP-C first hand, who assume the hardest and most important role of all – without your bravery, compassion and generosity, this work would not have been possible. Thank you.

Dinindu Senanayake – For introducing me to the Rosalind community and all you've done for the 300 level labs over the years. When I started here I was drowning, not waving. Seeing you make the shore made the swim possible – thank you for the sea of data to swim in. By the way, the opera has broken again. Would you mind looking into that…?

Jeff and Liam – Loyal members of the Zaibatsu. Daijoubu desu!

Margaret Carswell & Ian Carswell – You've given me so much more than I deserve. There used to be a dream here. Now, there is something greater.

Taron Orsborn – Komm, süßer Tod.

Dale Carswell – By your side.

Seb Morwiski – My guiding moonlight.

Penny McFetridge – A counterbalance to the world of research. Without your kindness and knowledge of New Zealand's wild places, it would be easy to forget the beauty of the outside world. Thank you for Waiheke, Wanaka and all the beauty of those memories I carry in my heart now.

Coffee – A bioinformatician is a tool for converting coffee into code.

Katie McFetridge – Through it all, you have been there for me. The sleepless, harrowed nights and the years of anxiety – you've been there, my source of strength and hope. You have been a keeper of the fire that has kept me on this path. I can only hope to now kindle your flame, as you have mine.

Table of Contents

Table of Figures

Table of Tables

Table of Scripts

List of Abbreviations

FA/FASTA	FASTA file format
AUC	Area under curve
AZD	Alzheimer's disease
BALL	Biochemical Algorithms Library
BAM	Binary Alignment/Map
BED	Browser Extensible Data
BQSR	Base quality score recalibration
BWA	Burrows-Wheeler Aligner
CNS	Central nervous system
CNV	Copy Number Variation
contig	Overlapping sequencing data
COPD	Chronic obstructive pulmonary disease
DAPI	4',6-diamidino-2-phenylindole
DP	Depth
DMEM	Dulbecco's Modified Eagle Medium
ER	Endoplasmic Reticulum
ESCRT	Endosomal sorting complexes required for transport
FASTQ	Format for storing nucleotide sequence and associated quality metric
FCS	Fetal calf serum
FIJI	Fiji Is Just ImageJ
FS	Fisher Strand
GATK	Genome Analysis Tool Kit
GBA	β-Glucocerebrosidase
GD	Gaucher disease
GQ	Genome Quality
hg38	Genome Reference Consortium Human Reference 38
gVCF	Genomic variant call format
HPBCD	hydroxypropyl-β-cyclodextrin
HDACi	Histone deacetylase inhibitor

INDEL	Insertion/Deletion
LDL	Low-density lipoprotein
LSD	Lysosomal storage disorder
MEM	Maximal exact matches
MMP	Matrix metalloproteinase
MMPi	Matrix metalloproteinase inhibitor
MQ	Root mean square mapping quality
MQRS	Mapping quality rank sum test
MVBs	Multi-vesicular bodies
MW	Molecular weight
ND	Neutral density
NGS	Next generation sequencing
NP-C1	Niemann-Pick type C1 disease
PA	Phosphatidic acid
PBS	Phosphate-buffered saline
PCR	Polymerase Chain Reaction
Penstrep	Penicillin Streptomycin
PE	Paired-end
PED	Pedigree file-format
PFA	Paraformaldehyde
PODKAT	Position-Dependent Kernel Association Test
QD	Quality by depth
RMS	Root mean square
ROS	Reactive oxygen species
RPM	Revolutions per minute
RPRS	Read position rank sum test
SAM	Sequence Alignment/Map
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOR	Strand odds ratio

SREBP	Sterol regulatory element binding protein pathway
SW	Smith-Waterman algorithm
TfR	Transferrin receptor
TGN	Trans-golgi network
Ti/Tv	Transition transversion ratio
U18	(3β)-3-[2-(Diethylamino)ethoxy]androst-5-en-17-one hydrochloride
VCF	Variant call format
VGSP	Vertical supranuclear gaze palsy
VQSR	Variant quality score recalibration
WES	Whole exome sequence
WGS	Whole genome sequence

1 LITERATURE REVIEW

1.1 Lysosomes

Lysosomes were first biochemically characterized in 1955 after Christian de Duve noted that enzymes such as acid phosphatase of rat liver were latent in their activity as well as particle bound. This discovery led to the recognition of lysosomal roles within autophagy, such as the degradation of macromolecules, as well insights into the structural organization of the cell (de Duve, 2005). It is now recognized that the lysosome is a catabolically active organelle, due to an acidic pH maintained around 4-5 via a vacuolar proton-ATPase pump. This pH is maintained in a steady-state via secondary ion movement through counter-ion channels preventing a membrane potential difference (Ishida et al., 2013). Lysosomes contain high concentrations of calcium as well as highly glycosylated lysosomal-associated membrane proteins (LAMPs), both of which are involved in fusion and docking events (Appelqvist et al., 2013; Luzio et al., 2007; Saftig and Klumperman, 2009).

Cargo from the cellular membrane can be internalized through a network of endocytic vesicles, early endosome and late endosome compartments to the lysosomal organelle. This internalization of cargo from the plasma membrane to early/late endosomes is known as the endocytic pathway, which results in the degradation of the cargo by the acid hydrolase enzymes contained within the lysosomal compartment. In brief, the endosome can i) directly fuse with the lysosome, ii) hand-off vesicles the lysosome via buds, iii) transfer cargo via transient contact, and iv) form hybrid organelles to accomplish complex fusion (Luzio et al., 2007).

Late endosomes contain a greater number of vesicles than early endosomes and are sometimes referred to as multi-vesicular bodies (Luzio et al., 2007). The creation of lumenal vesicles within the endosomal components (both early and late) and their uptake of ubiquitinated proteins involves endosomal sorting complexes required for transport (ESCRT)

complexes. After uptake, the protein to be degraded is deubiquitinated and sorted into a budding vesicle for transport to the lysosome (Bowers et al., 2006).

As well as lysosomal fusion with late endosomes, the lysosome can also interact with phagosomes, autophagosomes and the plasma membrane. These interactions can facilitate recycling of macromolecular components within the cell, or even membrane repair (Luzio et al., 2007). Taken together, it is clear the lysosome is a complex organelle capable of fusing with a myriad of other intracellular components with a diverse range of functions.

The simplest take on the function of the lysosome is the degradation and recycling of cellular components, which can occur via several different pathways. The late endosomal/lysosomal complex results in the degradation of the endosomal cargo via lysosomal acid hydrolase, which is released from the TGN and accepted via the mannose-6-phosphate receptor (Smith, 2014). Specifically, the lysosome plays a key role in several forms of autophagy (Platt et al., 2012; Saftig and Klumperman, 2009), including macroautophagy, microautophagy and chaperone-mediated autophagy. Macroautophagy begins with the creation of an autophagosome, a vesicle that encapsulates cellular debris, which fuses with the lysosome to form an autolysosome, facilitating degradation (Luzio et al., 2007; Mellman, 1996). Microautophagy, however, involves the invagination of the lysosomal membrane to allow the endocytosis of the surrounding cytosol (Platt et al., 2012). Chaperone-mediated autophagy is a

direct, selective process for proteins that is regulated by LAMP receptors.

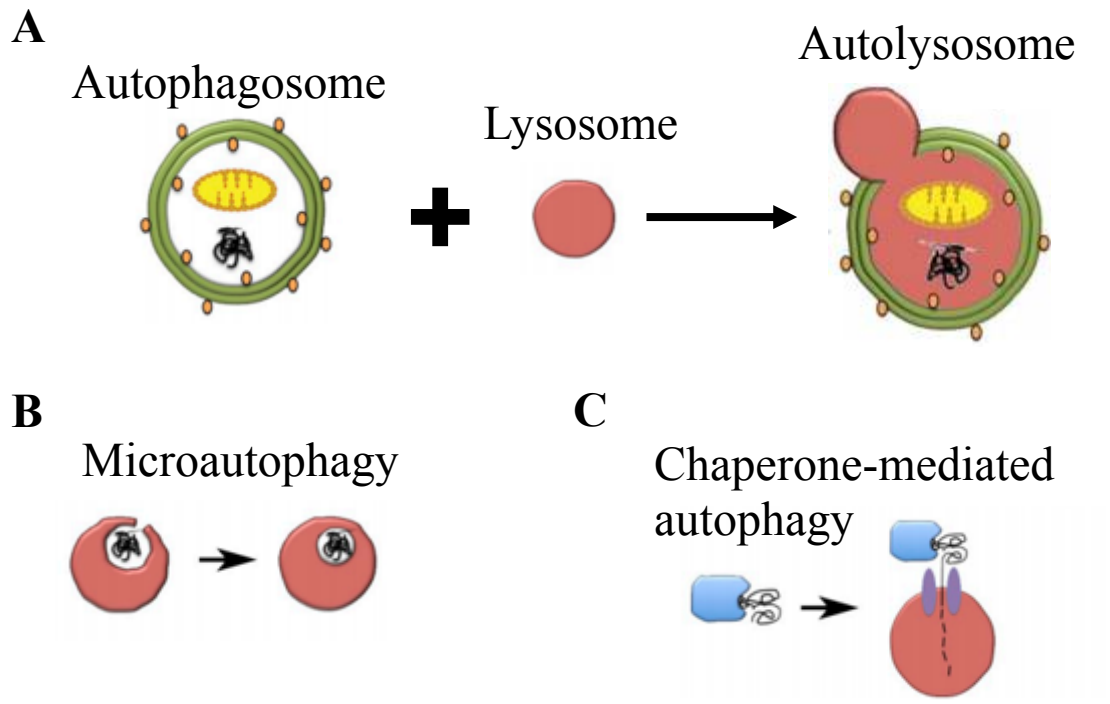


Figure 1.1 – **Three key forms of autophagy.** A) Macroautophagy involves the encapsulation of cellular material into autophagosomes. B) Microautophagy utilizes pinocytosis to capture cytosolic proteins. C) Chaperone-mediated autophagy uses lysosome associated membrane proteins as a receptor for chaperone proteins for selective delivery to lysosomes. Figure adapted from Platt et al., 2012.

The lysosome is involved in processes beyond degradation and recycling. The previously mentioned calcium reserves within the lysosome, along with the diverse range of acid hydrolases (including the cathepsin proteases), have roles in plasma membrane repair, remodelling, homeostasis and apoptosis (Kirkegaard and Jäättelä, 2009). Exocytosis of lysosomal enzymes can result in effects as wide ranging as extracellular matrix remodelling or initiating the lysosomal cell death pathway (Kirkegaard and Jäättelä, 2009; Luzio et al., 2007).

1.2 Lysosomal storage disorders

Defects in lysosomal function result in defects in cellular homeostasis and diseases in the case of humans (Appelqvist et al., 2013; Platt et al., 2012; Saftig and Klumperman, 2009; Vanier, 2014; Vitner et al., 2010). Genetic disorders of metabolism specifically relating to the

family of lysosomal storage diseases (LSDs) are inheritable diseases that, despite their individual rarity, are nonetheless clinically devastating to those affected and have a collective incidence of 1 in 5,000 live births (Fuller et al., 2006). As such, LSDs illustrate the vital functions of lysosomal proteins and hydrolases within the larger endosomal-lysosomal system. LSDs are commonly characterised by a lack of a functional lysosomal enzyme or protein, thereby disrupting lysosomal homeostasis and resulting in the accumulation of carbohydrates, lipids or proteins. As an inheritable genetic disease, a LSD usually presents clinically at a young age, however late-onset pathology is also possible (Platt et al., 2012; Vanier, 2014). As a family of 50 diseases, LSDs can be split into two broad categories based on whether the mutation affects a lysosomal enzyme or lysosomal membrane protein.

A genetic mutation resulting in deficiency of a functional lysosomal enzyme or insufficient enzymatic activity can lead to macromolecular accumulation and metabolic dysregulation arising from impaired endolysosomal efflux (Platt et al., 2012). A classic example of this is Gaucher disease (GD), the most common LSD with an estimated incidence of 5% of total LSD cases (Pinto et al., 2003). GD is the result of a defect in the lysosomal enzyme spell it out (GBA) with the major tissues affected being the bone marrow, liver and spleen (Platt et al., 2012). This is due to defective macrophage (Gaucher cells) infiltration into these organs (Gülhan et al., 2012; Lo et al., 2012). Another example of enzymatic deficiency is Niemann-Pick disease Type B (NP-B), where sphingomyelinase defects result in accumulation of sphingomyelin within the liver of sufferers (Brady et al., 1966).

For the second category of LSD, recall that lysosomes are of acidic pH and carry vesicles that contain hydrolytic enzymes, with a membrane that includes proton ATP-dependent pumps, along with other transport systems (Vellodi, 2005) - these play a significant role in the sorting, recycling and digestion of endocytosed material within the cellular body. These include lysosomal membrane proteins that are glycosylated (Appelqvist et al., 2013) and

have functions as diverse as trafficking of substrates, distribution of cargo, and exocytosis. Niemann-Pick disease Type C1 (NP-C1) is an example where a non-specific defect in NPC1, a membrane-bound protein involved in cholesterol transport, results in aggregates of cholesterol and other sphingolipids (Li et al., 2015; Park et al., 2003).

1.2.1 Clinical phenotypes

Despite the family of LSDs being relatively diverse with accumulation of glycogen, mucopolysaccharides or lipofuscins due to a defective enzyme or membrane protein, there is a classical description of LSDs as a family of paediatric neurodegenerative diseases (Wraith, 2002). Few LSDs lack some form of CNS pathology, as storage defects of metabolites affect various brain regions vulnerable to storage defects (Vitner et al., 2010). Furthermore, dysregulation of these storage pathways can lead to neuro-inflammation and activation of astrocytes from oxidative stress, altered calcium homeostasis and trafficking defects in the brain, capitulating LSDs as a neurodegenerative disorder. Notably, there is some overlap among the pathways defective in multiple LSDs such as conserved pathways that regulate lysosomal exocytosis, SL storage and available acidic calcium (Platt et al., 2012, 2014).

1.3 Niemann-Pick Disease

Niemann-Pick is a lysosomal storage disorder grouped into three main types: NP-A, NP-B, NP-C and NP-D (Crocker, 1961). The cause of dysfunction differs between the major types of Niemann-Pick (*i.e.*, mutations in separate genes resulting in different downstream effects). However, two phenotypes are shared between all four types: toxic intracellular accumulation of lipids and disrupted sphingomyelin homeostasis. NP-A disease is caused by a mutation in a sphingomyelinase enzyme responsible for ceramide biosynthesis, while NP-B disease is caused by a mutation in the acid sphingomyelinase gene. NP-C disease is caused by mutations in the lysosomal membrane NPC1 protein or the soluble NPC2 in the lysosomal lumen. NP-D disease is caused from a specific transversion in the NPC1 gene causing NP-C

disease (Greer et al., 1998). As the focus of this thesis is NP-C disease, additional background is now provided for NP-C.

1.3.1 Niemann-Pick Disease Type C

The defective NPC1 or NPC2 protein confers 95% and 5% of NP-C cases, respectively. There are more than 300 mutations in these genes to date (Li et al., 2016; Millat et al., 1999), of which any will result in disrupted trafficking of cholesterol and sphingolipids leading to pathological accumulation of these lipids in the lysosome (Platt et al., 2012; Vanier, 2014), although there is no genotype-phenotype correlation for these mutations (Runz et al., 2008). This is largely due to the interacting roles of NPC1 and NPC2 whereby NPC2 binds cholesterol in the lumen and effluxes it out of the lysosome by "handing it off" to NPC1 in the membrane (Kwon et al., 2009).

The cholesterol accumulation is better characterized than the sphingolipid accumulation. During normal cholesterol metabolism, cholesterol is produced in the endoplasmic reticulum (ER) when the cell recognizes cholesterol levels are low. One of the destinations for cholesterol is the late endosome/lysosome where it is degraded as well as recycled back to the ER. Cholesterol transport from the late endosome/lysosome to the ER relies upon the action of the NPC2-NPC1 hand-off of cholesterol, which results in the down-regulation of NPC1 expression via negative feedback from the sterol regulatory element binding protein pathway (SREBP) (Garver et al., 2008). In NP-C disease, failure of the cholesterol transport (facilitated by the NP-C1 protein) to reach the sterol reserves of the cell results in the failure to inhibit the SREBP pathway (causing a 'false impression' that cellular levels are low), leading to pathological accumulation of macromolecules within the cell.

Sphingosine, sphingomyelin, lactosylceramide, glucosylceramide, and gangliosides (GM1, GM2 and GM3) are all sphingolipids known to accumulate in NPC disease; however the pathways underlying the accumulation are not fully understood. It is well characterized that

sphingolipids have multiple roles within the cell and that their accumulation alters cell function (Kacher and Futerman, 2006). For example, glucosylceramide accumulation in GD results in the accumulation of additional sphingolipids such as ceramide, a major regulator of transport and signalling in eukaryotic cells (Hein et al., 2007). Interestingly, it is currently debated whether cholesterol or sphingolipids are the primary offending metabolite causing NP-C disease (Lloyd-Evans and Platt, 2010).

The resulting lysosomal dysfunction and lipid accumulation have cascading effects on other cellular pathways, potentially including innate immune system activation (Csepeggi et al., 2011; Platt et al., 2012). This activation arises from the detrimental effects of cellular aggregation, such as increased reactive oxygen species (ROS) and inflammatory or apoptotic signals arising from the macromolecule build up (Kacher and Futerman, 2006). Furthermore, altered calcium homeostasis has been implicated in classical neurological diseases such as Alzheimer's disease as well as LSDs (Bodennec et al., 2002; Mattson and Chan, 2003), although the mechanism of altered homeostasis is distinct across diseases. Finally, pro-inflammatory mediators have been shown to be elevated in LSDs, although the link between sphingolipids and pro-inflammatory mediators is unclear (Kacher and Futerman, 2006).

1.3.2 Neurovisceral aspects of NP-C

The primary organs affected in NP-C disease are the liver and brain, the organs most sensitive to changes in cholesterol homeostasis. Cholesterol is made in the liver, while the brain contains a separate pool of cholesterol that represents 20-30% of all lipids in the brain (Pfrieger, 2003). Moreover, initial onset of symptoms occurs across a wide range of ages, such that the disease is best viewed as different clinical forms categorised by age (Table *1.1*, Vanier, 2010). The neonatal (or perinatal) form of disease is characterized by prolonged jaundice and/or hepatosplenomegaly that precedes neurological symptoms. Early infantile (roughly two months to two years old) involves missing developmental marks or reversion of learnt motor

skills. Late infantile (three to six years) onset consists of gait difficulties and usually symptoms include VGSP, although at an early stage. The juvenile category of six to fifteen years is also regarded as the 'classical' presentation of NP-C, for this is the most common form of the disease across most countries. VGSP is often an initial symptom, and difficulties in learning along with cataplexy (often caused by laughter) are also common symptoms. Finally, the adult presentation of NP-C is heterogeneous in presentation, however psychiatric symptoms followed by cognitive and motor decline is a common set of symptoms.

Table 1.1 – **Major types of NP-C disease as defined by Vanier, 2010.**

Neonatal	Early infantile	Late infantile	Juvenile	Adolescent/Adult
Foetal ascites, icterus, cholestasis, hepatosplenomegaly, pulmonary alveolar lipoproteinosis	Delay in motor development, hypotonia	Ambulation issues, cataplexy, speech delays, VGSP	Seizures, ataxia, cataplexy, VGSP	Psychiatric problems, ataxia, dementia

NP-C disease, like most members of the family of LSDs, has neurological components that all patients will develop over the course of disease progression. However, NP-C also includes systemic visceral involvement of liver, spleen and pulmonary tissues (Patterson et al., 2012; Vanier, 2010). This systemic component of disease occurs prior to neurologic symptoms, and is considered in classifying disease forms categorized by age of onset of neurological involvement (Wraith and Imrie, 2009). In brief, 85% of NP-C patients experience a systemic component of disease prior to development of neurological symptoms (Vanier, 2010). The onset of neurological symptoms consist of impaired motor functions, gait issues, cataplexy, ataxia, vertical supranuclear gaze palsy (VGSP) (Solomon et al., 2005) and eventually psychiatric disturbances. Overall, it has been reported that the age of onset of neurological

symptoms is correlated with progression of the aforementioned neurological forms of the disease, and acts as a general predictor of disease progression and patient lifespan (Wraith and Imrie, 2009).

1.3.3 Pulmonary consequences of NP-C

Pursuant to the gastrointestinal and neurodegenerative symptoms, pulmonary consequences have also been reported (Gülhan et al., 2012; Sheth et al., 2017). In fact, these pulmonary consequences are also critical as a common cause of death for NP-C patients is respiratory failure (*e.g.*, pneumonia). Foam cell accumulation in the lungs has been reported in the feline model of NP-C disease. (Lowenthal et al., 1990) Furthermore, this pulmonary feature of NP-C felines is observed in both NPC1-deficient and NPC2-deficient mice (Roszell et al., 2013). Both NP-C1 and NP-C2 are clinically heterogeneous with pulmonary involvement, with the potential for recurrent pulmonary infection. This is hypothesised to be due to impaired macrophage NP-C2 protein expression resulting in inert surfactant within the lung (Griese et al., 2010; Sheth et al., 2017).

1.3.4 Laboratory diagnosis of NP-C using the filipin test

Given symptoms such as jaundice, splenomegaly and ataxia that are presented in many diseases, NP-C disease is undoubtedly challenging to diagnose. In addition, the clinical presentation of NP-C disease is highly heterogeneous regarding presentation of symptoms as well as the age of when the patient is first diagnosed with NP-C is highly variable (Vanier, 2010). As such, diagnosis must be conducted in the clinic as well as the laboratory.

A robust laboratory test for the archetypical NP-C phenotype is direct evaluation of the lysosomal accumulation of unesterified cholesterol in NP-C patient fibroblasts. This is accomplished via the filipin stain test (Vanier and Latour, 2015a). Filipin is a polyene antifungal secreted by *Streptomyces filipinensis* that binds to the ergosterol found in fungal cell walls. Because filipin is selective for unesterified cholesterol and not sterol esters, it can be

used to visualize accumulation of unesterified cholesterol within cells using fluorescent microscopy. This test accurately demonstrates impaired cholesterol transport along the endocytic pathway of the cell, and is considered the gold-standard diagnostic tool for diagnosing a suspected NP-C patient (Vanier, 2010). If the filipin test is found to be positive, targeted sequencing of the NPC1 and NPC2 genes can occur to fully confirm the diagnosis (*i.e.*, a mutation in NPC1 or NPC2).

1.3.5 Clinical diagnosis of NP-C using severity scoring

Clinical diagnosis of NP-C makes use of a disease severity scoring scale (Iturriaga et al., 2006; Yanjanin et al., 2010). This scale considers an evaluation of the patient's motor skills, strength of muscle and movement, swallowing testing and eye movement. Eye movement is examined in particular because saccadic eye movements are an early neurological symptom in NP-C (Abel et al., 2012). As neurological symptoms are progressive and predictive of disease outcomes, it is possible to generate a pathological model for NP-C disease (Yanjanin et al., 2010). This clinical severity scale examines eye movement, speech, motor skills, ambulation, swallowing, cognition, hearing, memory, seizure history on a 0-5 point scale and additive modifiers to the model (*e.g.*, hyperreflexia, cataplexy, psychiatric). This scale has proven to be a valuable tool for increasing diagnostic capability and monitoring disease progression (Shin et al., 2011; Vanier, 2010). Relative to this thesis, this scale also serves as a reminder that the clinical presentation of NP-C is heterogeneous and varied.

1.3.6 Potential treatments for NP-C

1.3.6.1 Miglustat

One perspective on treating lysosomal glycolipid accumulation is via partial inhibition of glycolipid biosynthesis pathways. Miglustat, originally used to treat GD, is a compound that inhibits glucosylceramide synthase and reduces the catabolic burden on the cell in a LSD context (Platt et al., 1994). An important feature of miglustat treatment is the capability to

cross-over the blood-brain barrier (Patterson et al., 2007), unlike direct lysosomal enzyme replacement therapy (Platt et al., 2012). However, miglustat is not therapeutic in the liver. Miglustat is approved to treat NP-C disease in many but not all countries (*e.g.*, it was not approved by the FDA in the USA).

1.3.6.2 Cyclodextrin

Cyclodextrins are oligosaccharides with structure and chemistry that complexes with many compounds including cholesterol (Del Valle, 2004), hence cyclodextrins are commonly used as vehicles for other drugs in treatment trials. This was actually a factor in the development of cyclodextrin as a promising therapy to treat NP-C disease, whereby the cyclodextrin vehicle was equally therapeutic in *Npc1*$^{-/-}$ mice compared the cyclodextrin vehicle with the candidate allopregnanolone therapy (Davidson et al., 2009) that was previously identified (Griffin et al., 2004). Cyclodextrin has repeatedly been shown to increase the lifespan and delay the onset of NP-C disease in the *Npc1*$^{-/-}$ mouse model as well as the feline model (Liu et al., 2009; Vite et al., 2015). A fully enrolled clinical trial (Phase 2/3) in humans is ongoing (ClinicalTrials.gov identifier NCT02534844) with 2-hydroxypropyl-β-cyclodextrin (HPBCD). However, one drawback to this approach is the complex structure of cyclodextrin family prevents crossing of the blood-brain barrier, thus requiring sedation and intrathecal (spinal tap) administration of the drug. Sucampo Pharmaceuticals, Inc. recently acquired Vtesse, Inc., the company responsible for advancing this clinical trial, for $200 million (http://www.cydanco.com/). Considering the prevalence of NP-C worldwide, this acquisition implies Sucampo perceives value in this cyclodextrin vehicle within a broader scope, especially considering the shared mechanisms NP-C has with other diseases (Platt et al., 2014).

1.3.6.3 Arimoclomol

Arimoclomol is a hydroxylamine drug that induces a heat shock protein response (Pratt et al., 2012) by interacting with heat shock protein 70 (Hsp70). This activity alleviates NP-C

disease symptoms by the stabilization of lysosomes via the interaction between Hsp70 and bis(monoacylglycero)phosphate (BMP) (Kirkegaard et al., 2010). Hence, defective lysosomal stability in NP disease is theorised to be corrected via this mode of action. Clinical trials (Phase 2) involving arimoclomol and human NP-C patients are currently underway (ClinicalTrials.gov identifier NCT02612129). Orphazyme is the pharmaceutical company heading this drug trial (http://nnpdf.org/research/clinical-trials/orphazyme-clinical-trial/, https://www.orphazyme.com/clinical-trials-aidnpc).

1.3.6.4 Vorinostat

Vorinostat is a histone deacetylase inhibitor that is already approved to treat cutaneous T-cell lymphoma, and has completed a clinical trial (Phase 1/2, ClinicalTrials.gov identifier NCT02124083) after being initially identified as a candidate therapy in the yeast model of NP-C disease, a result that was further confirmed when vorinostat reduced filipin accumulation in NP-C patient fibroblasts (Munkacsi et al., 2011; Pipalia et al 2011).) The rationale behind this treatment arises from the reported upregulation of NPC1 in NP-C patient fibroblasts (Pipalia et al., 2011, 2017), and the concept that epigenetic consequences of reduced histone deacetylation would partially rescue the NP-C phenotype (Munkacsi et al., 2016). Promising future developments will likely involve combined treatments to increase efficacy, such as the triple combination of vorinostat complexed with HPBCD in polyethylene glycol as demonstrated by Alam et al., 2016.

1.4 Exome sequencing as a tool for genetic modifier discovery

Whole genome sequencing of entire patient cohorts remains, at least for the short future, cost-prohibitive for identifying disease-causing genes and disease-modifying genes (Park and Kim, 2016). An alternative approach is to consider an enriched-subset of the human genome that is likely to impact the disease of interest, or allow identification of potentially relevant regions that can then be selectively sequenced. The human exome is a proven subset of the

human genome with specific relevance to disease (Bamshad et al., 2010, 2011). While it is true that the exome covers only a tiny fraction of the whole genome, roughly 2% (Bamshad et al., 2011), this 2% corresponds to the protein-coding sequences of the human genome and as such includes 85% of DNA sequence variants known to cause human diseases. For example, it is well established that DNA sequence variants in the exome alter the structure and function of proteins underlying Mendelian diseases (Bamshad et al., 2010, 2011; Gilissen et al., 2011; McInerney-Leo et al., 2013).

A simple definition of a disease-causing gene is a mutation that results in the presentation of the disease (Genin et al., 2008). In contrast, a genetic modifier of a disease is a mutation outside of the disease-causing gene that contributes to a disease phenotype but does not on its own confer the disease diagnosis (Cutting, 2010). The discovery of genetic modifiers to disease by exome sequencing is an extension to the methods described above (Cirulli et al., 2010; Cooper and Shendure, 2011; Alazami et al., 2015; Esslinger et al., 2017). In the case of NP-C disease, a genetic modifier would be a mutation outside of the NPC1 or NPC2 disease-causing genes.

Therefore, discovering a genetic modifier to a disease requires a more nuanced approach than the usual comparison of a healthy population (unaffected) and a patient cohort (affected) routinely used to identify a disease-causing gene. In contrast, the identification of genetic modifiers requires that the candidate modifier is associated with clinical severity or variability. Moreover, a genetic modifier of a disease does not necessarily have to be unique from the healthy population (*i.e.*, there is no longer an obvious segregation between the affected and the unaffected populations). This applies even in the cases of ‘simple’ monogenic diseases where there may be more than one modifier to said disease that interact in a multi-factorial fashion. Furthermore, variations within the genomic structure itself can contribute to inherited

disease and their clinical presentation (Beckmann et al., 2007), such as copy number variation (CNV).

1.4.1 Identifying causal variants using sibling pairs

Identification of variants that are causal agents of rare monogenic disorders is traditionally difficult due to factors such as low sample size and heterogeneous clinical presentations that result in decreased statistical power for older, more traditional genetic screens such as positional cloning. An effective compromise arises in the study of whole-exome sequencing of rare patient cohorts, precisely because the exome acts as an enriched subset of the genome when it comes to predicting effects on protein function. As a markedly simplified example, examining the whole-exome sequences of a sibling pair with a disease and comparing it to a database of healthy variants found within the healthy population enables the filtration of a candidate list of disease-causing genes down to potential causal variants (Bamshad et al., 2011; Cooper and Shendure, 2011).

This approach can be improved upon further by considering familial sets, and especially sibling pairs. Sibling pairs are a useful approach to study design with regards to sequence data, as they can improve statistical power to detect rare variants in an association test. This arises from the fact that disease-associated variants are likely to segregate within familial pairs (Zhu et al., 2010), with affected siblings sharing the causal variant. Such an approach requires that your affected/unaffected or discordant sibling pairs are sequenced with the same targeted genomic region, with collective testing of rare variants (Feng et al., 2011).

1.4.2 Sibling-pair analyses to identify genetic modifiers

One must carefully consider their approach when analysing whole exome sequence data for putative disease modifiers. An initial constraint that aids inquiry is to consider the familial or pedigree history of a patient cohort, where possible. This enables, with appropriate clinical co-variates, to segregate modifiers along your trait(s) in question. Without familial data, one

can identify putative modifiers (such as those reported in mevalonate kinase deficiency) (Marcuzzi et al., 2016), however statistical power and low sample numbers will prevent a clear correlation with disease phenotype. For example, the loci modifying the onset of cancer in patients with Gaucher disease was examined in sibling pairs (Lo et al., 2012). Whole exomes were sequenced from two affected siblings, sequences were compared against control databases, and homozygosity mapping was used to identify a novel mutation associated with increased cancer risk. These examples, and others (Luzon-Toro et al., 2015; Reddy et al., 2017; Schil et al., 2016; Tucci et al., 2014), demonstrate the suitability and potential for whole exome sequence analysis in the discovery of genetic modifiers to disease.

The clinical phenotype and study design must be clearly defined in order to allow valid assumptions to facilitate the discovery of genetic modifiers. One can quantitate the onset of the clinical phenotype by metrics such as age at diagnosis, or quantitate the severity of disease progression. The population to be studied must also be clearly defined – a good first choice is fraternal twins (identical twins would not be useful to identify modifiers if they have divergent severity, for obvious reasons) or sibling-based pairs, as this enables the comparison of the clinical phenotype chosen above with related patients with minimal environmental variation (*i.e.*, siblings are raised in the same house) and the unrelated control population. This also facilitates comparisons of variation between and within familial pairs (*e.g.*, determining the inheritance of the genetic modifier).

1.4.3 Intricacies to identifying genetic modifiers

An immediate distinction to make clear when searching for associations of rare genetic modifiers to disease severity is the expectation that disease severity will not be explained by one statistically significant genetic modifier. This is simply because within phenotypically-discordant sibling pairs, there will be a single modifier variant in one and not the other. Instead, one must examine the variants not on a *per variant* basis, but a *per gene* basis. To elucidate

this, even if each case sample has a different rare variant present from one another that is not present within their control, but this different variant is within the same gene – that can form an argument for significant association. This argument follows from the idea that even different variants within the same gene or genetic pathway can result in a similar functional change, resulting in the same phenotype (or disease modification).

Grouping variants together is also a means to increase statistical power. For example,Morris and Zeggini (Morris and Zeggini, 2010) coined a "functional unit" where the functional unit is the gene or pathway in question. With this notion, it is possible to discover genetic modifiers of disease severity in a small-sample cohort. For example, variants in APOE*E2 were associated with a delayed age of onset in PSEN1 Alzheimer's disease (Vélez et al., 2016).

1.5 Aims and hypotheses

The extensive variation in the onset and progression of NPC disease implies the presence of disease-modifying variants. While NP-C disease has a monogenic cause, there is still much to be elucidated regarding its broad clinical spectrum and the variation in quality of life among patients, including within sibling-pairs that share the same disease-causing NPC mutation. The overall hypothesis of this thesis is that there are genetic modifiers of NP-C disease severity. Specifically, the thesis aims are the following:

1. To sequence the exomes of five/six sibling pairs affected with NP-C disease.
2. To identify genetic modifiers common to all NP-C patients.
3. To identify genetic modifiers of disease severity that are unique within sibling pairs and segregating with disease severity.

2 EXOME SEQUENCE ANALYSIS OF SIBLING PAIRS WITH NP-C

2.1 Introduction

NP-C disease is an autosomal recessive, monogenic disease with an estimated prevalence of roughly 1:104,000 live births (Jahnova et al., 2014; Wassif et al., 2016). NP-C disease is caused by mutations in the NPC1 (95% of the cases) or NPC2 (5% of the cases) gene (Vanier, 2010). However, the actual population prevalence is probably higher, due to the lack of clinical awareness, difficulty in testing for the condition and heterogeneous presentation of the disorder (Wassif et al., 2016). There are four major types of NP-C disease based on the age of onset (define four types), with the classical presentation being the juvenile category of 6 to 15 years, development of gastrointestinal problems followed by dementia and ataxia, and loss of life before adolescence (Vanier 2010). This heterogeneity of NP-C disease is such that even siblings with the same causative mutation can have varied ages of onset and severity of symptoms (Imrie et al., 2007; Iturriaga et al., 2006; Patterson et al., 2012; Vanier, 2010). It is still not well-described how individuals with the same causal monogenic mutation can have such different clinical outcomes.

The onset and progression of any disease, NP-C disease included, is a combination of genetic and environmental factors. Environmental factors can include drug treatments or a specific diet to ameliorate disease, while genetic factors can be the disease-causing gene and any genetic modifiers extrinsic to the disease-causing genes. A genetic modifier of a disease is a mutation outside of the disease-causing gene that contributes to a disease phenotype but does not on its own confer the disease (Genin et al., 2008). As there are siblings affected with NP-C disease exhibiting divergent disease severity and these siblings were raised in the same households with similar diets and attempted treatments were either consistently applied to both siblings or none at all, we hypothesize that NP-C disease severity is a consequence of genetic

modifiers. Genetic modifiers have previously been identified in model systems to study NP-C disease such as the EAF1 and YAF9 histone acetylase genes in the yeast model (Munkacsi et al., 2011), the transmembrane protein TMEM97 in cholesterol-manipulated HeLa cells (Bartz et al., 2009), the small heat shock protein HSPB1 in NP-C mice Purkinje cells (Chung et al., 2016) and in human cohorts, with ApoE polymorphisms in fifteen NP-C patient cell lines (Fu et al., 2012). A systematic and unbiased genomic screen for genetic modifiers of NP-C has not yet been conducted directly in human patients.

Sequencing of only the protein-coding portion of the genome (the exome) has previously been used to identify modifiers of disease, such as in cystic fibrosis, Alzheimer's disease and Limb-Girdle muscular dystrophy (Emond et al., 2015; Fu et al., 2012; Reddy et al., 2017). When combined with kindred pairs, it is possible to detect variants that segregate with the clinical phenotype in question, given appropriate scoring and statistical measures (Amin et al., 2017; Hu et al., 2017). A genetic modifier does not have to be removed from the unaffected or control population, for the variant could be modifying the disease-causing gene/pathway. To further strengthen the power of your sib-pair analysis, functional groups of variants belonging to a 'unit' (gene, pathway *etc.*) with individual variants assigned weights are then tested for associations between phenotype and these variants per group (Morris and Zeggini, 2010; Zhu et al., 2010).

In this chapter, to identify putative genetic modifiers of NP-C disease severity we performed whole exome sequencing of an NP-C patient cohort (n = 10) consisting of five sibling pairs exhibiting divergent disease severity. The raw sequence reads were processed, variants were then annotated and filtered to determine those that segregated within sibling pairs and throughout the NP-C cohort.

2.2 Materials and Methods

2.2.1 Patients

NP-C patients with verified NPC1 gene mutations and positive filipin stain tests were sequenced with consent according to the Institutional Review Board of Columbia University Medical Center, in collaboration with Dr. Stephen Sturley. Patients had one other sibling in the cohort with the additional feature that every sibling-pair showed divergent disease severity. The clinical characteristics were provided by physicians as part of the study.

2.2.2 Exome sequencing

Exome sequences were selectively sequenced from genomic DNA (Figure 2.1). First, genomic DNA was extracted from each NP-C patient (peripheral lymphocytes or immortalized fibroblasts), sheared via sonication into fragments of random lengths and adaptors were linked to the fragments to form a library of potentially overlapping reads as described previously for a paired-end cluster generation kit (PE-401-1001) on an Illumina HiSeq2000. To specifically target exonic regions, biotin is hybridized to these fragments within a sea of complimentary sequences to the adaptors. Following a wash, the exonic fragments are recovered via a biotin-based pulldown, whereupon standard amplification and massively parallel sequencing follows. The creation of a library, enriched for protein-coding regions, enables the mapping and calling of potential genetic modifiers to disease.

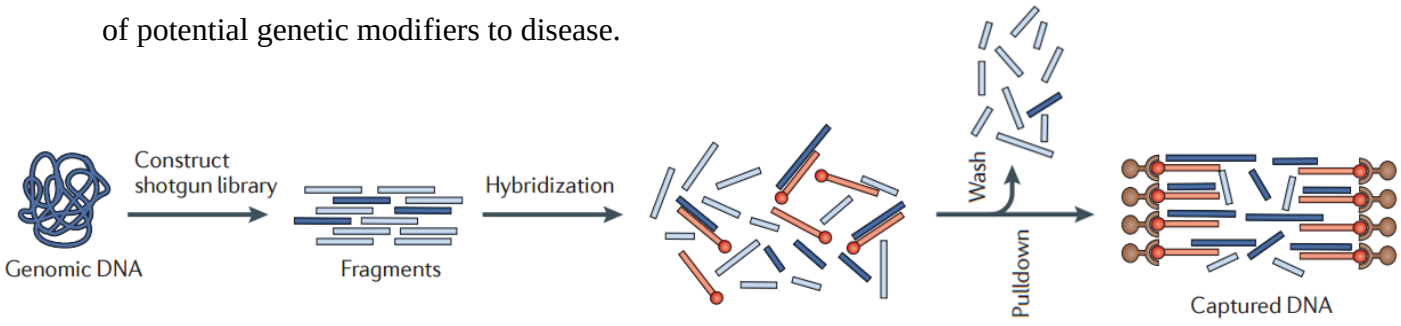


Figure 2.1 – **Exome sequencing pipeline.** Exons are shown in dark blue, biotinylated DNA baits are shown in orange. Figure adapted from Bamshad et al.

2.2.2.1 Read mapping and variant calling overview

Interleaved raw sequence reads in FASTQ format (Cock et al., 2010) were mapped to the GRCh38 Genome Reference Consortium Human Reference 38 (GCA_000001405.2) assembly of the human genome (https://www.ncbi.nlm.nih.gov/genome/?term=txid9606[orgn]) using the BWA-MEM alignment algorithm (Li and Durbin, 2010). Sequence calls were filtered with ≥10x coverage and integrated parameters as described below. Sequence calls were compared to HapMap and other resources as described by the Broad Institute (McKenna et al., 2010). Annotation of variants was performed using SnpEff based on NCBI and UCSC databases, such as dbSNP.

2.2.2.2 Trimming adapter sequences

Adapter sequences were removed from FASTQ sequence data using Trimmomatic (Bolger et al., 2014) based on Phred scores (Illumina) in the paired-end mode to conserve read pairs and find PCR primer fragments. Adapter sequences were trimmed using the following small script:

```
#!/bin/bash
IF=PATH TO INPUT FORWARD FILE (R1)
IR=PATH TO INPUT REVERSE FILE (R2)
OFP=PATH TO OUTPUT FORWARD PAIRED FILE
OFU=PATH TO OUTPUT FORWARD UNPAIRED FILE
ORP=PATH TO OUTPUT REVERSE PAIRED FILE
ORU=PATH TO OUTPUT REVERSE UNPAIRED FILE
java -jar trimmomatic-0.36.jar PE -phred33 $IF $IR $OFP $OFU $ORP $ORU
ILLUMINACLIP:EXOME_PIPE/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Script 2.1 – **Trimmomatic removal of adapter sequences.**

2.2.2.3 Interleaving

Forward and reverse reads within the same FASTQ file were interleaved together via the following Python script:

```
#!/usr/bin/env python
# encoding: utf-8

import sys
import argparse

def interface():
    parser = argparse.ArgumentParser()

    parser.add_argument('--rm-short-reads',
                        type=int,
                        help='Minimum number of base pairs \
                        either R1 or R2 read must be.')


    parser.add_argument('LEFT_INPUT',
                         type=argparse.FileType('r'),
                         default=sys.stdin,
                         nargs='?',
                         help='R1 reads.')

    parser.add_argument('RIGHT_INPUT',
                         type=argparse.FileType('r'),
                         default=sys.stdin,
                         nargs='?',
                         help='R2 reads.')

    parser.add_argument('INTERLEAVED_OUTPUT',
                         type=argparse.FileType('w'),
                         default=sys.stdout,
                         nargs='?',
                         help='Alignment file.')

    args = parser.parse_args()
    return args


def process_reads(args):

    left = args.LEFT_INPUT
    right = args.RIGHT_INPUT
    fout = args.INTERLEAVED_OUTPUT

    while 1:

        # process the first file
        left_id = left.readline()
        if not left_id: break
        left_seq = left.readline()
        left_plus = left.readline()
        left_quals = left.readline()

        # process the second file
        right_id = right.readline()
        right_seq = right.readline()
        right_plus = right.readline()
        right_quals = right.readline()
```

Script 2.2 – **Interleaving paired-end reads.** Takes two FASQ files as input and outputs a single, interleaved FASTQ file. Script adapted from https://gist.github.com/ngcrawford.

As each sample was run on separate lanes, an additional step was required to handle multiple libraries separately. This was accomplished using Script 2.2, taking the form “python interleave_fastq.py FWD.fastq REV.fastq OUT.fastq”:

```
#!/bin/bash
CURRENT_FILE=1
WRK_DIR=/media/chemgen/Elements1/DMF/L1
OUT_DIR=/media/chemgen/Elements1/DMF
pstr="[=============================================================================]"

cd $WRK_DIR

# count total number of files in dir
t=$(ls -1 | wc -l)

# divide by two for true file count of fwd and rev reads
tc=$(($t/2))

while [ $CURRENT_FILE -le 9 ]; do
      z=$(echo *_00*$CURRENT_FILE*) # both R1 and R2 reads as arguments
here
      python /media/chemgen/Elements1/scripts/interleave_fastq.py $z
$OUT_DIR/DMF_L1_CAGATC_Lib$CURRENT_FILE.fastq # interleave_fastq.py R1 R2
OUTPUT
      CURRENT_FILE=$((CURRENT_FILE=CURRENT_FILE+1))
      count=$(($CURRENT_FILE-1))
      pd=$(( $count * 73 / $tc ))
      printf "\r%3d.%1d%% %.${pd}s" $(( $count * 100 / $tc )) $(( ($count *
1000 / $tc) % 10 )) $pstr
      if [ "$CURRENT_FILE" -eq "$tc" ]; then
            break
      fi
done

while [ $CURRENT_FILE -le $tc ]; do
      z=$(echo *_0*$CURRENT_FILE*) # both R1 and R2 reads as arguments here
      python /media/chemgen/Elements1/scripts/interleave_fastq.py $z
$OUT_DIR/DMF_L1_CAGATC_Lib$CURRENT_FILE.fastq # interleave_fastq.py R1 R2
OUTPUT
      CURRENT_FILE=$((CURRENT_FILE=CURRENT_FILE+1))
      count=$(($CURRENT_FILE-1))
      pd=$(( $count * 73 / $tc ))
      printf "\r%3d.%1d%% %.${pd}s" $(( $count * 100 / $tc )) $(( ($count *
1000 / $tc) % 10 )) $pstr
done
```

Script 2.3 – **Interleaving individual lanes.** This example demonstrates taking the paired-end reads from lane 1 of a single sample, DMF, and interleaving them. This process is repeated for each lane (L1, L2, …, LN).

2.2.2.4 Mapping

Mapping was performed via the Burrows-Wheeler Alignment Tool using the BWA-MEM alignment algorithm (Li and Durbin, 2010) run in paired-end mode for Picard compatibility:

```
WRK_DIR=$GLOBAL_SCRATCH/AAA.TRIM
cd $WRK_DIR
module load bwa-kit
for f in *.fastq; do
STEM=$(basename "${f}" .fastq);
bwa mem -Mp $GLOBAL_SCRATCH/hg38.fa $f > ${STEM}.sam
done
```

Script 2.4 – **Mapping read data to GRCh38.** This script converted the reads from FASTQ format to SAM format. The 'mem' argument specifies that alignments will be initiated via maximal exact matches, and then extended according to the Smith-Waterman method (Smith and Waterman, 1981). The '-M' parameter (short for Mark) is used to ensure Picard compatibility further on in the pipeline, as Picard requires shorter split hits to be specifically marked as such. As our input FASTQ files have been interleaved (see Script 2.2), we know that the 2n and 2n+1 reads form a read pair and can tell BWA to operate in paired end mode via the '-p' parameter.

2.2.2.5 Sorting and realignment

Sequence alignment/map (SAM) files were sorted by coordinate and converted to binary alignment/map (BAM) format (Li et al., 2009a). Realignment around regions with insertions and deletions was not performed, as downstream variant calling was performed via HaplotypeCaller (McKenna et al., 2010), which works through *de novo* assembly, rendering

this realignment requirement of older, locus-based variant callers obsolete.

```
module load picard-tools
PICARD=/srv/global/scratch/groups/sbs/picard-tools-2.1.0/picard.jar
for dir in $GLOBAL_SCRATCH/AAA.TRIM/*sam;do # ALL .sam FILES IN DIR
RECURSIVELY
name=${dir%.*}

java -jar $PICARD SortSam \
      I=$dir \
      O=$name.bam \
      SORT_ORDER=coordinate
java -jar $PICARD BuildBamIndex \
      I=$name.bam
done
```

Script 2.5 – **Sorting SAM files via coordinate.** We build a BAM file with an index (.bai) for use later in the pipeline. A .bam file is the aligned sequence data, whereas the .bai file is the index file which allows quick navigation of a .bam file.

2.2.2.6 Add read group headers

Read groups (unique identifiers for a collection of reads) were defined in the header of BAM files in order to be processed within the GATK pipeline (McKenna et al., 2010):

```
module load picard-tools
PICARD=/srv/global/scratch/groups/sbs/picard-tools-2.1.0/picard.jar
shopt -s globstar
for dir in $GLOBAL_SCRATCH/AAA.TRIM/14/MERGE/*bam;do # dir == full path to
ALL files in AAA.TRIM
# e.g.
/srv/global/scratch/carsweshau/AAA.TRIM/9/L8.ACTGAT/Andy9_L8_ACTGAT_Lib9.ba
m
[[ -d $dir ]] && continue # if directory then skip
if [ ${dir: -4} == ".bam" ]; then
name=${dir%.*} # removes extension
filename=${dir##*/} # removes path prefix
filenoext=${name##*/}
sep=$(echo $filenoext | tr "_" "\n")
infor=( $sep ) # ${infor[2]} corresponds to ACTGAT in the above example
sample=${infor[0]}
unit=${infor[1]}
library=${infor[3]}

java -jar $PICARD AddOrReplaceReadGroups INPUT=${dir} OUTPUT=${name}_RG.bam
RGID=$filenoext RGLB=$library RGPL=illumina RGPU=$unit RGSM=$sample
fi
done
```

Script 2.6 – **Add or replace read groups.**

2.2.2.7 Marking duplicates

Duplicate reads that arose from the sequencing process per read group were marked, removing these molecular duplicates from downstream analysis. This step was also used to merge multiple BAM files per sample into a single BAM file per sample, with each collection of reads from a single library run per lane defined as a read group. This step results in our multiple libraries being correctly merged into a single BAM file per sample, while preserving pertinent information like lane number and read group.

```
module load picard-tools
PICARD=/srv/global/scratch/groups/sbs/picard-tools-2.1.0/picard.jar
shopt -s globstar
dir=$(echo "$GLOBAL_SCRATCH/AAA.TRIM/**")
name=${dir%.*}
java -jar $PICARD MarkDuplicates \
      I="$dir" \
      O=${name}_dedup.bam \
      METRICS_FILE=metrics.txt
```

Script 2.7 – **Mark molecular duplicates.**

2.2.2.8 Base quality recalibration

The final step in pre-processing was completed by recalibrating the individual base quality score assigned during each sequence read. This base quality score recalibration (BQSR) adjusted the base error reported by the sequencing platform. This was based on cycle quality, dinucleotide quality and global differences between the empirical and the reported score as well as bin specific shift quality. This recalibration resulted in reported read quality scores that better matched empirical quality, reduced error by dinucleotide and machine

cycling, and improved the accuracy of downstream variant calls (DePristo et al., 2011).

```
module load gatk
GATK=/srv/global/scratch/groups/sbs/GATK/3.5.0/GenomeAnalysisTK.jar
ANDY_COUNT=1
while [ $ANDY_COUNT -le 10 ]; do
for f in $GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/*dedup.bam; do
java -jar $GATK \
   -T BaseRecalibrator \
   -R $GLOBAL_SCRATCH/hg38.fa \
   -I $f \
   -knownSites $GLOBAL_SCRATCH/dbsnp_148_sorted.vcf \
   -o
$GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/Andy${ANDY_COUNT}_recal_data.table
java -jar $GATK \
   -T PrintReads \
   -R $GLOBAL_SCRATCH/hg38.fa \
   -I $f \
   -BQSR
$GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/Andy${ANDY_COUNT}_recal_data.table \
   -o
$GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/ANDY_${ANDY_COUNT}.merged.dedup.recal.
bam
done
ANDY_COUNT=$((ANDY_COUNT=ANDY_COUNT+1))
done
```

Script 2.8 – **Base recalibration.** Known sites SNP database obtained from ftp://ftp.ncbi.nih.gov/snp/.

2.2.2.9 Generating callable loci

As sequencing of our subjects was completed in 2011, the target BED regions (the targeted capture co-ordinates) were not provided for this HiSeq2000 exome project. As such, the CallableLoci module was used to generate a target list of intervals empirically:

```
module load gatk
GATK=/srv/global/scratch/groups/sbs/GATK/3.5.0/GenomeAnalysisTK.jar
ANDY_COUNT=1
while [ $ANDY_COUNT -le 13 ]; do
for f in $GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/*.recal.bam; do
java -jar $GATK \
     -T CallableLoci \
     -R $GLOBAL_SCRATCH/hg38.fa \
     -I $f \
     -summary table.txt \
     -o $GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/call_sites.bed
done
ANDY_COUNT=$((ANDY_COUNT=ANDY_COUNT+1))
done
```

Script 2.9 – **Callable loci.** This module provides estimates for the coverage at each locus of interest, returning summary intervals tagged with certain states (PASS, NO_COVERAGE, LOW_COVERAGE, EXCESSIVE_COVERAGE, POOR_MAPPING_QUALITY).

The list of callable regions within the BED file format was then converted to an interval list via the following script:

```
java -jar $PICARD BedToIntervalList I=call_sites.bed
SD=$GLOBAL_SCRATCH/hg38.dict
O=$GLOBAL_SCRATCH/AAA.TRIM/call_sites.interval_list
```

Script 2.10 – **Bed to interval list.** A simple file format conversion to ensure compatibility with the GATK interval parameter. From this, we can determine the sequence intervals to operate over, even without the original target regions.

2.2.2.10 Re-ordering contigs

An important constraint when using the GATK is ensuring the BAM file contig order matches the coordinate order of the reference, along with any other database files. To ensure this, the following custom perl script was employed to sort a supplied VCF file by reference genome:

```
#!/usr/bin/perl
open(DICT,$dict_file) or die "Can't open $dict_file!\n";
my @contig_order;
my $c=0;
while(<DICT>)
{
if($_=~ /\@SQ/)
      {
      my ($contig) = $_ =~ /SN:(\S+)/;
      $contig_order[$c]=$contig;
      ++$c;
      #print $contig,"\n";
      }
}
close(DICT);
open(VCF,$vcf_file) or die "Can't open $vcf_file!\n";

my %vcf_hash;
my $header;

while(<VCF>)
{
if($_=~/^#/){ $header .= $_; }
chomp($_);

my @data = split(/\t/,$_);
my $contig = $data[0];
my $start = $data[1];
my $variant = $data[4]."to".$data[5];
my $line = $_;

$vcf_hash{$contig}{$start}{$variant}=$line;

}
close(VCF);
print $header;

foreach my $contig (@contig_order) # sort by contig order
      {
      #print $contig,"\n";
      foreach my $start (sort {$a <=> $b} keys
%{$vcf_hash{$contig}}) # sort numerically by coordinates
            {
            #print $start,"\n";
            foreach my $variant (keys
%{$vcf_hash{$contig}{$start}})
                  {
                  print
$vcf_hash{$contig}{$start}{$variant},"\n";
                  }
            }     }
```

Script 2.11 – **Sort VCF.**

Custom perl script for sorting VCFs, copyright German Gaston Leparc.

Operating over intervals

By default, the GATK operates over the genomic intervals defined by the reference contigs. For exome sequencing projects, iterating over the entire genome is redundant, so we specified the intervals to process via the '-L' parameter. This intervals parameter can accept a list of intervals defined within a file (see Script 2.10).

2.2.2.11 Genotyping

The processed BAM files from 2.2.2.8 were examined for sites that vary in relation to the reference genome. HaplotypeCaller (McKenna et al., 2010) was use to target regions likely to be sites of significant variation, determining the allelic likelihood for those sites. Then using probabilistic methods, a genotype was supplied for that sample (including an estimate for how homozygous the sites were to the supplied reference) and presented in a genomic variant call format file (gVCF):

```
module load gatk
module load samtools
GATK=/srv/global/scratch/groups/sbs/GATK/3.5.0/GenomeAnalysisTK.jar
ANDY_COUNT=1
while [ $ANDY_COUNT -le 13 ]; do
for f in $GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/*.recal.bam; do
SAMPLE=Andy${ANDY_COUNT}
if [ $ANDY_COUNT -eq 10 ]; then
      SAMPLE="DHK"
fi
if [ $ANDY_COUNT -eq 11 ]; then
      SAMPLE="SAK"
fi
if [ $ANDY_COUNT -eq 12 ]; then
      SAMPLE="TAF"
fi
if [ $ANDY_COUNT -eq 13 ]; then
      SAMPLE="TEF"
fi
#SAMPLE_NAME=$(samtools view -H $f | grep '@RG')
java -jar $GATK \
      -T HaplotypeCaller \
      -R $GLOBAL_SCRATCH/hg38.fa \
      -I $f \
      -o
$GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/ANDY_${ANDY_COUNT}.raw.snps.indels.g.v
cf \
      --sample_name $SAMPLE \
      --emitRefConfidence GVCF \
      --dbsnp $GLOBAL_SCRATCH/1000G_phase1.snps.high_confidence.hg38.pl.vcf
\
      -L $GLOBAL_SCRATCH/AAA.TRIM/$ANDY_COUNT/call_sites.interval_list
done
ANDY_COUNT=$((ANDY_COUNT=ANDY_COUNT+1))
done
```

Script 2.12 – **Haplotype caller.** Estimates likely regions of variation within the provided sequence using the SW algorithm. Our dbsnp file has been sorted via coordinate to ensure contig compatibility with reference.

The produced collection of gVCFs were then loaded into GenotypeVCF (Van der Auwera et al., 2013) to generate a list of raw SNPs and indels. By combining our sample gVCFs into a cohort gVCF, we increased sensitivity to detect variation in individually difficult sites:

```
module load gatk
GATK=/srv/global/scratch/groups/sbs/GATK/3.5.0/GenomeAnalysisTK.jar
java -jar $GATK \
   -T GenotypeGVCFs \
   -R $GLOBAL_SCRATCH/hg38.fa \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/1/ANDY_1.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/2/ANDY_2.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/3/ANDY_3.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/4/ANDY_4.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/5/ANDY_5.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/6/ANDY_6.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/7/ANDY_7.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/8/ANDY_8.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/8/ANDY_9.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/10/ANDY_10.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/11/ANDY_11.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/12/ANDY_12.raw.snps.indels.g.vcf \
   --variant $GLOBAL_SCRATCH/AAA.TRIM/13/ANDY_13.raw.snps.indels.g.vcf \
   -o $GLOBAL_SCRATCH/raw_pooled_GVCF.vcf
```

Script 2.13 – **Genotype GVCFs.** This tool intelligently handles multiple samples with a joint aggregation step, merging each position of the selected –variants and re-estimating the genotype likelihoods.

2.2.3 Variant discovery

Variants were identified from our collection of raw variants using a multi-step process. To attempt to mitigate the inherent weaknesses of manual filtration, we adopted an integrative and selective manual filtration approach. Custom call sites for each sample were generated via CallableLoci (Van der Auwera et al., 2013) and used to exclude any regions outside this interval list. Furthermore, the normal distribution of variants was quantified for individual samples using a custom filter thresholds set. Annotation values were plotted against each other to generate clusters of variants, and from that a facsimile of multi-dimensional annotation

information was generated. This was used to inform which hard filter thresholds were ultimately used in variant filtration.

2.2.3.1 Transition/Transversion ratio

The Ti/Tv ratio was calculated as described previously (Kristina Strandberg and Salter, 2004). In brief, raw SNP calls were partitioned into tranches, and transitions and transversions were identified using VariantEval (McKenna et al., 2010).

2.2.3.2 Filtration of variants using SnpEff/SnpSift

Variants were annotated and filtered using the joint SnpEff/SnpSift package (Cingolani et al., 2012). Filtration was accomplished via user-designed boolean expressions of arbitrary complexity. Filtration in particular focused on excluding non-coding/synonymous variants, excluding common variants (defined as > 0.5% AF, see 2.2.5 for reasoning behind this threshold) and prioritising nonsense, missense and splice-site mutations of high or moderate predicted impact.

```
cat raw_pooled_GVCF.vcf \
  | java -jar SnpSift.jar filter \
    "((ANN[*].IMPACT = 'HIGH') | (ANN[*].IMPACT = 'MODERATE'))" \
  > filtered_pooled_GVCF.vcf
```

Script 2.14 – **SnpSift filtration of variants based off predicted impact.** This simple expression filters the annotated variants based off a user defined expression, in this case whether the predicted impact of the variant has a predicted impact score (PHRED, etc) of high or moderate. More complicated filter expressions can be constructed as required.

2.2.3.3 Exome wide association using PLINK

Association between genotypic and phenotypic data was investigated using PLINK (Purcell et al., 2007). Required inputs were pedigree and genotype information in .ped format, paired with a variant information file (.map). This information enabled the generation of a binary bi-allelic genotype table (confusingly given the same file extension as the browser extensible data format, .bed), an extended variant information file (.bim) and an information file containing patient sample family data (.fam). In our case, the .fam file included the discordant sibling phenotype values (see 2.2.3.4) of disease severity. Given this phenotype, a one degree of freedom chi-square allelic test for association of variants to disease severity was performed.

2.2.3.4 Discordant sibling pair analysis

Sibling pairs were considered to be discordant if they could be segregated along the binary trait of disease severity. This was categorized as being either high functioning NP-C1 or severely affected NP-C1. This clinical information along with genotype information was encoded in PED format, for use in down-stream analysis.

2.2.3.5 Copy number variation

CNV was detected using the copy-number caller Conifer (Krumm et al., 2012). Conifer uses the FASTQ files with raw sequencing data to quickly align 36bp fragments to your targeted regions, calculating read depth from the aligned reads. From this, singular value decomposition is applied to allow the calculation of a 15-exon average to determine sources of variation to the exons per sample. This allows the calculation of the relative copy number of an exon in a sample through the calculation of a dot product (Krumm et al., 2012).

2.2.4 Cell culture

Primary neurons were isolated by Remy Schneider (PhD student, VUW) from (Npc1nmf164) as previously described (Hilgenberg and Smith, 2007), and these cells were maintained in neuronal conditioned media (Neurobasal Medium, Gibco 21103-049). Immortalized fibroblasts derived from NPC patients (GMO3123 and GMO9503) were obtained from the Coriell Institute. GMO3123 carries missense mutations in exons 6 and 21 of the NPC1 gene resulting in impaired cholesterol esterification and 62% impaired sphingomyelinase activity compared to normal fibroblasts. GMO9503 was the control fibroblast since this cell line was derived from a healthy, age- and ethnically-matched person. Fibroblasts were maintained in media comprised of DMEM (ThermoFisher), 10% FCS, GlutaMax (ThermoFisher) and penicillin streptomycin (ThermoFisher). All cell lines were incubated at 37ºC with 5% CO_2.

Cells were passaged with removal of media followed by replacement with the cell-dissociation enzyme reagent TrypLE™ (ThermoFisher). After cell dissociation, these cells were centrifuged in fresh media at 1500 rpm to allow pellet formation. Following resuspension in media, 10 µL of cells were mixed with 10 µL of trypan blue, to allow an estimation of cell density through use of a haemocytometer. From this, cell passage followed in a 96-well plate at a density of $1x10^3$ cells/100 µL of media for treatment experiments.

2.2.4.1 Filipin staining in fibroblasts

Filipin (Sigma F9765) was prepared at 25 mg/mL in DMSO. In a 96 well plate, cells were washed twice with 3x PBS, 100 µl 1% PFA was added, incubated for 20 min, and washed twice with 3x PBS. Then 100 µL of 50 µg/mL filipin in PBS was added to each well in the dark, incubated for 2 h in the dark on a moving platform, washed twice with 3x PBS, and imaged using fluorescent microscopy.

2.2.4.2 Filipin staining in primary neurons

Primary neurons were fixed with 3% sucrose, 4% PFA in PBS for enhanced structural preservation of neurons. In 96 well plates, half of the media was replaced with with an equal amount of fixative, incubated at room temperature for 20 min with orbital rotation at 50 rpm, and washed three times with ice cold PBS. Then glycine (1.5 mg/mL in PBS) was added to to quench aldehyde groups and remove auto-fluorescence, and incubated at room temperature for 10 min. Next 100 µL of 50 µg/mL filipin in PBS was added to each well in the dark, incubated for 2 h in the dark on a moving platform, washed twice with 3x PBS, permeabilization buffer (1.5 mL donkey serum, 9 µg saponin and 13.5 mL 1x PBS) was added, washed twice with 3x PBS, and imaged using fluorescent microscopy.

2.2.4.3 Fluorescent microscopy

For primary neurons, VECTASHIELD Antifade Mounting Medium (Vector Laboratories) (4 µL) was applied to protect against photobleaching. Cells in contact with a wet coverslip (optimised via contact with a Kim-Wipe orthogonally) were interfaced with coverslip facing up. Nail polish was used to seal the cover slip to a microscope slide. Imaging of both fibroblasts and neurons was performed with the Olympus BX63 fluorescence upright microscope. Image pre-processing was performed using cellSens Standard software (Olympus Life Science Solutions). Filipin fluorescent images were taken with neutral density filters ND25, ND6 and a fluorometer DAPI filter (excitation wavelength 350 nm, emission wavelength 470 nm). Cell fluorescence was quantified using FIJI (ImageJ), using the reported integrated density to calculate corrected total cell fluorescence.

2.2.5 Association testing

Testing for significant association between genomic regions including variants and the clinical trait of disease severity was performed with the R software package PODKAT. In

particular, we outline the specific assumptions chosen to increase statistical power for this small sample association test. Where possible, neutral variants were filtered out, as this can decrease statistical power, even if damaging or protective variants are present. Furthermore, the binary trait of disease severity was selected as the clinical phenotype with the strongest signal for association to our genomic information.

2.2.6 Structural modelling

Structural modelling of proteins was performed via BALLView 1.4 (Moll et al., 2005), a molecular modelling framework that utilizes the biochemical algorithms library (Hildebrandt et al., 2010). Protein structures were downloaded from the RCSB Protein Data Bank (http://www.rcsb.org) *e.g.* MARCH8 PDB ID: 2d8s.

2.3 RESULTS

2.3.1 Exome sequence annotation

The onset and progression of NP-C is highly heterogeneous (Vanier, 2010). Since there is heterogeneity between siblings with the same disease-causing mutation, we hypothesized that there are genetic variants modifying disease severity. Genetic modifiers have previously been identified (Deutsch et al., 2016; Fu et al., 2012; Liao et al., 2015; Malnar et al., 2014), but only in the limits of the laboratory (*i.e.*, the genetic modifiers were defined by ability to reduce filipin accumulation in cell culture). To identify modifiers and associate these modifiers with clinical symptoms, we sequenced the exomes of five sib-pairs wherein each the disease-causing NPC1 mutation was shared yet the disease severity was divergent (Table 2.1).

Table 2.1 – **NP-C sibling-pair patient cohort with clinical covariates.**

Patient number / Gender	Sibship	Disease Severity at 15 years old	Age at neuro. symptoms	Seizures score†	Ambulation score†	Speech score†	Swallowing score†	NPC1 var.‡
1/M	A	Functional	17	0	2	3	3	p.I1061T/ p.P1007A
2/M		RIP	8	5	5	3	5	
3/M	B	High func.	25	0	0	0	0	p.G922X/ p.V378A
4/F		RIP	18	0	5	5	5	
5/F	C	RIP	8	3	-	-	-	p.I1061T
6/M		High func.	5	3	-	-	-	
7/F	D	Func.	27	-	4	-	4	p.I1061T/ p.A108H
8/F		Severe	5	0	5	5	5	
9/F	E	Severe	12	5	4	3	3	p.I1061T/ p.P1007A
10/M		High func.	15	0	0	0	0	

†: Incomplete availability for diagnostic severity scores, if completely unavailable value was denoted with '-'. Scores range from 1-5, *i.e.* an ambulation score of 1 would mean clumsiness, whereas a 5 would translate to wheelchair bound.

‡: Genomic coordinates prior to cross-mapping.

The discovery of high quality variants called from high-throughput sequencing data is well established, with several frameworks previously described (DePristo et al., 2011; McKenna et al., 2010). To accomplish variant calling, raw reads were mapped, aligned, deduped and recalibrated via a calculated per base error model. Between 60-80 Mb of raw

sequence was generated for each NPC patient. Multiple libraries of raw sequence reads (forward and reverse) per patient sample in FASTQ format were interleaved and individually pre-processed before merging into one BAM file per sample (Figure 2.2A). Each sample was then genotyped and underwent variant discovery (Figure 2.2B). Raw variants were filtered via quality metrics and integrative steps (Figure 2.2C). This error model and subsequent processing steps were performed through the best practises pipeline of the GATK (McKenna et al., 2010).

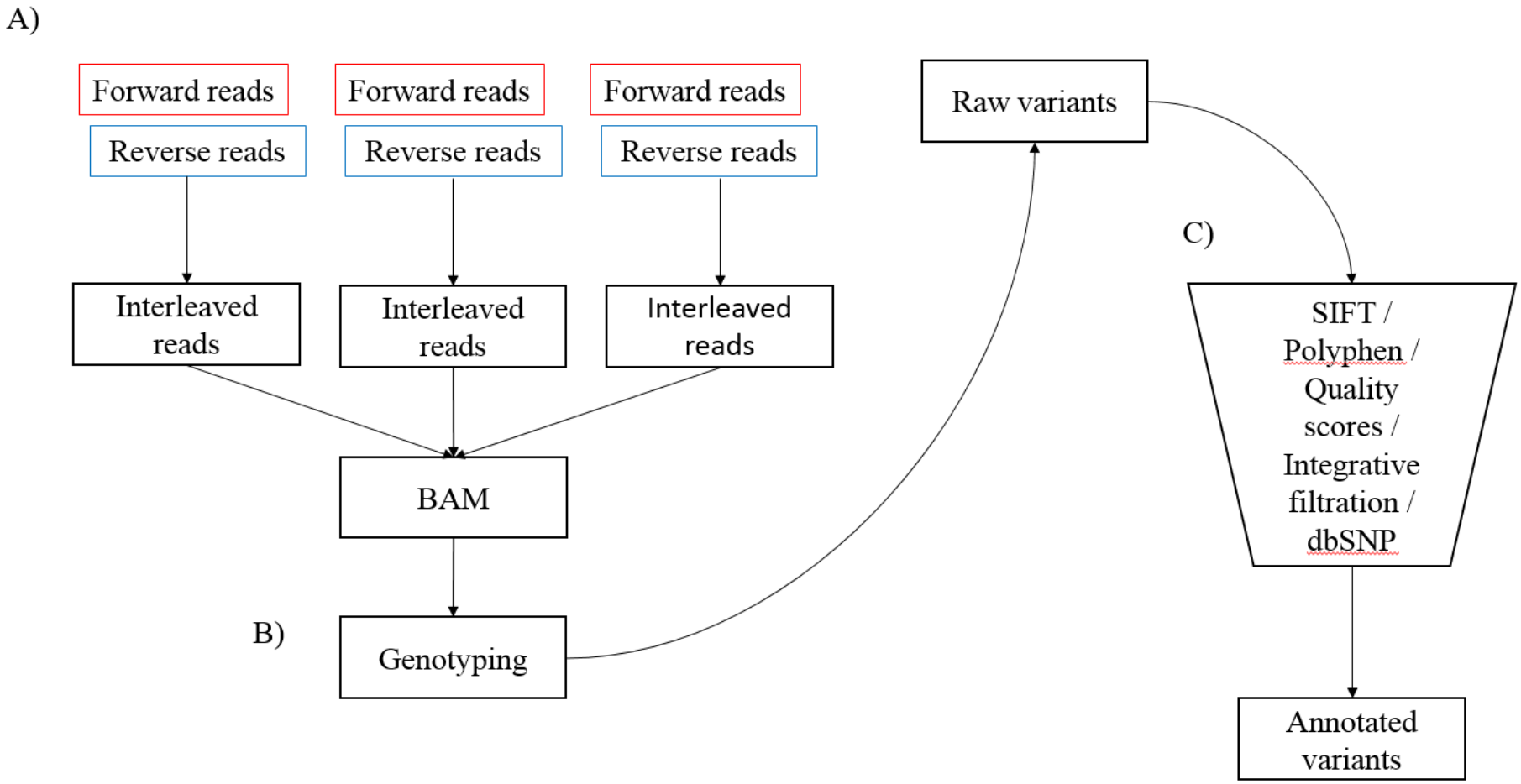


Figure 2.2 – O**verview for generation of variants from WES data.**

2.3.1.1 Base quality scores

Raw sequence reads were examined using the FastQC tool (Andrews, 2010), a quality control tool for high-throughput sequence data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Of the quality metrics assessed, all samples passed per base/per tile sequence quality, sequence content, duplication levels, length distribution and adapter content, while approximately half of the samples threw a warning for per sequence GC content and failed the Kmer content check. For the first 99 bp of every 100 bp, the quality scores were between 28-40 (Figure 2.3), a range that is acceptable for further analysis (Van der Auwera et al., 2013), as a decrease in mean base quality as more bases are read is to be expected.

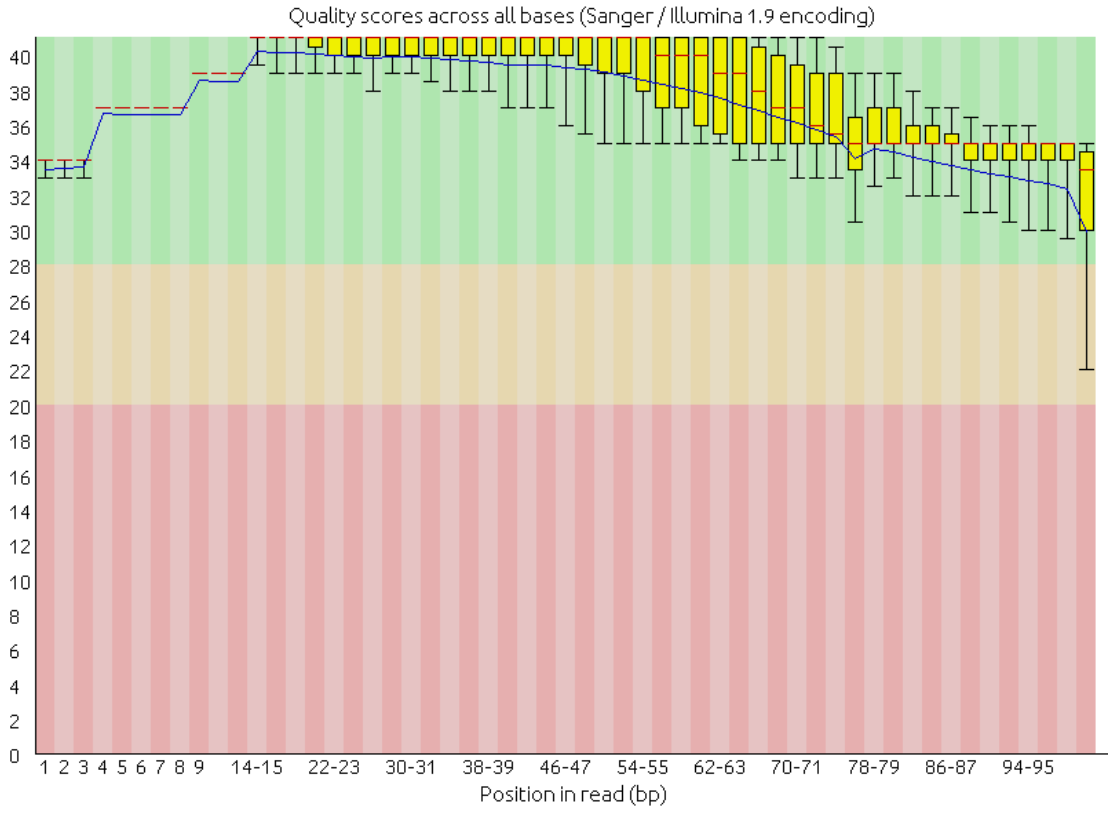


Figure 2.3 – **Representative quality score across all bases for raw sequence reads from Patient 1.**

2.3.1.2 GC content warning

Those sequence reads that generated a warning all had profiles where a warning constitutes that greater than 15% of the reads deviated from the normal distribution (Figure 2.4). An immediate feature is the shifted distribution, which represents a systemic bias independent of base position – this was not a dire problem as FastQC builds the reference distribution from the observed data, independent of the true GC content of the exome. Of more concern is the sharp peak around 59% mean GC content that disrupts an otherwise smooth distribution (Figure 2.4), suggesting a specific contaminant, either adapter dimers or some other over-represented sequence.

However, these metrics passed quality control, suggesting that this peak is not caused by an over-represented sequence (Figure 2.6) or the adapter dimers tested (Figure 2.5). To assess this, we utilized trimmomatic (Bolger et al., 2014) to remove other potential adapter sequences, although this did not remove the peak in those sequence reads affected. The absence of Illumina adapters (Figure 2.5) removed the possibility that this is the cause of the deviation from the expected distribution of GC content. An additional feature of the over-represented sequence quality check is the fact that the first 5 bp of the read are noiser than the subsequent calls, a phenomenon that was taken into consideration when using tools that only use the start of the read (such as the Kmer content module).

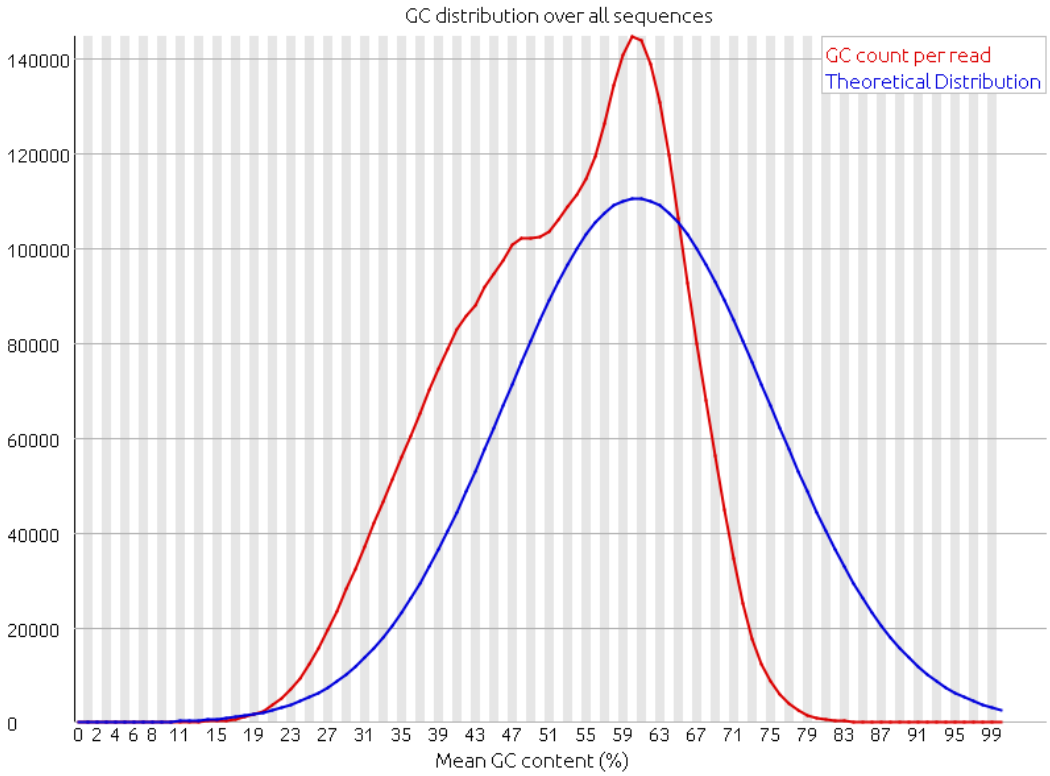


Figure 2.4 – **Sharp GC peak away from theoretical normal distribution of GC content.**

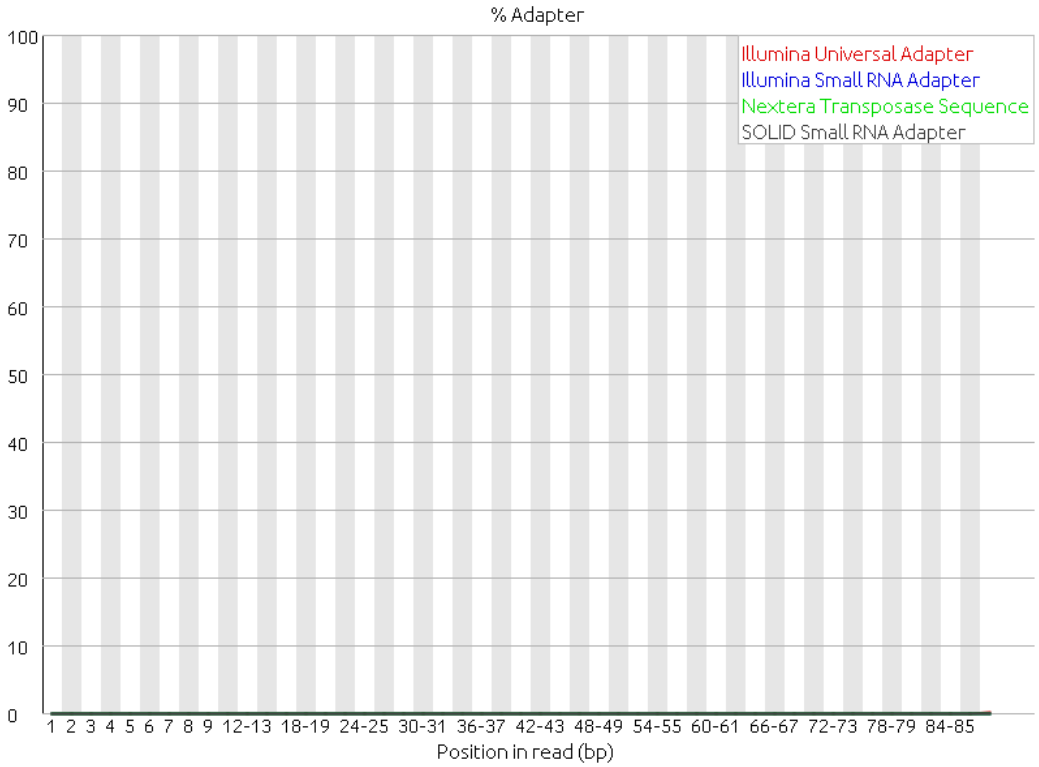


Figure 2.5 – **Absence of adapter dimers.**

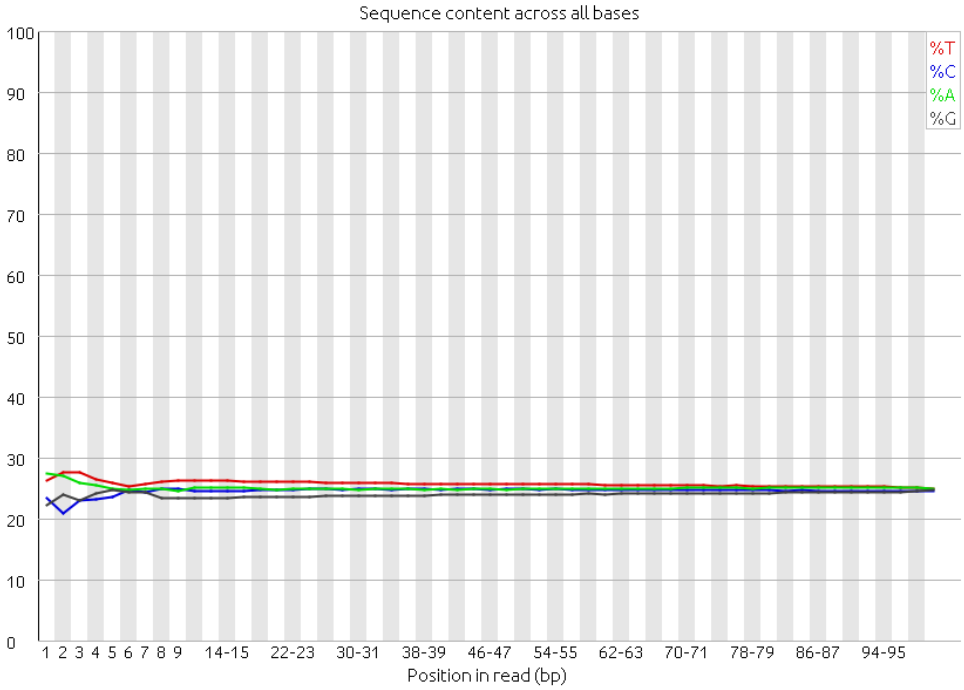


Figure 2.6 – **Per base sequence content is not overrepresented.**

2.3.1.3 Kmer content

The Kmer content module of FastQC makes a number of assumptions about the provided sequence data, the first of which is that any small sequence fragment does not have a positional bias (Schwartz et al., 2011). The Kmer module reasons that even if there is a biologically sound reason for Kmer enrichment or depletion, then that biological bias should affect all read positions within a sequence equally, and reports any Kmers with positional enrichment. However, this means that even if on a per base sequence content or overrepresented sequence metric there is not an issue, Kmers from those sequences can still be flagged by this module. This would appear as sharp peaks at single points in the sequence, which is what we observed for six different 6-mer sequences (**Error! Reference source not ound.**). The most prominent peak for all 6-mer sequences was identified between 40-47 bp in each read. As only 2% of the total sequence library was run using this module with the rest

extrapolated, Kmer bias may also be triggered from sequence libraries with random priming due to incomplete sampling.

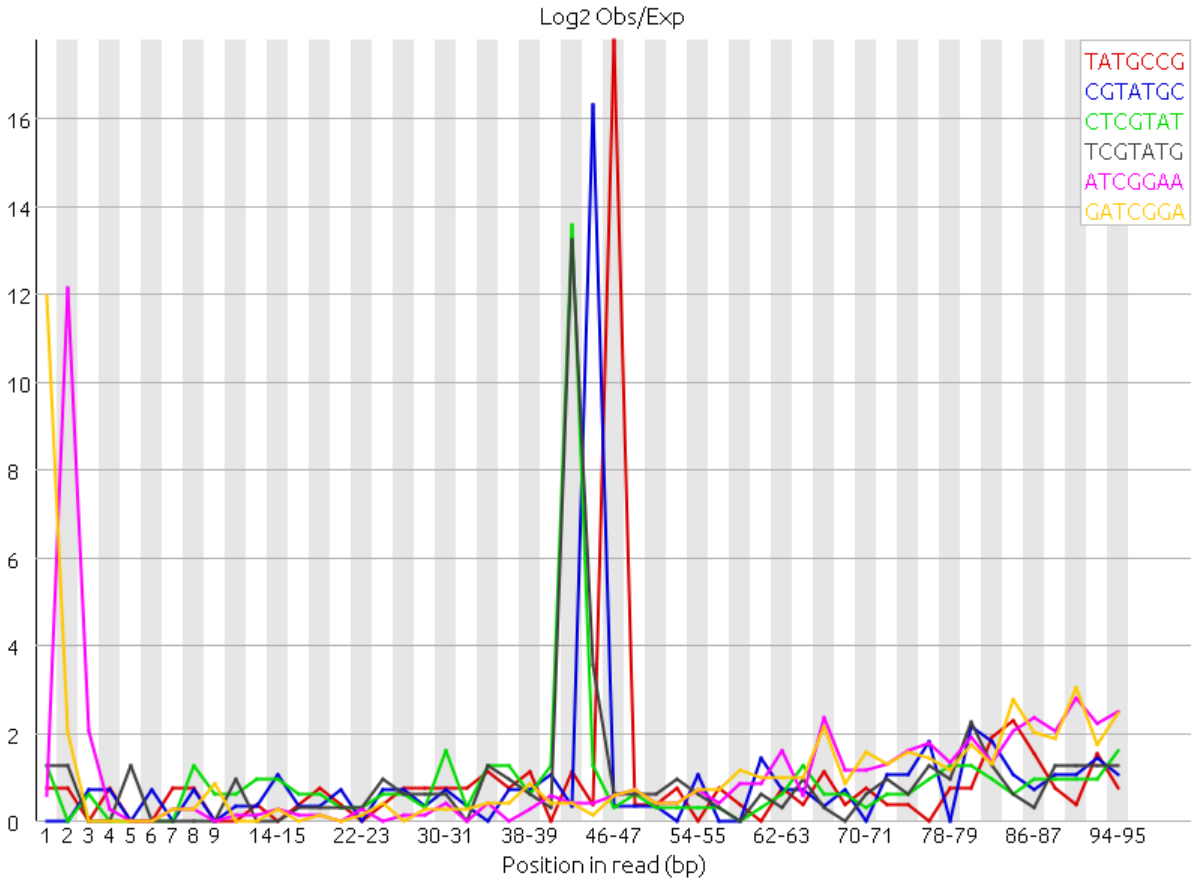


Figure 2.7 – **Deviations from even Kmer coverage.**

2.3.1.4 Raw read quality control

Given the above outputs, the given sample libraries are high in per base sequence quality scores (Figure 2.3), with per base sequence content smooth and overlapping (Figure 2.6), a single shifted normal distribution with a sharp GC peak (Figure 2.4) and no present adapter dimers (Figure 2.5). In the context of the whole quality control, the Kmer module result can be safely disregarded. Of more concern is the GC content peak that is not explained by adapter sequences or overrepresented sequences (Figure 2.5, Figure 2.6). A potential explanation is viewing the read distribution as a bimodal distribution with the shallow peak

corresponding to intronic content and the sharp peak exonic content (Figure 2.4), suggesting that there were still some reads mapped to introns while our library was targeted to exons.

2.3.1.5 Processed data quality control

To quantify the processed reads of our patient exomes, density plots against the exome annotation values were generated. In these plots, area under the curve corresponds to the probability of the annotation containing that value (Figure 2.8; Figure 2.9; Figure 2.10; Figure 2.11; Figure 2.12; Figure 2.13; Figure 2.14). The proportion of variants were lost when filtered via this annotation value (grey vertical line(s) in Figure 2.8; Figure 2.9; Figure 2.10; Figure 2.11; Figure 2.12; Figure 2.13; Figure 2.14). This style of plot can produce pleasing concordance with the biological underpinnings of the data set – for example, the bi-modal distribution of variant quality (when normalized to depth) corresponds to heterozygote and homozygote variants (McKenna et al., 2010). We normalized via depth to avoid accidental inflation of annotation values caused by deeper than average coverage. Each of the listed figures, when taken alone, visualized the likelihood of finding an annotation at that value. Our variant quality, when normalized by depth, resulted in a bimodal distribution corresponding to the zygosity of the called variants (Figure 2.8) – if we did not observe this, then that would be a red-flag. The density plot revealed that a significant portion of the called variants were below a depth of 10 (Figure 2.9), indicating we can then attempt to find parameters that target those variants that are both low in quality and low in depth. The strands odd ratio (Figure 2.10) had a long tail of failing variants above a SOR of 5 and fisher strand bias (Figure 2.11) showed failing variants with FSB beyond 100. These plots indicated that our called reads were not biased in terms of allelic count in forward and reverse reads. Figure 2.12) and quality score (Figure 2.13) is centred on zero, thus variant alleles were not different from reference alleles in terms of position of quality in some spurious way. Finally, the root mean square mapping quality was a sharp peak around 60, verifying that the majority of variants were of good quality.

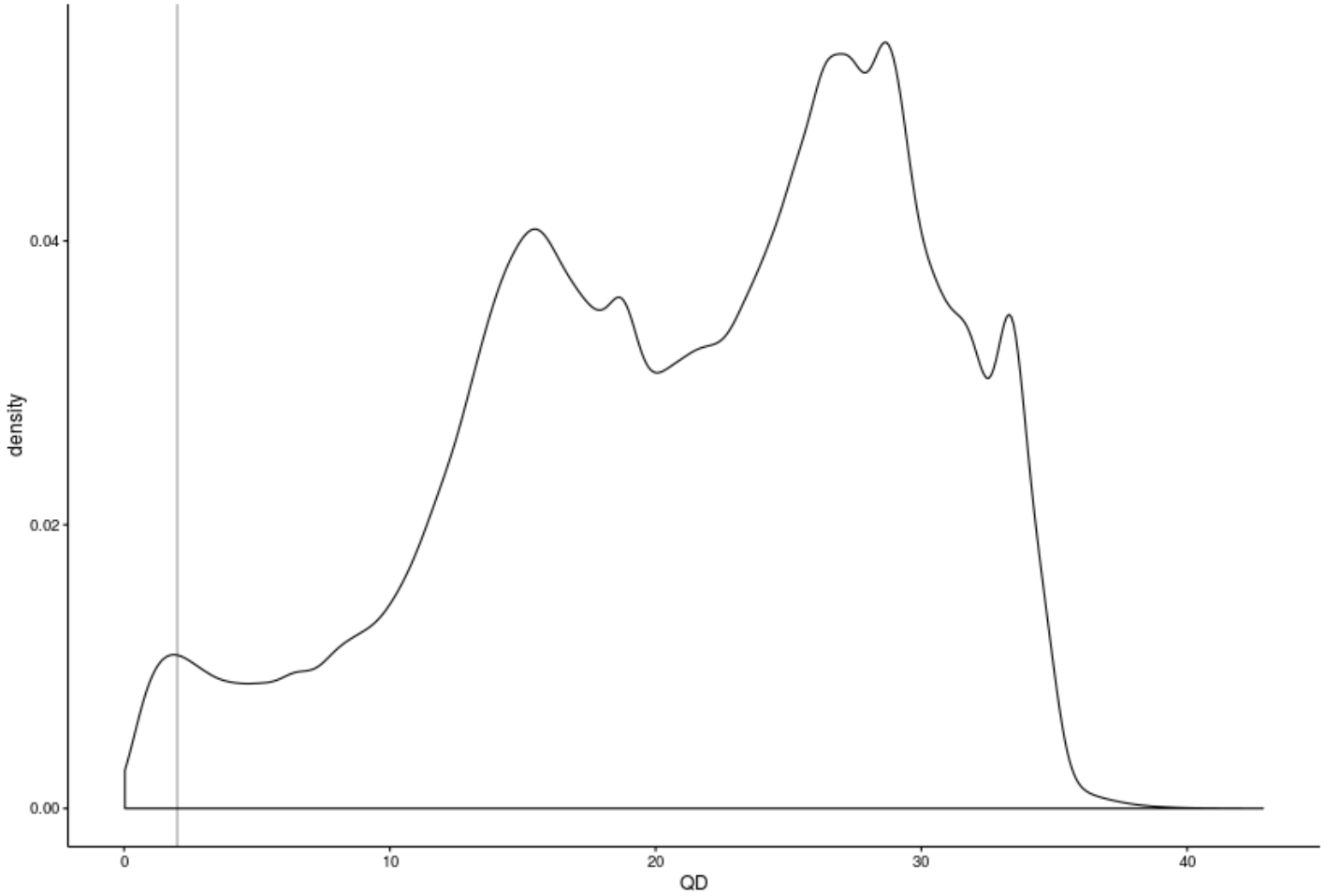


Figure 2.8 – **Density plot of exome variant quality normalized by depth.** The two large peaks around QD 15 and QD 28 correspond to heterozygote and homozygote variants, respectively. Any variant below the cut-off shown (QD < 2) would be filtered from the call set.

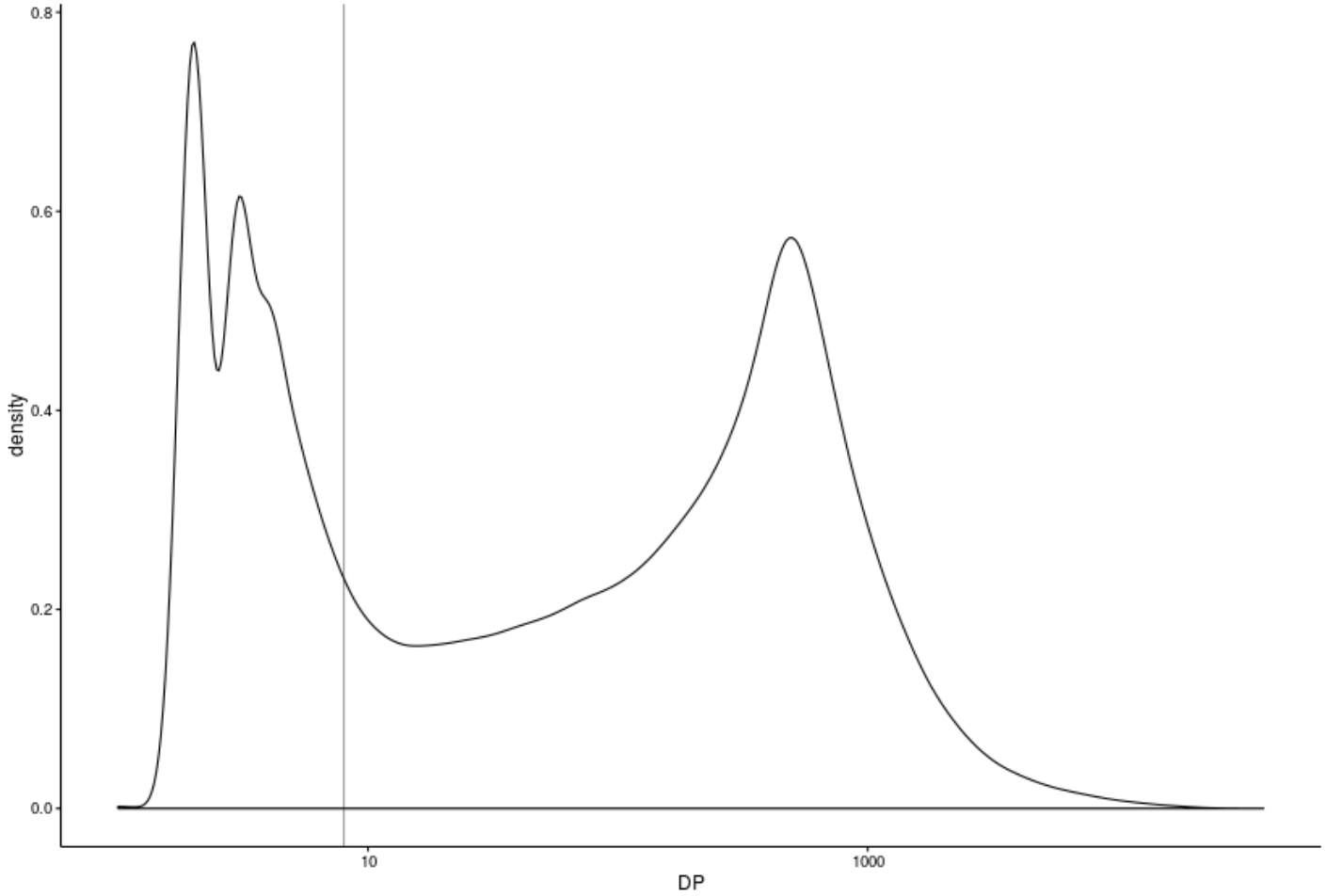


Figure 2.9 – **Density plot of exome depth.** Compare the plot of exome depth, DP, with that of quality normalized by depth, QD. Despite Carson et al. demonstrating that DP can be used to improve data quality in WES studies (Carson et al., 2014), it is not recommended to filter off DP alone in WES as the number of reads given an average coverage is not well-defined.

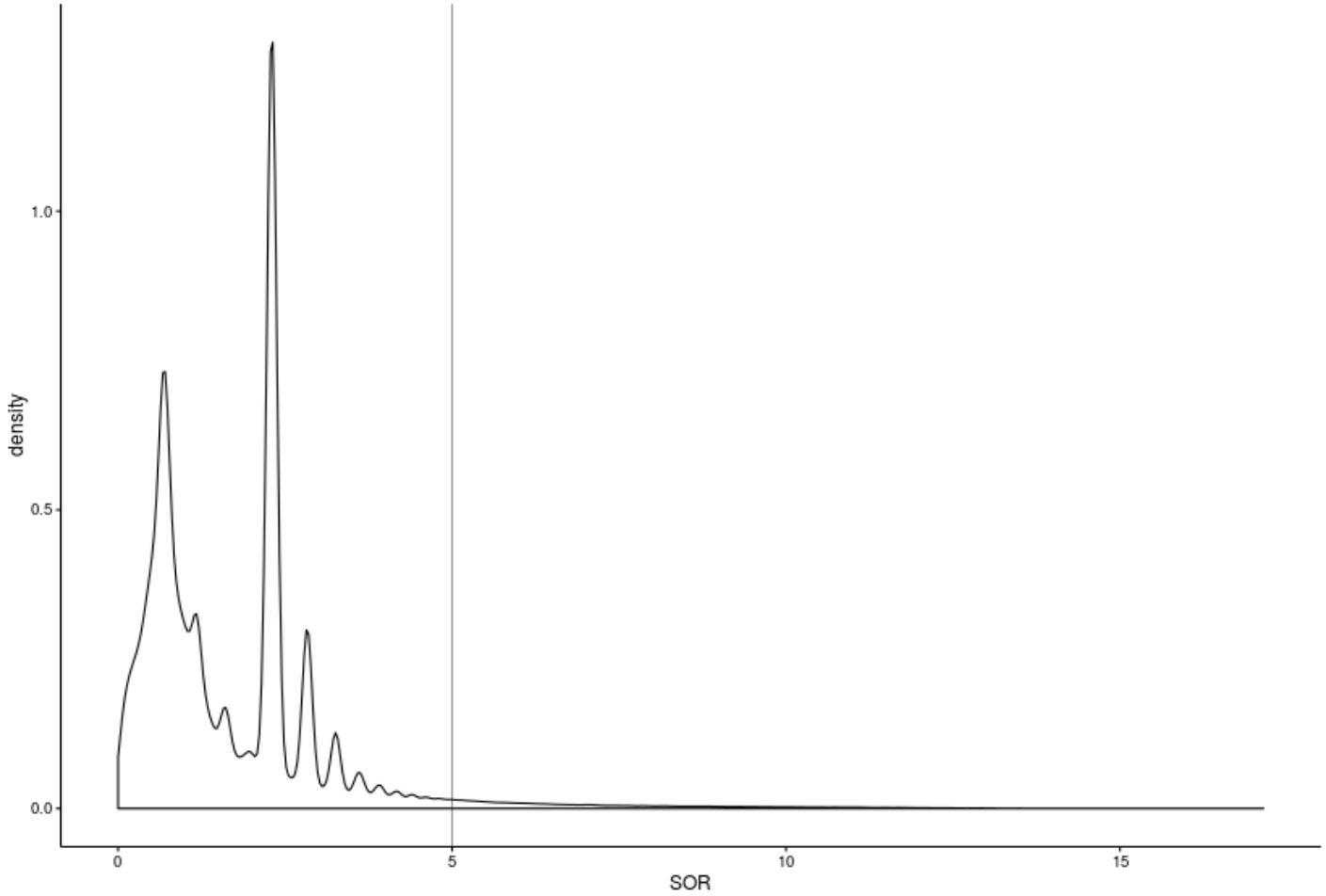


Figure 2.10 – **Density plot of exome strands odd ratio.** Strands odd ratio is a metric that considers the ratio between alleles of covering reads. In this example, a cut-off threshold of SOR > 5 will filter the long tail of likely failing variants.

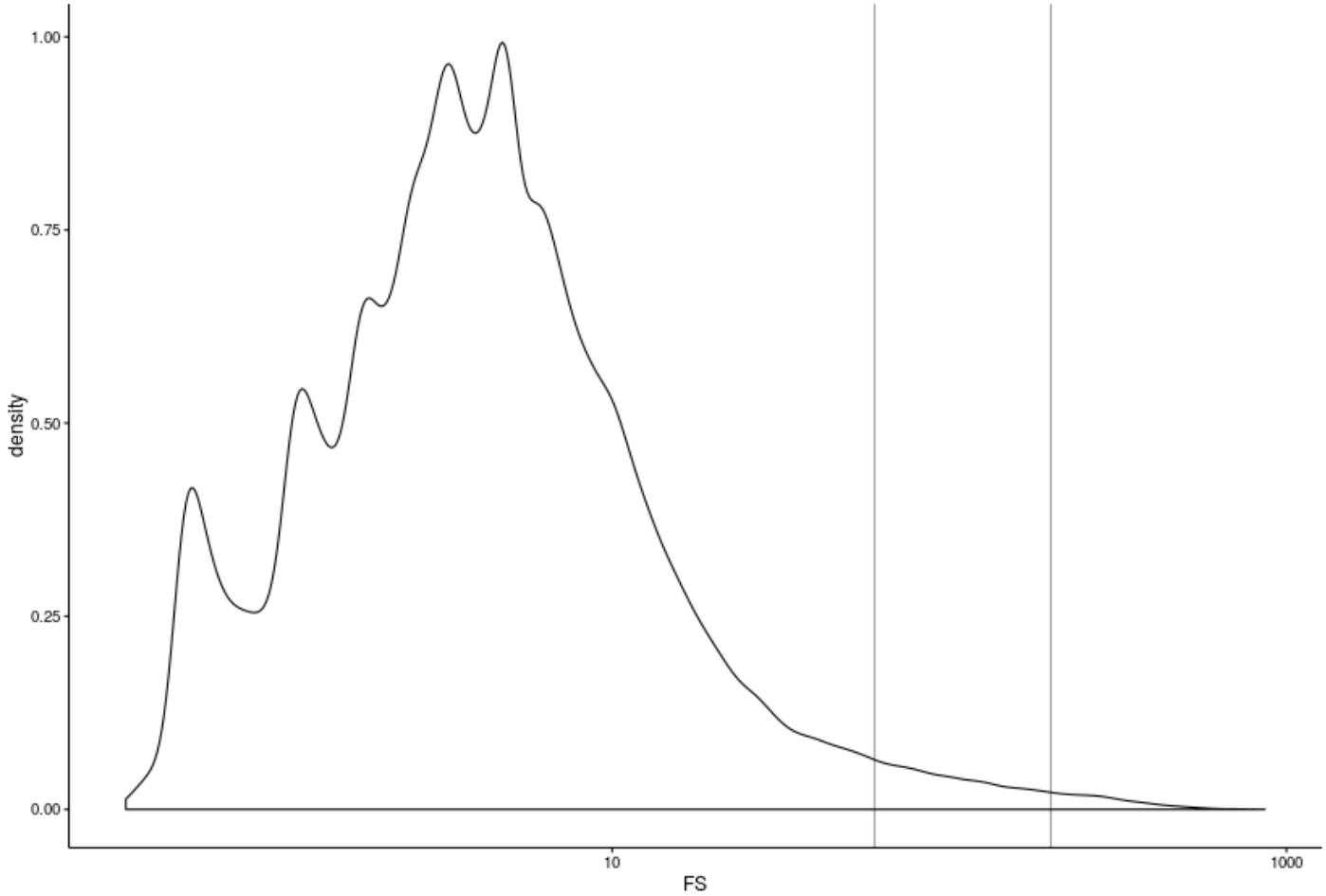


Figure 2.11 – **Density plot of exome Fisher Strand bias.** Fisher strand bias is the probability of a non-reference allele being observed greater or less than expected than the reference allele on the forward or reverse strand. In short, with no strand bias, FS will equal zero. FS is derived from the raw counts of allelic reads on both the forward and reverse strand, where those counts are then used as nominal variables in Fisher's Exact Test. An issue with filtering via a FS threshold is that variants that occur at the termination of exons are unfairly penalized, as they tend to having coverage in only a single direction.

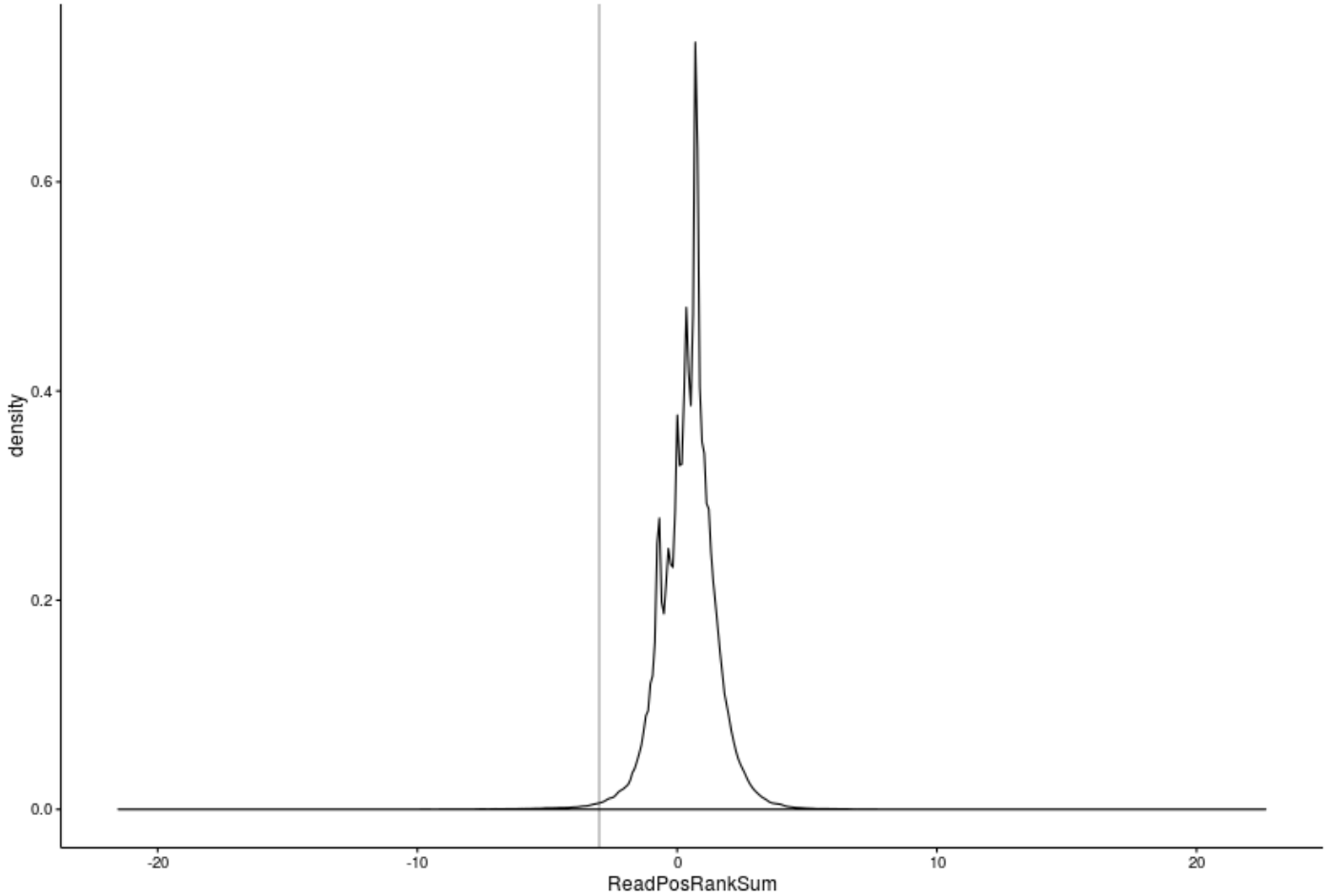


Figure 2.12 – **Density plot of exome read position rank score.** The read position rank sum test is a metric used to assay site position within reads, testing whether alleles alternative to the reference are only found at the ends of exons. If so, these are assumed to be the product of a sequencing error. The test metric is centred on zero – negative values represent alternative alleles at terminal positions more often than reference, positive values the opposite. Clearly, zero is preferred as it means no significant difference between the positions of alt and ref alleles.

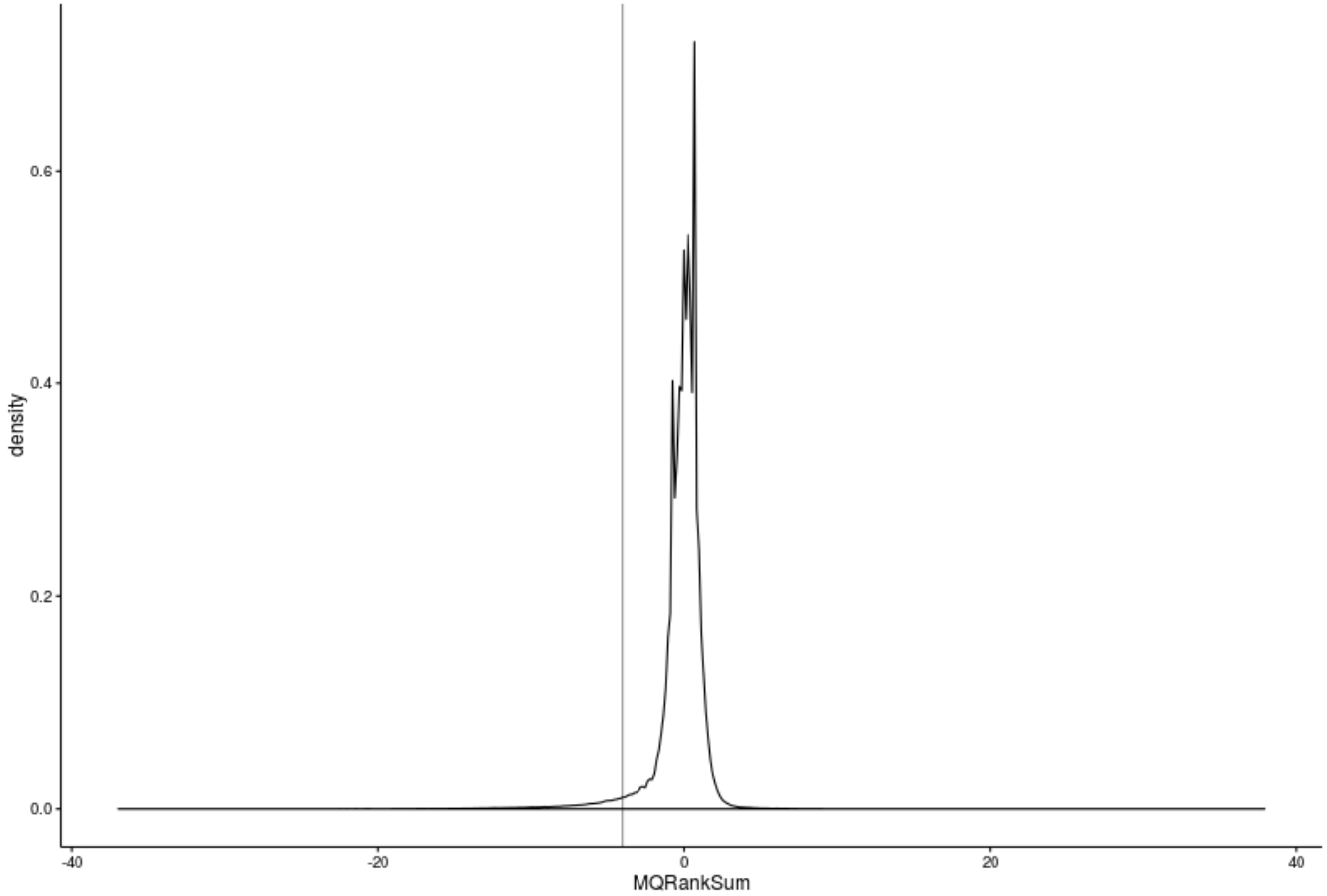


Figure 2.13 – **Density plot of exome mean quality rank score.** The mapping quality rank sum test compares the reported quality scores of the reads for the alternative and the reference allele. A value of zero means there is no difference between mapping quality scores for reference and alternative alleles. These plots with a narrow distribution around zero highlight the difficulties faced in using hard filters across a single dimension of data – it is difficult to exclude poor variants without excluding valid ones as well, and vice versa. If $MQ^{ALT} > MQ^{REF}$, then the reported MQRankSum score will be positive. If $MQ^{REF} > MQ^{ALT}$, then the score will be negative.

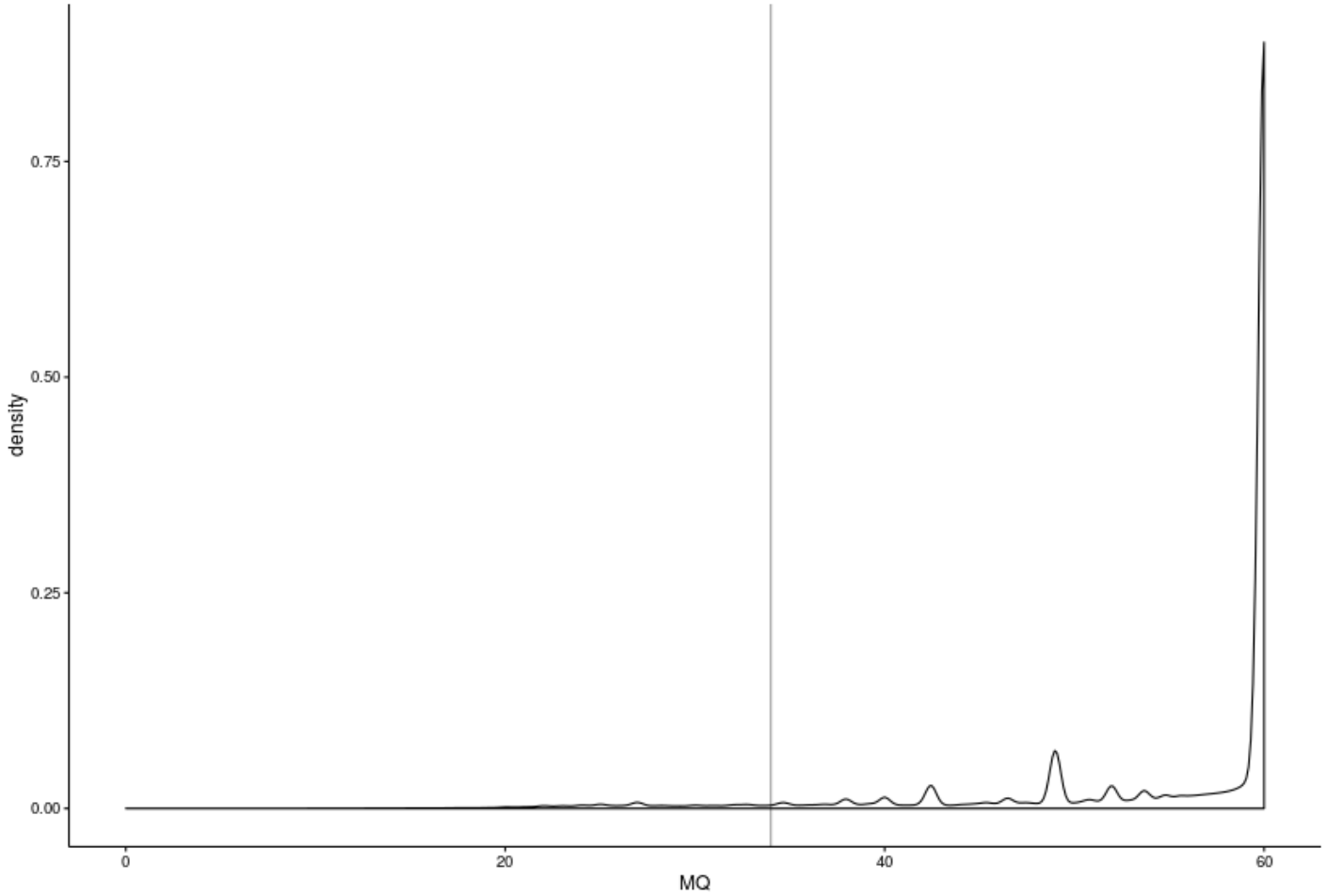


Figure 2.14 – **Density plot of root mean square mapping quality.** The Broad Institute recommends a hard filter threshold of MQ = 40. An alternative filter could be applied at MQ = 50 or higher. Because this statistic includes the standard deviation of the read mapping qualities, it includes the variation within the dataset. As such, the desired MQ value for analysis appropriate reads is around 60.

2.3.1.6 Integrative and selective filtration

The recommended approach to variant filtration after calling using the GATK pipeline is re-adjustment of variant call score via variant quality score recalibration (VQSR) (McKenna et al., 2010). This approach is desired because it offloads the difficult task of finding the optimal balance between tranche sensitivity (obtaining true positives) and tranche specificity (limiting false positives) from the researcher to an algorithmic approach. VQSR uses machine learning, where gold-standard datasets with extremely clear profiles (*e.g.*, 1000 Genomes, HapMap) are used as inputs to the algorithm, such that strong variant annotation sets and weak variant annotation sets are used to train the algorithm. This approach is powerful because the data can be considered from multiple-dimensions, an approach that is traditionally difficult. Dimensional analysis of data enables clusters of variants and ultimately results in more informed filtration of variants. Even if an individual annotation value for a variant is low, it might still be a true variant when considered in light of other annotation factors.

Unfortunately, due to the large number of variant sites required for the VQSR approach, it requires a sample size of at least 30 exomes (McKenna et al., 2010). As our cohort of 10 exomes fails this requirement, we were unable to employ the VQSR method and instead used manual filtration setting hard thresholds for various attributes of the calling set. If the value fails this threshold, it is culled from the putative variant VCF. It is easiest to demonstrate the reason this approach is flawed with a graphic showing randomly generated data that is filtered when using the (superior) VQSR method and those filtered using manual filtration (Figure 2.15); the area under the curve represents probability of observing an annotation at that particular value.

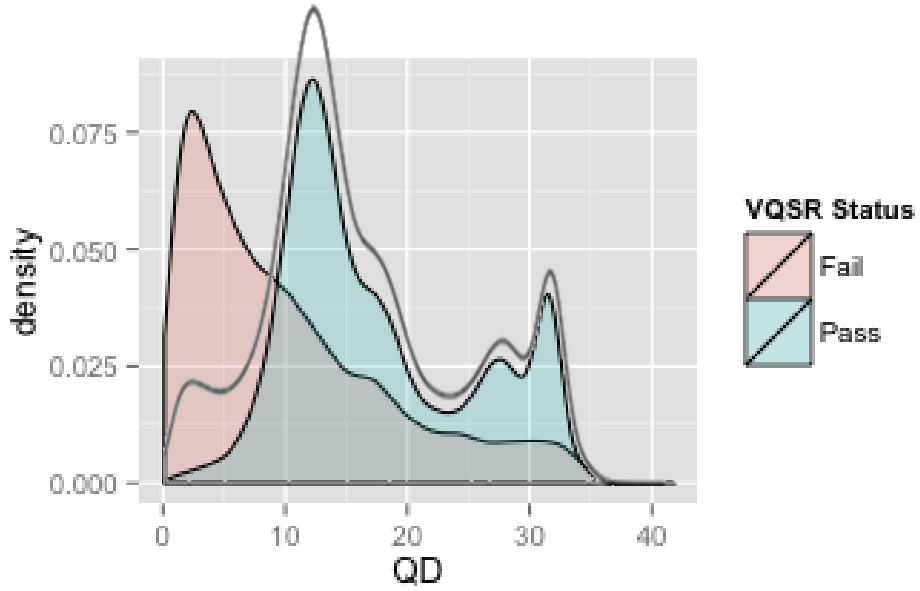


Figure 2.15 – **VQSR filtration density plot.** Red AUC represents failing 'variants', green AUC represents 'variants' that pass the VQSR filter. Curve with no fill is the 'pre-filtered set of variants' (i.e. the set of all hypothetical variants before VQSR). The y-axis represents proportion of variants.

After applying the VQSR method to all variants in our simulated cohort (uncoloured curve), we segregated variant calls that pass and fail (Figure 2.15). While it is true that a large number of variants that fail (red curve) have a QD value < 5, there was a significant tail of failing variants that had high QD values > 10. This simultaneously demonstrates the strength of observing data multi-dimensionally and the weaknesses of relying on a hard filter. If the threshold filter were culling any variant with a QD value < 5, then we were still passing a significant number of variants that ought to fail (for reasons found in the other annotation dimensions). Furthermore, we were failing some true variants (green curve) with QD values < 5. For these reasons, rather than 'selecting' a single manual threshold, we integrated these parameters together to filter our variants. While still inferior to the unavailable VQSR method, this approach attempts to recapitulate the clustered nature of traits with reliable variants.

Hence, by visualizing the annotation values of our called variants in density plots (Figure 2.8; Figure 2.9; Figure 2.10; Figure 2.11; Figure 2.12; Figure 2.13; Figure 2.14) and investigating the shortfalls of hard filtering a calling-set (Figure 2.15), we were able to optimize

the filter parameters for our individual patients (Table 2.2). All patients had a depth threshold set at >10, while GQ had a range between 92-98 as the hard filtration cutoff

Table 2.2 – **Summary of integrative filtration parameters.** *These parameters were selected to be sensitive to variant discovery.*

	Patient ID										
Filter parameters	1	2	3	4	5	6	7	8	9	10	ALL
Depth (DP)											10
Genome Quality (GQ)	93	92	93	98	92	92	94	94	92	92	-
Quality normalized by depth (QD)											10
Fisher Strand Bias (FS)											40
Root Mean Square Mapping Quality (MQ)											59
Strand Odds Ratio (SOR)											2
Mean Quality Rank Score (MQRS)											-4
Read Position Rank Sum (RPRS)											-3

Table 2.2 – **Summary of integrative filtration parameters.**

2.3.1.7 Ti/Tv ratio

The ratio of transitions (mutations from either a purine or pyrimidine nucleotide to the same kind of nucleotide) to transversions (mutations from one type of nucleotide to the other) in humans is a useful descriptive statistic to quickly check the quality of called variants (Kristina Strandberg and Salter, 2004). As there are twice as many possible transversions than there are transitions possible (Ebersberger et al., 2002), the transition:transversion (Ti/Tv) ratio has been consistently identified at 2.1 for whole genome sequencing and 2.8 for whole exome sequencing. A Ti/Tv ratio that deviates from the expected value is suggestive of false positives that were not filtered out.

The Ti/Tv ratio for our raw sequences was 2.16 (Table 2.3). Using an integrative approach, we improved the Ti/Tv ratio via the removal of confounding variants that distorted the ratio away from the expected ratio. As we refined parameters, the Ti/Tv ratio was altered

to the expected exome-wide level of 2.75 (Table 2.3; Figure 2.16). To visualize this, we utilized GATK to estimate the probability that each filtered variant was true and represented this as a continuous (non-discrete) measure. We then partitioned the SNP calls into tranches (or brackets) as determined by the Ti/Tv ratio of the SNPs in that tranche. This allowed us to pick the desired levels of specificity and sensitivity; the number of true variants increased as the number of false/poor variants increased. These results show that our called variants after integrative filtration were concordant with the biologically relevant, descriptive statistic of Ti/Tv. Furthermore, this approach enabled an empirical approach to selecting the desired level of specificity for true variant calls, while allowing a sensitivity threshold capable of facilitating discovery of putative genetic modifiers to NP-C disease. These results also enabled an estimation of the degree of improvement gained from integrative filtration: 10.89% increase in false positives removed and 58.98% increase in false negatives removed.

Filtration parameter	Ti/Tv
Raw	2.16
UCSC[1]	2.56
1000G[2]+dbSNP[3]	2.6
dbSNP	2.59
refseq.interval_list	2.62
DP10[4]	2.61
DP50	2.62
1000G+DP&&GQ[5]	2.696
1000G+DP&&GQ||DP||GQ	2.7
dbsnp148+above	2.72
1000G+DP&&GQ@300&94	2.74
1000G+DP&&GQ@400&97	2.75

Table 2.3 – **Effect of filtration parameters on Ti/Tv ratio.** Ti/Tv ratio is a statistic that is expected to be around 2.8 – 3.0 for whole exome sequences in humans. However, the exact value obtained depends on a large range of factors with respect to depth of coverage, quality of sequencing and technical errors. 1: University of California Santa Cruz; 2: 1000 Genomes; 3: Database SNP; 4: Depth of coverage; 5: Genome quality.

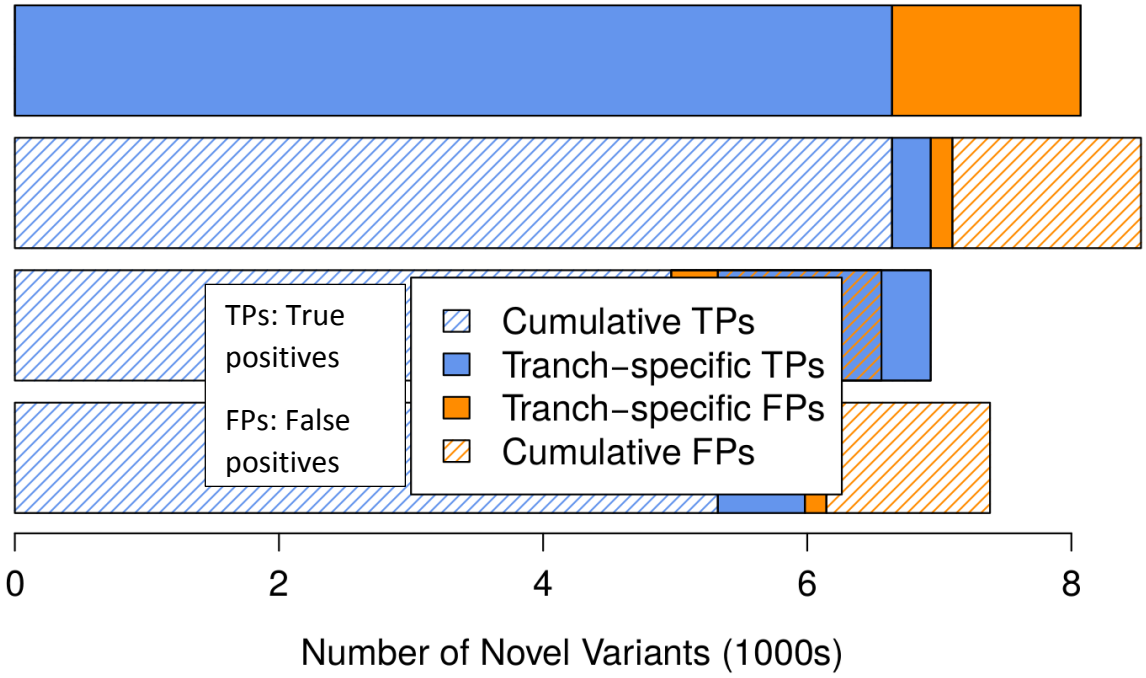


Figure 2.16 – **Tranche plot of NP-C exome SNP calls.** Each subsequent tranche is less specific, but increases sensitivity for novel variants, at the cost of introducing more false positives (FPs).

2.3.2 Identified putative modifiers within multiple functional categories

We next determined the genes of interest containing non-synonymous variants with predicted functional consequences (based on SIFT, MAPP, PolyPhen, etc.) within our whole exome data set, and whether these segregate across sibling-pair severity, or whether variants in genes of interest exist in the pooled subject cohort. We removed synonymous variants as well as variants that were predicted to be functionally benign while prioritizing stop gain/loss, missense and splicing variants. We anticipated that to modify disease severity, the genetic modifier should segregate cleanly between twins, so in this model a candidate variant would have to be present in one sibling and not the other, with greater weight given to variants that segregate in this manner across multiple kindreds. A homozygous variant against a heterozygous variant was considered to segregate cleanly in this manner, but only with a strict cutoff of 4-5 homozygous variants. We also considered an aggregation model, where each

patient's variants were filtered through frequency of incidence in our cohort and minor allele frequency (MAF) on a population level, with priority given to homozygous variants.

Table 2.4 – **Aggregate filtering of 10 exomes identifies candidate modifier genes**

Filter	Aggregate
Well-characterized alleles	22,367
MAF < 0.05%	2,922
PolyPhen/SIFT functional modelling in population	26

2.3.2.1 Twenty-six genes recovered across seven functional categories in the aggregate model

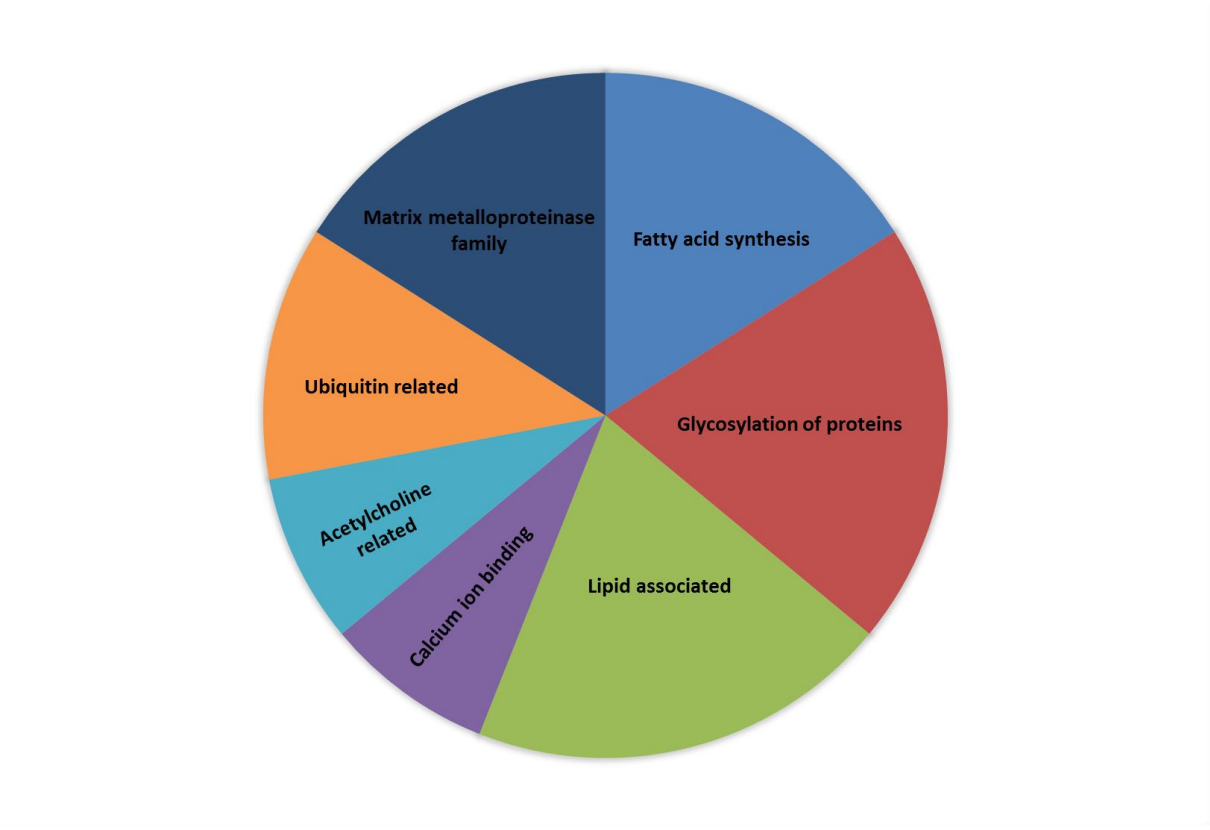


Figure 2.17 – **Functional categories of recovered genes.** Genes of potential interest were found to belong to seven broad, functional groups. Of greatest number of variants per functional group recovered to least: Lipid associated (5 genes), glycosylation of proteins (5 genes), fatty acid synthesis (4 genes), matrix metalloproteinase family (4 genes), ubiquitin related (4 genes), acetylcholine related (2 genes), calcium ion binding (2 genes).

2.3.2.2 Lipid associated

NP-C disease is by definition a disease caused by the defective transport of cholesterol and sphingolipids (Patterson et al., 2012). We identified variants in 5 lipid genes (APOBR, PCTP, XYLB, MROH8, PLEKHA2) in NP-C patients that were not identified in control populations (Table 2.5). These variants were identified as homozygous and heterozygous variants. APOBR is an apolipoprotein B48 receptor associated with macrophages and foam cell formation with lipoprotein uptake (Brown et al., 2000). Phosphatidylcholine transfer protein (PCTP) knockout mice have demonstrated defective lipid homeostasis (Kanno et al., 2007). D-Xylulokinase (XYLB) is an enzyme active in the liver and kidneys and is upstream of a key regulator of lipogenesis (Bunker et al., 2013). Obtaining a variant in MROH8 within a majority of our cohort is noteworthy, as this maestro heat like repeat family member has been previously identified as a susceptibility gene for Alzheimer's disease when hippocampal atrophy is used as the quantitative trait in a GWAS (Potkin et al., 2009). The splice donor variant in the pleckstrin homology domain containing A2 protein (PLEKHA2) present in 5 of 10 patients is of interest in the context of the other major functional groups, namely the MMP family, as it is fibronectin binding. Sphingolipid metabolism is clearly implicated in lysosomal storage disorders, with secondary sphingolipid accumulation a major feature of diseases like Gaucher, NP-C and Batten's diseases. Lipids regulate movement of late endocytic vesicles and in NP-C cholesterol-laden endocytic compartments have impaired motility (Lebrand et al., 2002). The SnpEff annotation predicts that the reported APOBR missense variant identified in NP-C patients (residue change valine to alanine) would modify the CLN3 gene, a transmembrane neuronal ceroid-lipofuscinosis gene involved in lysosomal function that is the causative gene for Batten disease (Cotman and Staropoli, 2012). The fact that a homozygous variant in an apolipoprotein receptor with association to foam cell formation and another LSD occurs in every NP-C patient in our cohort is an intriguing finding.

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
APOBR	rs40832	missense_variant	5 Hom	5 Hom	Apolipoprotein B48 receptor is a macrophage receptor that binds to the apolipoprotein B48 of TG-rich lipoproteins. If overwhelmed with elevated plasma triglyceride, this receptor may contribute to foam cell formation. This variant is predicted to act as a modifier for CLN3 (Batten's disease).
PCTP	rs2960062	structural_interaction_variant	5 Hom	5 Hom	Phosphatidylcholine Transfer Protein / StAR-Related Lipid Transfer Protein 2. Lipid binding and phosphatidylcholine transporter activity.
XYLB	rs818850	structural_interaction_variant	5 Hom	5 Hom	Member of a family of enzymes responsible for phosphotransferase activity, glucose metabolism and lipogenesis.
MROH8	rs11467214	frameshift_variant & stop_gained	5 Hom	3 Hom 1 Bi-allelic	Maestro heat-like repeat family member, function unknown but a genome-wide association study using hippocampal atrophy as the quantitative trait associated this gene with Alzheimer's disease (PMID:19668339).
PLEKHA2	rs76319743	splice_donor_variant	3 Hom	2 Hom	Pleckstrin Homology Domain with related pathways to IL-2 and Icos-IcoL Pathway in T-Helper Cell. Fibronectin binding and lipid binding.

Table 2.5 – **Lipid associated.**

2.3.2.3 Glycosylation of proteins

Lysosomes have a protective glycocalyx (glycoconjugate coating) to guard the membrane against the acidic enzymes within the lysosome. NPC1 is a glycoprotein required for cholesterol export from lysosomes (Maxfield and van Meer, 2010). Also, the major lysosomal membrane LAMP proteins are N- and O-linked glycosylated, which contribute to this protective lysosomal glycocalyx (Wilke et al., 2012). It has been shown that decreased levels of protein glycosylation lower the requirement for NPC1-mediated cholesterol export, and that inhibition of glycosylation reduces cholesterol accumulation in cases of NPC1 deficiency (Li et al., 2015).

We identified a variant in glycosylation genes that segregated within siblings (MUC5B) and 4 variants in glycosylation genes that were abundant in NP-C patients compared to control populations (FDFT1, SSPO, UCK1, RFT1) (Table 2.6). These variants are in genes that are functionally related to glycosylation and glycosylation pathways. Several of the affected genes have also been previously associated with neurological disorders, or pulmonary diseases such as chronic obstructive pulmonary disease. Notably, a homozygous structural interaction variant in FDFT1, a squalene synthetase in cholesterol biosynthesis, was identified across all patients. The major gel-forming mucin protein, MUC5B that is known to be up-regulated in COPD as the major contributor to lung mucus, is an especially noteworthy hit as the homozygous missense variant (residue change from leucine to proline) was found only in the more affected sibling in each sib-pair.

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
FDFT1	rs904011	structural_interaction_variant	5 Hom	5 Hom	Squalene synthetase, first specific enzyme in cholesterol biosynthesis. Diseases include disorders of glycosylation. Regulation of cholesterol biosynthesis by SREBP.
SSPO	rs397815440 & rs71194663	splice_acceptor_variant & splice_donor_variant	5 Hom	5 Hom	Sco-Spondin has related pathways with diseases associated with O-glycosylation and modulation of neuronal aggregation.
MUC5B	rs4963031	missense_variant	4 Het	5 Hom	Major gel-forming mucin protein, highly glycosylated macromolecular component of mucus. Major contributor to normal lung mucus. Up-regulated in COPD. Diseases associated with O-glycosylation of proteins.
UCK1	rs2296957	missense_variant	2 Het 2 Hom 1 Bi-allelic	1 Het 4 Hom	Uridine-Cytidine Kinase 1, involved in purine metabolism. Diseases associated include Kabuki syndrome. Downstream gene variant POMT1, an ER O-mannosyltransferase where defects in POMT1 result in Walker-Warburg syndrome. Related pathways are O-linked glycosylation.
RFT1	rs891368	3_prime_UTR_variant	1 Het 4 Hom	1 Het 3 Hom	Encodes an enzyme responsible for the translocation of intermediates in the N-glycosylation of proteins. N-glycan biosynthesis.

Table 2.6 – **Glycosylation of proteins.**

2.3.2.4 Matrix metalloproteinase family

The matrix metalloproteinase (MMP) family (calcium-dependent proteases, comprised of 25 genes) have functions in degradation of cartilage and bone, fibronectin binding, tissue repair/remodeling and inflammatory response (Page-McCaw et al., 2007). The MMP genes have been implicated in COPD (Li et al., 2009b), atherosclerosis (Williams et al., 2010) and multiple sclerosis (Škuljec et al., 2011). We identified a cluster of MMP-family variants on chromosome 11 and one variant in MMP-12 was abundant in NP-C patients compared to control populations (Table 2.7). These variants were mostly 'singletons' (*i.e.*, a unique variant found only in one sample, however all samples had variants within the genes of MMP1, MMP17 and MMP10). MMP10 and MMP17 degrade fibronectin, which is interesting given the context of the observed PLEKHA2 splice donor variant (Table 2.5). The exception to this was MMP12 with a homozygous frameshift insertion occurring in all 10 subjects. MMP12 has been previously associated with neurological diseases and pulmonary diseases including NP-C disease (Griese et al., 2010; Li et al., 2009b; Liao et al., 2015).

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
MMP10	multiple	varies	NA	NA	Breaks down fibronectin. Cluster of MMP genes on Chr11. Calcium ion binding.
MMP12	rs5794199	frameshift_insertion	5 Hom	5 Hom	Aneurysm formation, lung function and COPD. Cluster of MMP genes on Chr11. Calcium ion binding.
MMP1	multiple	varies	NA	NA	Diseases associated with this gene include COPD. Cluster of MMP genes on Chr11. Calcium ion binding.
MMP17	multiple	varies	NA	NA	This protein is unique among the membrane-type matrix metalloproteinases as it is anchored to the cell membrane via a glycosylphosphatidylinositol anchor. Degrades fibrin. Calcium ion binding.

Table 2.7 – **Matrix metalloproteinase family.**

2.3.2.5 Fatty acid synthesis

Dysregulation of fatty acids in LSDs can result in metabolic stress from insufficient catabolic intermediates (Platt et al., 2012). The broad range of enzymes and their intermediates within these pathways can result in cascade of cellular failures, ultimately resulting in the pathology of the LSD. We identified four variants in fatty acid-related genes (ACACB, LPIN2, ECHDC3, ACADS) that were abundant in NP-C patients compared to control populations (Table 2.8). Two rate-limiting genes in fatty acid uptake and fatty acid β-oxidation (ACACB and ACADS, respectively) had homozygous variants present in the cohort. In addition, LPIN2 and ECHDC3 function in fatty acid biosynthesis (Reue, 2009).

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
ACACB	rs11065772	structural_interaction_variant	5 Hom	5 Hom	Acetyl-CoA carboxylase-beta controls fatty acid oxidation via malonyl-CoA inhibition of carnitine-palmitoyl-CoA transferase I, the rate limiting step in fatty acid uptake and oxidation by mitochondria.
LPIN2	rs7980	downstream_gene_variant	1 Het 3 Hom	2 Het 2 Hom	Potential role in triglyceride metabolism and lipodystrophy. Fatty Acy-CoA biosynthesis.
ECHDC3	rs7899215	structural_interaction_variant	5 Hom	5 Hom	Enoyl-CoA Hydratase Domain Containing 3, related pathways include fatty acid biosynthesis.
ACADS	rs3914 & rs1799958	structural_interaction_variant	1 Hom 2 Het	1 Hom	Member of the acyl-CoA dehydrogenase family. This enzyme catalyzes the initial step of the mitochondrial fatty acid beta-oxidation pathway. Associated with SCAD deficiency (lipid storage myopathy), resulting in short-chain fatty acids not metabolizing properly.

Table 2.8 – **Fatty acid synthesis.**

2.3.2.6 Ubiquitin related

Depletion of cellular cholesterol has been shown to facilitate ubiquitylation of NPC1, while inhibition of the disassembly of the ESCRT complex resulted in aggregation of ubiquitylated NPC1 protein (Ohsaki et al., 2006). We identified the ubiquitin ligase MARCH8 gene with homozygous and heterozygous missense variants (residue change tyrosine to histidine) segregating in 4 of 5 sib-pairs, and variants in three genes that were abundant in NP-C patients compared to control populations (RAD23B, UHRF1, USP29). The family of membrane-bound E3 ubiquitin ligases includes MARCH8, a gene that has been shown to ubiquitinate the transferrin receptor (TfR) resulting in the subsequent lysosomal degradation of TfR (Fujita et al., 2013). Also potentially related to NP-C disease for which the HDAC inhibitor Vorinostat is a candidate therapy (Munkacsi et al., 2016), it is curious to see a structural interaction variant present in the RING-finger type E3 ubiquitin ligase UHRF1 involved in the recruitment of histone deacetylases.

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
RAD23B	rs3056494 & rs539541320 & rs5899731	frameshift_variant	Multiple Bi-allelic	Multiple Bi-allelic	Human orthologue of *S.cerevisiae* Rad23. Encoded protein elevates nucleotide excision activity of 3-methyladenine-DNA glycolase, and interacts with 26S proteasome, with possible role in ubiquitin mediated proteolytic pathway.
UHRF1	rs2123731	structural_interaction_variant	5 Hom	5 Hom	Recruits histone deacetylase to regulate gene expression. RING-finger type E3 ubiquitin ligase.
USP29	rs9973206	stop_gained	4 Hom	1 Het 4 Hom	USP29 has related pathways in Ubiquitin-Proteasome 2Dependent Proteolysis.
MARCH8	rs7908745	missense	5 Hom	4 Het 1 Hom	MARCH family of membrane-bound E3 ubiquitin ligases.

Table 2.9 – **Ubiquitin related.**

2.3.2.7 Calcium ion binding

Calcium, as an important signalling molecule for eukaryotic cells and it specifically has been shown to play an important role in the pathogenesis of neurological disorders including NP-C disease (Lloyd-Evans et al., 2008; Mattson and Chan, 2003). In neuronal models of Gaucher disease, accumulation of the sphingolipid glucosylceramide resulted in cultured neurons exhibiting altered calcium release, as well as elevated phosphatidylcholine (Bodennec et al., 2002). We identified variants in two calcium-related genes that were abundant in NP-C patients compared to control populations (Table 2.10). DPYSL2 is a member of the collapsin response mediator family involved in synaptic signalling via interaction with calcium, for which SNP variants have been associated with Alzheimer's disease (Lambert et al., 2013). PKD1L2 encodes a polycystin glycoprotein with a lipoxygenase, alpha-toxin domain involved in intracellular calcium homeostasis, where all 10 of our NP-C patients had a stop lost variant in this gene that results in a non-coding transcript variant.

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
DPYSL2	rs327222	structural_interaction_variant	5 Hom	5 Hom	Member of collapsin response mediator family. Plays a role in synaptic signalling through interactions with calcium channels. Diseases associated with DPYSL2 include Alzheimer's disease and schizophrenia.
PKD1L2	rs8054182	stop_lost	5 Hom	5 Hom	Polycystin protein. Includes a polycystin-1, lipoxygenase, alpha-toxin domain. Calcium ion binding.

Table 2.10 – **Calcium ion binding.**

2.3.2.8 Acetylcholine related

Acetylcholine is a neurotransmitter of interest to many neurodegenerative diseases (Holzgrabe et al., 2007; Tata et al., 2014). We identified 2 variants in acetylcholine-related genes (SLC18A3, CHAT). All patients in our cohort had two homozygous variants in SLC18A3, one missense variant and another 5' UTR variant. SLC18A3 is a member of the vesicular amine transporter family that transports acetylcholine into transport vesicles. Of particular note is that the 5' UTR variant was predicted by SnpEff to result in an upstream variant of the CHAT (Choline O-Acetyltransferase) gene, which also had a structural interaction variant in 9 of 10 patients. CHAT initiates the biosynthesis of acetylcholine for which genetic variation in this gene has previously been associated with increased Alzheimer's disease risk (Grünblatt et al., 2011; Ozturk et al., 2006).

Gene ID	dbSNP ID	Variant type	Zygosity in functional	Zygosity in severe	Description of encoded protein function
SLC18A3	rs1880675 & rs8187730	5_prime_UTR_variant & missense_variant	5 Hom	5 Hom	Member of the vesicular amine transporter family. Encoded protein transports acetylcholine into secretory vesicles. Results in upstream gene variant of CHAT (see below).
CHAT	rs8178992	structural_interaction_variant	5 Hom	4 Hom	Choline O-Acetyltransferase catalyzes the biosynthesis of the neurotransmitter acetylcholine. Polymorphisms in this gene have been associated with Alzheimer's disease and cognitive impairment.

Table 2.11 – **Acetylcholine related.**

2.3.3 MMP12 inhibitor reduces cholesterol accumulation in primary neurons of *Npc1*$^{-/-}$ mice

Given the association between MMP12 and lung disease (Li et al., 2009b) as well as the upregulation of MMP12 in NP-C astrocytes (Liao et al., 2015), we examined whether chemical inhibition of MMP12 would affect the cholesterol aggregation that is the hallmark of NP-C disease. These results would further test our identification of the rs5794199 variant in MMP12 on chromosome 11. As the *de-facto* method for diagnosing the classical NP-C phenotype, fluorescence arising from UV excitation of filipin staining unesterified cholesterol was used as a biological marker for NP-C disease. If unesterified cholesterol is aggregated within the neuronal cells, fluorescence will be greater than that observed in control cells.

To examine any potential effects of a loss of function arising from the frameshift insertion rs5794199 in MMP12 (Table 2.7), we treated primary murine neuronal cells with 500 nM MMP408, a selective MMP inhibitor for MMP12 (Li et al., 2009b). If treatment rescues the cholesterol accumulation that is the hallmark phenotype of NP-C, as measured by fluorescence levels of the filipin stain, we would expect a decrease in fluorescence. To quantify the relative levels of fluorescence, we normalized each category to the highest integrated density value as reported by FIJI (Schindelin et al., 2012), divided by cell number. These arbitrary fluorescent units (y-axis) were obtained from the average of 14 view-fields with biological triplicates. A ~40% reduction ($p < 0.001$) in relative fluorescence was observed in *Npc1*$^{-/-}$ murine neuronal cells after MMP408 treatment for 24 hours (Figure 2.19). A similar reduction in fluorescence (~50%, $p < 0.001$) was observed in the U18666A treatment of control neurons (Figure 2.20), which is an established pharmacological mimic of NP-C disease (Lu et

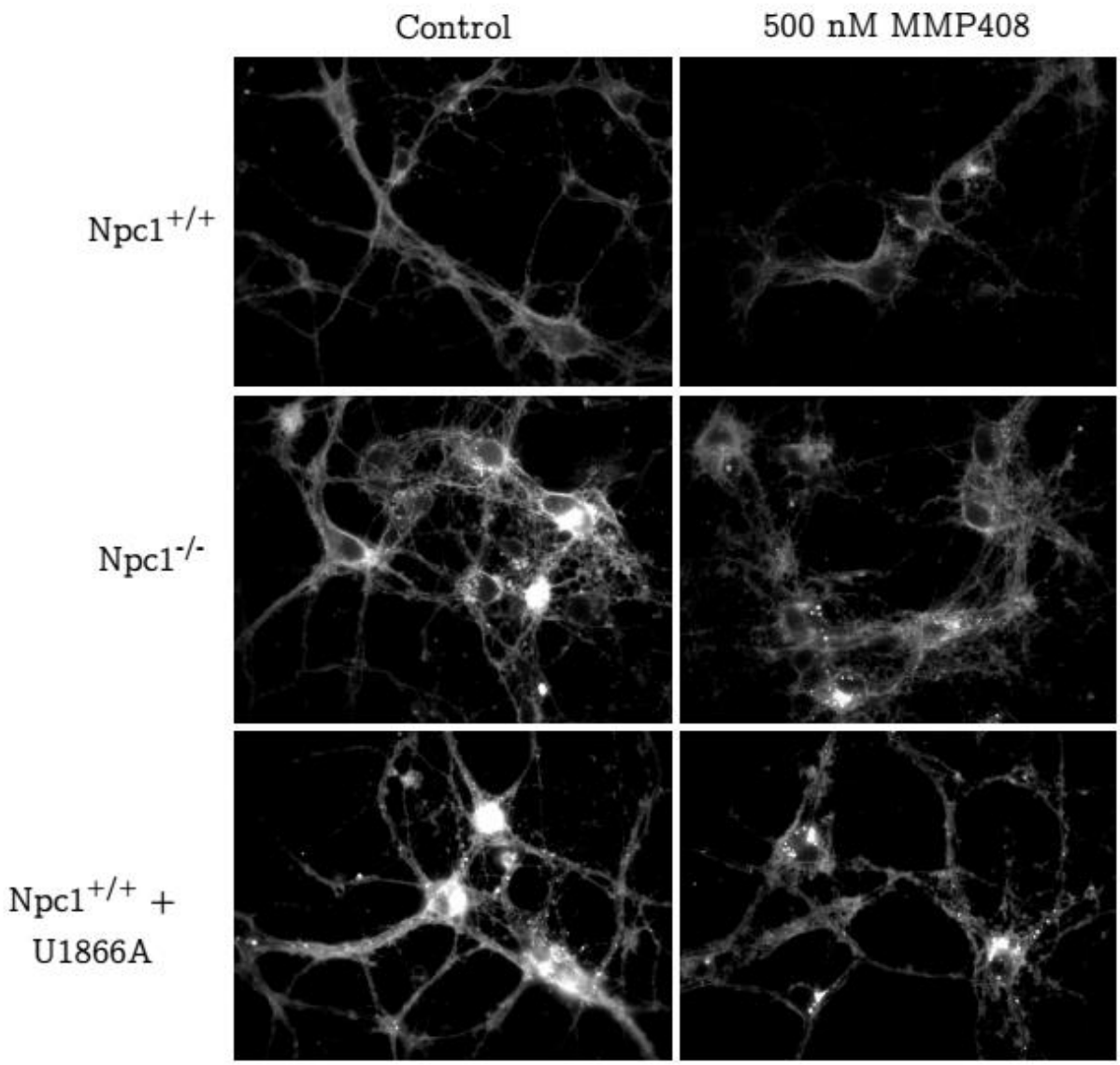


Figure 2.18 - **Flipin staining for un-esterifed cholesterol of murine primary neurons.** *WT, NP-C1 disease and U18-treatment groups, treated and untreated with a general MMP inhibitor. The MMPi groups showed decreased fluorescence compared with controls.*

al., 2015). These results suggest that NPC1 deficiency results in a pathological over-abundance or activity of MMP12, such that selective inhibition ameliorates cellular cholesterol levels.

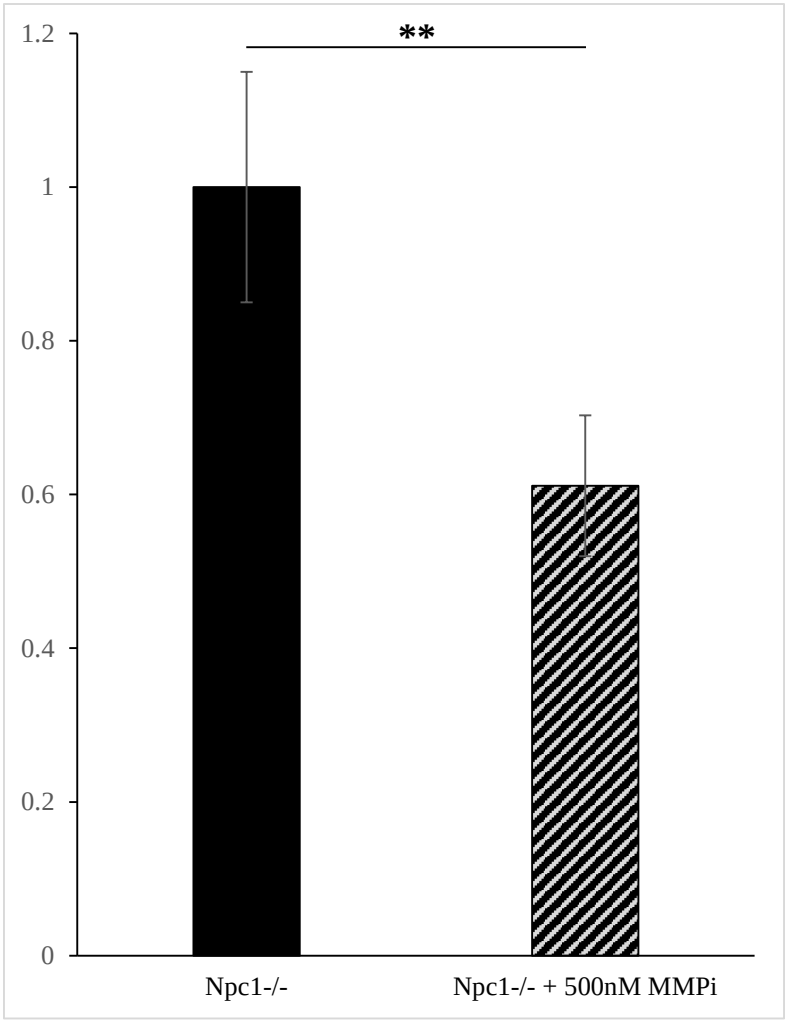


Figure 2.19 – **Quantification filipin (unesterified cholesterol) in Npc1-/- and Npc1-/- in the presence and absence of MMP408 treated fluorescence.**

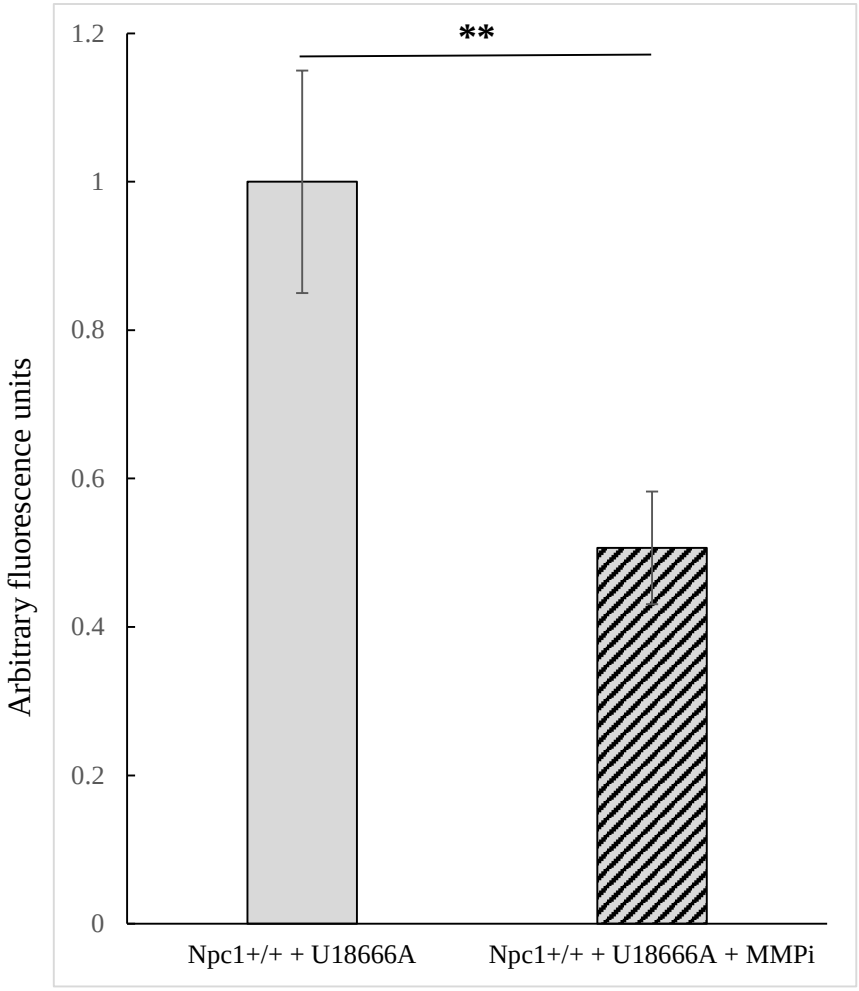


Figure 2.20 – **Quantification filipin (unesterified cholesterol) in U18 and U18 in the presence and absence of MMP408 treated fluorescence.**

2.3.3.1 Copy number variation at the MMP cluster on chromosome 11

As a preliminary investigation into our observation of the decreased fluorescence after treatment with MMP408, Control-FREEC, a program with the potential to detect copy number variation through depth of coverage analysis (Krumm et al., 2012), was used to calculate a B allele profile and estimate copy number alterations within the MMP12 region on chromosome 11. Given that we observed a rescue with MMP12 inhibition, we hypothesised that MMP12 was upregulated. Unfortunately, the signal to noise ratio precluded any conclusion from our copy number variation analysis (Figure 2.21).

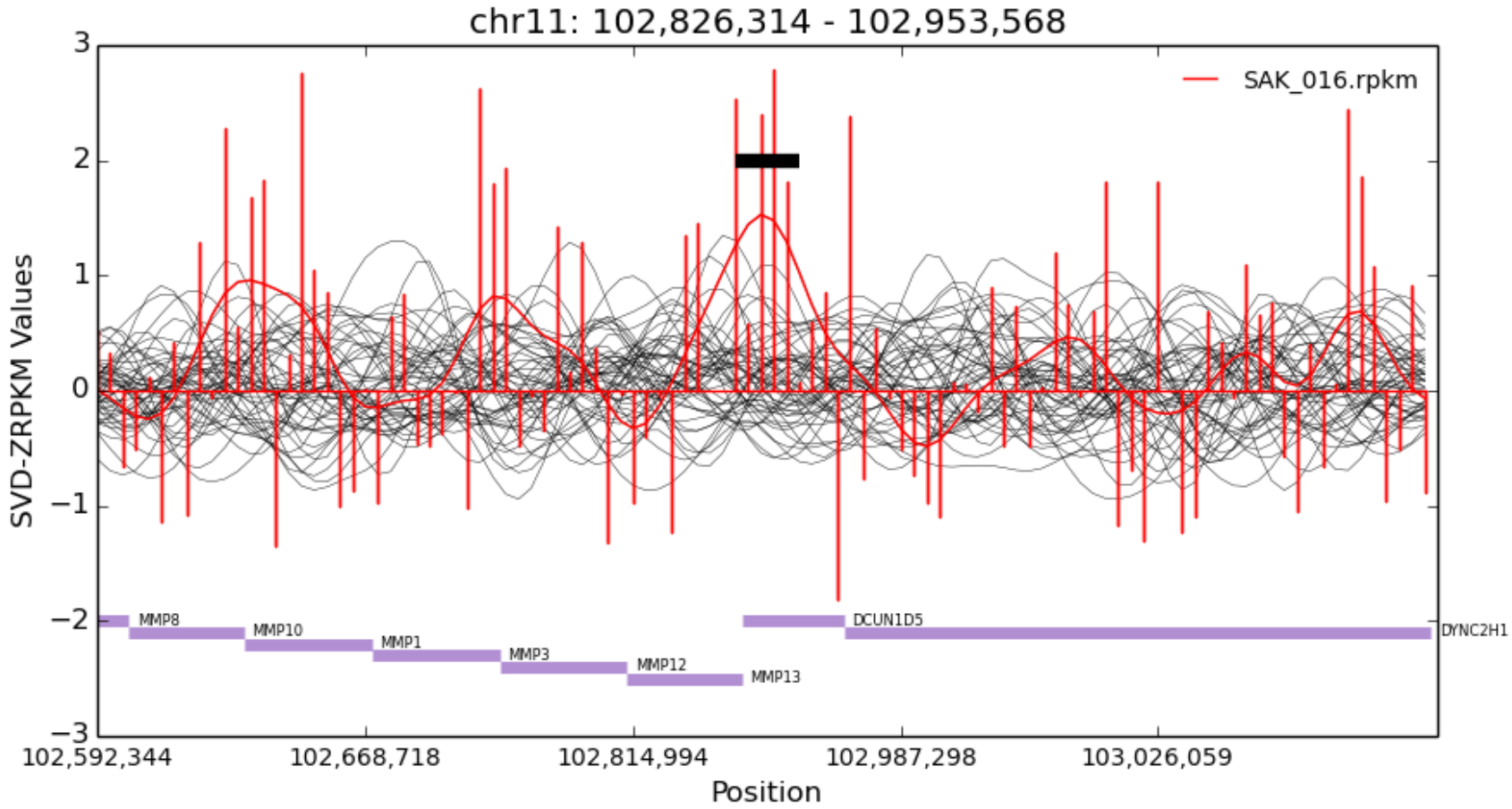


Figure 2.21 – **Conifer depth of coverage copy number profile around MMP cluster on chromosome 11.** Conifer returns discrete values – the red vertical lines are the actual values that represent the normalized and relative CNV of the exon in focus. The red curve is a loose fitting to the discrete values, while the grey curves are the other libraries. The black bar purportedly shows significance, but the signal to noise ratio is too weak to conclusively determine if there is CNV present at that location.

2.3.3.2 MARCH8 T95P SNP is conserved across multiple vertebrae taxa

We next examined sequence conservation on the rationale that conserved sequences have a higher probability of being deleterious if disrupted by a non-synonymous variant (Chun and Fay, 2009). In the aggregate model of NP-C patients, we identified a T95P variant in MARCH8, which is located within a highly conserved region of the RING-CH structure across multiple vertebrates (Figure 2.22). This region has been shown to be essential for the downregulation of transferrin receptor (TfR) via ubiquitination (Fujita et al., 2013), further strengthening the connection with NP-C disease for which TfR recycling is perturbed (Choudhury et al., 2004; McCaffrey et al., 2001). Taken together, we would predict that the T95P MARCH8 variant is likely to affect the functional activity of the MARCH8 protein and possibly modify the progression of NP-C disease.

```
MARCH8
                                                T95P
                                                ^
H.sapiens            351   PISPVSTSGDVCRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   400
P.troglodytes        351   PISPVSTSGDVCRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   400
M.mulatta            351   PISPVSTSGDVCRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   400
C.lupus              347   PVSPVSTSGDACRICHCEGDEESPLITPCRCTGSLHFVHQTCLQQWIKSS   396
B.taurus              78   -----------CRICHCEGDDESPLITPCRCTGSLHFVHQTCLQQWIKSS   116
M.musculus            76   -----------CRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   114
R.norvegicus          76   -----------CRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   114
G.gallus             346   PLSPVSASGDTCRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   395
D.rerio               75   -----------CRICHCEGDDESPLITPCHCTGSLRFVHQACLQQWIKSS   113
X.tropicalis          47   -----------CRICHCEGDDESPLITPCHCTGSLHFVHQACLQQWIKSS   85
```

Figure 2.22 – **SNP within conserved C-terminal membrane of MARCH8, that interacts with TfR.**

2.3.3.3 T95P mutation is in the conserved RING domain of MARCH8

To examine the potential for functional changes to the protein structure of MARCH8, the biochemical algorithms library (Hildebrandt et al., 2010) and visualization suite BALLView (Moll et al., 2005) were employed on the purified MARCH8 protein (PDB ID 2D8S). Cursory modelling of electrostatic forces showed that the orientations of the hydrogen bonds are not altered after introduction of the T95P mutation (Figure 2.23). Near the region of interest, T95P, the H-bonds were directed outwards from the conserved RING-CH structure (represented as the white backbone). This finding suggests that the overall function of the MARCH8 protein is unchanged as the binding pocket is unchanged, however there is still the potential for altered protein-protein or protein-cargo interactions along with the potential for additional intramolecular interactions. In addition to the T95P mutation, we also identified an Y226H mutation in MARCH8, but were not able to model structural changes arising from this mutation as the collection of solutions to the NMR structure had poor resolution outside the core protein and the Y226H mutation is outside this region (Figure 2.24). It is of note that the T95P mutation is upstream of a variant in the MARCH8-ALOX5 locus (rs970548) that has been associated with cholesterol regulation (Global Lipids Genetics Consortium et al., 2013). Figure 2.24B also reinforces the strong evolutionary conservation within this sequence region which is further denoted with a comparison of MARCH8 and MARCH1 (Figure 2.24), albeit we did not identify any variants in MARCH1 in any of our NP-C patients.

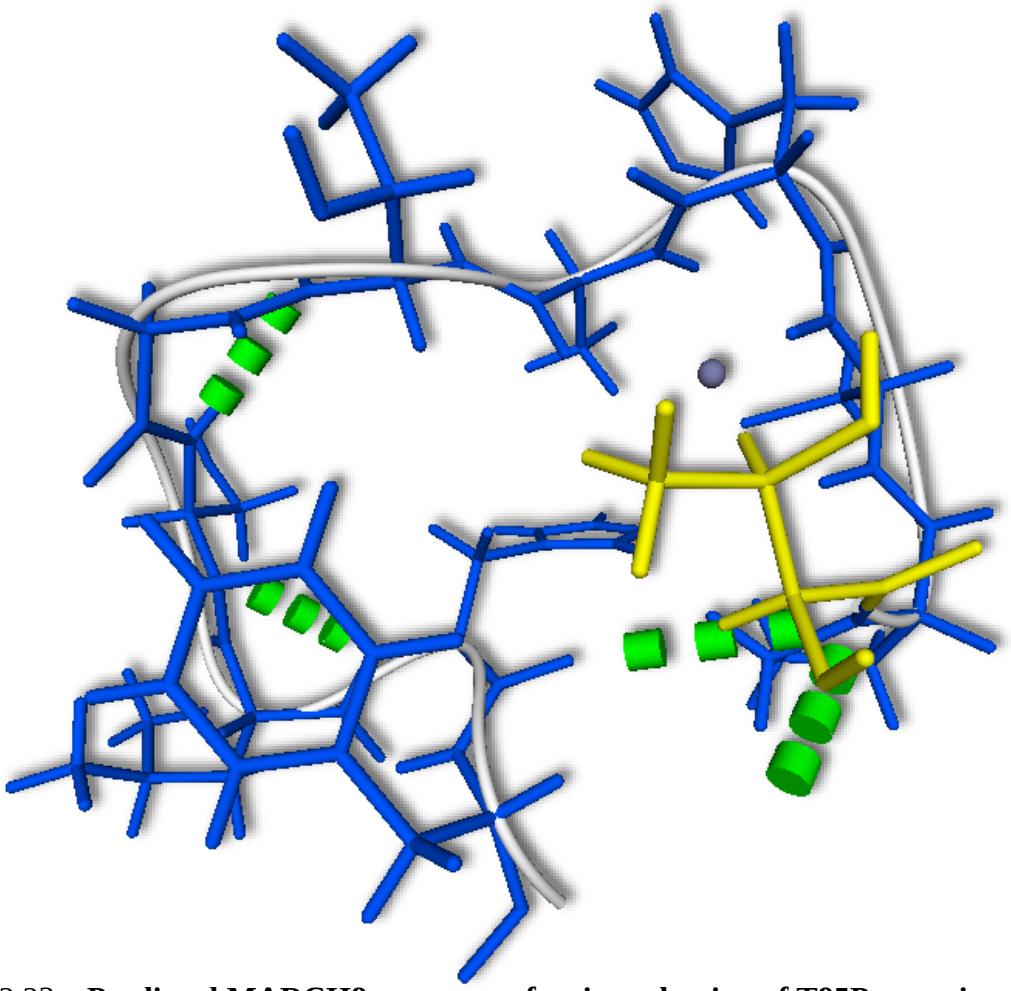
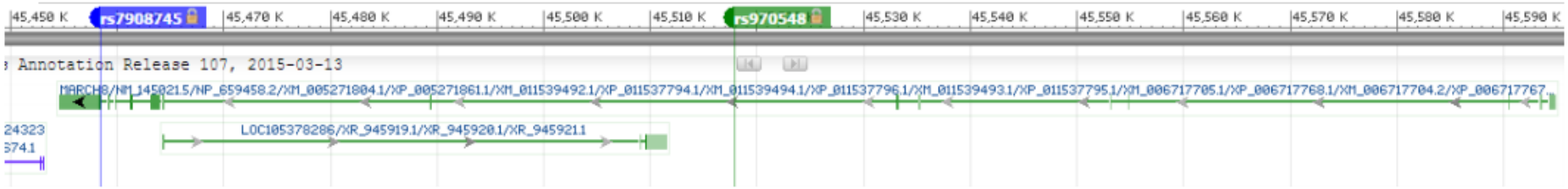


Figure 2.23 – **Predicted MARCH8 structure after introduction of T95P mutation.** Predicted structure from Biochemical Algorithms Library 1.4. Visualization from BALLView. Green cylinders represent hydrogen bonds, grey sphere is a zinc ion within binding pocket. Yellow backbone represents location of T95P SNP. Y226H not shown as it is within an unresolvable section of the protein structure data.

A



B

MSMPLHQISA IPSQDAISAR VYRSKTKEKE REEQNEKTLG HFMSHSSNIS KAGSPPSASA
PAPVSSFSRT SITPSSQDIC RICHCEGDDE SPLI**<<T95P>>**PCHCT GSLHFV*HQAC* LQQWIKSSDT
RCCELCKYEF IMETKLKPLR KWEKLQMTSS ERRKIMCSVT FHVIAITCVV WSLYVLIDRT
AEEIKQGQAT GILEWPFWTK LVVVAIGFTG GLLFMYVQCK VYVQLWKRLK AYNRVIYVQN
CPETSKKNIF EKSPLTEPNF ENKHG**<<Y266H>>**GICH SDTNSSCCTE PEDTGAEIIH V

C

MARCH8	P S S Q D I C R I C H C E G D D E S P L I T P C H C T G S L H F V H Q A C L Q Q W I K S S D T
MARCH1	P S T Q D I C R I C H C E G D E E S P L I T P C R C T G T L R F V H Q S C L H Q W I K S S D T

Figure 2.24 – **Single nucleotide variants and their predicted amino acid changes within MARCH8.** A) Location of SNV within MARCH8 found from WES analysis (rs7908745) compared with location of MARCH8-ALOX5 marker (rs970548) (Global Lipids Genetics Consortium et al., 2013). Non-synonymous variant early on in sequence is more likely to cause a non-functional protein. B) Amino acid sequence of MARCH8. Red text denotes homologous regions between MARCH8 and MARCH1 terminal domains. Blue text indicates transmembrane domains. Underlined regions are reported as responsible for association with TfR, while italicized residues are implicated in formation of protein structure (Fujita et al., 2013). Our reported variants are signified between <<>>, with reference amino acid, location and then amino acid change. C) Sequence comparison betweenMARCH8 and MARCH1.

2.3.4 MUC5B is significantly associated to disease severity

Given the clear segregation found in the aggregate model between severe and functional sib-pairs for the MUC5B variant rs4963031, we used PLINK to perform an exome-wide association test between the discordant sib-pairs with clinical severity scores for NP-C disease as statistical covariates. We found a significant association between disease severity and two variants (Figure 2.25). One gene was MUC5B, which we previously identified in the prior analysis (Table 2.6). The other gene was the Sialic Acid-Binding Immunoglobulin-Like Lectin 1 (SIGLEC1), a member of the immunoglobulin superfamily expressed by tissue macrophages (York et al., 2007). The results are highly suggestive that MUC5B is indeed a potential modifier of NP-C disease severity, as two separate forms of analysis (association testing with disease severity and segregation within sib-pairs) resolved MUC5B as a potential modifier out of a pool of 22,000.

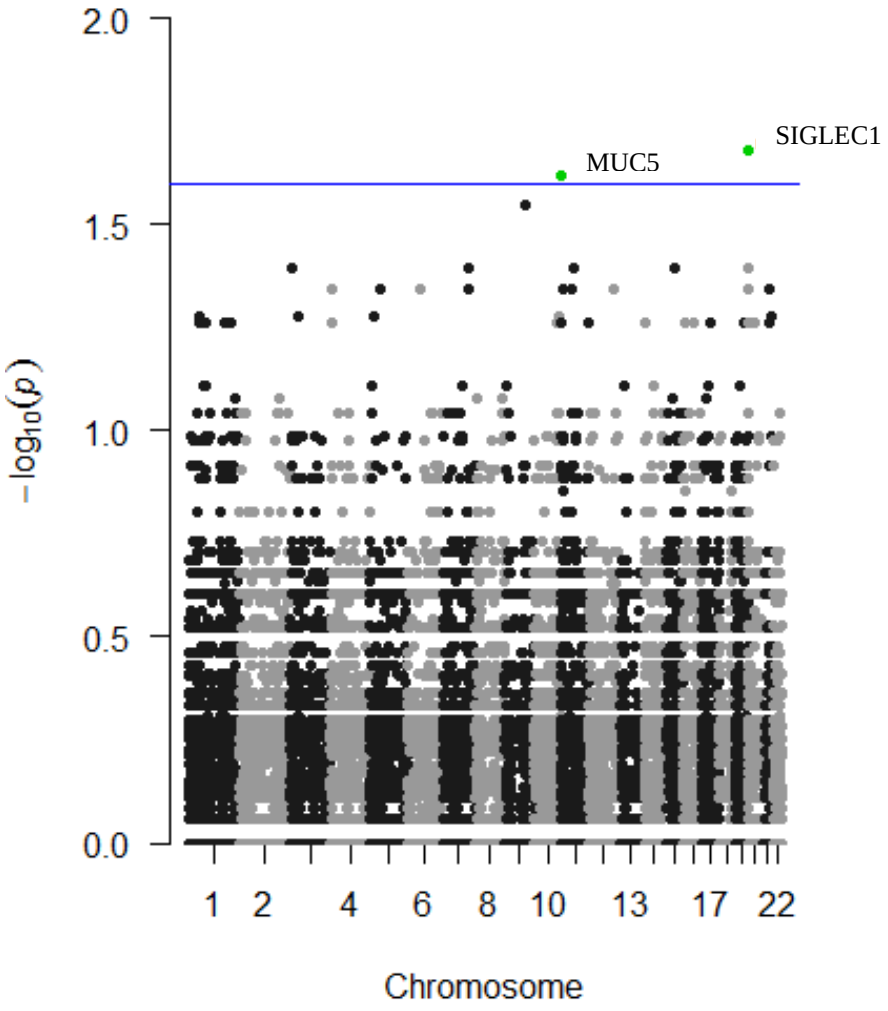


Figure 2.25 – **Manhattan plot of weakly associated genes to disease severity.**

2.4 Discussion

NP-C is a deadly neurological and systemic disorder lacking an FDA approved treatment. Being a paediatric disease, it is devastating to the families involved and a timely cure is required to alleviate the suffering caused by this disease. Progress in this area has been made (Davidson et al., 2009; Helquist et al., 2013; Munkacsi et al., 2016; Pipalia et al., 2011), however there is still much to be elucidated with respect to the mechanisms of rescue. In this context, discovery of genetic modifiers to NP-C by exome sequencing is an extension to the methods described previously (Cirulli et al., 2010; Cooper and Shendure, 2011; Alazami et al., 2015; Esslinger et al., 2017). Sib-pairs with divergent disease severity were sequenced with an Illumina HiSeq2000 with an average of ~30 Mb per patient. Filtering the aggregated patient cohort resulted in 26 variants, while homozygous variants that segregated across sib-pairs resulted in 2 variants.

2.4.1 Multifactorial model of modifiers in NP-C

Throughout our filtering processes, there was no set of criteria that reduced our pool of variants down to one modifier for all 10 NP-C patients, as in the case with discovery of causative genes (*e.g.*, Bamshad et al., 2010). However, this is to be expected, given the differences in discovery between causative genes and modifiers of genes (Cutting, 2010). Complex disease phenotypes are unlikely to be explained by a single genetic modifier because within a cohort of sibling pairs, the potential modifier variant in one sib-pair is likely to be unique and not found in the others.

One of these applied filters was minor allele frequency set at a threshold of 5%, or only looking at rare variants found in 5% or less of the population. This decision to look purely at rare variants was for several reasons. Firstly, there is some biological reasoning behind the assumption that common variants are less likely to associate with disease severity, decreasing

our statistical power (Morris and Zeggini, 2010). Furthermore, in order to correctly identify a common causal variant modifier as such, we would need either a large sample size (out of the question for rare disease research) or a very clear effect size arising from distinct segregation of the variant in question between sibling pairs, which is diminished with inclusion of common variants (Bansal et al., 2010).

An alternative approach to the filtering of variants would involve assuming the modifiers are common (*i.e.* MAF > 0.05) but only modifiers in NP-C patients due to an interaction with NPC1. This alternative approach of not excluding variants based off allele frequency would be challenging from a statistical perspective (grouping un-related SNPs can underpower your study), but approaches have been previously described (Bansal et al., 2010; Stitziel et al., 2011).

The collection of variants identified suggest that it is likely that disease severity in NP-C arises from a multifactorial cause based on variant burden, as opposed to a single genetic modifier. We can see this in the recovery of seven clear functional categories, with variants within having overlapping pathways or sites of interest. For example, MUC5B having association with pulmonary fibrosis (Peljto et al., 2013; Seibold et al., 2011) or ubiquitin-related MMP cellular invasion (Eisenach et al., 2012). The burden of each variant and its interaction with NP-C could explain the heterogeneous presentation of clinical symptoms. Precedence of interacting modifiers has been described previously with the discovery of interacting modifiers CAV2 and TMC6 variants that were significantly associated with risk for age of onset of chronic airway infection in cystic fibrosis (Emond et al., 2015). Emond et al. demonstrates the importance of using publically available exome sequence datasets (such as ExAC, (Lek et al., 2016)) as population controls to enable sufficient statistical power.

2.4.2 MMP-12 as a potential confounder in NP-C

Given the reported dysregulation within pulmonary tissue throughout NP-C disease progression (Griese et al., 2010; Roszell et al., 2013; Sheth et al., 2017), we then examined variants with a potential role in pulmonary tissue and NP-C. To begin, MMP12 protein has fibronectin as a substrate and is associated with tissue remodelling (Arikan et al., 2005). More pertinently, MMP12 has been shown to be essential for the development of emphysema in murine models (Hautamaki et al., 1997), as well as having an association with COPD (Li et al., 2009b). Specifically related to NP-C disease, microarray expression analysis with validation by qPCR confirmed increased expression of MMP12 in *Npc1*$^{-/-}$ mice liver tissue compared with controls (Cluzeau et al., 2012). Over-expression of immunity genes in *Npc*$^{-/-}$ mice has also been reported (with expression changes as high as ~80 fold), including MMP12 (Alam et al., 2012). This abundance of expressed MMP12 has been linked to axonal degeneration within a murine model of NP-C disease (Liao et al., 2015).

As 10 of 10 patients had a homozygous variant in MMP12, it cannot be the sole modifier of NP-C disease severity, as it did not segregate across the discordant sibling pairs. Furthermore, the association study did not detect MMP12 as potentially associated with disease severity, although this could have been a result from the study being underpowered (Bansal et al., 2010; Morris and Zeggini, 2010). These results suggest a potential role for members of the matrix metalloproteinase family, especially MMP12, in the outcomes for disease in relevant disease tissue of NP-C1 patients. Given the previous findings that MMP12 is upregulated in NP-C disease (Alam et al., 2012; Cluzeau et al., 2012; Liao et al., 2015) and our demonstration that MMP12 inhibition reduced the cholesterol accumulation in NP-C neurons (Figure 2.18), future work should examine the potential role the MMPs have on fibrotic remodelling of the diseased lung.

2.4.3 Role of pulmonary tissue related genes in NP-C

The recurring identification of genes with potential roles in pulmonary tissue (in particular, MMP12 and MUC5B) leads to the question – are these suitable targets for treatment? MMP12 deficiency has been shown to ameliorate the clinical outcome of COPD (Li et al., 2009b) in pulmonary tissue, however it has also been used as a target to treat demyelination within multiple sclerosis (Hansmann et al., 2012) and axonal degeneration caused by NPC1 deficiency (Liao et al., 2015). MUC5B has a common polymorphism associated with both idiopathic pulmonary fibrosis and pneumonia (Jiang et al., 2015; Seibold et al., 2011) that affects susceptibility and severity.

It is likely that treatment that targets one or both of these genes would aid in preventing death and/or severity in NP-C, but would not resolve the underlying issues arising from sphingolipid and secondary metabolite accumulation. It would be particularly informative to identify modifiers of the pulmonary aspects of NP-C disease since NP-C patients often die from pulmonary complications (*e.g.*, pneumonia). Our result showing reduced cholesterol accumulation in primary neurons suggests that MMP12 inhibition would be therapeutic in the brain and likely also in the lungs.

2.4.4 MARCH8 variants may mediate TfR defects in NP-C

MARCH8 is a gene of interest from our whole exome sequencing sibling pair analysis. Candidate genes from five pairs of siblings (n=10) were compared, and variants unique to each sibling were isolated. These unique variants were then compared against the unique variants of the individual in the whole sample pool. From this, five of five sibling pairs were found to have the MARCH8 variant, four of five of which were segregated between disease severities. In particular is that for those patients whose clinical outcome was known, all less clinically severe siblings had the MARCH8 variant.

The discovered T95P mutation in NP-C patients, while an interesting finding to report, requires a wealth of follow-up work to validate and confirm any potential role this variant in MARCH8 could have within clinical outcomes of NP-C disease. From an informatics perspective, much could be done to take a highly resolved MARCH8 protein structure and assess the possible influence the T95P variant could have on MARCH8 structure via detailed *in-silico* folding experiments. Furthermore, biological validation via immunohistochemistry to confirm the presence of the variation, or assessing the effect that recapitulating the SNP would have on ferritin levels within a cellular model of either NP-C1 or GD (Blendy, 2011).

What role could the transferrin receptor and associated pathways have with respect to the clinical outcomes in NP-C disease? To begin with, high levels of endosomal cholesterol in NP-C1 have been shown to disturb rab4-mediated recycling, drastically increasing transferrin recycling rates (Choudhury et al., 2004; Devlin et al., 2010; McCaffrey et al., 2001). Transferrin is necessary for iron delivery (Aisen, 2004) and plays a role in the production of ferritin. From this, one would expect a reduction in ferritin levels within tissue most affected by NP-C pathology and this is indeed the case (Christomanou et al., 2000).

Furthermore, it is known that diseases of cholesterol metabolism are similar in sometimes surprising ways (Platt et al., 2014). With this in mind, Gaucher disease (GD), another LSD that displays diverse clinical phenotypes between sibling pairs, shows aberrant ferritin (hyperferritinemia) levels (Lo et al., 2012). Intracellular iron concentration controls TfR expression levels via TfR mRNA stability (Fujita et al., 2013), resulting in a biological pathway where both ferritin and transferrin are important in the context of cholesterol accumulation in NP-C1 (Figure 2.26, also (Argüello et al., 2014)).

It would be most interesting to elucidate the potential role this MARCH8 variant could play within the clinical outcome of NP-C1, within the context of TfR mediated iron recycling.

In light of our results and the reported literature, we propose a preliminary model for the potential role of MARCH8 within the TfR defects found in NP-C disease (Figure 2.26). What role could MARCH8 have in the interaction between TfR, iron homeostasis and NP-C1 disease outcomes? MARCH8 can ubiquitinate TfR, leading to the reduction of TfR protein via degradation within the lysosome (Fujita et al., 2013). However, a genome-wide association study indicated an association between MARCH8-ALOX5 (marker rs970548) and cholesterol regulation (Global Lipids Genetics Consortium et al., 2013). Given this, the finding of a SNV causing a significant hydrophobic change within the conserved region of the terminal CT domain (proline often functions as a helix disruptor), suggests an effect on the neighbouring residues implicated in the RING-CH structure, with potential effects on protein/cargo interactions. Furthermore, another SNV downstream of the region associated with TfR (Figure 2.24), suggests the possibility for the clinical relevance of these MARCH8 variants for NP-C.

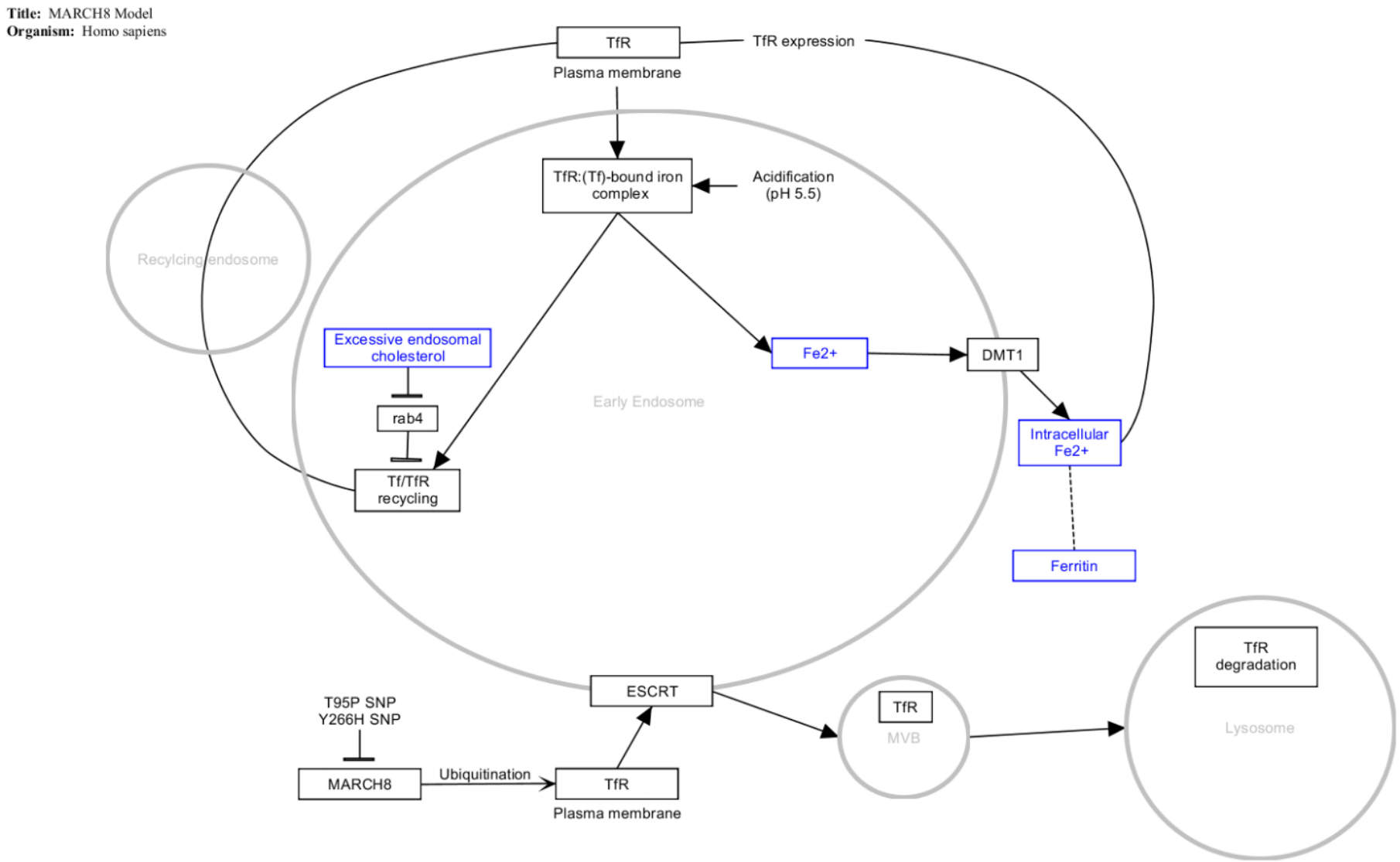


Figure 2.26 – **Proposed interaction of MARCH8 NP-C variants on TfR related pathways.** Cholesterol accumulation from NPC interfering with Rab4-mediated TfR/Tfn recycling rates potentially counter balanced by MARCH8 loss of function, increasing availability of TfR.

2.4.5 Future directions

In this thesis, only MMP12 had follow-up experiments performed in an attempt to validate the *in silico* findings from the exome sequences. Additional work is required to further assess the biological impact these reported sequence changes have on the NP-C disease phenotype. An approach for this would be utilizing the recent advances in CRISPR-Cas9 genome editing to recapitulate the identified SNPs in tissues of interest (neuronal, pulmonary, *etc.*) (Hsu et al., 2014; Konermann et al., 2015; Wang et al., 2014) and initially test for alterations in cholesterol accumulation in NP-C mutant cells (Vanier and Latour, 2015b). In addition to the NPC1-GeneX interactions that can be investigated using CRISPR-Cas9, this methodology can also be used to investigate more complicated interactions, such as interactions between more than two genes.

Another result that requires improvement is the findings from the Conifer CNV calling profile, with low signal to noise ratio in the calculation of copy number alteration via depth of coverage. However, this approach to accurately determine CNV is typically confounded by alterations to targeted exome capture efficiency or other factors that affect read distribution. Another copy-number caller, CopywriteR (Kuilman et al., 2015), resolves issues based on depth of coverage of captured exons by using the fact that off-target reads are uniform across the genome and using this to infer DOC and hence CNVs (Kuilman et al., 2015). Another approach is on target DOC calculations, such as Control-FREEC (Boeva et al., 2012). In brief, FREEC detects copy number alterations by construction of a B-allele frequency profile (here the B allele is defined as the alternative within the SNP database). This BAF profile can provide evidence for CNV either through loss of heterozygosity (indicating a loss event) or via allelic imbalance (indicating a copy number gain). This genomic information is then used to predict the copy number variation within each segment of the associated copy number profile. Using

other CNV/CNA callers with different approaches and combining their findings could resolve the structural variation in this region more accurately.

List of references

Abel, L.A., Bowman, E.A., Velakoulis, D., Fahey, M.C., Desmond, P., Macfarlane, M.D., Looi, J.C.L., Adamson, C.L., and Walterfang, M. (2012). Saccadic Eye Movement Characteristics in Adult Niemann-Pick Type C Disease: Relationships with Disease Severity and Brain Structural Measures. PLOS ONE *7*, e50947.

Aisen, P. (2004). Transferrin receptor 1. Int. J. Biochem. Cell Biol. *36*, 2137–2143.

Alam, M.S., Getz, M., Safeukui, I., Yi, S., Tamez, P., Shin, J., Velázquez, P., and Haldar, K. (2012). Genomic Expression Analyses Reveal Lysosomal, Innate Immunity Proteins, as Disease Correlates in Murine Models of a Lysosomal Storage Disorder. PLOS ONE *7*, e48273.

Alam, M.S., Getz, M., and Haldar, K. (2016). Chronic administration of an HDAC inhibitor treats both neurological and systemic Niemann-Pick type C disease in a mouse model. Sci. Transl. Med. *8*, 326ra23-326ra23.

Alazami, A.M., Patel, N., Shamseldin, H.E., Anazi, S., Al-Dosari, M.S., Alzahrani, F., Hijazi, H., Alshammari, M., Aldahmesh, M.A., Salih, M.A., et al. (2015). Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families. Cell Rep. *10*, 148–161.

Amin, N., Jovanova, O., Adams, H.H.H., Dehghan, A., Kavousi, M., Vernooij, M.W., Peeters, R.P., de Vrij, F.M.S., van der Lee, S.J., van Rooij, J.G.J., et al. (2017). Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. Mol. Psychiatry *22*, 537–543.

Appelqvist, H., Wäster, P., Kågedal, K., and Öllinger, K. (2013). The lysosome: from waste bag to potential therapeutic target. J. Mol. Cell Biol. *5*, 214–226.

Argüello, G., Martinez, P., Peña, J., Chen, O., Platt, F., Zanlungo, S., and González, M. (2014). Hepatic metabolic response to restricted copper intake in a Niemann-Pick C murine model. Met. Integr. Biometal Sci. *6*, 1527–1539.

Arikan, M.C., Shapiro, S.D., and Mariani, T.J. (2005). Induction of macrophage elastase (MMP-12) gene expression by statins. J. Cell. Physiol. *204*, 139–145.

Bamshad, M.J., Bigham, A.W., Buckingham, K.J., Dent, K.M., Huff, C.D., Jabs, E.W., Lee, C., Ng, S.B., Nickerson, D.A., Shannon, P.T., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30+.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. *11*, 773–785.

Bartz, F., Kern, L., Erz, D., Zhu, M., Gilbert, D., Meinhof, T., Wirkner, U., Erfle, H., Muckenthaler, M., Pepperkok, R., et al. (2009). Identification of Cholesterol-Regulating Genes by Targeted RNAi Screening. Cell Metab. *10*, 63–75.

Beckmann, J.S., Estivill, X., and Antonarakis, S.E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat. Rev. Genet. *8*, 639–646.

Blendy, J.A. (2011). Modeling Neuropsychiatric Disease-Relevant Human SNPs in Mice. Neuropsychopharmacology *36*, 364–365.

Bodennec, J., Pelled, D., Riebeling, C., Trajkovic, S., and Futerman, A.H. (2002). Phosphatidylcholine synthesis is elevated in neuronal models of Gaucher disease due to direct activation of CTP:phosphocholine cytidylyltransferase by glucosylceramide. FASEB J.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics *28*, 423–425.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma. Oxf. Engl. *30*, 2114–2120.

Bowers, K., Piper, S.C., Edeling, M.A., Gray, S.R., Owen, D.J., Lehner, P.J., and Luzio, J.P. (2006). Degradation of Endocytosed Epidermal Growth Factor and Virally Ubiquitinated Major Histocompatibility Complex Class I Is Independent of Mammalian ESCRTII. J. Biol. Chem. *281*, 5094–5105.

Brady, R.O., Kanfer, J.N., Mock, M.B., and Fredrickson, D.S. (1966). The metabolism of sphingomyelin. II. Evidence of an enzymatic deficiency in Niemann-Pick diseae. Proc. Natl. Acad. Sci. U. S. A. *55*, 366–369.

Brown, M.L., Ramprasad, M.P., Umeda, P.K., Tanaka, A., Kobayashi, Y., Watanabe, T., Shimoyamada, H., Kuo, W.-L., Li, R., Song, R., et al. (2000). A macrophage receptor for apolipoprotein B48: Cloning, expression, and atherosclerosis. Proc. Natl. Acad. Sci. *97*, 7488–7493.

Bunker, R.D., Bulloch, E.M.M., Dickson, J.M.J., Loomes, K.M., and Baker, E.N. (2013). Structure and Function of Human Xylulokinase, an Enzyme with Important Roles in Carbohydrate Metabolism. J. Biol. Chem. *288*, 1643–1652.

Carson, A.R., Smith, E.N., Matsui, H., Brækkan, S.K., Jepsen, K., Hansen, J.-B., and Frazer, K.A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics *15*, 125.

Choudhury, A., Sharma, D.K., Marks, D.L., and Pagano, R.E. (2004). Elevated endosomal cholesterol levels in Niemann-Pick cells inhibit rab4 and perturb membrane recycling. Mol. Biol. Cell *15*, 4500–4511.

Christomanou, H., Vanier, M.T., Santambrogio, P., Arosio, P., Kleijer, W.J., and Harzer, K. (2000). Deficient ferritin immunoreactivity in tissues from niemann-pick type C patients:

extension of findings to fetal tissues, H and L ferritin isoforms, but also one case of the rare Niemann-Pick C2 complementation group. Mol. Genet. Metab. *70*, 196–202.

Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res. *19*, 1553–1561.

Chung, C., Elrick, M.J., Dell'Orco, J.M., Qin, Z.S., Kalyana-Sundaram, S., Chinnaiyan, A.M., Shakkottai, V.G., and Lieberman, A.P. (2016). Heat Shock Protein Beta-1 Modifies Anterior to Posterior Purkinje Cell Vulnerability in a Mouse Model of Niemann-Pick Type C Disease. PLOS Genet. *12*, e1006042.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

Cirulli, E.T., Singh, A., Shianna, K.V., Ge, D., Smith, J.P., Maia, J.M., Heinzen, E.L., Goedert, J.J., and Goldstein, D.B. (2010). Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biol. *11*, R57.

Cluzeau, C.V.M., Watkins-Chow, D.E., Fu, R., Borate, B., Yanjanin, N., Dail, M.K., Davidson, C.D., Walkley, S.U., Ory, D.S., Wassif, C.A., et al. (2012). Microarray expression analysis and identification of serum biomarkers for Niemann–Pick disease, type C1. Hum. Mol. Genet. dds193.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. *38*, 1767–1771.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Cotman, S.L., and Staropoli, J.F. (2012). The juvenile Batten disease protein, CLN3, and its role in regulating anterograde and retrograde post-Golgi trafficking. Clin. Lipidol. *7*, 79–91.

Crocker, A.C. (1961). The Cerebral Defect in Tay-Sachs Disease and Niemann-Pick Disease*. J. Neurochem. *7*, 69–80.

Csepeggi, C., Jiang, M., Kojima, F., Crofford, L.J., and Frolov, A. (2011). Somatic Cell Plasticity and Niemann-Pick Type C2 Protein FIBROBLAST ACTIVATION. J. Biol. Chem. *286*, 2078–2087.

Cutting, G.R. (2010). Modifier genes in Mendelian disorders: the example of cystic fibrosis. Ann. N. Y. Acad. Sci. *1214*, 57–69.

Davidson, C.D., Ali, N.F., Micsenyi, M.C., Stephney, G., Renault, S., Dobrenis, K., Ory, D.S., Vanier, M.T., and Walkley, S.U. (2009). Chronic Cyclodextrin Treatment of Murine Niemann-Pick C Disease Ameliorates Neuronal Cholesterol and Glycosphingolipid Storage and Disease Progression. PLOS ONE *4*, e6951.

Del Valle, E.M.M. (2004). Cyclodextrins and their uses: a review. Process Biochem. *39*, 1033–1046.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Deutsch, G., Muralidhar, A., Le, E., Borbon, I.A., and Erickson, R.P. (2016). Extensive macrophage accumulation in young and old Niemann–Pick C1 model mice involves the alternative, M2, activation pathway and inhibition of macrophage apoptosis. Gene *578*, 242–250.

Devlin, C., Pipalia, N.H., Liao, X., Schuchman, E.H., Maxfield, F.R., and Tabas, I. (2010). Improvement in lipid and protein trafficking in Niemann-Pick C1 cells by correction of a secondary enzyme defect. Traffic Cph. Den. *11*, 601–615.

de Duve, C. (2005). The lysosome turns fifty. Nat. Cell Biol. *7*, 847–849.

Ebersberger, I., Metzler, D., Schwarz, C., and Pääbo, S. (2002). Genomewide Comparison of DNA Sequences between Humans and Chimpanzees. Am. J. Hum. Genet. *70*, 1490–1497.

Eisenach, P.A., de Sampaio, P.C., Murphy, G., and Roghi, C. (2012). Membrane Type 1 Matrix Metalloproteinase (MT1-MMP) Ubiquitination at Lys581 Increases Cellular Invasion through Type I Collagen. J. Biol. Chem. *287*, 11533–11545.

Emond, M.J., Louie, T., Emerson, J., Chong, J.X., Mathias, R.A., Knowles, M.R., Rieder, M.J., Tabor, H.K., Nickerson, D.A., Barnes, K.C., et al. (2015). Exome Sequencing of Phenotypic Extremes Identifies CAV2 and TMC6 as Interacting Modifiers of Chronic Pseudomonas aeruginosa Infection in Cystic Fibrosis. PLOS Genet. *11*, e1005273.

Esslinger, U., Garnier, S., Korniat, A., Proust, C., Kararigas, G., Müller-Nurasyid, M., Empana, J.-P., Morley, M.P., Perret, C., Stark, K., et al. (2017). Exome-wide association study reveals novel susceptibility genes to sporadic dilated cardiomyopathy. PLOS ONE *12*, e0172995.

Feng, T., Elston, R.C., and Zhu, X. (2011). A novel method to detect rare variants using both family and unrelated case-control data. BMC Proc. *5*, S80.

Fu, R., Yanjanin, N.M., Elrick, M.J., Ware, C., Lieberman, A.P., and Porter, F.D. (2012). ApolipoproteinE Genotype and Neurological Disease Onset in Niemann-Pick Disease, type C1. Am. J. Med. Genet. A. *158A*, 2775–2780.

Fujita, H., Iwabu, Y., Tokunaga, K., and Tanaka, Y. (2013). Membrane-associated RING-CH (MARCH) 8 mediates the ubiquitination and lysosomal degradation of the transferrin receptor. J. Cell Sci. *126*, 2798–2809.

Fuller, M., Meikle, P.J., and Hopwood, J.J. (2006). Epidemiology of lysosomal storage diseases: an overview. In Fabry Disease: Perspectives from 5 Years of FOS, A. Mehta, M. Beck, and G. Sunder-Plassmann, eds. (Oxford: Oxford PharmaGenesis), p.

Garver, W.S., Jelinek, D., Francis, G.A., and Murphy, B.D. (2008). The Niemann-Pick C1 gene is downregulated by feedback inhibition of the SREBP pathway in human fibroblasts. J. Lipid Res. *49*, 1090–1102.

Genin, E., Feingold, J., and Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease: strategies and difficulties. Hum. Genet. *124*, 357–368.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2011). Unlocking Mendelian disease using exome sequencing. Genome Biol. *12*, 228.

Global Lipids Genetics Consortium, Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.

Greer, W.L., Riddell, D.C., Gillan, T.L., Girouard, G.S., Sparrow, S.M., Byers, D.M., Dobson, M.J., and Neumann, P.E. (1998). The Nova Scotia (type D) form of Niemann-Pick disease is caused by a G3097-->T transversion in NPC1. Am. J. Hum. Genet. *63*, 52–54.

Griese, M., Brasch, F., Aldana, V., Cabrera, M., Goelnitz, U., Ikonen, E., Karam, B., Liebisch, G., Linder, M., Lohse, P., et al. (2010). Respiratory disease in Niemann-Pick type C2 is caused by pulmonary alveolar proteinosis. Clin. Genet. *77*, 119–130.

Griffin, L.D., Gong, W., Verot, L., and Mellon, S.H. (2004). Niemann–Pick type C disease involves disrupted neurosteroidogenesis and responds to allopregnanolone. Nat. Med. *10*, 704–711.

Grünblatt, E., Reif, A., Jungwirth, S., Galimberti, D., Weber, H., Scarpini, E., Sauer, C., Wichart, I., Rainer, M.K., Huber, K., et al. (2011). Genetic variation in the choline O-acetyltransferase gene in depression and Alzheimer's disease: The VITA and Milano studies. J. Psychiatr. Res. *45*, 1250–1256.

Gülhan, B., Özçelik, U., Gürakan, F., Güçer, Ş., Orhan, D., Cinel, G., Yalçın, E., Ersöz, D.D., Kiper, N., Yüce, A., et al. (2012). Different features of lung involvement in Niemann-Pick disease and Gaucher disease. Respir. Med. *106*, 1278–1285.

Hansmann, F., Herder, V., Kalkuhl, A., Haist, V., Zhang, N., Schaudien, D., Deschl, U., Baumgärtner, W., and Ulrich, R. (2012). Matrix metalloproteinase-12 deficiency ameliorates the clinical course and demyelination in Theiler's murine encephalomyelitis. Acta Neuropathol. (Berl.) *124*, 127–142.

Hautamaki, R.D., Kobayashi, D.K., Senior, R.M., and Shapiro, S.D. (1997). Requirement for Macrophage Elastase for Cigarette Smoke-Induced Emphysema in Mice. Science *277*, 2002–2004.

Hein, L.K., Meikle, P.J., Hopwood, J.J., and Fuller, M. (2007). Secondary sphingolipid accumulation in a macrophage model of Gaucher disease. Mol. Genet. Metab. *92*, 336–345.

Helquist, P., Maxfield, F.R., Wiech, N.L., and Wiest, O. (2013). Treatment of Niemann–Pick Type C Disease by Histone Deacetylase Inhibitors. Neurotherapeutics *10*, 688–697.

Hildebrandt, A., Dehof, A.K., Rurainski, A., Bertsch, A., Schumann, M., Toussaint, N.C., Moll, A., Stöckel, D., Nickels, S., Mueller, S.C., et al. (2010). BALL - biochemical algorithms library 1.3. BMC Bioinformatics *11*, 531.

Hilgenberg, L.G.W., and Smith, M.A. (2007). Preparation of Dissociated Mouse Cortical Neuron Cultures. J. Vis. Exp. JoVE.

Holzgrabe, U., Kapková, P., Alptüzün, V., Scheiber, J., and Kugelmann, E. (2007). Targeting acetylcholinesterase to treat neurodegeneration. Expert Opin. Ther. Targets *11*, 161–179.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Hu, H., Hübner, C., Lukacs, Z., Musante, L., Gill, E., Wienker, T.F., Ropers, H.-H., Knierim, E., and Schuelke, M. (2017). Klüver–Bucy syndrome associated with a recessive variant in HGSNAT in two siblings with Mucopolysaccharidosis type IIIC (Sanfilippo C). Eur. J. Hum. Genet. *25*, 253–256.

Imrie, J., Dasgupta, S., Besley, G.T.N., Harris, C., Heptinstall, L., Knight, S., Vanier, M.T., Fensom, A.H., Ward, C., Jacklin, E., et al. (2007). The natural history of Niemann–Pick disease type C in the UK. J. Inherit. Metab. Dis. *30*, 51–59.

Ishida, Y., Nayak, S., Mindell, J.A., and Grabe, M. (2013). A model of lysosomal pH regulation. J. Gen. Physiol. *141*, 705–720.

Iturriaga, C., Pineda, M., Fernández-Valero, E.M., Vanier, M.T., and Coll, M.J. (2006). Niemann–Pick C disease in Spain: Clinical spectrum and development of a disability scale. J. Neurol. Sci. *249*, 1–6.

Jahnova, H., Dvorakova, L., Vlaskova, H., Hulkova, H., Poupetova, H., Hrebicek, M., and Jesina, P. (2014). Observational, retrospective study of a large cohort of patients with Niemann-Pick disease type C in the Czech Republic: a surprisingly stable diagnostic rate spanning almost 40 years. Orphanet J. Rare Dis. *9*, 140.

Jiang, H., Hu, Y., Shang, L., Li, Y., Yang, L., and Chen, Y. (2015). Association between MUC5B polymorphism and susceptibility and severity of idiopathic pulmonary fibrosis. Int. J. Clin. Exp. Pathol. *8*, 14953–14958.

Kacher, Y., and Futerman, A.H. (2006). Genetic diseases of sphingolipid metabolism: Pathological mechanisms and therapeutic options. FEBS Lett. *580*, 5510–5517.

Kanno, K., Wu, M.K., Scapa, E.F., Roderick, S.L., and Cohen, D.E. (2007). Structure and function of phosphatidylcholine transfer protein (PC-TP)/StarD2. Biochim. Biophys. Acta *1771*, 654–662.

Kirkegaard, T., and Jäättelä, M. (2009). Lysosomal involvement in cell death and cancer. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1793*, 746–754.

Kirkegaard, T., Roth, A.G., Petersen, N.H.T., Mahalka, A.K., Olsen, O.D., Moilanen, I., Zylicz, A., Knudsen, J., Sandhoff, K., Arenz, C., et al. (2010). Hsp70 stabilizes lysosomes and reverts Niemann–Pick disease-associated lysosomal pathology. Nature *463*, 549–553.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Kristina Strandberg, A.K., and Salter, L.A. (2004). A comparison of methods for estimating the transition:transversion ratio from DNA sequences. Mol. Phylogenet. Evol. *32*, 495–503.

Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res. *22*, 1525–1532.

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., Nevedomskaya, E., Xu, G., de Ruiter, J., Lolkema, M.P., et al. (2015). CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol. *16*, 49.

Kwon, H.J., Abi-Mosleh, L., Wang, M.L., Deisenhofer, J., Goldstein, J.L., Brown, M.S., and Infante, R.E. (2009). Structure of N-terminal domain of NPC1 reveals distinct subdomains for binding and transfer of cholesterol. Cell *137*, 1213–1224.

Lambert, J.-C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A.L., Bis, J.C., Beecham, G.W., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.

Lebrand, C., Corti, M., Goodson, H., Cosson, P., Cavalli, V., Mayran, N., Fauré, J., and Gruenberg, J. (2002). Late endosome motility depends on lipids via the small GTPase Rab7. EMBO J. *21*, 1289–1300.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J., Deffieu, M.S., Lee, P.L., Saha, P., and Pfeffer, S.R. (2015). Glycosylation inhibition reduces cholesterol accumulation in NPC1 protein-deficient cells. Proc. Natl. Acad. Sci. *112*, 14876–14881.

Li, W., Li, J., Wu, Y., Wu, J., Hotchandani, R., Cunningham, K., McFadyen, I., Bard, J., Morgan, P., Schlerman, F., et al. (2009b). A Selective Matrix Metalloprotease 12 Inhibitor for Potential Treatment of Chronic Obstructive Pulmonary Disease (COPD): Discovery of (S)-2-(8-(Methoxycarbonylamino)dibenzo[b,d]furan-3-sulfonamido)-3-methylbutanoic acid (MMP408). J. Med. Chem. *52*, 1799–1802.

Li, X., Wang, J., Coutavas, E., Shi, H., Hao, Q., and Blobel, G. (2016). Structure of human Niemann–Pick C1 protein. Proc. Natl. Acad. Sci. *113*, 8212–8217.

Liao, G., Wang, Z., Lee, E., Moreno, S., Abuelnasr, O., Baudry, M., and Bi, X. (2015). Enhanced expression of matrix metalloproteinase-12 contributes to Npc1 deficiency-induced axonal degeneration. Exp. Neurol. *269*, 67–74.

Liu, B., Turley, S.D., Burns, D.K., Miller, A.M., Repa, J.J., and Dietschy, J.M. (2009). Reversal of defective lysosomal transport in NPC disease ameliorates liver dysfunction and neurodegeneration in the npc1−/− mouse. Proc. Natl. Acad. Sci. *106*, 2377–2382.

Lloyd-Evans, E., and Platt, F.M. (2010). Lipids on Trial: The Search for the Offending Metabolite in Niemann-Pick type C Disease. Traffic *11*, 419–428.

Lloyd-Evans, E., Morgan, A.J., He, X., Smith, D.A., Elliot-Smith, E., Sillence, D.J., Churchill, G.C., Schuchman, E.H., Galione, A., and Platt, F.M. (2008). Niemann-Pick disease type C1 is a sphingosine storage disease that causes deregulation of lysosomal calcium. Nat. Med. *14*, 1247–1255.

Lo, S.M., Choi, M., Liu, J., Jain, D., Boot, R.G., Kallemeijn, W.W., Aerts, J.M.F.G., Pashankar, F., Kupfer, G.M., Mane, S., et al. (2012). Phenotype diversity in type 1 Gaucher disease: discovering the genetic basis of Gaucher disease/hematologic malignancy phenotype by individual genome analysis. Blood *119*, 4731–4740.

Lowenthal, A.C., Cummings, J.F., Wenger, D.A., Thrall, M.A., Wood, P.A., and de Lahunta, A. (1990). Feline sphingolipidosis resembling Niemann-Pick disease type C. Acta Neuropathol. (Berl.) *81*, 189–197.

Lu, F., Liang, Q., Abi-Mosleh, L., Das, A., De Brabander, J.K., Goldstein, J.L., and Brown, M.S. (2015). Identification of NPC1 as the target of U18666A, an inhibitor of lysosomal cholesterol export and Ebola infection. eLife *4*.

Luzio, J.P., Pryor, P.R., and Bright, N.A. (2007). Lysosomes: fusion and function. Nat. Rev. Mol. Cell Biol. *8*, 622–632.

Luzon-Toro, B., Gui, H., Ruiz-Ferrer, M., Tang, S.M., Fernandez, R.M., Sham, P.C., Torroglosa, A., Tam, P.K.H., Espino-Paisan, L., Cherny, S.S., et al. (2015). Exome sequencing reveals a high genetic heterogeneity on familial Hirschsprung disease.

Malnar, M., Hecimovic, S., Mattsson, N., and Zetterberg, H. (2014). Bidirectional links between Alzheimer's disease and Niemann-Pick type C disease. Neurobiol. Dis. *72 Pt A*, 37–47.

Marcuzzi, A., Vozzi, D., Girardelli, M., Tricarico, P.M., Knowles, A., Crovella, S., Vuch, J., Tommasini, A., Piscianz, E., and Bianco, A.M. (2016). Putative modifier genes in mevalonate kinase deficiency. Mol. Med. Rep. *13*, 3181–3189.

Mattson, M.P., and Chan, S.L. (2003). Neuronal and glial calcium signaling in Alzheimer's disease. Cell Calcium *34*, 385–397.

Maxfield, F.R., and van Meer, G. (2010). Cholesterol, the central lipid of mammalian cells. Curr. Opin. Cell Biol. *22*, 422–429.

McCaffrey, M.W., Bielli, A., Cantalupo, G., Mora, S., Roberti, V., Santillo, M., Drummond, F., and Bucci, C. (2001). Rab4 affects both recycling and degradative endosomal trafficking. FEBS Lett. *495*, 21–30.

McInerney-Leo, A.M., Marshall, M.S., Gardiner, B., Coucke, P.J., Van Laer, L., Loeys, B.L., Summers, K.M., Symoens, S., West, J.A., West, M.J., et al. (2013). Whole exome sequencing is an efficient, sensitive and specific method of mutation detection in osteogenesis imperfecta and Marfan syndrome. BoneKEy Rep. *2*.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mellman, I. (1996). Endocytosis and Molecular Sorting. Annu. Rev. Cell Dev. Biol. *12*, 575–625.

Millat, G., Marçais, C., Rafi, M.A., Yamamoto, T., Morris, J.A., Pentchev, P.G., Ohno, K., Wenger, D.A., and Vanier, M.T. (1999). Niemann-Pick C1 Disease: The I1061T Substitution Is a Frequent Mutant Allele in Patients of Western European Descent and Correlates with a Classic Juvenile Phenotype. Am. J. Hum. Genet. *65*, 1321–1329.

Moll, A., Hildebrandt, A., Lenhof, H.-P., and Kohlbacher, O. (2005). BALLView: an object-oriented molecular visualization and modeling framework. J. Comput. Aided Mol. Des. *19*, 791–800.

Morris, A.P., and Zeggini, E. (2010). An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. Genet. Epidemiol. *34*, 188–193.

Munkacsi, A.B., Chen, F.W., Brinkman, M.A., Higaki, K., Gutiérrez, G.D., Chaudhari, J., Layer, J.V., Tong, A., Bard, M., Boone, C., et al. (2011). An "Exacerbate-reverse" Strategy in Yeast Identifies Histone Deacetylase Inhibition as a Correction for Cholesterol and Sphingolipid Transport Defects in Human Niemann-Pick Type C Disease. J. Biol. Chem. *286*, 23842–23851.

Munkacsi, A.B., Hammond, N., Schneider, R.T., Senanayake, D.S., Higaki, K., Lagutin, K., Bloor, S.J., Ory, D.S., Maue, R.A., Chen, F.W., et al. (2016). Normalization of Hepatic Homeostasis in the Npc1nmf164 Mouse Model of Niemann-Pick Type C Disease Treated with the Histone Deacetylase Inhibitor Vorinostat. J. Biol. Chem. jbc.M116.770578.

Ohsaki, Y., Sugimoto, Y., Suzuki, M., Hosokawa, H., Yoshimori, T., Davies, J.P., Ioannou, Y.A., Vanier, M.T., Ohno, K., and Ninomiya, H. (2006). Cholesterol depletion facilitates ubiquitylation of NPC1 and its association with SKD1/Vps4. J. Cell Sci. *119*, 2643–2653.

Ozturk, A., DeKosky, S.T., and Kamboh, M.I. (2006). Genetic variation in the choline acetyltransferase (CHAT) gene may be associated with the risk of Alzheimer's disease. Neurobiol. Aging *27*, 1440–1444.

Page-McCaw, A., Ewald, A.J., and Werb, Z. (2007). Matrix metalloproteinases and the regulation of tissue remodelling. Nat. Rev. Mol. Cell Biol. *8*, 221–233.

Park, S.T., and Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. Int. Neurourol. J. *20*, S76-83.

Park, W.D., O'Brien, J.F., Lundquist, P.A., Kraft, D.L., Vockley, C.W., Karnes, P.S., Patterson, M.C., and Snow, K. (2003). 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. *22*, 313–325.

Patterson, M.C., Vecchio, D., Prady, H., Abel, L., and Wraith, J.E. (2007). Miglustat for treatment of Niemann-Pick C disease: a randomised controlled study. Lancet Neurol. *6*, 765–772.

Patterson, M.C., Hendriksz, C.J., Walterfang, M., Sedel, F., Vanier, M.T., Wijburg, F., and NP-C Guidelines Working Group (2012). Recommendations for the diagnosis and management of Niemann-Pick disease type C: an update. Mol. Genet. Metab. *106*, 330–344.

Peljto, A.L., Zhang, Y., Fingerlin, T.E., Ma, S.-F., Garcia, J.G.N., Richards, T.J., Silveira, L.J., Lindell, K.O., Steele, M.P., Loyd, J.E., et al. (2013). Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. JAMA *309*, 2232–2239.

Pfrieger, F.W. (2003). Cholesterol homeostasis and function in neurons of the central nervous system. Cell. Mol. Life Sci. CMLS *60*, 1158–1171.

Pinto, R., Caseiro, C., Lemos, M., Lopes, L., Fontes, A., Ribeiro, H., Pinto, E., Silva, E., Rocha, S., Marcão, A., et al. (2003). Prevalence of lysosomal storage diseases in Portugal. Eur. J. Hum. Genet. *12*, 87–92.

Pipalia, N.H., Cosner, C.C., Huang, A., Chatterjee, A., Bourbon, P., Farley, N., Helquist, P., Wiest, O., and Maxfield, F.R. (2011). Histone deacetylase inhibitor treatment dramatically reduces cholesterol accumulation in Niemann-Pick type C1 mutant human fibroblasts. Proc. Natl. Acad. Sci. *108*, 5620–5625.

Pipalia, N.H., Subramanian, K., Mao, S., Balch, W.E., and Maxfield, F.R. (2016). Histone deacetylase inhibitors correct the cholesterol storage defect in most NPC1 mutant cells. bioRxiv 076695.

Pipalia, N.H., Subramanian, K., Mao, S., Ralph, H., Hutt, D.M., Scott, S.M., Balch, W.E., and Maxfield, F.R. (2017). Histone deacetylase inhibitors correct the cholesterol storage defect in most NPC1 mutant cells. J. Lipid Res. jlr.M072140.

Platt, F.M., Neises, G.R., Dwek, R.A., and Butters, T.D. (1994). N-butyldeoxynojirimycin is a novel inhibitor of glycolipid biosynthesis. J. Biol. Chem. *269*, 8362–8365.

Platt, F.M., Boland, B., and Spoel, A.C. van der (2012). Lysosomal storage disorders: The cellular impact of lysosomal dysfunction. J. Cell Biol. *199*, 723–734.

Platt, F.M., Wassif, C., Colaco, A., Dardis, A., Lloyd-Evans, E., Bembi, B., and Porter, F.D. (2014). Disorders of cholesterol metabolism and their unanticipated convergent mechanisms of disease. Annu. Rev. Genomics Hum. Genet. *15*, 173–194.

Potkin, S.G., Guffanti, G., Lakatos, A., Turner, J.A., Kruggel, F., Fallon, J.H., Saykin, A.J., Orro, A., Lupoli, S., Salvi, E., et al. (2009). Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. PloS One *4*, e6501.

Pratt, A.J., Getzoff, E.D., and Perry, J.J.P. (2012). Amyotrophic lateral sclerosis: update and new developments. Degener. Neurol. Neuromuscul. Dis. *2012*, 1–14.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

Reddy, H.M., Cho, K.-A., Lek, M., Estrella, E., Valkanas, E., Jones, M.D., Mitsuhashi, S., Darras, B.T., Amato, A.A., Lidov, H.G., et al. (2017). The sensitivity of exome sequencing in identifying pathogenic mutations for LGMD in the United States. J. Hum. Genet. *62*, 243–252.

Reue, K. (2009). The Lipin Family: Mutations and Metabolism. Curr. Opin. Lipidol. *20*, 165–170.

Roszell, B.R., Tao, J.-Q., Yu, K.J., Gao, L., Huang, S., Ning, Y., Feinstein, S.I., Vite, C.H., and Bates, S.R. (2013). Pulmonary Abnormalities in Animal Models Due to Niemann-Pick Type C1 (NPC1) or C2 (NPC2) Disease. PLOS ONE *8*, e67084.

Runz, H., Dolle, D., Schlitter, A.M., and Zschocke, J. (2008). NPC-db, a Niemann-Pick type C disease gene variation database. Hum. Mutat. *29*, 345–350.

Saftig, P., and Klumperman, J. (2009). Lysosome biogenesis and lysosomal membrane proteins: trafficking meets function. Nat. Rev. Mol. Cell Biol. *10*, 623–635.

Schil, K.V., Karlstetter, M., Aslanidis, A., Dannhausen, K., Azam, M., Qamar, R., Leroy, B.P., Depasse, F., Langmann, T., and Baere, E.D. (2016). Autosomal recessive retinitis pigmentosa with homozygous rhodopsin mutation E150K and non-coding cis-regulatory variants in CRX-binding regions of SAMD7. Sci. Rep. *6*, 21307.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Schwartz, S., Oren, R., and Ast, G. (2011). Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads. PLOS ONE *6*, e16685.

Seibold, M.A., Wise, A.L., Speer, M.C., Steele, M.P., Brown, K.K., Loyd, J.E., Fingerlin, T.E., Zhang, W., Gudmundsson, G., Groshong, S.D., et al. (2011). A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. N. Engl. J. Med. *364*, 1503–1512.

Sheth, J., Joseph, J.J., Shah, K., Muranjan, M., Mistri, M., and Sheth, F. (2017). Pulmonary manifestations in Niemann-Pick type C disease with mutations in NPC2 gene: case report and review of literature. BMC Med. Genet. *18*, 5.

Shin, J., Epperson, K., Yanjanin, N.M., Albus, J., Borgenheimer, L., Bott, N., Brennan, E., Castellanos, D., Cheng, M., Clark, M., et al. (2011). Defining natural history: assessment of the ability of college students to aid in characterizing clinical progression of Niemann-Pick disease, type C. PloS One *6*, e23666.

Škuljec, J., Gudi, V., Ulrich, R., Frichert, K., Yildiz, Ö., Pul, R., Voss, E.V., Wissel, K., Baumgärtner, W., and Stangel, M. (2011). Matrix Metalloproteinases and Their Tissue Inhibitors in Cuprizone-Induced Demyelination and Remyelination of Brain White and Gray Matter. J. Neuropathol. Exp. Neurol. *70*, 758–769.

Smith, M. (2014). Molecular Insights into Development in Humans: Studies in Normal Development and Birth Defects (World Scientific Publishing Co Inc).

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Solomon, D., Winkelman, A.C., Zee, D.S., Gray, L., and Büttner-Ennever, J. (2005). Niemann-Pick Type C Disease in Two Affected Sisters: Ocular Motor Recordings and Brain-Stem Neuropathology. Ann. N. Y. Acad. Sci. *1039*, 436–445.

Stitziel, N.O., Kiezun, A., and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. *12*, 227.

Tata, A.M., Velluto, L., D'Angelo, C., and Reale, M. (2014). Cholinergic system dysfunction and neurodegenerative diseases: cause or effect? CNS Neurol. Disord. Drug Targets *13*, 1294–1303.

Tucci, A., Liu, Y.-T., Preza, E., Pitceathly, R.D.S., Chalasani, A., Plagnol, V., Land, J.M., Trabzuni, D., Ryten, M., Jaunmuktane, Z., et al. (2014). Novel C12orf65 mutations in patients with axonal neuropathy and optic atrophy. J. Neurol. Neurosurg. Psychiatry *85*, 486–492.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *43*, 11.10.1-33.

Vanier, M.T. (2010). Niemann-Pick disease type C. Orphanet J. Rare Dis. *5*, 16.

Vanier, M.T. (2014). Complex lipid trafficking in Niemann-Pick disease type C. J. Inherit. Metab. Dis. *38*, 187–199.

Vanier, M.T., and Latour, P. (2015a). Chapter 18 - Laboratory diagnosis of Niemann–Pick disease type C: The filipin staining test. In Methods in Cell Biology, F.P. and N. Platt, ed. (Academic Press), pp. 357–375.

Vanier, M.T., and Latour, P. (2015b). Laboratory diagnosis of Niemann-Pick disease type C: the filipin staining test. Methods Cell Biol. *126*, 357–375.

Vélez, J.I., Lopera, F., Sepulveda-Falla, D., Patel, H.R., Johar, A.S., Chuah, A., Tobón, C., Rivera, D., Villegas, A., Cai, Y., et al. (2016). APOE*E2 allele delays age of onset in PSEN1 E280A Alzheimer's disease. Mol. Psychiatry *21*, 916–924.

Vellodi, A. (2005). Lysosomal storage disorders. Br. J. Haematol. *128*, 413–431.

Vite, C.H., Bagel, J.H., Swain, G.P., Prociuk, M., Sikora, T.U., Stein, V.M., O'Donnell, P., Ruane, T., Ward, S., Crooks, A., et al. (2015). Intracisternal cyclodextrin prevents cerebellar dysfunction and Purkinje cell death in feline Niemann-Pick type C1 disease. Sci. Transl. Med. *7*, 276ra26-276ra26.

Vitner, E.B., Platt, F.M., and Futerman, A.H. (2010). Common and Uncommon Pathogenic Cascades in Lysosomal Storage Diseases. J. Biol. Chem. *285*, 20423–20427.

Wang, Y., Liang, P., Lan, F., Wu, H., Lisowski, L., Gu, M., Hu, S., Kay, M.A., Urnov, F.D., Shinnawi, R., et al. (2014). Genome Editing of Isogenic Human Induced Pluripotent Stem Cells Recapitulates Long QT Phenotype for Drug Testing. J. Am. Coll. Cardiol. *64*, 451–459.

Wassif, C.A., Cross, J.L., Iben, J., Sanchez-Pulido, L., Cougnoux, A., Platt, F.M., Ory, D.S., Ponting, C.P., Bailey-Wilson, J.E., Biesecker, L.G., et al. (2016). High Incidence of Unrecognized Visceral/Neurological Late-onset Niemann-Pick Disease, type C1 Predicted by Analysis of Massively Parallel Sequencing Data Sets. Genet. Med. Off. J. Am. Coll. Med. Genet. *18*, 41–48.

Wilke, S., Krausze, J., and Büssow, K. (2012). Crystal structure of the conserved domain of the DC lysosomal associated membrane protein: implications for the lysosomal glycocalyx. BMC Biol. *10*, 62.

Williams, H., Johnson, J.L., Jackson, C.L., White, S.J., and George, S.J. (2010). MMP-7 mediates cleavage of N-cadherin and promotes smooth muscle cell apoptosis. Cardiovasc. Res. *87*, 137–146.

Wraith, J.E. (2002). Lysosomal disorders. Semin. Neonatol. *7*, 75–83.

Wraith, J.E., and Imrie, J. (2009). New therapies in the management of Niemann-Pick type C disease: clinical utility of miglustat. Ther. Clin. Risk Manag. *5*, 877–887.

Yanjanin, N.M., Vélez, J.I., Gropman, A., King, K., Bianconi, S.E., Conley, S.K., Brewer, C.C., Solomon, B., Pavan, W.J., Arcos-Burgos, M., et al. (2010). Linear Clinical Progression, Independent of Age of Onset, in Niemann-Pick Disease, type C. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *153B*, 132–140.

York, M.R., Nagai, T., Mangini, A.J., Lemaire, R., van Seventer, J.M., and Lafyatis, R. (2007). A macrophage marker, Siglec-1, is increased on circulating monocytes in patients with systemic sclerosis and induced by type I interferons and toll-like receptor agonists. Arthritis Rheum. *56*, 1010–1020.

Zhu, X., Feng, T., Li, Y., Lu, Q., and Elston, R.C. (2010). Detecting rare variants for complex traits using family and unrelated data. Genet. Epidemiol. *34*, 171–187.

PLINK 1.9.