Investigating postzygotic de novo mutations and somatic mosaicism in monozygotic twins discordant for complex disorders

Nirmal Vadgama

September 2017

Thesis submitted for the degree of Doctor of Philosophy

University College London, Institute of Neurology

Supervisors Prof John Hardy and Dr Jamal Nasir

Declaration

I, Nirmal Vadgama, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

If I was asked four years ago to sum up my thesis in one line, I may have said, ‘I’m writing to make sense of things that don’t make sense to me’. If asked now, I would say, ‘I know a thing or two about twins’. Like most research, this work was made possible through a joint contribution. While I accept the blame for any shortcomings, I would like to thank fellow colleagues for helping me along this long and tortuous, but nevertheless fruitful, path. Among them, I must single out Jamal Nasir for special mention – not only for being a supportive supervisor, but a great mentor and friend as well. I would also like to thank Alan Pittman for streamlining the exome sequencing data analysis, and Ruth Lovering for showing me the way of ‘gene ontology’. And of course, my principle supervisor John Hardy, for saying ‘yes’ to my PhD proposal when others (perhaps justifiably) rejected it for being too risky.

This work has also benefitted from the input of the following collaborators, in no particular order: Michael Simpson, Niranjanan Nirmalananthan, Robin Murray, Takeo Yoshikawa, Mark Kristiansen, Kerra Pearce, Sarah Morgan, Douglas Lamont, Peter De Rijik, David Gaze, Elliot Rees, George Kirov, Tomas Fitzgerald, Charles Lee and CZ Zhang.

It is custom here to also thank my family and friends, except none of them had a clue about what I was doing. However, I am grateful to them for at least pretending to be interested when I told them about this work, and buggering off when I needed to focus.

Abstract

Monozygotic (MZ) twins were long thought to be genetically identical, however recent studies have demonstrated genetic differences between them. To test the hypothesis that early post-twinning mutational events leads to phenotypic discordance, thirteen MZ twins discordant for a range of complex disorders were investigated at the genomic and proteomic level.

Whole-exome sequencing data was analysed using a union of VarScan2 and MuTect2 variant calling algorithms. Copy-number variation (CNV) analysis from Illumina HumanCore array data was also carried out using PennCNV and cnvPartition to identify structural variants that would not be detected by exome sequencing. All single nucleotide variants (SNVs), indels and CNVs were evaluated for functional consequence, evolutionary conservation, population frequency and overlap with known disease-susceptibility genes.

Twenty-two putative discordant SNVs and indels, but no discordant structural variants, were identified. Parent-offspring trio analysis was implemented to assess potential association of germline de novo mutations with susceptibility to disease. A rare, highly conserved de novo mutation in *RASD2* was detected in twins discordant for attention deficit hyperactivity disorder (ADHD). *RASD2* is enriched in the striatum and involved in the modulation of dopaminergic transmission. In the twins discordant for Tourette's syndrome, an inherited stop loss mutation was detected in *AADAC*, a known candidate gene for the disorder. Further, a de novo CNV duplication was identified in a twin pair discordant for ADHD overlapping *CD38*, a gene implicated in social amnesia and autism. When analysing the burden of shared CNVs among the twins, a rare hemizygous deletion in region 15q13.2 was detected in twins with schizophrenia, overlapping *ARHGAP11B*,

a human-specific gene involved in basal progenitor amplification and neocortex expansion.

To investigate potential downstream consequences of (epi)genetic mechanisms and underlying biochemical pathways, proteomic profiling of serum samples obtained from an MZ twin pair discordant for ischaemic stroke was analysed through a label-free pipeline. Biological processes overrepresented in the affected twin predominantly corresponded to stroke-related processes, including wound healing, blood coagulation and haemostasis. Further, a comparison of blood chemistries showed a >10- and >18-fold elevation of γ-glutamyltransferase (GGT) and erythrocyte sedimentation rate (ESR) levels respectively in the affected twin.

Table of Contents

List of Figures

List of Tables

Abbreviations

A

ADHD	Attention deficit hyperactivity disorder
ALS	Amyotrophic lateral sclerosis
ALSoD	Amyotrophic Lateral Sclerosis Online genetics Database

B

BAF	B allele frequency

C

CN	Copy number
CNV	Copy number variation

D

DA	Diamniotic
DC	Dichorionic
ddPCR	Droplet digital polymerase chain reaction
DECIPHER	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
dH_2O	Distilled water
DNP – 2,4 D	2,4-Dinitrophenol
DZ	Dizygotic

E

EBV	Epstein–Barr virus
EDTA	Ethylenediaminetetraacetic acid
EGF	Epidermal growth factor
ESR	Erythrocyte sedimentation rate

F

FTD	Frontotemporal dementia

G

GAPs	GTPase-activating proteins
GATK	Genome Analysis Toolkit
GERP++	Genomic Evolutionary Rate Profiling
GGT	Gamma-glytamyl transpeptidase
GTR	Genetic Testing Registry
GWAS	Genome-wide association study

H

h^2	Heritability
HPLC	High-performance liquid chromatography
HSP	Hereditary spastic paraplegia

I

IGV	Integrative Genomics Viewer
Indels	Insertion/deletion

L

LCLs	Lymphoblastoid cell lines
LLR	Log R ratio
LoF	Loss-of-function

M

MA	Monoamniotic
MAF	Minor allele frequency
MC	Monochorionic
METSIM	METabolic Syndrome In Men
MS	Mass spectrometry
mtDNA	Mitochondrial DNA
MZ	Monozygotic

N

NGS	Next-generation sequencing
NIH	National Institute of Heath
Numts	Nuclear mitochondrial DNA sequences

O

OCD	Obsessive compulsive disorder
OMIM	Online Mendelian Inheritance in Man

P

pLI	Loss Intolerance probability
PolyPhen	Polymorphism Phenotyping

R

rpPCR	Repeat-primed polymerase chain reaction
RVIS	Residual Variation Intolerance Score

S

SAM	Sequence Alignment/Map
SIFT	Scale Invariant Feature Transform
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SZDB	Schizophrenia database
SZGene	SchizophreniaGene

T

TBE	Tris/Borate/EDTA

V

VQSR	Variant Quality Score Recalibration

Chapter 1. Introduction

1.1 Overview

In this review, the continuing value of classical twin methods is considered from the modern perspective of molecular genetics before delving into a discussion on the impact of de novo mutations on human disease. Some of the popular myths and unquestioned assumptions about the twinning process are brought into focus, and a range of twin studies are presented. Methodological considerations and future strategies are also explored, while drawing from caveats of our own research. There has been a gradual shift in twin research, from focusing on the initial level of the DNA sequence, through to capturing the downstream effects of disease-causing mutations. Thus, novel strategies using omics technologies and their integration across multiple omics layers are briefly described in the context of twin research.

1.2 A thing or two about twins

1.2.1 History of twin research

Twins have captured the interest of researchers for centuries, and proposals to use them in empirical studies originate from as early as 415CE by Augustine of Hippo (Augustine, 2013). Galton published his classic article on twins in the 19th century (Galton, 1876) earning him the widely-held recognition of having been the first to iterate the classical twin method. According to Rende, Plomin and Vandenberg (1990), however, Galton did not make the distinction between MZ and dizygotic (DZ) twins, an understandable shortcoming nonetheless due to the ignorance of heredity at the time. Rende et al. (1990) conclude that Siemens (1924) and Merriman (1924) together discovered the genetic

identity of MZ and DZ twins, but the fundamental work of Poll (1915) and Weinberg (1901), and their forerunners, nor Bonnevie (1924), were not considered. Some of the obscurities in the origin of the method may have arisen from the possible language barriers of English-speaking scientists. Nevertheless Siemens, a dermatologist, made significant contributions to the systematic analysis of similarity of MZ and DZ twins, who independently formulated the twin rule of pathology: identical twins are more likely to be concordant for a heritable disease than non-identical twins, and concordance will be even lower in non-siblings (Siemens, 1924).

Siemens' study of moles (skin demarcations, not the animal) led him to come up with an innovative idea of merging correlation analysis and twin data. He correlated mole counts between co-twins and contrasted this correlation between MZ and DZ twin pairs. MZ twins, who share all – or nearly all – of their genetic material had a mole count correlation of 0.4. Whereas DZ twins, who are on average 50% genetically identical, had a correlation of only 0.2. The results demonstrated the importance of genetic factors to the variation in mole count. More generally, the genetic similarity of MZ twins is associated with their larger resemblance for the phenotype under study.

1.2.2 The classical twin design

The classical twin study compares phenotypic resemblances of MZ and DZ twins. Comparing the resemblance of MZ twins for a trait or disease with resemblance of DZ twins offers an initial estimate of the extent to which genetic variation determines phenotypic variation of that trait.

Hereditability (h^2) can be defined as the proportion of variance of a phenotype that is due to genetic influences between individuals in a population. It can be estimated from twice

the difference between MZ and DZ correlations – that is, $2(r_{MZ} - r_{DZ})$, where r_{MZ} and r_{DZ} are the estimated correlation coefficients of MZ and DZ twins, respectively.

In a meta-analysis of twin correlations, Polderman et al. (2015) reported variance components of all published twin studies of complex traits. They report that the heritability estimate for all 17,804 non-unique traits investigated is 49%. The largest heritability estimates were for traits classified under the ophthalmological domain (h^2 = 70%). Traits in the reproduction, environment and social values domains had the lowest heritability estimates.

In a population-based study of 1033 female-female twin pairs, the average MZ and DZ correlations for depression was reported to be around 0.4 and 0.2, respectively (Kendler, 1992); thus, the estimated heritability is calculated as approximately 40%. Further, in a meta-analysis of published twin studies of autism spectrum disorders (Tick et al., 2015), the meta-analysis correlations for MZ were almost perfect at 0.98, whereas DZ correlation was 0.53. Thus, the meta-analytic heritability estimates are considerable, ranging up to 90%.

The application of this type of analysis led to substantial changes in the way we think about the determinants of health and disease and the causes of individual differences in normal and abnormal behaviour. During the last couple of decades, a shift has taken place from strict environmental explanations to a more balanced view that recognises the importance of genes, for example in autism spectrum disorders in children, or in the development of dependence on nicotine, alcohol and other drugs in adults (Manuck and McCaffery, 2014). The recent advancement in DNA sequencing techniques has led to significant methodological improvements, allowing the comparison of twin genomes up to the base pair level. This is especially true with the co-evolution of next-generation

sequencing (NGS) platforms and DNA microarray strategies (Schuster, 2007). This has ineluctably spurred greater interest in the research of de novo mutations and cases of mosaicism in relation to MZ twins, including the mapping of disease genes in DZ twins (Tan et al., 2010).

1.2.3 The case co-twin design

Twins are particularly useful in case-control studies, and MZ twins form the ideal case-control study, as they are perfectly matched for genotype and family background. In an earlier use of this approach, the effect of vitamin C on the common cold was studied by administering it in one twin and a placebo in the co-twin. Contrary to popular belief, it was determined that vitamin C is no more effective than a placebo (Martin et al., 1982). In line with the Barker hypothesis (Hales and Barker, 2001), differences in birthweight in MZ and DZ twin pairs and at their association with differences in cardiovascular and metabolic parameters have been investigated. If these associations are due to shared genetic factors, difference scores are uncorrelated in MZ, but not in DZ twins. IJzerman, Stehouwer and Boomsma (2000) determined that the association between low birth weight and high blood pressure in later life seems to be mediated by common genes.

In a case–control design more relevant to this study, discordant MZ twins, only one of whom has a disease, are used to investigate which non-shared environmental influences are related to the disorder (Asbury et al., 2003). Such discordant MZ pairs might also form the perfect case-control design for gene-expression studies to distinguish between genes that are related to the causes of disease and genes that are expressed as a consequence of disease. Alternatively, such differential expression, congruent with disease discordance, might indicate causal genes that are differentially activated by epigenetic factors (Petronis et al., 2003).

More recently, the classical co-twin design has received much interest for studying molecular biology. Identifying genomic differences between twins and appreciating the subtle nuances in choosing the ideal analytical approach relies on an understanding of the mechanism and aetiology of the twinning process – and it is to this theme we now turn.

1.3 Mechanisms of twinning

1.3.1 Traditional models of twinning

Conventional wisdom dictates that DZ twins are the result of fertilisation of two distinct ova by two different spermatozoa, while MZ twins are the product of a single ovum and sperm that subsequently splits and forms two embryos (Hall, 2003).

The development of the chorion, the foetal placental membrane, begins around day three or four after conception, and the formation of the amnion occurs between day six and eight (Hall, 2003). The widely-held consensus of MZ twinning mechanisms are based on the yet unverified theory of postzygotic division of the conceptus. According to this model, the number of foetuses, chorions, and amnions depend on the time at which the embryo splits. In most DZ pregnancies, each embryo develops a separate chorion, amnion and placenta – although rare occurrences of other mechanisms have been reported (Hackmon, 2009).

Twins can be distinguished into four groups based on the zygosity and number of the placental membranes. In general, DZ pairs, who develop from the fertilisation of two ova, are dichorionic diamniotic (DC DA). MZ twins, however, may have a different arrangement of the adnexa (foetal membranes) depending on the timing of embryo division. Here, there are three possibilities; namely, monochorionic monoamniotic (MC

MA), monochorionic diamniotic (MC DA), and dichorionic diamniotic (DC DA). The latter state occurs if the embryo divides at an early stage, which would resemble the membrane constitution of DZ twins.

Various factors have been associated with DZ twinning, such as assisted reproductive technology pregnancies, genetic factors, and increased concentrations of follicle-stimulating hormone. The aetiology of MZ twinning is less clear. It used to be assumed that MZ twinning is a random process (Wong, Gottesman and Petronis, 2005), but it is now considered to be a consequence of prior events affecting the gamete. Several studies have demonstrated that MZ twinning is not a random teratogenic event, but rather familial, aluding to the idea that twinning itself may be an inherited phenomenon. Some reported cases appear to show twinning to have an autosomal dominant inheritance pattern (Machin, 2009a). Epigenetic mechanisms or even a twinning gene are thought to be involved (Machin, 2009b).

Proposed triggers for splitting include postzygotic mutations, abnormalities in cell surface proteins, skewed X-inactivation, and aberrations in the development of the zona pellucida (McNamara et al., 2016; Hall, 2003). It has even been postulated that genetic alterations in cells within the blastocyst could recognise each other as foreign, thereby using cell-recognition mechanisms to set up two separate cell masses, thus leading to the twinning process (Hall, 2003).

In vitro fertilisation has increased the incidence of MZ twins by two- to five-fold (McNamara et al., 2016), possibly owing to the handling, culture media, and/or microinjection procedures that damage, and thus split, the embryo. Of course, intrinsic irregularities associated with infertility may also be a factor, an idea that is consistent with the substantially increased rate of twin births reported among teenage girls and older

women – that is, at the beginning and end of reproductive age (Blickstein, Verhoeven and Keith, 1999; Alikani et al., 1994).

It may be worth mentioning, without straying too far from our primary focus, that induced delay of ovulation in rats (Butcher and Fugo, 1967), rabbits (Al Mufti and Bomse-Helmrich, 1979) and humans (Papiernik et al., 1979), as well as delayed artificial insemination in guinea pigs (Blandau and Young, 1939), are associated with overripeness ovopathy, intrauterine growth restriction and chromosomal aberrations of embryos and foetuses. Interestingly, in some of the experimental models, twinning was also observed. For instance, in the rabbit, twinning was evidenced by the observation of two early blastocysts inside the same intact zona pellucida (Al Mufti and Bomse-Helmrich, 1979).

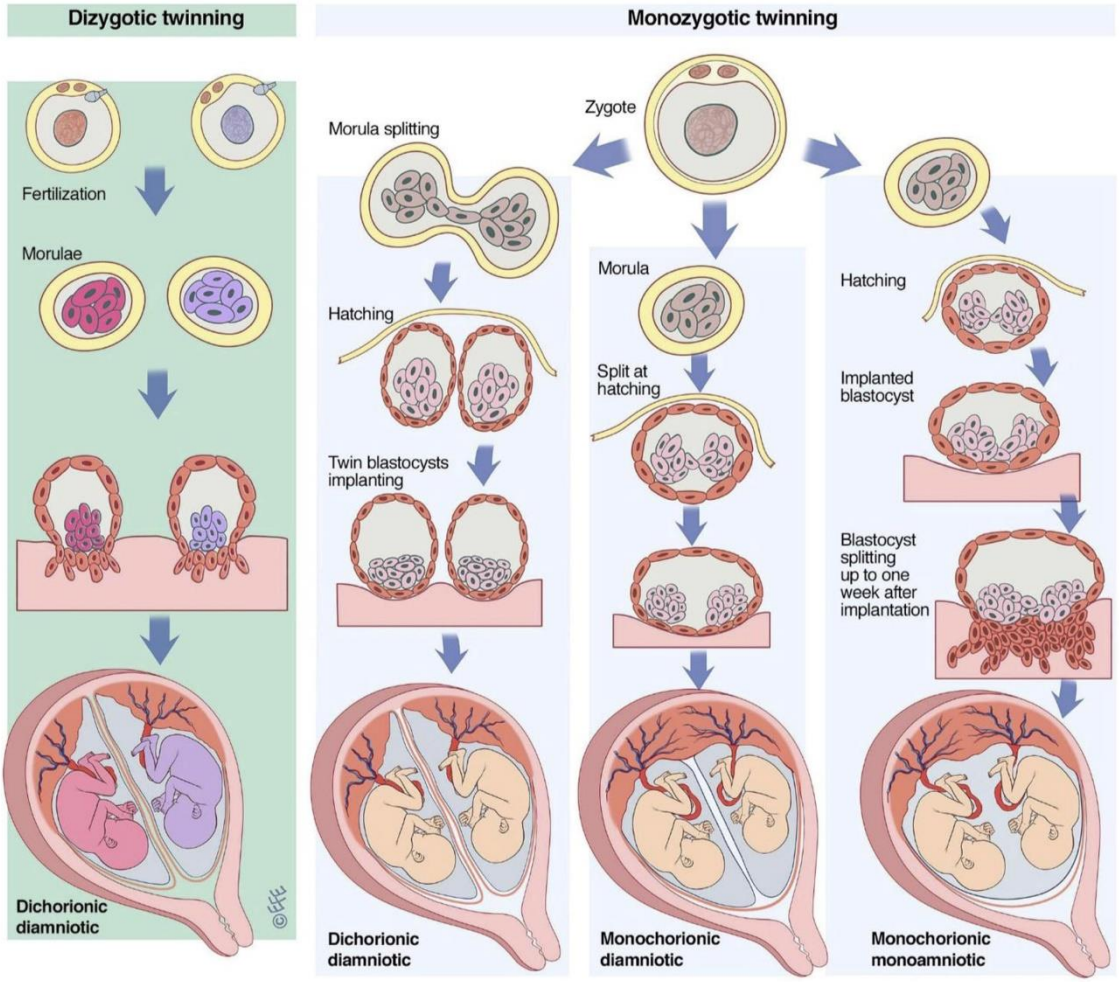


Figure 1.1. The traditional model of twinning. DZ twins are the product of 2 distinct fertilisation events, resulting in DC DA twins with each conceptus developing to become a genetically distinct individual. MZ twins result from postzygotic splitting of the product of a single fertilisation event. Splitting on days 1–3 (up to the morula stage) results in DC DA twins, on days 3–8 (during which blastocyst hatching occurs) in MC DA twins, on days 8–13 in MC MA twins, and if no split has occurred by day 13, in conjoined twins (not shown). In this diagram, 2 of the 3 oocyte-derived polar bodies are shown at the zygote stage. Figure from McNamara et al. (2016).

1.3.2 Zygosity myth-conceptions

In recruiting discordant MZ twins for our ongoing research in investigating postzygotic de novo mutations, we have come across an increasing number of participants who have been told they are DZ but have almost identical traits and physical features, or had been told they were MZ but present more like fraternal siblings (Vadgama et al, 2015). Some parents or twins had mistaken assumptions of zygosity diagnosis, acquired either

according to their own presuppositions or because they were given incorrect information by medical professionals.

Current practice guidelines for classifying twins as either MZ or DZ are prone to error, which are still largely based on chorionicity (the notion of whether the twins shared a placenta) and physical similarity (Boklage, 2009). Indeed, one of the popular twin-myths among the public is the idea that MZ twins must look exactly alike, or that like-sexed twins are invariably identical. Among medical professionals, a common error in zygosity assignment is the presumption that like-sexed twins are DZ if they were determined to be DC at birth. Results of DNA-based gold standard methods have shown that up to 25% of all like-sexed twin pairs classified as DZ, and 7% classified as MZ by chorionicity, turn out to be misclassifications (Forget-Dubois et al., 2003; Bamforth and Machin, 2004). Moreover, as mentioned previously, rare cases of MC DZ twins have been reported, thus undermining the reliability of current prenatal diagnosis based on chorionicity (Hackmon et al., 2009).

We propose the ethical importance and financial feasibility to carry out genetic testing using placental tissue or even buccal swabs in all like-sexed twins as a standard procedure under the National Health Service (Vadgama et al., 2015).

1.3.3 Challenging the convention

The traditional model of twinning is theoretical merely, not proven. It is based on our current understanding of embryological events, but it is widely endorsed because it seems plausible. Certainty and science are usually – or perhaps should always be – mutually exclusive. The orthodox view of twinning mechanisms is buttressed on pillars of ‘common knowledge’, incorrectly and submissively simplified by a disregard of attention and curiosity. In twin research, tracing the often-cited received opinions that are purported

to be facts to their origin, one finds them to be supported merely by repetition, rather than the verifiable observations critical to scientific integrity.

Nevertheless, the traditional models of twinning have not been unchallenged. For instance, the Spanish physician Gonzalo Herranz urges serious rethinking of the hitherto-unchallenged hypothesis of postzygotic splitting (Herranz, 2013). Four interconnected grounds of doubt may be found through his line of reasoning; namely, splitting has never been observed *in vitro*, factors that cleave the embryo have not been clearly defined, distinct embryos coexisting within a single zona pellucida is improbable, and postzygotic splitting becomes increasingly unlikely with the passage of time. Herranz's alternative theory of twinning is based on two postulates: first, MZ twinning transpires at the first cleavage division of the zygote; and second, the chorionicity and amnionicity status is determined by the degree of fusion of embryonic membranes within the zona pellucida (Figure 1.2).

However, this appraisal was not received without rebuttals. Denker (2013), for example, questioned Herranz's argument by noting that a lack of evidence may be the result of ethical limitations on research with human embryos. Denker also recapitulated the available data regarding twinning mechanisms in animals, the observed *in vivo* and *in vitro* differences in the zona pellucida, and the postzygotic developmental potential of cell lineages – factors which, he believed, Herranz overlooked. Denker concluded that both Herranz's fusion model and the traditional fission model remain unverified.

The mechanism of twinning remains poorly understood at the molecular level, warranting further research to determine what hypothesis, if either, is tenable. But it can only be done if our complacency with received opinion does not hinder rational inquiry. There is much

work to be done to address those unanswered questions buried under those unquestioned answers.

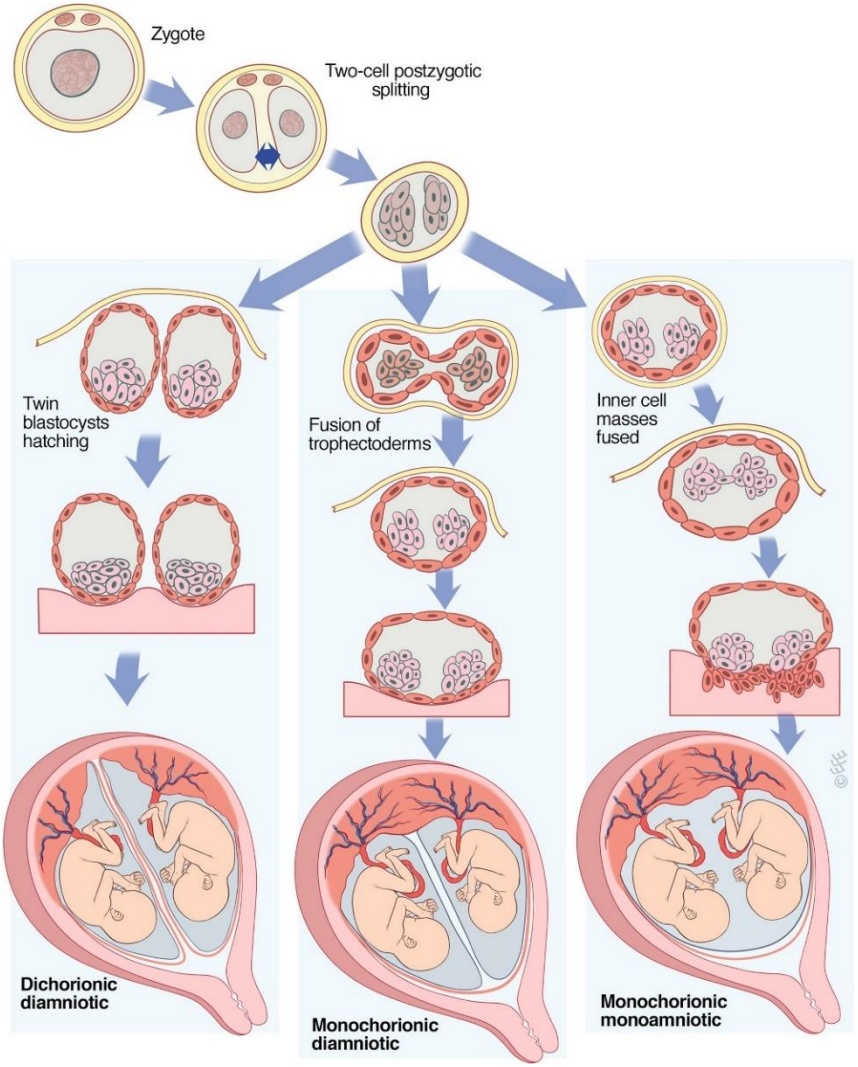


Figure 1.2. An alternative model of MZ twinning. In this model, splitting occurs at the postzygotic 2 cell stage, with each cell forming a distinct individual. If twin blastocysts hatch from the zona pellucida together, DC DA twins will result. If the 2 trophectoderms fuse before hatching and the inner cells masses are separated within the shared trophectoderm, MC DA twins will result. If the inner cell masses are fused and separated later, MC MA twins will result. Figure from McNamara et al. (2016).

1.4 Similar but not identical

As MZ twins frequently show discordance in some phenotypic aspects, the use of the term ‘identical twins’ is increasingly being relegated to the margins. Some cases of twin discordance might possibly be stochastic in origin. Transcriptional or translational

stochasticity will inevitably lead to variation, necessitated by the haphazard movements of molecules and the intricacy of their relations (Czyz et al., 2012). This can occur irrespective of identical environmental conditions (Petronis, 2006). But with a lack of understanding of causal mechanisms in operation, it is difficult to exclude environmental or genetic interactions when explaining discordance, such as for hair or eye colour, fingerprint profiles (suggesting twins cannot carry out the 'perfect crime', contrary to popular fiction), mirror image features (present in up to 25% of MZ twins), and major deformities (Czyz et al., 2012; Gringras and Chen, 2001). The most frequent are minor asymmetries of facial characteristics (ptosis, side of eruption of the first tooth, side of upsweep of the hair or eyebrow), handedness and side of wrinkles. These mirror-image features may be related to a disturbance of laterality.

Many mechanisms responsible for discordance in MZ twins have been described to date. The most commonly reported are mosaicism for chromosomal or single gene disorders, skewed X-inactivation, uniparental disomy, differential repeat expansion, differential telomere length reduction, asymmetric transmission of mitochondria, and epigenetic mechanisms such as genomic imprinting and inner cell mass damage of the blastocyst (Zwijnenburg, Meijers-Heijboer and Boomsma, 2010).

In the past few years, there have been comparatively more reports of epigenetic differences than genetic differences between MZ twins. This is perhaps owed to the limitations of NGS technologies in detecting genetic differences (a theme we will revisit later), or because studies have typically involved a search only in or around coding regions. Nevertheless, sequence differences between twins have been identified. In the following section, a review of the literature on de novo mutations is presented, with a particular focus on their detection, biological characterisation, and medical impact in MZ twins.

1.5 Finding the hay in a needlestack

1.5.1 Chromosomal aneuploidy and structural abnormalities

A de novo mutation is one that arises in one family member for the first time due to a genetic alteration in a gamete (an ovum or sperm) of one of the parents, or occurs postzygotically, in the fertilised egg itself. Effectively, all disease-causing aneuploidies arise as de novo events. The most recognised example for this is trisomy 21, first identified in 1959 as the cause of Down's syndrome (Lejeune, Gautier and Turpin, 1959).

Phenotypic discordance in MZ twins can result from chromosomal mosaicism. A discrepant karyotype in MZ twins can arise via two different mechanisms: A mitotic error arising before the twinning event, resulting in a mosaic showing a discrepant distribution of the two different cell lines between the two foetuses; or a mitotic error occurring after twinning, resulting in a mosaic chromosomal alteration in one foetus only (Campbell et al., 2015).

Only a select few aneuploidies are compatible with human life in the constitutional state; namely trisomy of chromosomes 13, 18, 21 and X, and monosomy of X. A wider range of aneuploidies has been observed in the mosaic state. These include monosomy 7, 18 and 21, and trisomy 7, 8, 9, 12, 14, 15, 16, 17, 20 and 22 – each with varying percentages of cells affected, prevalence and clinical features. Whole chromosomal aneuploidy occurs via nondisjunction with meiotic risk increasing with maternal age. Mosaic aneuploidy is the result of a combination of meiotic nondisjunction rescued by postzygotic loss or copy of a whole chromosome, which can in turn lead to uniparental disomy in the euploid cell line (Conlin et al., 2010) or alternatively due to postzygotic nondisjunction. Two other large-scale abnormalities that are observed in the mosaic state are ring chromosomes and isochromosomes.

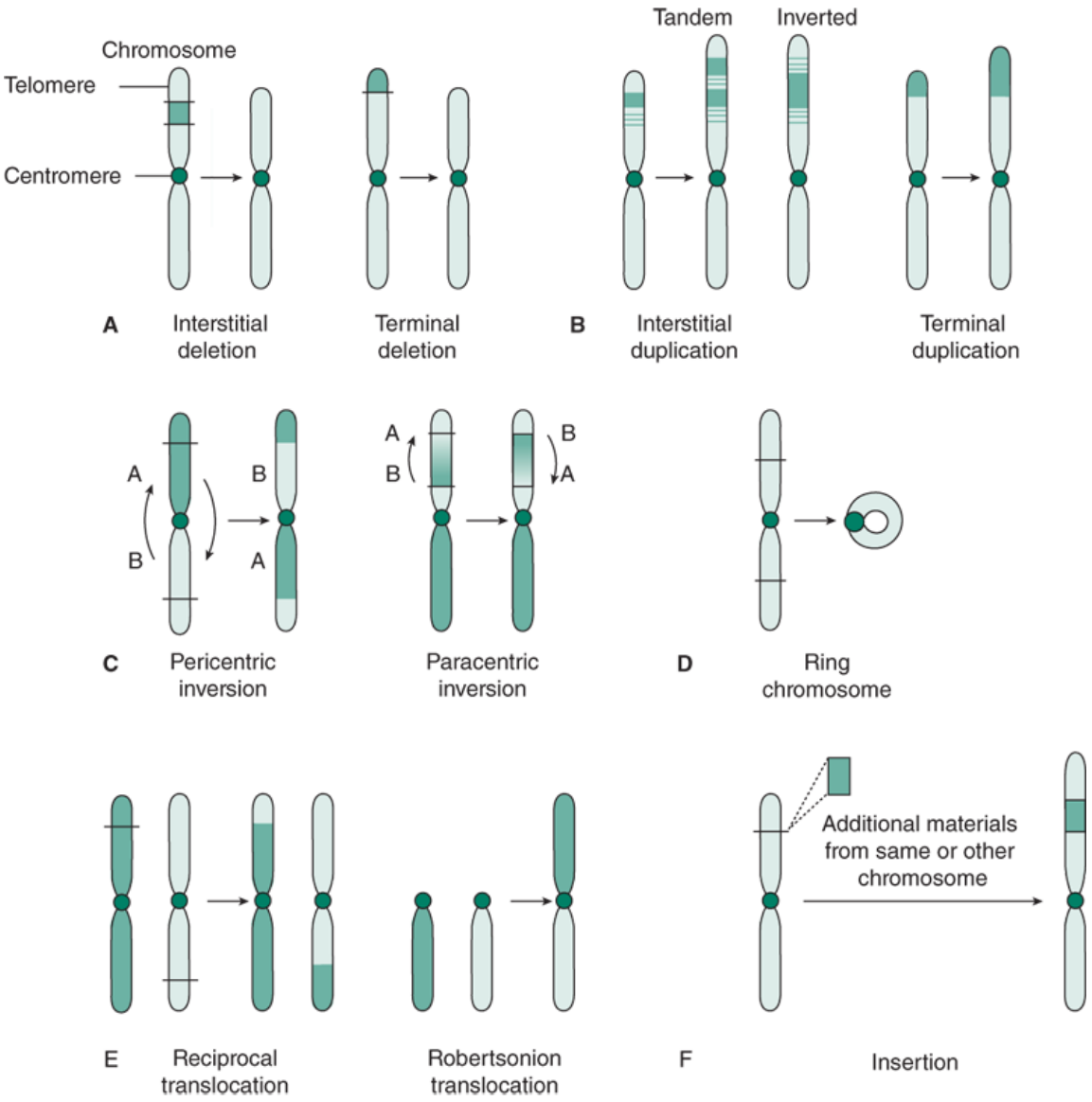


Figure 1.3. Examples of structural chromosomal abnormalities: deletion, duplication, inversion, ring chromosome, translocation, and insertion. Image from Abzug et al. (2014).

Grams, Rand and Norton (2011) report a case of iso(5p), resulting in tetrasomy (5p), detected prenatally in one foetus of an MC DA twin pair. The unaffected twin was found to be structurally and karyotypically normal. The mechanism of this rare nonmosaic iso(5p) event may be explained in three parts: first, the occurrence of trisomy 5 by meiosis II nondisjunction; second, a postzygotic rearrangement leading to the iso(5p) formation; third, the splitting in the twinning process leading to a loss of the isochromosome in one cell line – hence the discordant co-twins.

Some isochromosomes occur frequently enough that recognisable syndromes can be established, the most common being Xq (Pascual and Rosenberg, 2015). Another example includes that of 12p, which causes Pallister-Killian syndrome. Iso(12p) is only seen in the mosaic state, probably because such an aberration is lethal constitutionally (Campbell et al., 2015), but tolerated if eradicated from tissues vital to life (Raffel et al., 1986).

Timing of the division of cell linages in the twinning process can determine whether the MZ twins are completely discordant for the aneuploidy, or possess partial discordance due to the degree of somatic mosaicism. For example, in a study investigating twins discordant for trisomy 21, blood samples from a pair of MC DA twins showed 15% mosaicism for this aneuploidy in both twins (Egan et al., 2014). Remarkably, cheek and skin biopsies from the twins confirmed two distinct karyotypes: 47,XY,+21 in the affected twin and 46,XY in the unaffected twin. The presence of mosaicism in the blood samples of both twins was the consequence of their shared MC circulation. It was determined that postzygotic non-dysjunction during meiosis was the likely cause for the discordant somatic aneuploidy, and thus phenotype.

1.5.2 Single nucleotide variations

Recent studies based on observation of de novo SNV in parent-offspring trios have identified a low rate estimated to range between 0.82-1.70x10^{-8} mutations per base per generation (Dal et al., 2014; Kong et al., 2012; Campbell et al., 2012). Based on the mutation frequency of these studies, and given that there are approximately 3.0x10^{9} base pairs in the human genome (Human Genome Sequencing Consortium, 2004), it can be inferred that the average individual acquires 37 to 51 de novo single nucleotide mutations during development. Furthermore, the proportion of early postzygotic mutations to the overall human de novo SNV rate was estimated to be 0.34x10^{-8} and 0.04x10^{-8} in each twin of the MZ twin pair investigated (Dal et al., 2014), which translates into 1 to 10 single nucleotide mutations in the genome. This entails that although postzygotic SNVs are a relatively rare occurrence, they comprise a considerable proportion of the rate of de novo mutations in humans. Acuna-Hidalgo et al. (2015) have even shown that an important fraction of de novo mutations presumed to be germline in fact occurred either post-zygotically in the offspring, or were inherited because of low-level mosaicism in one of the parents. These mitotic events during embryogenesis resulting in somatic mosaicism may thus have an important and significant contribution in disease discordance observed in MZ twins. It is possible, however, that individual mutation rates vary considerably. This necessitates much larger studies if the true extent of variation in mutation rates are to be determined.

The first study to implement whole-genome sequencing on discordant twins was unsuccessful in finding reproducible differences to explain discordance for multiple sclerosis (Baranzini et al., 2010). The authors also investigated epigenetic and transcriptomic differences between the twins, but again, no differences in T cells were found (Baranzini, et al, 2010). This study should be considered merely exploratory, as

only three discordant twin pairs were investigated. Notwithstanding, multiple sclerosis has a number important environmental risk factors associated with its aetiology, including viral infection and migration (Ascherio, 2013).

Postzygotic de novo single nucleotide mutations have, nevertheless, been found to account for disease discordance in MZ twins. For instance, MZ twins discordant for neurofibromatosis type 1 were explained by the presence of a de novo *NF1* mutation in all somatic cells investigated in the affected twin, whereas somatic cells analysed from the unaffected twin were mosaic for the mutation (Kaplan et al., 2010), suggesting the distribution of the mutant allele across the various cells to be inadequate in manifesting the disorder. Other examples of de novo SNVs were found in MZ twins discordant for frontometaphyseal dysplasia (Robertson et al., 2006), Dravet's syndrome (Vadlamudi et al, 2010), Proteus syndrome (Lindhurst et al., 2011), schizophrenia (Tang et al., 2017), and neurodevelopmental disorders (Morimoto et al., 2017).

1.5.3 Copy number variations

Germline and somatic de novo genetic alterations have been implicated in human disease for decades. Over the past several years, genomic microarray technology has provided insight into the role of de novo CNVs in disease (Veltman and Brunner, 2012). Large CNVs typically occur at a low rate, arising at a frequency of only 0.01 to 0.02 events per generation (Veltman and Brunner, 2012; Kloosterman et al., 2015; Campbell and Eichler, 2013), but their disruptive effect on many genes contributes considerably to severe congenital malformations and neurodevelopmental disorders (Weischenfeldt et al., 2013).

Differences in large structural variants, particularly CNVs, have also been identified between clinically discordant MZ twins. In a MZ twin pair discordant for global developmental delay, non-allelic mitotic recombination at the 2p25.3 locus that occurred

during blastomeric divisions of a normal zygote, led to a variable copy number (CN) between the twins, with somatic mosaicism in the twin carrying the chromosomal anomaly (Rio et al., 2013). Several reports have mentioned twins discordant for the 22q11.2 microdeletion (Halder et al, 2012; Singh, 2002). This is the most common microdeletion syndrome; usually spontaneous, and characterised by a wide spectrum of clinical manifestations, it has an estimated incidence of one in 4,000 to 6,000 live births (Halder et al, 2012). CNV differences have also been observed in MZ twins discordant for Parkinson's disease (Bruder et al., 2008), congenital heart defects (Breckpot et al., 2012), and schizophrenia (Castellani et al., 2014).

1.5.4 Mitochondrial DNA

When we talk of the human genome, we often overlook the fact that humans possess at least two distinct genomes: nuclear DNA, which has been the centre of discussion thus far, and mitochondrial DNA (mtDNA), a circular structure of double-stranded DNA found inside mitochondria. Mitochondria are vital to cellular energy production, and thus crucial to healthy human physiology (Leem and Koh, 2012; Raimundo, 2014). mtDNA consists of merely 37 genes, and encode for tRNAs, rRNAs and proteins involved in cellular respiration.

Hence, disorders that are linked to tissues with high energy demands, such as brain or skeletal muscle, have been associated with defects in the mitochondria (Mattson, Gleichmann and Cheng, 2008; Rygiel et al., 2015). The pathogenesis of these disorders may involve mutations in either mtDNA or nuclear encoded mitochondrial genes (Rollins et al., 2009).

In contrast to the nuclear genome, double-stranded circular mtDNA is present in multiple copies within the same cell, and is exclusively maternally inherited. Among the hundred

to several thousand copies of mtDNA in a single cell, segregation of the mitochondrial populations during cell division can lead to mtDNA variation as they divide. The coexistence of different mtDNA sequences is known as heteroplasmy (Stewart and Chinnery, 2015); and homoplasmy, as one might have guessed, describes a cell that has a uniform collection of mtDNA.

Incomplete penetrance of several mitochondrial diseases can be attributed to heteroplasmy of pathogenic mtDNA mutations (Stewart and Chinnery, 2015; Taylor and Turnbull, 2005). Disease onset is (partly) dependent on whether the allele frequency of a mtDNA pathogenic variant exceeds a certain threshold (DiMauro and Schon, 2003).

mtDNA is known to have a higher mutation rate owing to apparent decreased replication fidelity (Song, Wheeler and Mathews, 2003; Lee and Johnson, 2006). mtDNA is known for its lack of protective histones and limited repair capacity, which can render it susceptible to insults that can overcome its replicative machinery or stimulate mitochondrial fusion causing mtDNA levels to differ. Various forms of cancer, toxin exposures, aging, and oxidative stress can either increase or decrease mtDNA CN, depending upon the nature of the damaging process (Wrede et al., 2015).

Deep sequencing is increasingly being used to detect genomic variation (Calvo et al., 2012). However, quantifying mtDNA heteroplasmy has been shown to be challenging (Yao, Kajigaya and Young, 2015). Mitochondrial variation can be misinterpreted given the presence of nuclear mitochondrial sequences (numts), which are highly similar nuclear fragments of the mitochondrial genome located on different chromosomes (Hazkani-Covo, Zeller and Martin, 2010). Studies have reported difficulties in estimating variation due to co-amplification of numts with the mtDNA (Parr et al., 2006; Bouhlal et al., 2013). Recent studies, however, have demonstrated significant improvements in

increasing the sensitivity of heteroplasmy detection, and NGS has been recognised as an effective method for the identification of low-levels of mtDNA heteroplasmy (Li et al., 2010; Kennedy et al., 2013; Kloss-Brandstätter et al., 2015).

Where differences in the nuclear genomes of MZ twins cannot be identified, it is possible that different levels of mtDNA heteroplasmy might explain the phenotypic discordance between the twins. Several small studies have ventured to investigate such differences in twins discordant for a range of complex traits including, sleep duration (Wrede et al., 2015), neurofibromatosis type 1 (Detjen et al., 2007) and schizophrenia (Li et al., 2017). On the most part, researchers concluded that the phenotypic discordance investigated was not attributable to the difference in mtDNA heteroplasmy.

1.6 Methodological considerations and future strategies

1.6.1 Detecting mosaicism

The above-mentioned findings point to the notion that somatic mosaicism may contribute to the clinical discordance between MZ twins. Thus, an optimal mosaicism search scenario might consist of investigations of cells in an early developmental stage and in the terminally differentiated tissues, and both nuclear DNA and mtDNA, to characterise entirely mosaicism phenotypic consequences in the examined individual.

Deep sequencing can detect genomic differences even with low allele frequency in somatic mosaicism between MZ twins (Li et al., 2013). Morimoto et al. (2017) concur. They identified variants in their study with an allele frequency of ~10%, which is beyond the limits of detection by conventional Sanger sequencing (Rohlin et al, 2009). The authors note, however, that apparent low-frequency allele mutations may in fact be

heterozygous mutations upon validation with Sanger sequencing. Due to possible alignment errors or nonspecific PCR amplification, heterozygous mutations may appear as low allele frequency mutations on NGS data analysis (Morimoto et al., 2017).

1.6.2 Tissue type matters

Reliably detecting de novo somatic mutations is more complex than calling de novo germline mutations, because somatic mutations will vary between tissue types and appear in percentages that are akin to false-positive sequencing rates. High-level mosaicism can result from early postzygotic mutations in an organism, leading to a wide distribution of the variant in many different tissues. In contrast, mutations that occur late during embryogenesis, or postnatally, can be present in only a select few somatic cells (Acuna-Hidalgo, Veltman and Hoischen, 2016). The difficulty in obtaining a wide variety of tissues in humans, due to ethical and practical reasons, makes a comprehensive study of somatic mosaicism difficult. Still, inter-tissue variation of somatic mutation frequencies has been determined by analysing routinely sampled tissues, such as blood, buccal epithelium, skin fibroblasts, hair follicles and urine. Lindhurst et al. (2011) were one of the first to detect somatic mosaicism in MZ twins using exome sequencing. The authors identified the cause of Proteus syndrome, in which a de novo mutation was found in *AKT1* in multiple tissues, but not in DNA from peripheral blood.

Another frequently used DNA source in genetic studies are lymphoblastoid cell lines (LCLs), which are created through *in vitro* infection of B-lymphocytes with Epstein-Barr virus (EBV), although their reliability in replicating a true variant in the donor remains controversial.

The co-twin study design is often used to identify de novo mutations, in which validation of the variant is an important step, especially when interpreting NGS data. Thus,

determining if the mutations identified in LCLs are real or an artefact can be challenging if false-positives are introduced in the study design. There is evidence to suggest that EBV promotes genetic instability in the host (Gruhne et al., 2009), and possibly causes mutations through integration and disintegration into the host's genome (Morissette and Flamand, 2010). This calls into question whether LCLs are a bona fide source of genomic DNA.

Large-scale studies, such as the 1000 Genomes project, have used LCL-derived DNA to characterise human genome variation on a population-based scale. Within this project, the genomes of parent–offspring trios were sequenced (Conrad et al., 2011). In one of the trio analyses, 35 de novo mutations were observed in both untransformed and LCL-derived DNA; however, over 600 de novo mutations were unique to the cell line DNA, which was subsequently validated. This demonstrates that most of the de novo mutations were caused by cell line transformation and culturing. It has therefore been recommended that cell lines should not be used for de novo mutation studies (Veltman and Brunner, 2012).

Other studies have shown that no significant bias is introduced by EBV-transformed B cells by comparing DNA from untransformed material from the same donor. For instance, Nickles et al. (2012) used whole-genome sequencing to assess the genomic signature of a LCL and determined that it is genetically indistinguishable from its genomic counterpart. Similarly, Londin et al. (2011) performed exome sequencing on a family of four individuals using DNA from peripheral blood and LCLs from each individual. The authors note that, although EBV transformation can result in low-level de novo mutations, LCLs remain an appropriate representation of the donor's genome. In summary, low passage LCLs may be a suitable choice in MZ twin research, although, considering caveats from our own research, high passage LCLs will likely introduce significant

anomalies rendering analysis difficult. It is up to the investigators' discretion to determine if LCLs are fit for purpose with the chosen study design, while bearing in mind that independent validation on DNA from uncultured sources is ideal.

1.6.3 Variant calling algorithms

1.6.3.1 Methods in SNV and indel analysis.

Although NGS has proven to be a powerful approach, there remain numerous technical challenges in obtaining an accurate and complete record of sequence variation from the copious data generated, and in converting raw sequence reads into biologically meaningful information (Pirooznia et al., 2014) (Figure 1.4). Granted accurately mapped and calibrated reads, identifying SNVs and indels, not to mention more complex variation such as, insertions, deletions, inversions, CNVs, and multiple base pair substitutions, requires complex statistical models and sophisticated bioinformatics tools (Pirooznia et al., 2014).

Several methods have recently been developed to enhance somatic mutation calling accuracy. Generally, these can be classified into two different types: 1) independent analysis for affected and unaffected tissue datasets followed by SNV type classification using a simple subtraction or a statistical significance test; and 2) simultaneous analysis for matched affected and unaffected tissue datasets using joint probability-based statistical approaches (Xu et al., 2014). Such variant callers belonging to the latter family include SomaticSniper, Strelka, VarScan2 and MuTect2. Unfortunately, agreement among different algorithms is rather low (O'Rawe et al., 2013), thus highlighting the different error models or assumptions underlying each algorithm. Taking the union of two or more variant callers will help reduce the probability of calling false positive SNVs and indels, and avoid biases from one particular caller. However, choosing the ideal tools for

variant calling will depend on a specific set of data type and experimental conditions (Xu et al., 2014).

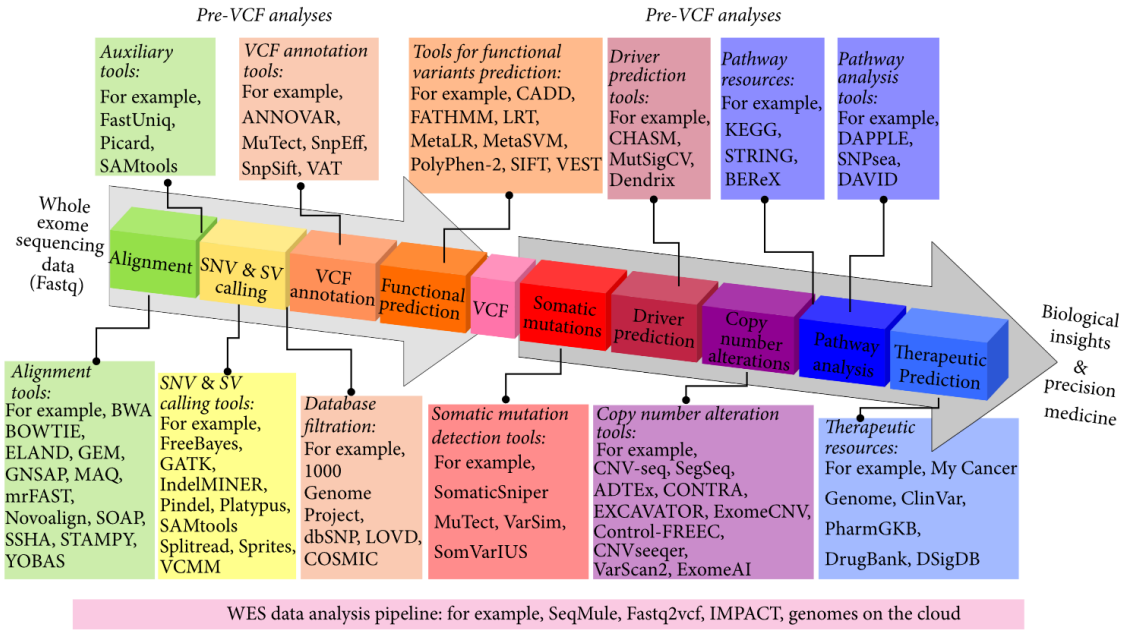


Figure 1.4. Whole-exome sequencing data analysis steps. Novel computational methods and tools have been developed to analyse the full spectrum of exome sequencing data, translating raw fastq files to biological insights and precision medicine. Figure from Hintzsche, Robinson and Tan (2016).

1.6.3.2 Methods in CNV analysis

Genotyping arrays have outplaced the use of comparative genomic hybridisation arrays in CNV detection uses. This is owed to microarray having a higher probe density and thus allowing better resolution of CNV breakpoints (Legault et al., 2015). In fact, there has been a significant upsurge over the past decade in the development of algorithms and technical resolution that have been applied across platforms and programs. Various software packages are available, including Birdsuite, dChip, cnvPartition, Genotyping Console and PennCNV. In a study assessing putative CNV calls made using multiple software, Castellani et al. (2014) suggest using a combination of three programs to optimally identify true CNV calls, giving a trade-off between sensitivity and specificity.

The authors also suggest the use of PennCNV over all other methods when the use of only one tool is preferable (Castellani et al., 2014).

SNP arrays have been used in twin research to look for CNV differences between co-twins. Several studies failed to validate discordant CNVs between the twins investigated. This could be due to a low sensitivity of the CN calling algorithm used. For non-mosaic (pre-twinning) CNV detection, these CN calling tools may be ideal. However, post-twinning CNVs may lead to mosaicism, where further methodological developments in CN calling algorithms would have to be applied.

In a longitudinal study investigating the age-related acquisition of somatic structural variants, four mosaic uniparental disomy events were identified in a cohort of elderly Danish twins, suggesting that somatic changes can result within different body tissues of the same individual (Koldby et al., 2016). Detection of non-mosaic somatic CNV differences between the MZ twins was also performed. Of these, however, only one CNV was eligible for experimental validation with qPCR, and it could not be validated. Numerous studies have searched for non-mosaic somatic CNV differences between MZ twins, with the overall findings being unfruitful. Although this is typically attributed to a small sample size, some studies have recruited a relatively large cohort of twins. For instance, Abdellaoui et al. (2015) searched for post-twinning de novo CNV mutations in 1,097 co-twins, but only two CNVs were validated with qPCR, both of which were present in the same individual. More recently, in a cohort of 100 twin pairs enriched for neurodevelopmental disorders, no postzygotic de novo CNVs were identified (Stamouli et al., 2017).

Another explanation for the lack of discordant CNVs between twins is that somatic CNVs are more likely to be mosaic, necessitating the use of apposite CN calling tools. This

postulate is in line with several studies, including, Magaard Koldby et al. (2016), Bruder et al (2008) and Forsberg et al. (2012). For microarray data, mosaic variants can be detected using the Mosaic Alteration tool (González et al., 2011), and estimation of the mosaic proportion of cells can be performed based on the study by Rodríguez-Santiago et al. (2010). Exome and genome sequencing can also be harnessed to detect mosaic structural abnormalities. A method called MrMosaic, which uses deviations in allele fraction and read coverage from NGS data to detect structural mosaic abnormalities, has proven to be successful (King et al., 2016).

1.6.4 Candidate variant validation

It has been suggested that without independent validation using bench confirmation techniques, such as real-time PCR (qPCR) or droplet digital PCR (ddPCR), CNV calls using computational methods should be at best considered tentative (Castellani et al., 2014). Further, incorporation of family data has been shown to help improve the quality of CNV calls alongside the use of multiple CNV calling methods (Castellani et al., 2014; Legault et al., 2015).

There are several reasons for why microarray CNV calling methods are prone to error, and thus why experimental validation is essential. Firstly, CNV calling using SNP arrays is a relatively new technology, and each platform has its own sensitivity and specificity; secondly, with SNP arrays, the sample in question is compared to reference samples or to a large reference cohort, so only the relative CN is determined; third, compared to whole-genome sequencing, SNP arrays give limited information on the location or orientation of a given CNV; finally, SNP arrays rely on amplification of DNA and measurements using (fluorescence) intensity, thus technical variation can influence experimental outcome (Brosens et al, 2016).

In terms of validating mosaic somatic SNVs and indels, other methods alongside Sanger sequencing are commonly used, including high-resolution melting analysis, allele-specific PCR, immunohistochemistry, pyrosequencing, and SNaPshot. These techniques are especially useful in candidate variant screening and result verification in the NGS era (Gajecka, 2015).

1.6.5 Catching the wave downstream

MZ twins that are interrogated at multiple omics levels or investigated under a variety of environmental conditions are offering remarkable insights. The following section will outline how the epigenome, transcriptome and proteome of MZ twins can be studied in an integrative manner to provide a more systematic exploration of gene-environmental interaction, and what current research in this area has to offer.

1.6.5.1 Epigenetics

Epigenesis, a term from which 'epigenetics' derives, is an early embryological theory postulating the development of an embryo from the successive differentiation and elaboration of an originally undifferentiated fertilised egg (as opposed to preformation, an outmoded developmental model that suggested complex organisms are already completely formed in the germ cell and develop merely by enlargement). The term 'epigenetics' was initially coined by developmental biologists who sought to explain the mechanism by which gene-gene and gene-environment interactions fashioned the phenotype of an individual during development (Youngson and Whitelaw, 2008). However, the term has found new meaning in modern parlance, even though a consensual definition is still lacking. Epigenetics can be generically described as alterations in genomic function that are controlled by potentially reversible heritable factors, without altering the DNA sequence.

In other words, epigenetic regulation of gene expression uses reversible modifications of DNA and chromatin structure to mediate the interaction of the genome with a variety of environmental factors and to generate changes in the patterns of gene expression in response to these factors. Such mechanisms include, DNA methylation, histone modifications, ATP-based chromatin remodelling, transcription factor-binding mechanisms and non-coding RNA-mediated gene silencing (Ketelaar, Hofstra and Hayden, 2011; Bell and Spector, 2011).

Epigenetic discordance can be observed to variable degrees in MZ twins, estimates for which have been shown to be affected by sample size, tissue type, age and CpG island selection (Czyz et al., 2012). Indeed, a major challenge in epigenetic research is that epigenetic differences are often tissue- and cell-specific, and sometimes it may be practically impossible to obtain the ideal tissue source. As previously discussed, MZ twins can be sub-classified depending on whether they share the same placenta or not. Chronicity, however, has been shown to affect epigenetic profiles between twins, where MC twins have been shown to have more disparate epigenetic profiles than DC twins (Kaminsky et al., 2009). Notwithstanding, most studies that investigate MZ twins have constituted them into a uniform group.

Epigenetic and transcriptomic modifications have been reported in twins discordant for a range of complex traits, including cancer (Heyn et al., 2012), Alzheimer's disease (Poulsen et al., 2007), autism spectrum disorders (Loke, Hannan and Craig, 2015), schizophrenia and bipolar disorder (Dempster et al., 2011), multiple sclerosis (Handunnetthi, Handel and Ramagopalan, 2010), pain sensitivity (Bell et al., 2014), psoriasis (Gervin et al., 2012), and type 1 diabetes (Stefan et al., 2014).

1.6.5.2 Proteomics

Where whole-genome and exome sequencing has been less beneficial, attention is increasingly switching to RNA sequencing and epigenetics to explore potentially-missed disease-causing mechanisms. Proteomics, however, has the potential to capture all the variations arising from genomic, transcriptomic and epigenetics changes.

Alternative splicing and post-translational protein modifications entail that the number of proteins can be two orders of magnitudes higher than the number of genes (Zierer et al., 2015). Proteomic techniques in current practice, such as immunoassays, protein arrays or mass spectrometry, are limited in that they can measure only a small fraction of the proteome. The most comprehensive analysis of the human proteome to date consists of over 18,000 proteins composed from 10,000 mass spectrometry experiments, across various tissues (Wilhelm et al., 2014)

Analysing the proteomes of human biological fluids (such as, serum, urine, saliva, synovial and cerebral spinal fluids) among individuals with a shared genetic background, but possibly different environmental and/or epigenetic influences, may cast welcoming new light on the identification of putative disease-associated biomarkers. Although comprehensive proteomics studies on discordant MZ twins are still lacking, the few studies that have implemented the co-twin design have provided some important insights. Disorders investigated include systemic autoimmune diseases (O'Hanlon et al., 2011), bipolar disorder (Kazuno et al., 2013), and chronic fatigue syndrome (Ciregia et al., 2013). In this thesis, a proteomic profiling of serum samples from MZ twins discordant for ischaemic stroke was analysed through a label-free pipeline. The finding of a distinct proteomic profile associated with ischaemic stroke raises the possibility of patient-centred diagnostic, prognostic and therapeutic strategies in future.

1.6.5.3 Metabolomics

The metabolome refers to the complete set of low molecular weight compounds in a sample. These compounds are the substrates and by-products of enzymatic reactions and have a direct effect on the phenotype of the cell. Thus, metabolomics aims at determining a sample's profile of these compounds at a specified time under specific environmental conditions. Similar to proteomics, there are, to date, no analytical methods at our disposal that can determine and quantify all metabolites in a single experiment. Remarkably, however, the Human Metabolome Database (Wishart et al., 2013) contains more than 40,000 distinct metabolites from different tissues. These low molecular weight compounds are the closest link to phenotype. Although genomics and proteomics have provided extensive information regarding the genotype, it may be difficult to elicit direct information about the clinical phenotype being investigated.

Several new omics technologies have emerged recently, which deserve to be mentioned but cannot be described in depth due to their branching themes lying beyond the scope of this thesis. These are glycomics (post-translational modifications), microbiomics, and phenomics (Zierer et al., 2015). The currently employed approach to data analysis in the omics era is to take the reductionist stance by focusing on individual factors. However, these emerging technologies have the capacity to broaden investigations into complex traits by an unprecedented scale. The 'holistic' approach, or systems biology, integrates data from different experiments to gain an understanding of the system as a whole (Zierer et al., 2015).

Large-scale multi-omics investigations into MZ twins are currently being carried out. For instance, the MuTher study, consisting of several hundred female twins, has been evaluated globally at the genome, transcriptome, metabolome, and microbiome levels

(Fizelova et al., 2016). The data has given valuable insights into the genetic control of molecular traits, biological pathways involved in metabolic syndrome, and the heritability of gut microbiota (Heinig et al., 2010). Another human reference population study, dubbed Metabolic Syndrome in Man (METSIM), consists of a cohort of about 10,000 Finnish men. Like the MuTher population, METSIM participants have been characterised clinically for a variety of metabolic and cardiovascular traits at the genomic, transcriptomic, and metabolomics levels (Civelek et al., 2017; Laakso et al., 2017; Fizelova et al., 2016).

1.7 Conclusion

The evidence presented above challenges the use of the expression 'identical twins', as MZ twins are not truly identical. Paradoxically, they can be very much alike, while being differentially affected by postzygotic (epi)genetic mutations and environmental influences, which polarise the genotype and phenotype within a given pair.

Whole-genome, epigenetic and transcriptomic profiling across multiple tissues and cell types in large cohorts of discordant MZ twins will continue to shed light on gene-environment interactions and the aetiology of complex traits. Other omics technologies, such as proteomics and metabolomics, can provide insights into downstream consequences of disease, and are increasingly being incorporated into twin research. Thus, future twin studies will likely be themed on 'multi-omics' – the integration of multiple types of omics data. This will require a conceptual shift in the research paradigm and inject new energy in the study of human genetics.

1.8 Hypothesis and thesis objectives

Discordant MZ twins provide a unique opportunity to study the genetics of complex disorders, where potential disease-causing variants are difficult to identify against a background of thousands of randomly occurring non-pathogenic polymorphisms found throughout the genome. The overwhelming majority of these randomly occurring polymorphisms, i.e. those inherited through the parental germline, will be identical between co-twins. Where MZ twins tend to share a similar environment, we believe a genetic cause is likely to explain their discordant phenotypes. Clearly, epigenetic mechanisms could also be responsible.

We have obtained DNA samples from thirteen MZ twins discordant for a variety of clinical phenotypes. Our proposed work provides a rare opportunity to untangle the genetic basis for complex disorders, and to correlate genetic variation with pathway analyses to gain insights into their molecular mechanisms. We hypothesise that phenotypic discordance between identical twins may involve postzygotic de novo mutations.

The aims and objectives can be outlined as follows:

- **Chapter 3:** To look for *C9orf72* repeat polymorphic variations between four MZ twin pairs discordant for ALS, using rpPCR and Southern blotting. To sequence 25 ALS-linked genes on an NGS panel, and confirm any potential SNPs found with Sanger sequencing.
- **Chapter 4**: To carry out whole-exome sequencing of discordant MZ twins and compare different bioinformatics methodologies for optimised filtering. To validate potential differences found with Sanger sequencing and other SNP genotyping assays capable of detecting low-level mosaicism. To cross-compare

the exome sequence data of DNA extracted from different tissue sources where available. To perform parent-offspring trio analysis, where parental DNA is available, to look for pre-twinning de novo mutations that are shared between the twins.

- **Chapter 5**: To determine if CNV differences exist in our cohort of MZ twin pairs, by unifying the results of two CNV calling methods. To validate potentially interesting CNVs with additional computational and experimental methods.
- **Chapter 6**: To perform cytogenetic, biochemical and proteomic profile comparisons of twins discordant for ischaemic stroke. To identify possible disease-associated biomarkers and provide insights into pathogenic mechanisms by carrying out pathway analysis on identified proteins.

Chapter 2. Materials and Methods

2.1 Description of twin pairs

Coriell Cell Repository DNA was obtained from five pairs of MZ twins discordant for ALS (218 and 318; 421 and 422; 242 and 243), Tourette's syndrome (489 and 490), and Parkinson's disease (PD821 and PD161). DNA was also obtained from the parents of the twins discordant for Tourette's syndrome (487 and 488). The Coriell Institute for Medical Research (Camden, NJ, USA) provide essential research reagents to the scientific community by establishing, verifying, maintaining, and distributing cell cultures and DNA derived from cell cultures. These collections are supported by funds from the National Institutes of Health (NIH) and several foundations.

Genomic DNA samples of two twin pairs discordant for schizophrenia were obtained; one pair from Professor Sir Robin Murray based at King's College London, Institute of Psychiatry (IP16 and IP17), and one pair from Dr Takeo Yoshikawa based at the RIKEN Institute, Brain Science Institute (RT1a and RT1b).

A further six discordant MZ twins were recruited, and DNA was extracted from blood and/or saliva samples for molecular/genetic analysis. These twins were discordant for stroke (HG and KG), ALS (LAS and SUS), lactase non-persistence (KEL and KIR), inclusion body myositis (AFF and UNAFF), dystonia (VF and LF), and ADHD (RP and OH). DNA from the parents of RP and OH was also obtained (DS and DV). Further details of subjects can be found on Table 2.1. Written informed consent was obtained from all participants prior to study entry (see Appendix A for the patient information sheet and consent form).

Subject	*Status*	*Diagnosis*	*Sex*	*Age*	*Age of onset*	*Ethnicity*	*DNA source*
LAS	Proband	ALS	F	71	67	Caucasian	Saliva
SUS	MZ twin	-	F	71	-	Caucasian	Saliva
218	Proband	ALS	M	54	50	Caucasian	LCL
318	MZ twin	-	M	54	-	Caucasian	LCL
421	Proband	ALS	F	58	55	Caucasian	LCL
422	MZ twin	-	F	58	-	Caucasian	LCL
242	Proband	ALS	M	35	34	Caucasian	LCL
243	MZ twin	-	M	35	-	Caucasian	LCL
KG(s) KG(b)	Proband	Stroke	F	62	56	Caucasian	Saliva Blood
HG(s) HG(b)	MZ twin	-	F	62	-	Caucasian	Saliva Blood
KEL	Proband	LNP	F	23	5	Caucasian	Saliva
KIR	MZ twin	-	F	23	-	Caucasian	Saliva
AFF	Proband	IBM	M	68	66	Caucasian	Blood, Saliva
UNAFF	MZ twin	-	M	68	-	Caucasian	Blood, Saliva
DS	Father	-	M	UN	-	Caucasian	Saliva
DV	Mother	-	F	UN	-	Afro-Caribbean	Saliva
RP	Proband	ADHD	M	10	7	Mixed	Saliva
OH	MZ twin	-	M	10	-	Mixed	Saliva
487	Father	TS, OCD, ADHD	M	44	6	American Indian	Blood
488	Mother	-	F	44	-	Caucasian	Blood
489	MZ twin	-	M	15	-	Mixed	Blood
490	Proband	TS	M	15	7	Mixed	Blood
PD821	Proband	PD	M	40	30	Caucasian	Blood
PD161	MZ twin	-	M	40	-	Caucasian	Blood
VF	Proband	Dystonia/HSP	F	46	34	Caucasian	Blood
LF	MZ twin	-	F	46	-	Caucasian	Blood
RT1a	MZ twin	SCPD	M	UN	UN	Asian	Blood
RT1b	Proband	SCZ	M	UN	UN	Asian	Blood
IP16	Proband	SCZ	M	UN	UN	Caucasian	Saliva
IP17	MZ twin	-	M	UN	-	Caucasian	Saliva

Table 2.1 Clinical characteristics and demographic information on the MZ twin cohort. UN = unknown; ALS = amyotrophic lateral sclerosis; LNP = lactase non-persistence; IBM = inclusion body myositis; ADHD = Attention deficit hyperactivity disorder; OCD = obsessive compulsive disorder; TS = Tourette's syndrome; PD = Parkinson's disease; HSP = hereditary spastic paraplegia; SCZ = schizophrenia; SCPD = schizotypal personality disorder.

2.2 Sample preparation and quantification

2.2.1 DNA extraction from saliva

The Oragene DNA Kit (DNA Genotek Inc., Kanata, Canada) was used to extract total DNA from saliva samples according to the manufacturer's instructions. Briefly, 500µL of the mixed saliva and Oragene solution was placed into a 1.5mL microcentrifuge tube and incubated for 2hrs at 50°C. 20µL of Oragene DNA purifier was added, and the tube was briefly vortexed. The solution was incubated on ice for 10mins, then centrifuged at room temperature for 10mins at 13,000rpm. The resulting supernatant was transferred into a new 1.5mL microcentrifuge tube, and the remaining pellet was discarded. 95% ethanol was added to the supernatant and the solution was gently mixed by inverting the tube several times; this was left at room temperature for 10mins to allow the DNA to precipitate. The solution was then centrifuged for 2mins at 13,000rpm, and the supernatant was carefully removed and discarded. For further purification, 200µL of 70% ethanol was added and then removed after 1min, taking care not to disturb the pellet. 50µL of Tris-EDTA buffer was added to the tube to dissolve the DNA pellet. The tube was briefly vortexed and incubated for 1hr at 50°C.

2.2.2 DNA extraction from blood

Blood from subjects was taken in ethylenediaminetetraacetic acid (EDTA) bottles and genomic DNA was extracted from whole blood using a FlexiGene kit (Quiagen) according to manufacturer's instructions. For each sample, 300µL of whole blood was mixed with 750µL of buffer (FG1) to lyse the cells. The sample was centrifuged for 20secs at 13,000rpm, and the supernatant was discarded leaving only the pellet. 150µL of protease-contain buffer (FG2) was added to the tube and then vortexed until the pellet became homogenised. The sample was centrifuged for 10secs and incubated at 65°C for

5secs in a heating block. 150µL of 100% isopropanolol was added to the tube and inverted several times to allow precipitation of the DNA. The sample was centrifuged for 3mins at 13,000rpm and the supernatant was discarded. 150µL of 70% ethanol was added to the sample and vortexed for 5secs before centrifuging for 3mins at 13,000rpm. The supernatant was discarded and the pellet air dried. 200µL of buffer (FG3) was added and the sample and briefly vortexed to dissolve the DNA.

2.2.3 DNA concentration and purity

A NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies) was used to measure genomic DNA concentration and purity. 1µL of DNase and RNase-free distilled deionised water (dH2O) was loaded onto the pedestal to obtain a blank reading. 1µL of DNA from all participants was individually loaded to determine the concentration. Absorbance measurements determine molecules absorbing at a specific wavelength. DNA will absorb at 260nm and will contribute to the total absorbance. For quality control, the 260/280nm and 260/230nm absorbance ratios are used to assess the purity of DNA. For DNA, a 260/280nm absorbance ratio of ~1.8 is generally accepted as adequate purity. The 260/230nm absorbance ratio is used as a secondary measure of nucleic acid purity, and a value within the range of 2.0-2.2 is considered acceptable. 1µL of all DNA samples were run on a 1.3% agarose gel to determine the relative concentrations and degree of potential degradation.

2.2.4 Monozygosity testing

All twin pairs were tested for their zygosity status before proceeding with the study. Analysis was performed using an in-house kit, consisting of eight microsatellite markers, an Amelogenin marker, and a SRY marker. DNA from blood was also screened using a commercial kit consisting of 16 markers, including fifteen STRs and Amelogenin,

according to the manufacturer's recommendations. This was carried out by John Short at St. George's NHS Trust.

2.3 Initial screen of known disease-associated variants

Some samples were screened for known pathogenic variants before being put forward for whole-exome sequencing analysis. A pathogenic hexanucleotide repeat expansion within the *C9orf72* gene has been identified as the major cause of ALS. Variation in the hexanucleotide repeat number was first assessed by repeat-primed PCR (rpPCR). A modified Southern blot method was used to confirm the rpPCR detected expansions where sufficient DNA allowed.

2.3.1 Repeat-primed PCR

To provide a qualitative assessment of the presence of an expanded $(GGGGCC)_n$ hexanucleotide repeat in *C9ORF72,* a rpPCR reaction was performed in four twin pairs discordant for ALS. Briefly, 100ng of genomic DNA were used as template in a final volume of 28μL containing 14μL of FastStart PCR Master Mix (Roche Applied Science, Indianapolis, IN, USA), and a final concentration of 0.18mM 7-deaza-dGTP (New England Biolabs Inc., Ipswich, MA, USA), 1x Q-Solution (Qiagen Inc., Valencia, CA, USA), 7% DMSO (Qiagen), 0.9mM $MgCl_2$ (Qiagen), 0.7μM reverse primer consisting of ~four GGGGCC repeats with an anchor tail, 1.4μM 6FAM-fluorescent labelled forward primer located 280bp telomeric to the repeat sequence, and 1.4μM anchor primer corresponding to the anchor tail of the reverse primer. Primers used for rpPCR are shown in Table 2.2 (Kobayashi et al., 2011; Warner et al., 1996). A touchdown PCR cycling program was used where the annealing temperature was gradually lowered from 70°C to 56°C in 2°C increments with a 3min extension time for each cycle.

The rpPCR is designed so that the reverse primer binds at different points within the repeat expansion to produce multiple amplicons of incrementally larger size. The lower concentration of this primer in the reaction means that it is exhausted during the initial PCR cycles, after which the anchor primer is preferentially used as the reverse primer. Fragment length analysis was performed on an ABI 3730xl genetic analyser (Applied Biosystems Inc., Foster City, CA, USA), and data analysed using GeneScan software (version 4, ABI). Repeat expansions produce a characteristic saw-tooth pattern with a 6-bp periodicity.

This protocol was carried out under the supervision of Gary Adamson at UCL, Institute of Neurology.

Primer name	***Primer sequence***	***Concentration***	***Modification***
ALSFTDf_(6FAM)	6-FAM-AGT CGC TAG AGG CGA AAG C	0.05 µmol	Modified DNA Oligos
ALSFTDr	TAC GCA TCC CAG TTT GAG ACG GGG GCC GGG GCC GGG GCC GGG G	0.05 µmol	Unmodified DNA Oligos
ALSFTDanchor	TAC GCA TCC CAG TTT GAG ACG	0.05 µmol	Unmodified DNA Oligos

Table 2.2. Primers used for rpPCR

2.3.2 Southern blotting

10µg DNA was digested with HindIII and XbaI overnight. DNA fragments were separated on a 0.9% TRIS-Borat-EDTA (TBE) agarose gel, transferred by alkali blotting onto an Amersham Hybond NTM-XL membrane (GE Healthcare, Fisher Scientific, Germany) and hybridised to a ^{32}P-labelled probe overnight. After washing, X-ray films were exposed for 4–6 days at −80°C. BstEII digested lambda DNA was used as a size marker for estimating repeat lengths. Minimal repeat sizes were used for calculation. This protocol was carried out by John Polke at UCL, Institute of Neurology.

2.3.3 Next-generation sequencing panels

2.3.3.1 Amyotrophic lateral sclerosis

Genomic DNA from one twin pair discordant for ALS (LAS and SUS) was processed on a gene panel designed by Morgan et al., 2015, which uses the Illumina TruSeq Custom Amplicon implemented on an Illumina MiSeq platform. This panel utilises PCR amplicon-based target enrichment and screens for variants in 25 ALS disease genes. Sequence analysis involved the full exomes of 10 genes strongly implicated in ALS, and specific genomic areas where disease-causing mutations cluster in 15 other minor or unproven ALS-linked genes. Probes were created using Illumina TruSeq custom amplicon assay DesignStudio v1.6 (http://www.illumina.com/applications/designstudio.ilmn).

2.3.3.2 Hereditary spastic paraplegia

LF presented with sudden onset for HSP at 34 years with apparent recurrent 20min dystonic cramps in her right leg. This was triggered by exercise, driving, caffeine and alcohol. Her identical twin sister (VF) and 16-year-old son may have similar features of spasticity. A diagnosis of complicated hereditary spastic paraparesis superimposed with dystonia was made. The patient did not respond well to L-Dopa, and there was no diurnal variation in presenting symptoms. Sequencing and analysis of genomic DNA from the twin affected with HSP (LF) was carried out on a multi-gene panel at Sheffield Children's NHS Foundation Trust, using the Agilent SureSelect Neurogenetic panel Version 1 with Illumina MiSeq Analysis pipeline: Version 2. LF was also screened for GLUT1 deficiency syndrome and DYT1 early-onset primary dystonia.

2.3.3.3 Lactase non-persistence

Additionally, Sanger sequencing was used to detect known SNPs in twins discordant for lactase non-persistence (KIR and KEL) (methods described below).

2.4 Next-generation sequencing

2.4.1 Whole-exome sequencing

Each exome captured sequencing library was produced from either one of two exome capture kits and sequencing systems. For the first set of samples (n=16), genomic DNA (3µg) from each sample was sheared to a mean fragment size of 150bp, with the use of focused acoustic technology (Covaris). These fragments were end repaired, 3' adenylated, and ligated with Illumina paired-end sequencing adapters according to the SureSelect Human All Exon V4 (Agilent Technologies) exome enrichment kit protocol. Between each of these steps, samples were purified using Agencourt AMPure XP beads to remove non-specific DNA. The sequence library was hybridised with biotinylated 120bp RNA library baits in PCR plates on a thermocycler at 65^{o}C for 24hrs. Streptavidin-coated superparamagnetic beads were used to preserve DNA bound to the RNA probes, and the unbound fraction was discarded using SureSelect wash buffers. After eluting the exome-enriched pool of DNA, a low-cycle PCR was used to amplify the specific region of DNA. This was followed by sequencing on a HiSeq2000 (Illumina) with 100bp paired-end reads.

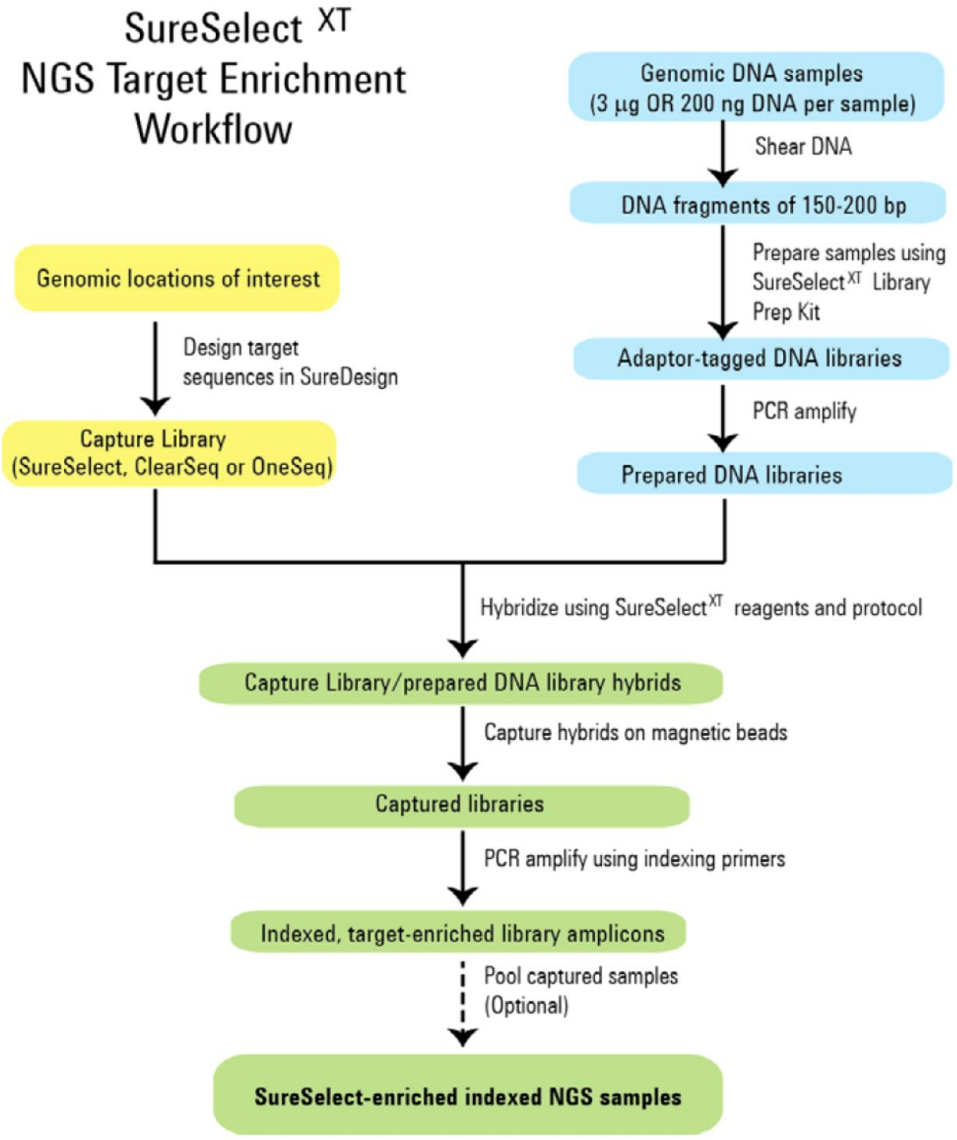


Figure 2.1. Overall target-enriched sequencing sample preparation workflow. Briefly, genomic DNA is fragmented, denatured, and hybridised with capture oligos during library preparation for high-throughput sequencing. The captured sequences are then enriched with streptavidin-conjugated paramagnetic beads and further amplified before being subjected to Illumina sequencing (figure from Agilent's protocol).

2.4.2 Pipeline 1

Sequence reads were aligned to the reference human genome (UCSC Genome Browser hg19, GRCh37) using Novoalign (Novocraft Technologies). Over 7.0Gb of sequences were uniquely aligned for each subject, where >85% of the coding bases of the GENCODE-defined exome were represented by at least 20 reads. Duplicate reads, resulting from PCR clonality or optical duplicates, and reads mapping to multiple

locations, were excluded from downstream analysis. Depth and breadth of sequence coverage was calculated using custom scripts and the BedTools package (Quinlan and Hall, 2010). Single-nucleotide substitutions and small insertions or deletions (indels) were identified with SAMtools (Li et al., 2009) and were annotated with the ANNOVAR software package.

Pairwise comparisons were carried out for all the twin pairs. Sites with alleleic representation of >15% in one twin, that have no representation in the corresponding twin, were identified using the VarScan v2.2.3 package. After exclusion of repetitive regions of the genome by retrieving repetitive DNA tracks from the UCSC genome browser (Karolchik et al., 2013), candidate discrepant sites were manually reviewed in both twins through inspection of the alignments using the Integrative Genomics Viewer (IGV) software v2.1 (Robinson et al., 2011). The variants that passed this stringent multistep filtering criteria would be validated by Sanger sequencing, and SNP assays capable of detecting low-level mosaicism (see below). For the first set of DNA samples, sequence capture, sequencing, alignment and variant calling was performed by Michael Simpson at King's College London, Division of Genetics and Molecular Medicine.

2.4.3 Pipeline 2

Other filtering approaches applicable to finding genetic differences between MZ twins were explored. Peter De Rijk (University of Antwerp, Department of Molecular Genetics) was recruited into the study on a collaborative basis to analyse the data using a software pipeline developed in-house called GenomeComb (http://genomecomb.sourceforge.net/). By using the query language provided, it can be used to compare, annotate and filter the results of NGS data, and SNVs can be ranked based on their probability of being an error (Reumers et al., 2012). The pipeline used fastq-mcf v1.1.2 (https://code.google.com/p/ea-

utils/) for adapter clipping. Reads were then aligned to the reference human genome (hg19) using the Burrows-Wheeler algorithm, version 0.7.5a (Li and Durbin, 2009), and the resulting sequence alignment map (SAM) files were converted to Binary Alignment and Mapping (BAM) files using SAMtools (Li et al., 2009). The resulting BAM files were sorted, and duplicates were removed using Picard, version 1.87 (http://broadinstitute.github.io/picard/). Realignment in the neighbouring regions of indels was performed with GenomeAnalysisToolKit (GATK), version 2.4–9 (McKenna et al., 2010). Variants with a coverage ≥5 were called using both GATK and SAMtools, version 0.1.19-44428cd (Li et al., 2009). At this initial stage, positions with a coverage <5 or a score <30 were considered unsequenced. The resulting variant sets of different individuals were combined and annotated using GenomeComb. The GenomeComb query tool was then used to select variants that differed between the twins, but complied to strict quality filtering rules:

Variants in regions with a clustering of many variants were removed, as well as in regions of known microsatellites, simple tandem repeats and segmental duplications. Variants were also removed if the coverage at the variant site was <15 or the quality score <70 in either sample. Furthermore, variants were retained only if the genotype call by GATK and SAMtools was the same, and were consequently combined and annotated with a variety of different databases (in-house exomes, dbSNP, 1000 Genomes Project).

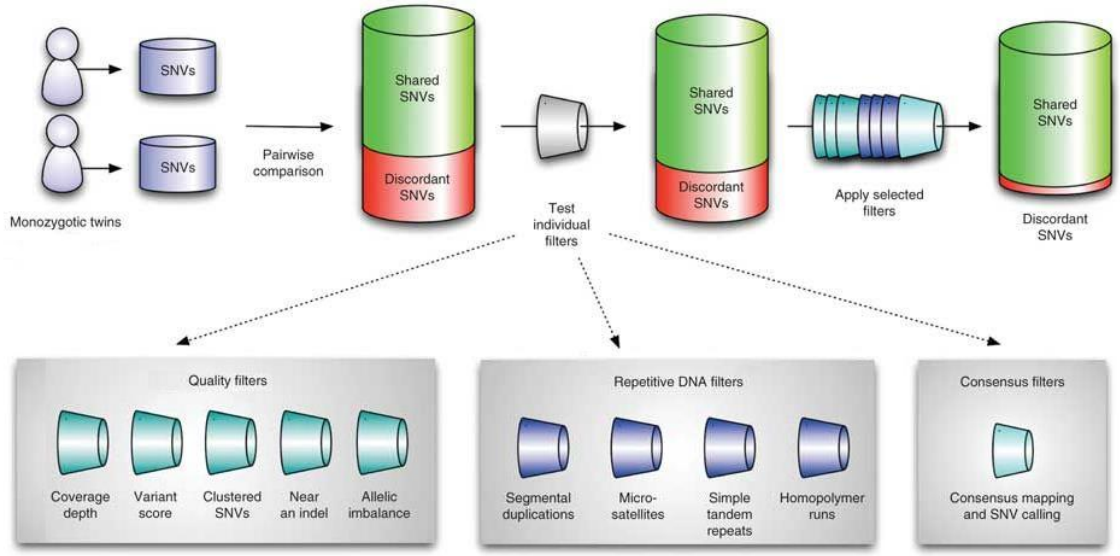


Figure 2.2. All discordant and shared SNVs between co-twins were calculated. As most discordant SNVs are assumed to be false positive errors, each filter was systematically tested to reduce as many discordances, while retaining as many shared variants as possible. Three types of filters were applied: 1) Quality filters, to remove regions of inferior sequencing quality, 2) repetitive DNA filters, to eliminate errors due to incorrect mapping to the reference genome, and 3) consensus filters, to retain SNVs identified with different sequence mapping and variation calling algorithms. Individual filters were combined to remove a maximum number of discordant variants between the twin exomes (adapted from Reumers et al., 2012).

2.4.4 Pipeline 3

As we recruited more discordant twins in our study, updated bioinformatic approaches were made available, with higher specificity in detecting mosaic variants. Thus, a novel pipeline developed by Alan Pittman at UCL Institute of Neurology was used as our main in-silico prioritisation tool. Previously-generated and new FASTQ files underwent streamlined analysis.

2.4.4.1 DNA library construction of remaining samples

Whole-exome sequencing libraries for the next set of samples (n=16) were prepared with Agilent SureSelect V6 and sequenced on an Illumina HiSeq3000 using a 75-bp paired-end reads protocol. This was carried out by Deborah Hughes at UCL, Institute of Neurology.

2.4.4.2 Whole-exome capture

Alignment of the previous and newly-sequenced exomes to the human reference genome (UCSC hg19), and variant calling and annotation was performed with an in-house pipeline developed by Alan Pittman. Briefly, this involves alignment with NovoAlign, removal of PCR-duplicates with Picard Tools followed by (sample-paired) local realignment around indels and germline variant calling with HaplotypeCaller according to the GATK best practices.

Potentially mosaic variants were identified with GATK MuTect2 (version 2.0) and VarScan2 (version 2.4.3), using each pair as reference to one-another (described below). The raw list of SNVs and indels were then filtered using ANNOVAR. Variants in splicing regions, 5'UTR, 3'UTR and protein-coding regions, such as missense, frameshift, stop-loss and stop gain mutations, were considered. Priority was given to rare variants (<1% in public databases, including 1000 Genomes project, NHLBI Exome Variant Server, Complete Genomics 69, and Exome Aggregation Consortium). Furthermore, we have an in-house set of approximately six thousand exomes encompassing controls, rare diseases for cross-checking any shortlisted candidate variants, and for sequencing artefact removal.

2.4.4.3 Comparison of two somatic variant calling methods

VarScan2 and MuTect2 algorithm tools were used to identify de novo variants using the somatic mutation calling method. The union of SNVs called by both variant callers was taken forward for tiering, filtering and manual review. This method was used on discordant MZ twins by treating the affected twin as the 'tumour' sample and the unaffected twin as the 'normal' sample (and viscera to detect somatic mutations present in the unaffected twin but not in the affected twin). MuTect2 uses a Bayesian classifier

approach to detect somatic mutations with very low allele fractions, requiring only few supporting reads, followed by carefully tuned filters that ensure high specificity. In this study, MuTect2 was run under the High-Confidence mode with its default parameter settings. Low quality sequenced data was first removed, followed by variant detection in the 'tumour' sample using a Bayesian classifier. A filtering step was then applied, which removes false positive variants caused by sequencing artefacts. Finally, variants are classified as somatic or germline by a second Bayesian classifier (Cibulskis et al, 2013).

VarScan2 reads BAM files from 'tumour' and 'normal' samples simultaneously to heuristically call a genotype at positions achieving certain thresholds of coverage and quality. It uses a one-tailed Fisher's exact test to calculate the significance of the difference in allele frequencies of the normal and tumour sample based on the number of reads supporting each allele. We used a cut-off value of Fisher's P-value <0.05. If the resulting p-value meets this significance threshold, variants are classified as somatic (if the tumour call was different from the normal and the normal call was the same as the reference), loss of heterozygosity (if the tumour variant call was not heterozygous but the normal variant call was heterozygous), or unknown (if the tumour call was different from the normal call and both calls were different from the reference). The variant is classified as germline if the difference does not meet the significance threshold.

In summary, VarScan2 provides sensitive detection of high-quality somatic SNVs, whereas MuTect2 provides sensitive detection of low allelic-fraction. The compatibility of the output VCF files between different methods was examined using Microsoft Excel. Somatic variants which were common to both algorithms were retained for downstream analysis.

Tools	*Version*	*URL*	*Remark*	*Release date*
VarScan	2.4.3	http://dkoboldt.github.io/varscan	Sensitive detection of high-quality somatic SNVs	Dec. 2016
MuTect2	2.0	https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php	Sensitive detection of low allelic-fraction	Nov. 2015

Table 2.3. Details of algorithm tools for somatic SNV detection within NGS data

2.4.5 DNA variant and gene prioritisation

Putative discordant variants called by both MuTect2 and VarScan2 were further filtered according to low variant quality (VQS<90), common variants (MAF>0.01) as reported in public databases (1000g, ExAC, cg69), and variants in regions containing segmental duplications. This stringent filtering criteria provided a manageable list for evaluation of candidate variant sites; thus, variants with genomic locations in exonic, 5'UTR, 3'UTR, splice site and promoter regions were retained.

For the detection of concordant variants, in addition to the filtering criteria above, variants were analysed for their potential deleterious effects using the polymorphism phenotyping v2 (PolyPhen2) and Sorting Tolerant from Intolerant (SIFT) algorithms (Kumar et al., 2009; Adzhubei et al., 2010). Within the prioritised variants, those harbouring truncating mutations or mutations predicted to be damaging were considered the most promising candidates. The priority order of variants, from most to least damaging, were as follows: frameshift, nonsense, splice site, missense and non-stop. All missense variants predicted to be benign were removed.

Where parental DNA was available, provisional postzygotic de novo mutations identified in the twins were excluded if they were detected in either parents. To identify germline

de novo mutations shared between the twins, parent-offspring trio analysis was performed. Contrary to single sample calling, where samples are analysed individually, joint genotyping was performed on all samples according to GATK best practices. Analysing variants simultaneously across all samples has several advantages, including 1) Being able to better distinguish between homozygous reference sites and sites with missing data; 2) Having higher sensitivity for low-frequency variants. Joint calling enables the 'rescuing' of genotype calls at sites where there's low coverage but other samples within the call set have a confident variant at that location; 3) Being able to more efficiently filter out false positives. Studies have shown that GATK's Variant Quality Score Recalibration (VQSR) provides better calling accuracy than simply using hard filtering (Pirooznia et al., 2014). VQSR builds a Gaussian mixture model by looking at the annotation values over a subset of the input call set, then by using machine learning algorithms, determines the annotation profile of good and bad calls, and evaluates all input variants. Joint calling provides a large enough dataset for accurate error modelling and ensures that filtering is applied uniformly across all samples.

The data were filtered in Excel to identify concordant mutations in co-twins, but absent in the parents and other samples in the dataset. In the twins discordant for Tourette's syndrome (489 and 490), because the father (487) was also affected, inherited pathogenic mutations from the father were also investigated.

PubMed, Online Mendelian Inheritance in Man (OMIM), NIH Genetic Testing Registry (GTR) and DisGeNET were reviewed for previous publications regarding candidate genes. In addition, gene databases for the disorders investigated in this study were also searched, including Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD), ALSGene, PDGene, Schizophrenia Database (SZDB), Schizophrenia Gene (SZGene).

All identified genes for each disorder were pooled together to form a comprehensive list, and was cross-checked with germline variants in each twin pair (Figure 2.3).

With the possibility of finding potentially-damaging rare variants in novel genes, after prediction of functional effects of the selected variants, all non-exonic variants and variants indicated as synonymous were removed (Figure 2.3). Genomic evolutionary rate profiling (GERP++) scores, a measure of evolutionary constraint at each base derived by aligning 29 mammalian genomes, were also used to estimate the conservation of each variant. Negative scores indicate a lack of conservation and high positive scores indicate the most conserved nucleotide positions among multiple species (Davydov et al., 2010).

All concordant and discordant variants between twin siblings were manually reviewed in IGV prior to validation (see below).

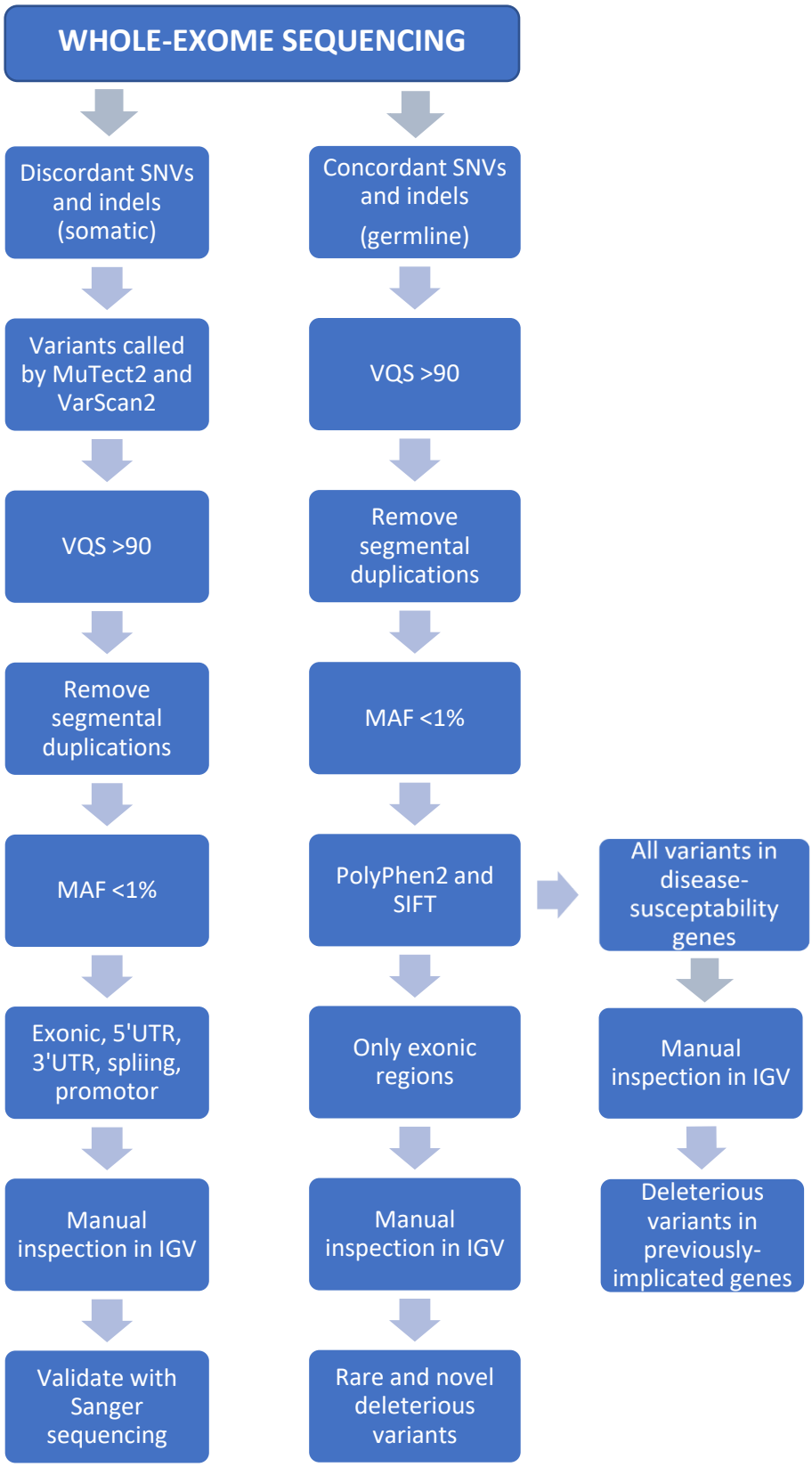


Figure 2.3. General overview of the methods used to detect discordant and concordant variants in MZ twins. The MAF filter was not applicable for twins discordant for lactase non-persistence, as common polymorphisms are associated with the condition, and Tourette's syndrome, where modifier variants were more likely to play a role.

2.4.6 Genome-wide SNP genotyping

2.4.6.1 Wet lab processing

300ng of high-quality genomic DNA from each subject was whole-genome amplified overnight at 37ºC for 20-24hrs in a deep well plate, then fragmented at 37ºC for 1hr15mins in a hybridisation oven, precipitated and resuspended in hybridisation buffer. Samples were denatured then taken from the plate and loaded onto the chips using a liquid handling robot (Freedom Evo, Tecan Ltd, Switzerland). Hybridisation took place overnight for 16-20hrs at 48°C.

The process of single base extension and staining was carried out by the liquid handling robot. The probes on the chip were extended by a single hapten-labelled dideoxynucleotide (ddNTP) base complementary to the hybridised DNA. ddATP and ddTTP bases were labelled with DNP (2,4- Dinitrophenol), whereas ddCTP and ddGTP were labelled with Biotin. The DNA samples were then stripped off the chip using formamide. The staining procedure involves signal amplification by multi-layer immunohistochemical staining. The haptens were detected simultaneously by Streptavidin and an anti-DNP primary antibody conjugated to green and red fluorophores respectively (STM reagent, Illumina). They were then counterstained with biotinylated anti-steptavidin and a DNP-labelled secondary antibody to the anti-DNP primary antibody (ATM reagent, Illumina) to amplify the fluorescent signals. The last layer of stain was the STM, containing the fluorophores to allow signal detection. Finally, the stained chips were coated in nail varnish to protect the dyes, and scanned using the iScan scanner with autoloader (Illumina Inc, San Diego, USA).

2.4.6.2 Initial data analysis and quality control

The data were initially analysed using the Illumina Genomestudio software. This generates genotypes, and CN and loss of heterozygosity data (cnvPartition v3.1.6, Illumina). Quality control checks were performed to assess the data quality. Samples were assessed for their call rate, which should be >98% and >99% average across the batch. Specially designed control probes were also checked. Every array contained both sample dependent and sample independent control probes. Sample independent probes assess the quality of the processing, and sample dependent probes also assess the quality of the DNA. The B-allele frequency (BAF) plots and CN analysis results were checked to identify potentially contaminated samples. The BAF plot would show more than three modes if the sample had been contaminated. A noisy BAF plot may also suggest degradation of the DNA sample. CN data that looks to be duplicated for the whole genome also suggests contamination with at least one other DNA sample. The wet lab processing and initial data analysis was run under the supervision by Kerra Pearce at UCL Genomics, Institute of Child Health.

2.4.7 Copy number variant detection

Further analysis was carried out by Elliot Rees at Cardiff University, Institute of Psychological Medicine and Clinical Neurosciences. For all samples, Log R Ratios (LRR) and BAFs were generated using Illumina Genome Studio software (v2011.1) and used to call CNVs with PennCNV (Wang et al., 2007). CNV calling was performed following the standard protocol and adjusting for GC content. Samples were excluded if they were found to be an outlier for any one of the following QC metrics: LRR standard deviation, BAF drift, wave factor and total number of CNVs called per person. The LRR represents a measure of magnitude of combined fluorescence-intensity signals, and the

BAF denotes the relative ratio of fluorescence signals from one allelic probe compared with another. Duplications can be identified by an increase of LRR and the occurrence of four clusters in BAF. Consequently, a deletion is characterised by a decrease of LRR and lack of heterozygosity (at 0.5) in BAF. CNVs from samples that passed QC were joined together if the distance separating them was <50% of their combined length using an in-house developed open source program (http://x004.psycm.uwcm.ac.uk/~dobril/combine_CNVs/). CNVs were then excluded if they were covered by <3 probes. After CNV merging, the remaining CNVs were visually re-evaluated using the GenomeStudio genotyping module. All CNV coordinates are according to UCSC build 37/hg19.

I used cnvPartition as the secondary CNV detection algorithm using the following default parameters:

Confidence Threshold	|35
Detect extended homozygosity	|True
Exclude intensity only	|False
GC wave adjust	|False
Include sex chromosomes	|True
Minimum homozygous region size	|1000000
Minimum probe count	|3

The CNVs detected by PennCNV and cnvPartition were detected on autosomes only and were based on at least three consecutive probes. Here, these CNVs are referred to as non-mosaic somatic CNVs; that is, acquired somatic CNVs that are present in a sufficiently high proportion of cells to be detected by the applied algorithms.

The evaluation of the non-mosaic structural variants was based on predefined and structured criteria and consisted of the steps shown below. The unique CNVs remaining after each step are listed in Table 5.2.

1. After CNV detection by PennCNV, CNVs of the same type (CN = 0, 1, 2, 3 or 4) were merged if they overlapped with at least 50% of the length of the smaller CNV.

2. Samples from MZ twins, and parents where applicable, were compared. Only CNVs found to be discordant between paired samples, or concordant but overlapping known disease-susceptibility genes, were of main interest.

3. Discordant CNVs were evaluated by visual inspection of LRR and BAF plots in GenomeStudio. If both twins had the same signal intensities, the CNV was classified as concordant. If there was insufficient evidence, the CN call was disregarded.

4. CNVs were independently detected again by cnvPartition in GenomeStudio. CNVs of the same type (deletion or duplication) were merged if they overlapped with at least 50% of the length of the smaller CNV.

5. Only CNVs detected by both algorithms were retained for further analysis. CNVs called by PennCNV were disposed of if the CNV calls made by cnvPartition did not confirm the finding.

6. The LRR and BAF plots of remaining CNVs were visually inspected to select the best candidates for ddPCR validation.

2.4.7.1 Comparison of CNV/LOH differences between twins.

The output files from cnvPartition and PennCNV were converted into BED format files using a Perl script (see http://genome.ucsc.edu/FAQ/FAQformat.html#format1 for details). The files contain no headers but all take the same format: ‘chr|startPosition

(bp)|endPosition(bp)'. BED files are useful as they can be imported as a custom track into the UCSC Genome Browser, allowing one to see all the genes and other features that coincide with the features detailed in the BED file (in this instance, the features in the BED files are the CNV and LOH regions). They can also be viewed and manipulated in Notepad or Excel. The BED files for each pair of twins were compared using BEDTools v2.17, an open source suite of command line operated tools for comparison of BED files (http://bedtools.readthedocs.org/en/latest/content/overview.html). Three comparisons were carried out using the 'intersect' tool:

1. Find features present in twin1 but not in twin2 (filename: {twin1}_not_{twin2}.bed)
2. Find features present in twin2 but not in twin1 (filename: {twin2}_not_{twin1}.bed)
3. Find features that overlap between the two twins (filename: {twin1}_overlap_{twin2}.bed"). This file reports the original feature in twin1 (chr|startPosition(bp)|endPosition(bp)), then the original feature in twin2 (chr|startPosition(bp)|endPosition(bp)), and the final column contains the size of the overlap in bp.

The first two files contain regions of CNV and LOH, which are different between the two twins. The third 'overlap' file was run through a Perl script to remove the overlapping regions to leave just the regions specific to one twin or the other (filename: {twin1}_{twin2}_unique_regions_from_overlaps.bed). These three files, taken together, contained the CN and LOH differences between each pair of twins. CNVs were then annotated with gene information and compared to previously reported losses, gains, inversions or segmental duplications, and thus categorised as novel or benign polymorphism, using the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). Genes contained within CNVs were then searched on various databases, such as PubMed, OMIM and DisGeNET, to determine pathogenic relevance.

2.4.8 CNV analysis with ExomeDepth

ExomeDepth is an algorithm designed to use read depth data from exome sequencing analysis to call CNVs (Plagnol et al., 2012). It controls technical variability between samples, a feature which ordinarily complicates the analysis and creates spurious CNV calls.

The read count information was extracted from the individual BAM files using the R package Rsamtools. All reads were paired-end. Only reads with a Phred scaled mapping quality ≥20, distance of <1000bp from each other and in the correct orientation, were included. The location was defined by the middle location between the extreme ends of both paired reads. Exons closer than 50bp were merged into a single location owing to the inability to properly separate reads mapping to either of them.

Parameters for ExomeDepth were applied according to the instructions provided by the user guide. To set the threshold for sensitivity and specificity, the correlation between reference and tests count was set to 0.9898. It is advised that this correlation should be >0.97 to avoid a high false positive rate. There is an option, as used in the cancer field, of combining sequence data for healthy and tumour tissue. It is possible to utilise this function by pairing the affected and unaffected twin as 'tumour' and 'normal', respectively – thus replacing the test sample with the affected twin sample, and the reference sample with the unaffected twin sample. However, it would be statistically more viable to compare each sample independently against an aggregate reference of all exome samples. This would render shared CN calls between twin pairs, that are not present in other samples, as more likely to be real. This data was used to confirm CNVs detected by the SNP genotyping method. CN calls that were shared by all three calling algorithms (PennCNV, cnvPartition, and ExomeDepth) were considered high confidence CNVs.

2.5 Biochemical and label-free quantitative proteomics analysis

2.5.1 Cytogenetics

Before carrying out biochemical and proteomic analysis on the MZ twins discordant for ischaemic stroke (HG and KG), cytogenetic analysis of peripheral lymphocytes was performed according to standard protocols to rule out karyotypic abnormalities. Chromosomes from cultured peripheral blood lymphocytes were analysed on G-bands by trypsin using Giemsa (GTG)-banded metaphase spreads. A total of 20 metaphases were analysed for each twin, including karyotyping of 10 metaphases using light microscopy. The banding resolution was 450 bands for each sample. The constitutional karyotypes were described in accordance with the International System for Human Cytogenetic Nomenclature (Simons et al., 2013).

2.5.2 Blood chemistry

Blood samples were drawn from the twins by standard venesection into plain tubes and fluoride and EDTA anticoagulated tubes, just prior to lunch, and processed in parallel. A total of 59 individual markers were measured to check for renal, liver, glucose, cardiovascular, blood, immunological, and hormonal function. The data were compared with the corresponding reference intervals for the age group.

2.5.3 Blood collection for proteomic analysis

From each plasma sample, 2ml blood was allowed to clot at 4°C for at least 2hrs and then centrifuged at 1500g for 10mins to reach sediment the clotted cells. Plasma was then collected, divided into aliquots, and stored frozen at −80°C until the analysis was carried out. The control sample was created by taking 5µl of sample from each twin's serum after processing to generate peptides and pooling these together. As the control serum contains

all the features of both twin samples, it was used to align the data effectively – that is, to correct for any drift in the retention time of the peptides analysed.

2.5.4 Gel electrophoresis and in-gel digestion

1µl of a 1:4 dilution of each serum sample was run on SDS-PAGE using a 4-12% bis-tris gel and MOPS buffer system. Briefly 10ul of sample buffer and 4µl of reducing agent were added to each sample prior to addition of 25µl of milliQ water. The gel lanes from each serum sample were processed as follows. Samples were processed by in-gel reduction with 10mM dithiothreitol and alkylation with 50mM iodoacetamide prior to overnight (16hrs) trypsin digestion (Modified Sequencing Grade, Roche) at 30^{o}C. Peptides were extracted from the gel by 1% formic acid, then 100% acetonitrile, and dried in a SpeedVac concentrator (Thermo Scientific). The resulting peptides were then resuspended in 50µl 1% formic acid, centrifuged and transferred to the high-performance liquid chromatography (HPLC) vial. 7.5µl of each sample and pooled was used for analysis.

2.5.5 Sample separation

Technical replicates (3 x 7.5µl) of trypsin-digested peptides were separated using an UltiMate 3000 RSLCnano UHPLC System (Thermo Scientific) coupled to a LTQ Orbitrap Velos Pro (Thermo Scientific). The nLC buffers used were buffer A (0.1% formic acid), and buffer B (80% acetonitrile in 0.08% formic acid). Technical replicates (3 x 2.5µg) of each sample were initially trapped on an Acclaim PepMap100 C18, 5µM (100µM x 2cm nanoViper). After trap enrichment, peptides were separated on an Easy-Spray PepMap RSLC C18, 2µM column (75µM x 50cm nanoViper) (Thermo Scientific) with a linear gradient of 2-40% buffer B over 120mins with a constant flow of 0.3µL/min.

The UHPLC system was coupled to a LTQ Orbitrap Velos Pro (Thermo Scientific), via a nano electrospray ion source (Proxeon Biosystems). A Top 15 method was used to acquire Data Dependant Acquisition data. Briefly, a 60,000-resolution full-scan mass spectrometry (MS) survey spectra (m/z 335-1800) were attained in the Orbitrap with an automatic gain control target of 1,000,000 ions. Lock mass was set at 445.120024 and the spray voltage set to 1.8kV. This was followed by ion-trap MS/MS scans for the 15 most intense peptide ions.

Maximal filling times were 500ms for the full scans and 100ms for the MS/MS scans. Precursor ion charge state screening was enabled and all unassigned charge states, as well as singly charged species, were rejected. The dynamic exclusion was set to an exclusion list size of 500 with a maximum retention period of 45secs and an exclusion mass width of ±10 ppm. The lock mass option was enabled for survey scans to improve mass accuracy (Olsen et al, 2005). Data were acquired using the XCalibur software.

2.5.6 Abundance quantification

Orbitrap Velos Pro .RAW data files were imported into Progenesis LC-MS (version 4.0) for label-free differential analysis and subsequent identification and quantification of relative ion abundance ratios. Following alignment of MS data, principled component analysis and preliminary filtering (power <80% and P>0.05, corresponding to an ion score threshold of 35), datasets were exported from Progenesis as an .mgf file. These files were subsequently used to map the individual peptide sequences to corresponding UniProt identifiers using the UniProt database via Mascot Daemon (version 2.4.1). Enzyme specificity was set to trypsin, with fixed (quantitative) modifications set to carbamidomethyl cysteine (as it is assumed that all cysteines are modified so this does not alter the number of potential peptides). To increase the number of potential peptides,

parameters for the variable (non-quantitative) modifications were set to methionine oxidation, methionine dioxidation, protein N-acetylation, and Gln -> pyro-Glu. Other parameters used were as follows: Peptide mass tolerance, ±10ppm; fragment mass tolerance, ±0.6 Da; minimum peptide length, 6; and maximum missed cleavages, 2. Statistical P-values as shown in Tables 6.3, 6.4 and Appendix B were automatically generated using Progenesis software through a one way Anova on the ArcSinh transform of the normalised data.

2.5.7 Peptide/protein identification

The filtered Mascot search results were reimported into the Progenesis LC-MS and conflicts for peptide assignments at protein level were examined and resolved appropriately.

2.5.8 Pathway and network analysis

To explore the biological processes associated with the differentially expressed proteins, I performed Gene Ontology (GO) pathway analysis using the functional analysis tools VLAD (Richardson and Bult, 2015), PANTHER (Mi, Muruganujan and Thomas, 2012), and g:Profiler, gOST (Reimand et al., 2007). For the main analysis, g:Profiler was used to look for overrepresentation of GO terms in the differentially expressed protein lists, relative to the whole human proteome. The Ensembl 87 GO annotation dataset was used with the associated ontology file, released on 13 December 2016.

Only proteins with an average of >1.5 fold higher level in both runs for each twin were used to seed each network, using Cytoscape (Smoot et al., 2010) version 3.3.0. The following parameters were applied: Hypergeometric test, Benjamini & Hochberg False Discovery Rate correction, significance level 0.05 and the human proteome as the

reference set. The network view was modified using the following options: NetworkAnalyzer: ‘map nodes size’, degree, low values to small sizes; edit, ‘remove duplicate edges’ and selectively removed all edges describing identical protein binding interactions; layout, ‘Allegro Fruchterman-Reingold’, all nodes. The GO terms associated with this network were then identified using GOlorize (Garcia et al., 2006) with the BinGO plugin (Maere, Heymans and Kuiper, 2005) within the Cytoscape tool, and including the GO term ontology (10th February 2017) and the 13th March 2017 gene association files (a combination of two 166 gaf files goa_human, goa_human_isoform).

2.6 Validation techniques

To determine the genotype of the two of the most common SNPs known to be associated with lactase non-persistence, and confirm candidate mosaic discordant SNVs found between co-twins, PCR and Sanger sequencing was employed. Various SNP assays were employed for variants that had low-level mosaicism, potentially undetectable with conventional Sanger sequencing. CNVs found between co-twins were validated using ddPCR, and gene expression analysis was performed to confirm the up- and down-regulated proteins found in twins discordant for ischaemic stroke.

2.6.1 PCR and gel electrophoresis

Primers were designed using Primer 3 software (Koressaar and Remm, 2007). PCR amplifications of the genes of interest were performed in a Mastercycler pro Thermal Cycler (Eppendorf). PCR reactions were performed in a 10µL volume containing 1µL of genomic DNA (10–30ng/µL), 0.1µL of DreamTaq DNA Polymerase (Thermo Scientific), 1.0µL of 10X DreamTaq Green Buffer (Thermo Scientific), 0.4µL of a dNTP mix containing dATP, dTTP, dCTP, and dGTP at a concentration of 5mM (Thermo

Scientific), 0.4µL of each primer (Sigma-Aldrich) at a concentration of 5pmol, and distilled deionised water. The reaction conditions were as follows: 95°C for 3mins, followed by 40 cycles at 94°C for 30secs, 54-61°C for 30secs, 72°C for 30secs, and a final extension at 72°C for 7mins. PCR products were separated by electrophoresis in 1.3% agarose gels made with 1x Tris-acetate-EDTA buffer (Alpha laboratories) in order to confirm the amplification of the desired PCR product. These were then purified following the Exonuclease I (New England Biolabs, US) - Shrimp Alkaline Phosphatase (Affymetrix, US) (ExoSAP) protocol for sequencing reactions. The mixtures were incubated at 37°C for 5mins to digest the remaining primers and dNTPs into nucleosides and inorganic phosphate followed by deactivation of the ExoSAP in the Mastercycler pro Thermal Cycler at 95°C for 5mins.

2.6.2 Sanger sequencing

Sequencing was performed bidirectionally using the BigDye Terminator Chemistry (Sequenase v3.1 Cycle Sequencing Kit; Applied Biosystems, CA, USA). The sequencing reactions were performed in a 10µL volume containing 1.0µL of primer (5pmol), 2.0µL of the PCR product, 2.0µL of 5x BigDye Buffer, 0.5µL of BigDye Terminators and 4.5µL of distilled deionised water. The BigDye XTerminator Purification Kit (Applied Biosystems, CA, USA) was used to eradicate unincorporated dye terminators and free salts by adding 20µL of the SAM solution and 5µL of the XTerminator to each of the 10µL sequencing reactions. The mixture was placed on a shaker at full speed for 30min and then centrifuged at 13,000rpm for 2mins. A 3130xl automated DNA sequencer (Applied Biosystems) was used to run the samples. The sequencing data was analysed using the CodonCode Aligner v.4.2.4 (CodonCode Co., USA) and Finch TV (GeoSpiza, Seattle, WA). Wild-type sequence was derived from the Ensembl Genome Browser (Wellcome Trust Sanger Institute, Cambridge, UK).

2.6.3 SNP validation using CloneJET PCR Cloning Kit

PCR amplified fragments resolved on 1.3% agarose gel were extracted using the GeneJET Gel Extraction Kit (Fermentas, Thermo Fischer Scientific). The separated DNA fragment was visualised using a UV-transilluminator, and quickly excised to minimise DNA damage due to UV exposure. After weighing, a 1:1 ratio of Binding Buffer volume (500µL) to the gel slice weight (500mg) was added. The gel mixture was incubated at 55oC for 10mins, whilst intermittently inverting the tube. The solution was transferred to the GeneJET Purification Column and centrifuged at 13,000rpm for 1min. The flow-through was discarded. 700µL of Wash Buffer was added and centrifuged for 1min at 13,000prm. The flow-through was again discarded and the empty column was centrifuged for 1min. The column was placed into a fresh 1.5mL microcentrifuge tube. 50µL of Elution Buffer was added to the column and centrifuged for 1min at 13,000rpm. The flow-through containing the purified DNA was collected.

To confirm the potential variant on EML5 in subject RT1b, and its absence in co-twin RT1a, the PCR product spanning this SNP was cloned for each of the twins separately. Amplified fragments were ligated to PCR vectors with CloneJET™ PCR Cloning Kit (Thermo Scientific) according to the manufacturer's instructions. The protocol for cloning PCR products with 3'-dA overhangs, generated by DreamTaq DNA polymerase, was followed. This involved setting up a blunting reaction on ice, containing 10µL of 2x reaction buffer, 1µL of the purified PCR product, 1µL of the DNA blunting enzyme, and 6µL of nuclease-free water. The mixture was briefly vortexed and centrifuged, and then incubated at 70oC for 5mins to inactivate the enzyme before being placed back on the ice. The ligation reaction was set up on ice by adding the following to the blunting reaction mixture: 1µL of pJET1.2/blunt cloning vector (50ng/µL) and 1µL of T4 DNA

ligase, thus making a total volume of 20µL. The ligation mixture was incubated at room temperature for 5mins and used directly for transformation.

For bacterial transfection, 10µL of the mixture was mixed with 100µL of HB101 E. coli competent cells and incubated on ice for 45mins. The mixture was then heat-shocked at 42°C for 2mins, put on ice again for 5mins, 1mL of LB medium added, and incubated at 37°C for 45mins at 450rpm. The bacteria were spun down for 4mins and resuspended in SOC medium. The pellet was cultured overnight (18hrs) at 37°C on an LB agar plate containing 100µg/mL of Ampicillin. The following day, a total of 96 individual bacterial colonies were randomly picked per twin for colony screening by PCR using the plasmid-derived pJET1.2 forward and reverse primers (Thermo Scientific). The resulting PCR products were subsequently Sanger sequenced.

2.6.4 SNP validation using KASP assay:

To determine the potential mosaic variant found in subject RT1b, DNA samples of both MZ twins and variant flanking sequence were submitted to LGC Genomics for Kompetitive Allele Specific PCR (KASP) assay design and genotyping. This uses fluorescent resonance energy transfer to quench fluorescence in reporter oligonucleotides until they are incorporated into allele-specific PCR products. Concordance between the other genotyping methods would provide validation of the SNP.

2.6.5 SNP validation using Sequenom MassARRAY assay

In parallel, DNA samples were further genotyped for the potential variant by NewGene using the Sequenome MassAssay genotyping (Sequenom, San Diego, CA, USA), according to the manufacturer's instructions (Fumagalli et al., 2010). Briefly, 2µL of template genomic DNA was amplified by multiplex PCR to extend wild-type and mutant DNA. This was followed by treatment with shrimp-alkaline-phosphatase to inactivate

unincorporated nucleotides. A primer extension reaction (iPLEX Pro) was then performed using mass-modified terminators, and the products were spotted on a SpectroCHIP (Sequenom). MALDI-TOF mass-spectrometry was used for allele-specific detection. Automated genotyping calls were generated using the MassARRAY RTTM software and were validated by manual review of the raw mass spectra.

2.6.6 RNA extraction

Gene expression analysis is currently being carried out to confirm the up- and down-regulated proteins found in KG. Future projects beyond the scope of this thesis would include performing RNA sequencing (whole-transcriptome shotgun sequencing) on twin pairs to determine downstream consequences of potential (epi)genetic alterations. Thus, whole blood samples of co-twins were collected in PAXgene Blood RNA Tubes (Qiagen, UK) and the total RNA was isolated using the PAXgene Blood RNA Kit (Qiagen, UK) according to manufacturer's instructions. The PAXgene Blood RNA tubes were centrifuged for 10mins at 4000rpm. The supernatant was removed by pipetting. 4ml of RNase-free water was added to the pellet and vortexed until the pellet had thoroughly dissolved, and then centrifuged for 10mins at 4000rpm. The supernatant was discarded. 350µl of resuspension buffer (BR1) was added and vortexed until the pellet had thoroughly dissolved. The sample was pipetted into a 1.5ml microcentrifuge tube, and 300µl binding buffer (BR2) and 40µl proteinase K was added. The contents were mixed by vortexing for 3secs and incubated for 10mins at 55°C using a shaker–incubator at 1,000rpm. The lysate was pipetted into a PAXgene Shredder column placed in a 2ml processing tube, and centrifuged for 3min at 15,000rpm. The supernatant was transferred to a 1.5ml microcentrifuge tube without disturbing the pellet. 350µl of 98% ethanol was added to the tubes, mixed by vortexing, and centrifuged briefly for 3secs. 700µl of the sample was added to the PAXgene RNA spin column placed in a 2ml processing tube and

centrifuged for 1min at 15,000rpm. The PAXgene column was placed in a new 2ml processing tube, and the old processing tube containing flow-through was discarded. This step was repeated of the remaining sample. 350µl of wash buffer (BR3) was added to the PAXgene spin column and centrifuged at 15,000rpm for 1min.

10µl DNase I (RNFD) stock solution was added to 70µl digestion buffer (RDD) in a 1.5ml microcentrifuge tube. Contents were mixed by flicking the tube, and centrifuged briefly to bring any residual liquid to the bottom. 80µl DNase I incubation mix was added onto the PAXgene spin column membrane and incubated at room temperature for 15mins. 350µl of BR3 was added to the PAXgene spin column and centrifuged for 1min at 15,000rpm. The flow through was discarded and 500µl of wash buffer 2 (BR4) was added into the spin column and centrifuged at 15,000rpm for 1min. After discarding the flow through, 500µl of BR4 was again added to the column and centrifuged at 15,000rpm for 2mins. The spin column was transferred to a 1.5ml elution tube, 40µl elution buffer (BR5) was pipetted to the centre of the column and the tube was centrifuged at 15,000rpm for 1min. Another 40µl of BR5 was added to the column and centrifuged at 15,000rpm to elute RNA. The eluate was incubated at 65°C for 5mins in shaker-incubator and then chilled immediately on ice. The RNA samples were diluted with 10mM Tris-HCl pH 7.5 for accurate quantification by absorbance at 260nm. The quality of RNA was checked using an Agilent Bioanalyzer and the stock RNA was stored at -80°C.

2.6.7 Identifying *ARHGAP11B* copy number using droplet digital PCR

All DNA samples from twin pairs were sent to The Jackson Laboratory for Genomic Medicine for CNV validation. To determine *ARHGAP11B* copy number, ddPCR was used with a QX200 instrument (Bio-Rad, Hercules, CA, USA) following the manufacturer's instructions. The gDNA from all twin pairs was first digested by EcoRI

(New England Biolabs) enzyme for 1hr at 37ºC. Digested gDNA (4 ng) was assayed per 20µl reaction. Initially, three primers and probes were designed bioinformatically using Primer3 (http://frodo.wi.mit.edu/). Individual primer sets were assayed by PCR and gel-electrophoresis to test for primer-dimers and non-specific product amplification, and one set was chosen for the study. The CN assay primers and probes for ARHGAP11B were as follows: ARHGAP11BF: AGCTACAGGTATGGAGACAG, ARHGAP11BR: TTAACGTAATTCACCTGCCC, ARHGAP11B MGB probe: FAM-AGAGAAGCTGATCATGTTTCAGCA) at a final concentration of 900 nM primers and 250 nM probe. Amplicon coordinates were chr15:30676855-30676927 (hg38). After droplet generation, the reaction mixes proceed to thermal cycling as 95 °C × 10 min (1 cycle), 94 °C × 30 s, and 60 °C × 60 s (40 cycles), 98 °C × 10 min (1 cycle), and 12 °C hold. After thermal cycling, plates were transferred to a droplet reader (Bio-Rad) that flows droplets single-file past a two-colour fluorescence detector. Differentiation between droplets that contain target and those that did not was achieved by applying a global fluorescence amplitude threshold in QuantaSoft (Bio-Rad). Confirmed CNV duplications had approximately 50% increase in the ratio of positive to negative droplets as did the reference channel. Conversely confirmed CNV deletions had approximately half the ratio of positive to negative droplets as did the reference channel.

2.7 Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org

ALSGene, http://www.alsgene.org/_green

ALSoD, http://alsod.iop.kcl.ac.uk/_green

com/public-data/69-Genomes

CompleteGenomics cg69database: http://www.completegenomics.

dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/

Decipher, http://decipher.sanger.ac.uk/

ExAC Browser, http://exac.broadinstitute.org/

Genome Analysis Toolkit (GATK), http://www.broadinstitute.org/gatk/

GenomeComb, http://genomecomb.sourceforge.net

G-profiler: http://biit.cs.ut.ee/gprofiler/index.cgi

KEGG, http://www.genome.jp/kegg/

MutationTaster, http://www.mutationtaster.org/

NHLBI Exome Sequencing Project (ESP) Exome Variant Server,

NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

PDGene, http://www.pdgene.org

PolyPhen-2, http://www.genetics.bwh.harvard.edu/pph2/

RefSeq, http://www.ncbi.nlm.niv.gov/RefSeq

SAMtools, http://samtools.sourceforge.net

SIFT, http://sift.bii.a-star.edu.sg/

SZDB, http://www.szdb.org

SZGene, http://www.szgene.org

Chapter 3. Investigating the genetics of amyotrophic lateral sclerosis

3.1 Overview

Amyotrophic lateral sclerosis (ALS) is a devastating neurogenerative disease of upper and lower motor neurons with no effective treatment (Kiernan et al., 2011), or definitive diagnostic test (Cirulli et al., 2015) – although multiple genes have been implicated in the disorder, including: *ALS2*, *ANG*, *DCTN1*, *FIG4*, *FUS*, *NEFH*, *SLC52A1*, *SLC52A2*, *SLC52A3*, *SOD1*, *TARDBP*, *UBQLN2*, *VAPB* and *VCP* (Meltz Steinberg et al., 2015; Morgan et al., 2015).

Familial ALS accounts for approximately 10% of cases, and is inherited in an autosomal dominant, autosomal recessive, or X-linked fashion. The remaining 90% of cases are believed to be sporadic (Renton et al., 2013; Cirulli et al., 2015). Of familial ALS, approximately 33% worldwide is due to hexanucleotide (GGGGCC) repeat expansions in the intron between non-coding exons 1a and 1b of *C9orf72*. These are also present in 8% of sporadic ALS cases, highlighting a major role for C9orf72 in neurodegeneration (Farg et al., 2014). It also accounts for approximately 25% of frontotemporal dementia (FTD), a disorder that has clinical, pathophysiological and genetic overlaps with ALS.

Using twin data to estimate the heritability of ALS, Al-Chalabi et al. (2010) found that 90% of the MZ twins investigated were discordant for the disorder. This high ALS discordant rate is further corroborated by studies where either eighteen of twenty-one (Graham et al., 1997), or all twelve (Dellefave et al., 2003) MZ twin pairs investigated were clinically discordant for ALS. A low disease concordance rate may point to

environmental influences as a possible explanation. Another possibility, however, is the occurrence of de novo mutations early in development of the affected twin.

In this study, genetic and clinical differences were compared between four ALS-discordant twin pairs. Three twin pairs were obtained from The Coriell Institute (catalogue IDs: ND08242 and ND08242; ND12421 and ND12422; ND12218 and ND14318). One twin pair (LAS and SUS) was referred to this study by Niranjanan Nirmalananthan (St George's University Hospitals NHS Foundation). Genetic screening of the hexanucleotide repeat expansion in *C9orf72* involved rpPCR and Southern blotting. The Coriell DNA samples were derived from immortalised lymphoblastoid cell lines (LCLs); DNA samples of LAS and SUS were obtained from buccal swab samples. To determine the sensitivity and specificity of repeat length calculation using LCL-derived DNA we additionally compared *C9orf72* repeat lengths in DNA derived from whole blood and DNA derived from LCLs. Samples from LAS and SUS were screened on a 25-ALS gene panel. Demographic and clinical information for each subject is summarised in Table 2.1.

3.2 Results

3.2.1 Repeat-primed PCR

C9orf72 hexanucleotide repeats were screened in four ALS-discordant twin pairs using rpPCR. One MZ twin pair (421 and 422) had abnormally-enlarged (>30) *C9orf72* repeat expansions, although they were reported to be discordant for the disease. A twin pair of 35-year-old males of European descent (242 and 243), had normal and equal numbers of *C9orf72* repeats (n=2). One twin pair (218 and 318) had normal, but different, numbers of *C9orf72* repeats (n=2 and n=8, respectively). An expansion size >30 is beyond the

resolution of rpPCR, thus southern blotting was used to confirm and quantify the expansion.

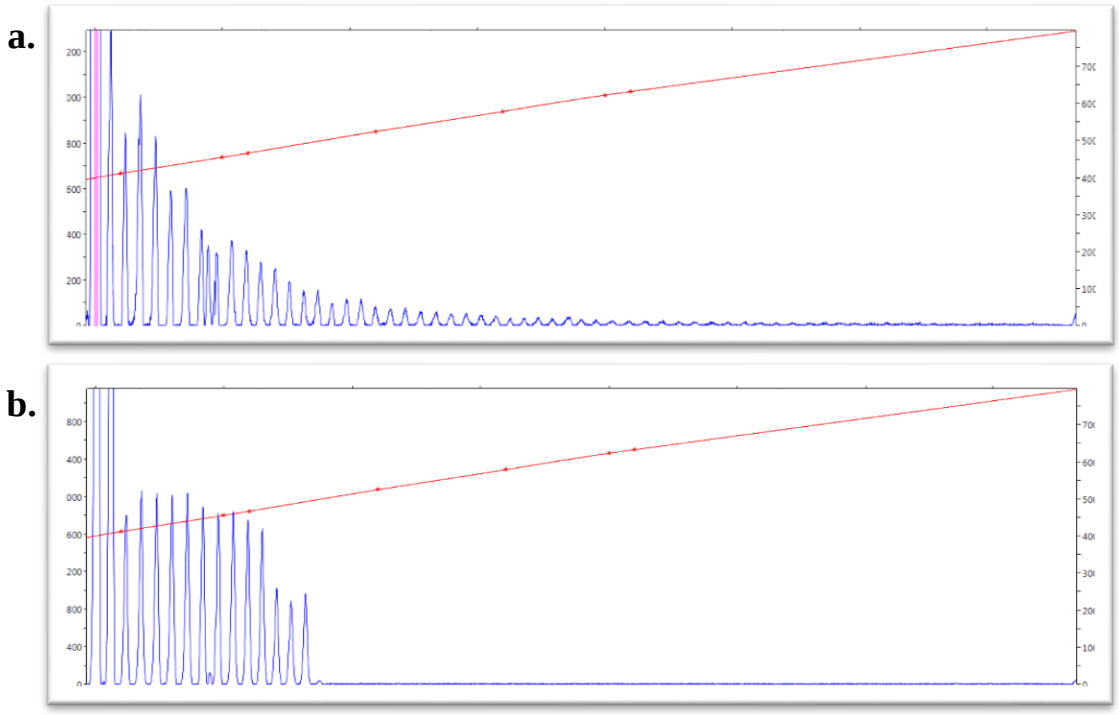


Figure 3.1. *C9orf72* repeat genotyping. a) The presence of a homozygous *C9orf72* hexanucleotide expansion in subject 421 is shown. The rpPCR result demonstrates a saw-tooth pattern, typical of a pathological expansion. Expansions are measurable up to 40-60 hexanucleotide repeats. b) A wild-type control result is shown for comparison. Fluorescence intensity is recorded on the vertical axis. DNA fragment peaks are sized based on the sizing curve produced from the points on the internal size standard, which is shown as consecutive red dots at 300, 340, 350, 400, 450, 490 and 500bp.

3.2.2 Southern blotting

Southern blotting was performed on Coriell DNA samples of three MZ twin pairs (242 and 243; 421 and 422; 218 and 318), using a modified protocol as previously described (DeJesus-Hernandez et al., 2011). Somatically unstable expanded repeats were confirmed in subjects 421 and 422 (Figure 3.2).

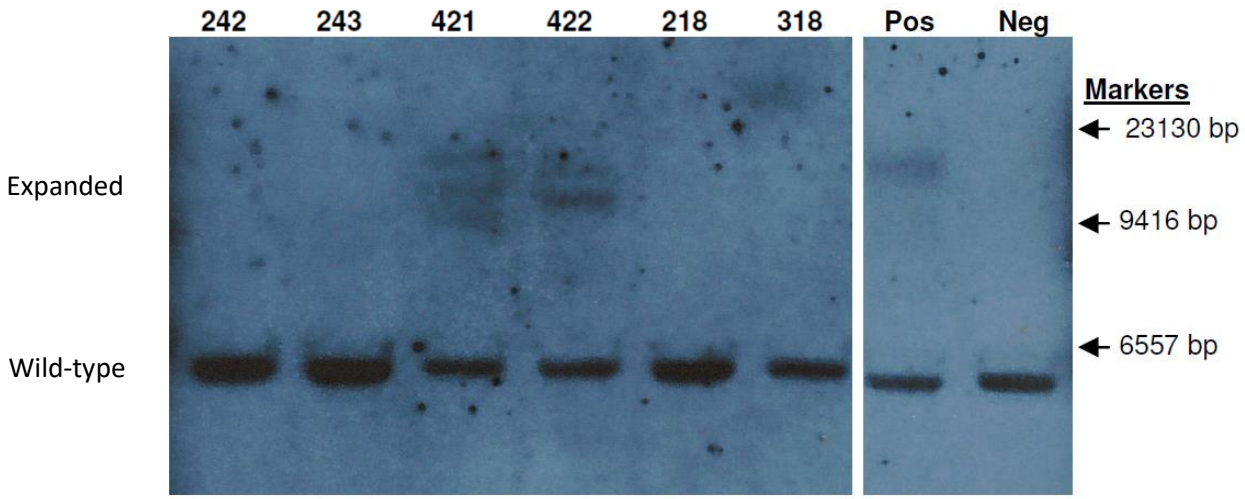


Figure 3.2. Southern blotting showing data from three ALS-discordant twin pairs. Positive for expansion was seen in twins 421 and 422. Typical LCL banding patterns can be seen and might represent pauciclonality of LCL DNA. Repeats associated with LCL DNA are smaller in size than repeats seen in peripheral-blood DNA. The positive control DNA source is from blood.

Total gDNA was digested with a restriction enzyme and hybridised with a single copy probe. A large band representing the C9orf72 repeat expansion was detected in samples 421 and 422. A 6.2 kb wild-type band was seen in all other samples. The pattern of the expansion in these samples is as expected for LCL-derived DNA. The multiple bands seen are likely to be smaller than when compared to the actual average size in blood. Somatic instability of the *C9orf72* repeat is known to occur in LCL-derived DNA compared with DNA derived from whole blood from the same patient. Accurate repeat length determination in immortalised cell lines is therefore difficult to ascertain (Hübers, et al., 2014). Southern blotting was not performed on samples LAS and SUS due to limited DNA sample availability, however rpPCR did not reveal expanded alleles.

Sample IDs	C9orf72 repeat no.	Southern blot
242	2	Wild type
243	2	Wild type
421	>30	Expansion
422	>30	Expansion
218	2	Wild type
318	8	Wild type
LAS	10	n/a
NAS	10	n/a

Table 3.1. Repeat length determination using rpPCR and Southern blotting.

3.2.3 NGS multigene panel

One twin pair (SUS and LAS) were further screened on a comprehensive assay containing 25 genes that have previously been implicated in ALS. Sequence analysis involved full exomes of known ALS-candidate genes (*BSCL2, CEP112, FUS, MATR3, OPTN, SOD1, SPG11, TARDBP, UBQLN2, and VCP*), and other genomic areas of potential importance (*ALS2, ANG, CHMP2B, DAO, DCTN1, FIG4, NEFH, PON1, PON2, PON3, PRPH, SETX, SQSTM1, VAPB, and VEGFA*) (Morgan et al., 2015). A possible mutation in *SOD1* (c.65A>G; p.E22G) was identified in both the twins; however, when viewed in IGV the region showed evidence of sequencing artefacts.

3.3 Discussion

The rpPCR method used in this study to screen for expanded *C9orf72* repeats can only accurately be determined for up to 60 repeats (Renton et al., 2011). Repeat expansions of >30 are considered pathological, and <20 repeats are considered normal. This range is based on a study of the repeat expansion in a Finnish cohort of 402 ALS cases and 478 controls. Of the ALS patients, rpPCR showed the repeat expansion to be in 28.1% cases with an average number of repeats of 53 (range 30 to 71). The control group had an average of 2 (range 0 to 22). Further, it was determined that the expansion accounts for 46.0% of familial ALS, and 21.1% of sporadic ALS in that Finnish population.

A definitive demarcation between normal and pathological repeat sizes is not known, however this study used the usual cut-off points described in previous studies in *C9orf72*-associated disorders (Majounie et al., 2012). All patients with a normal repeat range in our cohort had <20 units. As no subjects presented with the border-zone range of 20 to 30 units, there was no ambiguity between our normal and abnormal classification.

Three twin pairs had normal numbers of *C9orf72* repeats; of these, one pair (218 and 318) had different numbers of *C9orf72* repeats (n=2 and n=8). Although this would not cause phenotypic discordance, Dols-Icardo et al. (2014) found subtle differences in the repeat length between two MZ twins discordant for the disease. This suggests that stochastic expansion events do occur during cell division in this region, leading to somatic mosaicism, thereby contributing to the repeat expansion variation within and between the twins.

Abnormal *C9orf72* repeat expansions (>30) were found in one twin pair (421 and 422), yet they were reported to be discordant for the disorder. This finding was confirmed with southern blotting. Unfortunately, it has not been possible to obtain follow-up clinical

details for these twins, to find out if the initially unaffected twin developed ALS over time. This is nevertheless an interesting finding, as only one other ALS-discordant twin pair case is reported, where both twins have abnormally expanded *C9orf72* repeats (Xi et al., 2014). DNA is not available from other family members to confirm whether or not this is sporadic or familial *C9orf72* repeat-related ALS; however, clinical records have indicated an absence of ALS and other related neurodegenerative disorders in the family.

The affected twin (LAS) in this MZ twin cohort died at the age of 71, shortly after agreeing to partake in this study. The clinical phenotype in LAS was characterised by predominant lower motor neuron signs, with clinical onset at 67 years of age. Initial symptoms involved the distal muscles, predominantly of the lower limbs, with late bulbar and respiratory involvement. The ALS clinical course in the affected twin demonstrated slow progression with a long survival time. Interestingly, the unaffected twin (SUS) was asymptomatic two years after her twin was diagnosed. Thus, it remains possible that there has been a yet undiscovered de novo mutation event unique to the affected twin, or that both twins share a pathogenic mutation beyond the regions examined. If the latter, a possible explanation for the different intrafamilial phenotypic expression could involve putative modifier genes, epigenetic differences and/or environmental factors.

A limitation of the study is that the DNA samples obtained from Coriell were LCL derived. This poses a disadvantage for whole-exome sequencing analysis, as it has been shown that the process of cell line creation and culturing itself induces de novo mutations (Veltman and Brunner, 2012). Analysis of post-mortem brain biopsies would be more suitable to investigate neurodegenerative diseases. For LAS and SUS, DNA from mucosal epithelial cells was obtained, as an ectodermally-derived tissue would be preferable over peripheral blood when looking for CNS somatic variations.

In conclusion, a rare and interesting scenario has been identified, where one ALS-discordant MZ twin pair were concordant for the *C9orf72* repeat expansion. The phenotypic discordance observed between the twins in this cohort raises the possibility of environmental factors having an influence in disease manifestation. An environmental contribution in ALS has been determined to be in the order of 40% in previous twin studies (Al-Chalabi et al., 2010), and disease expression in patients with mutations in *SOD1* seem to be largely determined by the environment (Fogh et al., 2007). The phenotypic disparity between the twin pairs with similar *C9orf72* repeat sizes could also signify hitherto unexplored modifying genes, epigenetic mechanisms, or environmental causes.

Chapter 4. Whole-exome sequencing analysis

4.1 Overview

Over the past decade, investigations of de novo mutations have called into question the assumption that MZ twins are genetically identical. Indeed, it has been shown that the underlying genetic differences between co-twins may arise during embryonic development, giving rise to SNVs, indels, gene conversion, CNVs and postzygotic mitotic recombination. These variations have been proposed as potential genetic mechanisms resulting in discordant MZ twins (Ketelaar, Hofstra and Hayden, 2011).

Comparing MZ twins discordant for complex traits has been previously suggested for discovering disease-relevant variants in candidate genes (Mansilla et al., 2005). Chapter 1 outlines numerous studies that have discovered genetic differences between MZ twins that account for their discordant phenotypes. Such mechanisms include chromosomal mosaicism, as seen in trisomy 21 and 45,X0 (Nieuwint et al., 1999), and dominant gene mutations, which was first shown for Van der Woude and popliteal pterygium syndromes (Kondo et al., 2002). These post-twinning mutations would result in somatic mosaicism, a phenomenon defined as two or more genetically distinct populations of cells in an individual that were developed from a single fertilised egg (Freed, Stevens and Pevsner, 2014).

Comparing the genomes of discordant MZ twins signifies a promising opportunity for the search of novel candidate variants implicated in disease, which may ultimately narrow the conceptual gap of missing heritability. With the cost of current NGS sequencing methods dramatically lowering, whole genome- and exome-wide comparisons of MZ twins can be made more affordable and readily available. This circumvents the need to focus merely on a set of candidate genes, as previously suggested.

Whole-exome sequencing has proved to be a fruitful tool for identifying genetic causes of human diseases. This technique involves sequencing the annotated protein coding exons of the genome with nearby flanking intronic regions (Magne et al., 2014; Lin et al., 2012). Most known mutations causing high-penetrance genetic disorders are found in or adjacent to the coding regions, thereby justifying the use of whole-exome sequencing to explore potential genetic causes of complex disorders.

Considering the estimated somatic mutation rate is extremely low and that these variants can be obfuscated by the relatively high error rate of NGS (Aparicio and Huntsman, 2009; Kuhlenbäumer, Hullmann and Appenzeller, 2011), a highly sensitive filtering method with high specificity, sequence resolution and coverage should be implemented (Aparicio and Huntsman, 2009).

We performed whole-exome sequencing analysis of DNA from thirteen twin pairs discordant for a range complex disorders. We also performed whole-exome sequencing on DNA obtained from blood and saliva samples for a MZ twin pair discordant for stroke, and parent-offspring trio analysis using DNA obtained from the parents of the twins discordant for ADHD and Tourette's syndrome.

4.2 Results

4.2.1 Quality control and pre-analysis

Whole gDNA obtained was evaluated for quality and quantity by densitometry analysis using 1.2% agarose gel electrophoresis, as shown in Figure 4.1, and spectrophotometric measurement. Agarose gel electrophoresis showed that the molecular weight extracted was more than 10kb with uniform brightness and all samples had a 260/280 ratio of ~1.8, suggesting that the DNA was integrated, stable and that the extraction was successful.

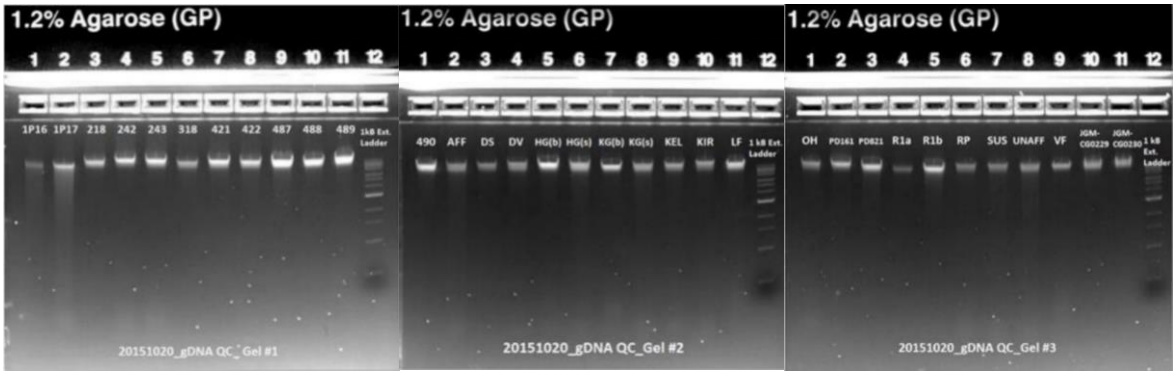


Figure 4.1. Agarose gel electrophoresis of gDNA. DNA samples had a tight band with minimal smearing, therefore passing quality control.

4.2.1.1 Hereditary spastic paraplegia

HSP-discordant twins (VF and LF) were recruited into the study through collaboration with Niranjanan Nirmalananthan (St George's University Hospitals NHS Foundation). VF was diagnosed with a complex form of HSP and was therefore assigned to an NGS panel of 41 HSP-linked genes. DNA sequencing was performed on the following genes: *AFG3L2, ALS2, AP5Z1, ATL1, B4GALNT1, BSCL2, C12orf65, CYP27A1, CYP2U1, CYP7B1, DDHD1, DDHD2, FA2H* (excluding exon 1), *FIG4, GBA2, GCH1, HSPD1, KIAA0196, KIF1A, KIFSA, L1CAM, MTPAP, NIPA1, PLP1, PSEN1, REEP1, RTN2,*

SACS, SIGMAR1, SLC16A2, SLC2A1, SPAST, SPG11, SPG20, SPG21, SPG7, VAMP1, VPS37A, WDR45, ZFYVE26, and *ZFYVE27.*

VF was also screened for GLUT1 deficiency syndrome and DYT1 early-onset primary dystonia. However, no clearly pathogenic mutations were detected in VF. This result reduces the likelihood that the symptoms seen in this individual are caused by pathogenic mutations in these genes, suggesting that other rare mutational mechanisms not detectable by this analysis may be present. The difficulty in determining the aetiological basis of complex movement disorders means that a diagnosis of HSP must currently rest on clinical grounds alone.

4.2.1.2 Lactase persistence

The *LCT* gene is 49.3 kb in length and located on the chromosome 2q21. It contains 17 exons and is translated into a 6 kb transcript (NCBI Reference Sequence NG_008104.1). There are at present five different functional alleles that have been associated with lactase persistence: -14010G>C (rs145946881); -13915T>G (rs41380347); -13910C>T (rs4988235); -13907C>G (rs41525747); and -22018G>A (rs182549). To determine the genotype of these alleles in the twins discordant for lactase persistence (KEL and KIR), approximately 300bp surrounding -13910C>T and ~200bp surrounding -22018G>A was analysed by Sanger sequencing. Both twins had an identical genotype for these SNPs, and no mosaicism could be seen in the sequence traces. The potential lactase non-persistence genotypes that have been reported in people of Northern European ancestry (Tishkoff et al., 2006; Ingram et al., 2008) were absent in the twins. The SNPs associated with lactase non-persistence in African populations and other genetically diverse groups were also absent (Friedrich et al., 2013; Ranciaro et al., 2014; Jones et al., 2013). These results suggest that both twins are genetically lactase-persistent. Known causes of secondary

lactase deficiency could be ruled out based on the clinical history of the twins, such as gastroenteritis, coeliac disease, Crohn's disease, ulcerative colitis, chemotherapy, or long courses of antibiotics. Thus, other yet unidentified SNPs associated with lactose intolerance could be present in the affected twin.

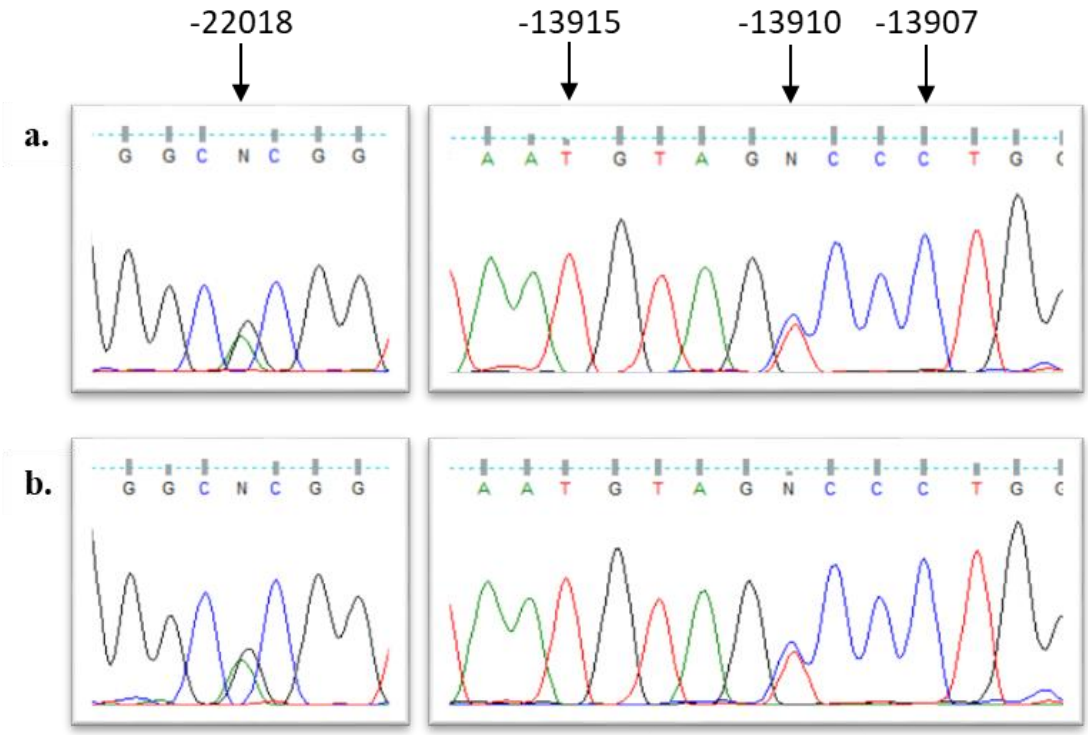


Figure 4.2. Sanger sequencing in individuals KIR (a) and KEL (b) reveals heterozygosity for variants -13910C>A and -22018G>A, and homozygosity for variants -13907C>G and -13915T>G. Variant -14010G>C was not checked using Sanger sequencing, however it was later confirmed to be homozygous for the G allele with exome sequencing and SNP array data.

4.2.2 Whole-exome sequencing analysis

Three different methodological approaches were applied to analyse the exome sequence data. We initially prioritised candidate discordant variants in seven MZ twin pairs (Pipeline 1). Three twin pairs were selected to explore an alternative method using GenomeComb (Pipeline 2). Finally, all samples were pooled together and re-analysed by combining two updated variant callers for increased specificity (Pipeline 3).

Samples	*Pipeline 1*	*Pipeline 2*	*Pipeline 3*
IP16 and IP17	✓	✓	✓
RT1a and RT1b	✓	✓	✓
PD821 and PD161	✓	✓	✓
HG(b) and KG(b)	✓	-	✓
HG(s) and KG(s))	✓	-	✓
VL and FL	✓	-	✓
KIR and KEL	✓	-	✓
AFF and UNAFF	✓	-	✓
LAS and SUS	-	-	✓
218 and 318	-	-	✓
421 and 422	-	-	✓
242 and 243	-	-	✓
RP and OH	-	-	✓
489 and 490	-	-	✓

Table 4.1. Comparing bioinformatics approaches for the analysis of MZ twins. The sequence read files were analysed using three pipelines: Pipeline 1 was developed by Michael Simpson, Pipeline 2 by Peter De Rijk, and Pipeline 3 by Alan Pittman.

4.2.2.1 Pipeline 1: VarScan2

On average, over 98% of the accessible regions were covered of the whole-exome reference. After short read alignment, variant calling and gene annotation, the exomes yielded a mean value of 25,082 total candidate variants, of which on average 171 were novel. Prior to filtering, pair-wise comparisons of exonic variants between co-twins contained on average 1,390 putative discordant variants unique to the affected or unaffected co-twin.

As the rate of somatic mutations has been estimated to be between 0.82-1.70x10-8 (Kong et al., 2012; Campbell et al., 2012; Dal et al., 2014), it was expected that a considerable proportion of the candidate discordant variants represented technical artefacts. This could be the result of errors occurring in the exome sequencing itself, or incorrect read alignments to the reference genome. The latter would be particularly common in regions

containing segmental duplications or repetitive elements, including genes with processed or unprocessed pseudogenes. Additional restrictive filters were applied to eliminate germline variants that were under-called in either co-twin due to low coverage at the end of reads, and false-positive variants due to various cryptic paralogous sequences. An in-house database with over 1,000 exome sequences was also used to further eliminate variants arising recurrently through misalignments, including ubiquitous variants that had perhaps resulted through recurring mutations at hypermutable sites – variants that would likely have little or no biological relevance. Further, only variants with a somatic allele frequency of >15% in one twin that had no representation in the other twin were retained. Application of this filtering criteria resulted in approximately 20-50 candidate discordant somatic variants between co-twins. Details of coverage statistics for the first seven MZ twin pairs are in Appendix C.

The short read genomic alignments for these potential variants were manually inspected using IGV, to remove additional artefacts that bypassed prior filtering parameters. These included variants located at the start and end position of reads, base quality scores that were less than 20 on average, homopolymer runs, and variants seen in genomic neighbourhoods with multiple nearby rare variants (suggestive of alignment artefacts caused by nearby indels). These variants were further inspected for intrinsic genome characteristics, such as segmental duplication, micro-satellites, and simple tandem repeats using online genome browsers Ensembl and UCSC.

After application of the initial filtering criteria and manual reviewing, each co-twin typically had no candidate somatic variants. However, for the MZ twins discordant for schizophrenia, a non-synonymous SNV on *EML5* (chr14:89151465; G>A substitution)

was found in subject RT1a with an allele frequency of 17.24%, which was absent in the co-twin (Figure 4.3).

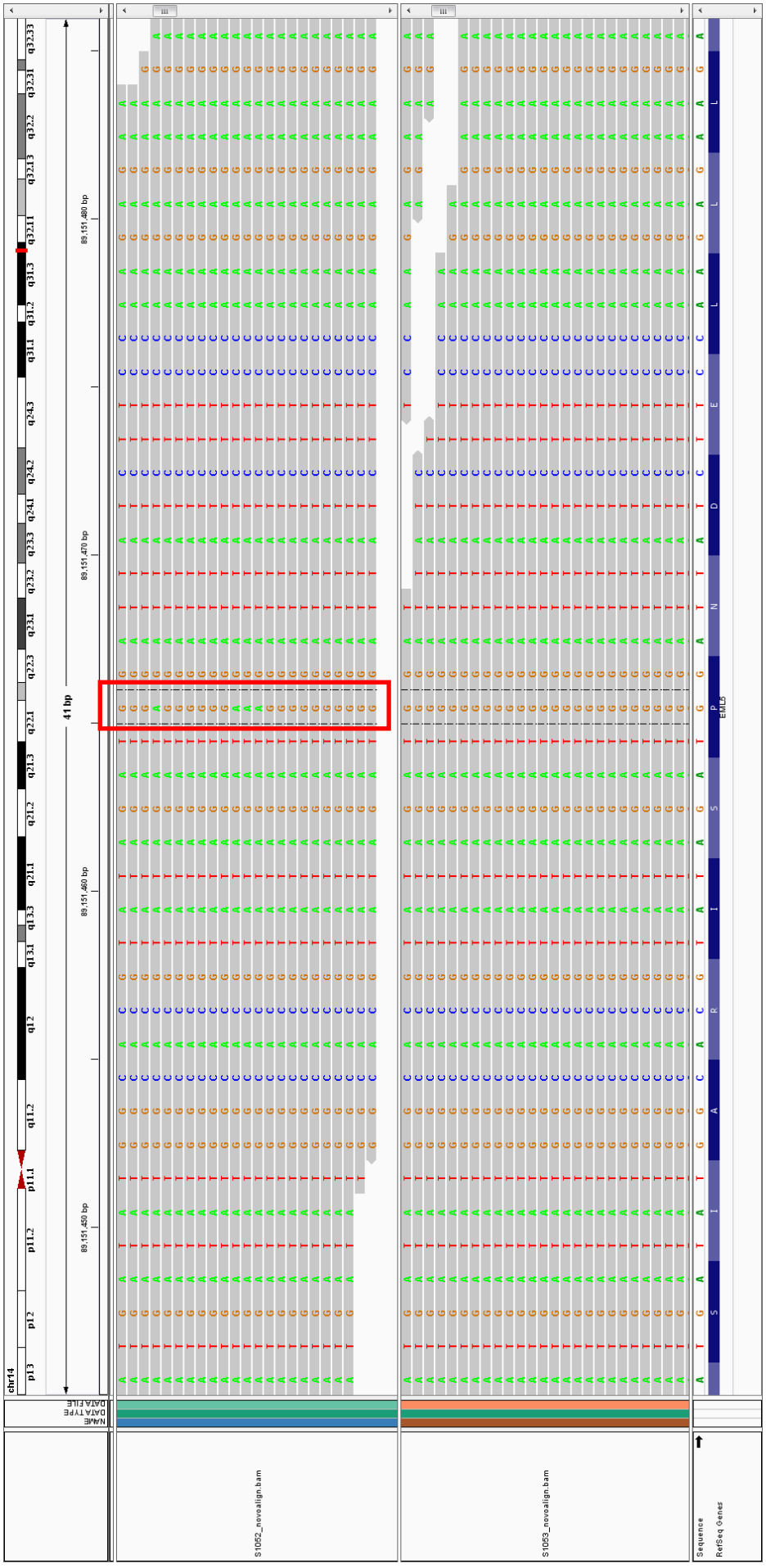


Figure 4. 3. Screenshot from IGV showing the potential mosaic variant on *EML5* in subject RT1b.

4.2.2.2 Pipeline 2: GenomeComb

Distinguishing true SNVs from errors in exome sequencing data remained challenging. Before analysing the rest of the dataset with Pipeline 1, more comprehensive bioinformatics approaches were explored to identify SNVs with high confidence. By testing a variety of different strategies to reduce errors, a refined method for the prioritisation of high confidence SNVs could be developed to analyse our twin cohort. At the time of this investigation, a study proposing an optimised filtering method that selectively reduces error rates in whole-genome sequencing was published (Reumers et al., 2012). The authors applied twelve individual filters in 1,048 combinations and determined the most effective in terms of having the highest specificity and sensitivity. This facilitated the identification of SNV differences in MZ twins discordant for schizophrenia (Reumers et al., 2012).

To test if this novel strategy could identify biologically relevant variants in whole-exome sequencing, a pairwise comparison between three MZ twin pairs from our cohort was made (IP16 and IP17; RT1a and RT1b; PD821 and PD161).

4.2.2.3 Assessment of accuracy filters

Pairwise comparisons between three twin pairs revealed that on average 98.1 (97.4-98.4) [median(0.25-0.75quartile)] of the reference genome was sequenced in co-twins, allowing, on average, the more than 96% concordance between co-twins.

As mentioned above, prior studies investigating genetic differences between twin pairs have illustrated that most discordant SNV calls are in fact false positives (Dal et al., 2014; Kong et al., 2012; Campbell et al., 2012). It is therefore assumed that most discordant SNVs in our twin exomes represent sequencing errors, and that shared SNVs reflect true

genetic variation. The filters were considered effective when at least twice as many discordant SNVs were removed compared to shared SNVs.

With GATK and SAMtools combined, an average of 127,325 variants were detected between the co-twins, of which 10,701 were discordant. Closer inspection of the sequence surrounding putative discordant SNVs revealed that many of them were due to an uncertain call in either twin or in regions where several other SNVs were clustered. In addition, many discordant SNVs were in genomic regions containing tandem repeats, segmental duplications or at positions with low (<15) or very high (>100) coverage. Thus, two types of accuracy filters were employed: filters that detect genomic regions with insufficient sequence quality, and filters that target genomic regions with repetitive DNA sequences.

Several parameters can be used to apply filters based on the quality of whole-exome sequencing data. These include the coverage depth of a given sequence, a quality score for every SNV that is detected (referred to as the 'variant score'), and the presence of clustered SNVs in the genomic region. For each of these parameters, thresholds at which a sequence can be considered high quality were defined. Standard receiver operating characteristic (ROC) curves and distribution analyses revealed that a coverage depth and variation score threshold of ≥15 and ≥70, respectively, were most optimal (Figures 4.4 and 4.5).

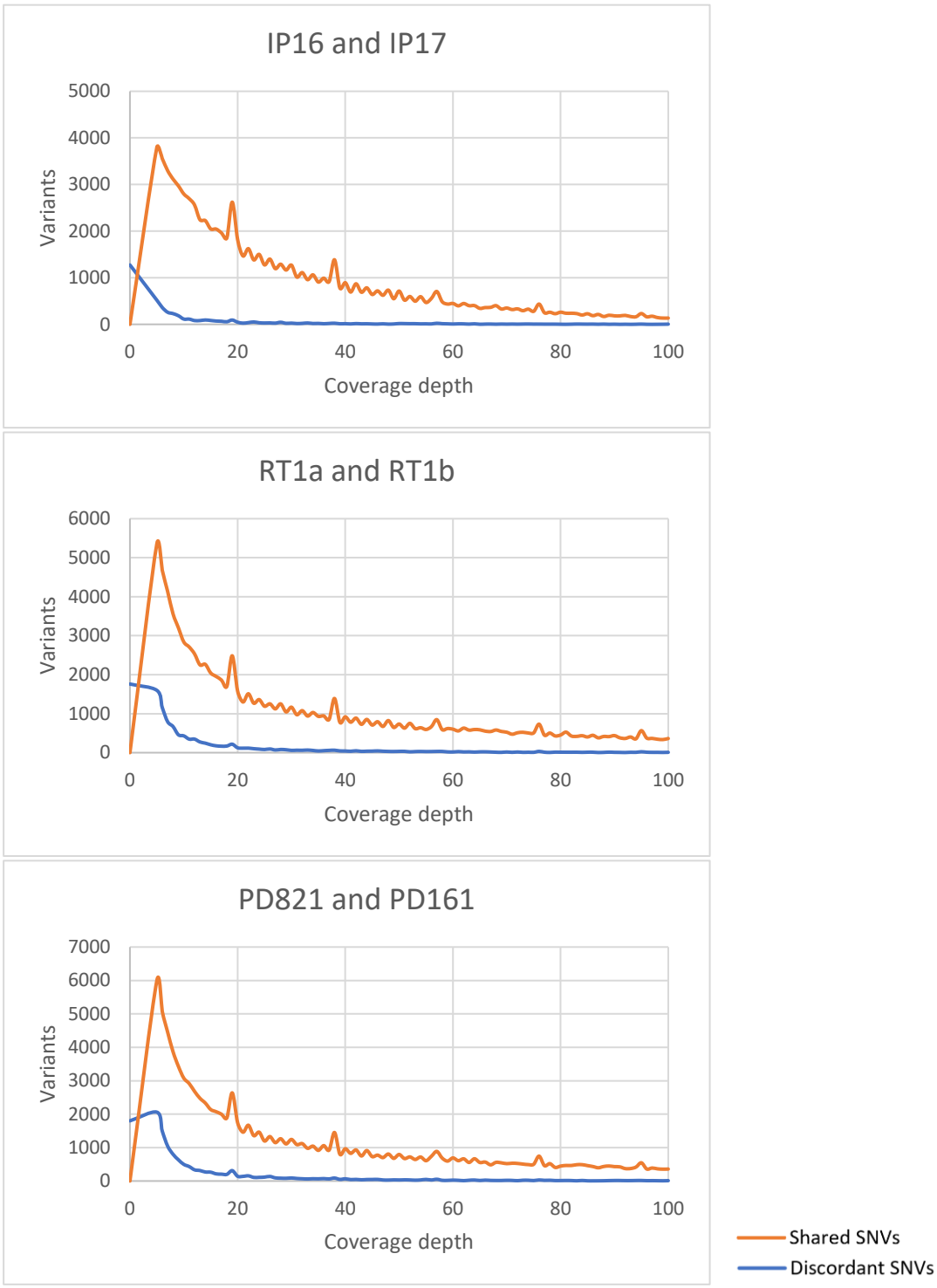


Figure 4.4. Coverage depth filter. The distribution graphs depict the frequency of shared and discordant SNVs against a coverage depth ranging from 1 to 100 in the three twin pairs analysed. A coverage depth threshold of ≥15 was used on all samples to remove sequencing artefacts from the data.

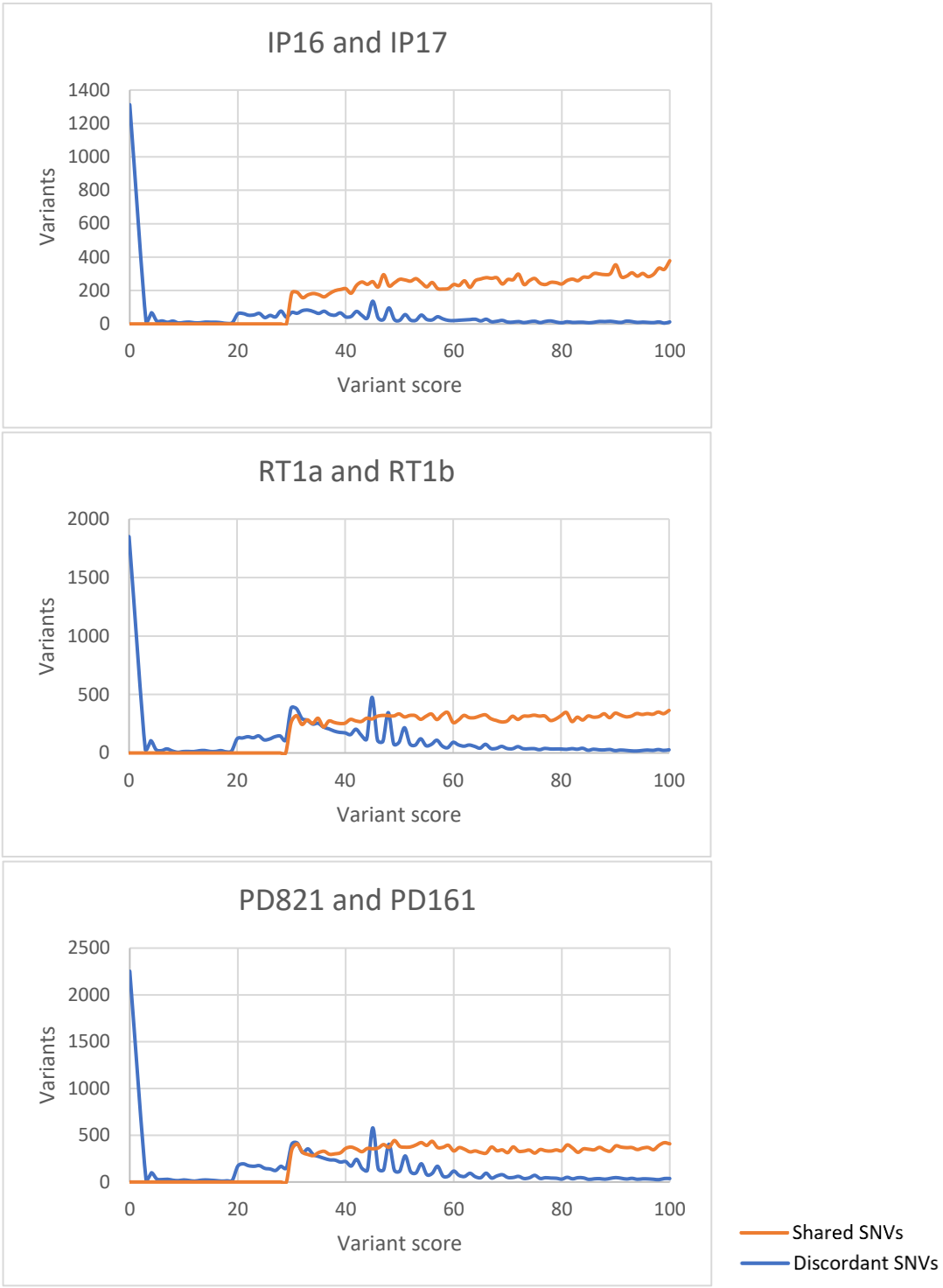


Figure 4.5. Variant score filter. The distribution graphs depict the frequency of shared and discordant SNVs against a variant score ranging from 1 to 100 in the three twin pairs. The data are difficult to interpret due to the irregular shape of the distributions. It is, nevertheless, clear that lower variant scores are characterised by a higher fraction of discordant SNVs. A variant score threshold of ≥70 was used on all samples to remove sequencing artefacts from the data.

The ratio of 'the proportion of discordant SNVs removed' to 'the proportion of shared SNVs removed' ($F_{discordant}/F_{shared}$) was calculated after application of individual filters. Each of the three individual quality filters was found to remove a $F_{discordant}/F_{shared}$ ratio of >2, suggesting that they were efficient in removing genotype calling errors. The 'variant score' and 'clustered SNV' filters discarded a relatively small mean percentage of the shared twin exomes, whereas the 'coverage depth' filter removed a more significant portion (6.6%, 13.5% and 28.3%, respectively). However, the combination of these three quality filters substantially lowered the average number of discordant SNVs between the three co-twins: 4,076 (38.09%) discordant SNVs and 97.816 (76.82%) shared SNVs remained, with a $F_{discordant}/F_{shared}$ ratio of 5.61.

Repetitive DNA sequences such as simple tandem repeats, microsatellites and segmental duplications often contain errors due to incorrect mapping to the reference genome. These regions, together with self-chained regions, interspersed repeats and low-complexity regions were identified using the UCSC genome browser database. Filters targeted at short repeats, such as tandem repeats and microsatellites were very selective ($F_{discordant}/F_{shared}$ >5) and removed only small portions of the exome for all three twin pairs. The segmental duplication filter was moderately successful, yielding a mean $F_{discordant}/F_{shared}$ of 2.77. In contrast, the RepeatMasker and self-chained filters were too restrictive while not being very selective ($F_{discordant}/F_{shared}$ <2). Moreover, the filter for homopolymer regions yielded a $F_{discordant}/F_{shared}$ ratio of <2 for all three twin pairs, thereby precluding the usefulness of this filter.

With cumulative application of the filters, the predicted error ratio dropped considerably for all twin pairs (Table 4.2). Intrinsic to this study design is the retrieval of many sequencing errors, mixed with a potential few real SNV differences. Thus, most of the

putative discordant SNVs even after application of the filters are expected to be false positives. However, the probability of finding a real difference after systematic optimised filtering is theoretically better.

Twin pair	Before cumulative application of filters			After cumulative application of filters		
	Discordant SNVs	*Shared SNVs*	*Per variant error rate*	*Discordant SNVs*	*Shared SNVs*	*Per variant error rate*
IP17 and IP17	4793 (4.81%)	94975 (95.19%)	4.80%	40 (0.08%)	51121 (99.92%)	0.08%
RT1a and RT1b	12397 (9.07%)	124168 (90.92%)	9.08%	48 (0.07%)	62350 (99.92%)	0.08%
PD821 and PD161	14913 (10.24%)	130730 (89.13%)	10.24%	73 (0.12%)	64313 (99.89%)	0.11%

Table 4.2. The effects of various quality and repetitive DNA filters were assessed on the total number of shared and discordant SNVs between MZ co-twins. Per variant error rate is the number of discordant SNVs divided by the total variants detected. This is considerably reduced after cumulative application of the filters.

When combining both quality and repetitive filters together, a total of only 161 putative discordant SNVs remained between the three twin pairs. However, after manually reviewing each variant by viewing the .bam files in IGV, none of the variants were deemed convincing. This suggests that although the filters based on quality and repetitive DNA regions considerably reduced the predicted error rate, there were still many false-positive SNVs remaining.

4.2.2.4 Pipeline 3: VarScan2 and MuTect2

Whole-exome sequencing data were reanalysed by means of VarScan2 and MuTect2 using the annotated variant and genotype attained by the Haplotype Caller-based analysis as reference to explore the possible occurrence of low-frequency variants compatible with a mosaicism state.

As there is a possibility of the unaffected twin having a de novo mutation that is not present in the affected twin, a reverse pairwise analysis was also performed where the affected twin was classified as the 'normal' sample and unaffected twin as the 'tumour' sample (Table 4.4).

The resulting discordant variants were further filtered by excluding those variants that were likely to be non-functional, e.g., synonymous variants and/or variants outside the exonic regions. There are exceptions to this rule, such as for the twin pair discordant for lactase non-persistence (KEL and KIR), where causally-linked variants are likely to be found in intronic regions, with an MAF greater than 0.01.

Consistent with the literature, DNA samples that were LCL-derived yielded a higher-than-average discordant call rate (218 and 318; 421 and 422; 242 and 243). To reduce potential for bias in these samples, only variants found in genes previously implicated in ALS were considered for downstream analysis.

Description	Twin pair	Total germline variants detected	Discordant MuTect2	Discordant VarScan2	Shared discordant variants	VQS >90	Not within segmental duplications	MAF <1%	Exonic and non-synonymous
ALS	LAS	91,434	6,964	5,964	555	63	33	16	4
	SUS	112,182							
ALS	218	97,037	18,302	12,465	10,847	7,857	7,616	651	159
	318	93,543							
ALS	421	85,247	23,796	822	520	257	248	53	11
	422	107,451							
ALS	242	92,502	2,851	2442	346	130	107	32	13
	243	96,855							
Stroke	KG(s)	103,110	858	619	64	4	2	0	0
	HG(s)	98,053							
	KG(b)	105,715	1,302	735	95	8	4	1	0
	HG(b)	105,223							
Lactose intolerance	KEL	94,388	1,573	1,186	116	18	8	2	1
	KIR	108,944							
Inclusion body myositis	AFF	103,017	902	515	58	0	0	0	0
	UNAFF	92,857							
ADHD	RP	100,518	1,895	2,173	115	16	8	0	0
	OH	100,085							
Tourette's syndrome	490	95,279	1,088	1,412	99	7	4	0	0
	489	87,799							
Parkinson's disease	PD821	102,750	1,498	1,148	111	7	7	0	0
	PD161	104,715							
Hereditary spastic paraplegia	LF	101,483	1,537	1,249	99	15	6	2	0
	VF	103,849							
Schizophrenia	RT1b	78,727	1,870	1,490	127	8	6	1	0
	RT1a	104,366							
Schizophrenia	IP16	87,063	1,246	1,243	95	10	8	2	0
	IP17	92,607							

Table 4.3. Filtering protocol for SNVs and indels used to identify differences of functional variants between twin siblings. Discordant variants that were not shared between MuTect2 and VarScan2 were filtered out, and the shared discordant calls were then further filtered according to our exclusion criteria: VQS <90, within segmental duplications (SDs), MAF >1% (as per 1000g, cg69, ExAC), exonic, and non-synonymous.

Description	Twin pair	Total germline variants detected	Discordant MuTect2	Discordant VarScan2	Shared discordant variants	VQS >90	Not within segmental duplications	MAF <1%	Exonic and non-synonymous
ALS	LAS	91,434	2,158	3,600	525	13	10	10	4
	SUS	112,182							
ALS	218	97,037	3,781	3,862	566	243	220	20	4
	318	93,543							
ALS	421	85,247	15,600	-	-	-	-	-	-
	422	107,451							
ALS	242	92,502	14,328	9,976	8,072	5,528	5,343	376	84
	243	96,855							
Stroke	KG(s)	103,110	1,430	914	100	13	8	1	0
	HG(s)	98,053							
	KG(b)	105,715	1,189	676	70	6	5	2	1
	HG(b)	105,223							
Lactose intolerance	KEL	94,388	745	558	85	7	4	4	1
	KIR	108,944							
Inclusion body myositis	AFF	103,017	1,688	1,300	92	11	7	3	0
	UNAFF	92,857							
ADHD	RP	100,518	1,933	2,175	109	13	7	0	0
	OH	100,085							
Tourette's syndrome	490	95,279	2,081	1,941	149	20	10	1	0
	489	87,799							
Parkinson's disease	PD821	102,750	1,276	1,082	106	8	7	2	1
	PD161	104,715							
Hereditary spastic paraplegia	LF	101,483	1,389	1,104	102	12	8	0	0
	VF	103,849							
Schizophrenia	RT1b	78,727	1,354	1,145	115	6	4	1	1
	RT1a	104,366							
Schizophrenia	IP16	87,063	860	912	20	0	0	0	0
	IP17	92,607							

Table 4.4. Somatic variants specific to the affected twin were assumed to be of biological significance, but to determine the total amount of somatic mutations, variants specific to the unaffected twin were also checked. Discordant variants that were not shared between MuTect2 and VarScan2 were filtered out, and the shared discordant calls were then further filtered according to our exclusion criteria: VQS <90, within segmental duplications (SDs), MAF >1% (as per 1000g, cg69, ExAC), exonic, and non-synonymous. The VarScan2 analysis of twins 421 and 422 failed due to technical issues that are currently being resolved. This data is currently omitted.

The bam files for each twin were loaded into IGV and the short read genomic alignments for the potential discordant variants were manually inspected to remove additional artefacts that bypassed prior filtering parameters. These included variants located at the start and end position of reads, base quality scores that were less than 20 on average, homopolymer runs, and variants seen in genomic neighbourhoods with multiple nearby rare variants (suggestive of alignment artefacts caused by nearby indels). These variants were further inspected for intrinsic genome characteristics, such as segmental duplication, micro-satellites, and simple tandem repeats using online genome browsers Ensembl and UCSC.

Twin pair	*Gene*	*Region*	*Type*	*Chr*	*Position*	*Genotype*		*Depth of coverage (proband)*		*Variant frequency (proband)*	*Depth of coverage (MZ twin)*		*Variant frequency (MZ twin)*
						Reference	**Variant**	**Reference**	**Variant**		**Reference**	**Variant**	
LAS and SUS	*DENND5B*	Exon	Stopgain	12	31632586	C	A	48	4	8%	12	0	0%
	LOXHD1	Exon	Nonsynonymous	18	44157785	G	A	150	9	6%	40	0	0%
	BTK	5'UTR	Substitution	X	100645610	C	A	58	0	0%	14	3	18%
	SLC26A1	Exon	Nonsynonymous	4	985253	C	A	70	0	0%	22	3	12%
218 and 318	*FBXO38*	Exon	Nonsynonymous	5	147805180	G	A	11	12	52%	12	0	0%
	RIT2	Exon	Nonsynonymous	18	40323567	C	T	15	11	42%	8	0	0%
	GRM6	Exon	Nonsynonymous	5	178417745	C	T	11	10	48%	14	0	0%
	PPARGC1A	Exon	Nonsynonymous	4	23814713	T	C	69	0	0%	18	5	22%
421 and 422	*KIAA1107*	Exon	Nonframeshift deletion	1	92647643	CTT	-	16	0	0%	26	4	14%
	C8orf48	Exon	Frameshift deletion	8	13425060	AG	-	13	0	0%	22	9	29%
	ATP6V1B2	Exon	Nonframeshift deletion	8	20054932	GATGCG GGG	-	12	0	0%	16	5	24%
	NR1H2	3'UTR	Deletion	19	50885933	C	-	15	0	0%	19	7	24%

	GS1-259H13.2	Exon	Frameshift deletion	7	99208104	G	-	11	0	0%	15	9	38%
	ACTR3C	5'UTR	Substitution	7	150020654	G	A	10	0	0%	10	8	44%
	KBTBD3	Exon	Nonsynonymous	11	105924526	A	G	20	0	0%	45	7	14%
242 and 243	*TUBGCP4*	Exon	Nonsynonymous	15	43678059	C	T	20	0	0%	43	15	26%
	TFIP11	Exon	Frameshift deletion	22	26888040	A	-	70	0	0%	83	48	36%
	PHKA2	Exon	Nonsynonymous	X	18924724	C	T	14	0	0%	28	3	10%
	GNL3L	Exon	Nonsynonymous	X	54581036	A	G	11	0	0%	20	15	42%
RT1a and RT1b	*EML5*	Exon	Nonsynonymous	14	89151456	G	A	53	0	0%	24	5	14%
KEL and KIR	*PLCB1*	Exon	Frameshift deletion	20	8637865	A	-	37	12	26%	60	0	0%
	TMEM87A	3'UTR	n/a	15	42503678	C	T	8	0	0%	13	5	28%

Table 4.5. Details of candidate discordant variants from whole-exome sequencing data after cumulative application of the filters and manual reviewing using IGV.

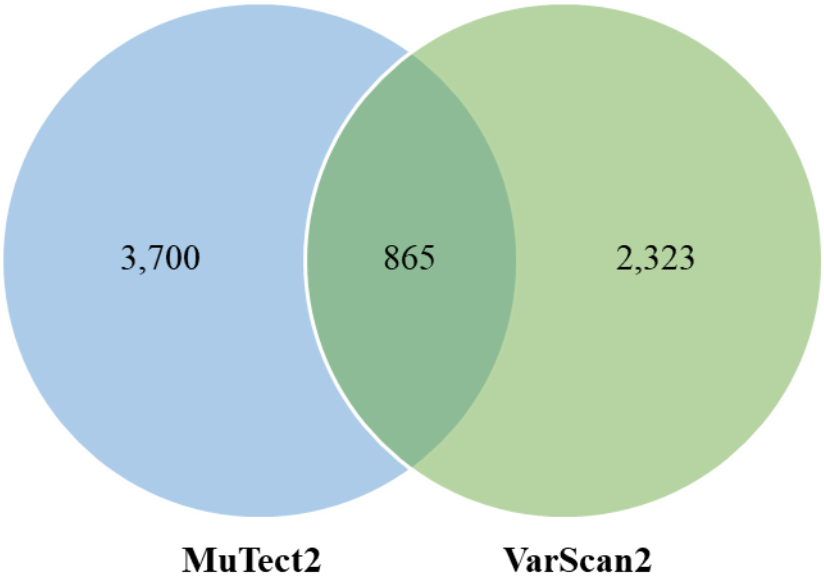


Figure 4.6. Venn diagram illustrating the overlap between MuTect2 and VarScan2. The two somatic mutation callers were used to detect differences between co-twins; only those that were shared were considered for downstream analysis. The figure shows the average number variants called in the total 28 pairwise comparisons.

4.2.3 Validation of twin-specific de novo SNVs

The candidate variant found on *EML5* (chr14:89151465; G>A substitution) in subject RT1a (Figure 4.3) had a somatic allele frequency of 17.24% according to pipeline 1. The same variant was detected in pipeline 3 with a 13.6% allele frequency (Table 4.5). In both cases, the variant was absent in the co-twin. When subjected to conventional Sanger sequencing, the electropherogram of the candidate variant on *EML5* showed no differences between the twins. The limitations of Sanger sequencing in its ability to detect low-level mosaicism has been previously documented (Rohlin et al., 2009). Constitutional heterozygosity is indicated on a Sanger sequence trace where the two nucleotides at a position are approximately of equal peak heights. Mosaicism lower than 20% is difficult to quantify, let alone detect, as the variant can be buried in the baseline

of the sequence trace. Thus, combinations of the conventional methods ought to be used to acquire an adequate sensitivity (Rohlin et al., 2009).

To verify this finding, three additional methods were used: Colony PCR (Figure 4.7), Sequenom MassARRAY genotyping (Figure 4.8) and Kompetitive Allele Specific PCR (KASP) genotyping (Figure 4.9). However, none of the employed SNP validation techniques confirmed the observed mosaic SNV, but rather showed identical wild-type genotypes. The variant seen in subject RT1a was determined to be a false-positive finding of high genotype quality, which was likely due to a PCR-induced mutation during the exome library preparation.

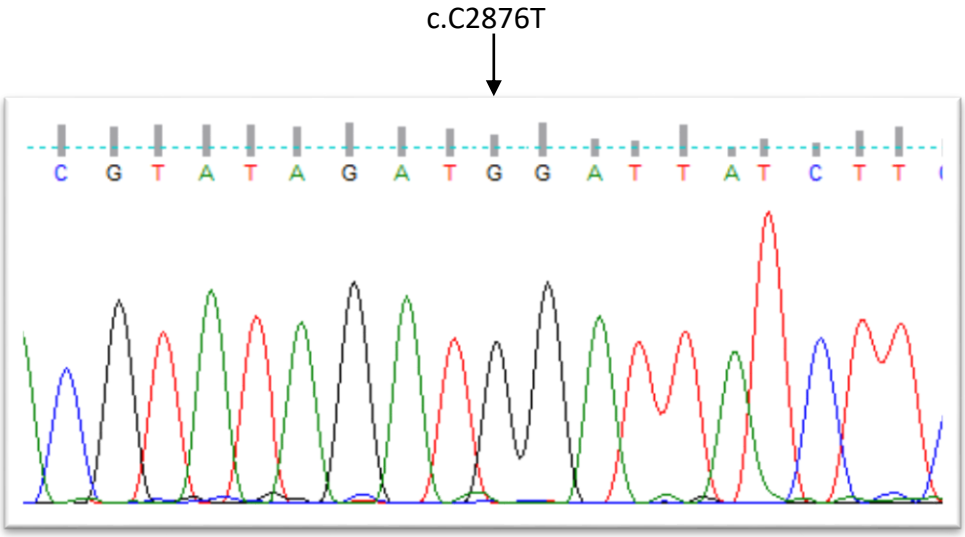


Figure 4.7. Bidirectional capillary Sanger sequencing was performed directly on bacterial colonies for the SNV on *EML5*. The picture shows one of the sequence traces (forward strand) of subject RT1a at position chr14:89151456. The PCR product spanning the variant was cloned for both twins. None of the 96 individual bacterial colonies that were randomly picked for colony screening showed the G>A substitution in either twin.

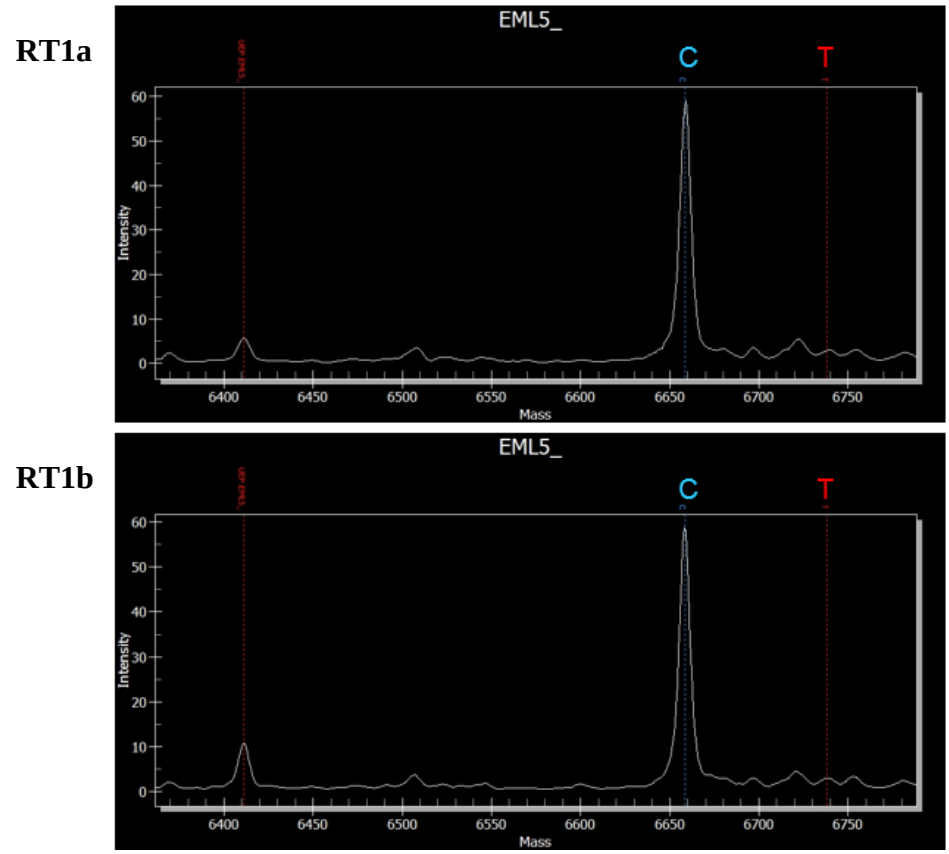


Figure 4.8. Sequenom MassARRAY genotyping was performed to validate the candidate variant (NewGene). The 'T' allele was not detected in either twin. The blue vertical dashed lines above the wave peak indicate the resulting genotypes, whereas the red ones indicate the expected location of the variant allele at the transcribed SNP.

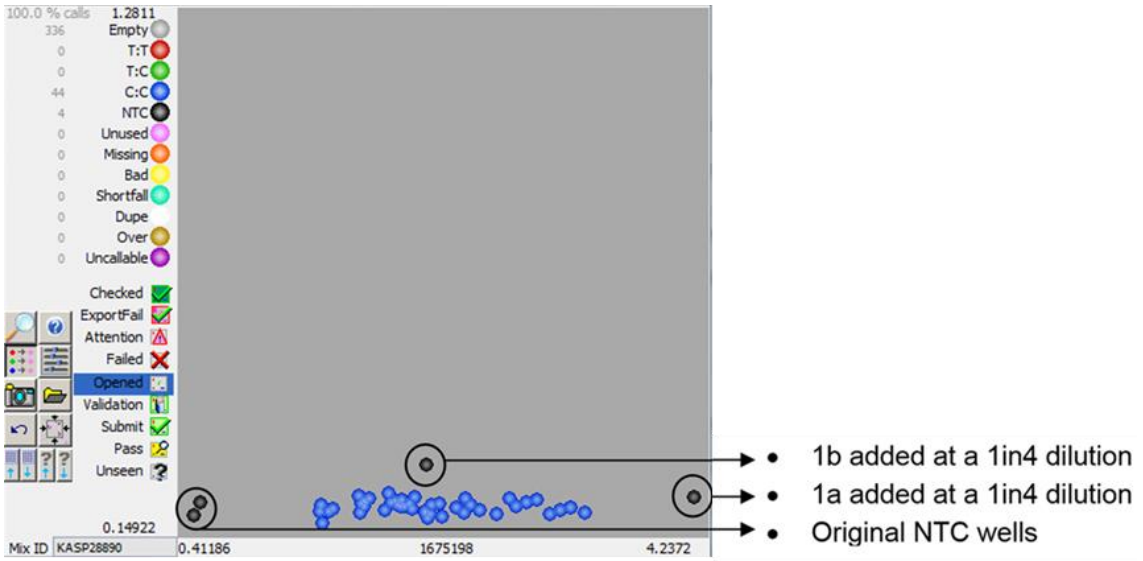


Figure 4.9. Genotyping of the SNV on EML5 was performed using competitive allele-specific PCR using KASP assays (LGC genomics). This method utilises fluorescent resonance energy transfer to quench fluorescence in reporter oligonucleotides until they are incorporated into allele-specific PCR products. The results also showed that the 'T' allele was not present in either twin.

4.2.4 Identifying pathogenic concordant variants

To test the hypothesis that rare, dominant or recessive variants could contribute to the complex disorders investigated, potentially damaging concordant exonic variants were examined. Several variants were identified in known disease-associated loci that could potentially explain disease onset according to a model taking into account the possibility of incomplete penetrance.

After application of the filtering criteria, each co-twin typically had 200 potentially damaging, rare concordant variants. All concordant variants were screened against lists of disease-specific susceptibility genes, which were obtained from various databases, including PubMed, OMIM, NIH GTR, DisGeNET, ALSoD, ALSGene, PDGene, SZDB, and SZGene. This produced a total of 113 variants in the twin cohort; however, by manually reviewing them in IGV we could remove mutations that are obvious artefacts of short-read sequencing and alignment. Mutations that looked real and were unambiguous were retained. These variants with their functional categories are shown in Table 4.6.

Homozygous variants with allele frequencies >1% are unlikely on their own to be pathogenic and were therefore filtered out of that data. Concordant homozygous recessive variants not found in public databases (i.e. novel variants) that also do not overlap segmental duplications or RepeatMasked sequence are not shown. Further studies are ongoing, particularly related to several genetic variants of high interest (not located in genes previously shown to be associated with human disease) that were found in both twins, and which may act as susceptibility factors in concert with putative interacting genetic and/or environmental factors.

Twin pair	*Gene*	*Region*	*Type*	*Chr*	*Position*	*RefSeq accession number*	*c.DNA position*	*Amino acid change*
LAS and SUS	*CNTN6*	5'UTR	n/a	3	1134367	NM_014461	c.-55326G>A	n/a
	CRIM1	Exonic	Nonsynonymous	2	36737128	NM_016441	c.G1504A	p.E502K
218 and 318	*AMPH*	Exonic	Nonsynonymous	7	38574539	NM_001635	c.C142T	p.R48W
	INA	Exonic	Nonsynonymous	10	105038008	NM_032727	c.A1040G	p.H347R
242 and 243	*DOC2B*	Exonic	Nonsynonymous	17	11873	NM_003585	c.C637T	p.R213W
	HFE	Exonic	Nonsynonymous	6	26093125	NM_139010	c.G289A	p.E97K
	VAPB	Exonic	Nonframeshift deletion	20	57016040	NM_004738	c.474_476del	p.158_159del
KEL and KIR	*LCT*	3'UTR	n/a	2	136545844	NM_002299	c.*50G>C	n/a
	LCT	Exonic	Nonsynonymous	2	136555659	NM_002299	c.A4916G	p.N1639S
	LCT	Intronic	n/a	2	136570613	n/a	n/a	n/a
	LCT	Intronic	n/a	2	136590558	n/a	n/a	n/a
	GLB1L	Intronic	n/a	2	220108217	n/a	n/a	n/a

PD821 and PD161	*HTR6*	Exonic	Nonsynonymous	1	20005151	NM_000871	c.C806T	p.T269M
	UNC5C	Exonic	Nonsynonymous	4	96141217	NM_003728	c.C1219T	p.R407C
LF and VF	*ZFR2*	Exonic	Nonsynonymous	19	3834902	NM_015174	c.G133A	p.V45M
490 and 489	*AADAC*	Exonic	Stop loss	3	151545958	NM_001086	c.T1198C	p.X400Q
	PDLIM5	Exonic	Nonsynonymous	4	95585202	NM_001256425	c.A800G	p.H267R
RP and OH	*DPP6*	Exonic	Nonsynonymous	7	154519496	NM_001290252	c.A461G	p.N154S
	CHD7	Exonic	Nonsynonymous	8	61777986	NM_017780	c.G8488A	p.A2830T
RT1a and RT1b	*SEMA6C*	Exonic	Nonsynonymous		151106475	NM_001178062	c.C1696T	p.P566S
	GABRR3	Exonic	Unknown	3	97753727	Unknown	Unknown	Unknown
	PCLO	Exonic	Nonframeshift insertion	7	82784833	NM_014510	c.1123_1124insCTCTTGGTCCTGCTAAGCCTCCAGCTC	p.Q375delinsPLGPAKPPAQ,PCLO
IP16 and IP17	*VLDLR*	Exonic	Nonsynonymous	9	2645718	NM_001018056	c.C1457G	p.A486G

Table 4.6. Concordant variants in known disease-linked genes.

4.2.5 Validation of the concordant variants

Based on Bayes theorem, the positive predictive value (PPV) of the pipeline depends on the number of mutations in the sample. In this context, the PPV can be defined as the percentage of variants identified by exome sequencing analysis that were validated as true variants – that is, [true positive/(true positive+false positive)]*100.

A high sensitivity and high specificity, whilst desirable, do not ensure a high PPV (which is also desirable). The PPV depends very much on the prevalence of de novo mutations. Validation rates from The Genome Institute tumour project ranged from 20% to >90%. This is because different types of cancer have different rates of mutation (Alexandrov et al., 2013; Meltz Steinberg et al., 2015). The validation rate was high for tumours that have a high somatic mutation prevalence (such as melanoma), and low for tumours that have a low mutation rate (such as acute myeloid leukaemia).

In MZ twins, it is expected that there will be a small number of de novo mutations (Kong et al., 2012, Campbell et al., 2012; Dal et al., 2014), rendering the PPV low relative to a tumour sample (with varying degrees depending on the type of tumour). Nevertheless, for germline variants, Meltz Steinberg et al. (2015) performed an evaluation of their own pipeline using a mixture of samples and found a sensitivity of 96.8% and a PPV of 93.1% for all SNVs. Considering that the concordant variants between co-twins were absent in all other samples, it would be extremely unlikely to obtain false-positives in the same gene location in both twin siblings. It was therefore deemed unnecessary – not to mention unfeasible – to validate all concordant variants with Sanger sequencing.

4.2.6 Parent-offspring trio analysis

A total of 217,290 variants were called in GATK's joint analysis. Variants shared by the MZ twins, but absent in their parents, were considered to be de novo germline mutations. After applying this initial exclusion criteria, a total of 424 and 412 putative de novo SNVs and indels were detected in twin pairs discordant for ADHD and Tourette's syndrome, respectively. Variants were further filtered as per similar parameters set for postzygotic de novo detection. Upon manual review in IGV most variants could be excluded on the basis that they were falsely miscalled in one of the parents. However, a nonsynonymous mutation in *RASD2*, a gene encoding for a GTP-binding protein Rhes on chromosome 22 (NM_014310:exon2:c.G170A:p.R57H), was found in the ADHD-discordant pair (Table 4.7). The variant is not reported in the dbSNP, 1000 Genomes, cg69 nor in the in-house database of 6,000 exomes. In the ExAC database containing more than 60,000 human exome data, the variant was found with an allele frequency of 8.13E-06 in the total population (allele count of 1/121112). The variant is also highly conserved across multiple species and predicted to be deleterious in online available bioinformatics tools.

	ADHD	*Tourette's syndrome*
Shared heterozygous variants in twins, but absent in parents	424	412
Remove segmental duplications	356	341
Variants in only exonic, splice site, 3'UTR and 5'UTR regions	98	84
Remove synonymous variants	70	60
Remove variants with MAF > 0.01	13	9
Variants remaining after manual review in IGV	1	0

Table 4.7. Filtering for germline de novo detection. A de novo mutation in *RASD2* was identified in the twins discordant for ADHD.

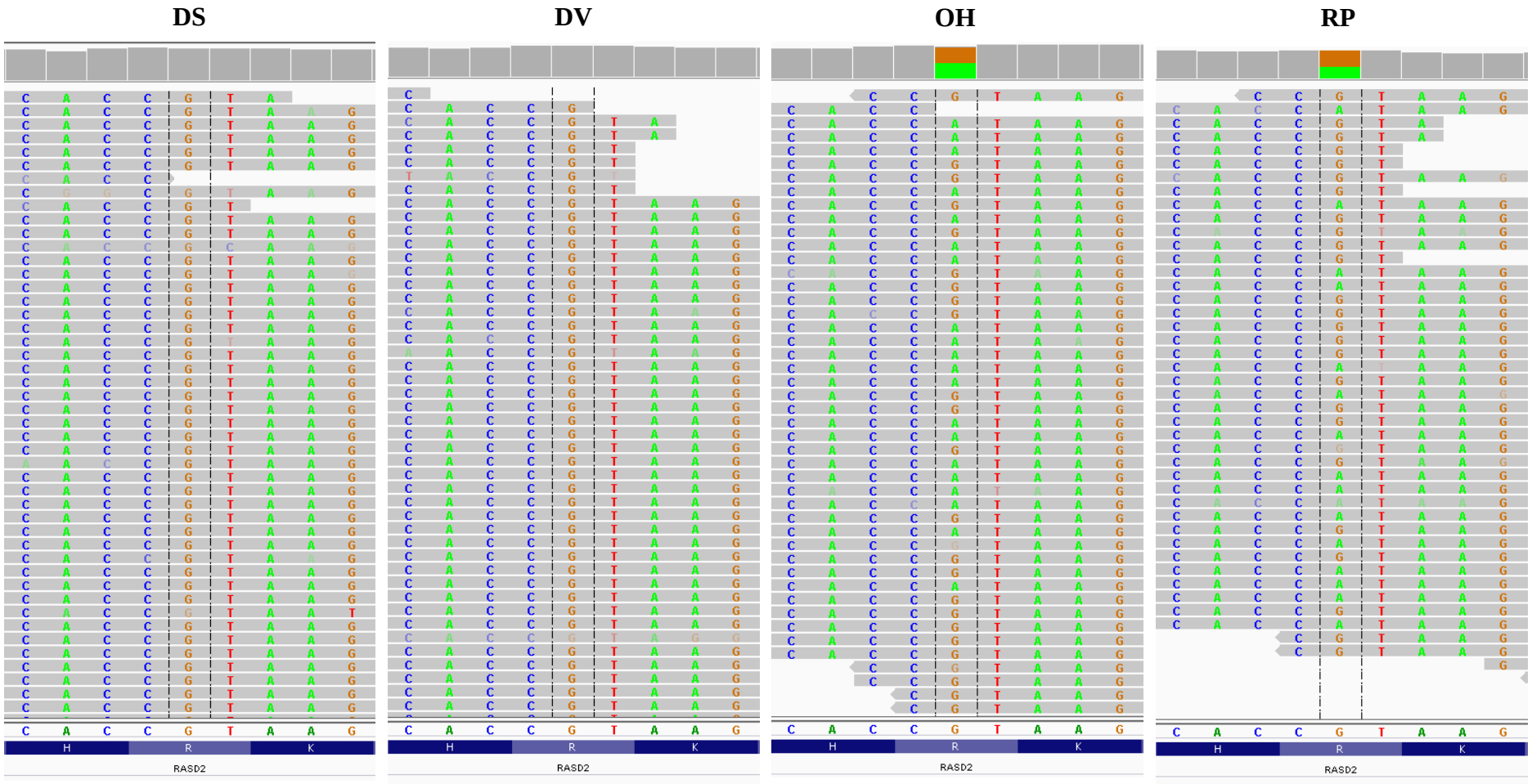


Figure 4.10. IGV screenshots showing a germline de novo mutation in the twin pair discordant for ADHD (OH and RP), which is absent in the parents (DS and DV). This region had a high depth of coverage, with the number of reads ranging from 70-100.

Because the father of the twins discordant for Tourette's syndrome also has the condition, it is likely that both twins had inherited variants associated with the disorder. We focused our attention on variants consistent with a dominant mode of inheritance – namely, variants that are homozygous or heterozygous in the affected father, absent in the mother, and heterozygous in the twins.

A filtering for rare or novel variants that were predicted to be damaging by at least one of the pathogenic prediction tools led to the identification of 41 variants shared between the twins. Only one of the variants found to be inherited from the father has previously been implicated in Tourette's syndrome, per our comprehensive list of 138 genes mined from various databases and search of literature. This was a stop loss mutation in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086:exon5:c.T1198C:p.X400Q). This variant was also found in the concordant variant analysis (Table 4.6).

4.2.7 Mitochondrial DNA analysis

We next tested the hypothesis that different levels of mtDNA heteroplasmy might account for the phenotypic discordance between the twins. After applying a minimum read count threshold of 10, a total of 399 shared and discordant variants were identified between the twins. These variants included 34 heteroplasmic variants and 365 homoplasmic variants. A total of 36 variants unique to either the affected or unaffected twin were verified using IGV. Among these, 23 were distributed on 12 genes throughout the mitochondrial genome, and 8 were localised at the hypervariable segments HV1 (16024–16383) and HV2 (57–372).

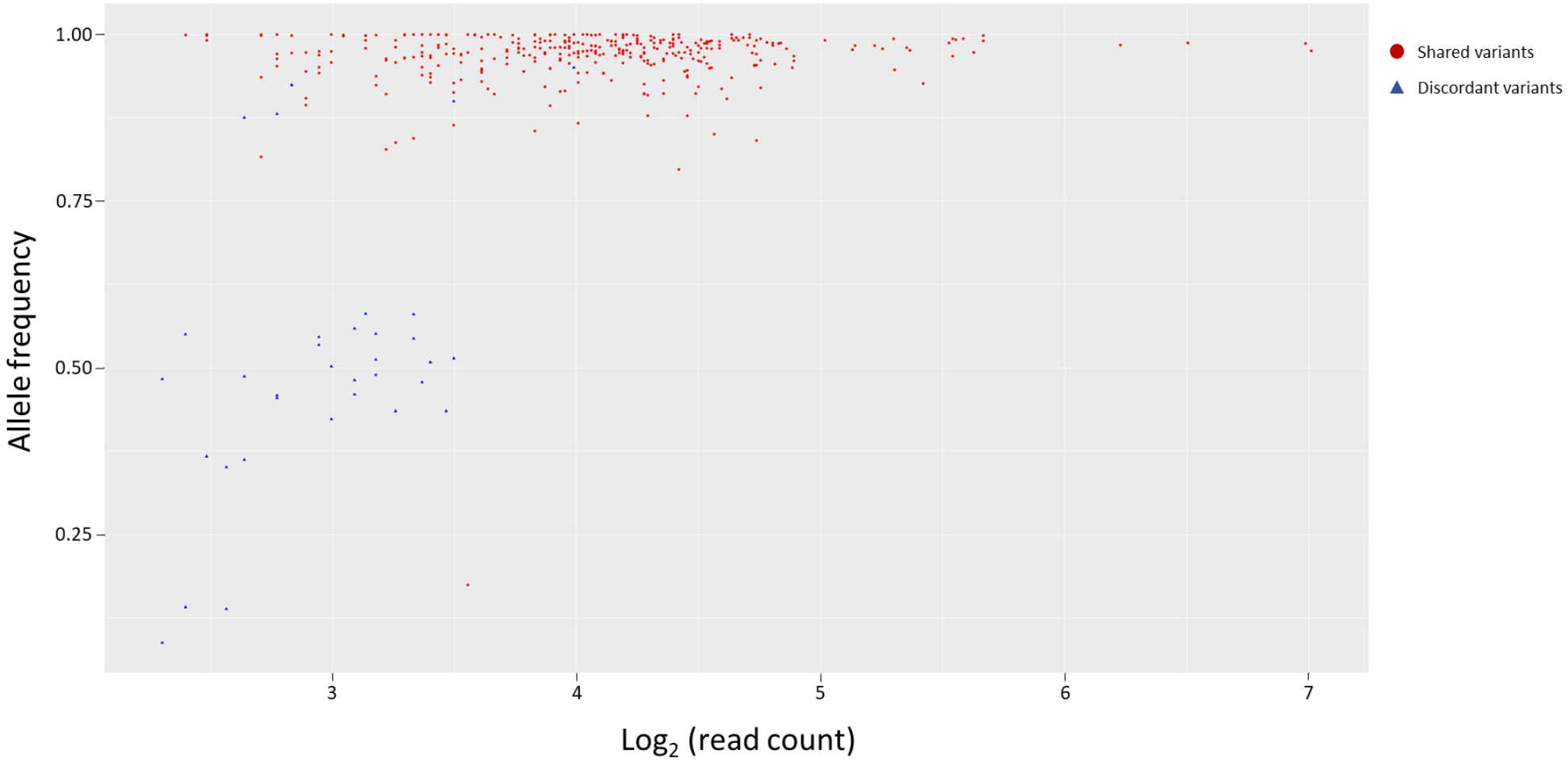


Figure 4.11. An allele frequency vs read count graph was plotted to get a visual representation of what variants were discordant between the twin pairs. Most of the variants clustered around 0.5 mark are from twin pairs 421 and 422. The DNA samples used for this pair are LCL-derived, and de novo mutations are known to be caused by the cell line transformation and culturing (Veltman and Brunner, 2012).

Twin pair	*Gene*	*Position*	*Genotype*		*Depth of coverage (proband)*		*Variant frequency (proband)*	*Depth of coverage (MZ twin)*		*Variant frequency (MZ twin)*
			Reference	**Variant**	**Reference**	**Variant**		**Reference**	**Variant**	
LAS and SUS	*MT-ND5*	13596	C	A	227	0	0%	136	10	9%
242 and 243	-	482	T	C	85	11	14%	93	0	0%

Table 4.8. Discordant mtDNA variants between twin pairs.

Most of the discordant variants came from twin pairs 421 and 422. These samples were excluded from further analysis due to the likelihood of artefacts created from transformed immortalised cell lines. Variants in hypervariable segments were also removed. This resulted in 2 high-confidence discordant variants between two twin pairs (Table 4.8). In twin 242 a variant was detected with 14% minor allele frequency in a non-coding region, which was not present in the unaffected twin. Further, a novel nonsynonymous mutation (c.1260A:p.S420R) was detected in twin SUS, which was not present in the co-twin affected with ALS.

No mtDNA sequence differences or different degrees of heteroplasmy were detected between any other individuals of the same pair. Thus, the phenotypic discordance of MZ twins with complex disorders investigated herein cannot be sufficiently explained by skewed distribution of mtDNA mutations.

4.3 Discussion

4.3.1 Assessment of variant calling pipelines

Advances in NGS technologies have improved our ability to characterise genomic sequence variation at a scale and resolution not previously possible. This has opened new avenues for studying how genetic variation contributes to human disease. A major challenge is how to process the copious information generated by the new technologies to yield high-quality data for downstream analyses. A variety of computational tools have been developed for this purpose.

However, each of these applications comes with a significant number of challenges. The most pertinent challenge is to determine which of the identified variants are errors, benign or pathogenic.

A whole-genome sequencing effort on four family members revealed, for instance, that sequencing errors are about 1,000 times more prevalent than true differences between the participants (Roach et al., 2010). A large proportion of these errors could be eliminated by sequencing a higher number of related individuals in this family. However, as mutations are expected to occur during every meiosis cycle, these de novo mutations may be missed when direct comparison to family members is used as a measure to decide which variants are errors. Furthermore, in many cases whole-genome sequences of related individuals may not be available or irrelevant to the study design. In such cases, other methods to distinguish sequencing errors from true variants are clearly needed. The situation becomes even more complex in the case of cancer genomes, in which numerous amplifications and deletion patterns are known to influence the effective depth of sequencing and alter the allelic balance for a given single nucleotide difference.

Thus far, large-scale validation with other sequencing methods has been used to independently validate SNVs detected in whole-genome sequences. For example, over 500 somatic variants in a tumour-normal sample from a lung cancer patient were validated using targeted mass spectrometry (Lee et al., 2010), whereas several other studies validated hundreds of SNVs using capillary Sanger sequencing (Pleasance et al., 2009; Friedberg, 2010; Dalgliesh et al., 2010; Reumers et al., 2012). This poses a considerable disadvantage as the independent validation of SNVs may quickly become as expensive and protracted as the analysis of the initial genome itself. To circumvent the need for such extensive validation analyses, sequencing projects in other smaller species have combined different technologies (e.g., long- and short-read technologies) to achieve more accurate sequencing (DiGuistini et al., 2009). At present, such an approach is, however, not cost- or time-effective for human genomes.

We hypothesised that prior to performing expensive validation or de novo sequencing experiments, in-depth bioinformatics analyses could be used to identify SNVs with high confidence. Thus, a reliable method for the prioritisation of high confidence SNVs was employed. Pipeline 2 was based on a) stringent sequence quality measures, b) the identification of error-prone regions in the genome and c) consensus calling of SNVs identified with different sequence mapping and variation calling algorithms. Using this approach, we could significantly lower the genome-wide error rate in MZ twin pairs, and provide more accurate theoretical estimates of the number of SNVs identified in their exomes.

Pipeline 3 was developed with the aim to rapidly identify and annotate variants. It utilises GATK for variant calling, which is based on the best practices for variant discovery analysis outlined by the Broad Institute. When comparing SNV calls from GATK and

SAMtools in Pipeline 2, it was found that GATK yields very high quality variant call data. Congruent with other studies (DePristo et al., 2011; Li et al., 2009), it was observed that recalibration of base quality scores before variant calling and realignment of mapped sequence reads around putative indels are important determinants of good performance.

Several previous studies have investigated factors that influence the accuracy of variant-calling algorithms with sequence data (DePristo et al., 2011; Liu et al., 2013; Li, Ruan and Durbin, 2008). One study sequenced 15 exomes from four families and processed the raw data using different alignment and variant-calling pipelines and found that there was low concordance between approaches (Liu et al., 2013). Another study used exome sequence data on 20 individuals and simulated whole-genome sequence data to compare different algorithms for variant-calling. Consistent with our results, this study found that GATK outperformed SAMtools, especially for low coverage data, and yielded the most accurate data with multi-sample calling (Ruan and Durbin, 2008).

Pipeline 1 used an older version of VarScan, and only showed variants with a somatic allele frequency of >15%. While this helped reduce putative discordant variants to a more manageable number, it may have eliminated variants with low level mosaicism. Interestingly, none of the putative discordant variants identified in Pipeline 2 were identified with the other methods, demonstrating the limitations of current variant calling algorithms. To detect low-level mosaic variants that were potentially missed using Pipeline 1, two tools were used in parallel in Pipeline 3, and the results were separately evaluated to attain the highest possible sensitivity and hence lowest false-negative ratio.

Using VarScan2 and MuTect2 to identify discordant somatic variants demonstrates an innovative technique in twin research. In contrast to other variant detection tools (including EBCall, JointSNVMix, LoFreq, SomaticSniper and Strelka), VarScan2 had a

lower false-positive rate and identified more true somatic SNVs (Wang et al., 2013; Huanget al., 2015). Moreover, VarScan2 was found to have the highest specificity (99.9998%) and sensitivity (97%), and was best at detecting variants present at low-allelic fractions (as low as 1%), and had perfect precision and recall for calling indel variants (Stead et al., 2013; Spencer et al., 2014; Kockan et al., 2016; Liu et al., 2016). However, MuTect2 identified more low coverage somatic variants (Cai et al., 2016). Although MuTect2 returned fewest candidates, it had excellent capability in both control of false calls and discovery of potential true positives. MuTect2 refers to dbSNP database to aid its classifier in discrimination between germline and somatic variants, and is shown to be efficacious. Although VarScan2 was as efficient in low-frequency mutations detection, it exhibited an advantage in discovering somatic SNVs with relatively high frequencies, which makes it a beneficial supplement of MuTect2 (Cai et al., 2016). As the two variant callers have their strengths and weaknesses, taking the union would avoid biases from either one.

It would be no surprise if the bioinformatics methods and protocols outlined herein soon become outdated. This is owed to the fact that the methodological approaches for processing NGS data are continually evolving. It is therefore important to stay abreast of these techniques, and evaluate which one would yield the highest quality data for the desired study design.

4.3.2 Discordant variants

Analysing the exome sequencing data in accordance with Pipeline 3 yielded twenty-two high-confidence discordant variants within the thirteen twin pairs investigated. Putative discordant variants could be excluded upon manual reviewing in IGV if they were deemed to be artefacts – that is, variants were either called incorrectly in one of the co-twins (false

positive), or incorrectly not called in one of the co-twins (false negative). False negatives are likely to be the result of sequence data issues or stochastic sampling bias in the twin exome that lacked the variant.

A nonsynonymous exonic variant was identified in *PPARGC1A* (NM_013261:exon9:c.A1829G:p.H610R), in the unaffected twin of the ALS-discordant pair 218 and 318. The affected twin (218), a fifty-four-year-old Caucasian male, experienced signs of ALS from the age of fifty. The site of symptom onset was in the upper limbs. Additional clinical features supporting the diagnosis were upper motor neuron signs in bulbar, cervical (upper limbs) and lumbosacral (lower limbs) regions, with definite lower motor neurone signs in bulbar and cervical regions. The electromyogram (EMG) studies, a test used to record the electrical activity of muscles, demonstrated acute cervical denervation. The affected twin was prescribed riluzole since the diagnosis was made.

The revised ALS Functional Rating Scale (ALSFRS-R) provides an estimate of the patient's degree of functional impairment based on a series of questions. Each question is graded from 1 to 4, and scores are summed to produce a reported score ranging from 0 to 48. The higher the score the more function is retained. At the time of diagnosis, the ALSFRS-R of the affected twin (218) was 41, however follow up data (duration of interval not known) showed a marked reduction to 27.

There was no family history of ALS or other motor neurone disorders, however his father and paternal grandfather had the diagnosis of Alzheimer's disease, and several paternal aunts and uncles were diagnosed with a non-specified form of dementia. Both twins were diagnosed with hypertension.

The clinical finding of signs suggestive of combined upper and lower motor neurone lesions that cannot be explained by any other disease process, together with progression compatible with a neurodegenerative disorder, led to a primary clinical diagnosis of ALS in the affected twin. However, the unaffected twin remained asymptomatic.

Paradoxically, the mutation occurring in the unaffected twin might explain their discordance. *PPARGC1A* encodes for peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PGC-1α), a nuclear transcriptional coactivator that plays a vital role in mitochondrial biogenesis, respiration, insulin secretion, gluconeogenesis, and muscle fibre-type switching (Lin et al., 2004; Albani et al., 2016). It is interesting to note that the role of *PPARGC1A* in ALS is well-documented (Qi et al., 2015; Albani et al., 2016; Bayer et al., 2017; Buck et al., 2017). This gene has also been related to molecular pathways involved in other neurological disorders, including Parkinson's, Huntington's and Alzheimer's disease, stroke and multiple sclerosis (Cui et al., 2006; Róna-Vörös and Weydt, 2010; Ghosh et al., 2015).

Mouse models of ALS and Duchenne muscular dystrophy have demonstrated that increased expression and activity of PGC-1α can lead to an improved clinical outcome (Handschin et al., 2007b; Liang et al., 2011) by augmenting mitochondrial biogenesis and reactive oxygen species detoxification (Austin and St-Pierre, 2012). ALS is in effect a motor neuronal disorder of a degenerative origin, and oxidative stress appears to play a key role in the pathogenesis.

Although recessive mutations tend to inactivate the affected gene and lead to a loss-of-function (LoF), dominant mutations often lead to a gain of function by, for example, increasing the activity or conferring a new activity of a given gene product. Thus, it is possible that the mutation in *PPARGC1A* in the unaffected twin has resulted in an

increased expression of the protein, thus providing a protective effect from an already predisposing illness in the twin pair. This is supported by the effects of PGC-1a in transgenic models of ALS, as well as human studies. For instance, in a genetic association study, Eschbach et al. (2013) found that SNPs in the brain-specific promoter region of PGC-1a modifies the clinical manifestation of human ALS. Remarkably, the modifier effect of PGC-1a on ALS onset and survival was in a strictly male-specific manner, and was more pronounced in humans than in mice (Eschbach et al., 2013).

Our finding has the potential to further elucidate the role of this antioxidant defence transcriptional coactivator in ALS. Validation of our findings using additional technologies, such as Sanger sequencing or allele-specific PCR capable of detecting low-level mosaicism, will be implemented as part of future studies, a lack of which herein is acknowledged as a limitation.

4.3.3 Concordant variants

In addition to identifying discordant variants, we sought to examine shared variants with predicted pathogenicity. This included rare homozygous and heterozygous variants, and those in known disease-susceptibility genes. Phenotypic discordance between twin pairs could be explained in a number of ways. For instance, if the variants exhibited incomplete penetrance, the unaffected twin could develop the disorder later (for example, Chapter 2 documents a shared pathogenic hexanucleotide repeat expansion in ALS-discordant twins 421 and 422). Oligogenic mechanisms may also be a factor.

Concordant variants in exonic, splice site, promotor, 5'UTR and 3'UTR regions were cross-compared against a list of disease-linked genes formulated via a comprehensive search of literature and gene databases. The number of concordant variants could be further reduced by removing those found in multiple other samples, repetitive sequences

or systematic mismapping of paralogous sequences. In total, 23 shared variants were identified in known disease-susceptibility genes.

As there will likely be a larger number of shared variants in novel genes (previously not related to the primary diagnosis of the affected twin), the search was restricted to only nonsynonymous variants in exonic regions, the data for which is not shown. The concordant variants found in exome sequencing should ideally be validated with an independent method, but this is impractical for the many variants identified. Nevertheless, as described above, it would be very unlikely for both twins to receive false-positive readings at the same gene location.

It may prove insightful to analyse the genes harbouring potentially pathogenic variants using pathway analysis tools, such as Ingenuity Pathway Analysis or Cytoscape. This would have the potential to identify major canonical pathways as well as networks overexpressed in the gene lists and describe the intricate relationships between the genes in the pathway. These pathway analysis tools are also able to identify potential disease associations through gene ontology classification. The lists of concordant variants identified within each twin pair will thus be of use to future genetic studies related to the disorders investigated.

4.3.4 De novo mutation detection in parent-offspring trio analysis

4.3.4.1 Attention deficit hyperactivity disorder

A nonsynonymous mutation (c.G170A:p.R57H) within the RASD family member 2 (*RASD2*) was found in both ADHD-discordant twins (OH and RP), but absent in their parents. *RASD2* belongs to the Ras superfamily of small GTPases and is enriched in the striatum and involved in the modulation of dopaminergic neurotransmission (Vitucci et

al., 2015). *RASD2* is located on chromosome 22q12.3, a region that harbours numerous susceptibility loci for psychosis (Potash et al, 2003), and has been suggested to be a vulnerability gene for neuropsychologically defined subgroups of schizophrenic patients (Liu et al., 2008). Currently, the co-twin has not been officially diagnosed but anecdotally has been showing clinical features of ADHD.

In a knockout mice model with targeted deletion of *Rasd2*, Vitucci et al. (2015) found that the absence of Rasd2 significantly increases the behavioural sensitivity to motor stimulation with administration of psychotomimetic drugs, such as amphetamine and phencyclidine. Based on these findings, and the postulate that *RASD2* influences prefronto-striatal phenotypes in humans, the authors hypothesise that a genetic mutation resulting in a reduction of this G-protein might play a role in cerebral circuitry dysfunction, resulting in exaggerated psychotomimetic drug responses and the development of specific phenotypes linked to schizophrenia-like symptoms (Vitucci et al., 2015). Interestingly, people who have ADHD (Kim-Cohen et al., 2003), or a family history of the disorder (Larsson et al., 2013), are more likely to develop psychotic disorders. The co-occurrence of ADHD and schizophrenia is suggested to be due to shared genetic factors, rather than aetiologically distinct subsyndromes (Larsson et al., 2013). Moreover, stimulants, such as amphetamines, are the most common pharmacological treatment for ADHD (Bolea-Alamañac et al., 2014). It is not known if the affected twin (RP) was prescribed medication for the condition.

4.3.4.2 Tourette's syndrome

For an individual to be diagnosed with Tourette's syndrome, they must display multiple motor tics (e.g. blinking or shrugging the shoulders) and at least one vocal tic (e.g. humming, throat clearing, or shouting out a phrase), both of which must be ongoing for

at least one year. Tics are abrupt, rapid, non-rhythmic, persistent, stereotyped motor movements or vocalisations. Individuals who have either motor or verbal tics (but not both) for more than a year are given diagnoses of chronic motor tics or chronic verbal tics, respectively.

The forty-four-year-old father (487) of American Indian descent was given a diagnosis of Tourette's syndrome at seven years of age, but experienced motor and vocal tics from the age of six. This follows the typical development of Tourette's syndrome, which manifests in early childhood with symptoms peaking before puberty. Tourette's syndrome can often co-occur with other neuropsychiatric disorders, such as ADHD and obsessive-compulsive disorder (OCD), and it is often these co-occurring conditions that bring affected individuals to seek medical attention. In line with this, the father had a secondary lifetime clinical diagnosis of OCD (onset nine years) and ADHD (onset twelve years). Although there is no prior family history of Tourette's syndrome, OCD or ADHD, his mother had a diagnosis of anxiety disorder.

The affected twin (490) similarly experienced tics from the age of seven, and was given a diagnosis of Tourette disorder at eight. No other conditions (such as ADHD or OCD) were evident in the affected twin by fifteen years of age, when DNA samples were collected from the family. The unaffected twin (489), moreover, remained asymptomatic for all conditions. The twins have a fourteen-year-old brother who is also asymptomatic (DNA samples for this individual was not obtained). The forty-four-year-old mother (488) of Caucasian descent has no primary lifetime clinical diagnosis or family history of related disorders. The racial category in the medical notes classify the twins as 'Caucasian', however it is evident that twins have mixed heritage parents.

No pathogenic germline de novo mutations were identified in the twin pair. However, a shared stop loss mutation in *AADAC*, a gene encoding for arylacetamide deacetylase on chromosome 3 (NM_001086:exon5:c.T1198C:p.X400Q), was inherited from the father.

In a recent study by Bertelsen et al. (2016), the authors identified eight patients with overlapping deletions of *AADAC* in a Danish cohort of 243 patients with Tourette's syndrome and 1571 ancestry-matched control subjects. Although there was no statistical significance in this cohort, the gene remained their most promising candidate. The authors proceeded to investigate cohorts from an additional five countries; namely, Iceland, the Netherlands, Hungary, Germany, and Italy. A final meta-analysis of 1181 patients and 118,730 control subjects from these countries determined the association to be significant. Further, functional studies demonstrated that *AADAC* is expressed in several brain regions previously implicated in the pathophysiology of Tourette's syndrome, including the Purkinje cell layer of the human cerebellum (Bertelsen et al., 2016). The CNV overlapping *AADAC* is the only one to date that has been successfully associated with Tourette's syndrome. A previous CNV study found a tentative link between the gene *FHITT* and the disorder, however on closer inspection this association was determined to be the result of population stratification (Fernandez et al., 2012). Bertelsen et al. (2016) avoided this Type 1 error, which was caused by within-population differences in ancestry between cases and control subjects, by ensuring uniform ancestry within each population.

The BrainSpan Atlas of the Developing Human Brain provides RNA sequencing and exon microarray data of various cortical and subcortical structures across the full course of human brain development (http://www.brainspan.org). Remarkably, the publicly-available transcriptome profiling data illustrates that *AADAC* expression peaks in the striatum between birth and adolescence. This is consistent with the typical clinical time

course of tic onset, and indeed the age of onset of Tourette's syndrome in the father and the affected twin investigated herein.

Considering the above evidence, *AADAC* remains an interesting candidate susceptibility gene for Tourette's syndrome. The stop loss mutation detected in the father and twins warrant functional studies to investigate the role of this gene in the pathogenesis of this disorder.

4.3.5 Mitochondrial DNA analysis

The mitochondrial genome, unlike nuclear DNA, is located in mitochondria and is exclusively inherited from the mother through the cytoplasm (Falkenberg, Larsson and Gustafsson, 2007). Previous studies have suggested an uneven distribution of mitochondria at the time of embryo separation, which starts from the two-cell stage and leads to the differential pattern of mtDNA heterogeneity between daughter cells (Lee et al., 2012). These findings indicated that the variability of mtDNA heterogeneity may exist between identical twins.

Thus, we also investigated differences of the mtDNA variants and levels of heteroplasmy between the twin pairs. However, no clear evidence was found for an mtDNA contribution to the phenotypic difference in these MZ twins, suggesting that heteroplasmy of common mtDNA variants might significantly contribute to the risk of the complex traits investigated herein.

However, there were an unusually high number of discordant variants in one twin pair (421 and 422), mainly detected in the affected twin. This is consistent with the somatic variant analysis on this twin pair (Tables 4.3 and 4.4), whose DNA samples, as mentioned previously, were derived from EBV-transformed B-lymphocytes. Other studies have shown a significant bias in using LCL-derived DNA, as de novo mutations can result

from cell line transformation and culturing (Veltman and Brunner, 2012). It has been shown that EBV promotes genetic instability in the host (Gruhne et al., 2009), and can cause mutations through integration and disintegration into the host's genome (Morissette and Flamand, 2010).

This preliminary analysis will be part of a larger project looking at sequence variations in the mitochondrial genome of over 5,000 individuals, where a more comprehensive analysis will take place. Albeit, the findings of the present analysis agree with previous studies concerning the role of mtDNA heteroplasmy in discordant MZ twins with other human disorders (Detjen et al., 2007), indicating that factors, such as epigenetic changes and environmental factors, may play a more important role in shaping the discordance between MZ twins with these conditions.

4.3.6 Considerations and limitations

A major strength of this study is the presence of discordant MZ twins for a range of clinical phenotypes. We included the genomes of parents of RP and OH, and 489 and 490 for parent-offspring trio analysis. We also cross-compared DNA obtained from blood and saliva in subjects HG and KG for exome sequencing analysis.

Ideally, DNA obtained directly from affected tissues would be more reliable in detecting de novo mutations and somatic mosaicism, especially if the timing of mutagenesis occurred later in development. However, this was impractical as the nature of most of the disorders investigated in this study entails obtaining brain, muscle or bowel biopsies.

Ectodermally-derived buccal epithelial cells have a closer embryological connection to brain tissue than leukocytes from blood. The epithelial lining of the mouth and the nervous system are both ectodermal derivatives, whereas the mesoderm ultimately gives rise to muscle, vascular tissue and the haematopoietic system (Freshney, 1987). This was

corroborated in a study looking at the aetiological basis of DNA methylation in psychiatric disorders. DNA methylation in saliva resembled patterns from each of the different brain regions more closely than methylation in blood (Smith et al., 2014). Thus, our choice of tissue samples did not pose a major challenge in mosaicism detection.

Although the identification of high-confidence discordant de novo variants between co-twins have the potential to offer novel insight into the extensive aetiological heterogeneity implicated in complex traits, they remain to be independently validated. Discordant variants were not identified in several twin pairs. The low rate of de novo mutations within the exonic regions of the genome is consistent with other published data (Li et al., 2014; Dal et al., 2014). Where no differences were found, it is therefore possible that the analysed exomes are truly genetically identical between these co-twins. However, while we have obtained good data quality and applied thorough analytical methodologies, it remains the case that we may have overlooked existing differences in somatic mutation detection.

One reason for this is that exome sequencing covers a mere 1-2% of the genome, and does not assess variation in noncoding gene regulatory elements, apart from when they are near exonic regions. However, our hypothesis of postzygotic mutations to explain twin discordance renders exome sequencing highly appropriate, given that phenotypic differences are more likely to be caused by nonsynonymous mutations in protein-coding regions of the genome. Nevertheless, it remains possible that discordant somatic mutations affecting the phenotype in our twin cohort may be found in regions outside the accessible exonic regions of the genome.

4.3.7 Summary

The discordant MZ twin strategy employed in this study to identify candidate genes involved in the pathogenesis of complex traits is reasonable and practical. Notwithstanding, the actual search for discordant variants is not an easy task. For some of the twin pairs, no convincing variants were found to explain their discordant phenotypes. This suggests that genetic alterations in these twins might lie outside the accessible regions, or have yet uninvestigated epigenetic differences.

Several studies failed to find somatic mutations using NGS technology between discordant MZ twins, such as for multiple sclerosis (Baranzini et al, 2010), VACTERL association (Solomon et al, 2013), congenital renal agenesis (Jin et al, 2014), Crohn's disease (Peterson et al, 2014), congenital heart defect and epilepsy (Chaiyasap et al, 2014), congenital hypothyroidism due to thyroid dysgenesis (Magne et al, 2015), and ALS (Meltz Steinberg et al, 2015). And in studies where differences were found, they were masked by a substantial number of false positives. For instance, Reumers et al. (2012) used whole-genome sequencing on MZ twins discordant for schizophrenia, and of the 846 potential discordant variants found post-filtering, only two were confirmed by Sanger sequencing. This filtering method was employed in our Pipeline 2 analysis, which confirmed the high miscall error rate of whole-genome and -exome sequencing. Indeed, Neilson et al. (2011) found the error rate of the Illumina platform to be approximately 1%. Considering this, and given that there are ~3 billion base pairs in the human genome, one can anticipate to find ~30 million miscalled positions by whole-genome sequencing. However, with the rapidly evolving bioinformatics tools at our disposal and advances in genomic technologies, a comprehensive search for somatic mosaicism is becoming realistic.

This study supports the polygenic nature of the complex disorders investigated and the threshold model for their manifestation. Taking the union of MuTect2 and VarScan2 to identify somatic variants offers a proof of concept for assessing the genetic aetiology of complex traits in discordant MZ twins. As well as individual-specific variants, various potentially pathogenic concordant SNVs were identified in disease susceptibility genes. While the latter may suggest a reason for disease manifestation in the affected twins, it does not explain their discordant phenotypes. Future work will involve investigating other potential contributory factors, such as epigenetic changes, discordant SNVs in non-coding regions of the genome, oligogenic mechanisms, environmental agents, or tissue-specific somatic mutations.

Chapter 5. Copy number variation in discordant monozygotic twins

5.1 Overview

In Chapter 4 we explored methodological approaches in identifying germline and somatic variants and their theoretical role in complex disorders. This chapter documents high-density microarray analysis with the aim to identify large genomic structural rearrangements of potential clinical significance.

Redon et al. (2006) defined a CNV as a section of DNA that is ≥1 kb and present at variable CN when compared to a reference genome. CNVs are structural variations that comprise deletions, insertions, duplications and complex multi-site variants, which alter the diploid status of DNA. They have been found in all human populations and are widespread throughout the genome (Freeman, 2006). Previous studies have demonstrated that postzygotic CNVs can cause discordant phenotypes in MZ twin pairs that otherwise are genetically identical (Bruder et al., 2008).

There is now overwhelming evidence that CNVs play a significant role in normal population variation and evolution (Ionita-Laza et al., 2009; Wong et al., 2007; McCarroll et al., 2008) and significant CNV associations have been found to underlie disease. Although postzygotic de novo CNVs have been detected in discordant MZ twins, many studies testing this hypothesis have reported no or limited evidence (Abdellaoui et al., 2015; Laplana et al., 2014) Additionally, there are reports of CN differences in MZ twin pairs without discordant phenotypes (Abdellaoui et al., 2015; Magnusson et al., 2016).

Some CNVs represent recurrent polymorphisms occurring at a low rate in the population and have no known clinical significance. However, is it estimated that 4.8–9.5% of the genome contributes to CNV, many of which are in areas of known genes and have the potential to impact phenotype (Zarrei et al., 2015; Iafrate et al., 2004)). The average number of CNVs per person ranges from 10-50 based on several reports using current SNP-based microarrays (Iafrate et al., 2004; McCarroll et al., 2008), with an estimated rate of de novo CNVs lying between 0.01 and 0.02 events per generation (Acuna-Hidalgo, Veltman and Hoischen, 2016).

In this study, we used a SNP microarray platform and various computational methods to identify CNVs in thirteen pairs of MZ twins with varying phenotypic discordances for complex disorders. With the aim to identify potential genetic factors that influence disease manifestation, it was hypothesised that shared or discordant CNVs could increase the risk of disease onset even in discordant twin pairs. To investigate this hypothesis, the burden of rare CNVs and CNVs overlapping disease-linked genes were analysed.

5.2 Results

DNA from the entire twin cohort (n=26) and two sets of parents (n=4) recruited in this study were processed on the Illumina HumanCore BeadChip 12v1. All processing was carried out in accordance with the Infinium HD Ultra Assay protocol (Rev B, 2010, Illumina Inc, San Diego, USA).

5.2.1 CNVs affecting disease-susceptibility genes

After CNV merging, a total of four discordant *de novo* CNV duplications were found in three subjects – namely, chr1:231711489-231803604 in UNAFF; chr4:15776181-15840839 and chr4:139790626-139915883 in RP; and chr8:105981846-106108593 in 490. However manual inspection revealed that these variants are shared with each of the corresponding co-twins. In other words, the apparent discordant CNVs were undercalled (false negative) in the twin that originally didn't have the CNV called by pennCNV. CNVs were called if they are covered by ≥10 probes, but in some instances, they spanned <10 probes so were filtered out of the data. Next, we lowered the probe threshold to ≥3 to detect smaller CNVs that would potentially be filtered out of the data. As this is expected to result in a higher frequency of false positive calls, CNVs were also called using cnvPartition, and CN segments were only included in further analysis if the CN calls agreed between both algorithms. The results obtained from SNP array analysis are summarised in Table 5.1.

These putative CNVs were compared against exome sequencing CNV calls. Only samples from Pipeline 1 (Table 4.1) were analysed using ExomeDepth. ExomeDepth yielded approximately 130 CNVs per sample, of which about 90% calls per sample are known in the population. This estimate was determined by comparing the probe locations matching with published CNVs (Conrad et al., 2009; Durbin et al., 2010). Most of the

samples had around a 1:1 deletions/duplications ratio. CNVs that were consistent between both SNP array and exome sequencing platforms were considered for experimental validation using ddPCR.

We focused on subsets of genes that are associated with known phenotypes in disease databases such as OMIM and DisGeNET, or genes that are intolerant to LoF mutations based on the Residual Variation Intolerance Score (RVIS) or the probability of being loss-of-function intolerant (pLI) score (Petrovski et al., 2013) (Table 5.2). An RVIS <0.0 means that a given gene has less common functional variation than expected, and is referred to as 'intolerant'; whereas an RVIS >0.0 indicates that a gene has more common functional variation. Genes with high pLI scores ($pLI \geq 0.9$) are extremely LoF intolerant, whereas genes with low pLI scores ($pLI \leq 0.1$) are LoF tolerant.

Disease status	ID	Chr	Start position	End position	CN 1(0or1) 2(3or4)	Size (bp)	No of probes	Genes
ALS	LAS	2	203553836	203665782	2	111946	8	*FAM117B, ICA1L*
		12	31266287	31406907	2	140620	19	*OVOS2*
	SUS	2	203553836	203665782	2	111946	8	*FAM117B, ICA1L*
		12	31296219	31406907	2	110688	15	*OVOS2*
ALS	218	3	160154258	160156933	2	2675	8	*TRIM59*
		21	14669931	14844368	1	174437	9	*FGF7P2, C21orf110*
	318	3	160154258	160156933	2	2675	8	*TRIM59*
		21	14669931	14844368	1	174437	9	*FGF7P2, C21orf110*
ALS	421	-	-	-	-	-	-	-
	422	-	-	-	-	-	-	-
ALS	242	4	116880079	117095295	1	215216	14	-
		4	131965131	132177937	2	212806	17	-
		22	33875161	33928049	2	52888	7	*LARGE1*
	243	4	116880079	117095295	1	215216	14	-
		4	131965131	132218117	2	252986	19	-
		22	33875161	33928049	2	52888	7	*LARGE1*
Stroke	KG(s)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	18	-
		20	23677829	23725019	2	47190	8	
	HG(s)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106154088	2	76230	16	
		20	23671740	23725019	2	53279	9	
	KG(b)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	12	-
		20	23671740	23725019	2	53279	9	
	HG(b)	4	111079337	111173378	2	94041	12	*ELOVL6*
		12	106077858	106177127	2	99269	10	-
		20	23677829	23725019	2	47190	8	
Lactose Intolerance	KEL	1	49580287	49672039	1	91752	8	*AGBL4*
		19	6904195	7103542	2	199347	40	*ZNF557, MBD3L2, MBD3L5, ADGRE4P, FLJ25758, MBD3L4, MBD3L3, ADGRE1*
	KIR	1	49580287	49672039	1	91752	8	*AGBL4*
		19	6904195	7103542	2	199347	40	*ZNF557, MBD3L2, MBD3L5, ADGRE4P, FLJ25758, MBD3L4, MBD3L3, ADGRE1*

Inclusion body myositis	AFF	1	231730121	231767890	2	37769	7	*DISC1*
		7	69966192	70022011	2	55819	8	*AUTS2*
		7	70137165	70295629	2	158464	19	*AUTS2*
		19	54731679	54740705	2	9026	5	*LILRB3*
	UNAFF	1	231711489	231803604	2	92115	11	*DISC1*
		7	69966192	70022011	2	55819	8	*AUTS2*
		7	70014509	70295629	2	281120	31	*AUTS2*
		19	54731679	54740705	2	9026	5	*LILRB3*
ADHD	DS	4	161953729	162002921	2	49192	6	-
	DV	-	-	-	-	-	-	-
	RP	4	139790626	139915883	2	125257	16	-
		4	15776181	15840839	2	64658	10	*CD38*
		4	161953729	162002921	2	49192	6	-
		13	64378676	64390470	2	11794	3	-
	OH	4	139790626	139915883	2	125257	9	-
		4	15776181	15840839	2	64658	10	*CD38*
		4	161953729	162002921	2	49192	6	-
		13	64346536	64390470	2	43934	5	-
Tourette's Syndrome	487	2	90108545	90109261	1	716	5	-
	488	2	90108545	90109261	1	716	5	-
		3	173259356	173289281	2	29925	4	*NLGN1*
		8	105981846	106021780	2	39934	4	
		8	106095659	106115255	2	19596	6	
		10	45218841	45359483	2	140642	14	
		12	8035139	8101326	2	66187	9	*SLC2A3, NANOGP1, NECAP1*
		22	22313954	22560977	2	247023	37	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
	490	2	89987044	90109261	1	122217	8	*IGKV2D-29, IGKV2D-28, IGKV2D-26, IGKV3D-20*
		3	173259356	173289281	2	29925	4	*NLGN1*
		5	113429984	113435957	1	5973	4	-
		8	105981846	106108593	2	126747	14	-
		10	45218841	45359483	2	140642	14	-
		22	22313954	22560977	2	247023	33	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
	489	2	89987044	90109261	1	122217	8	*IGKV2D-29, IGKV2D-28, IGKV2D-26, IGKV3D-20*
		3	173259356	173289281	2	29925	4	*NLGN1*

		5	113429984	113435957	1	5973	4	-
		8	105981846	106052343	2	70497	9	-
		10	45218841	45359483	2	140642	14	-
		22	22313954	22550078	2	236124	30	*IGLV4-69, IGLV4-60, IGLV8-61, IGLV6-57, TOP3B*
Parkinson's disease	PD821	3	173259356	173289281	2	29925	4	*NLGN1*
		17	34458934	34461869	2	2935	4	-
	PD161	3	173259356	173289281	2	29925	4	*NLGN1*
		17	34450463	34461869	2	11406	5	-
Dystonia/HSP	VF	1	248789519	248813267	1	23748	9	*OR2T11, OR2T35, OR2T27*
		2	40447587	40509154	1	61567	14	*SLC8A1*
		22	25669569	25875573	2	206004	7	*LRP5L*
	LF	1	248789519	248813267	1	23748	9	*OR2T11, OR2T35, OR2T27*
		2	40447587	40509154	1	61567	14	*SLC8A1*
		22	25669569	25905668	2	236099	10	*LRP5L*
Schizophrenia	RT1b	4	161861640	161924832	2	63192	6	-
		13	23548470	23586366	2	37896	18	-
		15	30950529	31088443	1	137914	9	*ARHGAP11B*
		15	32513176	32514341	2	1165	4	-
		22	49565404	49570587	2	5183	6	-
	RT1a	4	161861640	161924832	2	63192	6	-
		13	23548470	23586366	2	37896	18	-
		15	30950529	31088443	1	137914	9	*ARHGAP11B*
		15	32513176	32514341	2	1165	4	-
		22	49566426	49570587	2	4161	5	-
Schizophrenia	IP16	14	32204263	32563640	2	359377	40	*NUBPL, ARHGAP5*
		17	34450463	34461869	2	11406	5	-
		19	54731679	54844626	2	112947	17	*LILRA6, LILRB5, LILRB2, LILRA3, LILRA5*
	IP17	14	32164373	32563640	2	399267	41	*NUBPL, ARHGAP5*
		17	34450463	34461869	2	11406	5	-
		19	54749011	54844626	2	95615	12	*LILRA6, LILRB5, LILRB2, LILRA3, LILRA5*

Table 5.1. CNVs identified by PennCNV and cnvPartition where merged and manually screened in GenomeStudio. False negative CNV calls are included in the list. CN 1 = duplication, CN 2 = deletion. Putative de novo CNVs are highlighted in yellow.

Diagnosis of affected twin	Gene	Description	Disease association (DisGeNET)	Gene ontology (functioning of gene products)	RVIS	pLI	%HI
ALS	*LARGE1*	LARGE xylosyl- and glucuronyltransferase 1	Congenital muscular dystrophy; mental retardation syndromes;	GO:0005515 [protein binding] GO:0005794 [Golgi apparatus] GO:0006044 [N-acetylglucosamine metabolic process] GO:0006486 [protein glycosylation] GO:0006688 [glycosphingolipid biosynthetic process] GO:0008375 [acetylglucosaminyltransferase activity] GO:0009101 [glycoprotein biosynthetic process] GO:0015020 [glucuronosyltransferase activity] GO:0016757 [transferase activity, transferring glycosyl groups] GO:0030145 [manganese ion binding] GO:0030173 [integral component of Golgi membrane] GO:0035269 [protein O-linked mannosylation] GO:0042285 [xylosyltransferase activity] GO:0043231 [intracellular membrane-bounded organelle] GO:0043403 [skeletal muscle tissue regeneration] GO:0046716 [muscle cell cellular homeostasis] GO:0060538 [skeletal muscle organ development]	-1.11 (10.98%)	0.96	4.10
HSP	*SLC8A1*	Solute carrier family 8 member A1	Myocardial ischemia; status epilepticus;	GO:0002026 [regulation of the force of heart contraction] GO:0002027 [regulation of heart rate] GO:0005432 [calcium:sodium antiporter activity]	-1.41 (6.88%)	0.99	6.09

				GO:0005509 [calcium ion binding] GO:0005515 [protein binding] GO:0005516 [calmodulin binding] GO:0005829 [cytosol] GO:0005886 [plasma membrane] GO:0005887 [integral component of plasma membrane] GO:0006811 [ion transport] GO:0006816 [calcium ion transport] GO:0006883 [cellular sodium ion homeostasis] GO:0006936 [muscle contraction] GO:0007154 [cell communication] GO:0007596 [blood coagulation] GO:0008092 [cytoskeletal protein binding] GO:0010649 [regulation of cell communication by electrical coupling] GO:0010881 [regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion] GO:0010882 [regulation of cardiac muscle contraction by calcium ion signaling] GO:0014704 [intercalated disc] GO:0014829 [vascular smooth muscle contraction] GO:0016020 [membrane] GO:0016021 [integral component of membrane] GO:0030018 [Z disc] GO:0030315 [T-tubule] GO:0030501 [positive regulation of bone mineralization] GO:0030506 [ankyrin binding] GO:0034614 [cellular response to reactive oxygen			

				species] GO:0035725 [sodium ion transmembrane transport] GO:0035994 [response to muscle stretch] GO:0042383 [sarcolemma] GO:0044325 [ion channel binding] GO:0044557 [relaxation of smooth muscle] GO:0051481 [negative regulation of cytosolic calcium ion concentration] GO:0055013 [cardiac muscle cell development] GO:0055074 [calcium ion homeostasis] GO:0055085 [transmembrane transport] GO:0055119 [relaxation of cardiac muscle] GO:0060048 [cardiac muscle contraction] GO:0060401 [cytosolic calcium ion transport] GO:0060402 [calcium ion transport into cytosol] GO:0070509 [calcium ion import] GO:0070588 [calcium ion transmembrane transport] GO:0071313 [cellular response to caffeine] GO:0071436 [sodium ion export] GO:0086012 [membrane depolarization during cardiac muscle cell action potential] GO:0086064 [cell communication by electrical coupling involved in cardiac conduction] GO:0097369 [sodium ion import] GO:0098735 [positive regulation of the force of heart contraction] GO:1901660 [calcium ion export]			
Ischaemic stroke	*ELOVL6*	ELOVL fatty acid elongase 6	Heart failure; Psoriasis; obesity; atherosclerosis; atopic dermatitis	GO:0005515 [protein binding] GO:0005783 [endoplasmic reticulum] GO:0005789 [endoplasmic reticulum membrane]	-0.27 (36.97%)	0.97	3.68

				GO:0016021 [integral component of membrane] GO:0016747 [transferase activity, transferring acyl groups other than amino-acyl groups] GO:0019367 [fatty acid elongation, saturated fatty acid] GO:0019432 [triglyceride biosynthetic process] GO:0030176 [integral component of endoplasmic reticulum membrane] GO:0030497 [fatty acid elongation] GO:0035338 [long-chain fatty-acyl-CoA biosynthetic process] GO:0042759 [long-chain fatty acid biosynthetic process] GO:0044255 [cellular lipid metabolic process] GO:0044281 [small molecule metabolic process]			
Lactose intolerance	*MBD3L2*	Methyl-CpG binding domain protein 3 like 2	Stomach neoplasms	GO:0000122 [negative regulation of transcription from RNA polymerase II promoter] GO:0005634 [nucleus] GO:0006346 [methylation-dependent chromatin silencing] GO:0008327 [methyl-CpG binding]	n/a	n/a	95.74
	ADGRE4P	Adhesion G protein-coupled receptor E4, pseudogene	Colorectal cancer	n/a	n/a	n/a	n/a
	ADGRE1	Adhesion G protein-coupled receptor E1	Liver cirrhosis; secondary periodontitis	GO:0002250 [adaptive immune response] GO:0004888 [transmembrane signalling receptor activity] GO:0004930 [G-protein coupled receptor activity] GO:0005509 [calcium ion binding] GO:0005515 [protein binding]	2.21 (97.76%)	0.00	90.69

				GO:0005887 [integral component of plasma membrane] GO:0007155 [cell adhesion] GO:0007166 [cell surface receptor signalling pathway] GO:0007186 [G-protein coupled receptor signalling pathway] GO:0016020 [membrane] GO:0016021 [integral component of membrane]			
ADHD	*CD38*	CD38 molecule	Autism spectrum disorders	GO:0001666 [response to hypoxia] GO:0003953 [NAD+ nucleosidase activity] GO:0005634 [nucleus] GO:0005829 [cytosol] GO:0005886 [plasma membrane] GO:0007165 [signal transduction] GO:0007204 [positive regulation of cytosolic calcium ion concentration] GO:0007565 [female pregnancy] GO:0008152 [metabolic process] GO:0009725 [response to hormone] GO:0009986 [cell surface] GO:0016020 [membrane] GO:0016021 [integral component of membrane] GO:0016740 [transferase activity] GO:0016798 [hydrolase activity, acting on glycosyl bonds] GO:0016849 [phosphorus-oxygen lyase activity] GO:0030307 [positive regulation of cell growth] GO:0030890 [positive regulation of B cell proliferation] GO:0032024 [positive regulation of insulin secretion]	0.47 (70.44%)	0.00	77.57

				GO:0032355 [response to estradiol] GO:0032526 [response to retinoic acid] GO:0032570 [response to progesterone] GO:0033194 [response to hydroperoxide] GO:0034097 [response to cytokine] GO:0042493 [response to drug] GO:0043066 [negative regulation of apoptotic process] GO:0043231 [intracellular membrane-bounded organelle] GO:0045779 [negative regulation of bone resorption] GO:0045892 [negative regulation of transcription, DNA-templated] GO:0045893 [positive regulation of transcription, DNA-templated] GO:0045907 [positive regulation of vasoconstriction] GO:0050135 [NAD(P)+ nucleosidase activity] GO:0050853 [B cell receptor signaling pathway] GO:0060292 [long term synaptic depression] GO:0070062 [extracellular exosome] GO:0070555 [response to interleukin-1] GO:0097190 [apoptotic signaling pathway]			
Tourette's syndrome	*TOP3B*	DNA topoisomerase III beta	Cognitive impairment; schizophrenia	GO:0003677 [DNA binding] GO:0003916 [DNA topoisomerase activity] GO:0003917 [DNA topoisomerase type I activity] GO:0005515 [protein binding] GO:0005634 [nucleus] GO:0005654 [nucleoplasm] GO:0006265 [DNA topological change] GO:0044822 [poly(A) RNA binding]	-0.34 (33.91%)	0.11	15.86

Tourette's syndrome/Parkinson's disease	*NLGN1*	Neuroligin 1	Autistic disorder; schizophrenia; bipolar disorder; depression; memory impairment	GO:0002087 [regulation of respiratory gaseous exchange by neurological system process] GO:0004872 [receptor activity] GO:0005515 [protein binding] GO:0005794 [Golgi apparatus] GO:0005886 [plasma membrane] GO:0005887 [integral component of plasma membrane] GO:0006605 [protein targeting] GO:0007157 [heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules] GO:0007158 [neuron cell-cell adhesion] GO:0007399 [nervous system development] GO:0007416 [synapse assembly] GO:0009897 [external side of plasma membrane] GO:0009986 [cell surface] GO:0010841 [positive regulation of circadian sleep/wake cycle, wakefulness] GO:0014069 [postsynaptic density] GO:0016080 [synaptic vesicle targeting] GO:0016339 [calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules] GO:0017146 [NMDA selective glutamate receptor complex] GO:0023041 [neuronal signal transduction] GO:0030054 [cell junction] GO:0030165 [PDZ domain binding] GO:0030425 [dendrite] GO:0031175 [neuron projection development] GO:0032230 [positive regulation of synaptic transmission, GABAergic]	-1.26 (8.92%)	0.76	1.37

				GO:0032433 [filopodium tip] GO:0035418 [protein localization to synapse] GO:0042043 [neurexin family protein binding] GO:0043197 [dendritic spine] GO:0043198 [dendritic shaft] GO:0045184 [establishment of protein localization] GO:0045202 [synapse] GO:0045211 [postsynaptic membrane] GO:0045664 [regulation of neuron differentiation] GO:0046983 [protein dimerization activity] GO:0048488 [synaptic vesicle endocytosis] GO:0048489 [synaptic vesicle transport] GO:0048511 [rhythmic process] GO:0048789 [cytoskeletal matrix organization at active zone] GO:0050804 [modulation of synaptic transmission] GO:0050808 [synapse organization] GO:0050839 [cell adhesion molecule binding] GO:0051260 [protein homooligomerization] GO:0051290 [protein heterotetramerization] GO:0051491 [positive regulation of filopodium assembly] GO:0051965 [positive regulation of synapse assembly] GO:0051968 [positive regulation of synaptic transmission, glutamatergic] GO:0060076 [excitatory synapse] GO:0060291 [long-term synaptic potentiation] GO:0060999 [positive regulation of dendritic			

				spine development] GO:0061002 [negative regulation of dendritic spine morphogenesis] GO:0072553 [terminal button organization] GO:0097091 [synaptic vesicle clustering] GO:0097104 [postsynaptic membrane assembly] GO:0097105 [presynaptic membrane assembly] GO:0097110 [scaffold protein binding] GO:0097113 [AMPA glutamate receptor clustering] GO:0097114 [NMDA glutamate receptor clustering] GO:0097115 [neurexin clustering involved in presynaptic membrane assembly] GO:0097119 [postsynaptic density protein 95 clustering] GO:0097120 [receptor localization to synapse] GO:0097481 [neuronal postsynaptic density] GO:1900029 [positive regulation of ruffle assembly] GO:1900244 [positive regulation of synaptic vesicle endocytosis] GO:1902474 [positive regulation of protein localization to synapse] GO:1902533 [positive regulation of intracellular signal transduction] GO:2000302 [positive regulation of synaptic vesicle exocytosis] GO:2000310 [regulation of N-methyl-D-aspartate selective glutamate receptor activity] GO:2000311 [regulation of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate			

				selective glutamate receptor activity] GO:2000463 [positive regulation of excitatory postsynaptic potential] GO:2000809 [positive regulation of synaptic vesicle clustering]			
Inclusion body myositis	*AUTS2*	AUTS2, activator of transcription and developmental regulator	Neurological disorders; acute lymphoblastic leukaemia; aging of the skin; early-onset androgenetic alopecia; cancer	GO:0003674 [molecular function] GO:0003682 [chromatin binding] GO:0005515 [protein binding] GO:0005575 [cellular component] GO:0005634 [nucleus] GO:0008150 [biological process] GO:0045944 [positive regulation of transcription from RNA polymerase II promoter] GO:0051571 [positive regulation of histone H3-K4 methylation] GO:0060013 [righting reflex] GO:0098582 [innate vocalization behaviour] GO:2000620 [positive regulation of histone H4-K16 acetylation]	-1.94 (3.32%)	1.00	0.51
	LILRB3	Leukocyte immunoglobulin like receptor B3	Lung diseases; Takayasu's arteritis	GO:0002250 [adaptive immune response] GO:0004872 [receptor activity] GO:0004888 [transmembrane signalling receptor activity] GO:0005515 [protein binding] GO:0005887 [integral component of plasma membrane] GO:0006952 [defence response] GO:0007166 [cell surface receptor signalling pathway] GO:0045671 [negative regulation of osteoclast differentiation]	0.79 (81.49%)	0.00	97.17

Schizophrenia	*ARHGAP11B*	Rho GTPase activating protein 11B	15q13.3 microdeletion syndrome	GO:0005829 [cytosol] GO:0007165 [signal transduction] GO:0007264 [small GTPase mediated signal transduction] GO:0021987 [cerebral cortex development] GO:0051056 [regulation of small GTPase mediated signal transduction]	0.07 (52.75%)	0.07	81.20
	ARHGAP5	Rho GTPase activating protein 5	Liver carcinoma; breast cancer	GO:0003924 [GTPase activity] GO:0005096 [GTPase activator activity] GO:0005515 [protein binding] GO:0005525 [GTP binding] GO:0005737 [cytoplasm] GO:0005829 [cytosol] GO:0007155 [cell adhesion] GO:0007165 [signal transduction] GO:0007264 [small GTPase mediated signal transduction] GO:0007266 [Rho protein signal transduction] GO:0016020 [membrane] GO:0042169 [SH2 domain binding] GO:0043547 [positive regulation of GTPase activity] GO:0051056 [regulation of small GTPase mediated signal transduction]	-1.31 (8.10%)	0.99	8.15

Table 5.2. Selected genes within CNV regions of interest. Genes that are in the RefSeq database (http://www.ncib.nlm.nih.gov/gene) and Ensembl database (http://www.ensembl.org) are reported. Gene ontology data from the Gene Ontology Project (http://www.geneontology.org). RVIS v4 is constructed on the ExAC v2 data release (Petrovski et al., 2013).

5.2.1.1 Inclusion body myositis (AFF and UNAFF)

The twin pair discordant for inclusion body myositis yielded four CNVs. This included a 92115bp CNV duplication containing *DISC1* (chr1:231711489-231803604), located in the region 1q42.2; two CNV duplications overlapping *AUTS2*, which is likely a larger CNV falsely called as two (chr7:69966192-70295629); and finally, a CNV duplication (chr19:54731679-54740705) in the region 19q13.42 overlapping *LILRB3*. An important paralog of this gene is *LILRB1*, which has been associated with idiopathic inflammatory myopathies (Schleinitz et al., 2008). Although *DISC1* and *AUTS2* are known to be strongly implicated in schizophrenia (Ayalew et al., 2012) and autism (Oksenberg and Ahituv, 2013) respectively, neither twin presented with psychiatric symptoms to our knowledge. CNVs overlapping *AUTS2* and *LILRB3* were validated with ExomeDepth.

5.2.1.2 Attention deficit hyperactivity disorder (RP and OH)

Three putative de novo CNVs were detected in both twins, but were absent in the parents (Table 5.1). The CNV duplication on chromosome 4p15.32 covers >85% proximal of *CD38*, a gene encoding for a multifunctional ectonucleotidase involved in signal transduction, cell adhesion and calcium signalling. Interestingly, *CD38* has previously been implicated in ADHD (Ebstein et al., 2014), including social memory, amnesia and autism spectrum disorder (Higashida et al., 2012). This gene was called as discordant, however manual reviewing revealed that it was falsely undercalled in the unaffected twin. These CNVs were not analysed by ExomeDepth or experimentally validated.

5.2.1.3 Tourette's syndrome (489 and 490)

The twins discordant for Tourette's syndrome inherited four CNVs from the mother. An apparent CNV deletion (CN=1) in the mother and father (chr2:90108545-90109261) was also present in the twins, but was expanded by 121,501bp. This resulted in a loss of immunoglobulin kappa variable genes in the twins. CNVs have been found to undergo modification in size when transmitted from parent to offspring (South et al., 2008). However, this CNV should be at best regarded as tentative, as it is in close proximity to the centromere. Various CNV analysis protocols recommend removing called CNVs in HLA regions, and genomic regions that are near centromeres and telomeres. Conversely, enrichments of germline CNVs near assembly gaps and in regions of low-mappability have shown to be reliable (Monlong et al., 2015), as they can be the result of reduced selection pressure, rather than faults of the detection tools.

The multiple CNVs in the twins include maternally-inherited duplications of *NLGN1* and *TOP3B.* A CNV duplication on chromosome 22q11.22 (22313954-22550078) overlapped >90% of *TOP3B,* a gene that has been implicated in neurodevelopmental disorders (Stoll et al., 2013). Further, a CNV duplication on chromosome 3q26.31 contained *NLGN1,* a gene involved in forming excitatory synapses and maintaining synaptic plasticity (Hoy et al., 2013).

The affected twin was diagnosed with Tourette's syndrome at the age of seven. Interestingly, the father was also diagnosed with Tourette's syndrome at the same age, in addition to OCD and ADHD. The mother, however, was reported to be asymptomatic. Moreover, both twins have a putative de novo CNV deletion in a non-genic region (chromosome 5q22.3; location 113429984-113435957). These CNVs were not computationally or experimentally validated.

5.2.1.4 Ischaemic stroke (KG and HG)

DNA from this twin pair was extracted from blood ($HG_{(b)}$ and $KG_{(b)}$) and saliva ($HG_{(s)}$ and $KG_{(s)}$). A CNV duplication on chromosome 4q25 was found in both twins, containing *ELOVL6*. No intra-tissue CNV differences were detected in the SNP array analysis and ExomeDepth analysis confirmed this CNV. *ELOVL6* plays an important role in fat metabolism and insulin sensitivity (Matsuzaka and Shimano, 2009), and has been associated with heart failure, obesity, atherosclerosis, psoriasis, and atopic dermatitis (Uchida, 2011). These twins have a family history of hypertension, depression and psoriasis, and where diagnosed with atopic dermatitis (HG) and seborrheic dermatitis capitis (KG). Chapter 6 contains a detailed clinical history of these twins.

5.2.1.5 Lactose intolerance (KIR and KEL)

A large CNV duplication spanning 40 probes was detected on chromosome 19p13.2 (6904195-7103542) in both twins. Interestingly, the CNV spans genes that are related to gastrointestinal disorders, including *MBD3L2* (stomach neoplasms), *ADGRE4P* (colorectal cancer metastatic), and *ADGRE1* (liver cirrhosis). Lactose intolerance has shared genes with colorectal cancer (*LCT, CASR, GLB1*), and liver cirrhosis (*CASR and GLB1*) (Andrzej et al., 2015). This CNV was validated with ExomeDepth analysis.

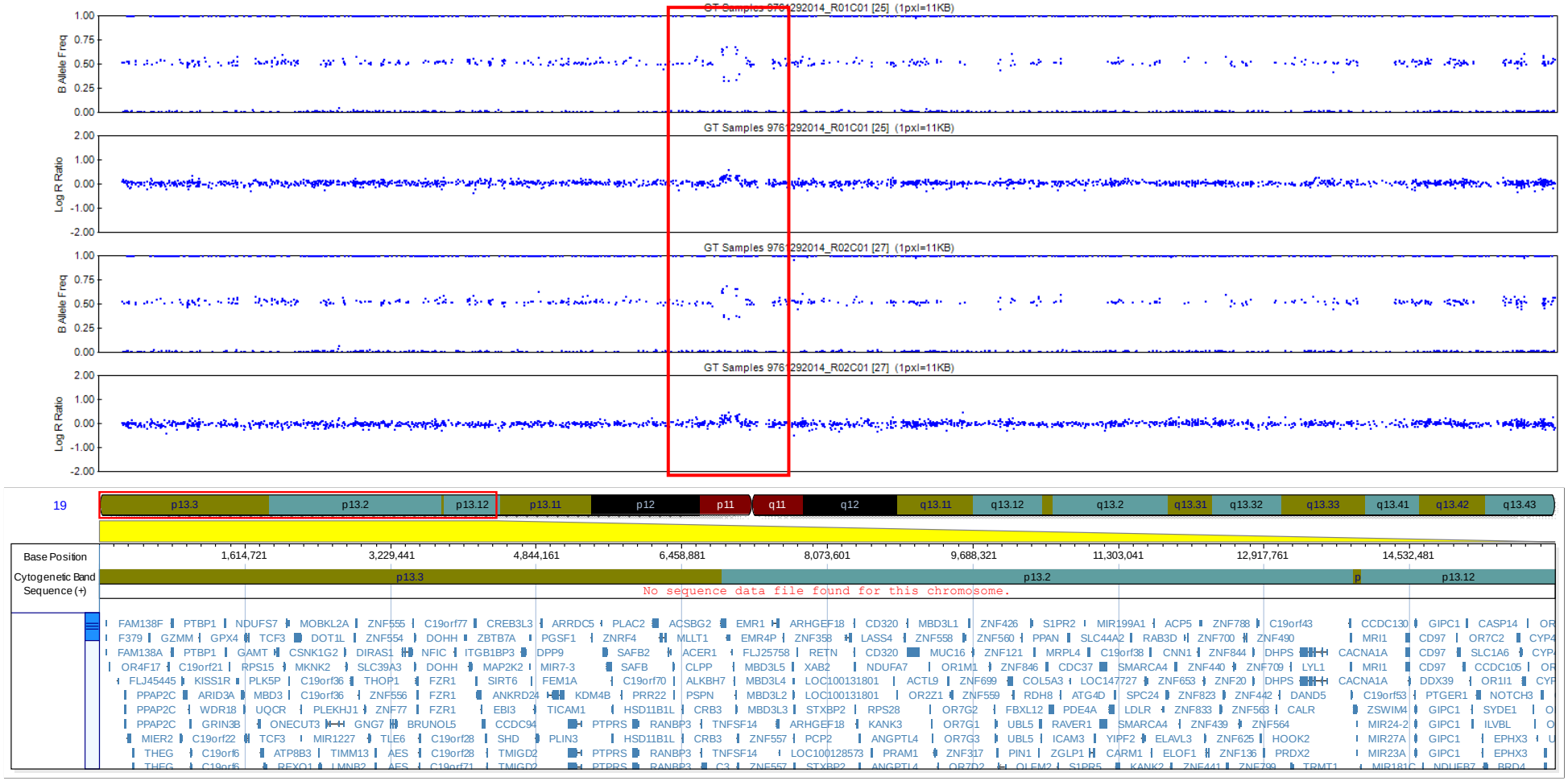


Figure 5.1. Results of chromosomal microarray analysis in KEL and KIR. A duplication (CN = 3) is depicted by the BAF plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The red rectangle contains the identical 199 kb duplicated genomic region on chromosome 19p13.2.

5.2.1.6 Hereditary spastic paraplegia (VF and LF)

VF was diagnosed with predominant HSP superimposed with dystonia. Her identical twin sister (LF) and sixteen-year-old son later presented with similar, but milder, symptoms. Further clinical details of VF can be found in section 4.2.1.1. A shared hemizygous deletion (CN = 1) of ~60 kb on chromosome 2p22.1 was found in both twins, containing *SLC8A1* (see Figure 5.1). This gene was not confirmed with ExomeDepth. This could be due to the limitations of exome sequencing CNV detection tools in being able to detect smaller CNVs that range between 1-4 exons (Yao et al., 2017). Unfortunately, DNA of the proband's son, who was also affected, was not available for segregation analysis.

Figure 5.2. Results of chromosomal microarray analysis in VF and LF. A hemizygous deletion (CN = 1) is depicted as a loss of heterozygotes in the BAF plot and loss of signal intensity in the LRR plot. The red rectangle contains the identical 62 kb deleted genomic region on chromosome 2p22.1.

5.2.1.7 Schizophrenia (IP16 and IP17)

The largest CNV identified by SNP array analysis was a 399 kb duplication on chromosome 14q12 (32164373-32563640) overlapping genes *NUBPL* and *ARHGAP5*. This was confirmed with exome sequencing CNV calling using ExomeDepth. According to the Database of ClinGen Dosage Sensitivity Map, the haploinsufficiency score of 8.15% (high rank = 0-10%) and pLI score of 0.99 suggests that duplication of *ARHGAP5* may be pathogenic. *ARHGAP5* also appears to be intolerant of variation with a genic intolerance score of -1.31, while being among the top 8% of most intolerant genes.

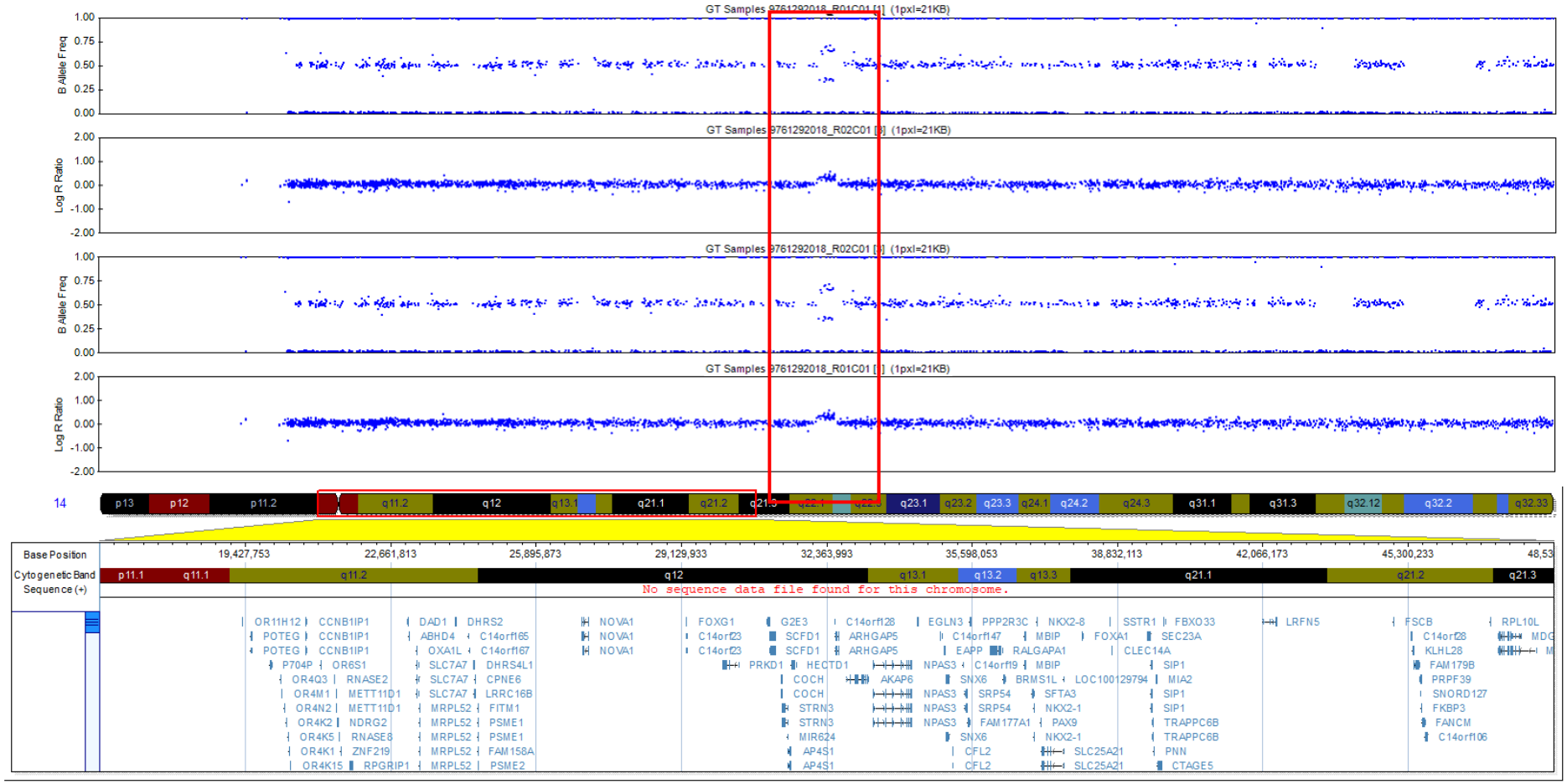


Figure 5.3. Results of chromosomal microarray analysis in IP16 and IP17. A duplication (CN = 3) is depicted by the BAF plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The red rectangle contains the identical 359 kb duplicated genomic region on chromosome 14q12.

5.2.1.8 Schizophrenia (RT1a and RT1b)

Both SNP array and exome sequencing CNV calling revealed a 138 kb hemizygous deletion in 15q13.2 (chr15:30950529-31088443) (Figure 5.1) covering *ARHGAP11B* in both twins. Due to lack of probe coverage in this region, it was difficult to ascertain an accurate size of the deletion. The CNV could potentially have extended to left-flanking neighbouring genes, including *CHRFAM7A.*

ARHGAP11B was of particular interest as gene ontology terms include cerebral cortex development. It promotes development and evolutionary expansion of the brain neocortex and it is able to promote amplification of basal progenitors in the subventricular zone, producing more neurons during foetal corticogenesis. CNV deletions containing *ARHGAP11B* has previously been associated with schizophrenia (Levinson et al., 2011).

The segment overlaps with CNVs recorded in the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). There are various syndromes, including schizophrenia, associated with this gene in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensemble Resources (DECIPHER), but the region deleted in this case is unique in that it encompasses the genes that were not associated with schizophrenia on their own. ClinVar has 93 records that mention *ARHGAP11B*; 38 of which are deletions, and of these 36 are considered pathogenic. No variants are confined to only *ARHGAP11B.*

Figure 5.4. Results of chromosomal microarray analysis in RT1a and RT1b. A hemizygous deletion (CN = 1) is depicted as a loss of heterozygotes in the BAF plot and loss of signal intensity in the LRR plot. The red rectangle contains the identical 138 kb deleted genomic region on chromosome 15q13.2. The full extent of the deletion cannot be determined due to the absent probes in the left flanking region.

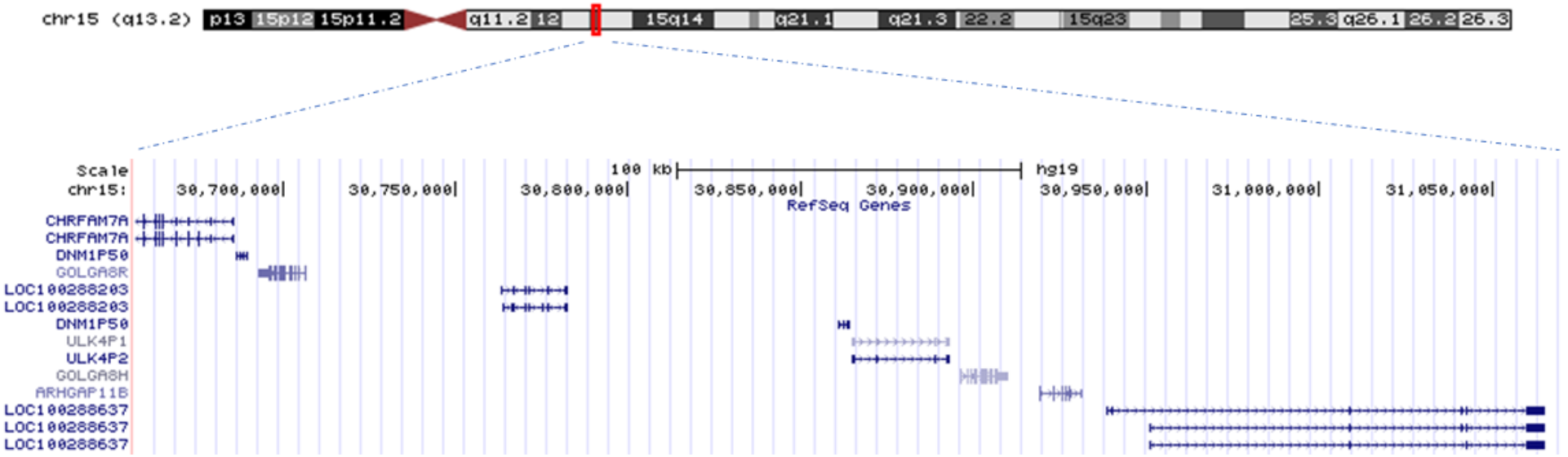


Figure 5.5. Visual representation of protein-coding RefSeq genes that could potentially be included in the CN deletion detected by the SNP array. Screen capture of the deleted region in 15q13.2 from UCSC Genome Browser GRCh37/hg19.

5.2.2 CNV validation with ddPCR

ddPCR was performed to validate the 138 kb CNV deletion spanning *ARHGAP11B* in the twins with schizophrenia (RT1a and RT1b). This technique uses an emulsion PCR approach to create many thousands of microvolume PCR reactions, with droplets counted at endpoint for presence or absence of the test and reference PCR products (Ottolini et al., 2014). Although the microarray and exome sequencing analyses confirmed the CNV in these twins, ddPCR results were inconclusive (Table 5.3). Due to the relatively low amounts of DNA required for the ddPCR protocol (~10ng), and the highly complex nature of this region, experiments were carried out on all twin pairs in our cohort.

	Gated by NTC			Gated by Positive Cluster			Copy number change
Sample ID	**Well 1**	**Well 2**	**Avg. CN**	**Well 1**	**Well 2**	**Avg. CN**	
NTC	No Amp.	No Amp.	No Amp.	No Amp.	No Amp.	No Amp.	**N/A**
NA12878	2.35	2.80	2.58	1.81	2.22	2.02	**Gain**
NA10851	1.94	2.09	2.02	1.00	1.10	1.05	**Normal**
NA11892	1.96	2.14	2.05	0.95	0.96	0.96	**Normal**
NA11894	2.17	2.12	2.15	1.00	1.00	1.00	**Normal**
1P16	1.43	1.40	1.42	0.01	0.04	0.03	**Loss**
1P17	1.70	1.52	1.61	0.07	0.08	0.07	**Loss**
218	1.96	2.06	2.01	0.95	0.97	0.96	**Normal**
318	2.08	1.97	2.03	0.98	0.93	0.95	**Normal**
242	1.52	1.35	1.44	0.04	0.05	0.05	**Loss**
243	1.54	1.46	1.50	0.04	0.06	0.05	**Loss**
421	2.13	2.14	2.14	0.94	1.02	0.98	**Normal**
422	2.18	1.95	2.07	1.11	0.98	1.04	**Normal**
487	2.65	2.78	2.72	1.93	2.09	2.01	**Gain**
488	2.37	1.92	2.15	1.09	0.91	1.00	**Normal**
489	2.59	3.03	2.81	1.92	2.33	2.13	**Gain**
490	2.82	2.67	2.75	2.12	1.99	2.06	**Gain**
AFF	2.24	2.14	2.19	1.04	1.12	1.08	**Normal**
UNAFF	1.97	2.22	2.10	0.86	1.00	0.93	**Normal**

DS	2.13	2.30	2.22	0.97	1.11	1.04	**Normal**
DV	2.28	2.27	2.28	1.13	1.06	1.10	**Normal**
RP	2.15	2.24	2.20	0.99	1.13	1.06	**Normal**
OH	2.11	2.09	2.10	1.07	1.10	1.09	**Normal**
HG(b)	2.29	2.49	2.39	1.03	1.29	1.16	**Normal**
HG(s)	2.30	2.08	2.19	1.03	1.04	1.04	**Normal**
KG(b)	2.12	2.17	2.15	1.10	1.00	1.05	**Normal**
KG(s)	2.06	2.21	2.14	0.98	1.14	1.06	**Normal**
KEL	1.58	1.56	1.57	0.09	0.10	0.09	**Loss**
KIR	1.88	1.59	1.74	0.10	0.06	0.08	**Loss**
PD161	2.10	2.09	2.10	1.06	1.00	1.03	**Normal**
PD821	2.86	2.42	2.64	1.33	1.21	1.27	**Gain**
RT1a	1.62	1.83	1.73	1.07	1.23	1.15	**Normal or loss**
RT1b	1.76	1.76	1.76	1.08	1.12	1.10	**Normal or loss**
LF	2.43	2.25	2.34	1.20	1.09	1.15	**Gain**
VF	2.03	2.07	2.05	0.99	0.93	0.96	**Normal**
SUS	1.61	1.50	1.56	0.07	0.03	0.05	**Loss**

Table 5. 3. Summary of ddPCR results. Green and red colours indicate a gain and loss in CN, respectively. The yellow highlight shows the results for twins with schizophrenia, in whom the CNV deletion was detected by exome sequencing and SNP array CNV calling. One twin (LAS) was not processed due to limited DNA availability. Samples NA12878, NA10851, NA11892, NA11894 are individual controls from the 1000 Human Genome Project. It was assumed that they will have a CN of 2. NTC = No template control.

There are several reasons why the ddPCR results might be unreliable. The frequency of the *ARHGAP11B* deletion in our twin cohort is incongruent to the frequencies that have been observed in other studies with larger sample sizes. For instance, loss of *ARHGAP11B* was detected in 8 of 1257 patients with autism spectrum disorder (0.64%) and in 4 of 1577 controls (0.25%) (Leblond et al., 2012). The fact that ddPCR detected a deletion in 4 of 13 twin pairs in our cohort (31%), suggests that the probes used may not be specific. The twins discordant for Parkinson's disease, moreover, were discordant for the CN, and the microarray and exome sequencing CNV calling did not detect this deletion in any of the twin pairs except RT1a and RT1b.

Nevertheless, it is interesting to note that the apparent deletion was detected in the other twins discordant for schizophrenia (IP16 and IP17), And in RT1a and RT1b, the results were borderline when gated by no template control (mean values 1.73 and 1.76, respectively).

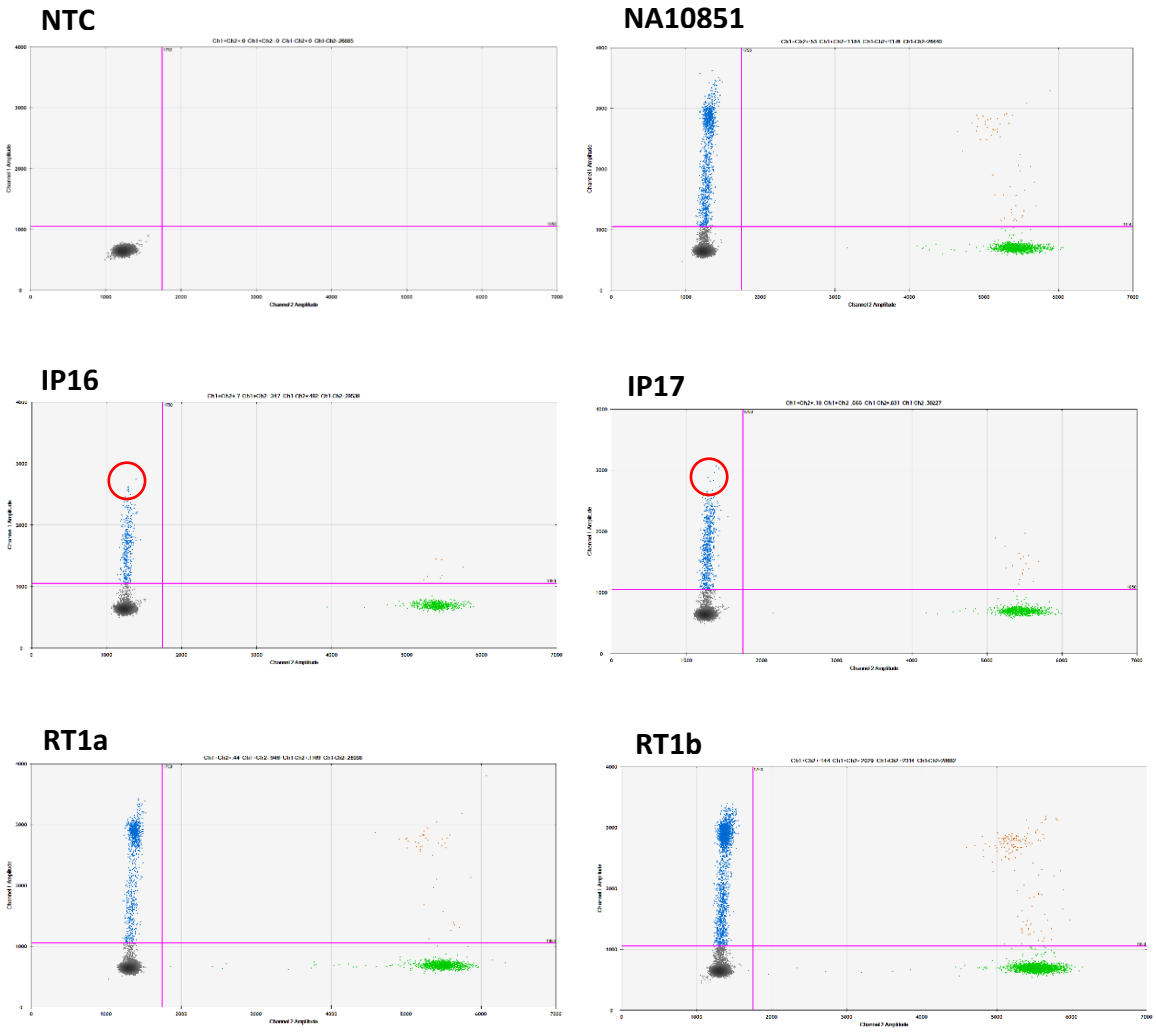


Figure 5.6. 2-D fluorescence amplitude plot shows duplicate wells of each twin sample with the *ARHGAP11B* assay. The black cluster on the plot represents the negative droplets, the blue cluster represents the droplets that are positive for *ARHGAP11B*. Red circle indicates a deletion.

5.3 Discussion

CNV analysis was carried out to investigate potential structural alterations to support the exome sequencing data. After CNV merging and manual reviewing of CNV calls, all CNVs between co-twins were found to be concordant. Although these results argue against a role of large structural rearrangements as a molecular aetiology of the observed clinical differences between the MZ twins, it is interesting to note that some of the genes within concordant CNVs have previously been implicated in the corresponding disorder of the affected twin.

5.3.1 Clinical significance of detected CNVs

5.3.1.1 Tourette's syndrome: What makes you tic?

The CNV duplication overlapping *TOP3B* (chr22:22313954-22560977) that was inherited from the mother may have contributed to an existing predisposition inherited from the father, and thus contributed to disease manifestation in the affected twin. This 'two-hit' hypothesis of the development of neuropsychiatric disorders may explain several observations in this case (Maynard et al., 2001).

It is likely that both twins (489 and 490) have a predisposition to develop Tourette's syndrome from the father's side. An inherited stop loss mutation (c.T1198C:p.X400Q) in *AADAC*, a candidate gene for Tourette's syndrome (Bertelsen et al., 2016), was identified in the previous chapter. It is possible that the CNV duplication overlapping *TOP3B* and *NLGN1* adds to this predisposition, and thus has a role to play in the manifestation of the disorder. *TOP3B* has previously been implicated in neurodevelopmental disorders (Stoll et al., 2013), and CNVs containing *NLGN1* have previously been identified in individuals

with OCD, Tourette's syndrome and autism (Pinto et al., 2014; Marshall et al., 2008; Gazzellone et al., 2016).

The unaffected twin (489) remained asymptomatic at the time of DNA sample collection. It is not uncommon, however, for the unaffected twin in a discordant pair to have a delayed or mild manifestation of the disease in question. Also, where there's a familial involvement, the second hit may not be needed depending on the nature of the predisposition. Environmental influences can also add to the familial predisposition and play a role in disease manifestation.

5.3.1.2 Inclusion body myositis

Previous studies strongly suggest that there is an active interaction between immune and muscle cells in the muscle microenvironment in myositis (Zheng et al., 2012). Depending on the cell type, this results in either stimulation or inhibition of immune cells. MHC class I is highly expressed in myositis muscle fibres, and the receptors that interact with MHC class I in myositis are undefined (Hirayasu and Arase, 2015). An important paralog of *LILRB3*, leukocyte immunoglobulin-like receptor 1, has been shown to be highly expressed in inflammatory cells in muscle, suggesting active communication between immune cells, such as antigen-presenting cells, and skeletal muscle (Hirayasu and Arase, 2015). The duplication of *LILRB3* in these twins could potentially be significant, although the pathophysiology of myositis and the role of inflammation in muscle fibre damage remain poorly understood at the molecular level.

5.3.1.3 Hereditary spastic paraplegia

SLC8A1 contributes to Ca^{2+} transport during excitation-contraction coupling in muscle, and is required for normal embryonic heart development and the onset of heart contractions (Shimizu et al., 2016). With a pLI score of 0.99 and HIS of 6.09%, *SLC8A1* is very intolerant to variation. Based on the allele frequencies from the NHLBI-ESP6500 data set, this gene has a RVIS of -1.41, and is among the 7% most intolerant of human genes (Petrovski et al., 2013; http://genic-intolerance.org). However, this gene has not been linked to dystonia, HSP or other movement disorders.

5.3.1.4 Ischaemic stroke: The clot thickens

ELOVL6 is regulated by a variety of methods by the cells, determined though expression and direct inhibition. Saito et al. (2011) note that expression of *Elovl6* is directly regulated by sterol regulatory element-binding protein-1a, -1c, and -2, including other lipogenic enzymes such as fatty acid synthase and stearoyl coenzyme A desaturase. Suppression of *Elovl6* expressioncan be brought about by dietary n-3 polyunsaturated fatty acids, such as eicosapentaenoate and docosahexaenoate acid (Kumadaki et al., 2008).

ELOVL6 is a rate-limiting long-chain fatty acid elongase and specifically catalyses the elongation of saturated and monounsaturated fatty acids, providing important constituents of triglycerides, esterified cholesterol, and phospholipids (Saito et al., 2011). It has been proposed that inhibition of *ELOVL6* could be a potential therapeutic target for a variety of cardiovascular diseases, such as insulin resistance, diabetes, and atherosclerosis (Saito et al., 2011). It is therefore possible that the duplication of this gene has led to an increase in the levels of this enzyme and this has contributed to the increased stoke risk experience by KG.

5.3.1.5 Schizophrenia: Mind the GAP.

A deletion in *ARHGAP11B* (CN = 1) and a duplication in *ARHGAP5* (CN = 3) were identified in two twin pairs discordant for schizophrenia. *ARHGAP* gene products belong to the Rho family of GTP-binding proteins, which are involved in membrane/cytoskeletal reorganisation events. There are approximately 80 distinct RhoGAP domain-containing proteins that are encoded in human DNA.

These proteins cycle between active GTP-bound and inactive GDP-bound forms. Activation to the GTP-bound state is mediated by specific guanine nucleotide dissociation stimulators, whereas acceleration of GTP hydrolysis can be accomplished by GTPase-activating proteins (GAPs). GTPases are important regulators of signalling pathways that link growth factors and/or their receptors to adhesions and associated structures (Potkin et al., 2008), and consequently influence the shape and migration of cells.

Ligand binding to the epidermal growth factor (EGF) receptor initiates numerous signalling pathways; one of these is mediated by Ras, which leads to enhanced cell proliferation. EGF receptor signalling can also induce cell motility, mitosis, differentiation and protein secretion (Wells, 1999). The EGF receptor family of receptors are localised on subventricular neural progenitor cells in foetal and adult brains, and these progenitors give rise to forebrain neurons in development and after injury in the adult (Fallon et al., 2000).

The *ARHGAP5* gene product (a GTPase-activating protein for Rho family members) is linked to Ras, and thus to EGF receptor-mediated proliferation, migration and differentiation of forebrain progenitors. Therefore, a *ARHGAP5* duplication in an MZ twin pair discordant for schizophrenia might point to an aetiological basis, because schizophrenia has been linked to altered prenatal neurogenesis of cortical neurons

(Akbarian, 1993). In addition, *ARHGAP5* and *ARHGAP11B* are contained within regions 14q12 and 15q13.2, respectively, which have previously been associated with schizophrenia (Lavedan et al., 2008: Chen et al., 2016).

Of special mention is the gene *ARHGAP11B*, which resides on chromosome 15q13.2, one of the most complex and unstable loci in the human genome. Several neurodevelopmental disorders have been linked to structural variants in this and nearby regions (Antonacci et al., 2014; El-Hattab et al., 2009; Shinawi et al., 2009). *ARHGAP11B* arose from partial duplication of *ARHGAP11A* in the human lineage, approximately one million years after divergence from chimpanzees, but before divergence from Neanderthals (Florio et al., 2015). This led to the formation of large and complex human-specific segmental duplications, mediating recurrent rearrangements contributing to 15q13.3 microdeletion syndrome associated with intellectual disability, epilepsy and schizophrenia (Dennis and Eichler, 2016). Remarkably, all events were related to the chromosome 15 core duplicon containing *GOLGA*, suggesting that these sequences have a fundamental role in the cycles of chromosomal rearrangement and segmental duplication expansions (Antonacci et al., 2014). *GOLGA* sequences might possess favoured sites for microhomology-mediated break-induced replication, mechanisms which may induce segmental duplication formations (Payen et al., 2008).

ARHGAP11B is, to date, the only human-specific gene shown to promote basal progenitor generation and proliferation, including cortical plate augmentation and gyrification induction. However, unlike *ARHGAP11A*, which encodes a RhoGAP, *ARHGAP11B* does not exhibit RhoGAP activity. This is owed to a single C>G base substitution in ancestral *ARHGAP11B* producing a new splice donor site in the modern version of the gene, resulting in GAP domain truncation with the addition of a human-specific C-terminal tail.

Florio et al. (2016) propose that this small genomic alteration led to increased basal progenitors during neocortex development, and plays an important role in the evolutionary expansion of the human neocortex.

Several factors could account for why the CNV identified by SNP array and exome sequencing analysis was not validated with ddPCR. The highly complex segmental-duplication architecture surrounding this region could potentially result in many sites of rearrangement and these conformations may vary among individuals. This provides the potential for alternate sites of rearrangement. This has been observed for the Angelman and Prader-Willi syndromes, for which uncharacteristic breakpoints have been mapped to other duplication structures (Locke et al. 2004). Other possibilities include genomic misassembly of these regions and incorrect mapping of the ddPCR probes used for CNV validation. Finally, if nonallelic homologous recombination is the underlying mechanism of rearrangement in this region, then additional repair or recombination events could result in more complex rearrangements. Nonetheless, this study shows that segmental duplications play an important role in normal variation as well as in genomic disease, defining hotspots of rearrangement that are susceptible to variation among the normal population. We propose that regions 15q13.2-13.3 are exceptional candidate sites that may be associated with neuropsychiatric disorders, and this preliminary study provides the necessary baseline to begin future studies on disease populations.

The duplicated 8 exons of *ARHGAP11A* is almost identical to the paralogous sequence of *ARHGAP11B*, and thus is not completely queried in high throughput genetic studies. Indeed, variations in this region have flown below the radar of available genome-wide technologies, which likely has downplayed its hypothesised associations with neurodevelopmental disorders. Because of the genomic complexity of the region, the

extent of human structural diversity and breakpoints of most rearrangement events are poorly understood at the molecular genetic level. Moreover, the wide expression of *ARHGAP11B*, its multiple functions and modes of regulation – not to mention its absence in non-human animals – present challenges for its study in disease.

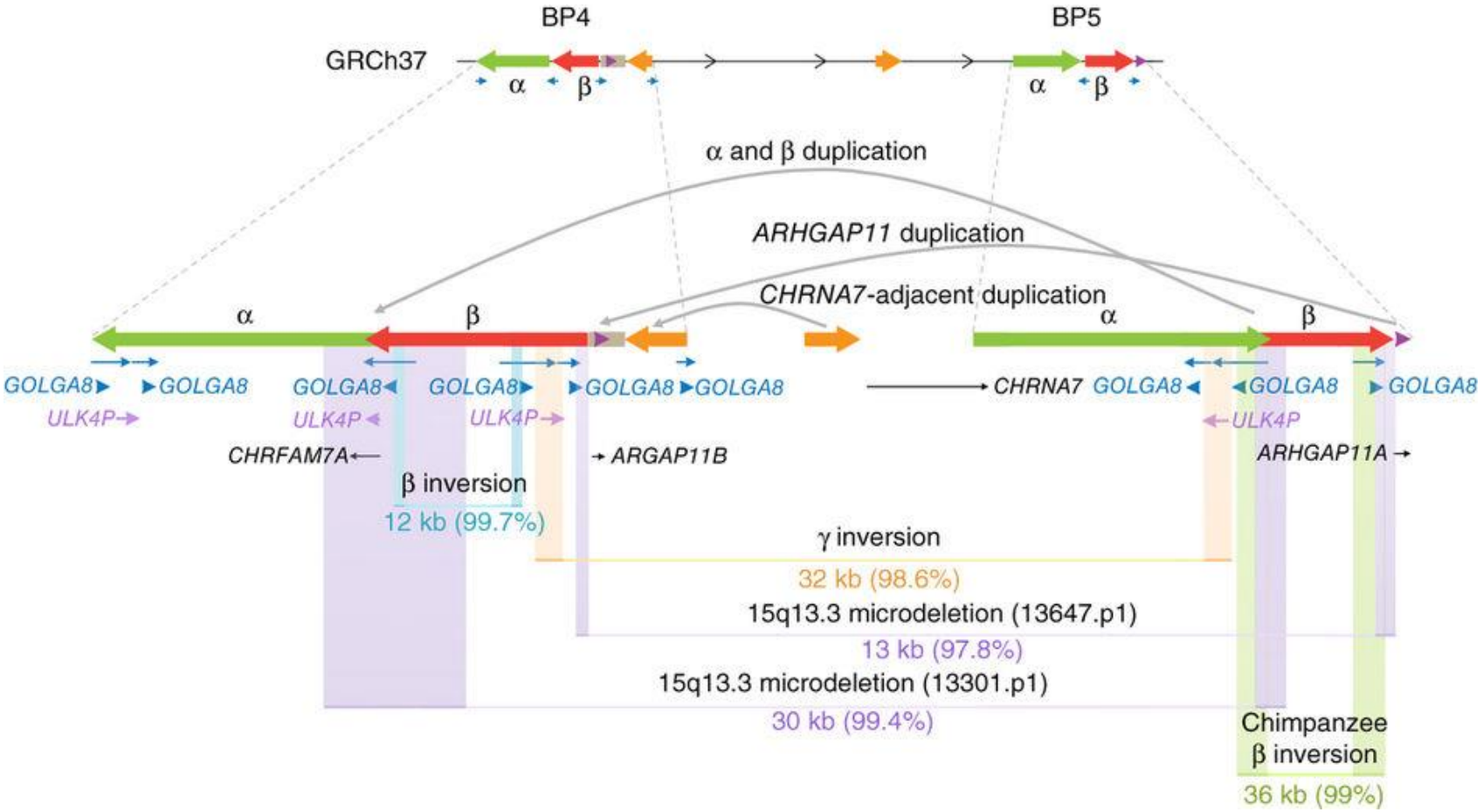


Figure 5.7. A schematic diagram showing eight independent rearrangements at the 15q13.2/13.3 region. Coloured boxes indicate the breakpoints identified for each rearrangement. The size and percent similarity of the paralogous sequences at the rearrangement breakpoints are shown. Figure from Antonacci et al. (2014).

Several RhoGAPs have been linked to schizophrenia. For example, a study reported an association between variation in the *ARHGAP32* gene, a neuron-associated GTPase-activating protein (Okabe et al., 2003), and schizophrenia and schizotypal personality traits (Ohi et al., 2012). *ARHGAP33* regulates synapse development and autistic-like behaviour (Schuster et al., 2015). A missense polymorphism in *CHN2* (also known as *ARHGAP3*) has been associated with schizophrenia in men (Hashimoto et al., 2005).

In a genome-wide association study (GWAS) from the Han Chinese population, Wong et al (2014) identified a schizophrenia susceptibility locus on Xq28, which harbours the gene *ARHGAP4* (Wong et al., 2013). In a study that used differential brain imaging activation patterns to determine candidate genes associated in schizophrenia, two candidate genes, *RSRC1* and *ARHGAP18*, were identified, both of which have a function in prenatal brain development.

Considering the above findings, we propose that both *ARHGAP5* and *ARHGAP11B* are interesting novel candidate genes for schizophrenia.

5.3.2 Limitations

It has become apparent that CNVs are common in human populations and play a significant role in the aetiology of complex disorders (Maiti et al., 2011). However, there are methodological challenges in identifying disease-specific CNVs and establishing their mechanism of action in disease causation.

It is possible that in the array-based analysis we missed true causative CNVs because our filtering criteria involved analysis with both PennCNV and cnvPartition. Using multiple software programs to identify CNVs has the advantage of increasing the specificity of CNV calls, but will have an undesirable decrease in sensitivity. With the HumanCore-24

BeadChips array having a median inter-marker spacing of 10kb, it is possible that we have missed smaller CNVs, or that mutational events are present involving small DNA changes. However, exome sequencing would have covered some of the areas of the genome that are not detectable *per se* using a SNP-array approach.

Computational prediction of CNVs from exome sequence data remains challenging. There are several algorithms that are commonly used, each having different sensitivities, but with a shared limitation in being able to detect small CNVs. Additionally, exonic CNV discovery using microarrays has limitations in that some exons would not be covered due to the low probe density in those regions. Thus, the rationale for combining a computational prediction algorithm for exome sequencing data and a high-resolution SNP-array is to increase the sensitivity and specificity of CN detection.

Moreover, array-based technologies have different degrees of genome coverage and there are many CNV calling algorithms. Although the Illumina HumanCore BeadChip 12v1 appears to meet most of the criteria in terms of coverage for CNV calling, a gold standard algorithm for data analysis has not been determined (Castellani et al., 2014; Zhang et al., 2011). Considering these limitations, whole-genome sequencing would have been an ideal choice to look for structural variants, as well as SNVs and indels, as it provides a more uniform and complete coverage of coding regions than current exome products (Meltz Steinberg et al., 2015).

The literature contains different results regarding the role that CNVs play in the aetiological basis of phenotypic discordance observed in MZ twins (Bloom et al., 2013; Ehli et al., 2012; Ono et al., 2010). This study adds to the growing literature that postzygotic de novo CN events between twins are rare, as no reproducible CNV

differences were found to explain the phenotypic discordance in the MZ twin pairs investigated.

This could be attributable to a relatively small sample size. Due to the statistical rarity of discordant MZ twins, this study only assessed thirteen twin pairs. Another explanation is that somatic CNVs are more likely to be present as mosaics. This hypothesis is in line with previous studies (Bruder et al., 2008; Forsberg et al., 2012; Magaard Koldby et al., 2016). However, the statistics of identifying somatically mosaic (incomplete) CNVs by SNP concordance analysis requires further methodological development beyond the scope of this study.

5.3.3 Conclusion

In summary, no CNV differences between co-twins or tissue types within a single individual were found. While this study did not provide decisive evidence for somatic genetic alterations as a mechanism for phenotypic variability, it highlights the need for further genome-wide research into the identification of modifying factors in these disorders. It also illustrates that shared CNVs overlapping disease-susceptibility genes can increase the risk for complex disorders, even in discordant MZ twins.

Four pre-twinning de novo CNVs were identified by the CN calling algorithms used, but were not experimentally validated. Three were found to not overlap any genes or regulatory regions. The tentative de novo CNV duplication found in the ADHD-discordant twin pair overlapped >85% proximal of *CD38*, and will be validated as part of a future study. This data suggests that post-zygotic CNVs may not play a significant role in twin discordance. Nevertheless, the shared CNVs detected herein are important, as some germline and even common population CNVs may predispose individuals to various phenotypes. The ability to better predict the varying degrees of pathogenicity of

structural variants and elucidating the role of modifier genes could significantly improve medical management and genetic counselling.

Chapter 6. Biochemical and proteomic analysis of twins discordant for ischaemic stroke

6.1 Overview

Stroke is a common disorder with significant morbidity and mortality, and complex aetiology involving both environmental and genetic risk factors. Although some of the major risk factors for stroke, such as smoking and hypertension, are well-documented, the underlying genetic and detailed molecular mechanisms remain largely elusive (The Lancet Neurology, 2013). Identification of biomarkers for stroke may help determine the aetiology, pathogenesis, clinical diagnosis and prognosis, and predict the outcome of pharmacological intervention such as thrombolysis.

Ischaemic stroke is reported to be misdiagnosed in more than 1 in 7 patients during the initial presentation in the emergency department, and increases to two thirds in cases where only nontraditional symptoms are present (Lever et al., 2013). It is important to distinguish between ischaemic and haemorrhagic stroke, and to identify other mimics such as structural brain lesions, epilepsy and migraine (Cordonnier and Leys, 2008). Although neuroimaging investigations such as computed tomography (CT) and magnetic resonance imaging (MRI) are upheld as the current diagnostic paradigm, they have several shortcomings. Ischaemic stroke can often be isodense on CT making it difficult to distinguish from haemorrhagic stroke. Stroke diagnosis is, therefore, reliant on probability rather than certainty. MRI is more reliable than CT for stroke diagnosis, but it is more expensive and not often routinely available at presentation. This delay, along with the potential for misdiagnosis, opens the door for molecular biomarkers, in the forms

of proteins, lipids and metabolites, to support the clinical diagnosis of stroke, as well as to identify patients at risk and provide prognostic indicators.

There are currently several candidate biomarkers that have been identified for stroke, but none are readily used in clinical practice as none have sufficient evidence (Jickling and Sharp, 2015). Thus, further studies and research are warranted.

A pair of MZ twins (KG and HG) discordant for ischaemic stroke were recruited in this study. The 62-year-old sisters of Caucasian descent have lived together since birth suggesting they share very similar environmental risk factors. They have been with the same medical practice for at least 50 years, and there is a family history of hypertension.

KG presented with a stroke in 2007 (aged 55 years). This has left her with persistent right-sided weakness affecting her balance, speech, and walking. In 2007 an MRI scan confirmed a left ganglion infarct, ischaemic changes in both cerebral hemispheres, and a small right temporal infarct. A coincidental finding of a right-sided intracavernous internal carotid artery aneurysm measuring 9 mm was made in 2007, prior to the stroke. This has been managed conservatively ever since, consisting of routine periodic follow-up imaging with MR angiography or CT angiography. Patients with significant aneurysm growth are strongly considered for interventional treatment, however KG refused this intervention despite hers measuring 10 mm in 2009. KG has not smoked for 7 years since the stroke, whereas HG, despite sharing the same risk factors and still smoking, has not had a stroke.

In addition, KG was diagnosed with a single patch of alopecia areata in 1995 (aged 44), and more recently with seborrhoeic dermatitis capitis in 2012 (aged 60). HG was diagnosed with atopic dermatitis in 2004 (aged 53). These problems might be linked to a

family history of psoriasis (in father) and suggest an autoimmune disorder in the family (see Appendix D for a detailed medical history of the twins).

This chapter outlines the clinical, cytogenetic, biochemical and proteomic profile comparisons between an MZ twin pair discordant for ischaemic stroke and cerebral aneurysm.

6.2 Results

6.2.1 Cytogenetics

Chromosome analysis was performed on GTG banded metaphases from synchronised peripheral lymphocyte cultures using standard procedures. GTG-banded chromosomes of peripheral blood lymphocytes showed a pericentric inversion on chromosome 5 in both KG and HG, with breakpoints at 5p13.1 and 5q11.2 (Figure 6.1). This rare pericentric inversion has been documented in previous studies, albeit with slightly different breakpoints (Concolino, 2002). The breakpoint in the short arm of chromosome 5 in this report is not in the cri du chat syndrome critical region, mapped in 5p15.3 for high pitched cry and in 5p15.2 for the remaining features (Beemer et al., 2008; Goodart, Butler and Overhauser, 1996). Concolino (2002) reports a family with the karyotype inv(5)(p15.1q11.2) that segregates with benign familial neonatal convulsions. Affected individuals were intellectually normal but had a history of convulsions in infancy and afebrile seizures later in childhood. Interestingly, KG was diagnosed with epilepsy and HG, although not given this diagnosis, had a marked delta response to photic stimulation (Appendix D).

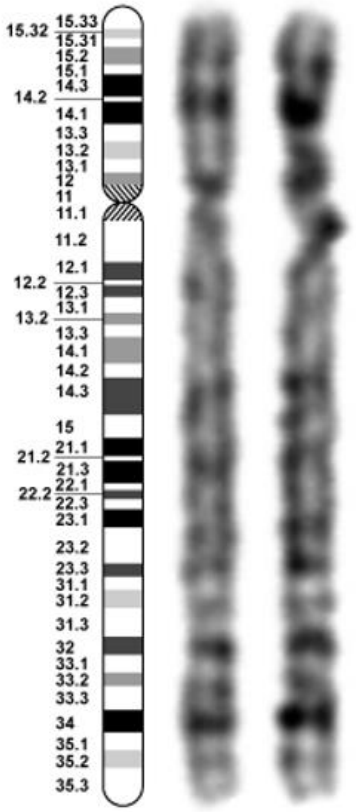


Figure 6.1. Idiogram of chromosome 5 and representative G-banded karyotypes of KG. The normal chromosome 5 is depicted on the left, and the aberrant chromosome 5 with pericentric inversion, inv(5)(p13.1q11.2), is on the right.

6.2.2 Blood biochemistry

To investigate potential causes or effects of the stroke, a detailed biochemical comparison was undertaken on blood samples from the twins. Blood samples were drawn from the twins by standard venesection into plain tubes and fluoride and EDTA anticoagulated tubes, just prior to lunch, and processed in parallel. A total of 59 individual markers were measured to check for renal, liver, glucose, cardiovascular, blood, immunological, and hormonal function (Table 6.1). The data were compared with the corresponding reference intervals for the age group.

Most of the results were within the normal range and similar between the sisters except for g-glutamyltransferase (GGT), an enzyme present in liver, but normally used to indicate alcohol-related liver disease, and ESR. A greater than 10-fold increase in GGT was identified in KG (244 U/L) compared to HG (23 U/L), the normal reference range being 8-40 U/L in females. An even greater difference of >18-fold was observed for ESR (HG=2mm/h; KG=36mm/h; reference interval=1-20 mm/h). Review of the medical notes for both the patients established that KG had elevated levels of GGT (81 U/L) significantly above the normal range, as far back as 1993 (Figure 6.2), although these values continued to increase up to when she had stroke and beyond. By contrast, GGT readings for HG have been near constant over the same period, and always within the normal range. The medical notes also reveal significantly elevated ESR in KG as far back as 2009 (27 mm/h) and 2008 (55 mm/h) following the stroke. These findings were subsequently published (Vadgama, et al., 2015).

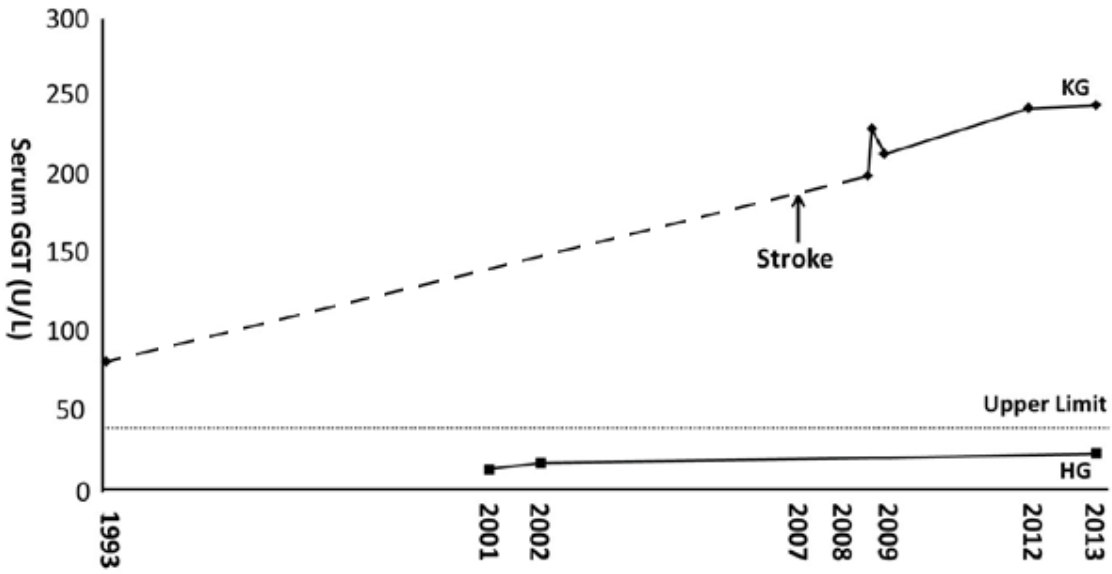


Figure 6.2. A comparison of variation in serum GGT over time. GGT readings were collected from patients' notes. KG has had high GGT values since 1993, prior to the onset of stroke in 2007. The unaffected twin, HG, has had consistently low GGT values over time. The dotted horizontal line depicts the upper limit of the reference interval.

Parameter	Units	HG	KG	Reference interval	Fold difference
			Clinical Biochemistry		
GLUCOSE FLUORIDE	mmol/L	4.0	6.6	3.0 — 6.0 (fasting)	1.7
SODIUM	mmol/L	138	134	133 — 146	1.0
POTASSIUM	mmol/L	4.2	4.6	3.5 — 5.3	1.1
CHLORIDE	mmol/L	93	96	95 — 108	1.0
BICARBONATE	mmol/L	31	25	22 — 29	1.2
UREA	mmol/L	4.3	4.0	2.5 — 7.8	1.1
CREATININE	µmol/L	52	51	60 — 110	1.0
BILIRUBIN	µmol/L	5	3	0 — 21	1.7
ALANINE TRANSAMINASE	U/L	25	24	0 — 52	1.0
ALKALINE PHOSPHATASE	U/L	97	184	30 — 130	1.9
ALBUMIN	g/L	40	35	35 — 50	1.1
GAMMA GT	IU/L	23	244	0 — 38	10.6
CALCIUM	mmol/L	2.50	2.24	2.20 — 2.60	1.1
PHOSPHATE	mmol/L	2.08	1.27	0.80 — 1.50	1.6
ADJUSTED CALCIUM	mmol/L	2.50	2.34	2.20 — 2.60	1.1
CREATINE KINASE	U/L	83	54	40 — 320	1.5
CHOLESTEROL	mmol/L	4.4	3.7	0.90 — 2.20	1.2
TRIGLYCERIDE	mmol/L	1.14	0.56	0.80 — 2.00	2.0
CHOLESTEROL HDL	mmol/L	1.24	2.22	3.3 — 5.2	1.8
CHOLESTEROL LDL (CALC)	mmol/L	2.6	1.2	<3	2.2
TOTAL CHOLESTEROL/HDL RATIO		3.5	1.7	0.0 — 5.0	2.1
NON HDL CHOLESTEROL	mmol/L	3.2	1.5		2.1
C REACTIVE PROTEIN	mg/L	11.9	25.8	0.0 — 10.0	2.2
FREE T4	pmol/L	16.5	13.8	10.0 — 23.0	1.2
TSH	mU/L	0.89	1.03	0.40 — 5.00	1.2
VITAMIN B12	ng/L	372	446	180 — 1000	1.2
FOLATE	µg/L	5.6	13.9	5.0 — 10.0	2.5
			Serum hormones and natriuretic peptide		
PROLACTIN	mIU/L	66	123	86 — 324	1.9
LH	IU/L	21.1	17.1	>15	1.2
FSH	IU/L	60.9	52.0	>30	1.2
OESTRADIOL	pmol/L	100	61	<90	1.6
PROGESTERONE	nmol/L	2	<1	<2	-
TESTOSTERONE	nmol/L	2.2	1.6	0.4 — 2.7	1.4
CARDIAC TROPONIN I	ng/L	<17	<17	0 — 50	-
NT-PRO BNP	ng/L	100	192		1.9
			Haematology		
HB	g/L	144.0	105.0	130 — 180	1.4
WBC	10^9/L	17.1	7.9	4.0 — 11.0	2.2
PLATELET	10^9/L	244	207	150 — 450	1.2
MCV	fL	87.2	86.3	80 — 97	1.0
ESR	mm/h	2	36	1 — 20	18.0
NEUTROPHIL	10^9/L	13.8	6.5	1.7 — 8.0	2.1
LYMPHOCYTE	10^9/L	2.2	0.6	1.0 — 4.0	3.7
MONOCYTE	10^9/L	0.7	0.4	0.24 — 1.1	1.8
EOSINOPHIL	10^9/L	0.3	0.4	0.1 — 0.8	1.3
BASOPHIL	10^9/L	0.1	0.1	0.0 — 0.3	1.0
HCT		0.47	0.34	0.41 — 0.52	1.4
RBC	10^{12}/L	5.34	3.96	4.5 — 6.0	1.3
MCH	pg	26.9	26.5	27 — 33	1.0
MCHC	g/L	308	307	320 — 365	1.0
RDW	%	14.0	14.4	11.5 — 15.0	1.0
LUC	10^9/L	0.1	0.1	0.0 — 0.4	1.0
%Hypo RBC	%	13.4	11.2		1.2
Hb (calc)	g/L	142.0	105.0		1.4
MPXI		1.5	^3.3	-20 — 100	2.2
CD4	%	55	60		1.1
CD3 TruCount	10^9/L	1.565	0.421		3.7
CD4 TruCount	10^9/L	1.151	0.328		3.5
CD8 TruCount	10^9/L	0.357	0.092		3.9
TruCount Lymphs	10^9/L	2.171	0.550		3.9

Table 6.1. A comparison of blood markers between MZ twins discordant for stroke. Blood samples were taken at the same time prior to lunch, and processed in parallel. GGT and ESR are significantly elevated in the affected, relative to the unaffected twin (Vadgama et al, 2015).

6.2.3 Label-free proteomics analysis

Preliminary analysis using unbiased, label-free proteomics technology identified potentially new protein biomarkers that are related to the pathophysiology of stroke. Proteins were extracted from serum samples collected from each twin. Following staining of the gel, a prominent band at approximately 60 kDa, corresponding mainly to albumin, was cut out from the gel to enhance the detection of less abundant proteins as albumin represents approximately 55% of total cellular protein (Walker, Hall and Hurst, 1990) (Figure 6.3). The proteins were eluted from the gel for proteomic studies, as described (see Materials and Methods).

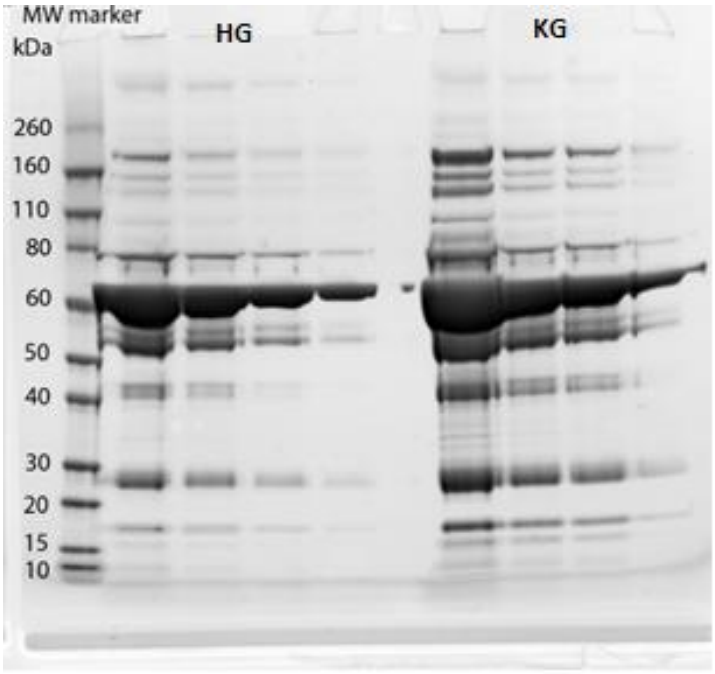


Figure 6.3. SDS-PAGE gel analysis. The 2 serum samples (KG and HG) were run on a 4-12% bis-tris with MOPS running buffer in decreasing concentrations: 0.5ul, 0.1ul, 0.05ul and 0.025ul (i.e. 1X, 1/5X, 1/10X and 1/20X dilutions). For the serum loading there appears to be more protein in the KG sample - this is based on the intensity of the stain used (Colloidal coomassie blue from Invitrogen) when compared to the first serum: HG.

The proteomes of the MZ twin pair discordant for stroke were quantified and compared at a post-symptomatic stage of the disorder. The proteins were first ranked according to the fold difference between the twins, including proteins that were present in only the affected or unaffected twin (data not shown).

The functional analysis tool PANTHER and VLAD was used for statistical analysis. PANTHER is a database of gene families and gene attributes (Mi, Muruganujan and Thomas, 2012), which provides tools for functional analysis of lists of genes or proteins. VLAD is a functional analysis tool (Richardson and Bult, 2015) which uses the most up-to-date GO annotation files. Gene lists were analysed statistically by enrichment analysis on 33 of the proteins that had a fold difference of >1.5 between the twins. The enrichment analysis takes a list of genes that have been ascribed with a numerical value, then finds functional classes for which the genes of that class have values that are non-randomly selected from the genome-wide distribution of values. PANTHER includes a Bonferroni correction to counteract the problem of multiple comparisons, whereas VLAD does not include any corrections. The summary table shows an interesting enrichment of blood coagulation, wound healing and inflammatory response, and a list of genes that have previously been implicated in ischaemic stroke (Table 6.2).

Term	Panther	VLAD			
	P-value	P	k	M	Genes
***GO:1903034**: regulation of response to wounding*	9.02E-15	1.01E-19	15	526	*A2M, AGT, APOA1, APOD, APOE, C5, CFH, CFI, F12, F2, HRG, PLG, S100A9, SERPINC1, THBS1*
***GO:0080134**: regulation of response to stress*	7.92E-11	2.97E-16	17	1400	*A2M, AGT, APOA1, APOD, APOE, C5, CFH, CFI, F12, F2, HP, HPX, HRG, PLG, S100A9, SERPINC1, THBS1*
***GO:0006950**: response to stress*	7.47E-10	9.27E-15	24	4990	*A2M, AGT, APOA1, APOD, APOE, C5, CD5L, CFH, CFI, F12, F2, HBA1, HBB, HBD, HP, HRG, IGHG2, ORM1, PLG, S100A9, SERPINA3, SERPINC1, THBS1, VWF*
***GO:0050727**: regulation of inflammatory response*	6.02E-10	1.03E-14	11	367	*A2M, AGT, APOA1, APOD, APOE, C5, CFH, CFI, F12, S100A9, SERPINC1*
***GO:0031347**: regulation of defence response*	5.78E-08	4.10E-13	12	704	*A2M, AGT, APOA1, APOD, APOE, C5, CFH, CFI, F12, HPX, S100A9, SERPINC1*
***GO:0009611**: response to wounding*	1.09E-07	5.61E-13	13	951	*A2M, APOA1, APOD, APOE, F12, F2, HBB, HBD, HRG, PLG, SERPINC1, THBS1, VWF*
***GO:0007596**: blood coagulation*	8.75E-07	1.11E-12	11	564	*A2M, APOA1, F12, F2, HBB, HBD, HRG, PLG, SERPINC1, THBS1, VWF*
***GO:0042060**: wound healing*	5.68E-07	2.21E-12	12	813	*A2M, APOA1, APOD, F12, F2, HBB, HBD, HRG, PLG, SERPINC1, THBS1, VWF*
***GO:0030193**: regulation of blood coagulation*	2.17E-07	1.49E-11	7	125	*APOE, F12, F2, HRG, PLG, SERPINC1, THBS1*
***GO:0050896**: response to stimulus*	4.64E-04	6.35E-08	26	12252	*A2M, AGT, APOA1, APOD, APOE, C5, CD5L, CFH, CFI, DSG1, F12, F2, HBA1, HBB, HBD, HP, HRG, IGHG2, ORM1, PLG, S100A9, SERPINA3, SERPINA7, SERPINC1, THBS1, VWF*

Table 6.2. Summary findings of PANTHER analysis with results compared to VLAD analysis. P-values for PANTHER analysis include Bonferroni correction. For the VLAD analysis, the P-value scoring method was selected, thus displaying the node's scores as a triple (P, k, M). P = the node's P-value; k = the number of genes in the query set annotated to that node; M = the number of genes in the database annotated to that node.

In the preliminary analysis above, only one run was performed per sample as a proof of principle. This generated semi-quantitative data as no technical replicates were performed, and both the higher and lower level proteins were included in a single analysis, rather than separating them. Theoretically, this is a valid approach to take because some proteins might positively regulate a process, whereas others negatively regulate a process but they would still both regulate the same process.

For a more comprehensive analysis, and with the aim to overcome the limit of reproducibility in the serum preparation, two additional runs were separately performed, each containing three technical replicates of control serum and serums of both twins (Figure 6.4). Thus 18 samples in total were analysed in nano-scale liquid chromatographic tandem mass spectrometry (nLC-MS/MS).

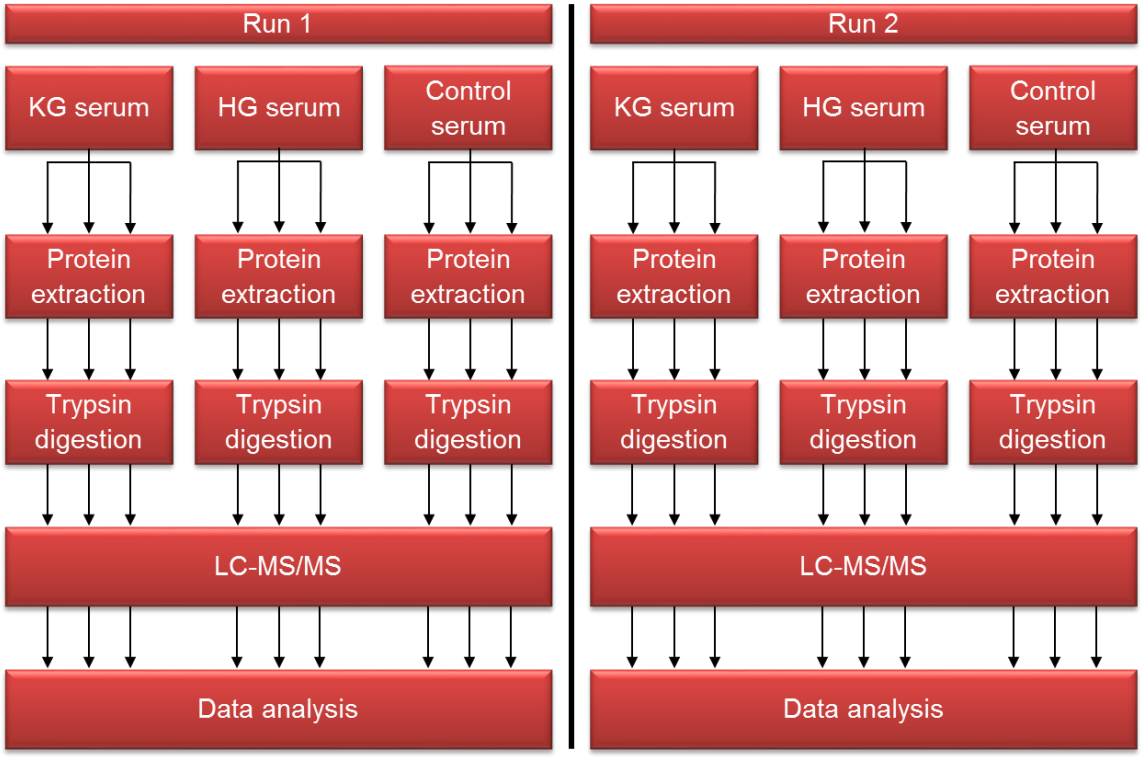


Figure 6.4. Schematic diagram illustrating optimised experimental design for the second label-free proteomics comparison of the proteome of a pair of discordant MZ twins.

To identify the proteins that were differentially expressed between the affected and unaffected twin blood serum, ANOVA test was performed in Progenesis LC-MS that identified 179 and 209 proteins ($P<0.05$) in runs 1 and 2 respectively.

Several of the proteins detected at higher levels in the serum of the affected twin had previously been associated with biological pathways relevant to ischaemic stroke, including haemoglobin subunit alpha 1 (Huang et al., 2009), alpha-2-macroglobulin (Nezu et al., 2013), hyaluronan binding protein 2 (Hanson et al., 2012), S100A8 (Ziegler et al., 2007) (Table 6.3). In addition, fibulin 1, an extracellular matrix protein associated with arterial stiffness (Cangemi et al., 2011), was present at a much higher level in the affected twin and had the highest fold difference average. Furthermore, the low serum levels of desmoplakin in the affected twin (and therefore identified as being at high levels in HG serum), has been inversely associated with ischaemic stroke (López-Farré et al., 2012) (Table 6.4).

Gene symbol	*Approved gene name*	*UniProt ID*	*Run 1*		*Run 2*		*Average fold difference*
			fold difference	**P-value**	**fold difference**	**P-value**	
FBLN1	fibulin 1	P23142	2.79	8.27E-05	23.96	4.49E-02	13.37
CA1	carbonic anhydrase 1	P00915	5.64	1.12E-05	4.84	5.89E-08	5.24
PIGR	polymeric immunoglobulin receptor	P01833	8.29	3.97E-02	1.96	7.36E-05	5.12
IGKV1-5	immunoglobulin kappa variable 1-5	P01602	2.22	1.31E-06	5.14	1.06E-05	3.68
GPLD1	glycosylphosphatidylinositol specific phospholipase D1	P80108	3.85	6.56E-06	2.08	1.82E-04	2.96
IGKV3-20	immunoglobulin kappa variable 3-20	P01620	2.38	3.47E-07	2.88	3.09E-03	2.63
PZP	PZP, alpha-2-macroglobulin like	P20742	2.91	4.16E-06	2.03	2.43E-05	2.47
C5	complement C5	P01031	3.44	4.77E-07	1.39	7.35E-05	2.42
APOL1	apolipoprotein L1	O14791	1.47	3.61E-03	2.94	2.17E-04	2.21
HBA1	hemoglobin subunit alpha 1	P69905	1.64	6.91E-07	2.69	5.39E-09	2.17
A2M	alpha-2-macroglobulin	P01023	2.78	3.70E-09	1.48	1.00E-05	2.13
HABP2	hyaluronan binding protein 2	Q14520	2.17	8.05E-04	1.89	1.60E-02	2.03
IGLV3-19	immunoglobulin lambda variable 3-19	P01714	1.54	2.92E-06	2.14	4.06E-07	1.84
C4BPA	complement component 4 binding protein alpha	P04003	2.20	1.05E-08	1.39	3.56E-04	1.80
C7	complement C7	P10643	2.06	1.38E-04	1.49	1.06E-02	1.78
FCN3	ficolin 3	O75636	1.76	6.90E-03	1.69	3.64E-04	1.72

S100A8	S100 calcium binding protein A8	P05109	1.43	1.59E-02	1.89	1.27E-02	1.66
CD5L	CD5 molecule like	O43866	1.16	6.30E-03	1.91	2.36E-05	1.53
F2	coagulation factor II, thrombin	P00734	1.29	1.20E-03	1.72	4.26E-05	1.50

Table 6.3. The top proteins identified at high levels in the affected twin compared with the unaffected twin (ranked by fold change) that were present in both runs. A fold change threshold of >1.5-fold was used as a cut off. Fold difference values in red indicate proteins which had been associated with stroke or stroke risk factors by other studies.

Gene name	*Protein name*	*UniProt ID*	*Run 1*		*Run 2*		*Average fold difference*
			fold difference	**P-value**	**fold difference**	**P-value**	
KRT17	keratin 17	Q04695	13.33	4.42E-05	4.08	1.70E-05	8.71
SPTB	spectrin beta, erythrocytic	P11277	13.81	5.55E-07	2.60	7.52E-06	8.20
SERPINA4	serpin family A member 4	P29622	6.95	3.33E-03	1.25	2.05E-02	4.10
KRT16	keratin 16	P08779	4.60	1.47E-06	2.99	6.10E-05	3.79
DSP	desmoplakin	P15924	2.72	4.11E-05	1.82	1.45E-02	2.27
IGHG4	immunoglobulin heavy constant gamma 4 (G4m marker)	P01861	2.80	2.71E-06	1.60	1.18E-03	2.20
DSG1	desmoglein 1	Q02413	1.95	1.69E-06	1.99	2.96E-03	1.97
KRT9	keratin 9	P35527	2.21	4.44E-06	1.60	3.29E-05	1.91
CPN1	carboxypeptidase N subunit 1	P15169	2.10	8.88E-03	1.50	4.96E-03	1.80
IGHA2	immunoglobulin heavy constant alpha 2 (A2m marker)	P01877	1.80	1.59E-05	1.55	1.34E-03	1.68
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	P04406	1.68	2.62E-03	1.63	2.27E-03	1.65
HPX	hemopexin	P02790	1.25	6.11E-02	1.81	2.15E-05	1.53
IGHA1	immunoglobulin heavy constant alpha 1	P01876	1.45	2.37E-04	1.58	6.44E-04	1.52

Table 6.4. The top proteins identified at high levels in the unaffected twin compared with the affected twin (ranked by fold change) that were present in both runs. A fold change threshold of >1.5-fold was used as a cut off. Fold difference values in red indicate proteins which had been associated with stroke or stroke risk factors by other studies.

6.2.4 Functional analysis:

Functional analysis was performed independently on all 4 datasets of differentially expressed proteins, to classify the proteins according to 'biological process', 'molecular function' and 'cellular component'. The g:Profiler web server (Reimand et al., 2007; http://biit.cs.ut.ee/gprofiler/) comprises several tools to perform functional enrichment analysis and mine additional information. The statistical method used by this tool, g:SCS (Set Counts and Sizes, Reimand et al., 2007) takes into account the ontology structure supporting the annotations and enables significantly enriched GO terms to be identified within short lists of proteins (down to lists of only 11 proteins).

Proteins were only included in the analysis if they were identified as at higher levels in either the affected or unaffected twin in both runs, or only one run. Proteins with high levels in one run, but low levels in the equivalent other run were excluded from the analysis. After applying this exclusion criteria, the number of proteins remaining were 125 and 156 in runs 1 and 2, respectively. Some of the protein identifiers associated with the differentially expressed peptides did not have any associated GO terms, and thus were not included in the g:Profiler analysis. In most cases these peptides were predicted to be derived from an immunoglobulin molecule. Overall, 95 and 139 proteins, in runs 1 and 2 respectively, were included in the analysis.

The GO terms enriched in the serum of each twin were relevant to very different biological pathways. The 'biological process' GO categories overrepresented in the affected twin related to cytolysis, response to stress, and blood coagulation (Table 6.5 and Appendix B). With between 15-30% of the proteins differentially overexpressed in KG associated with wound healing and/or blood coagulation. In HG, the enriched GO terms were almost exclusively related to skin development (Table 6.6 and Appendix B).

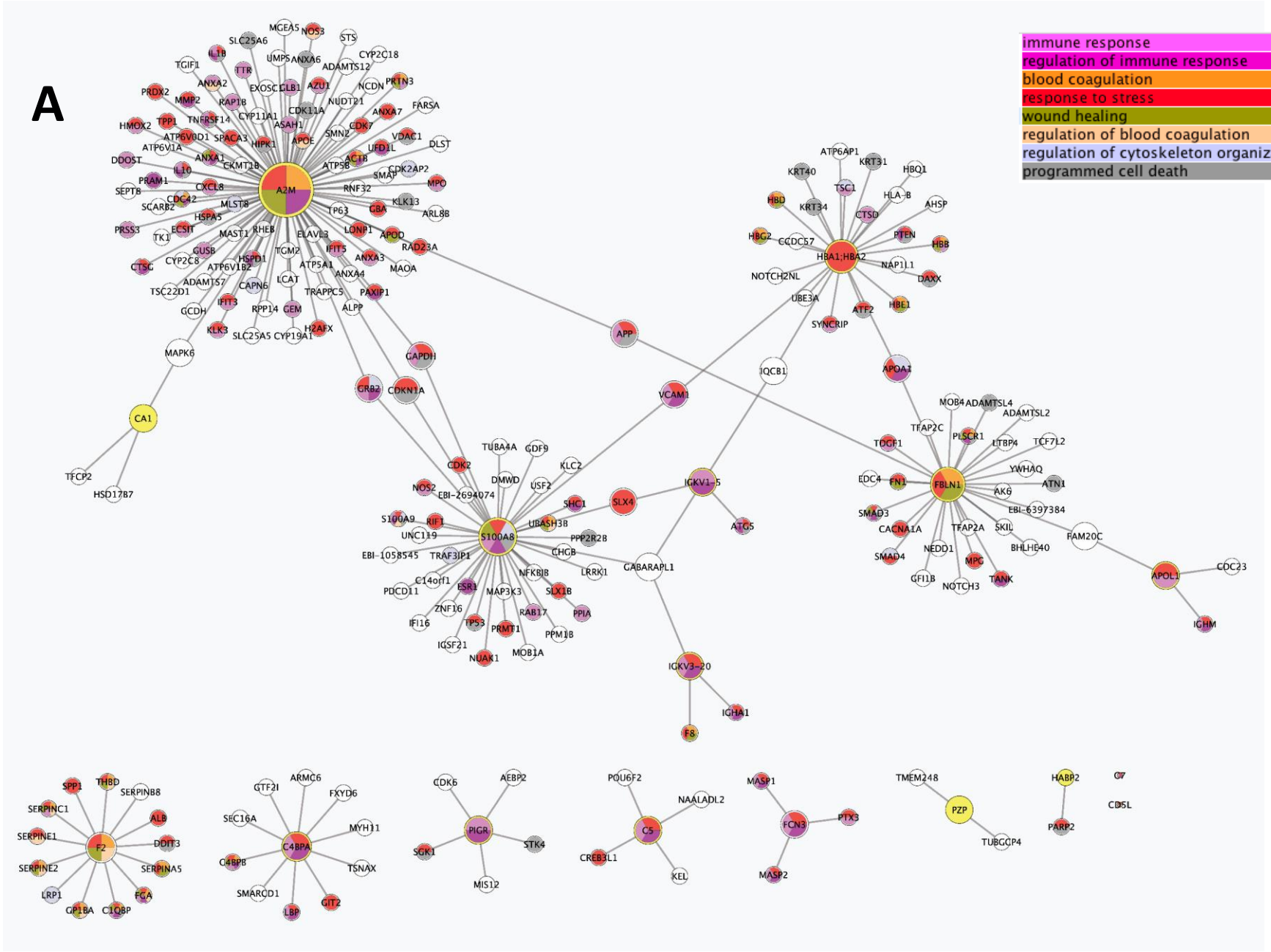
GO term	P-value	S	T	P-value	S	T
		t=50	t=18,711		t=77	t=18,711
	Run 1			Run 2		
response to stress	1.59E-11	35	3579	1.28E-08	42	3579
wound healing	5.06E-11	17	540	1.32E-05	15	540
haemostasis	7.89E-10	14	357	6.87E-05	12	357
blood coagulation	6.51E-10	14	352	5.88E-05	12	352
fibrin clot formation	5.18E-12	8	26	2.75E-08	7	26
lipoprotein particle	1.42E-04	5	39	3.43E-05	6	39

Table 6.5. Selection of enriched GO terms associated with the proteins present at high levels in the affected twin. GO terms were identified as significantly enriched in the high level proteins in KG following g:Profiler analysis. P-values <0.05 are considered as significantly enriched. **S** = number of protein identifiers (IDs) in both the study dataset and GO term group, **T** = number of human protein IDs associated with the GO term, **t** = number of protein IDs in study or GO datasets.

GO term	P-value	S	T	P-value	S	T
		t=45	t=18,711		t=62	t=18,711
	Run 1			Run 2		
epidermis development	5.01E-06	10	297	1.55E-02	8	297
skin development	1.52E-05	9	242	2.57E-04	9	242
establishment of skin barrier	3.04E-02	3	18	8.50E-04	4	18
intermediate filament cytoskeleton	1.15E-12	14	247	3.96E-03	8	247

Table 6.6. Selection of enriched GO terms associated with the proteins at high levels in the unaffected twin. GO terms were identified as significantly enriched in the high level proteins in HG following g:Profiler analysis. P-values <0.05 are considered as significantly enriched. **S** = number of protein identifiers (IDs) in both the study dataset and GO term group, **T** = number of human protein IDs associated with the GO term, **t** = number of protein IDs in study or GO datasets.

A network analysis of the overexpressed proteins in each twin was undertaken. Only proteins which were overexpressed in both runs and with an average fold difference >1.5 were included in the analysis (Tables 6.3 and 6.4). This analysis demonstrated that 8 of the proteins overexpressed in KG's serum can be located within a single network, with 11 proteins not connected to any network. Further investigation is need to clarify whether any of these unconnected proteins are associated. All but one of the 13 proteins overexpressed in the unaffected twin's serum were associated with a close network (Figure 6.5).


A
immune response
regulation of immune response
blood coagulation
response to stress
wound healing
regulation of blood coagulation
regulation of cytoskeleton organization
programmed cell death
A2M
HBA1;HBA2
FBLN1
S100A8
APP
GAPDH
VCAM1
IGKV1-5
IGKV3-20
C5
C4BPA
F2
PIGR
FCN3
PZP
APOL1
CA1
MAPK6
GRB2
CDKN1A
SLX4
IQCB1
APOA1
GABARAPL1
F8
IGHA1
IGHM
TUBGCP4
PARP2
CD5L

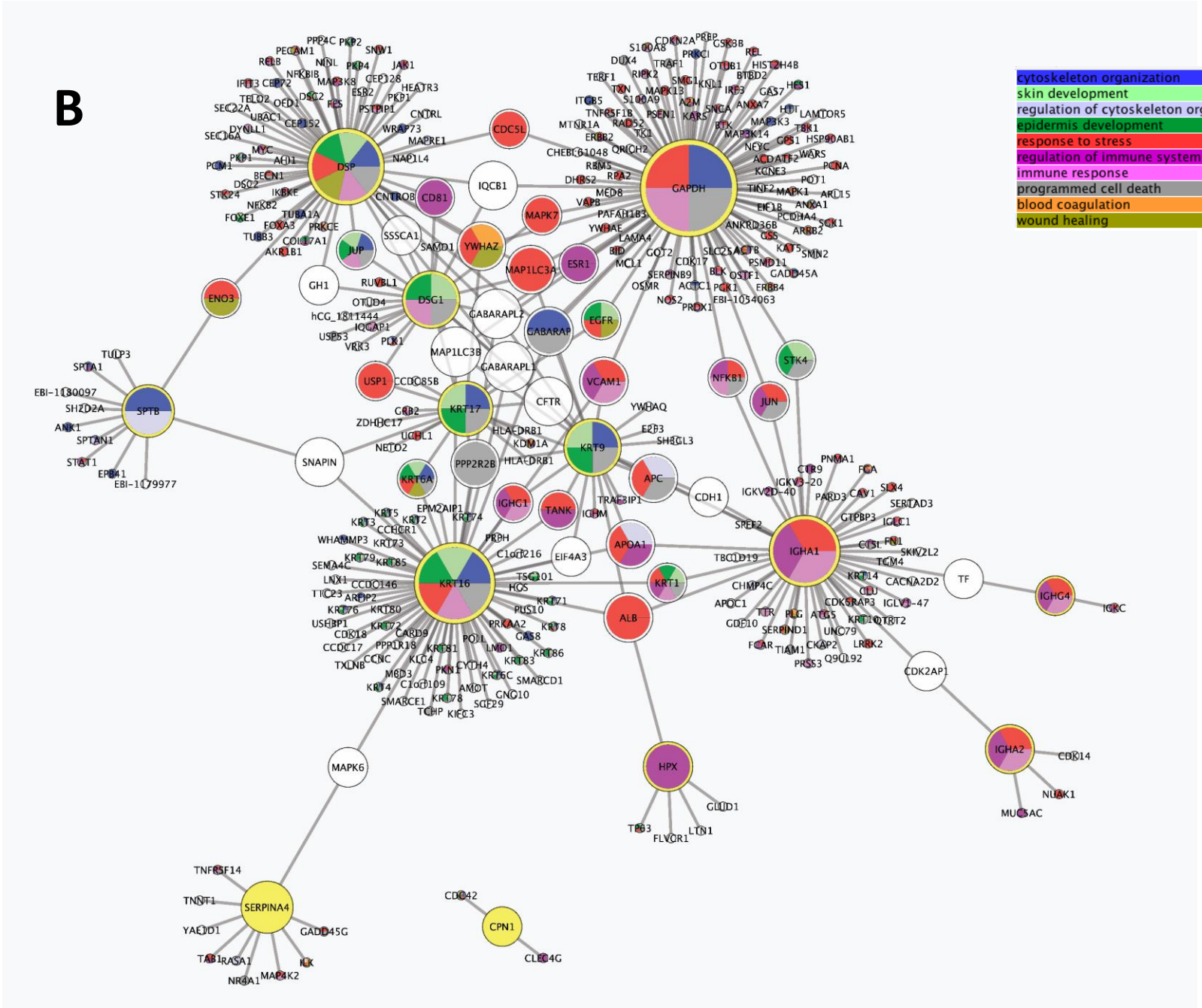

B
cytoskeleton organization
skin development
regulation of cytoskeleton organization
epidermis development
response to stress
regulation of immune system process
immune response
programmed cell death
blood coagulation
wound healing
DSP
GAPDH
DSG1
KRT17
KRT9
KRT16
KRT6A
SPTB
JUP
YWHAZ
MAP1LC3A
ESR1
EGFR
VCAM1
NFKB1
STK4
JUN
APC
CDH1
IGHA1
IGHG4
IGKC
TF
APOA1
ALB
HPX
IGHA2
CDK2AP1
SERPINA4
MAPK6
CPN1
CDC42
CLEC4G
ENO3
SNAPIN
CDC5L
IQCB1
MAPK7
CD81
SSSCA1
SAMD1
GABARAP
GABARAPL2
MAP1LC3B
GABARAPL1
CFTR
PPP2R2B
IGHG1
TANK
EIF4A3
KRT1
PKP2
SNW1
PKP4
JAK1
CEP128
HEATR3
PKP1
CNTRL
WRAP73
MAPRE1
NAP1L4
CNTROB
TUBA1A
PRKCE
COL17A1
AKR1B1
RUVBL1
OTUD4
IQGAP1
PLK1
VRK3
USP53
hCG_1811444
USP1
CCDC85B
GRB2
ZDHHC17
UCHL1
NETO2
HLA-DRB1
KDM1A
YWHAQ
EZR
SH3GL3
TRAF3IP1
IGHM
PRPH
C1orf216
TSG101
HOS
PLS10
PRKAA2
GAS8
PQI1
LMO1
KRT86
KRT83
SMARCD1
KRT6C
PKN1
CYTH4
AMOT
GNG10
SGF29
KIF3
TCHP
SMARCE1
KRT78
C1orf109
KRT14
CCNC
KLC4
TXLNB
M6D3
KRT81
PPP1R18
CDK18
CCDC17
CARD9
KRT72
USH1P1
KRT76
KRT80
ARPP2
TTC23
LNX1
CCDC146
SEMA4C
KRT79
KRT85
WHAMMP3
KRT73
KRT3
KRT5
KRT2
KRT74
EPM2AIP1
CCHCR1
KRT71
KRT18
KRT8
CHMP4C
APOC1
TTR
PLG
ATG5
GDF10
SERPIND1
UNC79
FCAR
TIAM1
CKAP2
LRRK2
PRSS3
Q9Q192
CDK5RAP3
IGLV3-47
KRT10
KRT2
CACNA2D2
KRT14
CLU
TBC1D19
TCM4
SKIV2L2
CCSI
FN1
SERTAD3
GTPBP3
IGLC1
SPEF2
PARD3
CAV1
SLK4
IGKV2D-40
IGKV3-20
CTR9
PNMA1
FCA
TULP3
SPTA1
EBI-1130097
SHZD2A
ANK1
SPTAN1
STAT1
EPB41
EBI-1179977
HPX
GLUD1
TF63
FLVCR1
LTN1
TNFRSF14
TNNT1
YAED1
GADD45G
TAB1
RASA1
NR4A1
MAP4K2
NLK
CDK14
NUAK1
MUC5AC
CDKN2A
PREP
CSK3B
REL
S100A8
PRKCI
OTUB1
HIST2H4B
DUX4
TRAF1
TERF1
RIPK2
SMG1
KNL1
BTBD2
HES1
TXN
MAPK13
IRF3
GAS7
S100A9
A2M
SNCA
ANXA7
HIT
ITGB5
TNFRSF1B
PSEN1
KARS
LAMTOR5
MTNR1A
RAD52
BTK
MAP3K3
TBK1
ERBB2
TK1
MAP3K14
GOS1
HSP90AB1
CHEBI:61048
ORC6
NEYC
WARS
RBM5
ACOT7
PCNA
DHRS2
RPA2
KCNE3
POT1
MED8
TNF2
MAPK1
ARL15
VAPB
EIF1B
ANXA1
PAFAH1B3
PCDHA4
YWHAE
ANKRD36B
SCK1
LAMA4
GSS
ARRB2
BID
GOT2
SLC25A37
KAT5
SMN2
MCL1
CDK17
PSMD11
SERPINB9
BLK
OSTF1
GADD45A
OSMR
ACTC1
PCK1
ERBB4
NOS2
EBI-1054063
PRDX1
PECAM1
RELB
NFKBIB
NINL
MAP3K8
IRF3
CEP72
TEL02
OFD1
DSC2
FES
SEC22A
UBAC1
PSTPIP1
PKP1
CEP152
DYNLL1
SEC16A
MYC
PKP1
AHI1
PCM1
BECN1
DSC2
IKBKE
STK24
NFKB2
FOXE1
FOXA3
TUBB3
PRKCE
COL17A1
AKR1B1
JUP
GH1
ENO3
SSSCA1
SAMD1
YWHAZ
MAP1LC3A
ESR1
EGFR
GABARAP
GABARAPL2
MAP1LC3B
GABARAPL1
CFTR
VCAM1
NFKB1
STK4
JUN
APC
CDH1
IGHA1
IGHG4
IGKC
TF
APOA1
ALB
HPX
IGHA2
CDK2AP1
SERPINA4
MAPK6
CPN1
CDC42
CLEC4G
ENO3
SNAPIN
CDC5L
IQCB1
MAPK7
CD81
SSSCA1
SAMD1
GABARAP
GABARAPL2
MAP1LC3B
GABARAPL1
CFTR
PPP2R2B
IGHG1
TANK
EIF4A3
KRT1
PKP2
SNW1
PKP4
JAK1
CEP128
HEATR3
PKP1
CNTRL
WRAP73
MAPRE1
NAP1L4
CNTROB
TUBA1A
PRKCE
COL17A1
AKR1B1
RUVBL1
OTUD4
IQGAP1
PLK1
VRK3
USP53
hCG_1811444
USP1
CCDC85B
GRB2
ZDHHC17
UCHL1
NETO2
HLA-DRB1
KDM1A
YWHAQ
EZR
SH3GL3
TRAF3IP1
IGHM
PRPH
C1orf216
TSG101
HOS
PLS10
PRKAA2
GAS8
PQI1
LMO1
KRT86
KRT83
SMARCD1
KRT6C
PKN1
CYTH4
AMOT
GNG10
SGF29
KIF3
TCHP
SMARCE1
KRT78
C1orf109
KRT14
CCNC
KLC4
TXLNB
M6D3
KRT81
PPP1R18
CDK18
CCDC17
CARD9
KRT72
USH1P1
KRT76
KRT80
ARPP2
TTC23
LNX1
CCDC146
SEMA4C
KRT79
KRT85
WHAMMP3
KRT73
KRT3
KRT5
KRT2
KRT74
EPM2AIP1
CCHCR1
KRT71
KRT18
KRT8
CHMP4C
APOC1
TTR
PLG
ATG5
GDF10
SERPIND1
UNC79
FCAR
TIAM1
CKAP2
LRRK2
PRSS3
Q9Q192
CDK5RAP3
IGLV3-47
KRT10
KRT2
CACNA2D2
KRT14
CLU
TBC1D19
TCM4
SKIV2L2
CCSI
FN1
SERTAD3
GTPBP3
IGLC1
SPEF2
PARD3
CAV1
SLK4
IGKV2D-40
IGKV3-20
CTR9
PNMA1
FCA
TULP3
SPTA1
EBI-1130097
SHZD2A
ANK1
SPTAN1
STAT1
EPB41
EBI-1179977
HPX
GLUD1
TF63
FLVCR1
LTN1
TNFRSF14
TNNT1
YAED1
GADD45G
TAB1
RASA1
NR4A1
MAP4K2
NLK
CDK14
NUAK1
MUC5AC
CDKN2A
PREP
CSK3B
REL
S100A8
PRKCI
OTUB1
HIST2H4B
DUX4
TRAF1
TERF1
RIPK2
SMG1
KNL1
BTBD2
HES1
TXN
MAPK13
IRF3
GAS7
S100A9
A2M
SNCA
ANXA7
HIT
ITGB5
TNFRSF1B
PSEN1
KARS
LAMTOR5
MTNR1A
RAD52
BTK
MAP3K3
TBK1
ERBB2
TK1
MAP3K14
GOS1
HSP90AB1
CHEBI:61048
ORC6
NEYC
WARS
RBM5
ACOT7
PCNA
DHRS2
RPA2
KCNE3
POT1
MED8
TNF2
MAPK1
ARL15
VAPB
EIF1B
ANXA1
PAFAH1B3
PCDHA4
YWHAE
ANKRD36B
SCK1
LAMA4
GSS
ARRB2
BID
GOT2
SLC25A37
KAT5
SMN2
MCL1
CDK17
PSMD11
SERPINB9
BLK
OSTF1
GADD45A
OSMR
ACTC1
PCK1
ERBB4
NOS2
EBI-1054063
PRDX1

Figure 6.5. Predicted protein networks associated with the proteins upregulated in the serum of the affected and unaffected twins. Cytoscape networks were constructed using **(A)** 19 proteins upregulated in the affected twin's serum (Table 6.3) and **(B)** 13 proteins upregulated in the unaffected twin's serum (Table 6.4). Nodes with a yellow outline indicate the seed proteins; other nodes indicate interacting proteins. Edges describe experimentally supported interactions. The superimposed GO terms associated with this network were identified using GOlorize with the BinGO plugin.

6.3 Discussion

6.3.1 The implications of raised GGT and ESR

Several factors have been linked to an increase in levels of serum GGT, including smoking, excessive alcohol consumption, and long-term treatment with the anticonvulsant phenobarbital (Whitfield et al., 2002; Lippi et al., 2008). In women over the age of 40, alcohol intake is further influenced by higher than average BMI (Danielsson et al., 2013). However, as neither twin drinks appreciable amounts of alcohol, the effect of alcohol consumption is unlikely to be significant. In addition, since both twins have smoked in the past, this factor is not considered to be of any significance, especially considering KG stopped smoking 7 years ago, and the unaffected twin (HG) continues to smoke. The drug history of both sisters has been carefully reviewed from medical records spanning more than 50 years, and only KG has taken anticonvulsants. However, her GGT values are higher than the average GGT values for patients taking the same dose of phenobarbital (Lippi et al., 2008). Phenytoin, another anticonvulsant taken by KG, has been reported to cause elevation in serum levels of GGT in children (Whitfield et al., 1972). However, the study in which this was documented (Whitfield et al., 1972), did not contain sufficient data to differentiate between the effects of phenytoin and barbiturate drugs such as phenobarbital, so this effect may be negligible, and once again the GGT value for KG is significantly higher than the average GGT values reported in that study.

Alterations in GGT even within the normal range are significantly correlated with ischemic stroke (Fraser et al., 2007). A genome-wide association study identified 26 loci associated with GGT levels (Chambers et al., 2011), and several of these loci, including *EPHA2*, *GCKR*, *MLX1PL*, *SOX9*, *NEDD4L* and *FUT2* have previously been reported to be involved in stroke based on other genetic studies.

GGT, an enzyme present on the surface of most cell types, is responsible for the catabolism of the antioxidant, glutathione, and the subsequent generation of reactive oxygen species (H_2O_2), thus acting as a marker for oxidative stress. It can also form complexes with various components in the blood, including circulating lipoproteins, which can carry it inside atheromatous plaques. Colocalisation of GGT with oxidised LDL and $CD68^+$ foam cells appear to be involved in the pathogenesis of atherosclerosis (Lippi et al., 2008; Franzini et al 2009), and GGT activity may also reflect endothelial function (Yu et al., 2007). Furthermore, GGT may also play a role in the proliferation and rupture of plaques (Franzini et al 2009), and it has been suggested that some of the circulating GGT may originate from such plaques (Whitfield et al., 2002). However, to what extent GGT is involved in processes leading to atherosclerosis, or to what extent it is a consequence of such pathologies, remains to be determined.

Raised ESR as a marker of poor prognosis following an ischemic stroke has been reported in previous studies (Nikanfar et al 2012; Swartz et al., 2005), but whether this is due to GGT elevation, or vice-versa, is not known. We believe GGT and ESR are useful biomarkers for ischemic stroke and GGT is a potential useful drug target. Screening patients at risk of ischemic stroke for GGT and ESR should be relatively inexpensive, and GGT inhibitors are readily available. Some of these have been shown to be effective in

reducing GGT and reducing the risk of cardiovascular events, although large scale randomised trials are needed to assess this more directly.

In summary, GGT has previously been implicated in stroke. However, it is known to be influenced by various genetic as well as environmental risk factors. We have been able to systematically exclude the main environmental risk factors, including smoking, alcohol intake, and certain medications. It was initially hypothesised that a de novo mutation in *GGT* could explain the stroke, but exome sequencing analysis of the twins excluded this. However, the possibility of a mutation elsewhere affecting *GGT* expression remains a possibility.

6.3.2 Label-free proteomic analysis

Plasma proteomes between MZ twins discordant for ischaemic stroke were compared to identify putative disease-associated markers in the affected twin. This approach provided a snapshot of proteins expressed in blood and differences in protein expression between the affected and unaffected twin. Proteomics has the potential to bridge the gap between genetics and biochemistry, identify functional biomarkers for complex conditions, and provide important insights into the mechanisms of disease. Like other high-throughput 'omics' technologies, proteomics can capture changes in thousands of variables simultaneously and has proved useful in biomarker research (Frantzi et al., 2014).

Candidate biomarkers found in this study have the potential to be used clinically for routine screening (diagnostic), guiding the right course of treatment (therapeutic), and/or predicting the outcome of the condition (prognostic).

Overall, 95 and 139 proteins in runs 1 and 2 were identified as differentially expressed and associated with GO terms. With a fold-difference average threshold of >1.5, 19

proteins were identified with higher levels in KG's serum in both runs, and 13 in HG's serum in both runs. Despite the differences in the proteins identified in each run, several GO terms were significantly enriched consistently within either the high or low protein lists, suggesting that they share common biological pathways. Although there are also GO terms which do not show consistency (Appendix B).

g:Profiler pathway analysis has identified an interesting enrichment of blood coagulation, wound healing, clot formation and platelet activation, all of which have previously been implicated in ischaemic stroke. This adds to our recently published study on a comparison of blood chemistries of this MZ twin pair, showing a >10- and >18-fold elevation of γ-glutamyltransferase (GGT) and erythrocyte sedimentation rate (ESR) levels respectively in the affected twin (Vadgama et al., 2015).

Moreover, the upregulation of fibulin 1 expression in the affected twin suggests that it is a novel candidate biomarker for ischaemic stroke. Fibulin 1 is a secreted calcium-binding glycoprotein (Argraves, 1990) that stabilises extracellular matrix integrity surrounding vascular smooth muscle. It is known to play important roles in wound repair (Liu et al., 2016) and mediates platelet adhesion by cross-linking with fibrinogen, forming part of a general mechanism by which platelets interact with exposed subendothelial matrices following vascular injury (Godyna et al., 1996).

GO terms associated with the proteins that were present at high levels in the unaffected twin (HG) were almost exclusively related to skin development. Interestingly, KG was diagnosed with seborrheic dermatitis capitis, whereas HG was diagnosed with atopic dermatitis. Atopic dermatitis is characterised by pruritus, erythema, induration, and scale (Siegfried and Hebert, 2015), and often affects the hands, feet, face, inside of the elbows and behind the knees. Although seborrheic dermatitis is another inflammatory skin

condition, this mostly affects sebum-rich skin, and in KG's case it was only affecting her scalp. In addition, seborrheic dermatitis is associated with overgrowth of fungi, typically Malassezia species, on sebum-rich skin (Siegfried and Hebert, 2015), whereas atopic dermatitis has two main pathological processes; namely, impaired skin barrier function and immune dysfunction (Oswald, 2017). Both disorders are characterised by spontaneous remissions and exacerbations, but atopic dermatitis has the potential to affect a larger area of skin than seborrheic dermatitis. The enrichment of 'skin development' GO terms associated with the serum of the unaffected twin suggests that HG was experiencing an episode of atopic dermatitis at the time the blood sample was taken.

Several of the proteins present at high levels in the affected twin (Table 6.3) have previously been identified as associated with ischaemic stroke. These include:

6.3.2.1 HBA, hemoglobin subunit alpha 1

Huang et al. (2009) found that hemoglobin α and β-chains were differentially expressed between stroke patients and controls ($P<0.0001$), with a sensitivity of 70.2% and the specificity of 85.3%. The authors conclude that serum free hemoglobin may serve as a potential biomarker for the diagnosis of acute ischaemic stroke. In our study, only the subunit alpha 1 was identified at high levels in the patient; however, this does support the suggestion that this protein could be a useful biomarker for ischaemic stroke.

6.3.2.2 A2M, alpha-2-macroglobulin

Alpha-2-macroglobulin is a cytokine transporter and protease inhibitor. It enhances procoagulant properties by the neutralisation of plasmin, plasminogen activators and metalloproteinases (Nezu et al., 2013). In addition, it functions as an inhibitor of fibrinolysis by inhibiting the activity of plasmin and kallikrein (Boer et al., 1993).

Interestingly, tissue kallikrein activity can be inhibited by kallistatin (SERPINA4), a serine proteinase inhibitor which was found to be at low levels in KG's serum (see below). Alpha-2-macroglobulin is also acts as an inhibitor of coagulation by inhibiting thrombin (Boer et al., 1993), which may, at least in part, explain why prothrombin (F2) was present at high levels in KG's serum. Nezu et al. (2013) identified alpha-2-macroglobulin as an acute ischaemic stroke biomarker, and Gori et al. (2017) found that alpha-2-macroglobulin was a factor significantly associated with poor outcome (death within three months), suggesting that it contributes to cerebral damage after ischaemic stroke and thrombolysis. The authors conclude that high levels of alpha-2-macroglobulin may be a useful prognostic biomarker to identify patients with an increased probability of death after the administration of tissue plasminogen activator (Gori et al., 2017).

6.3.2.3 HABP2, hyaluronan binding protein 2

Hyaluronan binding protein 2, also known as factor VII-activating protease, is a plasma serine protease that is predominantly produced in the liver (Römisch et al., 2000). Hyaluronan binding protein 2 was first thought to activate pro-urokinase plasminogen activator (PLAU) and coagulation factor VII (F7, Römisch et al., 2000). More recent studies have demonstrated that factor VII is in fact a weak substrate of hyaluronan-binding protein 2, pointing to additional substrates that have a more significant interaction (Hanson et al., 2012; Kanse et al., 2011). Hyaluronan binding protein 2 has been shown to inhibit tissue factor pathway inhibitor in clotting assays, and has been purported to also regulate the signalling functions of tissue factor (thromboplastin) (Kanse et al., 2011). This, and the fact that the enzyme has high homology with fibrinolytic and coagulation enzymes, demonstrates the role of hyaluronan binding protein 2 in haemostasis (Hanson et al., 2012).

Hyaluronan binding protein 2 is found to accumulate in lipid-rich areas within the necrotic core of atherosclerotic plaques, but not in normal vessels (Kannemeier et al., 2004); and a single nucleotide polymorphism in the gene, which attenuates its capacity to activate pro-urokinase, is associated with carotid stenosis (Willeit et al., 2003). In a more recent study, increased hyaluronan binding protein 2 antigen levels and activity were independently associated with all main aetiologic subtypes of ischaemic stroke, suggesting a potential role in the pathophysiology of the condition through prothrombotic mechanisms (Hanson et al., 2012).

6.3.2.4 S100A8, Protein S100-A8

Several studies have demonstrated a detrimental role of Toll-like receptors in cerebral ischaemia (Ziegler et al., 2007). S100A8 has been characterised as an endogenous ligand which activates TLR-signalling (Vogl et al., 2007), and has been shown to promote atherogenesis in mice (Schiopu and Cotoi, 2013). Thus, it has been postulated that upregulation and signalling of S100A8 contributes to neuroinflammation and the progression of ischaemic damage (Ziegler et al., 2007).

Wang et al. (2014) identified a molecular pathway of thrombosis that involves platelet S100A8, suggesting that targeting this protein has potential for treating atherothrombotic disorders, including myocardial infarction and stroke. Interestingly, blockers for the S100A8/A9 heterodimer (calprotectin) have been developed and are approved for clinical testing (Schiopu and Cotoi, 2013).

Several of the proteins that were at low levels in the affected twin could play an important role in the pathogenesis of ischaemic stroke. These include:

6.3.2.5 SERPINA4, serpin family A member 4 (kallistatin)

Kallistatin is a human serine proteinase inhibitor that binds strongly to tissue kallikrein, preventing the production of kinins (bradykinin and kallidin). These are potent vasodilators that are released by cleaving the kininogen substrate (Chao et al., 1996).

Animal models and studies involving human participants have shown that reduced levels of tissue kallikrein are associated with hypertension, restenosis, stroke, and cardiovascular and renal disease (Chao and Chao, 2005). The major function of tissue kallikrein is blood pressure regulation through the kinin B2 receptor (BDKRB2), ultimately providing a protective role against ischaemic stroke-induced injuries, and cardiovascular and renal dysfunction (Chao et al., 2006).

Kallistatin has also been found to have direct actions independent of its binding to tissue kallikrein, such as reduction of blood pressure, stimulation of neointima formation and inhibition of angiogenesis and tumour growth (Miao et al. 2000, 2002). Paradoxically, while kallistatin inhibits the liberation of kinins from the kininogens, thus preventing vasodilatation, it is itself a potent vasodilator. However, kallistatin-mediated vasodilation is thought to be unrelated to the tissue kallikrein-kinin system (Chao et al., 1997). Moreover, it has been shown that serum kallistatin levels are considerably reduced in hypertension, diabetes, cardiovascular and renal injury (Chao et al., 2015).

Collectively, the downregulation of kallistatin in KG is consistent with the literature and adds weight to its potential usefulness as a molecular biomarker for patients with cardiovascular disorders.

6.3.2.6 DSP, desmoplakin

Desmoplakin is an important component of desmosome structures in cardiac muscle and epidermal cells (Mueller and Franke, 1983; Stokes, 2007), and abnormalities in its expression results in cardiomyocyte death, changes in lipid metabolism, and defects in cardiac development (Yang et al. 2006). In a proteomics study set out to identify novel plasma biomarkers associated with vascular recurrence in post-stroke patients, it was found that patients taking statins for 3 months following the ischaemic stroke, and who did not experience a vascular recurrence during the follow-up period, had higher plasma levels of desmoplakin. This suggests that desmoplakin may be protective against a new vascular event, lending its usefulness as a prognostic biomarker (López-Farré et al., 2012). Decreased levels of plasma desmoplakin in KG could be a risk factor for the reoccurrence of vascular events. In addition, desmoglein 1 is also present at low levels in KG serum. Desmoplakin and desmoglein 1 associate as a heterodimer to form the desmosome, a key adhesion molecule, present at cell-cell junctions and is also associated with keratin filaments.

6.3.3 Limitations of approaches to identify proteomic biomarkers:

Proteomics has led to the identification of several serum biomarkers for a wide range of conditions, including cancer, autoimmune diseases, infectious diseases, stroke, schizophrenia, and bipolar disorder (Ray et al., 2011; O'Hanlon et al., 2011; Kazuno et al., 2013). Despite methodological advances, several biological and technological challenges remain for serum proteomics. These include the detection of a wide dynamic range of protein concentrations, low-abundance proteins being masked by high-abundance proteins (such as albumin and immunoglobulins), various interfering compounds such as salts, and not to mention, lack of reproducibility. Serum proteome

analysis can also be obfuscated by variations in storage, collection and handling procedures (Ray et al., 2011). In addition, there is great variability in protein expression between individuals, due to differences in lifestyle and dietary habits, among other factors. This introduces several confounding factors, which necessitates investigation of large sample sizes to exclude them.

The methodological approaches adopted in this study minimised some of the biological and technological limitations. MZ twins share identical genetic backgrounds and greater levels of consistency in their proteome, making them the ideal subjects for proteome analyses to identify differentially-expressed proteins.

This is the first study, to our knowledge, to apply a proteomics approach to analyse MZ twins discordant for stroke. The twins in this study have lived in the same environment since birth, which further reduces the interference of non-specific causal factors. However, a limitation of this approach is that the analysis only includes a single pair of MZ twins; thus, these results may not be applicable to stroke in general.

Further, protein levels were measured several years after the stroke, which may reflect changes associated with addressing the lesion (e.g. wound healing) and not necessarily predict the susceptibility to stroke. Despite the identification of a number proteins unique to each run, pathway analysis revealed that each list of differentially-expressed proteins includes representatives from common pathways. Both twins, despite having identical genomes, exhibited unique enriched GO terms. In KG, biological processes were predominantly related to wound healing and blood coagulation, whereas in HG processes relevant to epidermis development were enriched. Disturbances in epigenetic mechanisms can lead to changes in gene expression, which can in turn result in dysfunction or disease. Therefore, epigenetic factors may explain these differences.

6.3.4 Conclusion

The pathogenesis of ischaemic stroke is very complex, involving multiple molecular mechanisms (Xing et al., 2012). This makes it unlikely for a single biomarker to sufficiently reflect the underlying complexity. Thus, multiple biomarkers combined in a panel may be required to capture the dysregulation of several processes which converge to lead to susceptibility to an ischaemic event, thereby improving the diagnostic sensitivity and specificity.

Despite the above limitations, many of the candidate biomarkers and molecular pathways identified in our study are consistent with previous studies related to ischaemic stroke. The present findings also suggest future new targets that may be relevant to the pathology of ischaemic stroke.

The exact role of fibulin 1 in ischaemic stroke not clear. Further investigations into increased expression of fibulin 1 in stroke patients may add weight to its potential usefulness as a biomarker; and large-scale proteomic studies may further elucidate the common biochemical pathways and mechanisms associated with ischaemic stroke.

Future directions

Follow up from current work

This study identified somatic and germline de novo variants in a sample of MZ twins discordant for a range of complex disorders. Based on their biological functions and previous published data, the results are compatible with their potential role in the aetiology of the complex disorders investigated. This warrants independent validation using additional technologies such as Sanger sequencing, which will be considered as part of ongoing studies and its lack of herein is acknowledged as a limitation.

Computational analysis of CNVs using the already-generated exome sequencing data from all samples will be performed and compared with the microarray CNV analysis. CNVs that are confirmed by the two platforms will be put forward for experimental validation using ddPCR. The next step would be to determine the functional significance of validated CNVs and the impact of these variations on disease in larger patient cohorts.

Future studies would also involve validating the missense variant in *PPARGC1A* (c.A1829G) identified in the unaffected twin of the ALS-discordant twin pair (318), and determining whether the mutated form of the PGC-1a protein (p.H610R) affects expression at the RNA and protein level, and confirm in genetic studies with a larger patient cohort. Considering the role of this antioxidant defence transcriptional coactivator in ALS, targeting this pathway pharmacologically could be of therapeutic benefit.

Parent-offspring trio analysis revealed a rare, highly conserved de novo mutation in *RASD2* (c.G170A:p.R57H) in the ADHD-discordant twins, and an inherited stop loss mutation in *AADAC* (c.T1198C:p.X400Q) in twins discordant for Tourette's syndrome. Functional assays should be performed to elucidate the biological implications of these

missense mutations in cell culture and ultimately animal studies – including subcellular localisation of normal verses mutant protein, RNA and protein expression.

In addition to identifying discordant variants, rare concordant, potentially-damaging variants were also sought from the exome sequencing data. Biological processes associated with these shared variants in each twin pair will be identified with Gene Ontology classification using the functional analysis tool g:Profiler. Pathway analysis will determine the combined effect of estimated polygenic factors on the disorders investigated.

The ddPCR experiment was not able to unequivocally validate or refute the presence of the 138 kb CNV deletion in the twins with schizophrenia (RT1a and RT1b). This highly complex duplicated region on chromosome 15q13.2-13.3 likely prevented adequate probe specificity. This necessitates setting up robust genotyping assays capable of deducing both the sequence content and structure of this region. Unfortunately, functional studies assaying duplications are not well established. While CRISPR/Cas9 technology has facilitated knockouts within human induced pluripotent stem cells, the high degree of sequence identity between paralogs renders such a task challenging, and may even promote the formation of large structural rearrangements. Whole-genome sequencing has the potential to assay *ARHGAP11B* variation in disease cohorts. Although there are limitations in using short reads to assay variation between highly similar paralogs, recent advances in synthetic long read methods via barcoding (for example, 10X Genomics) may pave the way for improved variant calling and characterising the complexity of structural alterations in this region.

Future work

Our strategy of investigating MZ discordant twins collectively using genomic, proteomic and bioinformatics approaches was successful. More discordant MZ twins have since been recruited in the study and DNA samples have been obtained; these will be analysed using the strategies described in this thesis. This includes twins that are discordant for ulcerative colitis and Crohn's disease. For the twins discordant for inclusion body myositis, a muscle biopsy has been obtained, and future investigations will include performing multi-omic analyses to determine potential mechanisms that contribute to the discordant phenotype in those twins.

Transcriptome and proteome quantification will also be performed on other recruited twins using multiple tissues to provide novel insights into post-transcriptional gene regulation. Further, functional studies will be carried out for any interesting biological findings that emerge out of (epi)genetic or proteomic analyses. Initially, these findings will be confirmed using real-time PCR on RNA samples that we have already obtained.

The biological functions of these proteins can be further characterised by over-expressing the genes, or using siRNA knockdown approaches in cell-culture models. Although this is beyond the scope of the present study, these findings could lead to screening small molecules that could ultimately be of therapeutic benefit.

Conclusion

The comedian Steven Wright wryly remarked, 'a conclusion is the part where you got tired of thinking'. I am not sure if he's right this time. It is inevitable that this work will have raised more questions than it answers, and generated complex epicycles of

additional hypotheses that need to be followed though in much greater detail and to a greater depth.

The primary hypothesis of the present study was stimulated by several recent studies that showed de novo mutations and somatic mosaicism to have a significant impact on disease aetiology. Using the case co-twin design, with the assumption that twins may acquire de novo mutations during ontogeny that causally relate to their discordant phenotypes, we endeavoured to search for such variants. New technologies have enabled unprecedented capabilities for genetic research and demonstrate that somatic mosaicism might be more representative of the true range of de novo mutations observed in disease, since many mutations are lethal when constitutional. This fact may in future change the way we look at the human genome and cause us to reassess the genetic predisposition for many complex traits.

Although twenty-two discordant SNVs and indels were identified between six twin pairs in our cohort, a significant portion may turn out to be false positives based on independent validation. If we are to assume that the variants are real, and indeed deleterious, it would add to the literature that postzygotic mutations and somatic mosaicism do play a role in MZ twin discordance, and help explain disease onset in six out of the thirteen twin pairs. However, no structural differences were detected between all the twin pairs investigated, suggesting that CNVs may not be a significant factor in twin discordancy. Regardless, our results are in line with published literature, which suggests that the rate of post-fertilisation mutations is low. Yet we cannot exclude the possibility of having missed existing somatic mutations in the twin cohort examined, and the search for an explanation may lie beyond mere genetic changes at the level of the DNA sequence.

This study presents a theoretical framework for mosaicism detection and indicates potential weaknesses of existing variant detection tools. The false positive and false negative results identified throughout the literature and within this study emphasises the limitations of current NGS technology in identifying true discordant variants. Future research directions include systematic whole-genome, methylome, transcriptome and proteome analyses on discordant MZ twins, to identify novel disease-causing candidate genes and elucidate the role of differential gene expression in complex disorders.

References

Abdellaoui, A., Ehli, E., Hottenga, J., Weber, Z., Mbarek, H., Willemsen, G., van Beijsterveldt, T., Brooks, A., Hudziak, J., Sullivan, P., de Geus, E., Davies, G. and Boomsma, D. (2015). CNV Concordance in 1,097 MZ Twin Pairs. *Twin Research and Human Genetics*, 18(01), pp.1-12.

Abzug, M., Deterding, R., Hay, W. and Levin, M. (2014). *Current diagnosis & treatment pediatrics*. New York, N.Y.: McGraw-Hill Education LLC.

Acuna-Hidalgo, R., Bo, T., Kwint, M., van de Vorst, M., Pinelli, M., Veltman, J., Hoischen, A., Vissers, L. and Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *The American Journal of Human Genetics*, 97(1), pp.67-74.

Acuna-Hidalgo, R., Veltman, J. and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1).

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A. and Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), pp.248-249.

Al Mufti, W. and Bomse-Helmrich, O. (1979). Etude expérimentale de la surmaturité ovocytaire et ses conséquences chez les rongeurs. *Contracept Fertil Sex*, 7:845–847.

Albani, D., Pupillo, E., Bianchi, E., Chierchia, A., Martines, R., Forloni, G. and Beghi, E. (2016). The role of single-nucleotide variants of the energy metabolism-linked genes *SIRT3*, *PPARGC1A* and *APOE* in amyotrophic lateral sclerosis risk. *Genes & Genetic Systems*, 91(6), pp.301-309.

Al-Chalabi, A., Fang, F., Hanby, M., Leigh, P., Shaw, C., Ye, W. and Rijsdijk, F. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(12), pp.1324-1326.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Aparicio, S., Behjati, S., Biankin, A., Bignell, G., Bolli, N., Borg, A., Børresen-Dale, A., Boyault, S., Burkhardt, B., Butler, A., Caldas, C., Davies, H., Desmedt, C., Eils, R., Eyfjörd, J., Foekens, J., Greaves, M., Hosoda, F.,

Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D., Jones, D., Knappskog, S., Kool, M., Lakhani, S., López-Otín, C., Martin, S., Munshi, N., Nakamura, H., Northcott, P., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J., Puente, X., Raine, K., Ramakrishna, M., Richardson, A., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T., Span, P., Teague, J., Totoki, Y., Tutt, A., Valdés-Mas, R., van Buuren, M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S., Siebert, R., Campo, E., Shibata, T., Pfister, S., Campbell, P. and Stratton, M. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415-421.

Alikani, M., Noyes, N., Cohen, J. and Rosenwaks, Z. (1994). Fertilization and early embrology: Monozygotic twinning in the human is associated with the zona pellucida architecture. *Human Reproduction*, 9(7), pp.1318-1321.

Andrzej, P., Piotr, M., Borun, P., Skrzypczak-Zielinska, M., Wojciechowska-Lacka, A., Godlewski, D. and Banasiewicz, T. (2015). Influence of lactose intolerance on colorectal cancer incidence in the Polish population. *Hereditary Cancer in Clinical Practice*, 13(S1).

Antonacci, F., Dennis, M., Huddleston, J., Sudmant, P., Steinberg, K., Rosenfeld, J., Miroballo, M., Graves, T., Vives, L., Malig, M., Denman, L., Raja, A., Stuart, A., Tang, J., Munson, B., Shaffer, L., Amemiya, C., Wilson, R. and Eichler, E. (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nature Genetics*, 46(12), pp.1293-1302.

Aparicio, S. and Huntsman, D. (2009). Does massively parallel DNA resequencing signify the end of histopathology as we know it?. *The Journal of Pathology*, 39: 195-202

Argraves, W. (1990). Fibulin is an extracellular matrix and plasma glycoprotein with repeated domain structure. The Journal of Cell Biology, 111(6), pp.3155-3164.

Asbury, K., Dunn, J., Pike, A. and Plomin, R. (2003). Nonshared Environmental Influences on Individual Differences in Early Behavioral Development: A Monozygotic Twin Differences Study. *Child Development*, 74(3), pp.933-943.

Ascherio, A. (2013). Environmental factors in multiple sclerosis. *Expert Review of Neurotherapeutics*, 13(sup2), pp.3-9.

Augustine (2013). *De civitate dei*. Warminster: Aris & Phillips.

Austin, S. and St-Pierre, J. (2012). PGC1a and mitochondrial metabolism - emerging concepts and relevance in ageing and neurodegenerative disorders. *Journal of Cell Science*, 125(21), pp.4963-4971.

Ayalew, M., Le-Niculescu, H., Levey, D., Jain, N., Changala, B., Patel, S., Winiger, E., Breier, A., Shekhar, A., Amdur, R., Koller, D., Nurnberger, J., Corvin, A., Geyer, M., Tsuang, M., Salomon, D., Schork, N., Fanous, A., O'Donovan, M. and Niculescu, A. (2012). Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Molecular Psychiatry*, 17(9), pp.887-905.

Bamforth, F. and Machin, G. (2004). Why Zygosity of Multiple Births is not Always Obvious: An Examination of Zygosity Testing Requests From Twins or Their Parents. *Twin Research*, 7(05), pp.406-411.

Baranzini, S., Mudge, J., van Velkinburgh, J., Khankhanian, P., Khrebtukova, I., Miller, N., Zhang, L., Farmer, A., Bell, C., Kim, R., May, G., Woodward, J., Caillier, S., McElroy, J., Gomez, R., Pando, M., Clendenen, L., Ganusova, E., Schilkey, F., Ramaraj, T., Khan, O., Huntley, J., Luo, S., Kwok, P., Wu, T., Schroth, G., Oksenberg, J., Hauser, S. and Kingsmore, S. (2010). Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, 464(7293), pp.1351-1356.

Bayer, H., Lang, K., Buck, E., Higelin, J., Barteczko, L., Pasquarelli, N., Sprissler, J., Lucas, T., Holzmann, K., Demestre, M., Lindenberg, K., Danzer, K., Boeckers, T., Ludolph, A., Dupuis, L., Weydt, P. and Witting, A. (2017). ALS-causing mutations differentially affect PGC-1α expression and function in the brain vs. peripheral tissues. *Neurobiology of Disease*, 97, pp.36-45.

Beemer, F., France, H., Rosina-Angelista, I., Gerards, L., Cats, B. and Guyt, R. (2008). Familial partial monosomy 5p and trisomy 5q; three cases due to paternal pericentric inversion 5 (p151q333). *Clinical Genetics*, 26(3), pp.209-215.

Bell, J. and Spector, T. (2011). A twin approach to unraveling epigenetics. *Trends in Genetics*, 27(3), pp.116-125.

Bell, J., Loomis, A., Butcher, L., Gao, F., Zhang, B., Hyde, C., Sun, J., Wu, H., Ward, K., Harris, J., Scollen, S., Davies, M., Schalkwyk, L., Mill, J., Ahmadi, K., Ainali, C., Barrett, A., Bataille, V., Bell, J., Buil, A., Deloukas, P., Dermitzakis, E., Dimas, A., Durbin, R., Glass, D., Grundberg, E., Hassanali, N., Hedman, A., Ingle, C., Knowles, D., Krestyaninova, M., Lindgren, C., Lowe, C., McCarthy, M., Meduri, E., di Meglio, P., Min, J., Montgomery, S., Nestle, F., Nica, A., Nisbet, J., O'Rahilly, S., Parts, L., Potter, S., Sekowska, M., Shin, S., Small, K., Soranzo, N., Spector, T., Surdulescu, G., Travers, M., Tsaprouni, L., Tsoka, S., Wilk, A., Yang, T., Zondervan, K., Williams, F., Li, N., Deloukas, P., Beck, S., McMahon, S., Wang, J., John, S. and Spector, T. (2014). Differential methylation of the TRPA1 promoter in pain sensitivity. *Nature Communications*, 5.

Bertelsen, B., Stefánsson, H., Riff Jensen, L., Melchior, L., Mol Debes, N., Groth, C., Skov, L., Werge, T., Karagiannidis, I., Tarnok, Z., Barta, C., Nagy, P., Farkas, L., Brøndum-Nielsen, K., Rizzo, R., Gulisano, M., Rujescu, D., Kiemeney, L., Tosato, S., Nawaz, M., Ingason, A., Unnsteinsdottir, U., Steinberg, S., Ludvigsson, P., Stefansson, K., Kuss, A., Paschou, P., Cath, D., Hoekstra, P., Müller-Vahl, K., Stuhrmann, M., Silahtaroglu, A., Pfundt, R. and Tümer, Z. (2016). Association of AADAC Deletion and Gilles de la Tourette Syndrome in a Large European Cohort. *Biological Psychiatry*, 79(5), pp.383-391.

Blandau, R. and Young, W. (1939). The effects of delayed fertilization on the development of the guinea pig ovum. *American Journal of Anatomy*, 64(2), pp.303-329.

Blickstein, I., Verhoeven, H. and Keith, L. (1999). Zygotic Splitting after Assisted Reproduction. *New England Journal of Medicine*, 340(9), pp.738-739.

Bloom, R., Kähler, A., Collins, A., Chen, G., Cannon, T., Hultman, C. and Sullivan, P. (2013). Comprehensive analysis of copy number variation in monozygotic twins discordant for bipolar disorder or schizophrenia. *Schizophrenia Research*, 146(1-3), pp.289-290.

Boer, J. P., Creasey, A. A., Chang, A., Abbink, J. J., Roem, D., Eerenberg, A. J., Hack, C. E. & Taylor, F. B. (1993). Alpha-2-macroglobulin functions as an inhibitor of fibrinolytic, clotting, and neutrophilic proteinases in sepsis: studies using a baboon model. Infect. Immun., 61, pp. 5035–5043

Boklage, C. (2009). Traces of embryogenesis are the same in monozygotic and dizygotic twins: not compatible with double ovulation. *Human Reproduction*, 24(6), pp.1255-1266.

Bolea-Alamañac, B., Nutt, D., Adamou, M., Asherson, P., Bazire, S., Coghill, D., Heal, D., Müller, U., Nash, J., Santosh, P., Sayal, K., Sonuga-Barke, E. and Young, S. (2014). Evidence-based guidelines for the pharmacological management of attention deficit hyperactivity disorder: Update on recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, 28(3), pp.179-203.

Bonnevie, K. (1924). Studies on papillary patterns of human fingers. *Journal of Genetics*, 15(1), pp.1-111.

Bouhlal, Y., Martinez, S., Gong, H., Dumas, K. and Shieh, J. (2013). Twin mitochondrial sequence analysis. *Molecular Genetics & Genomic Medicine*, 1(3), pp.174-186.

Boukaftane Y., Khoris J., Moulard B., Salachas F., Meininger V., Malafosse A., Camu W., Rouleau G.A.. (1998). Identification of six novel SOD1 gene mutations in familial amyotrophic lateral sclerosis. *Canadian Journal of Neurological Sciences*, 25 (3), pp. 192-196

Breckpot, J., Thienpont, B., Gewillig, M., Allegaert, K., Vermeesch, J. and Devriendt, K. (2011). Differences in Copy Number Variation between Discordant Monozygotic Twins as a Model for Exploring Chromosomal Mosaicism in Congenital Heart Defects. *Molecular Syndromology*, 2(2), pp.81-87.

Brosens, E., Marsch, F., de Jong, E., Zaveri, H., Hilger, A., Choinitzki, V., Hölscher, A., Hoffmann, P., Herms, S., Boemers, T., Ure, B., Lacher, M., Ludwig, M., Eussen, B., van der Helm, R., Douben, H., Van Opstal, D., Wijnen, R., Beverloo, H., van Bever, Y., Brooks, A., IJsselstijn, H., Scott, D., Schumacher, J., Tibboel, D., Reutter, H. and de Klein, A. (2016). Copy number variations in 375 patients with oesophageal atresia and/or tracheoesophageal fistula. *European Journal of Human Genetics*, 24(12), pp.1715-1723.

Bruder, C., Piotrowski, A., Gijsbers, A., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E., Tiwari, H., Allison, D., Komorowski, J., van Ommen, G., Boomsma, D., Pedersen, N., den Dunnen, J., Wirdefeldt, K. and Dumanski, J. (2008). Phenotypically Concordant

and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *The American Journal of Human Genetics*, 82(3), pp.763-771.

Buck, E., Bayer, H., Lindenberg, K., Hanselmann, J., Pasquarelli, N., Ludolph, A., Weydt, P. and Witting, A. (2017). Comparison of Sirtuin 3 Levels in ALS and Huntington's Disease—Differential Effects in Human Tissue Samples vs. Transgenic Mouse Models. *Frontiers in Molecular Neuroscience*, 10.

Butcher, R. and Fugo, N. (1967). Overripeness and the Mammalian Ova. *Fertility and Sterility*, 18(3), pp.297-302.

Cai, L., Yuan, W., Zhang, Z., He, L. and Chou, K. (2016). In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, 6(1).

Calvo, S., Compton, A., Hershman, S., Lim, S., Lieber, D., Tucker, E., Laskowski, A., Garone, C., Liu, S., Jaffe, D., Christodoulou, J., Fletcher, J., Bruno, D., Goldblatt, J., DiMauro, S., Thorburn, D. and Mootha, V. (2012). Molecular Diagnosis of Infantile Mitochondrial Disease with Targeted Next-Generation Sequencing. *Science Translational Medicine*, 4(118), pp.118ra10-118ra10.

Campbell, C. and Eichler, E. (2013). Properties and rates of germline mutations in humans. *Trends in Genetics*, 29(10), pp.575-584.

Campbell, C., Chong, J., Malig, M., Ko, A., Dumont, B., Han, L., Vives, L., O'Roak, B., Sudmant, P., Shendure, J., Abney, M., Ober, C. and Eichler, E. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*, 44(11), pp.1277-1281.

Campbell, I., Shaw, C., Stankiewicz, P. and Lupski, J. (2015). Somatic mosaicism: implications for disease and transmission genetics. *Trends in Genetics*, 31(7), pp.382-392.

Cangemi, C., Skov, V., Poulsen, M., Funder, J., Twal, W., Gall, M., Hjortdal, V., Jespersen, M., Kruse, T., Aagard, J., Parving, H., Knudsen, S., Hoilund-Carlsen, P., Rossing, P., Henriksen, J., Argraves, W. and Rasmussen, L. (2011). Fibulin-1 Is a Marker for Arterial

Extracellular Matrix Alterations in Type 2 Diabetes. Clinical Chemistry, 57(11), pp.1556-1565.

Castellani, C., Melka, M., Wishart, A., Locke, M., Awamleh, Z., O'Reilly, R. and Singh, S. (2014). Biological relevance of CNV calling methods using familial relatedness including monozygotic twins. *BMC Bioinformatics*, 15(1), p.114.

Chaiyasap, P., Kulawonganunchai, S., Srichomthong, C., Tongsima, S., Suphapeetiporn, K. and Shotelersuk, V. (2014). Whole Genome and Exome Sequencing of Monozygotic Twins with Trisomy 21, Discordant for a Congenital Heart Defect and Epilepsy. *PLoS ONE*, 9(6), p.e100191.

Chambers, J., Zhang, W., Sehmi, J., Li, X., Wass, M., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S., Coin, L., Deng, G., Gieger, C., Heard-Costa, N., Hottenga, J., Kühnel, B., Kumar, V., Lagou, V., Liang, L., Luan, J., Vidal, P., Leach, I., O'Reilly, P., Peden, J., Rahmioglu, N., Soininen, P., Speliotes, E., Yuan, X., Thorleifsson, G., Alizadeh, B., Atwood, L., Borecki, I., Brown, M., Charoen, P., Cucca, F., Das, D., de Geus, E., Dixon, A., Döring, A., Ehret, G., Eyjolfsson, G., Farrall, M., Forouhi, N., Friedrich, N., Goessling, W., Gudbjartsson, D., Harris, T., Hartikainen, A., Heath, S., Hirschfield, G., Hofman, A., Homuth, G., Hyppönen, E., Janssen, H., Johnson, T., Kangas, A., Kema, I., Kühn, J., Lai, S., Lathrop, M., Lerch, M., Li, Y., Liang, T., Lin, J., Loos, R., Martin, N., Moffatt, M., Montgomery, G., Munroe, P., Musunuru, K., Nakamura, Y., O'Donnell, C., Olafsson, I., Penninx, B., Pouta, A., Prins, B., Prokopenko, I., Puls, R., Ruokonen, A., Savolainen, M., Schlessinger, D., Schouten, J., Seedorf, U., Sen-Chowdhry, S., Siminovitch, K., Smit, J., Spector, T., Tan, W., Teslovich, T., Tukiainen, T., Uitterlinden, A., Van der Klauw, M., Vasan, R., Wallace, C., Wallaschofski, H., Wichmann, H., Willemsen, G., Würtz, P., Xu, C., Yerges-Armstrong, L., Abecasis, G., Ahmadi, K., Boomsma, D., Caulfield, M., Cookson, W., van Duijn, C., Froguel, P., Matsuda, K., McCarthy, M., Meisinger, C., Mooser, V., Pietiläinen, K., Schumann, G., Snieder, H., Sternberg, M., Stolk, R., Thomas, H., Thorsteinsdottir, U., Uda, M., Waeber, G., Wareham, N., Waterworth, D., Watkins, H., Whitfield, J., Witteman, J., Wolffenbuttel, B., Fox, C., Ala-Korpela, M., Stefansson, K., Vollenweider, P., Völzke, H., Schadt, E., Scott, J., Järvelin, M., Elliott, P. and Kooner, J. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet*, 43(11), pp.1131-1138.

Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., et al. (2011) Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature Genet* 43 (11):1131-8.

Chao J, Bledsoe G, Chao L. (2015). Kallistatin: A novel biomarker for hypertension, organ injury and cancer. Austin Biomark & Diagn, 2(2)

Chao, J. and Chao, L. (2005). Kallikrein-kinin in stroke, cardiovascular and renal disease. Experimental Physiology, 90(3), pp.291-298.

Chao, J., Bledsoe, G., Yin, H. and Chao, L. (2006). The tissue kallikrein-kinin system protects against cardiovascular and renal diseases and ischemic stroke independently of blood pressure reduction. Biological Chemistry, 387(6).

Chao, J., Schmaier, A., Chen, L., Yang, Z. and Chao, L. (1996). Kallistatin, a novel human tissue kallikrein inhibitor: Levels in body fluids, blood cells, and tissues in health and disease. Journal of Laboratory and Clinical Medicine, 127(6), pp.612-620.

Chao, J., Stallone, J., Liang, Y., Chen, L., Wang, D. and Chao, L. (1997). Kallistatin is a potent new vasodilator. Journal of Clinical Investigation, 100(1), pp.11-17.

Chen, J., Calhoun, V., Perrone-Bizzozero, N., Pearlson, G., Sui, J., Du, Y. and Liu, J. (2016). A pilot study on commonality and specificity of copy number variants in schizophrenia and bipolar disorder. *Translational Psychiatry*, 6(5), p.e824.

Cibulskis, K., Lawrence, M., Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3), pp.213-219.

Ciregia, F., Giusti, L., Da Valle, Y., Donadio, E., Consensi, A., Giacomelli, C., Sernissi, F., Scarpellini, P., Maggi, F., Lucacchini, A. and Bazzichi, L. (2013). A multidisciplinary approach to study a couple of monozygotic twins discordant for the chronic fatigue syndrome: a focus on potential salivary biomarkers. *Journal of Translational Medicine*, 11(1), p.243.

Cirulli, E., Lasseigne, B., Petrovski, S., Sapp, P., Dion, P., Leblond, C., Couthouis, J., Lu, Y., Wang, Q., Krueger, B., Ren, Z., Keebler, J., Han, Y., Levy, S., Boone, B., Wimbish, J., Waite, L., Jones, A., Carulli, J., Day-Williams, A., Staropoli, J., Xin, W., Chesi, A., Raphael, A., McKenna-Yasek, D., Cady, J., Vianney de Jong, J., Kenna, K., Smith, B., Topp, S., Miller, J., Gkazi, A., Al-Chalabi, A., van den Berg, L., Veldink, J., Silani, V., Ticozzi, N., Shaw, C., Baloh, R., Appel, S., Simpson, E., Lagier-Tourenne, C., Pulst, S., Gibson, S., Trojanowski, J., Elman, L., McCluskey, L., Grossman, M., Shneider, N., Chung, W., Ravits, J., Glass, J., Sims, K., Van Deerlin, V., Maniatis, T., Hayes, S., Ordureau, A., Swarup, S., Landers, J., Baas, F., Allen, A., Bedlack, R., Harper, J., Gitler, A., Rouleau, G., Brown, R., Harms, M., Cooper, G., Harris, T., Myers, R. and Goldstein, D. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*.

Civelek, M., Wu, Y., Pan, C., Raulerson, C., Ko, A., He, A., Tilford, C., Saleem, N., Stančáková, A., Scott, L., Fuchsberger, C., Stringham, H., Jackson, A., Narisu, N., Chines, P., Small, K., Kuusisto, J., Parks, B., Pajukanta, P., Kirchgessner, T., Collins, F., Gargalovic, P., Boehnke, M., Laakso, M., Mohlke, K. and Lusis, A. (2017). Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. *The American Journal of Human Genetics*, 100(3), pp.428-443.

Concolino, D. (2002). Familial pericentric inversion of chromosome 5 in a family with benign neonatal convulsions. *Journal of Medical Genetics*, 39(3), pp.214-216.

Conlin, L., Thiel, B., Bonnemann, C., Medne, L., Ernst, L., Zackai, E., Deardorff, M., Krantz, I., Hakonarson, H. and Spinner, N. (2010). Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Human Molecular Genetics*, 19(7), pp.1263-1275.

Conrad, D., Keebler, J., DePristo, M., Lindsay, S., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C., Torroja, C., Garimella, K., Zilversmit, M., Cartwright, R., Rouleau, G., Daly, M., Stone, E., Hurles, M. and Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7), pp.712-714.

Conrad, D., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C., Kristiansson, K., MacArthur, D., MacDonald, J., Onyiah, I., Pang, A., Robson, S., Stirrups, K., Valsesia, A., Walter,

K., Wei, J., Tyler-Smith, C., Carter, N., Lee, C., Scherer, S. and Hurles, M. (2009). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289), pp.704-712.

Cordonnier, C. and Leys, D. (2008). Stroke: the bare essentials. Practical Neurology, 8(4), pp.263-272.

Cui, L., Jeong, H., Borovecki, F., Parkhurst, C., Tanese, N. and Krainc, D. (2006). Transcriptional Repression of PGC-1α by Mutant Huntingtin Leads to Mitochondrial Dysfunction and Neurodegeneration. *Cell*, 127(1), pp.59-69.

Czyz, W., Morahan, J., Ebers, G. and Ramagopalan, S. (2012). Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. *BMC Medicine*, 10(1).

Dal, G., Erguner, B., Sa ro lu, M., Yuksel, B., Onat, O., Alkan, C. and Ozcelik, T. (2014). Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *Journal of Medical Genetics*, 51(7), pp.455-459.

Dalgliesh, G., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C., Teague, J., Andrews, J., Barthorpe, S., Beare, D., Buck, G., Campbell, P., Forbes, S., Jia, M., Jones, D., Knott, H., Kok, C., Lau, K., Leroy, C., Lin, M., McBride, D., Maddison, M., Maguire, S., McLay, K., Menzies, A., Mironenko, T., Mulderrig, L., Mudie, L., O'Meara, S., Pleasance, E., Rajasingham, A., Shepherd, R., Smith, R., Stebbings, L., Stephens, P., Tang, G., Tarpey, P., Turrell, K., Dykema, K., Khoo, S., Petillo, D., Wondergem, B., Anema, J., Kahnoski, R., Teh, B., Stratton, M. and Futreal, P. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, 463(7279), pp.360-363.

Danielsson, J., Kangastupa, P., Laatikainen, T., Aalto, M. and Niemelä, O. (2013) Individual and Joint Impacts of Ethanol Use, BMI, Age and Gender on Serum Gamma-Glutamyltransferase Levels in Healthy Volunteers. *Int J Mol Sci* 14:11929-41.

Davydov, E., Goode, D., Sirota, M., Cooper, G., Sidow, A. and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, 6(12), p.e1001025.

DeJesus-Hernandez, M., Mackenzie, I., Boeve, B., Boxer, A., Baker, M., Rutherford, N., Nicholson, A., Finch, N., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G., Karydas, A., Seeley, W., Josephs, K., Coppola, G., Geschwind, D., Wszolek, Z., Feldman, H., Knopman, D., Petersen, R., Miller, B., Dickson, D., Boylan, K., Graff-Radford, N. and Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, 72(2), pp.245-256.

Dellefave L, Bangash MA, Siddique T. (2003). Pairwise concordance rates are similar in monozygotic and dizygotic twins for amyotrophic lateral sclerosis. *ALS and other motor neuron disorders*, 4(Suppl 1): 47–50

Dempster, E., Pidsley, R., Schalkwyk, L., Owens, S., Georgiades, A., Kane, F., Kalidindi, S., Picchioni, M., Kravariti, E., Toulopoulou, T., Murray, R. and Mill, J. (2011). Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Human Molecular Genetics*, 20(24), pp.4786-4796.

Denker, H. (2013). Comment on G. Herranz: The timing of monozygotic twinning: a criticism of the common model. Zygote (2013). *Zygote*, 23(02), pp.312-314.

Dennis, M. and Eichler, E. (2016). Human adaptation and evolution by segmental duplication. *Current Opinion in Genetics & Development*, 41, pp.44-52.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), pp.491-498.

Detjen, A., Tinschert, S., Kaufmann, D., Algermissen, B., Nürnberg, P. and Schuelke, M. (2007). Analysis of Mitochondrial DNA in Discordant Monozygotic Twins With Neurofibromatosis Type 1. *Twin Research and Human Genetics*, 10(03), pp.486-495.

DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S., Docking, T., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra, M., Hamelin, R., Bohlmann, J., Breuil, C. and Jones, S. (2009). De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology*, 10(9), p.R94.

DiMauro, S. and Schon, E. (2003). Mitochondrial Respiratory-Chain Diseases. *New England Journal of Medicine*, 348(26), pp.2656-2668.

Dols-Icardo, O., Garcia-Redondo, A., Rojas-Garcia, R., Sanchez-Valle, R., Noguera, A., Gomez-Tortosa, E., Pastor, P., Hernandez, I., Esteban-Perez, J., Suarez-Calvet, M., Anton-Aguirre, S., Amer, G., Ortega-Cubero, S., Blesa, R., Fortea, J., Alcolea, D., Capdevila, A., Antonell, A., Llado, A., Munoz-Blanco, J., Mora, J., Galan-Davila, L., Rodriguez De Rivera, F., Lleo, A. and Clarimon, J. (2013). Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia. *Human Molecular Genetics*, 23(3), pp.749-754.

Durbin, R., Altshuler, D., Durbin, R., Abecasis, G., Bentley, D., Chakravarti, A., Clark, A., Collins, F., De La Vega, F., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S., Gibbs, R., Knoppers, B., Lander, E., Lehrach, H., Mardis, E., McVean, G., Nickerson, D., Peltonen, L., Schafer, A., Sherry, S., Wang, J., Wilson, R., Gibbs, R., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Lander, E., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061-1073.

Ebstein, R., Monakhov, M., Lai, P. and Chew, S. (2014). Gene Expression and Human Personality Traits: Inverse Association with Novelty Seeking. *Messenger*, 3(1), pp.72-77.

Egan, E., Reidy, K., O'Brien, L., Erwin, R. and Umstad, M. (2013). The Outcome of Twin Pregnancies Discordant for Trisomy 21. *Twin Research and Human Genetics*, 17(01), pp.38-44.

Ehli, E., Abdellaoui, A., Hu, Y., Hottenga, J., Kattenberg, M., van Beijsterveldt, T., Bartels, M., Althoff, R., Xiao, X., Scheet, P., de Geus, E., Hudziak, J., Boomsma, D. and Davies, G. (2012). De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems. Eur J Hum Genet, 20(10), pp.1037-1043.

El-Hattab, A., Smolarek, T., Walker, M., Schorry, E., Immken, L., Patel, G., Abbott, M., Lanpher, B., Ou, Z., Kang, S., Patel, A., Scaglia, F., Lupski, J., Cheung, S. and Stankiewicz, P. (2009). Redefined genomic architecture in 15q24 directed by patient deletion/duplication breakpoint mapping. *Human Genetics*, 126(4), pp.589-602.

Eschbach, J., Schwalenstocker, B., Soyal, S., Bayer, H., Wiesner, D., Akimoto, C., Nilsson, A., Birve, A., Meyer, T., Dupuis, L., Danzer, K., Andersen, P., Witting, A., Ludolph, A., Patsch, W. and Weydt, P. (2013). PGC-1a is a male-specific disease modifier of human and experimental amyotrophic lateral sclerosis. *Human Molecular Genetics*, 22(17), pp.3477-3484.

Falkenberg, M., Larsson, N. and Gustafsson, C. (2007). DNA Replication and Transcription in Mammalian Mitochondria. *Annual Review of Biochemistry*, 76(1), pp.679-699.

Fallon, J., Reid, S., Kinyamu, R., Opole, I., Opole, R., Baratta, J., Korc, M., Endo, T., Duong, A., Nguyen, G., Karkehabadhi, M., Twardzik, D., Patel, S. and Loughlin, S. (2000). In vivo induction of massive proliferation, directed migration, and differentiation of neural cells in the adult mammalian brain. *Proceedings of the National Academy of Sciences*, 97(26), pp.14686-14691.

Farg, M., Sundaramoorthy, V., Sultana, J., Yang, S., Atkinson, R., Levina, V., Halloran, M., Gleeson, P., Blair, I., Soo, K., King, A. and Atkin, J. (2014). C9ORF72, implicated in amytrophic lateral sclerosis and frontotemporal dementia, regulates endosomal trafficking. Human Molecular Genetics, 23(13), pp.3579-3595.

Fernandez, T., Sanders, S., Yurkiewicz, I., Ercan-Sencicek, A., Kim, Y., Fishman, D., Raubeson, M., Song, Y., Yasuno, K., Ho, W., Bilguvar, K., Glessner, J., Chu, S., Leckman, J., King, R., Gilbert, D., Heiman, G., Tischfield, J., Hoekstra, P., Devlin, B., Hakonarson, H., Mane, S., Günel, M. and State, M. (2012). Rare Copy Number Variants in Tourette Syndrome Disrupt Genes in Histaminergic Pathways and Overlap with Autism. *Biological Psychiatry*, 71(5), pp.392-402.

Fizelova, M., Jauhiainen, R., Stančáková, A., Kuusisto, J. and Laakso, M. (2016). Finnish Diabetes Risk Score Is Associated with Impaired Insulin Secretion and Insulin Sensitivity, Drug-Treated Hypertension and Cardiovascular Disease: A Follow-Up Study of the METSIM Cohort. *PLOS ONE*, 11(11), p.e0166584.

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F., Peters, J., Guhr, E., Klemroth, S., Prufer, K., Kelso, J., Naumann, R., Nusslein, I., Dahl, A., Lachmann, R., Paabo, S. and Huttner, W. (2015). Human-specific

gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science*, 347(6229), pp.1465-1470.

Fogh, I., Rijsdijk, F., Andersen, P., Sham, P., Knight, J., Neale, B., McKenna-Yasek, D., Silani, V., Brown, R., Powell, J. and Al-Chalabi, A. (2007). Age at onset in sod1-mediated amyotrophic lateral sclerosis shows familiality. *Neurogenetics*, 8(3), pp.235-236.

Forget-Dubois, N., Pérusse, D., Turecki, G., Girard, A., Billette, J., Rouleau, G., Boivin, M., Malo, J. and Tremblay, R. (2003). Diagnosing Zygosity in Infant Twins: Physical Similarity, Genotyping, and Chorionicity. *Twin Research*, 6(6), pp.479-485.

Forsberg, L., Rasi, C., Razzaghian, H., Pakalapati, G., Waite, L., Thilbeault, K., Ronowicz, A., Wineinger, N., Tiwari, H., Boomsma, D., Westerman, M., Harris, J., Lyle, R., Essand, M., Eriksson, F., Assimes, T., Iribarren, C., Strachan, E., O'Hanlon, T., Rider, L., Miller, F., Giedraitis, V., Lannfelt, L., Ingelsson, M., Piotrowski, A., Pedersen, N., Absher, D. and Dumanski, J. (2012). Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells. *The American Journal of Human Genetics*, 90(2), pp.217-228.

Frantzi, M., Bhat, A. and Latosinska, A. (2014). Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clinical and Translational Medicine, 3(1), p.7.

Franzini M, Corti A, Martinelli B, Del Corso A, Emdin M, Parenti GF, et al. (2009) □-glutamyltransferase activity in human atherosclerotic plaques--biochemical similarities with the circulating enzyme. *Atherosclerosis* 202:119-27.

Fraser, A., Harris, R., Sattar, N., Ebrahim, S., Smith, G.D. and Lawlor, D.A. (2007) Gamma-Glutamyltransferase is associated with incident vascular events independently of alcohol intake: analysis of the british women's heart and health safety study and meta-analysis. *Arterioscler Thromb Vasc Biol* 27: 2729-2735.

Freed, D., Stevens, E. and Pevsner, J. (2014). Somatic Mosaicism in the Human Genome. *Genes*, 5(4), pp.1064-1094.

Freeman, J. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, 16(8), pp.949-961.

Freshney, R. (1987). *Culture of animal cells*. New York: A.R. Liss.

Friedberg, E. (2010). A comprehensive catalogue of somatic mutations in cancer genomes. *DNA Repair*, 9(4), pp.468-469.

Friedrich, D., Santos, S., Ribeiro-dos-Santos, Ã. and Hutz, M. (2013). Several Different Lactase Persistence Associated Alleles and High Diversity of the Lactase Gene in the Admixed Brazilian Population. *PLoS ONE*, 8(10).

Gajecka, M. (2015). Unrevealed mosaicism in the next-generation sequencing era. *Molecular Genetics and Genomics*, 291(2), pp.513-530.

Galton, F. (1876). The History of Twins, as a Criterion of the Relative Powers of Nature and Nurture. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 5, p.391.

Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B. and Aittokallio, T. (2006). GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, 23(3), pp.394-396.

Gazzellone, M., Zarrei, M., Burton, C., Walker, S., Uddin, M., Shaheen, S., Coste, J., Rajendram, R., Schachter, R., Colasanto, M., Hanna, G., Rosenberg, D., Soreni, N., Fitzgerald, K., Marshall, C., Buchanan, J., Merico, D., Arnold, P. and Scherer, S. (2016). Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. *Journal of Neurodevelopmental Disorders*, 8(1).

Gervin, K., Vigeland, M., Mattingsdal, M., Hammerø, M., Nygård, H., Olsen, A., Brandt, I., Harris, J., Undlien, D. and Lyle, R. (2012). DNA Methylation and Gene Expression Changes in Monozygotic Twins Discordant for Psoriasis: Identification of Epigenetically Dysregulated Genes. *PLoS Genetics*, 8(1), p.e1002454.

Ghosh, A., Jana, M., Modi, K., Gonzalez, F., Sims, K., Berry-Kravis, E. and Pahan, K. (2015). Activation of Peroxisome Proliferator-activated Receptor α Induces Lysosomal Biogenesis in Brain Cells. *Journal of Biological Chemistry*, 290(16), pp.10309-10324.

Godyna S, Diaz-Ricart M, Argraves WS (1996) Fibulin-1 mediates platelet adhesion via a bridge of fibrinogen. Blood 88: 2569–2577

González, J., Rodríguez-Santiago, B., Cáceres, A., Pique-Regi, R., Rothman, N., Chanock, S., Armengol, L. and Pérez-Jurado, L. (2011). A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics*, 12(1), p.166.

Goodart, S., Butler, M. and Overhauser, J. (1996). Familial double pericentric inversion of chromosome 5 with some features of cri-du-chat syndrome. *Human Genetics*, 97(6), pp.802-807.

Gori, A., Giusti, B., Piccardi, B., Nencini, P., Palumbo, V., Nesi, M., Nucera, A., Pracucci, G., Tonelli, P., Innocenti, E., Sereni, A., Sticchi, E., Toni, D., Bovi, P., Guidotti, M., Tola, M., Consoli, D., Micieli, G., Tassi, R., Orlandi, G., Sessa, M., Perini, F., Delodovici, M., Zedde, M., Massaro, F., Abbate, R. and Inzitari, D. (2017). Inflammatory and metalloproteinases profiles predict three-month poor outcomes in ischemic stroke treated with thrombolysis. Journal of Cerebral Blood Flow & Metabolism, pp.0271678X1769557.

Graham, A., Macdonald, A. and Hawkes, C. (1997). British motor neuron disease twin study. *Journal of Neurology, Neurosurgery & Psychiatry*, 62(6), pp.562-569.

Grams, S., Rand, L. and Norton, M. (2011). Complete isochromosome 5p in one fetus of a monochorionic twin pair. *Prenatal Diagnosis*, 31(6), pp.605-607.

Gringras, P. and Chen, W. (2001). Mechanisms for differences in monozygous twins. *Early Human Development*, 64(2), pp.105-117.

Gruhne, B., Kamranvar, S., Masucci, M. and Sompallae, R. (2009). EBV and genomic instability—A new look at the role of the virus in the pathogenesis of Burkitt's lymphoma. *Seminars in Cancer Biology*, 19(6), pp.394-400.

Hackmon, R., Jormark, S., Cheng, V., O'Reilly Green, C. and Divon, M. (2009). Monochorionic dizygotic twins in a spontaneous pregnancy: a rare case report. *The Journal of Maternal-Fetal & Neonatal Medicine*, 22(8), pp.708-710.

Halder, A., Jain, M., Chaudhary, I. and Varma, B. (2012). Chromosome 22q11.2 microdeletion in monozygotic twins with discordant phenotype and deletion size. *Molecular Cytogenetics*, 5(1), p.13.

Hales, C. and Barker, D. (2001). The thrifty phenotype hypothesis. *British Medical Bulletin*, 60(1), pp.5-20.

Hall, J. (2003). Twinning. *The Lancet*, 362(9385), pp.735-743.

Handschin, C., Kobayashi, Y., Chin, S., Seale, P., Campbell, K. and Spiegelman, B. (2007). PGC-1a regulates the neuromuscular junction program and ameliorates Duchenne muscular dystrophy. *Genes & Development*, 21(7), pp.770-783.

Handunnetthi, L., Handel, A. and Ramagopalan, S. (2010). Contribution of genetic, epigenetic and transcriptomic differences to twin discordance in multiple sclerosis. *Expert Review of Neurotherapeutics*, 10(9), pp.1379-1381.

Hanson, E., Kanse, S., Joshi, A., Jood, K., Nilsson, S., Blomstrand, C. And Jern, C. (2012). Plasma factor VII-activating protease antigen levels and activity are increased in ischemic stroke. Journal of Thrombosis and Haemostasis, 10(5), pp.848-856.

Hashimoto, R., Yoshida, M., Ozaki, N., Yamanouchi, Y., Iwata, N., Suzuki, T., Kitajima, T., Tatsumi, M., Kamijima, K. and Kunugi, H. (2005). A missense polymorphism (H204R) of a Rho GTPase-activating protein, the chimerin 2 gene, is associated with schizophrenia in men. *Schizophrenia Research*, 73(2-3), pp.383-385.

Hazkani-Covo, E., Zeller, R. and Martin, W. (2010). Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLoS Genetics*, 6(2), p.e1000834.

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S., Bauerfeind, A., Hummel, O., Lee, Y., Paskas, S., Rintisch, C., Saar, K., Cooper, J., Buchan, R., Gray, E., Cyster, J., Braund, P., Gracey, J., Krishnan, U., Moore, J., Nelson, C., Pollard, H., Attwood, T., Crisp-Hihn, A., Foad, N., Jolley, J., Lloyd-Jones, H., Muir, D., Murray, E., O'Leary, K., Rankin, A., Sambrook, J., Godfroy, T., Brocheton, J., Proust, C., Schmitz, G., Heimerl, S., Lugauer, I., Belz, S., Gulde, S., Linsel-Nitschke, P., Sager, H., Schroeder, L., Lundmark, P., Syvannen, A., Neudert, J., Scholz, M., Deloukas, P., Gray, E., Gwilliams, R., Niblett., D., Erdmann, J., Hengstenberg, C.,

Maouche, S., Ouwehand, W., Rice, C., Samani, N., Schunkert, H., Goodall, A., Schulz, H., Roider, H., Vingron, M., Blankenberg, S., Münzel, T., Zeller, T., Szymczak, S., Ziegler, A., Tiret, L., Smyth, D., Pravenec, M., Aitman, T., Cambien, F., Clayton, D., Todd, J., Hubner, N. and Cook, S. (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467(7314), pp.460-464.

Herranz, G. (2013). The timing of monozygotic twinning: a criticism of the common model. *Zygote*, 23(01), pp.27-40.

Heyn, H., Carmona, F., Gomez, A., Ferreira, H., Bell, J., Sayols, S., Ward, K., Stefansson, O., Moran, S., Sandoval, J., Eyfjord, J., Spector, T. and Esteller, M. (2012). DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. *Carcinogenesis*, 34(1), pp.102-108.

Higashida, H., Yokoyama, S., Huang, J., Liu, L., Ma, W., Akther, S., Higashida, C., Kikuchi, M., Minabe, Y. and Munesue, T. (2012). Social memory, amnesia, and autism: Brain oxytocin secretion is regulated by NAD+ metabolites and single nucleotide polymorphisms of CD38. *Neurochemistry International*, 61(6), pp.828-838.

Hintzsche, J., Robinson, W. and Tan, A. (2016). A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. *International Journal of Genomics*, 2016, pp.1-16.

Hirayasu, K. and Arase, H. (2015). Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. *Journal of Human Genetics*, 60(11), pp.703-708.

Hoy, J., Haeger, P., Constable, J., Arias, R., McCallum, R., Kyweriga, M., Davis, L., Schnell, E., Wehr, M., Castillo, P. and Washbourne, P. (2013). Neuroligin1 Drives Synaptic and Behavioral Maturation through Intracellular Interactions. *Journal of Neuroscience*, 33(22), pp.9364-9384.

Huang, H., Mullikin, J. and Hansen, N. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, 16(1).

Huang, P., Lo, L., Chen, Y., Lin, R., Shiea, J. and Liu, C. (2009). Serum free hemoglobin as a novel potential biomarker for acute ischemic stroke. Journal of Neurology, 256(4), pp.625-631.

Hübers, A., Marroquin, N., Schmoll, B., Vielhaber, S., Just, M., Mayer, B., Högel, J., Dorst, J., Mertens, T., Just, W., Aulitzky, A., Wais, V., Ludolph, A., Kubisch, C., Weishaupt, J. and Volk, A. (2014). Polymerase chain reaction and Southern blot-based analysis of the C9orf72 hexanucleotide repeat in different motor neuron diseases. *Neurobiology of Aging*, 35(5), pp.1214.e1-1214.e6.

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931-945.

Iafrate, A., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S. and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9), pp.949-951.

IJzerman, R., Stehouwer, C. and Boomsma, D. (2000). Evidence for Genetic Factors Explaining the Birth Weight-Blood Pressure Relation : Analysis in Twins. *Hypertension*, 36(6), pp.1008-1012.

Imelfort, M., Batley, J., Grimmond, S. and Edwards, D. (2009). Genome Sequencing Approaches and Successes. *Plant Genomics*, pp.345-358.

Ingram, C., Mulcare, C., Itan, Y., Thomas, M. and Swallow, D. (2008). Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics*, 124(6), pp.579-591.

Ionita-Laza, I., Rogers, A., Lange, C., Raby, B. and Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 93(1), pp.22-26.

Jickling, G. and Sharp, F. (2015). Biomarker Panels in Ischemic Stroke. Stroke, 46(3), pp.915-920.

Jin, M., Zhu, S., Hu, P., Liu, D., Li, Q., Li, Z., Zhang, X., Xie, Y. and Chen, X. (2014). Genomic and Epigenomic Analyses of Monozygotic Twins Discordant for Congenital Renal Agenesis. *American Journal of Kidney Diseases*, 64(1), pp.119-122.

Jones, B., Raga, T., Liebert, A., Zmarz, P., Bekele, E., Danielsen, E., Olsen, A., Bradman, N., Troelsen, J. and Swallow, D. (2013). Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. *The American Journal of Human Genetics*, 93(3), pp.538-544.

Kaminsky, Z., Tang, T., Wang, S., Ptak, C., Oh, G., Wong, A., Feldcamp, L., Virtanen, C., Halfvarson, J., Tysk, C., McRae, A., Visscher, P., Montgomery, G., Gottesman, I., Martin, N. and Petronis, A. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genetics*, 41(2), pp.240-245.

Kannemeier, C., Al-Fakhri, N., Preissner, KT., Kanse, SM. (2004). Factor VII activating protease (FSAP) inhibits growth factor-mediated cell proliferation and migration of vascular smooth muscle cells. The FASEB Journal, 18(6), pp.728-30

Kanse, S., Declerck, P., Ruf, W., Broze, G. and Etscheid, M. (2011). Factor VII-Activating Protease Promotes the Proteolysis and Inhibition of Tissue Factor Pathway Inhibitor. Arteriosclerosis, Thrombosis, and Vascular Biology, 32(2), pp.427-433.

Kaplan, L., Foster, R., Shen, Y., Parry, D., McMaster, M., O'Leary, M. and Gusella, J. (2010). Monozygotic twins discordant for neurofibromatosis 1. *Am. J. Med. Genet.*, 152A(3), pp.601-606.

Karolchik, D., Barber, G., Casper, J., Clawson, H., Cline, M., Diekhans, M., Dreszer, T., Fujita, P., Guruvadoo, L., Haeussler, M., Harte, R., Heitner, S., Hinrichs, A., Learned, K., Lee, B., Li, C., Raney, B., Rhead, B., Rosenbloom, K., Sloan, C., Speir, M., Zweig, A., Haussler, D., Kuhn, R. and Kent, W. (2013). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42(D1), pp.D764-D770.

Kazuno, A., Ohtawa, K., Otsuki, K., Usui, M., Sugawara, H., Okazaki, Y. and Kato, T. (2013). Proteomic Analysis of Lymphoblastoid Cells Derived from Monozygotic Twins Discordant for Bipolar Disorder: A Preliminary Study. PLoS ONE, 8(2), p.e53855.

Kendler, K. (1992). A Population-Based Twin Study of Major Depression in Women. *Archives of General Psychiatry*, 49(4), p.257.

Ketelaar, M., Hofstra, R. and Hayden, M. (2011). What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration. *Clinical Genetics*, 81(4), pp.325-333.

Kiernan, M., Vucic, S., Cheah, B., Turner, M., Eisen, A., Hardiman, O., Burrell, J. and Zoing, M. (2011). Amyotrophic lateral sclerosis. The Lancet, 377(9769), pp.942-955.

Kim-Cohen, J., Caspi, A., Moffitt, T., Harrington, H., Milne, B. and Poulton, R. (2003). Prior Juvenile Diagnoses in Adults With Mental Disorder. *Archives of General Psychiatry*, 60(7), p.709.

King, D., Sifrim, A., Fitzgerald, T., Rahbari, R., Hobson, E., Homfray, T., Mansour, S., Mehta, S., Shehla, M., Tomkins, S., Vasudevan, P. and Hurles, M. (2016). Detection of structural mosaicism from targeted and whole-genome sequencing data.

Kloosterman, W., Francioli, L., Hormozdiari, F., Marschall, T., Hehir-Kwa, J., Abdellaoui, A., Lameijer, E., Moed, M., Koval, V., Renkens, I., van Roosmalen, M., Arp, P., Karssen, L., Coe, B., Handsaker, R., Suchiman, E., Cuppen, E., Thung, D., McVey, M., Wendl, M., Uitterlinden, A., van Duijn, C., Swertz, M., Wijmenga, C., van Ommen, G., Slagboom, P., Boomsma, D., Schönhuth, A., Eichler, E., de Bakker, P., Ye, K. and Guryev, V. (2015). Characteristics of de novo structural changes in the human genome. *Genome Research*, 25(6), pp.792-801.

Kloss-Brandstätter, A., Weissensteiner, H., Erhart, G., Schäfer, G., Forer, L., Schönherr, S., Pacher, D., Seifarth, C., Stöckl, A., Fendt, L., Sottsas, I., Klocker, H., Huck, C., Rasse, M., Kronenberg, F. and Kloss, F. (2015). Validation of Next-Generation Sequencing of Entire Mitochondrial Genomes and the Diversity of Mitochondrial DNA Mutations in Oral Squamous Cell Carcinoma. *PLOS ONE*, 10(8), p.e0135643.

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H. and Koizumi, A. (2011). Expansion of Intronic GGCCTG Hexanucleotide Repeat in NOP56 Causes SCA36, a Type of Spinocerebellar Ataxia Accompanied by Motor Neuron Involvement. *The American Journal of Human Genetics*, 89(1), pp.121-130.

Kockan, C., Hach, F., Sarrafi, I., Bell, R., McConeghy, B., Beja, K., Haegert, A., Wyatt, A., Volik, S., Chi, K., Collins, C. and Sahinalp, S. (2016). SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics*, 33(1), pp.26-34.

Kondo, S., Schutte, B., Richardson, R., Bjork, B., Knight, A., Watanabe, Y., Howard, E., Ferreira de Lima, R., Daack-Hirsch, S., Sander, A., McDonald-McGinn, D., Zackai, E., Lammer, E., Aylsworth, A., Ardinger, H., Lidral, A., Pober, B., Moreno, L., Arcos-Burgos, M., Valencia, C., Houdayer, C., Bahuau, M., Moretti-Ferreira, D., Richieri-Costa, A., Dixon, M. and Murray, J. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature Genetics*, 32(2), pp.285-289.

Kong, A., Frigge, M., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W., Sigurdsson, G., Walters, G., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D., Helgason, A., Magnusson, O., Thorsteinsdottir, U. and Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412), pp.471-475.

Koressaar, T. and Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics*, 23(10), pp.1289-1291.

Kuhlenbäumer, G., Hullmann, J. and Appenzeller, S. (2011). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Human Mutation*, 32(2), pp.144-151.

Kumadaki, S., Matsuzaka, T., Kato, T., Yahagi, N., Yamamoto, T., Okada, S., Kobayashi, K., Takahashi, A., Yatoh, S., Suzuki, H., Yamada, N. and Shimano, H. (2008). Mouse Elovl-6 promoter is an SREBP target. *Biochemical and Biophysical Research Communications*, 368(2), pp.261-266.

Kumar, P., Henikoff, S. and Ng, P. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(8), pp.1073-1081.

Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A., Collins, F., Mohlke, K. and Boehnke, M. (2017). The Metabolic Syndrome in Men study: a

resource for studies of metabolic and cardiovascular diseases. *Journal of Lipid Research*, 58(3), pp.481-493.

Laplana, M., Royo, J., Aluja, A., López, R., Heine-Sunyer, D. and Fibla, J. (2014). Absence of Substantial Copy Number Differences in a Pair of Monozygotic Twins Discordant for Features of Autism Spectrum Disorder. *Case Reports in Genetics*, 2014, pp.1-9.

Larsson, H., Ryden, E., Boman, M., Langstrom, N., Lichtenstein, P. and Landen, M. (2013). Risk of bipolar disorder and schizophrenia in relatives of people with attention-deficit hyperactivity disorder. *The British Journal of Psychiatry*, 203(2), pp.103-106.

Lavedan, C., Licamele, L., Volpi, S., Hamilton, J., Heaton, C., Mack, K., Lannan, R., Thompson, A., Wolfgang, C. and Polymeropoulos, M. (2008). Association of the NPAS3 gene and five other loci with response to the antipsychotic iloperidone identified in a whole genome association study. *Molecular Psychiatry*, 14(8), pp.804-819.

Leblond, C., Heinrich, J., Delorme, R., Proepper, C., Betancur, C., Huguet, G., Konyukh, M., Chaste, P., Ey, E., Rastam, M., Anckarsäter, H., Nygren, G., Gillberg, I., Melke, J., Toro, R., Regnault, B., Fauchereau, F., Mercati, O., Lemière, N., Skuse, D., Poot, M., Holt, R., Monaco, A., Järvelä, I., Kantojärvi, K., Vanhala, R., Curran, S., Collier, D., Bolton, P., Chiocchetti, A., Klauck, S., Poustka, F., Freitag, C., Waltes, R., Kopp, M., Duketis, E., Bacchelli, E., Minopoli, F., Ruta, L., Battaglia, A., Mazzone, L., Maestrini, E., Sequeira, A., Oliveira, B., Vicente, A., Oliveira, G., Pinto, D., Scherer, S., Zelenika, D., Delepine, M., Lathrop, M., Bonneau, D., Guinchat, V., Devillard, F., Assouline, B., Mouren, M., Leboyer, M., Gillberg, C., Boeckers, T. and Bourgeron, T. (2012). Genetic and Functional Analyses of SHANK2 Mutations Suggest a Multiple Hit Model of Autism Spectrum Disorders. *PLoS Genetics*, 8(2), p.e1002521.

Lee, H. and Johnson, K. (2006). Fidelity of the Human Mitochondrial DNA Polymerase. *Journal of Biological Chemistry*, 281(47), pp.36236-36240.

Lee, H., Ma, H., Juanes, R., Tachibana, M., Sparman, M., Woodward, J., Ramsey, C., Xu, J., Kang, E., Amato, P., Mair, G., Steinborn, R. and Mitalipov, S. (2012). Rapid Mitochondrial DNA Segregation in Primate Preimplantation Embryos Precedes Somatic and Germline Bottleneck. *Cell Reports*, 1(5), pp.506-515.

Lee, W., Jiang, Z., Liu, J., Haverty, P., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K., Bhatt, D., Ha, C., Johnson, S., Kennemer, M., Mohan, S., Nazarenko, I., Watanabe, C., Sparks, A., Shames, D., Gentleman, R., de Sauvage, F., Stern, H., Pandita, A., Ballinger, D., Drmanac, R., Modrusan, Z., Seshagiri, S. and Zhang, Z. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297), pp.473-477.

Leem, J. and Koh, E. (2012). Interaction between Mitochondria and the Endoplasmic Reticulum: Implications for the Pathogenesis of Type 2 Diabetes Mellitus. *Experimental Diabetes Research*, 2012, pp.1-8.

Legault, M., Girard, S., Lemieux Perreault, L., Rouleau, G. and Dubé, M. (2015). Comparison of Sequencing Based CNV Discovery Methods Using Monozygotic Twin Quartets. *PLOS ONE*, 10(3), p.e0122287.

Lejeune J, Gautier M, Turpin R. (1959). Etude des chromosomes somatiques de neuf enfants mongoliens. [Study of somatic chromosomes from 9 mongoloid children]. *Comptes rendus Hebd des séances l'Académie des Sci*. 248:1721–2.

Lever, N., Nyström, K., Schindler, J., Halliday, J., Wira, C. and Funk, M. (2013). Missed Opportunities for Recognition of Ischemic Stroke in the Emergency Department. Journal of Emergency Nursing, 39(5), pp.434-439.

Levinson, D., Duan, J., Oh, S., Wang, K., Sanders, A., Shi, J., Zhang, N., Mowry, B., Olincy, A., Amin, F., Cloninger, C., Silverman, J., Buccola, N., Byerley, W., Black, D., Kendler, K., Freedman, R., Dudbridge, F., Pe'er, I., Hakonarson, H., Bergen, S., Fanous, A., Holmans, P. and Gejman, P. (2011). Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. *American Journal of Psychiatry*, 168(3), pp.302-316.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H., Bi, R., Fan, Y., Wu, Y., Tang, Y., Li, Z., He, Y., Zhou, J., Tang, J., Chen, X. and Yao, Y. (2016). mtDNA Heteroplasmy in Monozygotic Twins Discordant for Schizophrenia. *Molecular Neurobiology*, 54(6), pp.4343-4352.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Li, H., Ruan, J. and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), pp.1851-1858.

Li, R., Montpetit, A., Rousseau, M., Wu, S., Greenwood, C., Spector, T., Pollak, M., Polychronakos, C. and Richards, J. (2013). Somatic point mutations occurring early in development: a monozygotic twin study. *Journal of Medical Genetics*, 51(1), pp.28-34.

Liang, H., Ward, W., Jang, Y., Bhattacharya, A., Bokov, A., Li, Y., Jernigan, A., Richardson, A. and Van Remmen, H. (2011). PGC-1α protects neurons and alters disease progression in an amyotrophic lateral sclerosis mouse model. *Muscle & Nerve*, 44(6), pp.947-956.

Lin, J., Wu, P., Tarr, P., Lindenberg, K., St-Pierre, J., Zhang, C., Mootha, V., Jäger, S., Vianna, C., Reznick, R., Cui, L., Manieri, M., Donovan, M., Wu, Z., Cooper, M., Fan, M., Rohas, L., Zavacki, A., Cinti, S., Shulman, G., Lowell, B., Krainc, D. and Spiegelman, B. (2004). Defects in Adaptive Energy Metabolism with CNS-Linked Hyperactivity in PGC-1α Null Mice. *Cell*, 119(1), pp.121-135.

Lin, X., Tang, W., Ahmad, S., Lu, J., Colby, C., Zhu, J. and Yu, Q. (2012). Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. *Hearing Research*, 288(1-2), pp.67-76.

Lindhurst, M., Sapp, J., Teer, J., Johnston, J., Finn, E., Peters, K., Turner, J., Cannons, J., Bick, D., Blakemore, L., Blumhorst, C., Brockmann, K., Calder, P., Cherman, N., Deardorff, M., Everman, D., Golas, G., Greenstein, R., Kato, B., Keppler-Noreuil, K., Kuznetsov, S., Miyamoto, R., Newman, K., Ng, D., O'Brien, K., Rothenberg, S., Schwartzentruber, D., Singhal, V., Tirabosco, R., Upton, J., Wientroub, S., Zackai, E., Hoag, K., Whitewood-Neal, T., Robey, P., Schwartzberg, P., Darling, T., Tosi, L., Mullikin, J. and Biesecker, L. (2011). A Mosaic Activating Mutation in AKT1 Associated with the Proteus Syndrome. *New England Journal of Medicine*, 365(7), pp.611-619.

Lippi, G., Montagnana, M., Salvagno, G.L. and Guidi, G.C. (2008) Influence of stable, long-term treatment with phenobarbital on the activity of serum alanine aminotransferase and γ-glutamyltransferase. *Br J Biomed Sci* 65 (3):132-5.

Liu, G., Cooley, M., Jarnicki, A., Hsu, A., Nair, P., Haw, T., Fricker, M., Gellatly, S., Kim, R., Inman, M., Tjin, G., Wark, P., Walker, M., Horvat, J., Oliver, B., Argraves, W., Knight, D., Burgess, J. and Hansbro, P. (2016). Fibulin-1 regulates the pathogenesis of tissue remodeling in respiratory diseases. JCI Insight, 1(9).

Liu, X., Han, S., Wang, Z., Gelernter, J. and Yang, B. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, 8(9), p.e75619.

Liu, Y., Fann, C., Liu, C., Chen, W., Wu, J., Hung, S., Chen, C., Jou, Y., Liu, S., Hwang, T., Hsieh, M., Chang, C., Yang, W., Lin, J., Chou, F., Faraone, S., Tsuang, M. and Hwu, H. (2008). RASD2, MYH9, and CACNG2 Genes at Chromosome 22q12 Associated with the Subgroup of Schizophrenia with Non-Deficit in Sustained Attention and Executive Function. *Biological Psychiatry*, 64(9), pp.789-796.

Locke, D. (2004). BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *Journal of Medical Genetics*, 41(3), pp.175-182.

Loke, Y., Hannan, A. and Craig, J. (2015). The Role of Epigenetic Change in Autism Spectrum Disorders. *Frontiers in Neurology*, 6.

Londin, E., Keller, M., D'Andrea, M., Delgrosso, K., Ertel, A., Surrey, S. and Fortina, P. (2011). Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics*, 12(1).

López-Farré, A., Zamorano-León, J., Segura, A., Mateos-Cáceres, P., Modrego, J., Rodríguez-Sierra, P., Calatrava, L., Tamargo, J. and Macaya, C. (2012). Plasma desmoplakin I biomarker of vascular recurrence after ischemic stroke. Journal of Neurochemistry, 121(2), pp.314-325.

Machin, G. (2009a). Familial monozygotic twinning: A report of seven pedigrees. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 151C(2), pp.152-154.

Machin, G. (2009b). Non-identical monozygotic twins, intermediate twin types, zygosity testing, and the non-random nature of monozygotic twinning: A review. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 151C(2), pp.110-127.

Maere, S., Heymans, K. and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16), pp.3448-3449.

Magaard Koldby, K., Nygaard, M., Christensen, K. and Christiansen, L. (2016). Somatically acquired structural genetic differences: a longitudinal study of elderly Danish twins. *European Journal of Human Genetics*, 24(10), pp.1506-1510.

Magne, F., Serpa, R., Van Vliet, G., Samuels, M. and Deladoëy, J. (2014). Somatic Mutations Are Not Observed by Exome Sequencing of Lymphocyte DNA from Monozygotic Twins Discordant for Congenital Hypothyroidism due to Thyroid Dysgenesis. *Horm Res Paediatr*, 83(2), pp.79-85.

Magnusson, P., Lee, D., Chen, X., Szatkiewicz, J., Pramana, S., Teo, S., Sullivan, P., Feuk, L. and Pawitan, Y. (2016). One CNV Discordance in NRXN1 Observed Upon Genome-wide Screening in 38 Pairs of Adult Healthy Monozygotic Twins. *Twin Research and Human Genetics*, 19(02), pp.97-103.

Maiti, S., Kumar, K., Castellani, C., O'Reilly, R. and Singh, S. (2011). Ontogenetic De Novo Copy Number Variations (CNVs) as a Source of Genetic Individuality: Studies on Two Families with MZD Twins for Schizophrenia. *PLoS ONE*, 6(3), p.e17125.

Majounie, E., Renton, A., Mok, K., Dopper, E., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J., van Swieten, J., Abramzon, Y., Johnson, J., Sendtner, M., Pamphlett, R., Orrell, R., Mead, S., Sidle, K., Houlden, H., Rohrer, J., Morrison, K., Pall, H., Talbot, K., Ansorge, O., Hernandez, D., Arepalli, S., Sabatelli, M., Mora, G., Corbo, M., Giannini, F., Calvo, A., Englund, E., Borghero, G., Floris, G., Remes, A., Laaksovirta, H., McCluskey, L., Trojanowski, J., Van Deerlin, V., Schellenberg, G., Nalls, M., Drory, V., Lu, C., Yeh, T., Ishiura, H., Takahashi, Y., Tsuji, S., Le Ber, I., Brice, A., Drepper, C., Williams, N., Kirby, J., Shaw, P., Hardy, J., Tienari, P., Heutink, P., Morris, H., Pickering-Brown, S. and Traynor, B. (2012). Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *The Lancet Neurology*, 11(4), pp.323-330.

Mansilla, M., Kimani, J., Mitchell, L., Christensen, K., Boomsma, D., Daack-Hirsch, S., Nepomucena, B., Wyszynski, D., Felix, T., Martin, N. and Murray, J. (2005). Discordant MZ Twins With Cleft Lip and Palate: A Model for Identifying Genes in Complex Traits. *Twin Research and Human Genetics*, 8(01), pp.39-46.

Manuck, S. and McCaffery, J. (2014). Gene-Environment Interaction. *Annual Review of Psychology*, 65(1), pp.41-70.

Markus, H. (2011). Stroke genetics. *Human Molecular Genetics*, 20(R2), pp.R124-R131.

Marshall, C., Noor, A., Vincent, J., Lionel, A., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C., Vos, Y., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., Sloman, L., Summers, A., Gibbons, C., Teebi, A., Chitayat, D., Weksberg, R., Thompson, A., Vardy, C., Crosbie, V., Luscombe, S., Baatjes, R., Zwaigenbaum, L., Roberts, W., Fernandez, B., Szatmari, P. and Scherer, S. (2008). Structural Variation of Chromosomes in Autism Spectrum Disorder. *The American Journal of Human Genetics*, 82(2), pp.477-488.

Martin NG, Carr AB, Oakeshott JG, Clark P. (1982). Co-twin control studies: vitamin C and the common cold. *Prog Clin Biol Res*,103:365–73.

Matsuzaka, T. and Shimano, H. (2009). Elovl6: a new player in fatty acid metabolism and insulin sensitivity. *Journal of Molecular Medicine*, 87(4), pp.379-384.

Mattson, M., Gleichmann, M. and Cheng, A. (2008). Mitochondria in Neuroplasticity and Neurological Disorders. *Neuron*, 60(5), pp.748-766.

Maynard, T., Sikich, L., Lieberman, J. and LaMantia, A. (2001). Neural Development, Cell-Cell Signaling, and the "Two-Hit" Hypothesis of Schizophrenia. *Schizophrenia Bulletin*, 27(3), pp.457-476.

McCarroll, S., Kuruvilla, F., Korn, J., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M., de Bakker, P., Maller, J., Kirby, A., Elliott, A., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K., Rava, R., Daly, M., Gabriel, S. and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, 40(10), pp.1166-1174.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), pp.1297-1303.

McNamara, H., Kane, S., Craig, J., Short, R. and Umstad, M. (2016). A review of the mechanisms and evidence for typical and atypical twinning. *American Journal of Obstetrics and Gynecology*, 214(2), pp.172-191.

Meltz Steinberg, K., Nicholas, T., Koboldt, D., Yu, B., Mardis, E. and Pamphlett, R. (2015). Whole genome analyses reveal no pathogenetic single nucleotide or structural differences between monozygotic twins discordant for amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 16(5-6), pp.385-392.

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs*, 33(5), p.i-57.

Mi, H., Muruganujan, A. and Thomas, P. (2012). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1), pp.D377-D386.

Miao, R. (2002). Kallistatin is a new inhibitor of angiogenesis and tumor growth. Blood, 100(9), pp.3245-3252.

Miao, R., Murakami, H., Song, Q., Chao, L. and Chao, J. (2000). Kallistatin Stimulates Vascular Smooth Muscle Cell Proliferation and Migration In Vitro and Neointima Formation in Balloon-Injured Rat Artery. Circulation Research, 86(4), pp.418-424.

Monlong, J., Meloche, C., Rouleau, G., Cossette, P., Girard, S. and Bourque, G. (2015). Human copy number variants are enriched in regions of low-mappability. *bioRxiv* 034165

Morgan, S., Shoai, M., Fratta, P., Sidle, K., Orrell, R., Sweeney, M., Shatunov, A., Sproviero, W., Jones, A., Al-Chalabi, A., Malaspina, A., Houlden, H., Hardy, J. and Pittman, A. (2015). Investigation of next-generation sequencing technologies as a diagnostic tool for amyotrophic lateral sclerosis. *Neurobiology of Aging*, 36(3), pp.1600.e5-1600.e8.

Morimoto, Y., Ono, S., Imamura, A., Okazaki, Y., Kinoshita, A., Mishima, H., Nakane, H., Ozawa, H., Yoshiura, K. and Kurotaki, N. (2017). Deep sequencing reveals variations in somatic cell mosaic mutations between monozygotic twins with discordant psychiatric disease. *Human Genome Variation*, 4, p.17032.

Morissette, G. and Flamand, L. (2010). Herpesviruses and Chromosomal Integration. *Journal of Virology*, 84(23), pp.12100-12109.

Mueller, H. and Franke, W. (1983). Biochemical and immunological characterization of desmoplakins I and II, the major polypeptides of the desmosomal plaque. Journal of Molecular Biology, 163(4), pp.647-671.

Nezu, T., Hosomi, N., Aoki, S., Deguchi, K., Masugata, H., Ichihara, N., Ohyama, H., Ohtsuki, T., Kohno, M. and Matsumoto, M. (2013). Alpha2-macroglobulin as a promising biomarker for cerebral small vessel disease in acute ischemic stroke patients. Journal of Neurology, 260(10), pp.2642-2649.

Nickles, D., Madireddy, L., Yang, S., Khankhanian, P., Lincoln, S., Hauser, S., Oksenberg, J. and Baranzini, S. (2012). In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. *BMC Genomics*, 13(1), p.477.

Nielsen, R., Paul, J., Albrechtsen, A. and Song, Y. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, 12(6), pp.443-451.

Nieuwint, A., Van Zalen-Sprock, R., Hummel, P., Pals, G., Van Vugt, J., Van Der Harten, H., Heins, Y. and Madan, K. (1999). 'Identical' twins with discordant karyotypes. *Prenatal Diagnosis*, 19(1), pp.72-76.

Nikanfar, M., Shaafi, S., Hashemilar, M., Oskouli, D.S. and Goldust, M. (2012) Evaluating role of leukocytosis and high sedimentation rate as prognostic factors in acute ischemic cerebral strokes. *Pak. J. Biol. Sci.* 15 (8): 386-90.

O'Hanlon, T., Li, Z., Gan, L., Gourley, M., Rider, L. and Miller, F. (2011). Plasma proteomic profiles from disease-discordant monozygotic twins suggest that molecular pathways are shared in multiple systemic autoimmune diseases. Arthritis Research & Therapy, 13(6), p.R181.

O'Hanlon, T., Li, Z., Gan, L., Gourley, M., Rider, L. and Miller, F. (2011). Plasma proteomic profiles from disease-discordant monozygotic twins suggest that molecular pathways are shared in multiple systemic autoimmune diseases*. *Arthritis Research & Therapy*, 13(6), p.R181.

Ohi, K., Hashimoto, R., Nakazawa, T., Okada, T., Yasuda, Y., Yamamori, H., Fukumoto, M., Umeda-Yano, S., Iwase, M., Kazui, H., Yamamoto, T., Kano, M. and Takeda, M. (2012). The p250GAP Gene Is Associated with Risk for Schizophrenia and Schizotypal Personality Traits. *PLoS ONE*, 7(4), p.e35696.

Okabe, T., Nakamura, T., Nishimura, Y., Kohu, K., Ohwada, S., Morishita, Y. and Akiyama, T. (2003). RICS, a Novel GTPase-activating Protein for Cdc42 and Rac1, Is Involved in the β-Catenin-N-cadherin andN-Methyl-d-aspartate Receptor Signaling. *Journal of Biological Chemistry*, 278(11), pp.9920-9927.

Oksenberg, N. and Ahituv, N. (2013). The role of AUTS2 in neurodevelopment and human evolution. *Trends in Genetics*, 29(10), pp.600-608.

Olsen, J. (2005). Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. Molecular & Cellular Proteomics, 4(12), pp.2010-2021.

Ono, S., Imamura, A., Tasaki, S., Kurotaki, N., Ozawa, H., Yoshiura, K. and Okazaki, Y. (2010). Failure to Confirm CNVs as of Aetiological Significance in Twin Pairs Discordant for Schizophrenia. *Twin Research and Human Genetics*, 13(05), pp.455-460.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W., Wei, Z., Wang, K. and Lyon, G. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, 5(3), p.28.

Oswald, K. (2017). Atopic dermatitis: dupilumab and crisaborole could herald a new era in treatment. The Pharmaceutical Journal, 298(7898), pp.83-86.

Ottolini, B., Hornsby, M., Abujaber, R., MacArthur, J., Badge, R., Schwarzacher, T., Albertson, D., Bevins, C., Solnick, J. and Hollox, E. (2014). Evidence of Convergent Evolution in Humans and Macaques Supports an Adaptive Role for Copy Number

Variation of the β-Defensin-2 Gene. *Genome Biology and Evolution*, 6(11), pp.3025-3038.

Papiernik, E., Spira, A., Bomsel-Helmrich, O. and Lebel, S. (1979). Ovarian overripeness and intrauterine growth retardation. *The Lancet*, 314(8150), pp.1025-1026.

Parr, R., Maki, J., Reguly, B., Dakubo, G., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J. and Thayer, R. (2006). *BMC Genomics*, 7(1), p.185.

Pascual, J. and Rosenberg, R. (2015). *Rosenberg's molecular and genetic basis of neurological and psychiatric disease*. London, England: Academic Press.

Payen, C., Koszul, R., Dujon, B. and Fischer, G. (2008). Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms. *PLoS Genetics*, 4(9), p.e1000175.

Petersen, B., Spehlmann, M., Raedler, A., Stade, B., Thomsen, I., Rabionet, R., Rosenstiel, P., Schreiber, S. and Franke, A. (2014). Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics*, 15(1), p.564.

Petronis, A. (2006). Epigenetics and twins: three variations on the theme. *Trends in Genetics*, 22(7), pp.347-350.

Petronis, A., Gottesman, I., Kan, P., Kennedy, J., Basile, V., Paterson, A. and Popendikyte, V. (2003). Monozygotic Twins Exhibit Numerous Epigenetic Differences: Clues to Twin Discordance?. *Schizophrenia Bulletin*, 29(1), pp.169-178.

Petrovski, S., Wang, Q., Heinzen, E., Allen, A. and Goldstein, D. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, 9(8), p.e1003709.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., Vorstman, J., Thompson, A., Regan, R., Pilorge, M., Pellecchia, G., Pagnamenta, A., Oliveira, B., Marshall, C., Magalhaes, T., Lowe, J., Howe, J., Griswold, A., Gilbert, J., Duketis, E., Dombroski, B., De Jonge, M., Cuccaro, M., Crawford, E., Correia, C., Conroy, J., Conceição, I., Chiocchetti, A., Casey, J., Cai, G., Cabrol, C., Bolshakova, N., Bacchelli, E., Anney, R.,

Gallinger, S., Cotterchio, M., Casey, G., Zwaigenbaum, L., Wittemeyer, K., Wing, K., Wallace, S., van Engeland, H., Tryfon, A., Thomson, S., Soorya, L., Rogé, B., Roberts, et al. (2014). Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics*, 94(5), pp.677-694.

Pirooznia, M., Kramer, M., Parla, J., Goes, F., Potash, J., McCombie, W. and Zandi, P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, 8(1), p.14.

Plagnol, V., Curtis, J., Epstein, M., Mok, K., Stebbings, E., Grigoriadou, S., Wood, N., Hambleton, S., Burns, S., Thrasher, A., Kumararatne, D., Doffinger, R. and Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21), pp.2747-2754.

Pleasance, E., Stephens, P., O'Meara, S., McBride, D., Meynert, A., Jones, D., Lin, M., Beare, D., Lau, K., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H., Ordoñez, G., Mudie, L., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J., Mangion, J., Sun, Y., McLaughlin, S., Peckham, H., Tsung, E., Costa, G., Lee, C., Minna, J., Gazdar, A., Birney, E., Rhodes, M., McKernan, K., Stratton, M., Futreal, P. and Campbell, P. (2009). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278), pp.184-190.

Polderman, T., Benyamin, B., de Leeuw, C., Sullivan, P., van Bochoven, A., Visscher, P. and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), pp.702-709.

Poll, H. (1915). Über Vererbung beim Menschen. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 13(1), pp.299-299.

Potash, J., Zandi, P., Willour, V., Lan, T., Huo, Y., Avramopoulos, D., Shugart, Y., MacKinnon, D., Simpson, S., McMahon, F., DePaulo, J. and McInnis, M. (2003). Suggestive Linkage to Chromosomal Regions 13q31 and 22q12 in Families With Psychotic Bipolar Disorder. *American Journal of Psychiatry*, 160(4), pp.680-686.

Potkin, S., Turner, J., Fallon, J., Lakatos, A., Keator, D., Guffanti, G. and Macciardi, F. (2008). Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia. *Molecular Psychiatry*, 14(4), pp.416-428.

Poulsen, P., Esteller, M., Vaag, A. and Fraga, M. (2007). The Epigenetic Basis of Twin Discordance in Age-Related Diseases. *Pediatric Research*, 61(5 Part 2), pp.38R-42R.

Qi, Y., Yin, X., Wang, S., Jiang, H., Wang, X., Ren, M., Su, X., Lei, S. and Feng, H. (2015). PGC-1α Silencing Compounds the Perturbation of Mitochondrial Function Caused by Mutant SOD1 in Skeletal Muscle of ALS Mouse Model. *Frontiers in Aging Neuroscience*, 7.

Quinlan, A. and Hall, I. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841-842.

Raffel, L., Mohandas, T., Rimoin, D., Opitz, J. and Reynolds, J. (1986). Chromosomal mosaicism in the Killian/Teschler-Nicola syndrome. *American Journal of Medical Genetics*, 24(4), pp.607-611.

Raimundo, N. (2014). Mitochondrial pathology: stress signals from the energy factory. *Trends in Molecular Medicine*, 20(5), pp.282-292.

Ranciaro, A., Campbell, M., Hirbo, J., Ko, W., Froment, A., Anagnostou, P., Kotze, M., Ibrahim, M., Nyambo, T., Omar, S. and Tishkoff, S. (2014). Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics*, 94(4), pp.496-510.

Ray, S., Reddy, P., Jain, R., Gollapalli, K., Moiyadi, A. and Srivastava, S. (2011). Proteomic technologies for the identification of disease biomarkers in serum: Advances and challenges ahead. PROTEOMICS - Clinical Applications, 5(9-10), pp.559-559.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A., Chen, W., Cho, E., Dallaire, S., Freeman, J., González, J., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J., Marshall, C., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D., Estivill, X., Tyler-Smith, C., Carter, N., Aburatani, H., Lee, C., Jones, K., Scherer, S. and Hurles, M. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118), pp.444-454.

Reimand, J., Kull, M., Peterson, H., Hansen, J. and Vilo, J. (2007). g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Research, 35(Web Server), pp.W193-W200.

Rende, R., Plomin, R. and Vandenberg, S. (1990). Who discovered the twin method?. *Behavior Genetics*, 20(2), pp.277-285.

Renton, A., Chiò, A. and Traynor, B. (2013). State of play in amyotrophic lateral sclerosis genetics. Nature Neuroscience, 17(1), pp.17-23.

Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D. and Del-Favero, J. (2011). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology*, 30(1), pp.61-68.

Richardson, J. and Bult, C. (2015). Visual annotation display (VLAD): a tool for finding functional themes in lists of genes. *Mammalian Genome*, 26(9-10), pp.567-573.

Rio, E., Moseley, L., Purdam, C., Samiric, T., Kidgell, D., Pearce, A., Jaberzadeh, S. and Cook, J. (2013). The Pain of Tendinopathy: Physiological or Pathophysiological?. *Sports Medicine*, 44(1), pp.9-23.

Rio, M., Royer, G., Gobin, S., de Blois, M., Ozilou, C., Bernheim, A., Nizon, M., Munnich, A., Bonnefont, J., Romana, S., Vekemans, M., Turleau, C. and Malan, V. (2012). Monozygotic twins discordant for submicroscopic chromosomal anomalies in 2p25.3 region detected by array CGH. *Clin Genet*, 84(1), pp.31-36.

Roach, J., Glusman, G., Smit, A., Huff, C., Hubley, R., Shannon, P., Rowen, L., Pant, K., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L., Hood, L. and Galas, D. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, 328(5978), pp.636-639.

Robertson, S., Thompson, S., Morgan, T., Holder-Espinasse, M., Martinot-Duquenoy, V., Wilkie, A. and Manouvrier-Hanu, S. (2006). Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. *Eur J Hum Genet*, 14(5), pp.549-554.

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G. and Mesirov, J. (2011). Integrative genomics viewer. *Nat Biotechnol*, 29(1), pp.24-26.

Rodríguez-Santiago, B., Malats, N., Rothman, N., Armengol, L., Garcia-Closas, M., Kogevinas, M., Villa, O., Hutchinson, A., Earl, J., Marenne, G., Jacobs, K., Rico, D., Tardón, A., Carrato, A., Thomas, G., Valencia, A., Silverman, D., Real, F., Chanock, S. and Pérez-Jurado, L. (2010). Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome. *The American Journal of Human Genetics*, 87(1), pp.129-138.

Rohlin, A., Wernersson, J., Engwall, Y., Wiklund, L., Björk, J. and Nordling, M. (2009). Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. *Hum. Mutat.*, 30(6), pp.1012-1020.

Rollins, B., Martin, M., Sequeira, P., Moon, E., Morgan, L., Watson, S., Schatzberg, A., Akil, H., Myers, R., Jones, E., Wallace, D., Bunney, W. and Vawter, M. (2009). Mitochondrial Variants in Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *PLoS ONE*, 4(3), p.e4913.

Römisch, J., Vermöhlen, S., Feussner, A. and Stöhr, H. (2000). The FVII Activating Protease Cleaves Single-Chain Plasminogen Activators. Pathophysiology of Haemostasis and Thrombosis, 29(5), pp.292-299.

Rona-Voros, K. and Weydt, P. (2010). The Role of PGC-1α in the Pathogenesis of Neurodegenerative Disorders. *Current Drug Targets*, 999(999), pp.1-7.

Rygiel, K., Miller, J., Grady, J., Rocha, M., Taylor, R. and Turnbull, D. (2015). Mitochondrial and inflammatory changes in sporadic inclusion body myositis. *Neuropathology and Applied Neurobiology*, 41(3), pp.288-303.

Saito, R., Matsuzaka, T., Karasawa, T., Sekiya, M., Okada, N., Igarashi, M., Matsumori, R., Ishii, K., Nakagawa, Y., Iwasaki, H., Kobayashi, K., Yatoh, S., Takahashi, A., Sone, H., Suzuki, H., Yahagi, N., Yamada, N. and Shimano, H. (2011). Macrophage Elovl6 Deficiency Ameliorates Foam Cell Formation and Reduces Atherosclerosis in Low-Density Lipoprotein Receptor-Deficient Mice. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 31(9), pp.1973-1979.

Schiopu, A. and Cotoi, O. (2013). S100A8 and S100A9: DAMPs at the Crossroads between Innate Immunity, Traditional Risk Factors, and Cardiovascular Disease. Mediators of Inflammation, 2013, pp.1-10.

Schleinitz, N., Cognet, C., Guia, S., Laugier-Anfossi, F., Baratin, M., Pouget, J., Pelissier, J., Harle, J., Vivier, E. and Figarella-Branger, D. (2008). Expression of the CD85j (leukocyte Ig-like receptor 1, Ig-like transcript 2) receptor for class I major histocompatibility complex molecules in idiopathic inflammatory myopathies. *Arthritis & Rheumatism*, 58(10), pp.3216-3223.

Schuster, S. (2007). Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1), pp.16-18.

Schuster, S., Rivalan, M., Strauss, U., Stoenica, L., Trimbuch, T., Rademacher, N., Parthasarathy, S., Lajkó, D., Rosenmund, C., Shoichet, S., Winter, Y., Tarabykin, V. and Rosário, M. (2015). NOMA-GAP/ARHGAP33 regulates synapse development and autistic-like behavior in the mouse. *Molecular Psychiatry*, 20(9), pp.1120-1131.

Shimizu, C., Eleftherohorinou, H., Wright, V., Kim, J., Alphonse, M., Perry, J., Cimaz, R., Burgner, D., Dahdah, N., Hoang, L., Khor, C., Salgado, A., Tremoulet, A., Davila, S., Kuijpers, T., Hibberd, M., Johnson, T., Takahashi, A., Tsunoda, T., Kubo, M., Tanaka, T., Onouchi, Y., Yeung, R., Coin, L., Levin, M. and Burns, J. (2016). Genetic Variation in the SLC8A1 Calcium Signaling Pathway Is Associated With Susceptibility to Kawasaki Disease and Coronary Artery AbnormalitiesCLINICAL PERSPECTIVE. *Circulation: Cardiovascular Genetics*, 9(6), pp.559-568.

Shinawi, M., Patel, A., Panichkul, P., Zascavage, R., Peters, S. and Scaglia, F. (2009). The Xp contiguous deletion syndrome and autism. *American Journal of Medical Genetics Part A*, 149A(6), pp.1138-1148.

Siegfried, E. and Hebert, A. (2015). Diagnosis of Atopic Dermatitis: Mimics, Overlaps, and Complications. Journal of Clinical Medicine, 4(5), pp.884-917.

Siemens, H. (1924). Die zwillingspathologie. Berlin: J. Springer.

Simons, A., Shaffer, L. and Hastings, R. (2013). Cytogenetic Nomenclature: Changes in the ISCN 2013 Compared to the 2009 Edition. *Cytogenetic and Genome Research*, 141(1), pp.1-6.

Singh, S. (2002). Monozygotic twins with chromosome 22q11 deletion and discordant phenotypes: updates with an epigenetic hypothesis. *Journal of Medical Genetics*, 39(11), pp.71e-71.

Smith, A., Kilaru, V., Klengel, T., Mercer, K., Bradley, B., Conneely, K., Ressler, K. and Binder, E. (2014). DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am. J. Med. Genet.*, 168(1), pp.36-44.

Smoot, M., Ono, K., Ruscheinski, J., Wang, P. and Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3), pp.431-432.

Solomon, B., Bous, S., Bianconi, S. and Pineda-Alvarez, D. (2010). Consideration of VACTERL association in patients with trisomy 21. *Clinical Dysmorphology*, 19(4), pp.209-211.

Song, S., Wheeler, L. and Mathews, C. (2003). Deoxyribonucleotide Pool Imbalance Stimulates Deletions in HeLa Cell Mitochondrial DNA. *Journal of Biological Chemistry*, 278(45), pp.43893-43896.

Spencer, D., Tyagi, M., Vallania, F., Bredemeyer, A., Pfeifer, J., Mitra, R. and Duncavage, E. (2014). Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *The Journal of Molecular Diagnostics*, 16(1), pp.75-88.

Stamouli, S., Anderlid, B., Willfors, C., Thiruvahindrapuram, B., Wei, J., Berggren, S., Nordgren, A., Scherer, S., Lichtenstein, P., Tammimies, K. and Bolte, S. (2017). Copy Number Variation Analysis of 100 Twin Pairs Enriched for Neurodevelopmental Disorders.

Stead, L., Sutton, K., Taylor, G., Quirke, P. and Rabbitts, P. (2013). Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Human Mutation*, 34(10), pp.1432-1438.

Stefan, M., Zhang, W., Concepcion, E., Yi, Z. and Tomer, Y. (2014). DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. *Journal of Autoimmunity*, 50, pp.33-37.

Stewart, J. and Chinnery, P. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews Genetics*, 16(9), pp.530-542.

Stokes, D. (2007). Desmosomes from a structural perspective. Current Opinion in Cell Biology, 19(5), pp.565-571.

Stoll, G., Pietiläinen, O., Linder, B., Suvisaari, J., Brosi, C., Hennah, W., Leppä, V., Torniainen, M., Ripatti, S., Ala-Mello, S., Plöttner, O., Rehnström, K., Tuulio-Henriksson, A., Varilo, T., Tallila, J., Kristiansson, K., Isohanni, M., Kaprio, J., Eriksson, J., Raitakari, O., Lehtimäki, T., Jarvelin, M., Salomaa, V., Hurles, M., Stefansson, H., Peltonen, L., Sullivan, P., Paunio, T., Lönnqvist, J., Daly, M., Fischer, U., Freimer, N. and Palotie, A. (2013). Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. *Nature Neuroscience*, 16(9), pp.1228-1237.

Swartz, J.E., Jacobson, B.F., Connor, M.D., Bernstein, P.L. and Fritz, V.U. Erythrocyte sedimentation rate as a marker of inflammation and ongoing coagulation in stroke and transient ischaemic attack. (2005). *S Afr Med J*.95:607-612.

Tan, Q., Ohm Kyvik, K., Kruse, T. and Christensen, K. (2010). Dissecting complex phenotypes using the genomics of twins. *Functional & Integrative Genomics*, 10(3), pp.321-327.

Tang, J., Fan, Y., Li, H., Xiang, Q., Zhang, D., Li, Z., He, Y., Liao, Y., Wang, Y., He, F., Zhang, F., Shugart, Y., Liu, C., Tang, Y., Chan, R., Wang, C., Yao, Y. and Chen, X. (2017). Whole-genome sequencing of monozygotic twins discordant for schizophrenia indicates multiple genetic risk factors for schizophrenia. *Journal of Genetics and Genomics*, 44(6), pp.295-306.

Tanzi, R. and Bertram, L. (2005). Twenty Years of the Alzheimer's Disease Amyloid Hypothesis: A Genetic Perspective. *Cell*, 120(4), pp.545-555.

Taylor, R. and Turnbull, D. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics*, 6(5), pp.389-402.

The Lancet Neurology, (2013). Disparities in stroke: not just black and white. The Lancet Neurology, 12(7), p.623.

Tick, B., Bolton, P., Happé, F., Rutter, M. and Rijsdijk, F. (2015). Heritability of autism spectrum disorders: a meta-analysis of twin studies. *Journal of Child Psychology and Psychiatry*, 57(5), pp.585-595.

Tishkoff, S., Reed, F., Ranciaro, A., Voight, B., Babbitt, C., Silverman, J., Powell, K., Mortensen, H., Hirbo, J., Osman, M., Ibrahim, M., Omar, S., Lema, G., Nyambo, T., Ghori, J., Bumpstead, S., Pritchard, J., Wray, G. and Deloukas, P. (2006). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1), pp.31-40.

Turgut, O., Yilmaz, A., Yalta, K., Karadas, F., Yilmaz, M.B. (2006) γ-glutamyltransferase is a promising biomarker for cardiovascular risk. *Medical Hypothesis* 67: 1060-64.

Uchida, Y. (2011). The role of fatty acid elongation in epidermal structure and function. *Dermato-Endocrinology*, 3(2), pp.65-69.

Umstad, M., Short, R., Wilson, M. and Craig, J. (2012). Chimaeric twins: Why monochorionicity does not guarantee monozygosity. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 52(3), pp.305-307.

Vadgama, N., Gaze, D., Ranson, J., Hardy, J. and Nasir, J. (2015). Elevated γ-Glutamyltransferase and Erythrocyte Sedimentation Rate in Ischemic Stroke in Discordant Monozygotic Twin Study. International Journal of Stroke, 10(4), pp.E32-E33.

Vadgama, N., Nirmalananthan, N., Sadiq, M., Hardy, J. and Nasir, J. (2015). Identical non-identical twins and non-identical identical twins. *BMJ*, p.h6589.

Vadlamudi, L., Dibbens, L., Lawrence, K., Iona, X., McMahon, J., Murrell, W., Mackay-Sim, A., Scheffer, I. and Berkovic, S. (2010). Timing of De Novo Mutagenesis — A Twin

Study of Sodium-Channel Mutations. *New England Journal of Medicine*, 363(14), pp.1335-1340.

Veltman, J. and Brunner, H. (2012). De novo mutations in human genetic disease. *Nat Rev Genet*, 13(8), pp.565-575.

Vitucci, D., Di Giorgio, A., Napolitano, F., Pelosi, B., Blasi, G., Errico, F., Attrotto, M., Gelao, B., Fazio, L., Taurisano, P., Di Maio, A., Marsili, V., Pasqualetti, M., Bertolino, A. and Usiello, A. (2015). Rasd2 Modulates Prefronto-Striatal Phenotypes in Humans and 'Schizophrenia-Like Behaviors' in Mice. *Neuropsychopharmacology*, 41(3), pp.916-927.

Vogl, T., Tenbrock, K., Ludwig, S., Leukert, N., Ehrhardt, C., van Zoelen, M., Nacken, W., Foell, D., van der Poll, T., Sorg, C. and Roth, J. (2007). Mrp8 and Mrp14 are endogenous activators of Toll-like receptor 4, promoting lethal, endotoxin-induced shock. Nature Medicine, 13(9), pp.1042-1049.

Walker, H., Hall, W. and Hurst, J. (1990). *Clinical methods*. Boston: Butterworths.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., Hakonarson, H. and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11), pp.1665-1674.

Wang, L., Luo, H., Chen, X., Jiang, Y. and Huang, Q. (2014). Functional Characterization of S100A8 and S100A9 in Altering Monolayer Permeability of Human Umbilical Endothelial Cells. PLoS ONE, 9(3), p.e90472.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K., Pao, W. and Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine*, 5(10), p.91.

Warner, J., Barron, L., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D. and Brock, D. (1996). A general method for the detection of large CAG repeat expansions by fluorescent PCR. *Journal of Medical Genetics*, 33(12), pp.1022-1026.

Weinberg, W. (1901). Beiträge zur Physiologie und Pathologie der Mehrlingsgeburten beim Menschen. *Archiv für die Gesammte Physiologie des Menschen und der Thiere*, 88(6-8), pp.346-430.

Weischenfeldt, J., Symmons, O., Spitz, F. and Korbel, J. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2), pp.125-138.

Wells, A. (1999). EGF receptor. *The International Journal of Biochemistry & Cell Biology*, 31(6), pp.637-643.

Whitfield, J., Zhu, G., Nestler, J., Heath, A. and Martin, N. (2002). Genetic Covariation between Serum γ-Glutamyltransferase Activity and Cardiovascular Risk Factors. *Clinical Chemistry*, 48(9), pp.1426-1431.

Whitfield, J.B., Pounder, R.E., Neale, G. and Moss, D.W. (1972) Serum γ-glytamyl transpeptidase activity in liver disease. *Gut* 13:702-8.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A., Lieberenz, M., Savitski, M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J., Bantscheff, M., Gerstmair, A., Faerber, F. and Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp.582-587.

Willeit, J., Kiechl. S., Weimer. T., Mair. A., Santer. P., Wiedermann. CJ., Roemisch, J. (2003). Marburg I Polymorphism of Factor VII-Activating Protease: A Prominent Risk Predictor of Carotid Stenosis. Circulation, 107(5), pp.667-670.

Wishart, D., Jewison, T., Guo, A., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R. and Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1), pp.D801-D807.

Wong, A., Gottesman, I. and Petronis, A. (2005). Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics*, 14(suppl_1), pp.R11-R18.

Wong, E., So, H., Li, M., Wang, Q., Butler, A., Paul, B., Wu, H., Hui, T., Choi, S., So, M., Garcia-Barcelo, M., McAlonan, G., Chen, E., Cheung, E., Chan, R., Purcell, S., Cherny, S., Chen, R., Li, T. and Sham, P. (2013). Common Variants on Xq28 Conferring Risk of Schizophrenia in Han Chinese. *Schizophrenia Bulletin*, 40(4), pp.777-786.

Wong, K., deLeeuw, R., Dosanjh, N., Kimm, L., Cheng, Z., Horsman, D., MacAulay, C., Ng, R., Brown, C., Eichler, E. and Lam, W. (2007). A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *The American Journal of Human Genetics*, 80(1), pp.91-104.

Wrede, J., Mengel-From, J., Buchwald, D., Vitiello, M., Bamshad, M., Noonan, C., Christiansen, L., Christensen, K. and Watson, N. (2015). Mitochondrial DNA Copy Number in Sleep Duration Discordant Monozygotic Twins. *Sleep*, 38(10), pp.1655-1658.

Xi, Z., Yunusova, Y., van Blitterswijk, M., Dib, S., Ghani, M., Moreno, D., Sato, C., Liang, Y., Singleton, A., Robertson, J., Rademakers, R., Zinman, L. and Rogaeva, E. (2014). Identical twins with the C9orf72 repeat expansion are discordant for ALS. Neurology, 83(16), pp.1476-1478.

Xing, C., Arai, K., Lo, E. and Hommel, M. (2012). Pathophysiologic Cascades in Ischemic Stroke. International Journal of Stroke, 7(5), pp.378-385.

Xing, C., Torres-Caban, M., Wang, T., Lu, Q., Xing, G. and Elston, R. (2007). Linkage studies of catechol-O-methyltransferase (COMT) and dopamine-beta-hydroxylase (DBH) cDNA expression levels. *BMC Proceedings*, 1(Suppl 1), p.S95.

Xu, H., DiCarlo, J., Satya, R., Peng, Q. and Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, 15(1), p.244.

Yang, Z. (2006). Desmosomal Dysfunction due to Mutations in Desmoplakin Causes Arrhythmogenic Right Ventricular Dysplasia/Cardiomyopathy. Circulation Research, 99(6), pp.646-655.

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., Wang, J. and Shen, Y. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*, 10(1).

Yao, Y., Kajigaya, S. and Young, N. (2015). Mitochondrial DNA mutations in single human blood cells. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 779, pp.68-77.

Youngson, N. and Whitelaw, E. (2008). Transgenerational Epigenetic Effects. *Annual Review of Genomics and Human Genetics*, 9(1), pp.233-257.

Yu, C., Kastin, A.J., Ding, Y., Pan, W. (2007) Gamma glutamyl transpeptidase is a dynamic indicator of endothelial response to stroke. *Exp Neurol* 203 (1):116-22.

Zarrei, M., MacDonald, J., Merico, D. and Scherer, S. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3), pp.172-183.

Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E. and Liu, C. (2011). Accuracy of CNV Detection from GWAS Data. *PLoS ONE*, 6(1), p.e14511.

Zheng, J., Umikawa, M., Cui, C., Li, J., Chen, X., Zhang, C., Huynh, H., Kang, X., Silvany, R., Wan, X., Ye, J., Cantó, A., Chen, S., Wang, H., Ward, E. and Zhang, C. (2012). Inhibitory receptors bind ANGPTLs and support blood stem cells and leukaemia development. *Nature*, 485(7400), pp.656-660.

Ziegler, G., Harhausen, D., Schepers, C., Hoffmann, O., Röhr, C., Prinz, V., König, J., Lehrach, H., Nietfeld, W. and Trendelenburg, G. (2007). TLR2 has a detrimental role in mouse transient focal cerebral ischemia. Biochemical and Biophysical Research Communications, 359(3), pp.574-579.

Zierer, J., Menni, C., Kastenmüller, G. and Spector, T. (2015). Integration of 'omics' data in aging research: from biomarkers to systems biology. *Aging Cell*, 14(6), pp.933-944.

Zwijnenburg, P., Meijers-Heijboer, H. and Boomsma, D. (2010). Identical but not the same: The value of discordant monozygotic twins in genetic research. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153B(6), pp.1134-49.

Appendix A: Patient information sheet & consent form

TITLE: *Genetic investigation of identical twins by next-generation sequencing.*

The focus of our research is to determine the extent of genetic similarity between identical twins, and we would like to invite you to take part in our research study. Before making a decision, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and ask us if there is anything that is not clear, or if you would like more information.

PURPOSE

To make this research possible, we are gathering genetic material (DNA) from identical twins. Since you appear to meet this criterion, we are contacting you to see if you would be willing to contribute a saliva specimen for use in this research.

There is considerable evidence from medical research to suggest that identical twins can differ genetically. Because these are likely to involve complex genetic mechanisms, it presents novel challenges to determine the extent to which this occurs. **Your contribution could therefore be of major importance in the field of medical genetics.**

PROCEDURES

If you agree to participate, we will arrange to collect a small saliva sample (2 mL) from you. DNA will be extracted from the saliva samples and used for scientific research. DNA and clinical information collected from your GP will be stored at St. George's, University of London. This will include information about your age, sex, and clinical status. **Your DNA and clinical data will be stored in a coded way to keep your identity confidential.**

RISKS

There are no more than minimal medical risks associated with this research. These samples would be for research purposes only and will be stored indefinitely; no diagnostic genetic testing will be performed without your consent. It should be noted that this is research only and for that reason will not have any insurance implications.

CONFIDENTIALITY

All information which is collected during the course of the research will be kept strictly confidential. We will take the following steps to ensure confidentiality. A research number will be assigned to you, and your name will not be used. The only people who will have access to your individual identity are Dr Jamal Nasir and staff. **The results from the analysis of your DNA will not be released or shared in any way with your**

relatives, with insurance companies, or any third party not involved in research. When results of this study are published, your name will not be used. We are happy to update you with the outcome of the research if you wish to be informed.

PARTICIPATION IS VOLUNTARY

You do not have to contribute to this study if you don't want to. **You have the right to leave the study at any time without giving any reason, and without penalty.** If you wish to leave the study, please contact Dr. Jamal Nasir or Mr. Nirmal Vadgama.

WHO HAS REVIEWED THE STUDY?

This study has been reviewed by the Research Degrees Committee and other senior medical and academic staff at St. George's, University of London. It has also been reviewed by the Eleanor Peel Trust and St George's Hospital NHS Trust Charitable Foundation Medical Research Committee who are funding the research.

CONTACTS

If you have any questions about the study please contact:

Dr. Jamal Nasir
Senior Lecturer in Functional Genetics
Division of Biomedical Sciences (BMS)
Room 2.142C
Jenner Wing
St. George's University of London
Cranmer Terrace, London
SW17 0RE

Email: jnasir@sgul.ac.uk
Tel. 0208 725 1064
Fax 0208 725 1039

Nirmal Vadgama BSc(Hons), MRes
Division of Biomedical Sciences (BMS)
Room 2.140F
Jenner Wing
St. George's University of London
Cranmer Terrace, London
SW17 0RE

Email: nvadgama@sgul.ac.uk
Tel. 0207 737 2661
Mob. 0786 422 0709

CONSENT FORM

Title of Project*: Genetic investigation of identical twins by next-generation sequencing*

Please ✓box as appropriate

1) I confirm that I have read and understand the information sheet dated _____ for the above study and have had the opportunity to ask questions. ☐

2) It may not be possible to identify an alteration in these genes using current technology but it may be possible in the future. I am therefore happy for the sample to be stored for future testing. ☐

3) I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason, without my medical care or legal right being affected. ☐

4) I understand that sections of my medical notes may be looked at by responsible individuals from St. George's, University of London where it is relevant to my taking part in research. I give permission for these individuals to have access to my records. ☐

5) I would like to receive any results relevant to me. ☐

6) I agree to take part in the above study. ☐

You will get a copy of this consent form to keep once it is returned to us.

Signature of Subject..... Date.....

Name of Subject.....

Signature of Investigator..... Date.....

Name of Investigator.....

Appendix B: Raw data for proteomic analysis

Table 1: Differentially expressed proteins identified in serum of the affected (KG) or unaffected (HG) twin. List of proteins identified by proteomic analysis of serum from 2 individuals: KG (affected) run 1 and run 2 datasets; HG (unaffected) run 1 and run 2 datasets, used in the g:profiler analysis.

HG Run 1		HG Run 2	
UniProt ID	HGNC symbol	UniProt ID	HGNC symbol
A0M8Q6	IGLC7	A0M8Q6	IGLC7
P01597	IGKV1-39	B7U540	KCNJ18
P01611	IGKV1D-12	O15078	CEP290
P01617	IGKV2D-28	O43929	ORC4
P01619	IGKV3-20	P00751	CFB
P01700	IGLV1-47	P02787	TF
P01717	IGLV3-25	P02790	HPX
P01742	IGHV1-69	P03952	KLKB1
P01762	IGHV3-11	P04004	VTN
P01766	IGHV3-13	P04114	APOB
P01770	IGHV3-30	P04206	IGKV3-20
P01771	IGHV3-33	P04217	A1BG
P01773	IGHV3-30	P04264	KRT1
P01775	IGHV3-23	P04406	GAPDH
P01779	IGHV3-23	P04433	IGKV3-11
P01781	IGHV3-7	P06396	GSN
P01814	IGHV2-70	P06733	ENO1
P01861	IGHG4	P07225	PROS1
P01876	IGHA1	P08697	SERPINF2
P01877	IGHA2	P08779	KRT16
P02533	KRT14	P0C0L5	C4B
P02538	KRT6A	P0C0S5	H2AFZ
P02747	C1QC	P11277	SPTB
P02750	LRG1	P12259	F5
P02768	ALB	P13645	KRT10
P02790	HPX	P13671	C6
P04004	VTN	P13805	TNNT1
P04114	APOB	P15924	DSP
P04207	IGKV3-15	P18428	LBP
P04209	IGLV2-14	P19012	KRT15
P04259	KRT6B	P19013	KRT4
P04264	KRT1	P19823	ITIH2
P04406	GAPDH	P19827	ITIH1
P04430	IGKV1-16	P22792	CPN2
P04433	IGKV3-11	P23246	SFPQ
P05154	SERPINA5	P29622	SERPINA4
P05787	KRT8	P35443	THBS4
P07358	C8B	P35527	KRT9
P08697	SERPINF2	P35858	IGFALS
P08779	KRT16	P43652	AFM
P0C0L5	C4B	P62736	ACTA2

KG Run 1		KG Run 2	
UniProt ID	HGNC symbol	UniProt ID	HGNC symbol
O14791	APOL1	A6NJ16	IGHV4OR15-8
O43866	CD5L	O14791	APOL1
O75636	FCN3	O43866	CD5L
O75882	ATRN	O75122	CLASP2
P00450	CP	O75636	FCN3
P00734	F2	P00734	F2
P00739	HPR	P00736	C1R
P00747	PLG	P00738	HP
P00748	F12	P00739	HPR
P00915	CA1	P00742	F10
P01008	SERPINC1	P00747	PLG
P01023	A2M	P00915	CA1
P01024	C3	P00918	CA2
P01031	C5	P01009	SERPINA1
P01598	IGKV1-5	P01011	SERPINA3
P01599	IGKV1-17	P01019	AGT
P01602	IGKV1-5	P01023	A2M
P01603	IGKV1-33	P01024	C3
P01604	IGKV1-5	P01031	C5
P01613	IGKV1D-33	P01042	KNG1
P01620	IGKV3-20	P01591	JCHAIN
P01625	IGKV4-1	P01598	IGKV1-5
P01714	IGLV3-19	P01600	IGKV1-39
P01833	PIGR	P01602	IGKV1-5
P01834	IGKC	P01610	IGKV1-17
P01871	IGHM	P02774	GC
P02647	APOA1	P04003	C4BPA
P02649	APOE	P04208	IGLV1-47
P02671	FGA	P04211	IGLV7-43
P02741	CRP	P04220	BOT
P02743	APCS	P04275	VWF
P02751	FN1	P05090	APOD
P02788	LTF	P05109	S100A8
P04003	C4BPA	P05155	SERPING1
P04196	HRG	P05156	CFI
P04275	VWF	P05452	CLEC3B
P05109	S100A8	P05543	SERPINA7
P05155	SERPING1	P06309	IGKV2D-28
P05160	F13B	P06312	IGKV4-1
P05164	MPO	P06331	IGHV4-34
P06312	IGKV4-1	P06727	APOA4

P0CG06	IGLC2	P62805	HIST1H4A
P11277	SPTB	Q02413	DSG1
P13645	KRT10	Q04695	KRT17
P13647	KRT5	Q06033	ITIH3
P15169	CPN1	Q16610	ECM1
P15924	DSP	Q5D862	FLG2
P23083	IGHV1-2	Q5SW24	DACT2
P27918	CFP	Q5T749	KPRP
P29622	SERPINA4	Q5VTE0	EEF1A1P5
P30740	SERPINB1	Q66K66	TMEM198
P35527	KRT9	Q6NXT2	H3F3C
P35542	SAA4	Q6UWP8	SBSN
P35908	KRT2	Q86TI0	TBC1D1
P62736	ACTA2	Q86YZ3	HRNR
Q01546	KRT76	Q8NFH3	NUP43
Q02413	DSG1	Q8WXI4	ACOT11
Q04695	KRT17	Q96M42	LINC00479
Q7Z794	KRT77	Q96P31	FCRL3
Q86YZ3	HRNR	Q96PD5	PGLYRP2
Q99856	ARID3A	Q9C0H6	KLHL4
Q9UBP0	SPAST	Q9NQL9	DMRT3
Q9Y6D5	ARFGEF	Q9NR34	MAN1C1
		Q9NXC5	MIOS
		Q9NZT1	CALML5
		Q9ULV0	MYO5B

P06681	C2	P07360	C8G
P07996	THBS1	P08185	SERPINA6
P08571	CD14	P08571	CD14
P08603	CFH	P09871	C1S
P09871	C1S	P0CG04	IGLC1
P0CG04	IGLC1	P0CG05	IGLC3
P10643	C7	P0DJI8	SAA1
P10909	CLU	P10643	C7
P12035	KRT3	P10909	CLU
P18136	IGKV3-20	P18136	IGKV3-20
P20742	PZP	P19652	ORM2
P23142	FBLN1	P20742	PZP
P51884	LUM	P20851	C4BPB
P68871	HBB	P21399	ACO1
P69905	HBA1	P23142	FBLN1
P80108	GPLD1	P25311	AZGP1
Q03591	CFHR1	P27169	PON1
Q14520	HABP2	P36955	SERPINF1
Q5TH69	ARFGEF3	P43234	CTSO
Q96KV7	WDR90	P49913	CAMP
Q9UL18	AGO1	P51884	LUM
		P53367	ARFIP1
		P68871	HBB
		P69905	HBA1
		P80108	GPLD1
		P80188	LCN2
		P80748	IGLV3-21
		Q00325	SLC25A3
		Q02223	TNFRSF17
		Q03591	CFHR1
		Q08380	LGALS3BP
		Q14520	HABP2
		Q14624	ITIH4
		Q5T0N5	FNBP1L
		Q5T4S7	UBR4
		Q5VYK3	ECM29
		Q6P9F7	LRRC8B
		Q6VAB6	KSR2
		Q6ZT07	TBC1D9
		Q7Z2Z2	EFL1
		Q8TD19	NEK9
		Q8TF72	SHROOM3
		Q8WZ42	TTN
		Q92835	INPP5D
		Q99676	ZNF184
		Q9BZC7	ABCA2

Table 2: Enriched GO terms associated with the proteins at high levels in the serum of the affected (KG) or unaffected (HG) twin. GO terms were identified as significantly following g:Profiler analysis. P-values <0.05 are considered as significantly enriched. S = number of protein identifiers (IDs) in both the study dataset and GO term group, T = number of human protein IDs associated with the GO term, t = number of protein IDs in study or GO datasets. Rows highlighted in green are terms enriched in both KG (affected) datasets but not either of the HG (unaffected) datasets. Rows highlighted in orange are terms enriched in both HG (unaffected) datasets but not either of the KG (affected) datasets.

term ID	t type	t group	t name	t depth	T	KG - Run 1, t=50			KG - Run 2, t=77			HG - Run 1, t=45			HG - Run 2, t=62		
						p-value	t	S	p-value	t	S	p-value	t	S	p-value	t	S
GO:0045104	BP	4	intermediate filament cytoskeleton organization	1	40	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	1.45E-06	45	6	0.024	62	4
GO:0045103	BP	4	intermediate filament-based process	1	41	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	1.7E-06	45	6	0.0266	62	4
GO:0010876	BP	5	lipid localization	1	353	#N/A	#N/A	#N/A	0.0419	77	9	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0009892	BP	5	negative regulation of metabolic process	1	2477	#N/A	#N/A	#N/A	0.00265	77	27	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:1903027	BP	5	regulation of opsonization	1	2	#N/A	#N/A	#N/A	0.0492	77	2	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0010605	BP	5	negative regulation of macromolecule metabolic process	2	2286	#N/A	#N/A	#N/A	0.00755	77	25	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0050789	BP	5	regulation of biological process	2	10756	#N/A	#N/A	#N/A	0.0149	77	63	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043170	BP	5	macromolecule metabolic process	4	9000	#N/A	#N/A	#N/A	0.000825	77	59	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0009913	BP	6	epidermal cell differentiation	1	177	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.0116	45	6	#N/A	#N/A	#N/A
GO:0008544	BP	6	epidermis development	1	297	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	5.01E-06	45	10	0.0155	62	8
GO:0030216	BP	6	keratinocyte differentiation	2	127	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00167	45	6	#N/A	#N/A	#N/A
GO:0031424	BP	6	keratinization	3	50	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.0182	45	4	#N/A	#N/A	#N/A
GO:0042160	BP	8	lipoprotein modification	1	6	#N/A	#N/A	#N/A	0.00391	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0008152	BP	8	metabolic process	1	10975	0.0144	50	44	0.000746	77	66	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0034442	BP	8	regulation of lipoprotein oxidation	1	5	#N/A	#N/A	#N/A	0.00196	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0042161	BP	8	lipoprotein oxidation	2	6	#N/A	#N/A	#N/A	0.00391	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0050748	BP	8	negative regulation of lipoprotein metabolic process	3	7	#N/A	#N/A	#N/A	0.00682	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:1903318	BP	8	negative regulation of protein maturation	3	33	0.00415	50	4	0.0291	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0010955	BP	8	negative regulation of protein processing	4	33	0.00415	50	4	0.0291	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

GO:0043277	BP	8	apoptotic cell clearance	5	35	0.00529	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051179	BP	9	localization	1	5810	3.96E-05	50	35		6.37E-13	77	59		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:1902578	BP	9	single-organism localization	1	3031	4.95E-08	50	29		3.81E-11	77	42		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051234	BP	9	establishment of localization	2	4738	6.36E-07	50	34		7.3E-14	77	55		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0044765	BP	9	single-organism transport	2	2804	6.83E-09	50	29		2.31E-12	77	42		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0006909	BP	9	phagocytosis	3	264	3.93E-06	50	10		0.0377	77	8		0.00057	45	8		#N/A	#N/A	#N/A
GO:0046903	BP	9	secretion	3	1085	2.64E-05	50	16		1.01E-07	77	23		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0006810	BP	9	transport	3	4605	2.78E-07	50	34		1.85E-14	77	55		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0032940	BP	9	secretion by cell	4	952	0.000284	50	14		3.04E-06	77	20		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0016192	BP	9	vesicle-mediated transport	4	1492	7.52E-14	50	27		2.68E-15	77	35		7.31E-05	45	17		0.00251	62	18
GO:0006897	BP	9	endocytosis	5	655	1.16E-09	50	17		3.76E-08	77	19		8.98E-05	45	12		#N/A	#N/A	#N/A
GO:0006887	BP	9	exocytosis	5	405	0.000228	50	10		2.84E-06	77	14		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0006898	BP	9	receptor-mediated endocytosis	6	316	1.33E-06	50	11		7.52E-09	77	15		9.11E-06	45	10		#N/A	#N/A	#N/A
GO:0045055	BP	9	regulated exocytosis	6	285	8.21E-06	50	10		4.09E-07	77	13		#N/A	#N/A	#N/A		0.0115	62	8
GO:0002576	BP	9	platelet degranulation	7	106	4.47E-10	50	10		1.16E-12	77	13		#N/A	#N/A	#N/A		5.61E-06	62	8
GO:0007010	BP	12	cytoskeleton organization	1	1158	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.000119	45	15		#N/A	#N/A	#N/A
GO:0010035	BP	12	response to inorganic substance	1	469	#N/A	#N/A	#N/A		0.000164	77	13		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0045109	BP	12	intermediate filament organization	2	20	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		2.84E-06	45	5		#N/A	#N/A	#N/A
GO:0001816	BP	16	cytokine production	1	626	0.00146	50	11		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0002431	BP	16	Fc receptor mediated stimulatory signaling pathway	1	127	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00167	45	6		#N/A	#N/A	#N/A
GO:0038094	BP	16	Fc-gamma receptor signaling pathway	1	125	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00152	45	6		#N/A	#N/A	#N/A
GO:0050891	BP	16	multicellular organismal water homeostasis	1	60	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00423	62	5
GO:0030104	BP	16	water homeostasis	1	71	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00982	62	5
GO:0038096	BP	16	Fc-gamma receptor signaling pathway involved in phagocytosis	2	121	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00126	45	6		#N/A	#N/A	#N/A
GO:0002433	BP	16	immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	2	121	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00126	45	6		#N/A	#N/A	#N/A
GO:0050820	BP	16	positive regulation of coagulation	2	26	1.67E-05	50	5		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A

GO:1900048	BP	16	positive regulation of hemostasis	2	25	1.35E-05	50	5		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051240	BP	16	positive regulation of multicellular organismal process	2	1414	0.00595	50	15		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0001817	BP	16	regulation of cytokine production	2	568	0.00503	50	10		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0033561	BP	16	regulation of water loss via skin	2	20	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0424	45	3		0.00134	62	4
GO:0061436	BP	16	establishment of skin barrier	3	18	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0304	45	3		0.00085	62	4
GO:0030194	BP	16	positive regulation of blood coagulation	3	25	1.35E-05	50	5		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0090303	BP	16	positive regulation of wound healing	4	48	0.000415	50	5		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051918	BP	16	negative regulation of fibrinolysis	5	10	2.23E-05	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051649	BP	22	establishment of localization in cell	1	2109	#N/A	#N/A	#N/A		0.0232	77	23		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031589	BP	23	cell-substrate adhesion	1	307	0.00337	50	8		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0010810	BP	23	regulation of cell-substrate adhesion	1	177	0.00102	50	7		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050832	BP	29	defense response to fungus	1	33	0.00415	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0009620	BP	29	response to fungus	1	47	0.0176	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0065007	BP	38	biological regulation	1	11372	0.0101	50	45		0.0166	77	65		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050817	BP	38	coagulation	1	357	7.89E-10	50	14		6.87E-05	77	12		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0002376	BP	38	immune system process	1	2387	6.5E-15	50	33		2.83E-08	77	34		0.0123	45	18		#N/A	#N/A	#N/A
GO:0048519	BP	38	negative regulation of biological process	1	4572	0.016	50	27		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031324	BP	38	negative regulation of cellular metabolic process	1	2308	0.035	50	18		0.000636	77	27		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0048523	BP	38	negative regulation of cellular process	1	4251	0.0143	50	26		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0001869	BP	38	negative regulation of complement activation, lectin pathway	1	2	0.0167	50	2		0.0492	77	2		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0048518	BP	38	positive regulation of biological process	1	5194	8.8E-06	50	34		0.000584	77	43		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0019538	BP	38	protein metabolic process	1	5370	4.03E-06	50	35		1.25E-07	77	50		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0032268	BP	38	regulation of cellular protein metabolic process	1	2326	0.000384	50	21		0.0104	77	25		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051239	BP	38	regulation of multicellular organismal process	1	2631	0.00309	50	21		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0048583	BP	38	regulation of response to stimulus	1	3592	1.67E-10	50	34		0.0149	77	32		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050896	BP	38	response to stimulus	1	8162	1.8E-07	50	44		1.06E-05	77	59		0.0392	45	34		#N/A	#N/A	#N/A

GO:0007596	BP	38	blood coagulation	2	352	6.51E-10	50	14		5.88E-05	77	12		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0002252	BP	38	immune effector process	2	713	3.18E-10	50	18		1.36E-06	77	18		1.97E-06	45	14		#N/A	#N/A	#N/A
GO:0006955	BP	38	immune response	2	1530	1.02E-14	50	28		8.09E-08	77	27		0.000106	45	17		#N/A	#N/A	#N/A
GO:0043086	BP	38	negative regulation of catalytic activity	2	819	0.00037	50	13		0.003	77	15		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0032269	BP	38	negative regulation of cellular protein metabolic process	2	980	0.00285	50	13		0.000193	77	18		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050819	BP	38	negative regulation of coagulation	2	51	2.01E-09	50	8		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00186	62	5
GO:0002683	BP	38	negative regulation of immune system process	2	370	0.00122	50	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0051241	BP	38	negative regulation of multicellular organismal process	2	1001	0.00361	50	13		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0048585	BP	38	negative regulation of response to stimulus	2	1348	0.000082	50	17		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0002684	BP	38	positive regulation of immune system process	2	934	2.04E-10	50	20		3.02E-07	77	21		6.23E-05	45	14		#N/A	#N/A	#N/A
GO:0045862	BP	38	positive regulation of proteolysis	2	353	0.00949	50	8		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0048584	BP	38	positive regulation of response to stimulus	2	1994	1.05E-10	50	27		0.000134	77	26		0.0242	45	16		#N/A	#N/A	#N/A
GO:0072376	BP	38	protein activation cascade	2	131	1.42E-29	50	21		1.35E-26	77	22		9.38E-15	45	13		3.01E-05	62	8
GO:0051604	BP	38	protein maturation	2	266	9.52E-09	50	12		1.74E-07	77	13		#N/A	#N/A	#N/A		0.00688	62	8
GO:0006508	BP	38	proteolysis	2	1684	6.27E-16	50	30		1.45E-17	77	39		0.00262	45	16		#N/A	#N/A	#N/A
GO:0065008	BP	38	regulation of biological quality	2	3506	5.86E-05	50	27		0.000775	77	34		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050790	BP	38	regulation of catalytic activity	2	2328	0.00918	50	19		0.0105	77	25		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0050818	BP	38	regulation of coagulation	2	90	4.64E-09	50	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0316	62	5
GO:1900046	BP	38	regulation of hemostasis	2	85	2.73E-09	50	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0239	62	5
GO:0050776	BP	38	regulation of immune response	2	879	4.48E-12	50	21		7.56E-07	77	20		2.89E-05	45	14		#N/A	#N/A	#N/A
GO:0002682	BP	38	regulation of immune system process	2	1336	1.42E-10	50	23		6.06E-06	77	23		0.00502	45	14		#N/A	#N/A	#N/A
GO:0051246	BP	38	regulation of protein metabolic process	2	2489	7.4E-06	50	24		0.000197	77	29		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0032101	BP	38	regulation of response to external stimulus	2	703	9.37E-13	50	20		1.08E-06	77	18		#N/A	#N/A	#N/A		0.0454	62	11
GO:0080134	BP	38	regulation of response to stress	2	1272	5.76E-10	50	22		0.000396	77	20		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0009605	BP	38	response to external stimulus	2	2021	1.21E-08	50	25		8.31E-06	77	28		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0006950	BP	38	response to stress	2	3579	1.59E-11	50	35		1.28E-08	77	42		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A

GO:0002250	BP	38	adaptive immune response	3	390	8.04E-07	50	12	8.62E-10	77	17	2.59E-07	45	12	#N/A	#N/A	#N/A
GO:0072378	BP	38	blood coagulation, fibrin clot formation	3	26	5.18E-12	50	8	2.75E-08	77	7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0007597	BP	38	blood coagulation, intrinsic pathway	3	19	1.68E-08	50	6	3.05E-07	77	6	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0006956	BP	38	complement activation	3	108	4.34E-19	50	15	2.56E-19	77	17	6.9E-16	45	13	0.000175	62	7
GO:0006952	BP	38	defense response	3	1448	1.47E-16	50	29	8.72E-13	77	32	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0006959	BP	38	humoral immune response	3	223	1.77E-17	50	17	3.04E-15	77	18	2.74E-13	45	14	0.00184	62	8
GO:0045087	BP	38	innate immune response	3	791	1.85E-09	50	18	1.17E-07	77	20	0.0421	45	10	#N/A	#N/A	#N/A
GO:0002443	BP	38	leukocyte mediated immunity	3	332	2.23E-06	50	11	1.53E-08	77	15	8.11E-07	45	11	#N/A	#N/A	#N/A
GO:0030195	BP	38	negative regulation of blood coagulation	3	47	9.99E-10	50	8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00123	62	5
GO:1900047	BP	38	negative regulation of hemostasis	3	47	9.99E-10	50	8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00123	62	5
GO:0051346	BP	38	negative regulation of hydrolase activity	3	396	9.58E-07	50	12	2.12E-06	77	14	#N/A	#N/A	#N/A	0.0154	62	9
GO:0044092	BP	38	negative regulation of molecular function	3	1067	0.00734	50	13	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0051248	BP	38	negative regulation of protein metabolic process	3	1034	0.000781	50	14	1.26E-05	77	20	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0045861	BP	38	negative regulation of proteolysis	3	306	1.62E-05	50	10	1.64E-11	77	17	#N/A	#N/A	#N/A	0.00186	62	9
GO:0032102	BP	38	negative regulation of response to external stimulus	3	259	6.95E-09	50	12	0.00339	77	9	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:1903035	BP	38	negative regulation of response to wounding	3	68	3.38E-10	50	9	0.0259	77	5	#N/A	#N/A	#N/A	0.00792	62	5
GO:0030168	BP	38	platelet activation	3	163	0.011	50	6	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0050778	BP	38	positive regulation of immune response	3	664	5.67E-12	50	19	4.92E-09	77	20	7.78E-07	45	14	#N/A	#N/A	#N/A
GO:0032103	BP	38	positive regulation of response to external stimulus	3	264	0.0146	50	7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:1903036	BP	38	positive regulation of response to wounding	3	55	0.000831	50	5	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0016485	BP	38	protein processing	3	239	2.69E-09	50	12	4.56E-08	77	13	#N/A	#N/A	#N/A	0.0031	62	8
GO:0030193	BP	38	regulation of blood coagulation	3	85	2.73E-09	50	9	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.0239	62	5
GO:0050878	BP	38	regulation of body fluid levels	3	504	8.12E-08	50	14	0.00282	77	12	#N/A	#N/A	#N/A	0.00188	62	11
GO:0031347	BP	38	regulation of defense response	3	658	2.27E-07	50	15	0.000182	77	15	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0002920	BP	38	regulation of humoral immune response	3	51	2.01E-09	50	8	1.63E-09	77	9	0.000407	45	5	4.36E-05	62	6
GO:0051336	BP	38	regulation of hydrolase activity	3	1348	0.000541	50	16	0.00101	77	20	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0002697	BP	38	regulation of immune effector process	3	313	0.000299	50	9	0.016	77	9	#N/A	#N/A	#N/A	0.0228	62	8

GO:0050727	BP	38	regulation of inflammatory response	3	303	2.01E-09	50	13	6.25E-08	77	14	#N/A	#N/A	#N/A	0.00171	62	9
GO:2000257	BP	38	regulation of protein activation cascade	3	33	2.84E-13	50	9	2.19E-11	77	9	0.00334	45	4	0.000197	62	5
GO:1903317	BP	38	regulation of protein maturation	3	82	3.14E-11	50	10	3.77E-09	77	10	0.00447	45	5	7.05E-07	62	8
GO:0030162	BP	38	regulation of proteolysis	3	696	7.74E-13	50	20	6.61E-14	77	25	#N/A	#N/A	#N/A	0.000949	62	13
GO:1903034	BP	38	regulation of response to wounding	3	146	1.15E-08	50	10	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.023	62	6
GO:0009611	BP	38	response to wounding	3	644	5.59E-11	50	18	0.000138	77	15	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0002253	BP	38	activation of immune response	4	536	2.34E-12	50	18	9.27E-11	77	20	7.03E-07	45	13	#N/A	#N/A	#N/A
GO:0002460	BP	38	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	4	262	1.81E-07	50	11	5E-10	77	15	6.45E-08	45	11	#N/A	#N/A	#N/A
GO:0006957	BP	38	complement activation, alternative pathway	4	14	0.000105	50	4	0.000756	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0006958	BP	38	complement activation, classical pathway	4	84	5.84E-13	50	11	8.54E-16	77	14	1.58E-11	45	10	#N/A	#N/A	#N/A
GO:0001867	BP	38	complement activation, lectin pathway	4	9	0.00355	50	3	0.0163	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0042730	BP	38	fibrinolysis	4	27	1.27E-09	50	7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00477	62	4
GO:0007599	BP	38	hemostasis	4	357	7.89E-10	50	14	6.87E-05	77	12	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0006954	BP	38	inflammatory response	4	656	1.24E-14	50	21	2.8E-12	77	23	#N/A	#N/A	#N/A	0.00356	62	12
GO:0002449	BP	38	lymphocyte mediated immunity	4	262	1.81E-07	50	11	5E-10	77	15	6.45E-08	45	11	#N/A	#N/A	#N/A
GO:0002921	BP	38	negative regulation of humoral immune response	4	11	0.00695	50	3	0.000252	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0010466	BP	38	negative regulation of peptidase activity	4	243	0.000576	50	8	3.19E-09	77	14	#N/A	#N/A	#N/A	0.000266	62	9
GO:2000258	BP	38	negative regulation of protein activation cascade	4	8	0.00237	50	3	5.39E-05	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0061045	BP	38	negative regulation of wound healing	4	60	1.03E-10	50	9	0.0139	77	5	#N/A	#N/A	#N/A	0.00423	62	5
GO:0002673	BP	38	regulation of acute inflammatory response	4	72	8.04E-12	50	10	4.25E-08	77	9	#N/A	#N/A	#N/A	0.000358	62	6
GO:0052547	BP	38	regulation of peptidase activity	4	391	1.23E-05	50	11	1.23E-08	77	16	#N/A	#N/A	#N/A	0.00152	62	10
GO:0070613	BP	38	regulation of protein processing	4	81	2.77E-11	50	10	3.32E-09	77	10	0.00421	45	5	6.38E-07	62	8
GO:0061041	BP	38	regulation of wound healing	4	126	2.6E-09	50	10	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00983	62	6
GO:0042060	BP	38	wound healing	4	540	5.06E-11	50	17	1.32E-05	77	15	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0002526	BP	38	acute inflammatory response	5	136	1.63E-17	50	15	1.57E-17	77	17	0.00251	45	6	4.05E-05	62	8

GO:0019724	BP	38	B cell mediated immunity	5	155	5.93E-10	50	11	1.99E-13	77	15	2.06E-10	45	11	#N/A	#N/A	#N/A
GO:0045916	BP	38	negative regulation of complement activation	5	8	0.00237	50	3	5.39E-05	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0010951	BP	38	negative regulation of endopeptidase activity	5	233	0.000418	50	8	1.8E-09	77	14	#N/A	#N/A	#N/A	0.00256	62	8
GO:0030449	BP	38	regulation of complement activation	5	32	3.44E-11	50	8	1.6E-11	77	9	0.00294	45	4	0.000168	62	5
GO:0052548	BP	38	regulation of endopeptidase activity	5	369	0.0012	50	9	6.84E-08	77	15	#N/A	#N/A	#N/A	0.00869	62	9
GO:0051917	BP	38	regulation of fibrinolysis	5	14	5.19E-07	50	5	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0006953	BP	38	acute-phase response	6	48	0.0192	50	4	2.86E-06	77	7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0016064	BP	38	immunoglobulin mediated immune response	6	153	5.14E-10	50	11	1.63E-13	77	15	1.79E-10	45	11	#N/A	#N/A	#N/A
GO:0001868	BP	38	regulation of complement activation, lectin pathway	6	2	0.0167	50	2	0.0492	77	2	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0002455	BP	38	humoral immune response mediated by circulating immunoglobulin	7	94	2.13E-12	50	11	4.55E-15	77	14	7.3E-13	45	11	#N/A	#N/A	#N/A
GO:2000379	BP	41	positive regulation of reactive oxygen species metabolic process	1	83	0.00656	50	5	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0072593	BP	41	reactive oxygen species metabolic process	1	228	0.00554	50	7	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0019835	BP	47	cytolysis	1	29	0.00243	50	4	0.000334	77	5	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043062	BP	63	extracellular structure organization	1	332	0.00603	50	8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0030198	BP	63	extracellular matrix organization	2	331	0.0059	50	8	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043534	BP	70	blood vessel endothelial cell migration	1	73	0.00346	50	5	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043588	BP	73	skin development	1	242	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	1.52E-05	45	9	0.000257	62	9
GO:0031639	BP	81	plasminogen activation	1	18	0.0339	50	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0042744	BP	83	hydrogen peroxide catabolic process	1	22	#N/A	#N/A	#N/A	0.00539	77	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043691	BP	98	reverse cholesterol transport	1	18	0.0339	50	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0034384	BP	106	high-density lipoprotein particle clearance	1	9	0.00355	50	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0034443	BP	110	negative regulation of lipoprotein oxidation	1	4	#N/A	#N/A	#N/A	0.000786	77	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0097006	BP	119	regulation of plasma lipoprotein particle levels	1	55	0.0331	50	4	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0044788	BP	129	modulation by host of viral process	1	20	0.0472	50	3	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
GO:0043226	CC	18	organelle	1	12783	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	0.00059	62	59

GO:0043227	CC	18	membrane-bounded organelle	2	11980	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.00107	62	57
GO:0031410	CC	34	cytoplasmic vesicle	1	1733	0.00301	50	17		0.00325	77	22		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0005576	CC	34	extracellular region	1	4494	2.61E-24	50	48		6.78E-21	77	61		1.56E-19	45	42		2.92E-14	62	47
GO:0044421	CC	34	extracellular region part	1	3800	5.57E-26	50	47		9.16E-23	77	59		1.23E-16	45	38		2.11E-17	62	47
GO:0005577	CC	34	fibrinogen complex	1	9	1.34E-05	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0032994	CC	34	protein-lipid complex	1	41	0.000184	50	5		0.000047	77	6		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0030141	CC	34	secretory granule	1	355	2.73E-07	50	12		6.46E-05	77	12		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031983	CC	34	vesicle lumen	1	106	2.72E-17	50	14		1.05E-17	77	16		0.0159	45	5		4.89E-09	62	10
GO:0072562	CC	34	blood microparticle	2	136	1.27E-44	50	28		9.35E-44	77	31		4.8E-11	45	11		6.93E-16	62	15
GO:0060205	CC	34	cytoplasmic membrane-bounded vesicle lumen	2	105	2.36E-17	50	14		8.93E-18	77	16		0.0152	45	5		4.44E-09	62	10
GO:0044433	CC	34	cytoplasmic vesicle part	2	912	2.19E-06	50	16		6.45E-05	77	18		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0043230	CC	34	extracellular organelle	2	2764	3.75E-24	50	42		3.13E-22	77	52		6.79E-19	45	36		2.78E-22	62	46
GO:0005615	CC	34	extracellular space	2	1378	1.5E-38	50	43		2.8E-37	77	52		1.31E-12	45	24		1.2E-11	62	27
GO:1990777	CC	34	lipoprotein particle	2	39	0.000142	50	5		3.43E-05	77	6		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031091	CC	34	platelet alpha granule	2	74	7.5E-10	50	9		5.48E-08	77	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031093	CC	34	platelet alpha granule lumen	3	55	4.48E-11	50	9		3.36E-09	77	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0034774	CC	34	secretory granule lumen	2	86	5.16E-11	50	10		3.4E-12	77	12		#N/A	#N/A	#N/A		1.04E-06	62	8
GO:0031089	CC	22	platelet dense granule lumen	1	14	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0323	62	3
GO:0099503	CC	34	secretory vesicle	2	462	5.5E-06	50	12		0.00112	77	12		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031982	CC	34	vesicle	2	3994	5.69E-19	50	43		1.56E-16	77	54		1.34E-14	45	37		1.83E-16	62	47
GO:0071682	CC	34	endocytic vesicle lumen	3	17	1.6E-06	50	5		0.00178	77	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:1903561	CC	34	extracellular vesicle	3	2763	3.69E-24	50	42		3.07E-22	77	52		6.71E-19	45	36		2.73E-22	62	46
GO:0097708	CC	34	intracellular vesicle	3	1712	0.00254	50	17		0.00264	77	22		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0034358	CC	34	plasma lipoprotein particle	3	39	0.000142	50	5		3.43E-05	77	6		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0070062	CC	34	extracellular exosome	4	2748	2.95E-24	50	42		2.36E-22	77	52		5.55E-19	45	36		3.46E-21	62	45
GO:0034364	CC	34	high-density lipoprotein particle	4	26	1.67E-05	50	5		2.53E-06	77	6		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0034385	CC	34	triglyceride-rich lipoprotein particle	4	20	0.0472	50	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0034366	CC	34	spherical high-density lipoprotein particle	5	8	0.00237	50	3		0.0109	77	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A

GO:0034361	CC	34	very-low-density lipoprotein particle	5	20	0.0472	50	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0044430	CC	39	cytoskeletal part	1	1525	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.000101	45	17		#N/A	#N/A	#N/A
GO:0005856	CC	39	cytoskeleton	1	2038	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		3.17E-05	45	20		#N/A	#N/A	#N/A
GO:0099080	CC	39	supramolecular complex	1	735	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		1.79E-08	45	16		#N/A	#N/A	#N/A
GO:0099513	CC	39	polymeric cytoskeletal fiber	2	698	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		8.23E-09	45	16		#N/A	#N/A	#N/A
GO:0099081	CC	39	supramolecular polymer	2	712	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		1.11E-08	45	16		#N/A	#N/A	#N/A
GO:0099512	CC	39	supramolecular fiber	3	712	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		1.11E-08	45	16		#N/A	#N/A	#N/A
GO:0045095	CC	39	keratin filament	4	97	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		4.14E-09	45	9		#N/A	#N/A	#N/A
GO:0031838	CC	73	haptoglobin-hemoglobin complex	1	4	#N/A	#N/A	#N/A		0.000786	77	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0045111	CC	83	intermediate filament cytoskeleton	1	247	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		1.15E-12	45	14		0.00396	62	8
GO:0005882	CC	83	intermediate filament	2	203	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		7.27E-14	45	14		0.000901	62	8
GO:0005579	CC	88	membrane attack complex	1	7	#N/A	#N/A	#N/A		0.00682	77	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0031012	CC	133	extracellular matrix	1	527	0.0222	50	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		0.0219	62	10
GO:0004064	MF	3	arylesterase activity	1	6	#N/A	#N/A	#N/A		0.00391	77	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0070051	MF	4	fibrinogen binding	1	3	0.0499	50	2		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0003823	MF	20	antigen binding	1	129	#N/A	#N/A	#N/A		0.0446	77	6		2.1E-06	45	8		#N/A	#N/A	#N/A
GO:0071814	MF	22	protein-lipid complex binding	1	23	0.000916	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0071813	MF	22	lipoprotein particle binding	2	23	0.000916	50	4		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0008233	MF	29	peptidase activity	1	708	6.77E-05	50	13		0.000471	77	15		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0070011	MF	29	peptidase activity, acting on L-amino acid peptides	2	688	4.83E-05	50	13		0.000325	77	15		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:1901681	MF	32	sulfur compound binding	1	240	0.000524	50	8		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0098772	MF	37	molecular function regulator	1	1354	0.019	50	14		0.00492	77	19		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0030234	MF	37	enzyme regulator activity	2	957	0.0142	50	12		0.000771	77	17		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0005200	MF	37	structural constituent of cytoskeleton	2	110	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		2.57E-10	45	10		0.00445	62	6
GO:0004857	MF	37	enzyme inhibitor activity	3	381	0.0166	50	8		1.28E-06	77	14		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A
GO:0061134	MF	37	peptidase regulator activity	3	214	1.12E-05	50	9		2.54E-11	77	15		#N/A	#N/A	#N/A		0.0176	62	7
GO:0061135	MF	37	endopeptidase regulator activity	4	175	0.000941	50	7		3.43E-11	77	14		#N/A	#N/A	#N/A		0.00466	62	7
GO:0030414	MF	37	peptidase inhibitor activity	4	177	0.00102	50	7		4.02E-11	77	14		#N/A	#N/A	#N/A		0.00503	62	7

GO:0004866	MF	37	endopeptidase inhibitor activity	5	169	0.000743	50	7		2.1E-11	77	14		#N/A	#N/A	#N/A		0.00369	62	7	
GO:0004867	MF	37	serine-type endopeptidase inhibitor activity	6	95	0.0128	50	5		1.71E-08	77	10		#N/A	#N/A	#N/A		0.0411	62	5	
GO:0004175	MF	41	endopeptidase activity	1	478	5.99E-07	50	13		2.51E-06	77	15		0.0473	45	8		#N/A	#N/A	#N/A	
GO:0017171	MF	41	serine hydrolase activity	1	275	5.85E-10	50	13		1.7E-08	77	14		4.61E-05	45	9		#N/A	#N/A	#N/A	
GO:0004252	MF	41	serine-type endopeptidase activity	2	246	1.4E-10	50	13		3.77E-09	77	14		0.000332	45	8		#N/A	#N/A	#N/A	
GO:0008236	MF	41	serine-type peptidase activity	2	272	5.09E-10	50	13		1.47E-08	77	14		4.19E-05	45	9		#N/A	#N/A	#N/A	
GO:0008289	MF	58	lipid binding	1	654	#N/A	#N/A	#N/A		0.00712	77	13		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	
GO:0031720	MF	59	haptoglobin binding	1	3	0.0499	50	2		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	
GO:0005539	MF	61	glycosaminoglycan binding	1	204	7.39E-06	50	9		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	
GO:0008201	MF	61	heparin binding	1	157	1.95E-05	50	8		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	
GO:0005198	MF	96	structural molecule activity	1	706	#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		1.37E-07	45	15		#N/A	#N/A	#N/A	
GO:0030492	MF	112	hemoglobin binding	1	5	#N/A	#N/A	#N/A		0.00196	77	3		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	
GO:0005102	MF	122	receptor binding	1	1450	0.0414	50	14		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A		#N/A	#N/A	#N/A	

Appendix C: Coverage and variant statistics for exome sequencing

	IP16	IP17	RT1a	RT1b	PD821	PD161	HG(b)	KG(b)	HG(s)	KG(s)	VL	FL	KIR	KEL	AFF	UNAFF	Average
Total variants	23672	24209	24667	24892	24662	24754	25868	25924	25796	25587	25373	25461	25325	26052	24814	24259	25082.19
Heterozygous variants	14259	14652	13837	13973	14996	15034	16013	16082	15972	15836	15628	15684	15559	16065	15178	14822	15224.38
Homozygous variants	9413	9557	10830	10919	9666	9720	9855	9842	9824	9751	9745	9777	9766	9987	9636	9437	9857.81
Coding variants	21036	21433	21824	21968	21751	21797	22815	22878	22768	22636	22441	22489	22460	22955	21904	21483	22164.88
Heterozygous coding variants	12705	12985	12271	12352	13245	13245	14149	14219	14131	14037	13869	13892	13829	14167	13462	13185	13483.94
Homozygous coding variants	8331	8448	9553	9616	8506	8552	8666	8659	8637	8599	8572	8597	8631	8788	8442	8298	8680.94
Splice variants	2636	2776	2843	2924	2911	2957	3053	3046	3028	2951	2932	2972	2865	3097	2910	2776	2917.31
Heterozygous splice variants	1554	1667	1566	1621	1751	1789	1864	1863	1841	1799	1759	1792	1730	1898	1716	1637	1740.44
Homozygous splice variants	1082	1109	1277	1303	1160	1168	1189	1183	1187	1152	1173	1180	1135	1199	1194	1139	1176.88
Nonsynonymous SNVs	9490	9682	9864	9939	9758	9771	10239	10272	10209	10179	10104	10142	10251	10492	9874	9657	9995.19
Heterozygous nonsynonymous SNVs	5783	5927	5560	5618	5950	5945	6391	6436	6384	6362	6274	6285	6357	6526	6129	5986	6119.56
Homozygous nonsynonymous SNVs	3707	3755	4304	4321	3808	3826	3848	3836	3825	3817	3830	3857	3894	3966	3745	3671	3875.63
Synonymous SNVs	10603	10794	10991	11053	11017	11050	11533	11546	11519	11437	11302	11313	11207	11421	10977	10798	11160.06
Heterozygous synonymous SNVs	6359	6494	6211	6237	6734	6743	7142	7156	7126	7065	6980	6996	6898	7049	6738	6609	6783.56
Homozygous synonymous SNVs	4244	4300	4780	4816	4283	4307	4391	4390	4393	4372	4322	4317	4309	4372	4239	4189	4376.50
Stoploss SNVs	9	11	12	12	11	11	13	13	13	13	17	17	9	8	11	11	11.94
Heterozygous stoploss SNVs	5	7	7	7	6	6	12	12	12	12	13	13	4	3	8	8	8.44
Homozygous stoploss SNVs	4	4	5	5	5	5	1	1	1	1	4	4	5	5	3	3	3.50
Stopgain SNVs	72	75	71	73	59	59	67	67	67	71	79	78	76	81	80	77	72.00
Heterozygous stopgain SNVs	59	61	48	50	41	38	53	50	53	55	58	57	57	61	66	63	54.38
Homozygous stopgain SNVs	13	14	23	23	18	21	14	17	14	16	21	21	19	20	14	14	17.63
Deletions	218	230	228	229	220	221	221	218	219	212	219	216	230	229	235	239	224.00
Heterozygous deletions	136	138	133	132	130	130	126	125	122	121	133	129	132	131	137	143	131.13
Homozygous deletions	82	92	95	97	90	91	95	93	97	91	86	87	98	98	98	96	92.88
Insertions	200	196	196	193	189	192	212	213	207	205	191	195	206	209	218	212	202.13
Heterozygous insertions	119	112	101	97	94	94	117	116	115	114	99	97	103	106	118	111	107.06
Homozygous insertions	81	84	95	96	95	98	95	97	92	91	92	98	103	103	100	101	95.06
Frameshift deletions	63	67	61	61	67	71	64	67	63	61	61	58	80	79	75	74	67.00
Heterozygous frameshift deletions	33	33	27	28	33	36	30	32	27	28	30	28	38	35	37	38	32.06
Homozygous frameshift deletions	30	34	34	33	34	35	34	35	36	33	31	30	42	44	38	36	34.94
Frameshift insertions	62	57	58	57	57	57	64	63	64	64	52	51	58	57	64	62	59.19
Heterozygous frameshift insertions	30	27	25	23	25	24	28	30	30	31	20	18	23	22	24	22	25.13

Homozygous frameshift insertions	32	30	33	34	32	33	36	33	34	33	32	33	35	35	40	40	34.06
Ts/Tv ratio	3.02	2.98	2.96	2.96	2.97	2.95	2.95	2.94	2.95	2.97	3.01	3.01	2.94	2.92	2.93	2.96	2.96
Heterozygous Ts/Tv ratio	3.07	3.02	2.96	2.95	3.02	3.01	2.99	2.96	2.98	3.01	3.1	3.11	2.93	2.91	2.97	2.99	3.00
Homozygous Ts/Tv ratio	2.94	2.91	2.97	2.97	2.89	2.87	2.89	2.91	2.91	2.92	2.88	2.87	2.96	2.93	2.87	2.92	2.91
Novel Varients	172	177	350	348	117	122	136	134	138	138	150	158	149	167	152	140	171.75
Total reads	41146376	72641801	78243596	1.08E+08	90927756	99806001	83590179	85463775	85353802	61617739	80054992	97206924	35828951	81825854	1.09E+08	62983273	79627084
Mapped to target reads	32832986	57069814	57872331	79572806	62838744	69615504	59187597	61176445	60475872	43138989	56827493	69090065	25868488	52187435	79127354	45922530	57050278
Percentage	79.8	78.56	73.96	73.71	69.11	69.75	70.81	71.58	70.85	70.01	70.99	71.08	72.2	63.78	72.33	72.91	71.96438
Mapped to target reads plus 150bp	37617148	64933592	64403921	88750871	70582058	78281869	66042281	68086421	66821215	47627118	63630277	77563149	29615038	60055515	89501305	51432982	64059048
Percentage	91.42	89.39	82.31	82.21	77.62	78.43	79.01	79.67	78.29	77.29	79.48	79.79	82.66	73.39	81.82	81.66	80.9025
Mean coverage	59.74	104.47	108.06	148.35	117.03	129.66	110.97	114.84	113.83	81.33	106.07	128.78	47.66	96.16	146.62	85.8	106.2106
Accessible target bases	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618	33323618
Accessible target bases 1x	32443868	32557326	32694429	32759740	32795331	32804574	32990307	32991645	32974977	32949035	32939272	32962488	32906613	33021700	32764297	32649489	32825318
Percentage	97.36	97.7	98.11	98.31	98.41	98.44	99	99	98.95	98.88	98.85	98.92	98.75	99.09	98.32	97.98	98.50438
Accessible target bases 5x	31673420	32070797	32237744	32349466	32331876	32365474	32746246	32745348	32710645	32598597	32622333	32676388	32440357	32788860	32296165	31972762	32414155
Percentage	95.05	96.24	96.74	97.08	97.02	97.12	98.27	98.26	98.16	97.82	97.9	98.06	97.35	98.4	96.92	95.95	97.27125
Accessible target bases 10x	30419518	31478701	31857004	32078543	32000123	32068292	32489713	32491580	32428826	32134865	32302785	32413443	31477751	32524463	31891409	31109920	31947934
Percentage	91.29	94.46	95.6	96.26	96.03	96.23	97.5	97.5	97.31	96.43	96.94	97.27	94.46	97.6	95.7	93.36	95.87125
Target bases 20x	27085420	29917177	30908654	31488836	31222102	31403252	31729446	31757170	31643208	30691993	31426251	31746844	27932594	31650028	30854176	28843995	30643822
Percentage	81.28	89.78	92.75	94.49	93.69	94.24	95.22	95.3	94.96	92.1	94.31	95.27	83.82	94.98	92.59	86.56	91.95875

Appendix D: A detailed clinical case report of an MZ twin pair discordant for ischaemic stroke

KG and HG are 62-year-old, female, MZ twins of Caucasian descent. KG experienced a stroke in April 2007, resulting in right-sided weakness affecting her speech, writing, memory and balance, requiring her to use a stick. She also complained of sialorrhea and dribbling from the right side of her mouth. She has since experienced recurrent accidental falls. An MRI scan showed a left ganglion infarct, ischaemic changes in both cerebral hemispheres and a small right temporal infarct. There is still some persistent weakness in her right leg and none of the complications from stroke have completely resolved. A coincidental finding of a right sided intra-cavernous ICA cerebral aneurysm measuring 9 mm was made in January 2007 following CT and MRI scans to investigate diplopia resulting from right sided (lateral rectus) 6^{th} nerve palsy, first diagnosed in May 2006, for which she underwent right-sided Jensen's procedure. This has been managed conservatively ever since, measuring 10 mm in 2009.

HG also has a history of diplopia, first reported in 1969 aged 18, and was recently (2011) diagnosed with glaucoma. She also underwent a CT scan in 2011 as a result of left optic disc atrophy after walking into a telephone booth. This scan was normal with no evidence of intracranial pathology.

KG and HG were born as healthy babies, following a normal delivery with no obstetric complications, with approximately similar birth weights. In clinical examinations, the twins are described as hirsute. Currently, KG and HG weigh 83 kg and 70 kg respectively, with a height of 1.6 and 1.63 metres respectively, and BMIs of 32 and 26 respectively. KG's weight has fluctuated considerably from 65 kg to 84 kg since September 1998.

KG's heart size from a medical report in 2001 is described at the upper limit of normal, and she underwent a laproscopic choleocystectomy in 1993.

KG had meningococcal meningitis in June 1963, aged 12. She was briefly hospitalised as a result, with left sided hemiplegia which later resolved. An angiogram (reported 1st August 1963) '*showed no evidence of any space-occupying lesion in the brain, nor any sub-dural abcess*'. She had several left-sided Jacksonian epileptic attacks whilst in hospital, and EEG showed an abnormal record with evidence of acute right parietal disturbance. However, after several follow up scans and neurological examinations in August and September of 1963 and January 1964, it was concluded that KG had made a complete recovery. She appeared to remain seizure-free for several years until the age of 16 when she was diagnosed with epilepsy, which was reported to be due to the earlier meningitis. Her initial EEG findings were mostly within normal limits, except for high amplitude delta activity on closing the eyes during photic stimulation. According to the medical notes, KG had an average of 6 seizures per year between the ages of 16-21 years. Her last two reported seizures were in October 2002 in July 2003. On each of these occasions she was unresponsive for about a minute, but it was noted that no injury had occurred. Anecdotally, her medical records noted that the earlier descriptions of her seizures were entirely compatible with having experienced both complex partial seizures and generalised tonic-clonic seizures. An EEG was also performed on HG at around the time when KG was first diagnosed with epilepsy, and it was reported to be very similar, with a marked delta response to photic stimulation. Although HG had never been diagnosed with epilepsy, she had a further series of EEGs between 1968 and 1977 (aged 25) all of which were described as unstable and '*well outside normal limits*' suggesting '*some slight tendency to fits*'.

The twins were diagnosed with hypertension around 2002 and it appears there is a family history of hypertension from the mother's side of the family. Their brother, who is a smoker, also has hypertension. There is also a family history of colon cancer and depression (Figure D.1). Their mother died of colon cancer at the age of 77, and a paternal aunt is thought to have died from bowel or stomach cancer aged 84. KG had an excision of a non-malignant fibroadenoma from her right breast at age 34. Both the father and grandfather have suffered from depression. The father had a subdural haemorrage aged 88 years following a fall. The grandfather committed suicide, aged 55 (neurasthenia). Both sisters have suffered from depression and anxiety since their teens, but there is no evidence for suicide ideation. HG appears to have suffered from more severe depression, having received at least seven courses of electroconvulsive therapy (ECT), and was sectioned under Section 25 of the Mental Health Act in 1976 aged 25. HG also appears to have had recurrent feelings in her mid-twenties described in various medical reports at the time as '*lesbian thoughts*', '*lesbian feelings*', '*latent lesbian tendencies*' and '*she still gets an urge to go out with girls*'.

The sisters grew up believing they were MZ twins, but this has been confirmed by molecular analysis on DNA from saliva samples. In addition, cytogenetic analysis revealed that the sisters have a pericentric inversion on chromosome 5, with breakpoints at 15p13.1 and 15q11.2.

The sisters carry some of the risk factors associated with stroke. Both started smoking at the age of 16, and have been diagnosed with hypertension for 11 years. KG gave up smoking following the onset of stroke. Prior to that her smoking was restricted due to smoking restrictions at her place of work – she was an NHS receptionist for 37 years prior to the stroke. HG did not have restrictions at work and smoked throughout. She still

smokes up to 10 cigarettes a day. The sisters consume alcohol only on special occasions, but not excessively. KG was diagnosed with a single patch of alopecia areata in 1995 (aged 44), and more recently with seborrheoic dermatitis capitis in 2012 (aged 60). HG was diagnosed with atopic dermatitis in 2004 (aged 53). These problems might be linked to a family history of psoriasis (in father) and suggest an autoimmune disorder in the family (Figure D.1).

Described herein are MZ twins with a family history of hypertension, depression and psoriasis. However, they are discordant for stroke, and there is no obvious family history of stroke. Despite sharing the same risk factors and still smoking, HG has not suffered a stroke. To our knowledge, this is the first documented case of twins discordant for stroke. KG also has an intracranial aneurysm.

Remarkably, the twins share a uniquely similar set of circumstances. Both are right handed and enjoy a similar diet. These church-going, celibate spinsters have lived together in the same house since the age of 3, within a few miles from their place of birth. They have largely been part of the same medical practice for this period, mostly with the same doctor for the past 30 years. This is documented in over 300 pages of detailed medical notes spanning more than 50 years. This provides a unique opportunity for exploring the genetic, epigenetic, environmental, and therapeutic aspects of stroke, by searching for genomic, transcriptomic and proteomic differences between the twins.

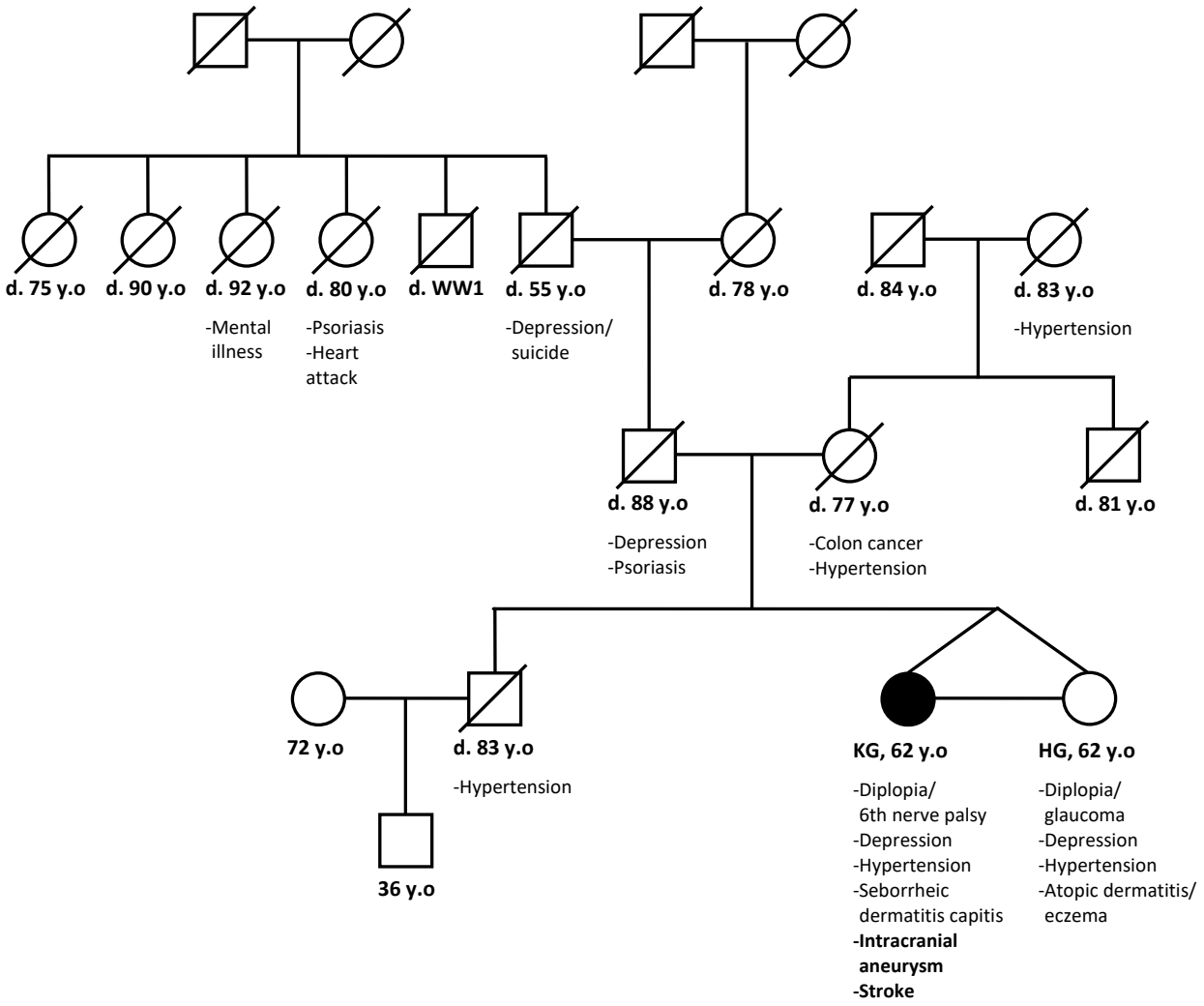


Figure D.1. Family tree of KG and HG. The twins are discordant for stroke and intracranial aneurysm, but there is a family history of hypertension and depression.

Publications

Manuscripts in preparation

- Vadgama, N., Lamont, D., Hardy, J., Nasir, J., Lovering, RC. (2018). **Distinct proteomic profiles in monozygotic twins discordant for ischaemic stroke identifies candidate disease markers**. *Manuscript submitted to Translational Stroke Research*.

- Vadgama, N., Simpson, M., Pittman, A., Niranjanan, M., Hardy, J., Nasir, J. (2018). **Investigating postzygotic de novo mutations in discordant monozygotic twins to explore disease-related pathways for complex disorders.** *Manuscript in preparation*.

- Mohamoud, H., Musharraf, J., Vadgama, N., Nasir, J. (2018) **A homozygous mutation in the *WIPI2* gene suggests failure in autophagy pathways in a novel multisystemic disorder.** *Manuscript in preparation*.

- Murphey, D., Hardy, J., Vadgama, N., Nasir, J., Pittman, A., et al. (2018). **Analysis of mitochondrial DNA from discarded archival next-generation sequencing data**. *Manuscript in preparation.*

- Mohamoud, H., Vadgama, N., Jelani, M., Hussain, M., Ahmed, S., Feng, Q., Ahmad, I., Ancla, R., Nasir, J., Wang, J., Al-Aama, J. (2018). **Whole-exome sequencing identifies digenic inheritance of *DEPP* and *HBB* genes in a congenital polycythemia family with pregnancy-induced deep vein thrombosis.** *Manuscript in preparation*.

Published manuscripts

- Mohamoud, H., Ahmed, S., Jelani, M., Alrayes, N., Childs, K., Vadgama, N., Almramhi, M., Al-Aama, J., Goodbourn, S. and Nasir, J. (2018). **A missense mutation in TRAPPC6A leads to build-up of the protein, in patients with a neurodevelopmental syndrome and dysmorphic features.** *Scientific Reports*, 8(1).

- Vadgama, N., Nirmalananthan, N., Sadiq, M., Hardy, J. and Nasir, J. (2015). **Identical non-identical twins and non-identical identical twins**. *BMJ*, p.h6589.

- Vadgama, N., Gaze, D., Ranson, J., Hardy, J. and Nasir, J. (2015). **Elevated γ-glutamyltransferase and erythrocyte sedimentation rate in ischemic stroke in discordant monozygotic twin study**. *International Journal of Stroke*, 10(4), pp.E32-E33.

- Alrayes, N., Mohamoud, H., Jelani, M., Ahmad, S., Vadgama, N., Bakur, K., Simpson, M., Al-Aama, J. and Nasir, J. (2015). **Truncating mutation in intracellular phospholipase A1 gene (DDHD2) in hereditary spastic paraplegia with intellectual disability (SPG54)**. *BMC Research Notes*, 8(1).

Presentations

- Vadgama, N. (2017). **Investigating complex disorders in an era of 'omics' technologies.** *Presented at Phytomedical Compounds for Diabetes and Diabetes-Related Complications Workshop, Mexico City, MX.*

- Vadgama, N., Simpson, M., Niranjanan, N., Gaze, D., Pearce, K., Kristiansen, M., De Rijk, P., Rees, E., Kirov, G., Pittman, A., Morgan, S., Lamont, D., Hardy, J., Nasir, J. (2017). **Investigating postzygotic de novo mutations and somatic mosaicism in monozygotic twins discordant for complex disorders.** *Presented at IoN-ICM 2017 Workshop. London, UK.*

- Vadgama, N, D. Lamont, J. Hardy, J. Nasir, R.C. Lovering. (2016). Abstract: **Label-free quantitative proteomic profiling of discordant monozygotic twin blood serum identifies Fibulin 1 as a candidate biomarker for ischaemic stroke**. *Presented at American Society of Human Genetics 2016 Annual Meeting. Vancouver, CA*

- Vadgama, N., Simpson, M., Niranjanan, N., Gaze, D., Pearce, K., Kristiansen, M., De Rijk, P., Rees, E., Kirov, G., Pittman, A., Morgan, S., Lamont, D., Hardy, J., Nasir, J. (2015). **Comparing genetic, proteomic and clinical profiles of discordant monozygotic twins.** *Presented at American Society of Human Genetics 2015 Annual Meeting. Baltimore, Md.*