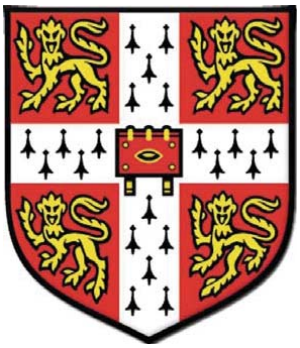
Genetic and Phenotypic Investigations into Developmental Disorders



Dr Wendy Dawn Jones
Wellcome Trust Sanger Institute
Newnham College,
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
July, 2017

Abstract

Genetic and Phenotypic Investigations into Developmental Disorders, Dr Wendy Dawn Jones

Genetic developmental disorders cause distress to families and substantial mortality, morbidity and costs to the health service. However not all genetic diseases have been discovered or had their genetic cause elucidated, and the phenotypic spectrum of many molecularly solved disorders is not fully understood.

Wiedemann-Steiner syndrome (WSS), resulting from mutations in *KMT2A*, is a multiple congenital-anomaly syndrome associated with hypertrichosis, intellectual disability and a distinctive facial appearance. In order to understand the broader spectrum of WSS I identified 84 individuals with WSS and mutations in *KMT2A* and performed a detailed phenotypic evaluation. My cohort is 15 times larger than the biggest cohort reported so far. I identified new phenotypic features, and defined the mutational spectrum and growth profile associated with WSS and mutations in *KMT2A*. In addition, I ran a clinician facial recognition experiment that confirmed WSS is distinguishable from other developmental disorders. To investigate the genetic architecture of hypertrichosis more generally, I assembled a cohort of 228 individuals with hypertrichosis. I showed by analysing their exome variant profiles that there is a burden of mutations in genes that play a role in maintaining the structure and function of chromatin in this group compared to other individuals with developmental disorders. I showed, in principle, grouping by hypertrichosis is a successful method for gene discovery.

Finally, I investigated autosomal recessive disease in 1080 individuals with developmental disorders in the DDD study, for which I generated a population matched control dataset using the parental untransmitted alleles. My work gives the first insight into the contribution of autosomal recessive disease to developmental disorders, by studying untransmitted haplotypes. The themes of this thesis include those important in current Clinical Genetics practice in the whole exome sequencing era: loss of function versus missense variants, the use of next generation sequencing to unravel the underlying causes of developmental disorders and the challenges of assigning pathogenicity to variants.

Declarations

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed 60,000 words.

Dr Wendy Dawn Jones, July 2017

Acknowledgements

First and foremost, I would like to thank my supervisors Jeff Barrett and Matt Hurles. In particular, for our punchy and exciting conversations about science, their highly constructive feedback and for inspiring me by their own achievements and ideas. Special thanks also to David Fitzpatrick for his advice with my study of individuals with Wiedemann-Steiner syndrome.

I also wish to thank all those colleagues and collaborators without whom this work would not have been possible, especially, Michael Simpson, Meriel McEntagart, Charu Despande and Hans Bjornsson. Special thanks also, to all the clinicians who recruited individuals to the DDD study and to the WiSH study and welcomed me into their clinics to meet individuals with Wiedemann-Steiner syndrome. I also thank the Hurles and Barrett Groups and the DDD study team for all their help and stimulating conversations, both about science and life. In particular, to Margriet, Art and TJ. I remain forever indebted to those who helped me in my journey to become a programmer, in particular, Tomas, Dan and TJ.

On a personal note, thanks to all my friends, old and new who helped, often without realising, with their love and support and sense of fun. To name but a few, thanks to Sharmeen, Kate, Angela, Elora, Alex, Laura and Lucy. To Cor and Jeannette for their support and understanding. To my grandfathers who taught me to aim high in life and my grandmothers who taught me about hard work and perseverance. To my wonderful parents for all their support and encouragement throughout my life. To David for making me laugh. To Rogier for everything, especially his endless support and understanding.

Finally, and most importantly, I would like to thank the individuals affected with developmental disorders and their families for kindly taking part in this research, without whom, this work would have not been possible.

Attributions

Below is a summary of the contributions of other scientists and clinicians to the work described in this dissertation. I carried out all of this work under the supervision of Dr Jeff Barrett and Dr Matt Hurles.

Chapter 2:

Patient recruitment to the Genotype-phenotype study was by Clinical Geneticists in the UK, Europe and worldwide. Local Clinicians filled out questionnaires and provided phenotype information. Dr Roman Laskowski performed the protein modelling experiments and created the figures associated with these and the figure legends. I was responsible for the execution of this investigation from start to finish. I managed the flow of patient data for all 84 patients and liaised with clinicians to obtain phenotype data. I was responsible for the clinician facial recognition survey. I analysed all the phenotypic and molecular data and performed the statistical analysis and was responsible for the clinical interpretation of findings.

Chapter 3:

Whole exome sequencing was carried out by the Wellcome Trust Sanger Institute (WTSI) Core facility. SNV and INDEL detection was carried out by Mr Martin Pollard. I received help with variant annotation from Stephen Clayton (VEP), Mr Martin Pollard and Mr TJ Singh. The clinical filtering programme was written by Dr Jeremy McRae based on code originally written by Dr Saeed Al Turki and Dr Jeff Barrett. Mr TJ Singh provided help with performing statistical analysis on missense variants and python programming. The programme used to calculate the burden of variants in chromatin genes was written by Dr Jeremy McRae. I was responsible for the execution of this investigation from start to finish. I designed the recruitment criteria and liaised with UK clinical genetics clinicians to identify individuals with relevant phenotypes within the DDD and recruit 20 trios from overseas. I analysed the exome sequencing data from the variant call format stage, writing my own QC

scripts and scripts to analyse the exome data to identify rare coding variants. I was responsible for reviewing all of the variants in confirmed disease genes and candidate genes, performing statistical analysis to analyse the significance of missense variants and performing the burden analysis for variants in chromatin genes.

Chapter 4:

Some parts of this investigation have been published (1, 2). Patient recruitment into the DDD study was carried out by Clinical Geneticists and Genetic Counsellors in the UK and Ireland. The DDG2P was devised by Professor David Fitzpatrick and is updated by Professor David Fitzpatrick and Dr Helen Firth. The Decipher team supported and ran the portal in Decipher into which clinical information was uploaded. Whole exome sequencing was carried out by the Wellcome Trust Sanger Institute (WTSI) Core facility. SNV and INDEL detection was carried out by the GAPI pipeline team at the WTSI. *De novo* mutation detection was carried out by the DDD Study Bioinformatics team. The DDD management team comprises Dr Matt Hurles, Dr Jeffrey Barrett, Dr Caroline Wright and Dr Helen Firth and Professor David Fitzpatrick, Consultants in Clinical Genetics.

The untransmitted diplotypes was an idea conceived by Dr Jeff Barrett. I had help with Perl programming from Dr Dan King, Dr Tomas Fitzgerald and Dr Ray Millar. The analysis looking at variant numbers to adjust QUAL score was carried out working with Dr Tomas Fitzgerald. Dr Tomas Fitzgerald carried out the Mann-Whitney test of the numbers of rare SNPs per sample in probands and untransmitted diplotypes and carried out the PCA. I was responsible for the untransmitted diplotypes dataset generation and analysis from start to finish. I was responsible for the QC analysis and wrote the scripts for untransmitted diplotype generation.

Publications

Resulting from this work:

Peer reviewed articles:

Fitzgerald TW*, Gerety SS*, **Jones WD***, van Kogelenberg M* et al. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015 ;519(7542):223-8.

Wright CF, Fitzgerald TW, **Jones WD**, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet. 2015;385(9975):1305-14.

King DA, **Jones WD**, Crow YJ, Dominiczak AF, Foster NA, Gaunt TR, et al. Mosaic structural variation in children with developmental disorders. Human molecular genetics. 2015.

Akawi N, McRae J, Ansari M, Balasubramanian M, Blyth M, Brady AF, *et al*. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nature genetics. 2015;47(11):1363-9.

McRae, JF, Clayton, S, Fitzgerald, Kaplanis J *et al*. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433-8.

Book chapters

Jones WD, Hypertrichosis Chapter. Oxford Desk Reference: Clinical Genetics (2nd edition) by Firth HV, Hurst JA. Oxford University Press. In press.

* Denotes joint first author.

Contents

List of Abbreviations

2D	Two dimensional
3D	Three dimensional
aCGH	Array comparative genomic hybridisation
bcf	Binary variant call format
BWA	Burrows-Wheeler Aligner
ADHD	Attention deficit hyperactivity disorder
ALT	Alternate allele
BAM	Binary alignment map
CdLS	Cornelia de Lange Syndrome
CM	Centimetres
CNV	Copy number variant
CpG	5'-C-phosphate-G-3'
DAM	Damaging
DD	Developmental disorder
DNA	Deoxyribonucleic acid
DDD	Deciphering developmental disorders
DDG2P	Developmental Disorder Gene2Phenotype
DNA	Deoxyribonucleic Acid
EMG	Electromyogram
ESP	Exome Sequencing Project
EURODIS	Rare Diseases Europe
ExAC	The Exome Aggregation Consortium
FDNA	Facial Dysmorphology novel analysis
FISH	Fluorescence *in situ* hybridisation
FORGE	Finding of rare disease genes
FUNC	Functional
GAPI	Genome analysis production informatics
GATK	Genome analysis toolkit
GO	Gene ontology
GRCh37	Genome Reference Consortium human genome (build 37)

HbF	Fetal haemoglobin / haemoglobin F
HET	Heterozygous
HOM	Homozygous
HPO	Human phenotype ontology
H3K4	Lysine 4 of histone H3
INDEL	Insertion or deletion
IGV	Integrative Genomics Viewer
Kg	Kilogram
LoF	Loss of function
MAF	Minor allele frequency
Mb	Megabase
MM	Millimetres
MRI	Magnetic resonance imaging
NA	Not available
NHS	National health service
OFC	Occipital frontal circumference
OMIM	Online Mendelian Inheritance in Man
PCA	Principle component analysis
PCR	Polymerase chain reaction
PEG	Percutaneous endoscopic gastrostomy
PEJ	Percutaneous endoscopic jejunostomy
PHD	Plant homeo-domain
PolyPhen	Polymorphism Phenotyping
PROB DAM	Probably damaging
QC	Quality control
QUAL	Variant quality score from GATK
REF	Reference allele
RNA	Ribonucleic acid
RT-PCR	Real time polymerase chain reaction
SD	Standard deviations
SNP	Single-nucleotide-polymorphism
SNV	Single nucleotide variant

SWI/SNF	Switch-Sucrose Non-Fermentable
TDT	Transmission disequilibrium test
UTR	Untranslated region
UK	United Kingdom
VCF	Variant call format
VEP	Variant Effect Predictor
VQSLOD	Variant quality score log-odds
VQSR	Variant Quality Score Recalibration
WISH	Wiedemann-Steiner syndrome or related phenotypes or hypertrichosis
WiSH-WES	Wiedemann-Steiner syndrome and hypertrichosis whole exome sequencing
WSS	Wiedemann-Steiner syndrome
WSSP	Wiedemann-Steiner syndrome phenotype
WTSI	Wellcome Trust Sanger Institute

Chapter 1

Introduction

1.1 Definition, incidence and impact

Developmental disorders are a diverse group of conditions that result in abnormal human development. They demonstrate variability, both within a single disorder and across different types of disorder. Some are life limiting, painful, severely debilitating or degenerative. Developmental disorders may be associated with congenital abnormalities, for example heart or brain malformations or with neurological, cognitive or behavioral phenotypes, for example hypotonia, delayed developmental milestones, intellectual disability or autism. Some developmental disorders have phenotypes known only to affect one organ, for example *NR2F2* (MIM 107773) mutations in congenital heart disease(3). However, many developmental disorders manifest with a multitude of variable phenotypic features affecting a variety of organ systems.

Many individuals with developmental disorders have intellectual disability either as part of a syndrome or as an isolated phenotype. Intellectual disability is defined as substantial impairment of cognitive and adaptive functions that has onset in childhood(4). Severity can range from mild to profound. Developmental disorders can also result from environmental causes, for example *in utero* exposures, trauma or infection. However, many developmental disorders have a genetic cause, in fact the majority of cases of severe intellectual disability are thought to be genetic(5). However, not all rare genetic diseases (including genetic developmental disorders) have been defined and had their underlying cause elucidated. Estimates from the pre-genomics era using human genome mutation rates and the number of essential genes are that there may be around 7750 –15,300 rare-disease-causing genes(6). For those that do receive a genetic diagnosis for their or their child's developmental disorder, the time period leading to diagnosis or 'diagnostic odyssey' may take several years. A study by Rare Diseases Europe (EURODIS) of 6000 families or individuals in 17 countries affected by 8 rare diseases (the majority developmental disorders) showed that for 25% of people the time to diagnosis was 5 to 30 years(7).

Rare diseases are life-threatening or chronically debilitating diseases with low prevalence. Prevalence of rare diseases is defined as less than 1 in 20,000 people in the United States of America and less than 1 in 2000 people in Europe [Commission of the European Communities(8). In Europe it is estimated that five to eight thousand

different rare diseases affect 6-8% of the population(9) Many rare diseases are genetic developmental disorders. Although these disorders are individually rare, collectively they are common and genetic rare diseases affect at least 1 in 50 individuals(10). Some of these genetic developmental disorders can present in the neonatal period, and around a third of these infants will succumb to their rare disease in their first year of life(11-13). In addition to the effects on affected individuals and their families, intellectual disability and other developmental disorders are associated with significant morbidity and mortality and pose enormous socio-economic costs(14-16). These costs include the indirect costs of productivity losses in workplaces or households that occurs when an individual with a developmental disorder is unable to work, or are limited in the amount or type of work they could do or dies prematurely(15).

1.2 A short history of genetic developmental disorders

1.2.1 Copy number change as a cause of developmental disorders

Over time improvements in technology have increased the possible number of genetic diagnoses we are able to make. Copy number change is defined as a gain or loss of chromosomal genetic material compared to the reference human genome. Multiple studies in control populations, have shown that there is tolerance for copy-number change in some regions of the genome(17-19). All humans are estimated to carry copy number variants (CNVs) and they are thought for the most part benign and part of normal variation. However, the effect of a CNV depends on whether it changes the relative location and or sequence of genomic DNA and CNVs are a well-established cause of developmental disorders.

Chromosomal causes of developmental disorders were first identified with the identification of the presence of an extra copy of chromosome 21 in individuals with Down syndrome in 1959(20). This discovery was made on karyotype analysis, a technique honed by Tjio and colleagues who discovered in 1956 that man has 46 chromosomes(21). This was followed, by the discovery of several other chromosome imbalances, including unbalanced translocations, marker chromosomes and large deletions and duplications as the cause of developmental disorders. Karyotyping is able to detect imbalances as small as 5 to 10 Mb and these explain 10-15% of intellectual disability. Karyotyping also has the ability to detect mosaicism (more than one cell

population deriving from a single zygote) a phenomenon apparent in a diverse range of human disorders including developmental disorders(22). Chromosomal mosaicism has been detected from the earliest stages of karyotype use(23).

Developed in the 1980s, Fluorescence in situ Hybridisation (FISH) uses fluorescent labelling to detect chromosome imbalances and provided an accurate method for identifying and confirming small deletions and duplications. The discovery of FISH led to the development of methods to detect subtelomeric chromosomal deletions and duplications ((24) and reviewed by Rudd(25)), these were found to cause 2-5% of previously unexplained intellectual disability(26, 27). More recently, the technique of optical mapping has been reported to successfully detect structural chromosomal variants(28, 29). Optical mapping approaches involve the construction of ordered restriction maps from individual molecules of genomic DNA using single-molecule measurements and computational analysis(30).

Chromosome Microarrays

Chromosome microarrays have increasingly become the 1st line copy number diagnostic test in the developed world(31). Microarray technology can detect smaller gains and losses of DNA sequence than karyotype analysis with a range in length from 1000bp in size. The commonest technologies used diagnostically are array comparative genomic hybridisation (aCGH), which uses fluorescent labelling for comparison to control DNA and single-nucleotide-polymorphism (SNP) arrays which uses fluorescence to label SNPs. Due in the main part to their ability to detect submicroscopic deletions and duplications, chromosome microarrays offer higher diagnosis rates than traditional karyotyping, with 15%–20% of individuals with developmental delay, intellectual disability, autistic spectrum disorder or multiple congenital abnormities receiving a diagnosis(31). However, chromosome microarrays are unable to detect truly balanced translocations, also high-resolution arrays are not in widespread use in a clinical diagnostic setting and small exonic duplications and deletions can go undetected(32).

1.2.2 Single gene causes of developmental disorders

In the 1970s Fred Sanger and colleagues developed a new method of sequencing deoxyribonucleic acid (DNA), the dideoxy chain-termination method(33). This technique

enabled the rapid and accurate sequencing of large stretches of DNA. The introduction of 'Sanger' sequencing together with the introduction and improvement of the polymerase chain reaction (PCR)(34, 35), and the genetic map or catalogue of polymorphisms and linkage methods transformed the diagnostic and research arena for genetic diseases in the 1990s.

Much of early gene discovery in developmental disorders was driven by linkage experiments requiring multiple affected family members. As the presence of female carriers could permit pedigree analysis, this led to the discovery of many X-linked disorders, including fragile X syndrome(36) and Rett syndrome(37). Many autosomal conditions for which gene discovery was possible (neurofibromatosis, myotonic dystrophy, Noonan-spectrum disorders and tuberous sclerosis) are characterised by variable intellectual disability, which increased the possibility of being able to study multiple affected family members, because reproduction was not impaired in all affected individuals by intellectual disability.

1.2.3 Genome wide sequencing approaches in developmental disorders

Sanger sequencing is reliable and robust and was the mainstay of sequencing technology used for over 25 years. However, it is time consuming, and in the diagnostic arena affords little more than targeted sequencing of one or two genes at a time. Second (Next) generational sequencing platforms for genome wide sequencing have become widely available since 2005 and have significantly reduced the cost of DNA sequencing relative to Sanger sequencing(38). These methods carry out massively parallel sequencing of small fragments of DNA from across the entire genome or exome to deliver results rapidly. Initial successes for whole exome sequencing in the clinical arena were a proof of principle analysis in Freeman Sheldon syndrome(39) and a diagnosis of a known condition (congenital chloride diarrhea) in an individual thought to have Bartter syndrome(40). The first developmental disorder of unknown cause unraveled by whole exome sequencing was Miller syndrome which was found to be caused by rare biallelic loss of function variants in *DHODH* (MIM 126064)*(41)*. This seminal paper also illustrated the possibility of incidental or unexpected findings in genome wide sequencing by identifying variants in a ciliary gene in two individuals with

bronchiectasis, recurrent lung infections and chronic obstructive pulmonary disease. Since this time, hundreds more developmental disorder genes have been discovered.

The limitations of genome wide sequencing techniques include the inability or difficulty to detect balanced translocations and variants in repetitive regions of the genome or in regions with highly homologous sequences elsewhere in the genome.

De novo mutations are an increasingly recognized cause of developmental disorders

Exome sequencing confirmed that many undiagnosed developmental disorders result from new germline mutations in autosomal dominant genes arising between generations, known as *de novo* dominant mutations. The first clinically well recognized disorder noted to arise as a *de novo* mutation was Kabuki Make-up syndrome(42). Following this *de novo* mutations were found to underlie a number of distinctive multiple anomaly syndromes including the Say Barber Biesecker type of Ohdo syndrome, Coffin Siris syndrome, and Wiedemann-Steiner syndrome(43-46).

Large projects and consortia

With the increasing widespread use of whole exome sequencing many collaborations have been formed including the nationwide project FORGE in Canada, which aims to discover new genes and identify mutations in known genes(47). In the UK, the Deciphering Developmental Disorders (DDD) Study, is a nationwide study which uses multiple complementary genome wide approaches to decipher the underlying genetic cause of developmental disorders with a trio design(48). I discuss the DDD study in detail in Chapter 2.

1.3 Diagnosing Developmental disorders

1.3.1. Different types of diagnoses for developmental disorders

There are different types or levels of diagnosis for genetic disorders including: clinical, biochemical, genetic (molecular or cytogenetic). A clinical diagnosis means a doctor has examined the individual and has decided their phenotype fits with a certain disorder. This may be specific, i.e. they think that they fit with one particular disorder, e.g. Kabuki-

Make up syndrome, or within a spectrum of disorders, e.g. they have a ciliopathy, i.e. one of a group of disorders that results from impaired ciliary function and have shared features. A clinical diagnosis enables a recurrence risk to be given for future pregnancies. However, there is a possibility that a clinical diagnosis may be incorrect, and thus any given recurrence figures may not be accurate. It also doesn't enable other relatives to be tested for the disorder or for a specific genetic test to be carried out in future pregnancies. Clinical assessment for so called 'dysmorphic features' has played a significant role in the diagnosis and understanding of developmental disorders and making clinical diagnoses. Dysmorphic features are defined as features unusual for a person's age and ethnicity, with Dysmorphology defined as the study of human congenital malformations and syndromes. With the number and rarity of conditions involved, genetics clinicians have traditionally worked together to share knowledge and help diagnose patients. This has led to the occurrence of multiple international meetings for discussing clinical cases and viewing images, on a local, regional and on an international arena for example the Smith Dysmorphology meeting in the USA and the Manchester Dysmorphology Conference. There are also databases set up to help make diagnoses based on phenotypic features, such as the London Medical database www.lmdatabases.com(49) and Possum www.possum.net.au(50).

An international group of clinicians worked together to publish stringent standardized human morphological terms with consensus definitions (and also highlight terms not acceptable for use), illustrating each term with a photograph(51-56). The aim of this work was to increase the accuracy of discussions between dysmorphologists and other specialists such as molecular geneticists and developmental biologists. This standardized list has been accepted as the morphological terms that should be utilized by the American Journal of Human Genetics(57). Ontologies have also been developed to record standardised phenotypic terms, such as the Human Phenotype Ontology (HPO) (58).

Biochemical diagnoses of genetic diseases are most commonly achieved in metabolic disorders where they are able to quantify the enzyme defect or other metabolic disturbance to confirm the diagnosis. Sometimes biochemical testing can be carried out in pregnancy to look for recurrence of metabolic disorders, however genetic testing is the

gold standard for prenatal testing. Also genetic testing can sometimes identify specific subtypes of the metabolic disorders which may direct management as certain subtypes respond better to treatment. For some genetic conditions there are metabolic or chemical tests available that may give evidence as to the diagnosis or carrier state of an individual for example measuring creatine-kinase in duchenne muscular dystrophy or haemoglobin H inclusion bodies in Alpha-Thalassemia X-Linked Intellectual Disability Syndrome **(**ATRX).

Finally, a genetic diagnosis implies having a genetic confirmation of the individual's disorder by identifying the genetic aberration (sequence variant, copy number variant or imprinting defect) that has caused the individual's disorder. This may be a molecular diagnosis, generally meaning a diagnosis achieved through single gene analysis or a cytogenetic diagnosis from microarray or karyotyping. Achieving a genetic diagnosis enables other relatives to be screened to see whether they are carriers for a disorder, it also offers an accurate test to be available in future pregnancies.

1.3.2 Barriers to diagnosing developmental disorders

Barriers that have prevented making a genetic diagnosis in developmental disorders include: the large number of disorders, the diversity of phenotypes associated with developmental disorders, the diversity of genes and mechanisms implicated and the variability of the disorders. Also for some disorders, the small number of families with multiple affected members available and the reproductive disadvantage of the disorder limits the opportunities for studying genes which have been inherited together with the family and which segregate with the disorder to determine the cause (linkage mapping).

1.4 Advantages to making a genetic diagnosis

1.4.1 Benefit to affected individuals, their families and society

For families, being without a genetic diagnosis for their child's or their own developmental disorder, can lead to distress, guilt and anxiety. With a diagnosis may come relief and an end to the diagnostic odyssey and uncertainty. Graungaard *et al* showed that families without a diagnosis find it hard to cope with an uncertain future(59). A study of families with fragile X syndrome showed most viewed having a diagnosis as a

benefit as opposed to a disadvantage(60). Making a genetic diagnosis also enables families to access specific information about their child's condition and join support groups, it may help achieve learning support at school or special educational services.

In pursuit of making a diagnosis, individuals may undergo multiple investigations, many of which are invasive or painful including, blood tests, skin or muscle biopsies, lumbar punctures, brain magnetic resonance imaging (MRI) scanning or an electromyogram (EMG). These investigations are more difficult to carry out in young children especially those with learning or behavioral difficulties and investigations which are straightforward for adults or older children, such as undergoing a brain MRI scan may require a general anesthetic. These investigations are often at a large cost to the healthcare service or provider and require significant time commitments for the individual, family members and caregivers and possibly require an inpatient hospital admission.

Without a confirmed molecular or cytogenetic diagnosis individuals may be given an incorrect clinical diagnosis which may in itself lead to morbidity. In the study by Rare Diseases Europe (EURODIS) for some individuals an incorrect diagnosis led to treatments, including surgery and psychiatric treatment based on an incorrect diagnosis(7).

A genetic diagnosis enables the tailoring of medical care to the individual's specific condition. There are few genetic disorders for which there are specific pharmacological treatments available. Developing treatments has been challenging as not all conditions have not been molecularly defined and with small numbers of affected patients dispersed across the world, securing pharmaceutical funding and setting up of trials is difficult. However, there may be management guidelines or screening recommended for later onset related comorbidities. A genetic diagnosis may also enable individuals take part in clinical trials or be put on a disease registers to be contacted should treatments become available in the future.

Achieving a genetic diagnosis also gives parents accurate information about recurrence risks in future pregnancies. This may enable them also to pursue pre-natal testing or pre-implantation genetic diagnosis in future pregnancies if this is something they would

like. It also enables screening of the wider family to find out whether they are also affected or carriers of the condition. Being a carrier may confer a significant offspring risk for women where the disorder is X-linked or for a consanguineous couple when a recessive disorder has been identified in the family.

Achieving a genetic diagnosis also allows for individuals and their families to take part in phenotypic and natural history studies if they choose. This leads to greater understanding of the natural history of these rare conditions, which will ultimately improve their management. Understanding the phenotypes of these disorders will also enable identification of end-points for clinical trials and help development treatments in the future.

Understanding genetic disorders and their causes, allows screening programs to be developed and implemented, for example cystic fibrosis testing in newborn infants in the UK to alleviate the clinical consequences of the disease by early treatment, and prenatal screening for trisomies in the UK to enable couples to terminate an affected pregnancy if this is something they choose.

Understanding the phenotype and genetic cause of developmental disorders can also give important insights into common and complex disorders. For example Gaucher's disease is a rare lysosomal storage disorder resulting from biallelic variants in the glucocerebrosidase (GBA) gene. Several individuals with this condition and their relatives were noted to have developed Parkinson's disease. Further investigation of this phenomenon showed that Parkinson's disease segregated with the mutant *GBA* alleles in the family(61). This led to several studies, including a multicenter collaborative study of over 5691 individuals with idiopathic Parkinson's disease(62) which confirmed that heterozygous mutant GBA alleles are common and important risk factors for not only Parkinson's disease but also dementia with Lewy bodies (reviewed by Siebert *et al(63))*. More recently mutant GBA alleles have also been shown to be important risk factors for developing multisystem atrophy(64).

Identifying the underlying genetic cause of developmental disorders also advances biology more generally by increasing understanding of molecular pathways and gene and protein function in health and disease in humans and other organisms. For example the discovery through linkage analysis that a heterozygous missense mutation in the

forkhead box P2 gene (FOXP2) resulted in a rare severe speech and language disorder in one three generational family(65-68) uncovered FOXP2 as vital not only for normal language development in humans but also for birdsong in songbirds(69).

1.5 Summary and justification for this investigation

In summary, developmental disorders cause significant mortality, morbidity, distress to families and costs to the health service. For families, diagnosing developmental disorders alleviates stress, possible guilt and anxiety and provides them with coping mechanisms. It also enables the tailoring of medical care to an individual's particular condition and may prevent further invasive diagnostic investigations. Diagnosing developmental disorders can also give us insights into common human diseases as well as advance scientific understanding of other species and the world more generally.

Although advances in sequencing technologies have led to a genetic revolution in knowledge and unraveled the cause of a number of genetic disorders, there remain many challenges at this rapidly moving time. Firstly, not all genetic diseases have been discovered or had their genetic cause elucidated. Secondly there is limited longitudinal phenotypic data available for individuals with molecularly confirmed genetic disorders, meaning that we don't fully understand the wider phenotypes of many of these conditions in particular the phenotypes of many disorders in adulthood. Thirdly, molecularly confirmed developmental disorders are highly variable and little progress has been made to understand this variability or to explain incomplete penetrance of some disorders. Fourthly there are few genetic disorders with available treatments, in order to facilitate the development of treatments in the future, disorders must be molecularly defined and phenotypically characterised over time to allow end points for clinical trials to be identified. Treating genetic disorders is one of the biggest challenges in medical genetics for the next century.

1.6 Outline of this dissertation

This investigation takes traditional and contemporary approaches to understanding developmental disorders:

In Chapter 2, I introduce the developmental disorder Wiedemann-Steiner syndrome (WSS), an autosomal dominant multiple congenital-anomaly syndrome associated with a distinctive facial appearance, developmental delay and hypertrichosis. I led the original gene discovery project for this disorder in 2012(46), by discovering that *de novo* variants in *MLL* (now called *KMT2A*) underlie WSS. Since this time, a number of case reports and small case series have been published, but the full phenotypic spectrum of this disorder is unknown. I investigate the wider phenotypic spectrum associated with WSS by reporting on the phenotype of 84 individuals with this disorder and *KMT2A* mutations.

The theme of Wiedemann-Steiner syndrome continues into Chapter 3, where I present a phenotypic investigation looking for variants in genes underlying developmental disorders associated with hypertrichosis or WSS or related phenotypes. I also present a burden analysis showing that there is a burden of variants in genes encoding chromatin modification enzymes in individuals with hypertrichosis and WSS like phenotypes.

Understanding the architecture of developmental disorders in general is the theme in Chapter 4, where I present the DDD Study and my contributions to this project, including an investigation into the contribution of recessive causation to developmental disorders.

In the Chapter 5 (the final chapter) I will highlight the themes running throughout this dissertation, including dominant versus recessive inheritance, loss of function versus missense variants, the use of next generation sequencing to unravel the underlying causes of developmental disorders and challenges in assigning pathogenicity to variants.

In summary, in this chapter I have introduced developmental disorders and the approaches I will be taking to investigate them. I have detailed how knowledge of genetic disease has improved over time with advancing technology, but that clinical assessment for dysmorphic features has often been vital in gene discovery and continues to be important in the interpretation of variants from broad approaches to sequencing (whole exome and whole genome sequencing). Finally, I have outlined the investigations I will present in this dissertation.

Chapter 2

Wiedemann-Steiner syndrome resulting from mutations in *KMT2A*: A Genotype-phenotype study

2.1 Aims

- **To investigate the phenotype of Wiedemann-Steiner syndrome (WSS) resulting from KMT2A mutations**
- **To investigate the spectrum of mutations in *KMT2A* associated with WSS**
- **To investigate how missense mutations potentially affect *KMT2A* function and cause WSS**
- **To investigate whether experienced clinical geneticists can distinguish the facial appearance of individuals with *KMT2A* mutations from individuals with undiagnosed developmental disorders**

2.2 Introduction

2.2.1 Wiedemann-Steiner syndrome (WSS) and motivation for this investigation

Wiedemann-Steiner syndrome (WSS) is an autosomal dominant multiple congenital-anomaly syndrome associated with a distinctive facial appearance, developmental delay and hypertrichosis(70-72). Since the discovery that *de novo* mutations in *MLL* (now called *KMT2A*) underlie WSS(46) only 15 further individuals with WSS and *KMT2A* mutations have been reported(73-81). These individuals were reported in single case reports or in small case series. Therefore, the full phenotypic and mutational spectrum of WSS remains unknown and as a result, individuals with WSS may not be correctly diagnosed and may have unidentified medical needs.

The ability of clinicians to recognize distinctive facial features has played an important role in the diagnosis of genetic syndromes. As more and more patients with developmental disorders undergo whole exome sequencing to achieve a diagnosis, there will be increasing numbers of individuals identified with missense variants in *KMT2A* in whom a diagnosis of WSS had not been suspected. It is not fully understood how missense variants in *KMT2A* cause WSS and there are missense variants in *KMT2A* in healthy individuals in control databases such as The Exome Aggregation Consortium (ExAC) control database. Therefore, it is vital that missense variants in *KMT2A* are interpreted accurately to correctly diagnose and manage individuals with WSS. Therefore, determining the extent to which clinicians can truly distinguish WSS from other developmental disorders would be useful knowledge when determining the

pathogenicity of *KMT2A* variants identified on whole exome sequencing, in particular missense variants.

2.2.2 Wiedemann-Steiner syndrome (WSS): Definition

In 1989, Wiedemann reported a boy with pre- and post- natal growth deficiency, developmental delay and a distinctive facial appearance(70). His facial features included a round, flat face, hypertelorism, a long philtrum, short palpebral fissures, low set ears and a high arched palate. In addition, he had strabismus and dilatation of the renal calyces(70).

Subsequently, in 2000 Steiner & Marques reported an eight-year-old girl with similar phenotypic features to the individual reported by Wiedemann(2). She had hypotonia, short stature, an unusual facial appearance and intellectual disability. Her facial features included mild synophrys, telecanthus, narrow and down-slanting palpebral fissures, a low nasal bridge, a long and flat philtrum and a thin upper lip. She had a sacral dimple and a high arched palate(71)(Figure 2-1). She had mild hypertrichosis (increased hair) of her arms, legs and back which became accentuated with age(71)(figure 2-2).

Figure 2-1: Facial appearance of the girl reported by Steiner and Marques in 2000(71)
Facial appearance demonstrates telecnathus, mild synophrys, narrow downslanting palpebral fissures, a low nasal bridge, a low nasal bridge and flat and long philtrum. Reproduced from Steiner CE, Marques AP. Growth deficiency, mental retardation and unusual facies. Clin Dysmorphol. 2000;9(2):155-6. Figure 1. Frontal view of the patient's face. Reproduced with permission, copyright Lippincott, Williams & Wilkins.

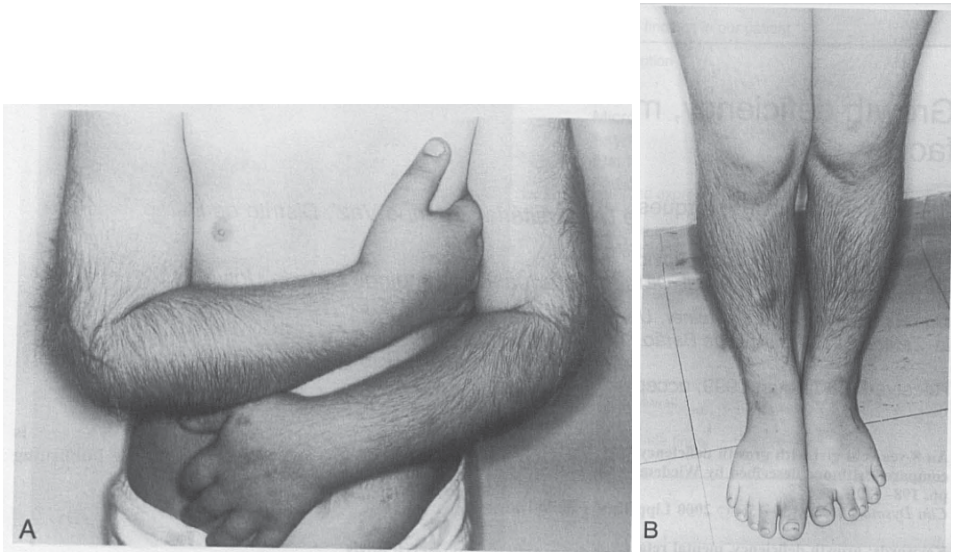
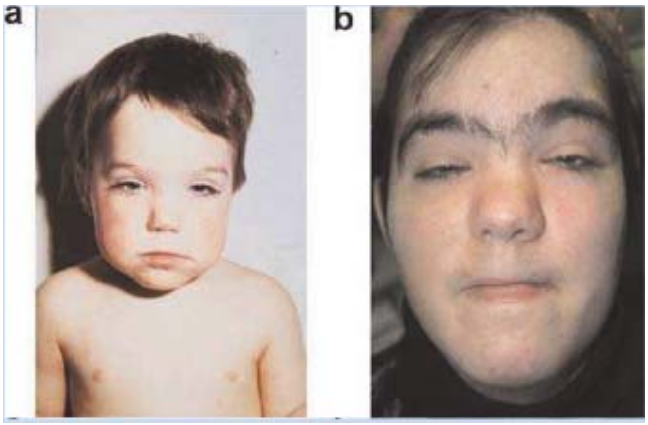


Figure 2-2: The arms and legs of the girl reported by Steiner and Marques in 2000(71)
The arms (A) and legs (B) showing hypertrichosis. Reproduced from Steiner CE, Marques AP. Growth deficiency, mental retardation and unusual facies. Reproduced from Clin Dysmorphol. 2000;9(2):155-6. Figure 2. Hypertrichosis on arms (a) and legs (b). Reproduced with permission, copyright Lippincott, Williams & Wilkins.

Koenig et *al* coined the name 'Wiedemann-Steiner syndrome' in a report on three individuals with a distinctive facial appearance and severe developmental delay(70-72)(Figure 2-3). They felt all three individuals had a similar phenotype to the individuals reported by both Wiedemann and Steiner(70, 71). Facial features common to the individuals reported by Koenig *et al* included: arched or thick eyebrows, hypertelorism, narrow palpebral fissures, and broad nasal bridge and tip(72). All three individuals developed hypertrichosis(72).

I

II



III

Figure 2-3: I-III: Facial appearance of the individuals reported by Koenig et al in 2010(72). Facial appearance demonstrates arched and or thick eyebrows, hypertelorism, narrow palpebral fissures, broad nasal bridge and tip. Reproduced from Koenig R, Meinecke P, Kuechler A, Schafer D, Muller D. Wiedemann-Steiner syndrome: three further cases. American journal of medical genetics Part A. 2010;152A(9):2372-5. Figure I a, b from Figure 1a, b: Patient 1 at the age of 2 7/12 and 12 4/12 years. Figure II a, b from Figure 3 a, b: Patient 2 at the age of 2 ½ years and 20 years. Figure III a, b from Figure 5: Patient 3 at the age of 19 months and 6 8/12 years. Reproduced with permission, copyright John Wiley and Sons.

2.2.3 *De Novo* mutations in *MLL* cause Wiedemann-Steiner syndrome

In 2012, my colleagues and I demonstrated that de novo mutations in the histone methyltransferase *MLL* (subsequently renamed as *KMT2A*)(82) underlie a distinct phenotype consistent with a diagnosis of WSS(46). This phenotype consisted of hypertrichosis cubiti (excessive hair of the elbow regions), short stature, intellectual disability and a distinctive facial appearance(46). Other associated features observed in the five reported individuals with *KMT2A* mutations were feeding difficulties, behavioral difficulties, skeletal abnormalities and cardiac defects.

2.2.4 15 further individuals with *KMT2A* mutations and WSS reported

Since the discovery that *de novo* mutations in KMT2A cause WSS there have been 15 further individuals with WSS and heterozygous *KMT2A* mutations reported in case reports and case series(73-81). These have increased the phenotypic spectrum of features associated with WSS and *KMT2A* mutations to include epilepsy, renal abnormalities, microphthalmia, congenital immune deficiency and premature eruption of dentition(73, 75, 76, 78). The largest case series reported since the initial gene discovery was by Miyake *et al*, who reported 5 individuals with *KMT2A* variants(74). As well as singletons, Dunkerton *et al* described monozygous twins with WSS and a heterozygous nonsense mutation in *KMT2A*; the twins shared the majority of their observed phenotypic features relating to WSS, differing only in subtle facial dysmorphism(78).

In addition to individuals reported in detailed case reports and case series *KMT2A* variants have been identified in six individuals in large consortia sequencing projects seeking to investigate epilepsy and autism. These include three individuals with epileptic encephalopathy, one individual with unclassified epilepsy (83, 84) and two individuals with autism(85). There is no information available as to whether these individuals have other phenotypic features of WSS.

2.2.5 The mutational spectrum of *KMT2A* mutations in WSS

The individuals reported in my previous work all had frameshift or nonsense mutations in *KMT2A*(46). Using real time PCR (RT-PCR), we showed reduced *MLL (KMT2A)* expression in primary skin fibroblast cells derived from an individual with a *de novo* frameshift mutation in *KMT2A* (c.6913del, p.Ser2305LeufsTer2) compared to unrelated healthy controls. We proposed haploinsufficiency as the disease mechanism(46) reporting their findings were consistent with the notion that transcripts arising from the mutant *MLL (KMT2A)* alleles are subject to nonsense-mediated decay. Subsequently, Strom *et al* expanded the mutational spectrum of *KMT2A* mutations associated with WSS to include a *de novo* variant predicted to affect splicing (c.4086+G>A) and a *de novo* missense mutation (c.4342T>C, p.Cys1448Arg)(76). Mendelsohn *et al* further expanded the observed mutational spectrum to include a multi-exon deletion (of exons 2 to 10) in a girl with WSS(73).

Stellaci *et al* proposed a possible genotype-phenotype correlation for *KMT2A* mutations. They reported a boy with congenital immunodeficiency with low levels of immunoglobulins(75) and severe epilepsy and a *de novo* missense mutation c.3481T>G (p.Cys1161Gly). The c.3481T>G mutation is predicted to disrupt a residue located within the cysteine-rich CXXC DNA binding domain of KMT2A. Stellacci *et al* noted that their individual and two of the other reported individuals who had missense mutations predicted to affect one of two functional domains of KMT2A (the CXXC DNA binding domain and the plant homeo-domain (PHD) zinc finger motif) had a tendency towards infections(74, 76). They proposed that missense mutations affecting these functional domains in a mechanism different from haploinsufficiency might specifically impact the transcriptional control of genes involved in the regulation of haematopoiesis and immune functions(75).

2.2.6 WSS overlaps phenotypically with other developmental disorders

A number of authors have reported phenotypic similarities between WSS and other developmental disorders resulting from mutations in genes that play a role in modifying chromatin structure (chromatin disorders) (46, 74, 77, 79). Jones *et al* and Miyake *et al*

reported similarities of the WSS phenotype to that of Kabuki Make-up syndrome, another congenital multiple anomaly syndrome(46, 74). In fact, three of the six individuals with KMT2A mutations reported by Miyake *et al* had initially been diagnosed with 'atypical' Kabuki Make-up syndrome(74). Kabuki Make-up syndrome results from mutations in the histone methyltransferase KMT2D or the lysine specific demethylase KDM6A(42, 86). KMT2D is a histone methyltransferase which adds trimethylation to histone H3 at lysine 4(87). KMT2D facilitates gene expression through acting as a transcriptional coactivator through interacting with transcriptional machinery at the promoters of target genes to facilitate gene expression(87). KDM6A is a demethylase that removes trimethylation from H3K27 a closed chromatin mark(88). Therefore, KMT2D and KDM6A have complementary functions and loss of function mutations in the genes encoding these enzymes leads to similar phenotypes(89). Yuan *et al* identified a heterozygous *de novo* nonsense mutation in an individual they felt had overlapping features with Cornelia de Lange Syndrome(CdLS)(77). CdLS results from mutations in the *HDAC8* gene (a histone deacetylase) or in genes encoding proteins affecting cohesin structure or function. CdLS is associated with hypertrichosis, short stature, limb defects and a distinctive facial appearance(90-94). Finally, Bramswig *et al* reported a child with a *de novo KMT2A* missense mutation who had overlapping features with Coffin Siris syndrome(79). Coffin Siris syndrome is a multiple congenital anomaly syndrome associated with hypertrichosis, developmental delay, agenesis of the corpus callosum and nail hypoplasia. Coffin Siris syndrome is caused by mutations in genes encoding components of the SWI/SNF chromatin remodelling complex(44, 45).

2.2.7 KMT2A is a histone methyl-transferase with multiple domains

KMT2A encodes the histone methyltransferase enzyme KMT2A, which is expressed in most cell types(82, 95). The KMT2A protein is a large (3,969 aa) multi-domain protein(96) which is one of a family of histone–lysine N-methyltransferase 2 (KMT2) proteins. The KMT2 proteins, including KMT2A, are highly conserved(97) and they play an important role in epigenetic regulation by methylating lysine 4 on the histone H3 tail (H3K4) to modify the structure of chromatin and promote DNA accessibility.

The nomenclature of the KMT2 family of proteins was updated in 2012; they were previously called the MLL family (82). The re-naming sought to alleviate confusion with the previous nomenclature given to these proteins: some enzymes had the same name, and the previous naming system did not reflect the complex function of these enzymes(82).

KMT2A generates mono-, di-, and trimethylated histone H3K4, through its SET domain and interaction with cofactors (Reviewed by Rao *et al(98))*. Mono-, di-, and trimethylated histone H3K4 have been shown to regulate chromatin-mediated transcription, including the transcription of multiple Hox and Wnt genes(99). The protein binding partners of KMT2A are discussed in more detail in Chapter 3: Gene discovery in hypertrichosis. The SET domain, which is responsible for KMT2A's histone-methyltransferase activity(100), is located at the C-terminus of the KMT2A protein (Figure 2-5). This is followed by, a WDR5 interaction (Win) motif, a nuclear receptor motif, a transactivation domain, a bromo domain, and plant homeodomain finger motifs(101). This is followed by a cysteine-rich CXXC domain which binds to unmethylated CpG dinucleotides(102). At the N-terminus there are three AT hooks, which are small DNA-binding protein motifs that also interact with chromatin(100, 102). KMT2A is proteolytically processed into N-terminal and C-terminal fragments at two independent sites(103). The cleaved fragments then form a stable complex that localizes to a sub-nuclear compartment(103).

2.2.8 Facial recognition investigations in dysmorphic syndromes

The delineation of WSS as a developmental disorder and its subsequent genetic characterisation relied on clinicians identifying the distinctive facial appearance of affected individuals(70-72). Clinicians have been recognising features unusual for age and ethnicity (dysmorphic features) and using them to help make genetic diagnoses since at least 1862 when Down syndrome was first described by Down(104). However, only a few investigations have been carried out to determine the accuracy with which clinicians are able to recognise individuals with specific genetic developmental disorders(105-108). No facial recognition experiments have been carried out in individuals with WSS and KMT2A mutations to confirm the facial appearance of

individuals with WSS can be distinguished by clinical geneticists from other developmental disorders.

Facial recognition experiments have given insight into the facial features used by clinicians to make certain diagnoses, highlighted disorders that can successfully be recognized by facial appearance as well as helped guide diagnostic criteria(105-108). These studies have also helped confirm that facial features of developmental disorders can change with age(107). The groups of clinicians used in these facial recognition experiments have included trained Dysmorphologists, and Hepatologists, who in general, would be expected to make few diagnoses based on the facial appearance of their patients. These experiments have involved clinicians being assessed at recognising the facial appearance of Alagille syndrome or Cornelia de Lange Syndrome (CdLS) (for a definition of CdLS, please see above)(105-108). Alagille syndrome is an autosomal dominant developmental disorder associated with cholestasis, cardiac anomalies, skeletal anomalies and other features.

In 2010, Rohatgi *et al* investigated the ability of Dysmorphologists to distinguish individuals with Cornelia de Lange syndrome (CdLS) from individuals with developmental disorders with overlapping features to CdLS, including fetal alcohol syndrome, Floating-Harbor syndrome and Kabuki Make-up syndrome(107). Floating-Harbor syndrome is named after the two hospitals in which the first individuals with the disorder were identified, namely the Boston Floating hospital and the Harbor General Hospital in California(109-111). They asked the clinicians to score each of the individuals as 'Classic', 'mild' or 'non-CdLS' as well as their certainty of the answer. The authors showed, that using facial photographs alone correct diagnoses were made in 90% of classic CdLS and 87% of non-CdLS individuals(107). However, the Dysmorphologists diagnosed only 54% of individuals classified as mild or variant CdLS correctly(107). The Dysmorphologists were asked to document the facial features they used in drawing their conclusions, this enabled the authors to assess the utility of specific features used to make a diagnosis of CdLS. For example, pencilled arched eyebrows, synophrys, long eyelashes, and thin upper lip were features helpful to making a diagnosis of CdLS in individuals with mild CdLS resulting from NIPBL mutations, however heavy straight eyebrows distracted from a diagnosis of CdLS in individuals with

SMC1A mutations(107). As a result of their findings Rohatgi *et al* suggested modifications to the previously used clinical diagnostic criteria for CdLS may be needed(107).

Computational approaches to facial recognition, including the use of machine learning have been developed to recognise individuals with CdLS and other developmental disorders from two dimensional (2D) photographs(112, 113). Ferry *et al* demonstrated that their machine learning driven method of 'Clinical Face Phenotype Space' was able to discriminate between syndromes with a similar accuracy to earlier methods which had utilised three dimensional (3D) image capture(113, 114).

More recently, Basel-Vanagaite *et al* repeated the experiment of Rohatgi *et al* using the same photographs of individuals with CdLS or overlapping disorders. They showed that the average detection rate of the automated Facial Dysmorphology Novel Analysis (FDNA) technology was 87%(112) compared to 77% by the Dysmorphologists in the study of Rohatgi *et al(107)*.

2.2.9 Questions studied in this investigation

There remain a number of unanswered questions about WSS resulting from *KMT2A* mutations, as the largest studies to date have contained only 5 individuals and no one has carried out a large phenotypic study of individuals with WSS. These include: What is the wider phenotypic and mutational spectrum associated with WSS resulting from *KMT2A* mutations? What is the incidence of WSS resulting from *KMT2A* mutations? And in an era where clinicians are interpreting variant findings from next generational sequencing analysis, is WSS caused by *KMT2A* mutations a facially recognisable developmental disorder? It is also not clear whether *KMT2A* mutations do underlie WSS as none of the original individuals reported under this classification have undergone sequencing.

2.3 Methods

2.3.1 Recruitment to the study was genotype or phenotype driven

I identified 84 individuals with *KMT2A* mutations through genotype- or phenotype-driven recruitment from 15 different countries. Genotype-driven recruitment involved the recruitment of individuals with a heterozygous *KMT2A* mutations and a phenotype consistent with WSS. These individuals were recruited either following diagnostic exome sequencing or targeted capillary sequencing of *KMT2A* in their local genetics centre or from a variety of separate research projects including the Deciphering Developmental Disorders (DDD) study, a study investigating individuals with atypical Cornelia de Lange syndrome and a study investigating individuals with Pierpont syndrome. The Deciphering Developmental disorders (DDD) study is collaboration between all the Clinical Genetics units in the United Kingdom and Ireland and the Wellcome Trust Sanger Institute(48), further information about the DDD study is given in Chapter 4: The Deciphering Developmental Disorders Study / Investigations into Autosomal Recessive Developmental Disorders. I considered variants to be pathogenic if they were predicted to result in a loss of protein function (nonsense, frameshift, splice donor and splice acceptor variants or multi-exonic deletions). I considered missense variants to be pathogenic if they were de novo and the phenotype fitted with the published phenotype associated with WSS or if the variant was identical to a variant that had been classified as pathogenic in another individual, or if an individual who had been diagnosed clinically as having WSS was subsequently found to have a *KMT2A* missense variant on sequencing.

For Phenotypic-driven recruitment, I identified 247 individuals with phenotypic features consistent with Wiedemann-Steiner syndrome (as assessed by a Clinical Geneticist) and / or with evidence of increased body hair. Individuals were recruited with one or both parents (duos or trios). Recruitment criteria, including the Human Phenotype Ontology(58) terms used to select patients with increased body hair are listed in given in Chapter 3: Gene discovery in hypertrichosis. 228 of the individuals underwent whole exome sequencing as part of the Deciphering Developmental Disorders study. The 20 remaining individuals were recruited from outside the UK and underwent whole exome sequencing separately at the Wellcome Trust Sanger Institute (WTSI). I identified

individuals with heterozygous pathogenic *KMT2A* mutations (this included variants predicted to result in loss of protein function and missense variants not present in the ExAC database) who were then carried forward for detailed phenotypic study. For DDD study individuals they were only carried forward for further study if the variant had already been clinically reported. For full methods, please see Chapter 3: Gene discovery in hypertrichosis.

2.3.2 Sequencing methods

Each of the different studies from which individuals were recruited or local diagnostic exome or *KMT2A* capillary sequencing had different sequencing methods. For individuals undergoing sequencing as part of the Deciphering Developmental Disorders (DDD) study or the WiSH Study, please see Chapter 3: Gene discovery in hypertrichosis for details of whole exome sequencing methods. Other individuals either underwent whole exome sequencing or targeted capillary sequencing of *KMT2A.*

For individuals not sequenced as part of the DDD study or the WiSH study the KMT2A variant data for each individual, including at least, the DNA sequence change and the predicted protein effect were provided by the patient's clinician. Transcript information and genomic co-ordinates were not available for every variant. I first obtained the correct DNA sequence change for each variant using Human Genome Variation Society (HGVS)(115) nomenclature system for the transcript ENST00000534358.1 and genomic co-ordinates for each variant. To do this I used Genome Reference Consortium human genome build 37 (GRCh37) in the Ensembl Variant Effect Predictor (VEP) programme(116) (release 88, March 2017). As a final check, I used Mutalyzer(117) to confirm the nomenclature of all variants was in line with HGVS nomenclature.

2.3.3 Phenotype analysis of individuals with WSS and *KMT2A* mutations

I devised a phenotype questionnaire which employed a general and targeted approach to capturing phenotype data of individuals with WSS and *KMT2A* mutations (Appendix 1). The questionnaire included general systems based questions, such as: "Urogenital abnormalities?", as well as a more targeted approach to gather negative data for key features of the disorder, for example: "Premature eruption of dentition?" I devised the

final two questions on the questionnaire (do they take any medications? And have they every been admitted to hospital or attended the emergency department?) to gather information on significant illnesses that may not have been elicited by the earlier part of the questionnaire.

To maximise clinical data given clinicians' busy clinical work-load, I calculated that a 2-sided A4 questionnaire was the optimum length. I used short bulleted questions for succinctness. In addition, to capture accurate 'negative data', I gave the options of: 'Yes', 'No', and 'Unknown', to enable clinicians to record that information about that specific feature was unknown. For examination findings, I gave the options of 'Yes', 'No', and 'Not Assessed' to enable clinicians to record that clinical findings had not been specifically looked for on clinical examination and to accurately quantify negative data. Devising the questionnaire was an iterative process and I added further questions, where necessary, as further phenotypic associations were determined.

All individuals entered the study with some growth and phenotypic information either provided by their local clinician via email, or with HPO coded phenotypic information from the Decipher database for the individuals recruited who were part of the DDD study(118). For further information about the Decipher database please see Chapter 4: The Deciphering Developmental Disorders Study / Investigations into Autosomal Recessive Developmental Disorders.

I visited 32/84 of the individuals in their local genetics centre with their Genetics clinician to carry out detailed clinical phenotyping. I asked the families questions based on the phenotype questionnaire and examined the affected individual with their local genetics clinician. 48 further individuals underwent more detailed phenotypic analysis by their local clinician completing the phenotype questionnaire I devised to capture their phenotypic data (appendix 2). Three further individuals did not have a phenotype questionnaire completed by their local clinician, including one individual in the DDD study and one individual who had been previously published in the medical literature for whom there was a significant amount of phenotypic information available.

In order to investigate the facial features of individuals with *KMT2A* mutations and WSS, for consistency, I reviewed facial photographs of 67 of the individuals with WSS and *KMT2A* mutations, including the individuals I had previously examined. I recorded the facial phenotypic features present. If I didn't feel I could confidently assess a phenotypic feature form the photograph, e.g. Synophrys may be difficult to assess in a blond haired individual, I recorded unknown. However, I also took into consideration any facial phenotypic features recorded by the local clinician or those I had recorded when examining each of the individuals in this case, e.g. if synophrys had been recorded on examining them I would record this.

2.3.4 Selection of missense variants for protein modelling.

To investigate how predicted missense mutations may potentially affect KMT2A function and cause WSS, I selected *de novo* missense mutations from my cohort and from the published literature to investigate the potential effect they may have on the 3D structure of KMT2A and how this may impair its biochemical function. It may be that one or more of these variants are pathogenic because they affect, however in this investigation I focused on the possible effect of these variants on the 3D protein structure of *KMT2A*. For the individuals from the medical literature with a phenotype consistent with WSS, I included only individuals reported in case reports or small case series who had *de novo* mutations and published photographs in order to give greater confidence that their KMT2A variant was pathogenic and their phenotype fitted with WSS.

In addition, I included two *de novo* missense variants from large consortia sequencing studies. This included one missense variant reported by de Rubeis *et al* in an individual with autism(85), and one missense variant identified in an individual with epilepsy reported by the Functional genomic variation in the epilepsies research project (EuroEPINOMICS-RES) Consortium *et al* (119). Although there was limited phenotypic information available about these individuals, as the phenotypes autism and epilepsy are observed in individuals with WSS I included them, however it is uncertain whether the phenotype of these individuals is consistent with WSS.

2.3.5 I selected control missense variants from the ExAC database

In order to compare the pathogenic missense mutations I had selected for protein modelling to background normal variation, I generated a control dataset of missense variants from the ExAC database(version 0.3)(120). I selected missense variants in the KMT2A transcript ENST00000534358 with a frequency of >1%. I selected only common variants (frequency >1%) in order to remove variants present in only a single individual and therefore reduce the chance of sequencing error. In addition to remove sequencing errors, I removed Non-PASS variants. In addition, I filtered out variants with multiple reference or alternate alleles.

2.3.6 Clinician recognition of facial features associated with WSS

To determine whether experienced Clinical Geneticists can distinguish the facial appearance of individuals with pathogenic *KMT2A* variants from *KMT2A*-negative individuals with developmental disorders I carried out a facial recognition study. I showed six Clinicians (four consultant Clinical Geneticists and two trainee clinical geneticists) 27 separate facial photographs consecutively in a projected Microsoft PowerPoint presentation. The size of this investigation was limited by the time that could be reasonably be requested of the clinicians for participation. Ideally a larger number of images than 27 would have been used. 19 of these photographs showed individuals with *KMT2A* variants and eight photographs showed randomly selected individuals with developmental disorders from the DDD study. Not all of the individuals in the facial recognition study were part of this cohort of 84 individuals. Also of the 19 individuals with KMT2A variants, 3 of these variants were missense variants identified on whole exome sequencing in individuals in whom a diagnosis of WSS was not certain, of these one was *de novo* and two were inherited. Given Clinical Geneticists are busy individuals, I decided that 27 photographs would be the optimal number of photographs to use to ensure an adequate number of clinicians would be in agreement to take part in the investigation.

I asked each of the clinicians to score each of the facial photographs from 0 (they do not feel the face resembles WSS at all) to 10 (it is a face that they would use as an exemplar of WSS for teaching/training).

2.3.7 Estimation of the incidence of Wiedemann-Steiner syndrome

In order to estimate the incidence of WSS I used the published SNV loss of function *de novo* mutation rate of *KMT2A*(121). I added this to the INDEL loss of function rate to give a per chromosome rate and then doubled this to obtain a per child rate. Finally, I took into account factors that may increase or decrease this rate to give an estimated range for incidence of WSS.

2.3.8 Ethical approval

Each family provided signed consent and ethical approval for this study was obtained from Guy's and St. Thomas' National Health Service (NHS) Foundation Trust local research ethics committee (ref.:08/H0802/84), ''Systematic Characterization of Genes in Inherited Disorders".

2.4 Results

2.4.1 I identified 84 individuals with *KMT2A* mutations

I identified 84 individuals with heterozygous *KMT2A* mutations and a phenotype consistent with WSS, including 73 mutations predicted to result in loss of function of the protein product, 10 missense mutations and one inframe deletion. My cohort consisted of 82 singletons and one set of monozygous twins. One of the individuals, a 51-year-old man (WSSP37), appeared to have a mosaic *KMT2A* frameshift mutation (c.3697delG p.Val1233LeufsTer2). All other 83 individuals from routine testing were understood to have germline *KMT2A* mutations, however deeper interrogation for mosaicism was not carried out. The age of individuals at their last clinical assessment ranged from 1 year 4 months of age to 51 years of age with a mean age of 10.99 years and a median age of 10.08 years. My cohort consisted of 37 males (44%) and 47 females (56%), which is not significantly different from balanced.

2.4.2 98.6% of mutations were *de novo* where inheritance was known and one mutation was inherited from a mosaic father

Where inheritance information was available (70 individuals), 69 mutations were identified as being d*e novo* (98.6%). I conclude from this that WSS is nearly completely penetrant and it is highly likely that the chance of reproduction is reduced in WSS individuals, this is likely due as least in part to learning difficulties and there may be other factors involved.

The one inherited mutation (c.3697delG p.Val1233LeufsTer2) in individual WSSP36 was identified from her 51-year-old father (WSSP37) who was mosaic for this mutation. For details of the sequencing method, please see Chapter 3: Gene discovery in hypertrichosis. Essentially, the mutation was not called in the exome variant profile from either the mother or father using GATK and therefore it appeared *de novo* in the proband. However, when I reviewed the Integrative Genomics Viewer (IGV) Plots for this trio, I detected that her father's variant profile showed 127 reads showing a G at position 118352491127 (73 forward reads, 54 reverse reads) and 11 reads showing a deletion at this position (Figure 2-4). The probands phenotype was consistent with WSS and her father also had some milder phenotypic features of WSS, as discussed further below.

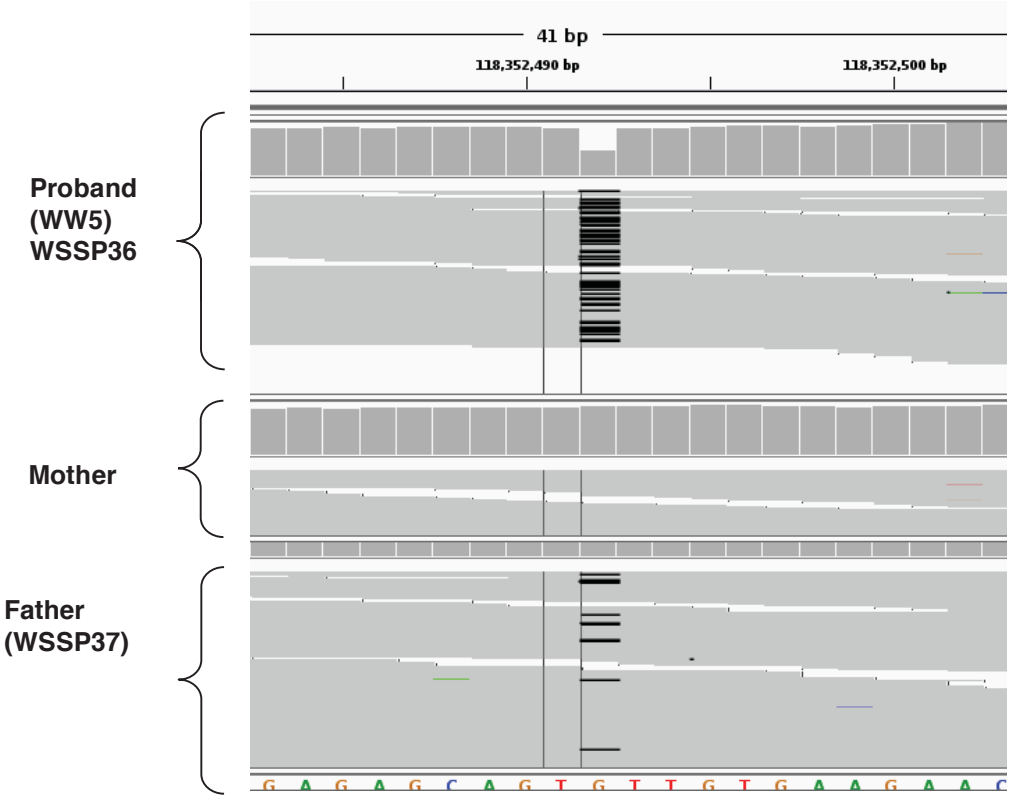


Figure 2-4: Integrative Genomics Viewer Plot showing Mosaic *KMT2A* variant
This Integrative Genomics Viewer (IGV) plot shows reads from proband WW5 (WSSP36) and her mother and father (WSSP37). The proband carries a single nucleotide deletion of a G (guanine) at position Chr11:118352491 (c.3697delG). This is predicted to result in the protein change p.Val1233LeufsTer2. The proband has 56 reads showing a G at position 118352491 (33 forward reads, 23 reverse reads) and 50 reads showing a deletion at this position. Her father has 127 reads showing a G at position 118352491127 (73 forward reads, 54 reverse reads) and 11 reads showing a deletion at this position, consistent with mosaicism.

2.4.3 I confirmed that *de novo* mutations in *KMT2A* cause Wiedemann-Steiner syndrome

None of the original patients reported under the classification of WSS by Wiedemann-Steiner or Koenig had been previously confirmed to have *KMT2A* mutations. Three of the individuals I present in this work (WSSP18, WSSP53 and WSSP58) were the three individuals reported by Koenig *et al*(72) under the classification of Wiedemann-Steiner syndrome (figure 2-3). All three individuals harbored a nonsense mutation in KMT2A (c.3518_3521delGCTT p.Cys1173Ter, c.3790C>T p.Arg1264Ter and c.4012+1G>C). Therefore, I have confirmed that 3 of the 5 individuals originally reported as having WSS have *KMT2A* mutations. Samples from the remaining two individuals originally reported by Wiedemann and Steiner were not available for sequencing.

2.4.4 There were 73 loss of function mutations, 10 missense mutations and one inframe deletion

I identified 36 frameshift mutations, 28 nonsense mutations, 10 missense mutations, 8 mutations predicted to affect splicing, one inframe deletion and one exonic deletion. These mutations are shown in Figure 2-5 with those falling in functional domains displayed also in Figures 2-6(A-C) and given in Appendix 1. There are two familial mutations, one in monozygous twins another in a daughter and her mosaic father. In addition, there are four recurrent mutations: c.2318dupC, p.Ser774ValfsTer12 (identified in WSSP56 and WSSP80), c.3460C>T, p.Arg1154Trp (identified in WSSP33 and WSSP84), c.3790C>T, p.Arg1264Ter(identified in WSSP30, WSSP47,WSSP53, WSSP62), and c.6379C>T p.Arg2127Ter (identified in WSSP59 and WSSP26).

The loss of function mutations are distributed throughout the gene with some clustering in and just after the CXXC zinc finger domain and in the PHD-like zinc-binding domain (zf-HC5HC2H) (figure 2-5). There is clustering of the missense mutations from our cohort and from published reports around the two zinc-finger domains (p=1e-8; this is the lowest value achievable from denovonear which is limited by the maximum number of iterations run, in this case 100 million simulations). The population control variants from the ExAC database are distributed along the length of the gene, with the majority located outside the recognized domain regions. I will discuss the missense variants in detail later in this chapter.

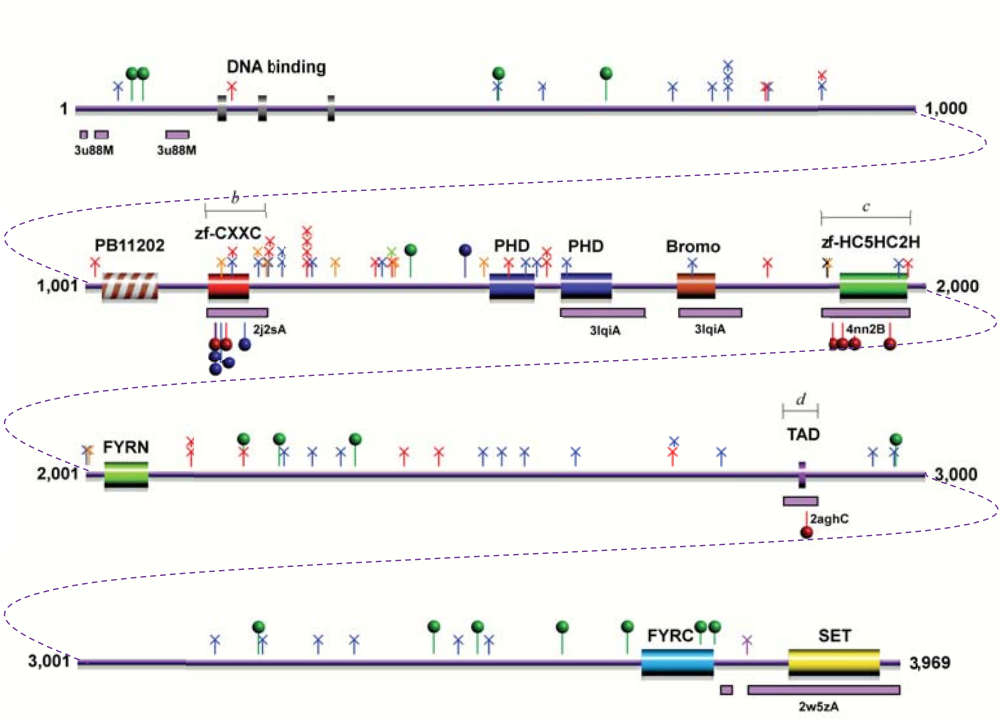


Figure 2-5: Distribution of *KMT2A* variants in individuals with WSS and control missense variants from The ExAC database

A schematic diagram of the KMT2A protein showing its Pfam domains, known 3D structures and KMT2A variants observed in individuals with WSS. The full length of the protein's 3,969 residues is represented by the purple line, cut into four lengths of up to 1,000 residues each. The cylinders along the line represent the Pfam sequence domains: PB11202 (an unclassified Pfam-B domain), zf-CXXC = CXXC zinc finger domain, PHD = PHD finger, zf-HC5HC2H = PHD-like zinc-binding domain, FYRN = F/Y-rich N-terminus, FYRC = F/Y-rich C-terminus, and SET = SET domain. The black-bordered purple bars underneath represent the 3D structural entries in the Protein Databank (PDB) corresponding to that part of the protein. The loss of function (nonsense, frameshift, splice donor and splice acceptor) variants are represented by crosses above the line. Blue crosses represent frameshift variants, red crosses represent nonsense variants, purple crosses represent in-frame deletions, orange crosses represent splice donor/acceptor variants and the green cross denotes the beginning of a multi-exonic deletion. Missense mutations are marked by coloured balls, these are below the line wherever they can be mapped onto a 3D structure. The red balls represent disease-associated missense mutations from this cohort. The blue balls represent *KMT2A* missense mutations reported in the literature. The green balls represent common (>1%) population control missense variants extracted from the ExAC database(120).

2.4.5 The growth profile in WSS resulting from *KMT2A* mutations

Information regarding birth weight was available 35 males and for 43 females (Figure 2-6 and 2-7). The majority of *KMT2A* positive individuals had weights within two standard deviations of the mean (38/43 females, 32/35 males). However, in infancy and childhood the growth charts show evidence for failure to thrive with height and weight in childhood being at the lower end of the normal range or less than two standard deviations (SDs) below the mean in the majority of individuals (Figure 2-8 to 2-11). 48% (16/33) males and 44% (19/43) females have stature >2SDs below the mean (short stature). With increasing age weight becomes more variable between *KMT2A* mutation positive individuals (Figure 2-10 and 2-11). 43% (3/7) of women over 17 years had a weight greater than 2 SDs above the population mean (figure 2-11). Head circumference in *KMT2A* mutation positive individuals was on or below the mean in all individuals, with 39% (12/31) males and 42.5% (17/40) females with a head circumference more than 2SDs below the mean (microcephaly) (figure 2-12 and 2-13).

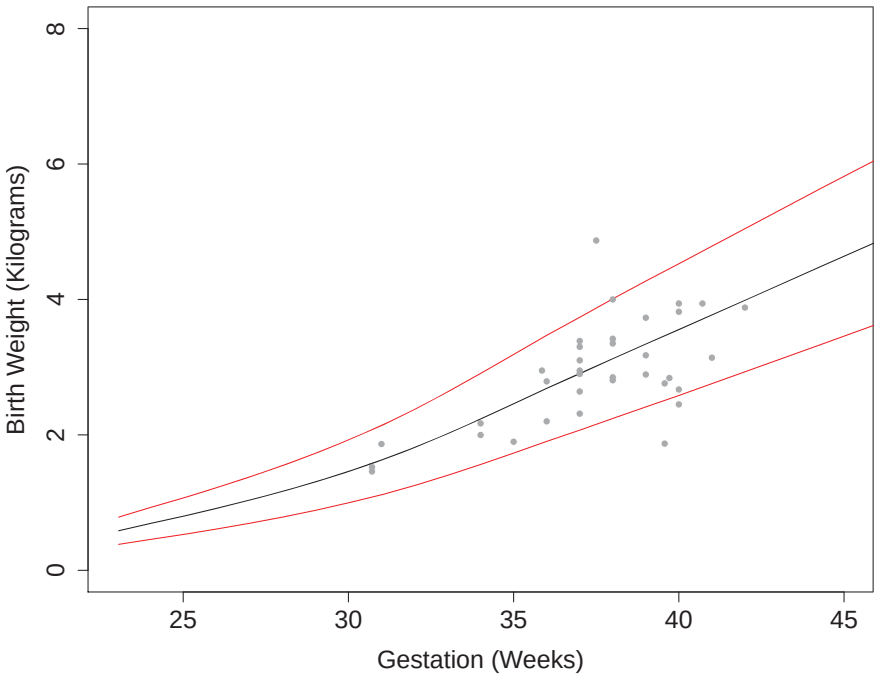


Figure 2-6: Birth weight of males with *KMT2A* mutations (n=35)

Each grey point represents a single individual's weight measurement in Kg at birth. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

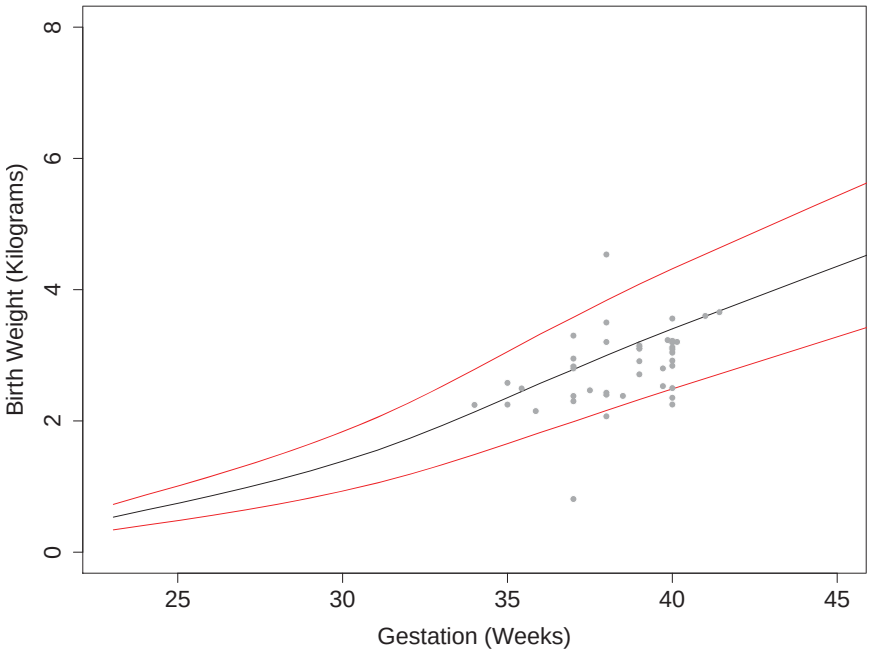


Figure 2-7: Birth weight of females with *KMT2A* mutations (n=43)

Each grey point represents a single individual's weight measurement in Kg at birth. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

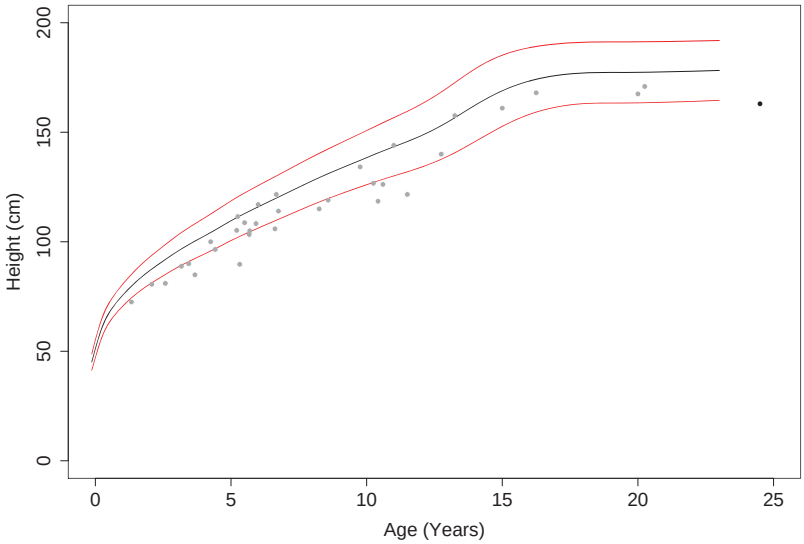


Figure 2-8: Height of males with *KMT2A* mutations (n=33)

Each grey point represents a single individual's height measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black point represents the height of the individual aged 51 years with a mosaic *KMT2A* mutation. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

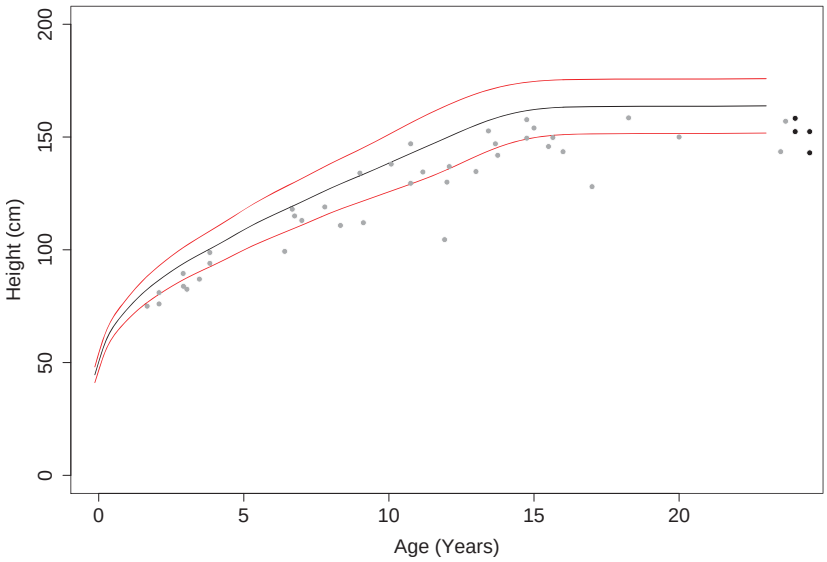


Figure 2-9: Height of females with *KMT2A* mutations (n=43)

Each grey or black point represents a single individual's height measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black dots represent individuals aged greater than 24 years old with jitter. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

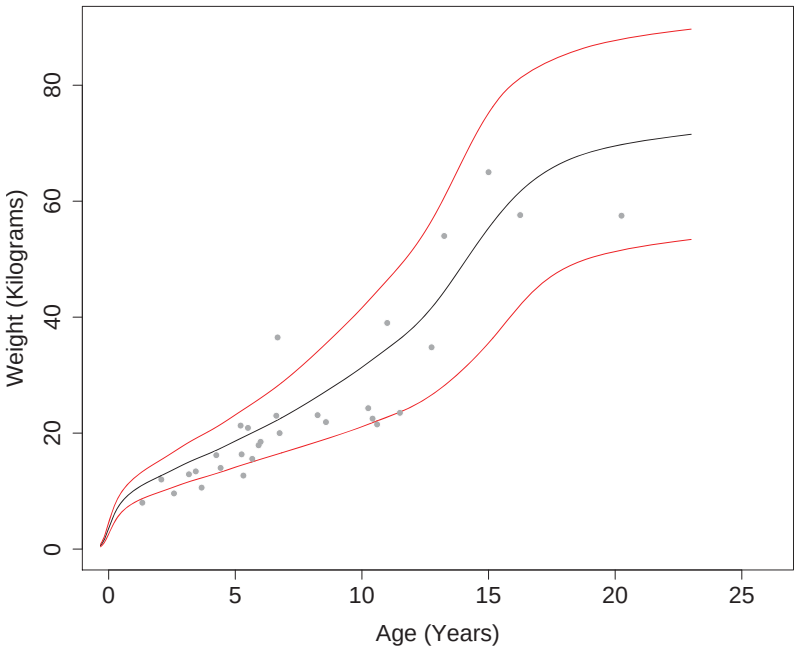


Figure 2-10: Weight of males with *KMT2A* mutations (n=30)

Each grey or black point represents a single individual's height measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

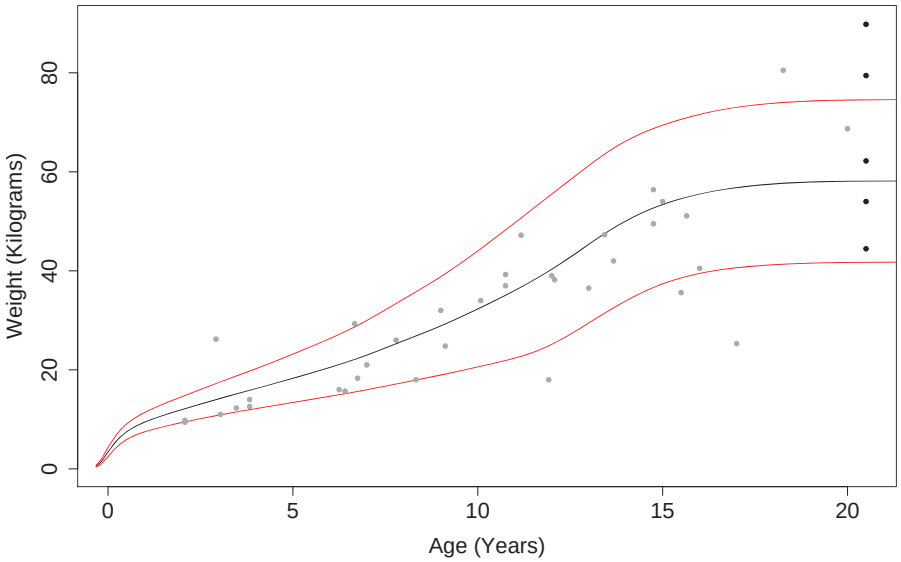


Figure 2-11: Weight of females with *KMT2A* mutations (n=40)

Each grey or black point represents a single individual's height measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black points represent individuals who are greater than 20.5 years old. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

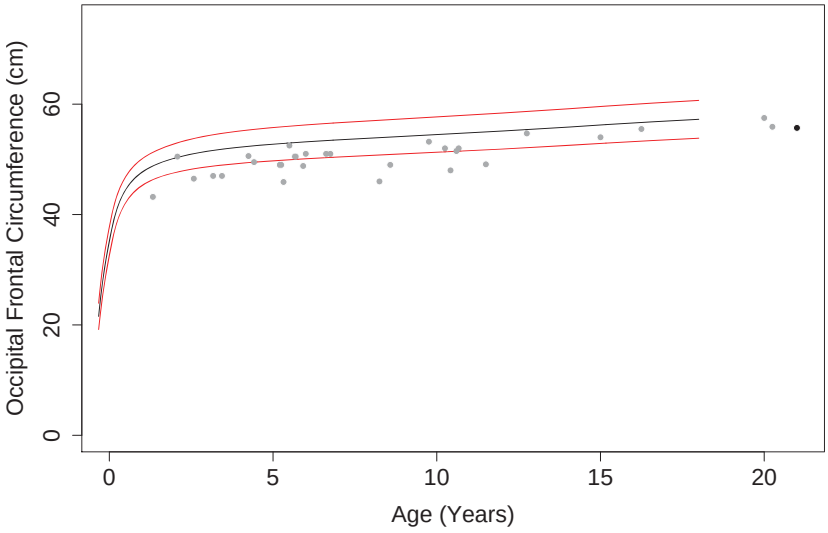


Figure 2-12: Occipital Frontal Circumference of males with *KMT2A* mutations (n=31)

Each grey point represents a single individual's Occipital frontal circumference (OFC) measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black point represents the height of the individual aged 51 years with a mosaic *KMT2A* mutation. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

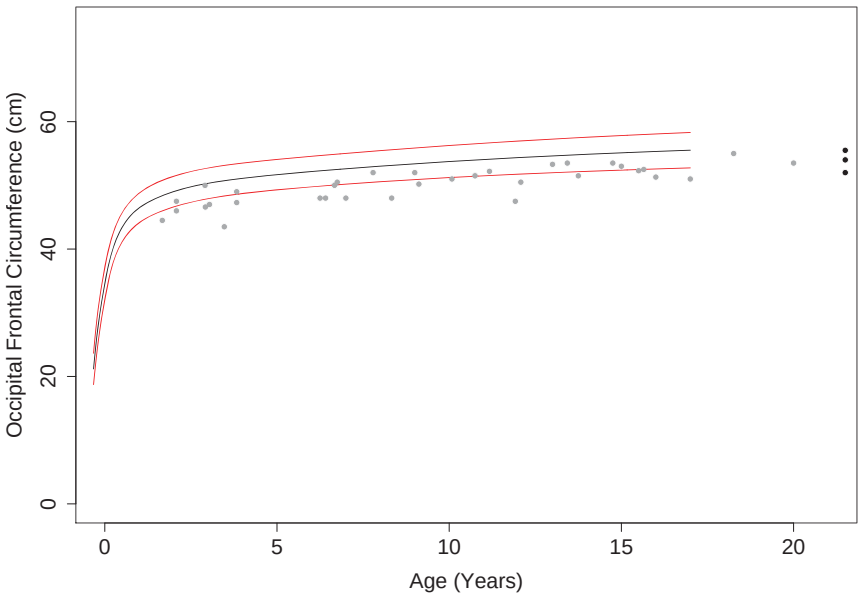


Figure 2-13: Occipital frontal circumference of females with *KMT2A* mutations (n=40)

Each grey or black point represents a single individual's height measurement. Where there was more than one data measurement available, data from when there was most recently a full set of OFC, weight and height data available. The black dots represent individuals aged greater than 24 years old with jitter. The black line represents the UK population mean, the red lines represent mean +/- 2 standard deviations (The British 1990 Growth Reference(122)).

2.4.6 Developmental delay, hypertrichosis, behavioral difficulties, and feeding difficulties were the commonest features

The most common phenotypic features in individuals with *KMT2A* mutations were developmental delay / intellectual disability (all individuals except the mosaic father), hypertrichosis (86%), behavioral difficulties (71%), feeding problems (75%, and requiring PEG or PEJ tube feeding 18%), hypotonia (51%) and constipation (50%), see table 2-1. In terms of facial features, the most commonly observed features were long eyelashes, narrow palpebral fissures, broad eyebrows, flat midface, wide nasal bridge, broad nasal tip and thin upper vermillion border (Table 2-1).

Human Phenotype Ontology Term (Where available)	Feature	Number of Individuals with feature / Number of individuals assessed for that feature	Percentage
Neurological			
HP:0001318	Muscular Hypotonia	43	**51%**
HP:0001250	Seizures	16	**19%**
HP:0011342	Mild global developmental delay	24/72	**33%**
	Mild to moderate global developmental delay	6/72	**8%**
HP:0011343	Moderate global developmental delay	31/72	**43%**
	Moderate to severe developmental delay	4/72	**5%**
HP:0011344	Severe global developmental delay	5/72	**7%**
HP:0012736	Profound global developmental delay	2/72	**3%**
	Mainstream/Regular School(Extra help)	20/61 (14/61)	**33%**
	Special Needs School	40/61	**66%**
HP:0000708	Behavioural abnormality	60	**71%**
	MRI Brain abnormality	14	**17%**
HP:0010832 (HP:0007328)	Abnormality of pain sensation	34	**40%**

HP:0002360	Sleep disturbance	34	**40%**
Gastrointestinal			
HP:0011968	Feeding difficulties	63	**75%**
HP:0011471	Gastrostomy tube feeding in infancy PEJ also	15	**18%**
HP:0011470	Nasogastric tube feeding in infancy (without needing PEG or PEJ feeding)	8	**10%**
HP:0002020	Gastroesophageal reflux	18	**21%**
HP:0002019	Constipation	42	**50%**
Cardiovascular			
HP3000001	Abnormal heart morphology	14	**17%**
Dermatological			
HP:0000998	Hypertrichosis	69	**86%**
HP:0004780	Elbow hypertrichosis	57	**72%**
HP:0011913 HP:0004532 HP:0011914	Lumbar hypertrichosis Sacral hypertrichosis Thoracic hypertrichosis	55	**80%**
	Hypertrichosis of the Lower limbs	54	**78%**
HP:0002219	Facial hypertrichosis	18	**36%**
Urogenital			
HP:0012210	Abnormal renal morphology	9	**10%**
HP:0000811	Abnormal external genitalia	9	**10%**
HP:0000140	Abnormality of the menstrual cycle	9/13	**69%**
Immunological			
HP:0002719	Increased frequency of infection	33	**39%**
Ophthalmological			
	Ophthalmological abnormality	42	**50%**
Ear Nose and Throat			

-	Otitis media with effusion / recurrent otitis media	13	**15%**
HP:0000365	Hearing impairment	11	**13%**
Mouth and palate			
HP:0006292	Abnormality of dental eruption	35	**42%**
Musculoskeletal			
HP:0000960	Sacral dimple	31	**37%**
-	Generalised muscular build	18	**21%**
HP:0007552	Abnormal subcutaneous fat tissue distribution	10	**12%**
	Abnormal planar fat pads	18	**21%**
	Swelling of feet or hands	22	**26%**
HP:0001212	Prominent fingertip pads	23	**27%**
HP:0006191	Deep palmar creases	5	**6%**
Features unusual for age and ethnicity			
HP:0000527	Long eyelashes	52/67	**78%**
	Narrow palpebral fissures	55/67	**82%**
	Eyebrow lateral flare	11/67	**16%**
HP:0011229	Broad eyebrow	56/67	**82%**
	Normal eyebrows	6/67	**9%**
HP:0000508	Ptosis	5/67	**7%**
HP:0000316	Hypertelorism	48/67	**72%**
HP:0040199	Flat midface	53/67	**79%**
HP:0000431	Wide nasal bridge	62/67	**93%**
HP:0000455	Broad nasal tip	64/67	**96%**
HP:0002263	Exaggerated cupids bow upper lip	10/67	**15%**

HP:0000233	Thin upper vermillion border	52/67	**78%**
HP:0012745	Short palpebral fissures	8/67	**12%**
HP:0000637	Long palpebral fissures	26/67	**39%**

Table 2-1: Phenotypic features of 84 individuals with WSS and *KMT2A* mutations

Table showing phenotypic features of 84 individuals with WSS and KMT2A mutations. Where only a proportion of individuals were assessed for a certain phenotypic feature this was stated in column 3, with the denominator reflecting the population assessed for each phenotypic feature. Where Human Phenotype Ontology (HPO)(58) terms are available these are used to code each phenotypic feature.

2.4.7 84% have mild to moderate developmental delay / intellectual disability

Amongst the WSS individuals for whom information about intellect was available 84% (61/72) of individuals were classified as having mild, moderate or mild-to-moderate developmental delay or learning difficulties. With 13% (9/72) individuals classified as having moderate-to-severe or severe developmental delay or learning difficulties and 3% (2/72%) of individuals having profound difficulties. In terms of motor milestones, the mean age of sitting was 11 months, walking was 22 months and mean age of first words was 19 months. Comparing these figures to the normal range adjusted for prematurity (sitting <12 months, walking <21 months and first words <21 months), shows mean age of sitting is within the adjusted normal range however the mean figures for walking and first words are outside the prematurity-adjusted-normal range and delayed.

In terms of education, where information about education is available 66% (40/61) of individuals attend a special needs school and 23% (14/61) of individuals require extra help in mainstream school. There is limited information available about adult outcomes. There are 10 individuals greater than 18 years old in the study. Of the four individuals for whom information is available. One individual works in a mainstream environment, two individuals work in sheltered environment and one individual carries out voluntary work. Three individuals live in sheltered living accommodation with minimal support.

71% (60/84) of individuals have behavioral difficulties. Commonly reported difficulties are abnormal fear / anxiety-related behavior, attention deficit hyperactivity disorder 7% (6/84), autism 26% (22/84), inflexible adherence to routines or rituals 10% (8/84) and anxiety 21% (18/84).

2.4.8 Seizures are associated with poor developmental outcomes

I sought to investigate the individuals with more severe difficulties (severe or profound developmental delay or learning difficulties) and assess whether they had any medical co-morbidities that resulted in their poor developmental outcome. In addition, I had noted in clinic that individuals with seizures often had more severe developmental delay. There is a significant association between seizures and an individual having severe or profound developmental delay / learning difficulties (p=0.0001) (Table 2-2). Other developmental disorders, such as tuberous sclerosis show worse developmental outcomes in individuals with seizures(123). It is possible that the seizures themselves slow developmental progress or affect brain development, however further investigations are needed to determine the cause for the association.

	Mild or moderate delay or learning difficulties	Severe or profound delay or learning difficulties
History of seizures		
Yes	10	6
No	67	1

Table 2-2: Relationship between seizure history and level of learning difficulties N=84

Table to show the association between seizure history and level of learning difficulties in 84 individuals with WSS resulting from KMT2A mutations.

2.4.9 Not all individuals with WSS and *KMT2A* mutations have hypertrichosis

We previously reported all five of our individuals had hypertrichosis(46), in particular referencing the increased hair in the elbow region observed in all individuals with *KMT2A* mutations. In subsequent case reports and case series only two individuals have been

reported not to have hypertrichosis(76, 81). I therefore sought to investigate the proportion of individuals with hypertrichosis. Hypertrichosis was reported in 69/84 (82%) of individuals in my cohort, with 77 clinicians or families asked on direct questioning about the presence of hypertrichosis. Increased hair of the elbow region was observed in 57/84 (68%) individuals, and on the back and legs in 55/84 (65%) and 54/84 (64%) individuals respectively. 18/84 (21%) individuals were noted to have facial hypertrichosis. The incidence of facial hypertrichosis has not previously been investigated in individuals with WSS and *KMT2A* mutations.

2.4.10 Imaging investigations in WSS caused by KMT2A mutations

17% (14/84) of individuals were reported to have congenital heart disease. The true incidence may be higher than this as 46/70 of the remaining individuals have not had an echocardiogram and not all structural cardiac defects produce symptoms in childhood nor can be detected by auscultation using a stethoscope. 11% (9/84) individuals had an abnormality of renal morphology, however only 7/75 of the remaining individuals have had a renal tract ultrasound, therefore abnormalities of renal morphology which can be asymptomatic may have not been detected.

40% of those individuals undergoing brain imaging (34/44) have a structural brain abnormality. However, again the true incidence may be higher than this as only 7 other individuals have had a brain MRI and for 44 individuals there is no information available about MRI brain imaging.

2.4.11 42% of individuals had an abnormality of dental eruption

35 (42%) of individuals were reported to have an abnormality of dental eruption. With the specific abnormalities reported as: advanced eruption of teeth (23), premature eruption of permanent teeth (7), premature loss of primary teeth (6) and persistence of primary teeth (2). Premature eruption of dentition was reported in the individual by Mendelsohn *et al(73)* who had several secondary teeth at four years of age. In this investigation, I expand the phenotype of dental eruption abnormalities to include persistence of primary teeth.

2.4.12 Sleep disturbance is common and may reflect disruption to circadian rhythm

34 individuals (40%) were reported to have sleep disturbance. Disrupted sleep patterns are observed in other developmental disorders such as Smith Magenis syndrome and can cause significant distress to families(124). Increasing evidence shows that chromatin remodelling events play a role in circadian regulation, reviewed by Aguilar-Arnal *et al(125)*. In particular, *KMT2A* has been shown to interact with the core circadian transcription factor complexes CLOCK-BMAL1 and PER-CRY(126). Therefore, the disruption of sleep in WSS may result from an underlying disruption of circadian rhythm. Individuals with WSS also have other phenotypic features that may contribute to sleep disturbance including gastro-oesophageal reflux, tendency towards otitis media with effusion and behavioural difficulties.

2.4.13 Recurrent infections are common in individuals with WSS

39% (33/84) of individuals were reported to have an increased frequency of infections. Increased frequency of infections has been reported previously (46, 74-76). (note the individual reported by Jones *et al* is also included in this cohort). The individual reported by Stellacci *et al* had congenital immune deficiency. The individual reported by Bramswig *et al* died from sepsis at 3 years of age(79). Stellacci *et al* suggested recurrent infections were more common in individuals with missense mutations than loss of function mutations, however I observed no difference in the frequency of increased infections in these groups (p=1.0). Similar disorders, such as Kabuki Make-up syndrome are also associated with an increased susceptibility to infection(127).

I propose the cause of the increased infections in WSS is likely to be multifactorial, with possible contributions resulting from: hypotonia, unsafe swallow, epilepsy, vesico ureteric-reflux and the increased incidence of otitis media. However, it is important to highlight individuals that may have a serious but treatable immune deficiency for further investigation as they may have a congenital immune deficiency. Further work is needed to investigate infections in individuals with WSS and determine the frequency of congenital immune deficiency to guide management and screening.

2.4.14 Other phenotypic features of WSS associated with KMT2A mutations

50% (42/84) of individuals had one or more ophthalmological abnormality, these included strabismus (18), myopia (10) hypermetropia (7) and myopia (10).

26% (22/84) of individuals were reported as having have swollen hands and or feet leading to the suspicion in some that the individual had lymphoedema, however, this has not been confirmed in any individuals in this cohort. In one individual (WSSP10), ultrasound duplex of the extremities was carried out at the age of 6 years to investigate his leg swelling. This was reported as showing an appearance that was felt to be more consistent with excessive subcutaneous fat (or homogeneous swollen subcutaneous fat), than edema. Other features not reported previously, include genital abnormalities in females.

In terms of previously unreported phenotypic features, 40% (34/84) of individuals report an abnormality of pain sensation with 23% (19/84) individuals defining this more specifically as reduced pain sensation. 18% (9/84) of individuals have genital abnormalities, although hypospadias has previously been reported our report expands the spectrum to include genital abnormalities in girls including cliteromegaly and labia minora agenesis. In addition, 2% (2/84) individuals have Raynauld's phenomenon.

2.4.15 A recurrent mutation (p.Arg1154Trp) is associated with seizures

There was no significant difference in the frequency of the most common phenotypic features of WSS (epilepsy, hypertrichosis, behavioral difficulties, feeding difficulties, hypotonia and constipation) between individuals with loss of function mutations versus those individuals with missense mutations. In addition, there was no significant difference between the incidence in the three groups of structural congenital abnormalities (cardiac, renal or brain lesions) sleep disturbance or infections between individuals with loss of function or missense mutations.

However, there is one recurrent missense mutation observed in three individuals and all of these individuals have seizures: c.3460C>T, p.Arg1154Trp. This is seen in two

individuals in this cohort WSSP33 and WSSP84 and in an individual reported by the Euro EPINOMICS consortium, a cohort study of individuals with classic encephalopathies including infantile spasms and Lennox Gastaut syndrome. All three individuals have a history of seizures. For WSSP33 the seizures started before the age of nine years, however no further information is available. WSSP84 had two generalised seizures in the first year of life associated with fever which were associated with a pathological EEG and developmental regression. WSSP33 and in an individual reported by the Euro EPINOMICS consortium, a cohort study of individuals with classic encephalopathies including infantile spasms and Lennox Gastaut syndrome(84). Further data is required to ascertain whether this is a chance association and whether the c.3460C>T, p.Arg1154Trp variant results in an increased frequency of seizures.

2.4.16 Father with mosaic *KMT2A* variant has a milder phenotype

The gentleman with the mosaic KMT2A variant (WSSP37), is of normal intelligence and has short stature. He was reported to have significant hypertrichosis with long hypertrichosis of his back as a child. Given developmental delay and learning difficulties are seen in all individuals with WSS and apparently, germline KMT2A mutations, it is likely because he carries the mutation in the mosaic form that he does not have more significant learning difficulties as a result of this.

2.4.17 The largest cluster of missense mutations lies in the CXXC zinc finger domain

The largest missense mutations cluster is located in the CXXC zinc finger domain (Figure 2-14). Zinc finger domains such as this, contain two zinc ions and selectively bind nonmethyl-CpG-DNA. Each zinc ion is tetrahedrally co-ordinated by four conserved cysteine residues. Some of the disease-associated *KMT2A* missense variants appear to disrupt the binding of one of the two zinc ions by affecting one of the eight co-ordinating cysteines. This would be expected to disrupt the domain's fold affecting DNA binding and recognition (Figure 2-15). Mutations resulting in a protein unfolding and being a target for proteolysis is seen in other disorders, such as ATRX syndrome caused by mutations in the *ATRX* gene(128). The effect of the Arg1154Trp mutation is less clear, however mutations of Arg1154 have been shown to abolish or significantly decrease

DNA binding but have no effect on global protein folding in model organisms(129). It is proposed that the detrimental effect on DNA binding may solely result from the removal of a functionally important side chain(129).

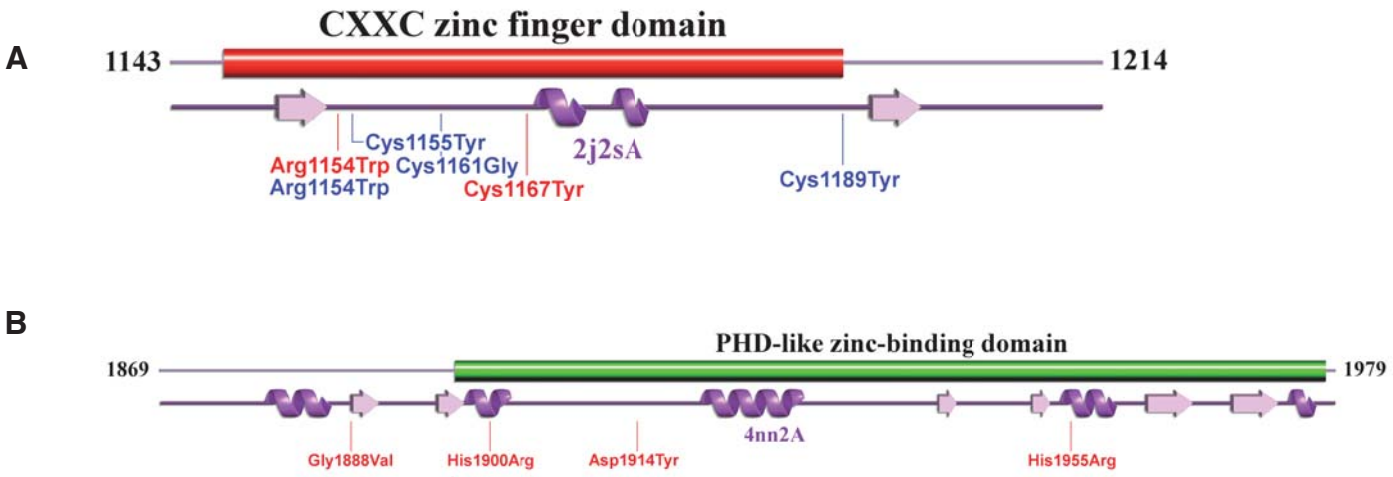


Figure 2-14: Distribution of *KMT2A* missense mutations

A: A zoomed in view of the region 1,143-1,214 corresponding to the 2j2s structure of the CXXC zing finger domain, and the most concentrated cluster of disease-associated KMT2A variants. The secondary structure of the protein is shown as arrows for beta strands and corkscrews for alpha-helices. Missense variants from this cohort are denoted in red. *KMT2A* missense variants reported in the literature are denoted in blue. The Arg1154Trp mutation is present three times, twice in this cohort and previously reported by the EuroEPINOMICS-RES Consortium (130). B.Residues 1,869-1.979. The KMT2A variants were plotted using the Canonical Uniprot protein sequence, this differs by three residues to the nomenclature of the KMT2A protein variants provided in Appendix 2.

The second largest cluster of disease-associated missense mutations is in and around the PHD-like zinc binding domain (Fig 2-16B). Although, there isn't a 3D structure of this domain in the PDB, there is one for a related protein, human Borjeson-Forssman-Lehmann associated protein, (PHF6) a transcriptional regulator that associates with ribosomal RNA promoters(131). Two of the de novo missense mutations, His1900Arg and His1955Arg directly affect zinc-binding residues and therefore are likely to disrupt the domain's fold and disable its DNA-binding ability.

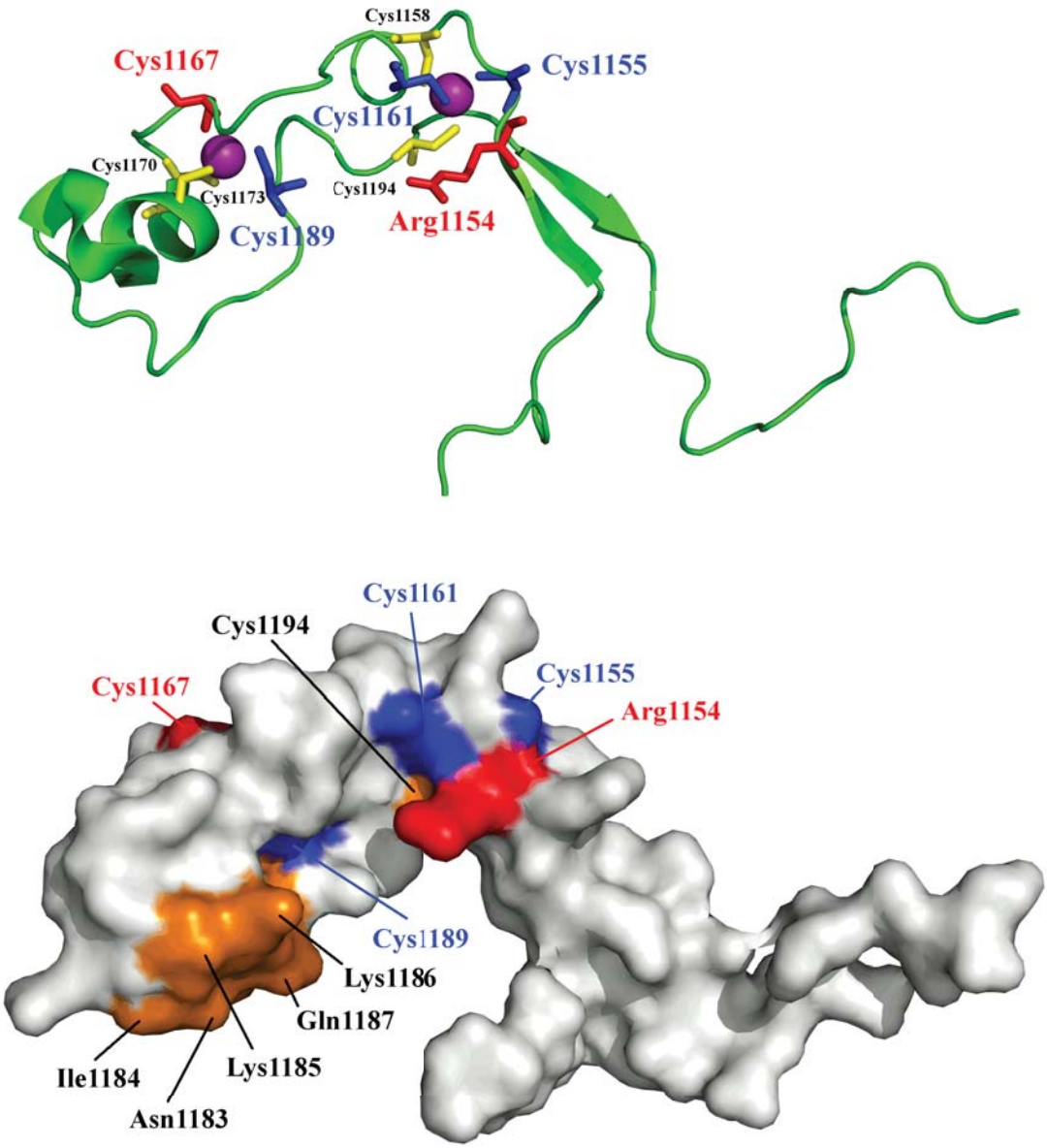


Figure 2-15: The 3D structure of the CXXC domain

A cartoon depiction of the CXXC domain, showing the two zinc ions (purple spheres), each coordinated by four cysteine residues (shown as yellow sticks, unless affected by a mutation when they are coloured red for Cys1167 from this cohort, and blue for Cys1155 and Cys1189 from the published literature). Also shown is Arg1154 (in red), mutations in which are identified as disease-causing in two individuals from this cohort and one individual reported by the EuroEPINOMICS-RES consortium. **b.** A representation of the domain's surface, with the residues thought to interact with DNA shown by the orange and red colouring. The red residue is Arg1154. Its mutation to the larger tryptophan is likely to interfere with DNA binding.

2.4.18 Clinicians can recognize patients with loss of function variants

After modelling the missense mutations, I then assessed whether clinicians can distinguish individuals with *KMT2A* mutations from control individuals with undiagnosed developmental disorders. Overall, the 6 clinicians scored the group of individuals with pathogenic loss of function *KMT2A* variants as more likely to have WSS than they scored the *KMT2A*-negative individuals selected at random from the DDD study (figure 2.16). The difference between the distributions of mean scores for the loss of function group versus the *KMT2A* negative group was significant ($p = 0.001664$, Two-sample Kolmogorov-Smirnov test). There was no significant difference in the distribution of scores between the trainee Clinical Geneticists (each with 1-5 years Dysmorphology experience) and the Consultant Clinical Geneticists (each with 5+ years Dysmorphology experience). Suggesting that years of experience are not necessary to be able to distinguish the facial features of individuals with WSS from those of individuals with developmental disorders.

2.4.19 There was a bimodal distribution for missense variants

The distribution of mean scores for the faces of individuals with missense variants was bimodal. When these scores were interrogated further it was obvious that the three individuals with the lowest scores were individuals in whom the diagnosis of WSS was being questioned (Figure 2-16). Two individuals had inherited missense variants from an affected parent. One individual had a de novo missense variant that was not felt to be causal as his phenotype did not fit with WSS.

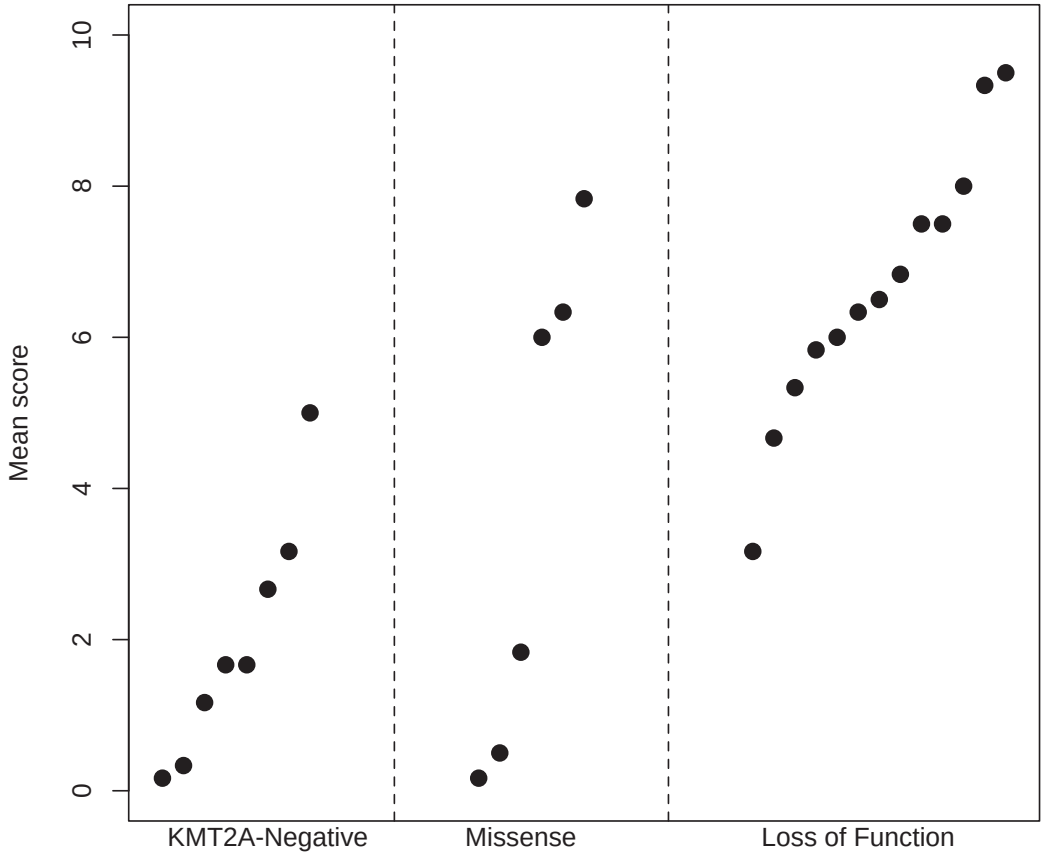


Figure 2-16: Scatterplot showing the mean score of all 6 clinicians for each 'face' by *KMT2A* variant status.

2.4.20 WSS has an estimated prevalence of 1 in 25,000 to 1 in 40,000

The SNV loss of function rate of KMT2A is 8.91935E-06(121). Adding this to the INDEL loss of function rate (gives a figure of 1.48E-05), this is a per transmission rate and therefore needs to be doubled to obtain a per child rate which is 2.96E-05. This equates to a prevalence of 1/34,000. Taking into account factors that may increase or decrease this rate (Table 2-3) I estimated that a prevalence of 1 in 25,000 to 1 in 40,000 would be appropriate for WSS.

Factors that might increase the birth prevalence of WSS from per child CNV and INDEL LoF rate	Factors that might decrease the birth prevalence of WSS from per child CNV and INDEL LoF rate
Include intragenic CNVs or INDELS too big to detect on exome sequencing are not included	Preferential spontaneous miscarriage
Other minor classes of LoF mutations, e.g. structural variants, intronic splicing mutations are not included	
Missense that would act as loss of function mutations are not included	

Table 2-3: Factors that might increase or decrease the birth prevalence of WSS from the per child CNV and INDEL LoF rate.

2.5 Discussion

2.5.1 Summary of findings

In summary, I have identified 84 individuals with *KMT2A* mutations from 82 unrelated families. Using data from these individuals I defined the *KMT2A* mutational spectrum and provided the first detailed evaluation of the features associated with *KMT2A*-associated WSS in a cohort more than 15 times larger than the largest previous report (N=5 individuals). I have successfully studied a monogenic disorder by collaborating with multiple large sequencing studies including the DDD study across 15 countries. I have demonstrated the key features of WSS, and broadened the phenotypic spectrum known to be associated with WSS as well as given insight into how missense mutations may cause WSS. In addition, I have confirmed the earlier findings of Jones *et al* that *de novo* mutations in *KMT2A* cause WSS(46, 72) as I report *KMT2A* mutations in all three of the original unscreened individuals reported under the classification of WSS by Koenig *et al*.

In this investigation, I have highlighted the key features of WSS caused by *KMT2A* mutations, namely developmental delay, hypertrichosis, behavioral difficulties, feeding problems, hypotonia and constipation. I have confirmed that sleep difficulties and frequent infections are also common in WSS individuals with *KMT2A* mutations. I have demonstrated the growth pattern of WSS, namely a normal birthweight followed to failure to thrive associated with short stature or height in the lower half of the normal range and a head circumference in the lower half of the normal range or microcephaly. I have

given the first evidence that adult women with WSS have a tendency towards obesity in adulthood. I have also demonstrated that not all individuals with WSS have hypertrichosis.

I have reported new features of WSS including female genital abnormities, abnormalities of pain sensation and Raynaulds phenomenon. I have heighted epilepsy as an important feature of WSS and that epilepsy is associated with poorer outcomes in WSS individuals. I have reported the first somatic mosaic individual with a *KMT2A* mutation and shown he has a milder than other individuals with germline *KMT2A* mutations and has normal intelligence. I have used standardised phenotypic terms (HPO terms) wherever possible throughout this phenotypic analysis to enable other researchers to readily access and understand my findings and perform meta-analyses in the future. I demonstrated through my facial recognition study that the facial appearance of individuals with WSS is distinguishable from that of other individuals with developmental disorders by experienced Clinical Geneticists and trainees in clinical genetics.

2.5.2 Limitations to this investigation

This study is limited to the populations of people living in the vicinity of Clinical Genetics services with the means to be reviewed by a Genetics doctor. As a result, some populations are under-represented in this study. In fact, 83/84 individuals are white. This study is therefore underpowered to investigate WSS in black and Asian or indigenous populations where there may be variability in facial and or other features. Other limitations to this study is that the negative data is not complete as observed from the phenotype Table 2-1. Clinicians are busy people and filling questionnaires is often performed quickly which is likely one contributing factor to the lack of negative data.
Finally, adults are under-represented in this cohort, in order to understand the difficulties faced by adults and know how to look after them from a medical point of view it is important to understand the adult phenotype of individuals with WSS. A further limitation is that in this study I didn't investigate the effect of the missense variants on splicing. Some of the missense variants in my investigation may be pathogenic due to an effect on splicing.

2.5.3 Interpretation of missense variants

In conditions where the predominant mutation mechanism is loss of function alleles, the interpretation of missense findings will remain challenging for many years to come. Without a functional or biochemical assay, current practice is to rely on the patient's clinical phenotype and the predicted effect of the mutation on the protein product. With time, functional work and recurrent mutations missense mutations will be further understood but in the meantime an element of caution needs to be exercised. Clinical phenotyping remains vital to interpret variants and investigations such as this are vital to help clinicians interpret variants found on next generational sequencing platforms. Facial recognition software may play an important role in this process in the future.

KMT2A is a large gene and it may be that phenotypes other than WSS are associated with missense variants in *KMT2A* as is observed with *CREBBP*. Mutations in CREBBP have long been established to be associated with the chromatin disorder Rubinstein Taybi syndrome which shares a number of similarities to WSS, but more recently other phenotypes have been associated with missense variants in *CREBBP(132)*. With the passage of time and with further individuals with developmental disorders undergoing exome sequencing, it will become apparent whether there are other phenotypes resulting from missense mutations in *KMT2A*.

2.5.4 Challenges for phenotypic investigations in the era of genomics

It's well recognised that with next generational sequencing approaches large amounts of sequencing data is generated. However, but given the increased rate of diagnosis with these platforms there will be large accompanying amounts of phenotype data for clinicians to manage in the study of novel disease genes or already recognised genes with broadening phenotypic spectra.

Phenotype study methods long employed by Clinical Geneticists largely consisting of transfer of information by secure email or letter will be put under strain by large sample sizes and the difficulties of managing more data points than ever before. The data management problem has been addressed by research studies such as the DDD and

100,000 genomes project which have incorporated online phenotyping, enabling clinicians to enter their own phenotype data coded by HPO terms into an online portal. This data is then standardised, less prone to error and much more manageable. More ideal yet would be for all clinical data in hospitals to all be recorded by standardised terms at the point of care (clinical review by a doctor or nurse). Then with the patient's consent if they were to join a research study this data set could simply be transferred to the research study. Maintaining genomic information is also important, and ensuring that mutations are recorded correctly in a changing world of genomic builds. This has been recognised by databases such as Decipher who store mutations in a manner which is easily updatable.

2.5.5 Future Directions

Unanswered medical questions

There a number of unanswered questions about WSS that need to be addressed in the future for the accurate management of the affected individuals. This includes (this list is not exhaustive): Further study of infections and immune function; behavioural analysis to help families manage challenging behaviour, assessment of specific learning difficulties to guide teachers with support in the classroom; investigations to difficulties experienced in adulthood with screening and medical management employed as appropriate; investigation of the aetiology of hand and foot swelling in early life. As with other complex developmental disorders, a multidisciplinary approach is most appropriate for the care of these individuals who have multiple medical needs.

To accurately counsel families about recurrence risks further study is needed to determine the true incidence of germline mosaicism. In addition, study of individuals with seemingly germline mutations for evidence of mosaicism may give insight to the individuals at the milder end of the phenotypic spectrum.

How do mutations in KMT2A cause WSS

Studying the effect of missense variants on KMT2A protein levels would help further understand which variants are disease causing and help understand how variants cause

disease. There are commercially available antibodies to KMT2A(133) and these could be used with Western blotting techniques to analyze KMT2A levels. Functional experiments in cells with *KMT2A* mutations in early development may play an important role to determine how mutant *KMT2A* alleles cause disease and develop targets for treatment. But also, to understand how the same mutations in *KMT2A* can be seen in solid tumors and in the germline with very different effects. Greater knowledge of KMT2A in disease will increase understanding of epigenetic regulation in general, and potentially advance knowledge of other chromatin disorders whose aetiology may have shared mechanisms.

Increased knowledge about the sites of KMT2A binding within the genome would also assist understanding of how mutations in *KMT2A* cause WSS. Current knowledge suggests that KMT2A may bind to both gene promoters and gene enhancers(134, 135), however further investigation into the location of KMT2A binding within the genome is required.

Understanding the variability of WSS

Further investigations are needed to fully understand how the difficulties associated with WSS can range from mild to profound. I have shown that there is an association with epilepsy and severe developmental outcomes. However, there may in addition be mutations in other genes involved with the second mutation acting as a modifier or resulting in the individual having two disorders. Interrogation of the exome variant profiles of individuals with WSS stratified for the level of their learning difficulties may help elucidate whether mutations in other genes are involved. Exome sequencing in individuals with extreme phenotypes recently proved successful in identifying *DCTN4* as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis(136) and a similar approach might be possible for WSS and individuals with profound learning difficulties or developmental delay. Emond *et al* in their investigation, which successfully identified DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis, studied only 91 individuals(1). However, the effect size for *DCTN4* was estimated to be relatively large, the collective minor allele frequency of the implicated variants was reasonably high, and the individuals were phenotypically well matched with the exception of the trait of interest. Identifying modifier genes in WSS is likely to be more complex than in cystic fibrosis. Firstly, because the phenotype of WSS is more

complex, making phenotype matching more difficult. Individuals with WSS have multiple problems affecting many more multiple organ systems than are affected in cystic fibrosis including the brain. The number of individuals needed to identify modifiers would depend on the effect size and minor allele frequency of modifier variants, but it would also require accurate phenotype matching with the exception of the phenotype of interest. Therefore, although Emond *et al* showed a sample size of around 100 is possible to identify a modifier in cystic fibrosis(1), a sample size in excess of this is likely to be necessary to identify genetic modifiers in WSS.

Given the association with epilepsy and worse developmental outcomes, to understand this further, mice with heterozygous *KMT2A* mutations could be stimulated to have seizures to see whether this impacted negatively on their development.

2.5.6 Summary of discussion

In summary, I have through molecular and clinical analysis identified 84 individuals with *KMT2A* mutations from 82 unrelated families. Using data from these families I defined the *KMT2A* mutational spectrum and provided the first detailed evaluation of the features associated with *KMT2A*-associated WSS in a cohort more than 15 times larger than the largest previous report (N=5 individuals). I have highlighted areas for further investigation in the future, including further immune and behavioural phenotyping, investigations into the variability of WSS and functional experiments to determine how mutations in *KMT2A* cause WSS. A multidisciplinary approach to the medical care of these individuals is vital for the care of these individuals who have multiple medical needs.

Chapter 3

Gene discovery in hypertrichosis

3.1 Aims

- **To investigate the genetic basis of developmental disorders associated with hypertrichosis using whole exome sequencing**
- **To identify new genes implicated in developmental disorders associated with hypertrichosis**
- **To seek a burden of variants in genes that play a role in maintaining chromatin structure or function in individuals with developmental disorders associated with hypertrichosis**

3.2 Introduction

3.2.1 Hypertrichosis and motivation for this investigation

Hypertrichosis is the growth of terminal hair in excess of what is expected given the individual's age, sex and ethnicity. Hypertrichosis has been reported in isolation and in association with developmental disorders including those associated with genes whose protein products affect the structure and function of chromatin (chromatin disorders) and inborn errors of metabolism. It has also been used as a key phenotypic feature to aid gene discovery(46, 137, 138), in particular in the monogenic condition Wiedemann-Steiner syndrome (WSS) which was introduced in Chapter 2. However, to my knowledge no one has used whole exome sequencing more broadly to investigate the genetic basis of hypertrichosis associated with developmental disorders.

3.2.2 Hypertrichosis and its causes

There are a number of different causes for hypertrichosis. It may be congenital or acquired, localised or generalised and associated with metabolic disorders. Several developmental disorders are reported in association with hypertrichosis and it has been used as a key phenotypic feature to drive gene discovery in some disorders.

There are a number of reported disorders of congenital widespread hypertrichosis(139, 140). In these disorders, hypertrichosis often involves the face, and sometimes spares

only the palms and soles. Unlike the multiple-congenital-anomaly disorders associated with excess hair (discussed below) these disorders tend not to be associated with learning difficulties or multiple-congenital anomalies. Hypertrichosis can be a localized finding. For example, localized spinal hypertrichosis has been reported in association with an underlying defect in the vertebrae, spinal cord or nerve roots (spinal dysgraphism) including myelomeningocele, dermal cyst or sinus or a subdural lipoma.

Hypertrichosis is seen in association with disorders resulting from inborn errors of metabolism. These include the mucopolysaccharidoses (disorders resulting from deficiency of or abnormal structure of lysosomal enzymes), and disorders associated with lipodystrophy, such as Berardinelli Seip congenital lipodystrophy and Donohue syndrome caused by mutations in the insulin receptor gene.

There are a number of multiple-congenital-anomaly syndromes associated with hypertrichosis(45, 46, 137, 138, 141). These disorders may have a distinctive facial appearance and many are associated with developmental delay. This group tend to have less dense and less extensive Congenital generalised hypertrichosis terminalis than the conditions listed above(45, 46, 86, 141). This group include a number of disorders associated with genes encoding proteins that interact with and modify the structure of chromatin(45, 46, 142) including WSS resulting from pathogenic variants in the histone methyl-transferase *KMT2A*(46).

Knowledge and investigation of protein binding partners has driven gene discovery in developmental disorders including those associated with chromatin modification(44, 86, 143). To our knowledge no one has investigated the hypothesis that mutations in the protein binding partners of *KMT2A* result in a similar phenotype to WSS.

KMT2A encodes the histone methyltransferase enzyme KMT2A, which is expressed in most cell types(82, 95). The KMT2A protein is a large (3,969 aa) multi-domain protein(96) which is one of a family of histone–lysine N-methyltransferase 2 (KMT2) proteins. The KMT2 proteins, including KMT2A, are highly conserved(97) and they play an important role in epigenetic regulation. KMT2A generates mono-, di-, and

trimethylated histone H3K4, through its SET domain and interaction with cofactors (reviewed by Rao et al(98)). The chromatin activity of the KMT2 enzymes is modified by subunits of large multimeric complexes in which they function (144-148). Each KMT2 enzyme each has a unique set of interacting proteins, however there are some proteins that are common to all of the protein binding complexes, these are WD repeat protein 5 (WDR5), retinoblastoma-binding protein 5 (RBBP5), ASH2L and DPY30(149). In addition, the multimeric complexes of KMT2A and KMT2B, also include the proteins menin, HCF1 and HCF2(reviewed by Rao et al(98)). There is evidence that KMT2A also interacts with proteins implicated in other developmental disorders with overlap with Wiedemann-Steiner syndrome such as the histone-acetyltransferase CREBBP(150). Heterozygous mutations in *CREBBP* underlie Rubinstein Taybi syndrome, a congenital multiple anomaly syndrome which is also associated with hypertrichosis.

3.2.3 Summary of Introduction to this investigation

Hypertrichosis is the excessive growth of hair in excess of what is expected given the individual's age, sex and ethnicity. Hypertrichosis has a number of underlying causes, it may be congenital or acquired, localised or generalised and associated with metabolic disorders. Several developmental disorders are reported in association with hypertrichosis and it has been used as a key phenotypic feature to drive gene discovery in some disorders. However, to our knowledge no studies have carried out whole exome sequencing in individuals with developmental disorders including hypertrichosis to investigate the genes implicated in increased body hair more generally.

Knowledge and investigation of protein binding partners has driven gene discovery in developmental disorders including those associated with chromatin modification. To our knowledge no one has investigated the hypothesis that mutations in the protein binding partners of *KMT2A* result in a similar phenotype to WSS.

3.3 Methods:

3.3.1 Individuals with WSS and or increased body hair were identified

I assembled a cohort of 247 individuals with phenotypic features consistent with WSS (as assessed by a Clinical Geneticist) or similar to WSS and / or with evidence of increased body hair (I referred to this phenotype as WISH: Wiedemann-Steiner syndrome or related phenotypes or hypertrichosis). Individuals were recruited as singletons or with one or both parents (duos or trios). Recruitment criteria, including the Human Phenotype Ontology(58) terms used to select patients with increased body hair are listed in table 3-1. 228 of the individuals underwent whole exome sequencing as part of the Deciphering Developmental Disorders study. The 19 remaining individuals were recruited from outside the UK and underwent whole exome sequencing separately at the Wellcome Trust Sanger Institute (WTSI) as part of the Wiedemann-Steiner and hypertrichosis whole exome sequencing (WiSH-WES) Study.

1. Individuals coded as having any of the following Human Phenotype Ontology (HPO) terms or with an affected parent with any of the following HPO terms
HP:0000998 Hypertrichosis HP:0002219 Facial hypertrichosis HP0004532 Sacral hypertrichosis HP:0004535 Anterior cervical hypertrichosis HP:0004540 Congenital generalised hypertrichosis HP:0004554 Generalised hypertrichosis HP:0004780 Elbow hypertrichosis HP:0011913 Lumbar hypertrichosis HP:0011914 Thoracic hypertrichosis HP: 0001007 Hirsutism HP: 0002230 Generalised hirsutism HP: 0009747 Lumbosacral hirsutism HP: 0009889 Localised hirsutism HP:0009937 Facial hirsutism HP: 0011335 Frontal hirsutism HP:0000664 Synophrys
2. And / or coded as any of the following in the gene test, additional comments or known syndrome section:
Wiedemann-Steiner Steiner WSS Hypertrichosis cubiti Hypertrichosis Hirsutism Hairy Wiedermann Stiener KMT2A MLL
3. And / or previously tested for mutations in the following genes: *KMT2A (MLL)*
4. And / or flagged by the local clinician as having a phenotype consistent with Wiedemann-Steiner syndrome

Table 3-1: Recruitment criteria for this study
These criteria were used select individuals to the current study. These included common misspellings of Wiedemann and Steiner.

3.3.2 Individuals underwent whole exome sequencing

For DDD study sequencing methods please see Chapter 4: The Deciphering Developmental Disorders Study / Investigations into Autosomal Recessive Developmental Disorders. For the 19 trios who underwent sequencing separately as part of the WiSH-WES study DNA samples were sent to the Wellcome Trust Sanger Institute, DNA Samples were sent to the Wellcome Trust Sanger Institute (WTSI). Whole exome sequencing of family trios was carried out using a custom Agilent exome capture kit: SureSelectXT Human All Exon V5, followed by paired end sequencing (75bp reads) on an Illumina HiSeq platform. Reads were mapped to the reference human genome GRCh37 (hs37d5) using BWA(151).

3.3.3 WiSH-WES Study: Variants underwent annotation, QC and filtering

Variants were called using the haplotype caller from GATK(152) version 3.2-2. Variants were annotated with Ensembl Variant Effect Predictor(153)v2.2 (VEP). I annotated the variants with frequencies from the 1000 genomes project (all populations), esp6500(154), Exac02(155), dbsnp138(156, 157), clinvar 20140929(158) using ANNOVAR(159) and vcftools(160). I annotated the variants with VQSR and VQSLOD from GATK(152, 161) using bcftools from the SAMtools set of utilities(162). Rare variants were defined as 'frequency less than 1% in ExAC and 1000 genomes and common variants as 'frequency greater than or equal to 1% in ExAC and / or 1000 genomes'. I wrote custom python scripts to generate quality control metrics for the exome variants. Where there were multiple ALTs (alternate alleles) I used only the first allele stated. I used only the first ALT variant frequency for the quality control analysis. I filtered variants using the VQSR PASS filter and removed those not passing filters. I calculated *KMT2A* coverage using BEDtools(163).

3.3.4 I analysed variant call format files using custom scripts

For the 19 individuals in the WiSH-WES study, I wrote custom scripts in python and analysed Variant call format (VCF) files to identify rare (minor allele frequency >=0.01)

functional and loss of function variants using custom scripts. I defined loss of function variants as: Disruptive, Stop gained, transcript ablation, splice donor variant, splice acceptor variant and frameshift variant. I defined functional variants as: Missense, inframe deletion, inframe insertion, coding sequence variant and stop lost.

For individuals in the DDD study I used clinical-filter (https://github.com/jeremymcrae/clinical-filter) to identify rare functional and loss of function variants. I annotated variants with sufficient evidence as being developmental disorder genes using the Deciphering Developmental Disorders Genotype to Phenotype database (DDG2P)(2). The DDG2P is discussed further in Chapter 4: The Deciphering Developmental Disorders Study / Investigations into Autosomal Recessive Developmental Disorders.

To assign pathogenicity to each variant, I reviewed each rare functional or loss of function variant in a DDG2P gene alongside the phenotype of the affected individual including photographs (where available). I took into account presence of variants in population databases, and PolyPhen scores for Missense variants, and previous reporting of that specific variant to determine the likely pathogenicity of the variant. If present, I took into account the local clinicians pathogenicity contribution score, which are assigned on the Decipher database(118) upon receiving results. However, these data are incomplete, as not all clinicians will have reviewed the individuals again following reporting of their variants yet. The possible pathogenicity contribution scores in the Decipher database are shown in table 3-2. I assigned the contribution score of the clinician to each variant where present. When there was no contribution score, I assigned each variant as pathogenic, possibly pathogenic, or not contributing to the individual's phenotype.

Pathogenicity	
Class 5	Definitely pathogenic: would offer predictive testing based on this finding, if appropriate
Class 4b	Probably pathogenic: likely to be causal but evidence not conclusive, would curtail other diagnostic investigations and would seek additional confirmatory evidence before offering predictive testing.
Class 4a	Possibly pathogenic: Reasonably likely to be causal but uncertainty would preclude offer of predictive testing
Class 3	Uncertain: Insufficient evidence to decide whether this is a causal or benign variant
Class 2	Likely benign Likely not to be causal or of little clinical significance
Class 1	Benign: Strong evidence that the variant is not pathogenic
Contribution	
Full	Variant fully explains the patient's whole phenotype
Partial	Variant either partially explains patient's whole phenotype or fully explains part of the patient's whole phenotype
Uncertain	Contribution to patient's phenotype is unknown
None	Variant has no discernible contribution to patient's phenotype

Table 3-2 Possible pathogenicity and contribution scores on Decipher
Each variant is scored by their local genetics clinician with a pathogenicity and contribution score.

3.3.4 Gene discovery for new genes implicated in hypertrichosis

In order to identify candidate genes for hypertrichosis associated disorders I analysed the exome variant profiles of individuals with no DDG2P gene variants or where the variants identified were not felt to contribute to the individual's phenotype. I analysed the exome profiles of all undiagnosed individuals in a sequential manner hypothesizing in turn that that the undiagnosed developmental disorders could result from a *de novo* mutation, biallelic variants, or an X-linked variant. To assign pathogenicity to a novel gene I used the DDG2P criteria (see Chapter 4: The Deciphering Developmental Disorders Study / Investigations into Autosomal Recessive Developmental Disorders). In combination with assessing for statistical significance in analyses comparing incidence of *de novo* mutations compared to expected rates, see below.

3.3.5 Modelling mutation rates in analysis of de novo mutations

I analysed the significance of de novo variants in candidate genes using the underlying mutation rate, using the method and data from Samocha *et al(164)* adapted by Singh *et al(165).* Briefly, the tri-nucleotide mutation rates for each gencode canonical transcript were adapted to generate a mutation rate for every class of variant. Additionally, data from PolyPhen(166) was incorporated to provide separate mutation rates for missense variants predicted to be probably damaging.

3.3.5 Variant Interpretation used a number of programs and websites

The following programs and websites were used to interpret the possible pathogenicity of variants:

Pubmed: (http://www.ncbi.nlm.nih.gov/pubmed)

OMIM: (http://www.ncbi.nlm.nih.gov/pubmed)

DECIPHER: (https://decipher.sanger.ac.uk/)

ClinVar: (http://www.ncbi.nlm.nih.gov/clinvar/)

Uniprot: (http://www.uniprot.org/)

Ensembl: (http://www.ensembl.org/index.html)

Database of genomic variants: (http://dgv.tcag.ca/dgv/app/home)

3.3.6 I identified genes encoding proteins that complex with KMT2A

There are a number of genes encoding proteins, which complex with KMT2A. that there is evidence that which bind or interact with KMT2A. These genes were identified from the literature as encoding proteins which are core complexing proteins of KMT2A (evidence reviewed by Rao *et al(98)*). These genes are as follows: *ASH2L, RBBP5, WDR5, DPY30, MEN1, HCFC1 (HCF1), HCFC2 (HCF2) and PSIP1.* Of these genes that encode proteins that complex with *KMT2A,* only two are currently implicated in disease. 5 prime or 3 prime UTR mutations in *HCFC1* are associated with X-linked non-syndromic intellectual disability and loss of function mutations are associated with a colobamin disorder(167, 168). Mutations in *MEN1* are associated with Multiple Endocrine neoplasia type 1 (MEN1), an autosomal dominant disorder characterised by increased susceptibility to endocrine tumours. Hypertrichosis is not a recognised feature of MEN1.

3.3.7 I selected 870 genes which have a function related to chromatin

In order to carry out a burden analysis to look for an excess in variants in genes whose product has a function relating to chromatin: 'chromatin genes' in individuals with hypertrichosis, I first defined a list of chromatin genes. I performed a general search for the term 'chromatin' in the Gene Ontology (GO) databases(169, 170) (http://geneontology.org/). I filtered the search to include genes and gene products, and selected only genes associated with the taxon *Homo sapiens*. Through this method, I identified 1422 chromatin gene entries, upon removing duplicates entries this gave a final list of 870 chromatin genes. Of these chromatin-related genes 142 genes (16%) are present in the DDG2P database of genes reported to be associated with developmental disorders. I recognised that this list of 'chromatin genes' would miss some genes with a function related to chromatin, and that it may include some genes erroneously. However, I concluded this approach would enable me to generate a list of chromatin related genes in a timely manner.

3.3.8 Estimating the burden of mutations in chromatin genes

I investigated as to whether there was a burden (an increased number above null expectation) of mutations in chromatin genes in the DDD hypertrichosis cohort. I did this by simulating the number of mutations in chromatin genes expected by chance given their mutation rates. This involved assigning mutations at random to genes according to their mutation rate for each individual in the DDD study (2407 males and 1886 females) from which the hypertrichosis cohort was selected.

3.4 Results:

3.4.1 Number of individuals recruited per recruitment criteria from DDD study

I identified 228 individuals who fulfilled the criteria for this analysis. These individuals included 19 trios, with the remainder of individuals as duos or singletons. There were five families with two affected siblings in the study and one family with four affected siblings in the study. A breakdown of the method of cohort entry is shown in table 3-1 for individuals recruited through the DDD study. The selection criteria resulted in 228 individuals being selected from the DDD study. Although 22 patients were highlighted by clinicians separately as having phenotypes consistent with WSS added only 4 individuals to the cohort after selection had been carried out based on the other entry criteria, as the phenotype data entered by these clinicians fulfilled the study entry criteria by itself. Adding probands, with affected parents whose parents were coded with the relevant HPO terms didn't result in the selection of any further individuals into the study (see table 3-3 and 3-4).

Method of Cohort Entry	Number of individuals
Clinician highlighted patient	**22**
Matching in gene test section	1
Matching in additional comments	14
Matching in syndrome box	17
HPO matches	215
Maternal HPO matches	3
Paternal HPO matches	5
Decipher terms (HPO matches) or free text	**247 (224 unique)**
Total unique individuals	**228**

Table 3-3 Number of individuals recruited per each recruitment criteria This includes clinicians highlighting patients and matches on Decipher terms or free text. Note each of these recruitment criteria show overlap, for example clinicians may highlight patients who are coded with HPO terms relating to hypertrichosis. In light of this only unique individuals are included in the total count of 228.

Human Phenotype Ontology (HPO) term	Number of times HPO term used
HP:0000998 Hypertrichosis	27
HP:0002219 Facial hypertrichosis	0
HP0004532 Sacral hypertrichosis	4
HP:0004535 Anterior cervical hypertrichosis	2
HP:0004540 Congenital generalised hypertrichosis	0
HP:0004554 Generalised hypertrichosis	10
HP:0004780 Elbow hypertrichosis	8
HP:0011913 Lumbar hypertrichosis	1
HP:0011914 Thoracic hypertrichosis	2
HP: 0001007 Hirsutism	25
HP: 0002230 Generalised hirsutism	26
HP: 0009747 Lumbosacral hirsutism	9
HP: 0009889 Localised hirsutism	13
HP:0009937 Facial hirsutism	3
HP: 0011335 Frontal hirsutism	6
HP: Synophrys	102

Table 3-4 Number of individuals recruited per Human Phenotype Ontology (HPO) term

3.4.2 Data from the WiSH-WES samples were good quality

I carried out quality control of the WISH-WES data by looking at the number and ratios of SNV and INDEL variants (Figure 3-1). The data were of good quality. SNVs were as expected with no significant outlying counts. (Figure 3-1).

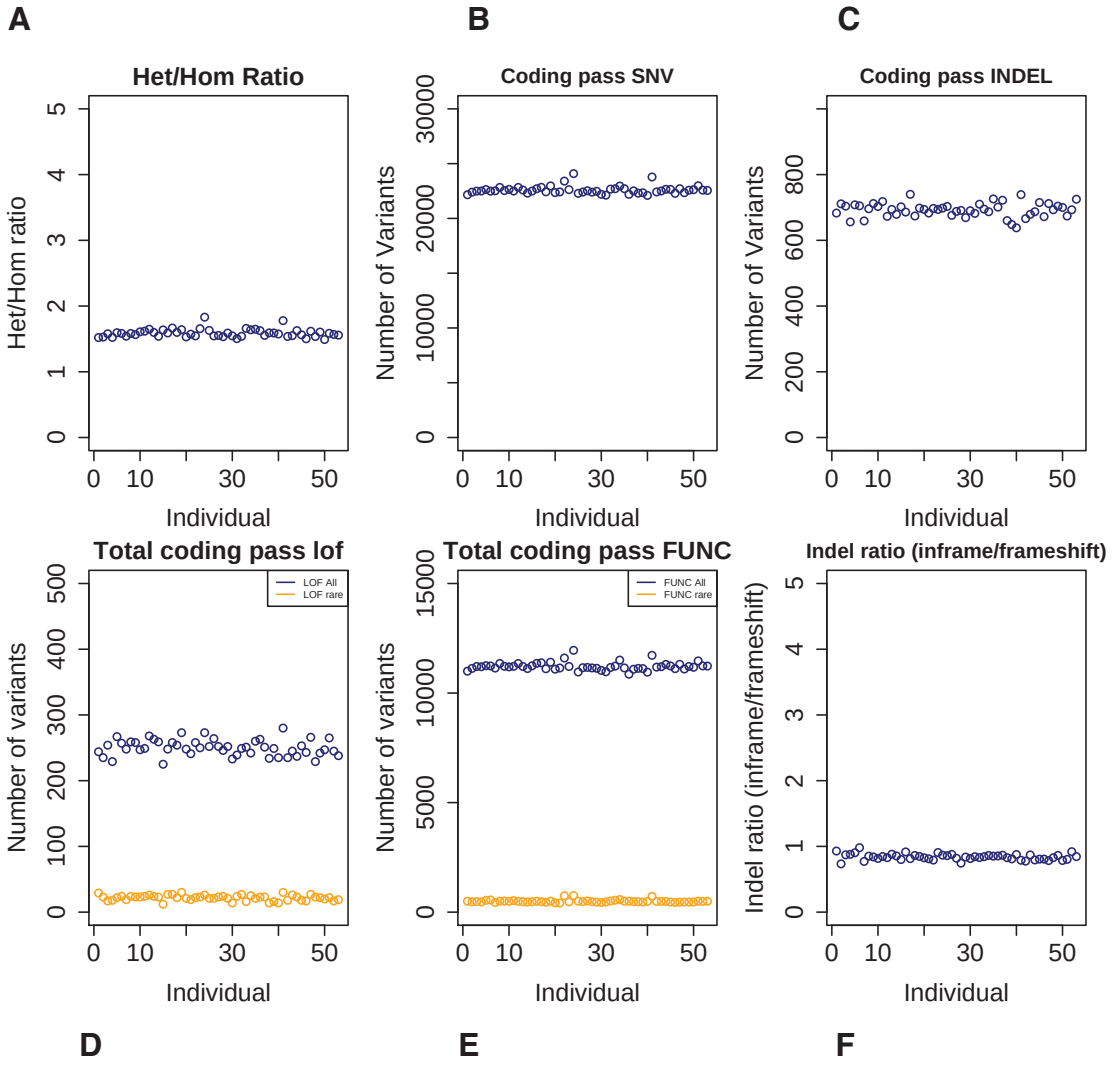


Figure 3-1: Quality control metrics for single nucleotide variants and INDELS in the WISH-WES cohort
(A) Ratio of heterozygous variants to homozygous variants. (B) Number of SNVs that lie within the coding region and have passed filters. (C) Number of INDELS that lie within the coding region and have passed filters. (D) Number of loss of function variants per sample, total variant number is shown in blue and rare variants (<1%) are shown in yellow. (E) Number of functional variants per sample, total variant number is shown in blue and rare variants (<1%) are shown in yellow.

3.4.4 *KMT2A* coverage in WiSH samples was good in WISH-WES samples

I showed also that coverage of KMT2A was good with 95.5% of bases having 10 or more reads and 97.75% of bases having 20 or more reads (Figure 3-2 and Figure 3-3).

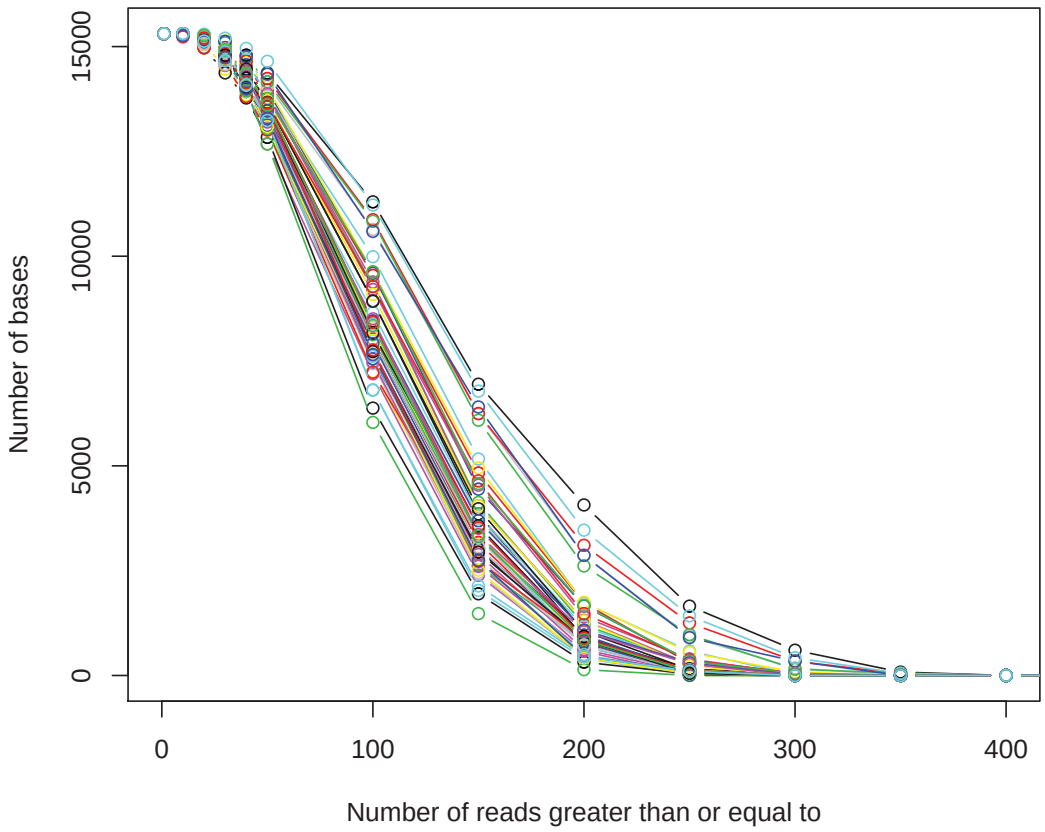


Figure 3-2: Coverage of KMT2A by exome sequencing reads by sample in the WISH-WES cohort
Number of bases covered by number of reads for parents and probands in the WISH-WES hypertrichosis cohort.

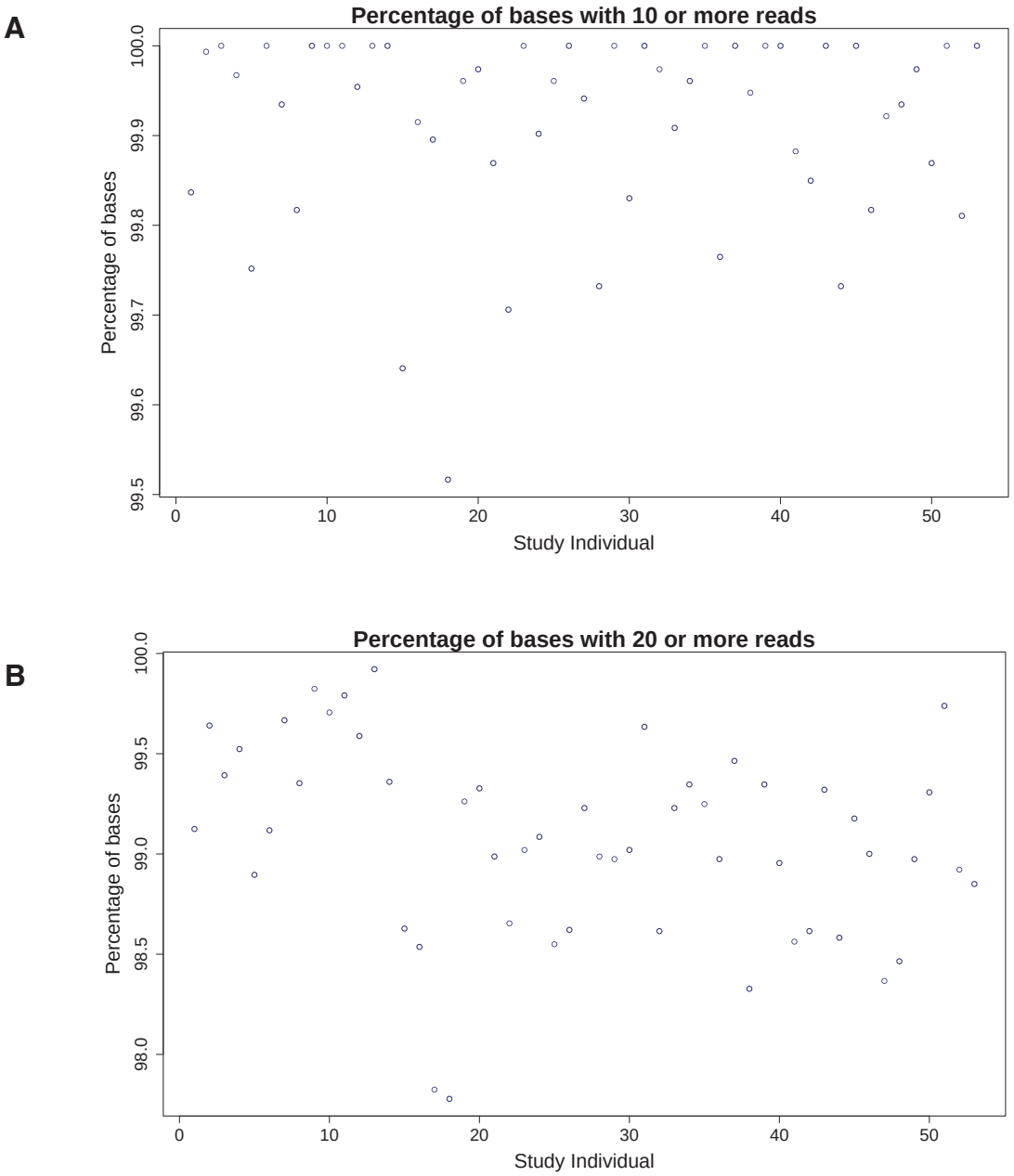


Figure 3-3: Coverage of KMT2A per study individual in the WISH-WES cohort
(A) Percentage of bases with 10 or more reads by study individual. (B) Percentage of bases with 20 or more reads by study individual.

3.4.3 28 WISH individuals had *KMT2A* mutations

I identified 27 individuals with rare functional or loss of function variants in *KMT2A* by analysing their filtered exome variant profiles. 20 of these variants were *de novo.* One individual from the WISH-WES cohort had inherited the *KMT2A* variant from her mosaic father (this is discussed in Chapter 2), for the other 6 individuals the inheritance of the KMT2A variant was not known as sequence information from their parents was not yet available. For details of the mutations, please see Appendix 2. One further individual (270606) was identified to have a *de novo* 2.11kb deletion within *KMT2A* by Convex analysis of his exome sequencing data. This intra-genic deletion includes four *KMT2A* exons.

I reviewed the phenotypic information and photographs (where available) for all 28 individuals with KMT2A variants, and concluded the phenotype was consistent with WSS in all 28 individuals. 12 of these individuals with *KMT2A* mutations or deletions were from the WISH-WES cohort (yield=12/19) and 16 individuals were from the DDD hypertrichosis cohort (yield=16/228). I then carried individuals whose mutation had been reported to their local clinician (22 individuals) forward for detailed phenotypic analysis. Details of this analysis is given in Chapter 2: Wiedemann-Steiner syndrome resulting from mutations in KMT2A: A Genotype-phenotype study.

3.4.5 Diagnostic yield of a pathogenic *de novo* mutation is 38%

I next investigated the DDD study cohort (220 individuals) for variants in genes other than *KMT2A* that could have caused their phenotype. I discovered that 47 individuals had rare functional or loss of function *de novo* mutations in autosomal or X-linked dominant and hemizygous (in males) DDG2P genes assessed as pathogenic by myself or their local clinician (table 3-5). For details of the mutations please see Appendix 2. Adding this figure to the 16 individuals with *de novo KMT2A* mutations gives 63 *de novo* mutations. Therefore the diagnostic yield of *de novo* mutations in autosomal dominant and X-linked dominant and hemizygous genes in males is 29%. The yield of pathogenic protein-truncating or missense *de novo* mutations in the same classes of DDG2P genes in 4293 individuals (each part of a trio) in the DDD study which was 23%. However, some individuals in this cohort had been identified by clinical geneticists as having WSS, a multiple congenital-anomaly syndrome associated with a distinctive facial appearance. I showed in Chapter 2 that WSS is a clinically recognisable disorder and that clinicians can reliably recognize this disorder based on facial appearance. Therefore, these hand-picked individuals may have falsely elevated the diagnostic yield as exome sequencing with good coverage of *KMT2A* in these individuals is effectively a diagnostic test. Removing the individuals with *KMT2A* mutations gives a diagnostic rate of 21%.

The *de novo* mutations were frequently found in chromatin genes: *ACTB, ARID1B* (7 individuals), *BCL11A, CREBBP, CTCF, DNMT3A* (2 individuals), *EP300, EYA1, MBD5*,), *SMAD4* (4 individuals), *SMARCA2* (2 individuals), *SMARCB1, HDAC8* (5 individuals), *PHF8 and SMC1A*. In total 30/47 (64%) of the pathogenic *de novo* mutations identified were in chromatin genes. This percentage is greater than the proportion of chromatin related genes which are reported to be associated with developmental disorders (16%), suggesting this cohort is enriched for mutations in genes which encode proteins with a role in maintaining chromatin structure and function. The chance that 64% of the cohort of 220 individuals carry a mutation in a random selection of X DDG2P genes (of the 142 DDG2P genes with a known role in maintaining chromatin structure or function only.

There are 666 dominant DDG2P genes, 116 of these are included in my list of genes which play a role in maintaining the structure and function of chromatin.

However, this investigation highlighted that two genes with a known role in chromatin function (*MECP2(171)* and *PHF6(172)*) were not identified as chromatin-related genes using the chromatin-related gene list, highlighting that the chromatin gene list used in this investigation doesn't include all known chromatin related genes.

Form of inheritance and mutation type	Genes carrying mutations (* = role in chromatin structure or function)
De novo mutations in dominant disease genes	*ABCC9 (2)* *ACTB** *ADNP (2)* *ARID1B* (7)* *BCL11A** *CBL* *COL4A3BP* *CREBBP** *CTCF** *DNMT3A* (2)* *DYRK1A* *EP300 *(2)* *EYA1** *HNRNPU* *MBD5** *MED13L(3)* *SCN2A* *SMAD*4* *SMARCA2 *(2)* *SMARCB1** *SYNGAP1* *TUBA1A*
Biallelic mutations in biallelic disease genes	*HACE1* *TMCO1**
De novo or inherited X-linked mutations (in X-linked dominant and hemizygous genes)	*DCX* *DDX3X (3)* *HDAC8* (5)* *IQSEC2* *MECP2* *PHF6* *PHF8** *SMC1A**
Inherited mutations in dominant disease genes	*ANKRD11* *ARID1B** *GRIN2A* *RAD21**
Mutations where the inheritance is not known in dominant disease genes	*ANKRD11 (2)* *ARID1B* (2)* *ASXL3 (2)* *EP300** *HNRNPU* *NIPBL** *SETD5** *WAC*(2)*

Table 3-5 Genes carrying mutations identified as pathogenic by mutation and inheritance type. * Denotes genes flagged as having a chromatin related function.

3.4.4 Four Individuals had heterozygous variants Inherited from an affected parent

I identified four inherited variants in individuals with affected parents which I assigned as pathogenic (table 3-5). These were in the genes *RAD21*, *ARID1B*, *GRIN2A* and *ANKRD11*. *RAD21* and *ARID1B* are both chromatin genes, (see table 3-5 and Appendix 2).

3.4.5 12 individuals had pathogenic heterozygous variants in dominant disease genes where inheritance was not known

I identified 12 individuals with mutations in dominant DDG2P genes that I assigned as pathogenic (table 3-5) and see Appendix 2 for details of the mutations. These genes included many of the genes in which de novo mutations had been identified. In addition, there were two individuals with variants in *WAC*, two individuals with variants in *ASXL3* and one individual with a variant in *HNRNPU*. 6/12 of these mutations were in chromatin genes.

3.4.6 14 Individuals had pathogenic mutations in X-linked DDG2P genes

I identified 14 individuals with mutations in DDG2P genes with an inheritance pattern of X-linked dominant or hemizygous genes in males DDG2P genes that I assigned to be pathogenic in causing their phenotype (table 3-5) see Appendix 2 for details of these mutations. Three of these genes are chromatin genes.

3.4.7 Two individuals had pathogenic biallelic variants in confirmed developmental disorder genes

I identified two individuals with rare loss of function biallelic variants in DDG2P genes that I assessed to be pathogenic (table 3-5). Individual 259339 has bilallelic frameshift mutations in *TMCO1*. *TMCO1* encodes a calcium selective channel, which plays a role in maintaining calcium homeostasis by preventing calcium stores from overfilling(173). Synophrys is a recognised feature of biallelic *TMCO1* mutations. The phenotypic features of individual 259339 include synophrys, and intellectual disability.

Individual 281381 has biallelic loss of function mutations (frameshift and nonsense) in HACE1. This is a recently discovered developmental disorders disease gene associated with intellectual disability and severe abnormalities of muscle tone including hypotonia, spasticity and dystonia(174). Four of the six individuals reported by Akawi *et al* (including individual 281381) had seizures(174). Therefore, the hypertrichosis could potentially be iatrogenic and as a result of seizure medication instead of a feature of the disease process itself.

3.4.8 Genes implicated in seizure disorders also feature in the list of genes associated with hypertrichosis

At least four of the DDG2P genes are associated with seizures: *DCX* and *DDX3X SCN2A and HACE1*. To my knowledge, none of these genes have been previously reported in association with hypertrichosis. Thus I propose that in these individuals their hypertrichosis could occur as a result of seizure medication instead of being a congenital phenomenon.

3.4.9 Variants in confirmed developmental disorder genes that are possibly pathogenic

I identified 14 heterozygous variants in dominant DDG2P genes, 6 biallelic variants in biallelic DDG2P genes and 3 variants in X-linked DDG2P genes that I assigned possibly pathogenic. A list of these genes is shown in Appendix 2. The challenges for assigning pathogenicity to many of these variants included. 1. There were no photographs for assessment of facial dysmorphic features. 2. The clinician had coded the variant as being of uncertain pathogenicity. 3. They were missense variants that had previously been unreported. 4. Inheritance information was not available.

3.4.10 Gene Discovery in the undiagnosed DDD individuals

I next investigated the undiagnosed individuals with hypertrichosis or features consistent with WSS in the DDD study with the aim of identifying new genes associated with developmental disorders. I selected all of the individuals who I hadn't identified as

having one or more pathogenic mutation(s) causing their phenotype (151 individuals). This did not include any individuals from the WISH-WES cohort.

I first investigated for the presence of *de novo* mutations in the same gene in two or more individuals. I identified no genes with loss of function variants in two or more individuals. I next investigated *de novo* missense variants and sought out genes containing variants PolyPhen scores of probably damaging in two or more individuals. I selected a PolyPhen score of probably damaging to increase the likelihood of the variant being damaging and therefore pathogenic. Two genes fulfilling these criteria were *ZMYND11* and *NR4A2.*

3.4.11 Missense variants in ZMYND11 are pathogenic

I identified two individuals with an identical *de novo* missense variant c.1798C>T p.Arg600Trp (ENST00000397962) in the X-linked gene *ZMYND11.* The variant is not present in the ExAC database (see table 3-6).

ID	SEX	CHR	POS	TRANS	CONSEQ	ALT/REF	GENO (P/M/F)	PPDNM	ExAC FREQ
262980	F	10	298399	ENST00000397962	Missense (ProbDAM/ DEL)	C/T	1/0/0	1	0
258442	M	10	298399	ENST00000397962	Missense (ProbDAM/ DEL)	C/T	1/0/0	1	0

Table 3-6 *De novo* missense variants in *NR4A2* identified in individuals with hypertrichosis. ID = Patient ID, CHR = chromosome, POS = Genomic position of the variant, TRANS = Transcript, CONSEQ = Predicted protein consequence, ALT/REF alternate base(s)/ reference base(s), GENO (P/M/F) = Genotype (Proband/Mother/Father), ppDNM = Posterior probability of the most likely DENOVO genotype configuration from DeNovoGear. (The range of ppDNM is 0-1, a value closer to 1 indicates higher probability of observing a denovo event at this position). ExAC FREQ = frequency of the variant in the ExAC database(120).

These individuals have overlapping phenotypes including developmental delay and synophrys. They both have short stature or stature in the lower range and weight below the normal range (See table 3-7). Looking more widely in the rest of the DDD (in the

4293 trios) there are three further individuals with de novo missense variants in *ZMYND11* (See Appendix 2 for details of these mutations).

ID	262980	258442
Height (SD)	-1.49	-3.29
Weight (SD)	-3.26	-2.58
OFC (SD)	-1.49	-1.39
HPO Terms	**HP:0001263 Global developmental delay** **HP:0000664 Synophrys** HP:0001999 Abnormal facial shape HP:0009916 Anisocoria HP:0000964 Eczema HP:0002020 Gastroesophageal reflux	HP:0009062 Infantile axial hypotonia **HP:0001263 Global developmental delay** **HP:0000664 Synophrys** HP:0000316 Hypertelorism HP:0000527 Long eyelashes HP:0006292 Abnormality of dental eruption HP:0006863 Severe expressive language delay HP:0003508 Proportionate short stature HP:0000377 Abnormality of the pinna HP:0007099 Arnold-Chiari type I malformation
Dysmorphic features	Photographs NA	Photographs NA
Other		Nuchal oedema 7mm detected on 20 week USS. No other abnormalities. Neck skin normal at birth. Premature loss of deciduous teeth

Table 3-7 The phenotype of individuals in the WiSH cohort with variants in *ZMYND11*. NA = Not available. Overlapping Human Phenotype Ontology terms are shown in bold.

Although *ZMYND11* is not in the DDG2P, there is sufficient evidence that it should be considered a confirmed disease gene(175, 176). Coe *et al* (175) reported five individuals with truncating mutations in *ZMYND11*, three of which were *de novo* and one was inherited from an affected father. Consistent phenotypic features across these individuals were mild developmental delay, behavioral difficulties and unusual facial features. Cobben *et al* (176) reported a *de novo* missense variant c.1798c / p.Arg600trp in a child with dysmorphic facial features, depressed and broad nasal bridge, hypopigmented eyebrows and lashes. He had severe developmental delay and feeding difficulties. *ZMYND11* lies in the smallest region of overlap of the 10p15.3 microdeletion syndrome and Coe *et al (175)* propose that it is the critical gene associated with the 10p15.3 microdeletion syndrome. Therefore, *ZMYND11* had been erroneously omitted from the DDG2P and was not a newly implicated disease gene. However, this finding proved in principle that grouping individuals with developmental disorders associated with hypertrichosis together is a successful strategy for identifying disease genes.

3.4.12 *NR4A2* is a candidate dominant gene

I identified two individuals with de novo missense variants in the gene *NR4A2* (table 3-8). Individual 267581 carried the mutation c.935G>A p.Arg312Gln (ENST00000339562) and individual 280657 carried the mutation c.866G>C p.Arg289Pro (ENST00000339562). Neither variant is present in the ExAC database. *NR4A2* encodes a nuclear receptor that acts as a transcriptional regulator(177). Zetterstrom *et* al showed that *NR4A2* homozygous knock out mice were hypoactive, were unable to make dopaminergic neurones in the brain and died soon after birth(178). The brains of heterozygous mice contained reduced dopamine levels but otherwise were reported to be healthy. Other nuclear receptors have been implicated in developmental disorders, including *NR2F2*, mutations in which are associated with congenital heart defects, in particular atrial ventricular septal defects(3).

Both individuals have synophrys, developmental delay, and other non-specific phenotypic features. Individual 267581 has short stature, individual 280657 is of normal stature. There are 10 individuals listed on the Decipher database with heterozygous deletions including *NR4A2* (https://decipher.sanger.ac.uk/). Two individuals have deletions that also include fewer than 5 other genes. The first individual (290757) has a 174.47kb deletion that encompasses *NR4A2* and partially deletes *GPD2* (MIM138430), this individual has a behavioral /psychiatric abnormality and delayed speech and language development. The second individual (296098) is reported to have cognitive impairment, as well as including *NR4A2* the deletion in this individual includes 4 other genes, three of which are protein coding: *ERMN*, *GALNT5* and *GPD2.* I next calculated the probability of two probably damaging missense mutations in *NR4A2* not arising by chance, however, this does not achieve significance ($P=9.2x10^{-6}$, where significance is $< 2x10^{-6}$). Therefore, based on the DDG2P criteria for a confirmed disease gene and statistical analysis for *de novo* mutations compared to expectation, *NR4A2* remains a candidate gene and further evidence is needed to assign pathogenicity to these variants.

ID	SEX	CHR	POS	CONSEQ	ALT/ REF	GENO (P/M/F)	PP DNM	ExAC FREQ
267581	**M**	**2**	**157185044**	**Missense** PolyPhen = PROB Dam SIFT=DEL	**C/G**	**1/0/0**	**1**	**0**
280657	**M**	**2**	**157184975**	**Missense** PolyPhen = PROB DAM SIFT=DEL	**C/T**	**1/0/0**	**0.999997**	**0**

Table 3-8 *De novo* missense variants in *NR4A2* identified in individuals with hypertrichosis. ID = Patient ID, CHR = chromosome, POS = Genomic position of the variant, TRANS = Transcript, CONSEQ = Predicted protein consequence, this refers to the transcript ENST00000339562. ALT/REF alternate base(s)/ reference base(s), GENO (P/M/F) = Genotype (Proband/Mother/Father), ppDNM = Posterior probability of the most likely DENOVO genotype configuration from DeNovoGear. (The range of ppDNM is 0-1, a value closer to 1 indicates higher probability of observing a denovo event at this position). ExAC FREQ = frequency of the variant in the ExAC database(120).

I carried out recessive analysis looking for biallelic variants in two or more individuals where at least one of the variants in each individual was loss of function, but I identified

no candidate biallelic variants. Larger sample sizes are needed to identify biallelic disease genes associated with hypertrichosis in this way.

3.4.13 *AKAP14* is a candidate X-Linked gene

Two male individuals were found to have a missense variant in the *AKAP14* gene located on the X chromosome (table 3-9). AKAP14 is an A-kinase anchoring protein, which has been shown to be associated with ciliary axonemes and likely plays a role in the signalling underlying ciliary beat frequency (179). There are no known human diseases associated with *AKAP14*, however there is some evidence that similar protein pathways are disrupted in autism(180). Other ciliary proteins are well known to be implicated in developmental disorders such as Bardet Beidl syndrome, and Varadi Papp syndrome. Ciliary disorders commonly include renal, brain, visual abnormalities and polydactyly and Bardet Biedl syndrome is associated with obesity.

Both individuals had inherited the variant from an unaffected mother. Both boys have synophrys, developmental delay and behavioral abnormalities, including autism in one individual and ADHD in the other. Both individuals have a weight above the normal range. Facially they both have a broad nasal bridge and tip and synophrys. Therefore, there is similarity in the phenotypes of these two individuals and some evidence for AKAP14 as a developmental disorder gene, however there is insufficient evidence that this is a disease causing disease gene at present as per the DDG2P guidelines for a assigning a confirmed DD gene, and it is therefore a candidate gene until there is further evidence of other individuals with variants in this gene.

ID	SEX	CHR	POS	TRANS	CONSEQ	ALT/REF	GENO (P/M/F)	ExAC FREQ
264181	M	X	119037493	ENST00000371431	Missense PolyPhen = PROB DAM SIFT=TOL	A/G	2/1/0	0
274098	M	X	119048820	ENST00000371431	Missense PolyPhen = PROB DAM SIFT=TOL	G/T	2/1/0	0

Table 3-9: AKAP14 variants identified in two male WiSH individuals. Both variants have been inherited from an unaffected mother. ID = Patient ID, CHR = chromosome, POS = Genomic position of the variant, TRANS = Transcript, CONSEQ = Predicted protein consequence, ALT/REF alternate base(s)/ reference base(s), GENO (P/M/F) = Genotype (Proband/Mother/Father), ppDNM = Posterior probability of the most likely DENOVO genotype configuration from DeNovoGear. (The range of ppDNM is 0-1, a value closer to 1 indicates higher probability of observing a denovo event at this position). ExAC FREQ = frequency of the variant in the ExAC database(120).

3.4.14 There is a burden of *de novo* variants in chromatin genes in individuals with WSS-like disorders and hypertrichosis

As WSS is a chromatin modification disorder and other disorders related to chromatin are associated with hypertrichosis, I next investigated whether there is a burden of variants in genes with a role in chromatin structure or function (chromatin genes) in all individuals with hypertrichosis or a phenotype similar to Wiedemann-Steiner syndrome (including those with and those without a diagnosis). Knowledge of the underlying architecture of developmental disorders associated with hypertrichosis would help drive gene discovery in the future.

I compared the number of observed mutations in chromatin genes in the 228 DDD study individuals with to the number of expected mutations using published *de novo* mutation rates(164). There was an increased number of mutations in chromatin genes above

expectation for both loss of function mutations and for loss of function mutations and functional mutations combined ($p = 3.8x10^{-7}$ and $p= 7.9x10^{-5}$ respectively).

3.4.15 Genes encoding proteins that form a complex with *KMT2A*

I next investigated for the presence of *de novo* variants in the genes encoding proteins that form a complex with *KMT2A* hypothesing that *de novo* mutations in these genes may cause similar phenotypes to WSS. There were no rare loss of function or missense variants in these genes in the WISH cohort. I therefore investigated the wider DDD study cohort of 7269 individuals. I identified zero *de novo* loss of function variants in any of these genes, and one *de novo* missense variant in *WDR5* (see table 3-10). Given this lack of variation, I concluded that these genes, given the multiple important complexes their protein products are involved in could be highly conserved essential genes, mutations in which might not be compatible with post natal life. If this was the case I would expect there to be little variation in control databases in these genes, and no evidence from copy number databases that CNVs involving these genes were present in disease or healthy populations.

ID	SEX	CHR	POS	REF	ALT	CONSEQ	GENE	ENST	PP DNM
277291	F	X	153219845	G	A	Synonymous variant	*HCFC1*	ENST00000310441	1
264916	M	9	15465456	C	G	3 prime UTR variant	*PSIP1*	ENST00000380733	0.999999
266639	M	9	137017122	C	T	Missense variant	*WDR5*	ENST00000358625	1

Table 3-10 Table to show the *de novo* variants in genes that encode proteins that form a complex with KMT2A. ID = Patient ID, CHR = chromosome, POS = Genomic position of the variant, TRANS = Transcript, CONSEQ = Predicted protein consequence, ALT/REF alternate base(s)/ reference base(s), GENO (P/M/F) = Genotype (Proband/Mother/Father), ppDNM = Posterior probability of the most likely DENOVO genotype configuration from DeNovoGear. (The range of ppDNM is 0-1, a value closer to 1 indicates higher probability of observing a denovo event at this position). ExAC FREQ = frequency of the variant in the ExAC database(120).

3.4.16 Variants in genes that encode proteins that complex with KMT2A are seen infrequently in population databases

I searched for further evidence of structural or sequence variants in the genes encoding proteins that bind KMT2A causing developmental disorders. I reviewed the Decipher database and identified no copy number variants (CNV)s encompassing these genes(181). I reviewed the population control database (ExAC)(120) and this showed only a small number of loss of function mutations in these genes (Table 3-11). With the exception of *DPY30*, where a high allele count of a loss of function variant suggests it is a sequencing error or common polymorphism. These findings give further evidence of these genes being conserved and that potentially some of them are essential genes, meaning certain variants in which are not compatible with life.

Gene	Total Number of unique LOFs in ExAC	Total allele count of LOF variants	Number of unique missense variants in EXaC	Total allele count of missense variants
ASH2L	3	4	135	1140
DPY30	8	23164	12	43
HCFC1	1	4	234	14187
HCFC2	4	6	128	2058
MEN1	0	0	115	115496
RBBP5	2	2	74	127
PSIP1	15	48	173	1336
WDR5	0	0	47	105

Table 3-11: Number of loss of function (LOF) and missense variants in genes encoding proteins that form complexes with KMT2A from the ExAC database.

3.4.17 Variants in KMT2A complex encoding genes are not present in fetal sequencing studies

I next looked for further evidence that the KMT2A complex encoding gene set contains essential genes. I analysed the Human brain expression data in the Brain Span Atlas of the developing Human Brain (http://www.brainspan.org) (182) and confirmed that all of the KMT2A complex encoding genes are expressed in the human brain, suggesting they could potentially have a role in development or neurological functioning. I next looked at whether variants in these genes were implicated in causing severe developmental abnormalities in foetuses which would give further evidence to them playing an essential role in development. A number of authors have carried out whole exome(183-186) or whole genome(187, 188) sequencing in foetuses or neonates with prenatal ultrasound abnormalities or abnormalities on post-mortem examination. However, I identified no structural variants or sequence variants in these KMT2A complex encoding genes as reported as being causal in these studies(183-188).

In order to look for further evidence that these genes were vital in early development, I next investigated whether there was any evidence that homozygous null mice for these genes had been successfully generated. Crabtree *et al* showed that MEN1 homozygous null mice died in utero at embryonic days 11.5 to 12.5, however, heterozygous mice developed features of Multiple endocrine neoplasia type 1*(189).* Although Stoller *et al* showed assessed *ASH2L* heterozygous knock out mice to be normal, they showed that *ASH2L* null embryos die during early gestation suggesting *ASH2L* is required for the earliest stages of embryogenesis(190). A similar pattern (heterozygous mice not displaying an obvious phenotype, however null or homozygous knock out mice showing embryonic lethality) has also been observed also for *DPY30(191).* Minocha *et al also* showed that knock out of the X-linked gene *HCFC1* was embryonically lethal in male mice. However female heterozygous knock-outs showed only a marginal but significant reduction in body length and lean mass(192). However, Sutherland *et al* studied mice with a homozygous gene trap insertion into *PSIP1* which produces a PSIP1 protein lacking important functional and conserved domains(193). They showed that the majority of mice died on day 1 postnatally, however interestingly a subset of mice survived. Surviving mice had skeletal defects including ectopic ribs, suggesting both that *PSIP1* may play a role in the control of HOX expression, but also that full length PSIP1 is not essential for cell survival postnatally(193). Thus, in summary, where evidence from mouse studies is available, homozygous knock out mice for these genes are generally lethal, giving further evidence that the genes that bind KMT2A are vital in early development.

My investigation suggests that these KMT2A complex encoding genes, are highly important in development, however there are still loss of function variants in some of these genes, suggesting some of these proteins may be more important than others. Further work is needed to elucidate whether variation in these genes or their regulatory regions play a role in developmental disorders, and if so what the mechanisms are for this. I have focused on haploinsufficiency being a potential disease mechanism for these genes, however other mechanisms may be implicated such recessive inheritance or non-coding variation in the case of *HCFC1*.

3.5 Discussion

3.5.1 Summary

To my knowledge no one has previously carried out whole exome sequencing in individuals with hypertrichosis likely due to heterogeneous causes. I have shown in principle this approach can successfully identify disease genes associated with hypertrichosis by identifying *ZMYND11* as a disease gene. I identified that the *de novo* diagnostic yield from carrying whole exome sequencing in individuals with hypertrichosis or WSS like phenotypes was 29% compared to the de novo diagnostic yield of the DDD more generally of 23%, however my cohort was enriched with individuals with a phenotype consistent with Wiedemann-Steiner syndrome, and removing KMT2A mutations, gave a diagnostic yield of 21%. I also showed that my hypertrichosis cohort was enriched for variants in chromatin genes. This suggests hypertrichosis is an indicator that an individual potentially has a chromatin disorder and may be more likely to harbor a diagnostic *de novo* mutation than individuals with developmental disorders more generally. In addition, there is some evidence that seizure disorders also feature in this group, it may be that iatrogenic hypertrichosis due to anti-epileptics result in other mechanisms for hypertrichosis in these individuals.

3.5.2 Grouping by hypertrichosis has proven in principle a successful strategy for gene discovery

My investigation highlighted *ZMYND11* as a disease gene in developmental disorders associated with hypertrichosis. Although this had been erroneously left out of the DDG2P, this discovery has shown in principle the cohort and investigation strategy can successfully discover new disease genes implicated in hypertrichosis associated phenotypes More numbers and larger cohort sizes are needed to discover more hypertrichosis associated genes in the future.

3.5.3 Hypertrichosis is an indicator that an individual's developmental disorder may result from chromatin dysregulation

Although many developmental disorders arise as the result of mutations in chromatin genes, few specific features have been identified as distinguishing individuals with chromatin disorders from those with disorders resulting from different aetiologies. Hypertrichosis may be a useful feature to guide the clinician as to there being an underlying abnormality with chromatin regulation.

HbF levels have been identified as a potential biomarker for chromatin disorders and have been found associated with missense and loss of function mutations in *BCL11A* and microdeletions encompassing and proximal to *BCL11A*(194, 195).

The combination of measuring HbF levels and assessing for hypertrichosis may be helpful in the future for identifying individuals with chromatin disorders who remain without a molecular diagnosis for their disorder, and those with intronic and mutations in functional gene elements that have eluded detecting by conventional sequencing methods. Identifying individuals with mutations in genes encoding chromatin modification is useful in terms of identifying shared problems and developing treatments.

Chapter 4

Investigations into Autosomal Recessive Developmental Disorders

The Deciphering Developmental Disorders Study

4.1 Aims:

1. **To investigate the underlying architecture of severe developmental disorders by carrying out burden analyses for evidence of autosomal recessive disease**

2. **To generate a population matched control dataset for studying individuals with developmental disorders using the untransmitted diplotypes from parent offspring trios**

3. **To contribute to the significant improvement of the diagnosis of children with developmental disorders as a clinician researcher working as a member of the Deciphering Developmental Disorders (DDD) analysis team.**

4.2 Introduction

4.2.1 Developmental disorders and motivation for this investigation

Developmental disorders are a diverse group of conditions that result in abnormal human development. Identifying the underlying genetic causes of these disorders has considerable benefit to affected individuals and their families, healthcare services and society. Strategies to unravel the causes of developmental disorders have been improving for decades, however despite decades of gene discovery efforts large numbers of families remain without a diagnosis for their disorder. A detailed background to developmental disorders is found in Chapter 1: Introduction.

The advent of next generation sequencing approaches has significantly improved discovery of the genetic cause of developmental disorders, however there are many more disorders to discover. Particular challenges to gene discovery in this current era of genomics include accurately assigning pathogenicity to variants, and establishing population matched, technically non-biased, phenotypically healthy control cohorts. In terms of developmental disorders, the contribution of the various types of autosomal disorder to developmental disorders as a whole is unknown.

4.2.2 The Deciphering Developmental Disorders (DDD) Study

With the increasingly widespread use of whole exome sequencing many local, national and international collaborations have been formed to share resources and combine

patient numbers to make diagnoses and facilitate new gene discovery. In the UK, the Deciphering Developmental disorders (DDD) study is a collaborative research study involving researchers from the Wellcome Trust Sanger Institute and Clinical Geneticists and Clinical Scientists from all the Clinical Genetics units in the UK and Ireland(48). The aim of the DDD study is to improve Clinical Genetic practice for children with developmental disorders. The entry criteria are severe, undiagnosed developmental disorders with the majority of individuals recruited having intellectual disability (see Table 4-1).

Inclusion Criteria	Exclusion Criteria
1. Neurodevelopmental disorder AND/OR 2. Congenital anomalies AND/OR 3. Abnormal growth parameters (height, weight, OFC, 2 items >3sd, 1 item >4sd) AND/OR 4. Dysmorphic features AND/OR 5. Unusual behavioral phenotype AND/OR 6. Genetic disorder of significant impact for which the molecular basis is currently unknown (affected family members)	1. Adults with capacity in Scotland 2. Terminations and stillbirths 3. Children with a known molecular diagnosis

Table 4-1 Recruitment criteria for the Deciphering Developmental Disorders (DDD) Study. Inclusion criteria and exclusion criteria for the DDD study.

The DDD study has a recruitment network of 180 clinicians recruiting from 24 regional genetics services throughout the UK and republic of Ireland. The DDD study uses whole exome sequencing in trios (proband and both parents) to make diagnoses and to facilitate the discovery of new genes implicated in developmental disorders. To enable consistent phenotyping using standardised terms, Probands and their parents undergo detailed clinical phenotyping using Human Phenotype Ontology (HPO)(58) terms (see also Chapter 1: Introduction) by their local clinician. This phenotypic information is entered into a portal in the Decipher database(118) alongside anthropometric data and information about family history, birth history, pregnancy and neuro-imaging. Decipher

facilitates further gene discovery through recording variation in a standardised updatable manner and making this available to clinicians worldwide(118).

A clinician curated database is used in order to facilitate the feedback of causal gene variants within the DDD Study. The Development Disorder Genotype-2-Phenotype Database (DDG2P) is a database of published genotype-phenotype relationships for genes associated with developmental disorders(196). The DDG2P was curated from data obtained from UniProt, OMIM and a systematic screen of the *American Journal of Human Genetics* and *Nature Genetics* since 2005. The DDG2P is updated regularly to incorporate new developmental disorder genes as they are published, or further evidence about the relationship of a gene with a developmental disorder. The DDG2P is categorised into the level of certainty that the gene causes developmental disease (confirmed, probable or possible), the mechanism of associated mutations (e.g. loss-of function, activating) and the allelic status associated with disease (e.g. monoallelic, biallelic) see Table 4-2.

Category	**Choices**
level of evidence for Developmental disorder association	Confirmed DD gene, Probable DD gene, Possible DD gene, Not DD gene, IF gene, DD and IF gene
Inheritance mode	Monoallelic, Biallelic, Both, Imprinted, Digenic, Hemizygous, X-linked dominant, Mosaic, Mitochondrial, Uncertain
Mutation type	Loss of function, All missense/in-frame, Dominant negative, Activating, Increased gene dosage, Cis-regulatory or promoter mutation, Uncertain

Table 4-2: Summary of the curation categories for genes associated with developmental disorders in the Development Disorder Genotype-2-Phenotype Database DDG2P clinician curated database. DD = Developmental disorder, IF = Incidental finding.

The DDD study has a bioinformatics pipeline to filter and flag variants in DDG2P genes for clinical reporting. In addition, multiple analyses are carried out to drive discovery of new genes, including analyses aimed at discovering new genes underlying specific modes of inheritance such as dominant disorders and recessive discovers. One of my key roles in the DDD study was to investigate autosomal recessive inheritance in the first 1133 trios.

4.2.3. Recessive gene discovery: A short history

Homozygosity (or autozygosity) mapping in consanguineous families has been a powerful approach to identify the cause of rare autosomal recessive conditions(197, 198). Consanguinity, usually defined in Clinical Genetics as a union between a couple who are second cousins or closer(199), is common in many cultures(199, 200). Consanguinity increases the coefficient of inbreeding (proportion of the genome which is identical or homozygous by descent) and therefore increases the likelihood of pathogenic mutations in a homoallelic state. Homozygosity (or autozygosity) mapping has been modified and improved in line with advances in technology. The underlying principle is that a hypothesis-free genome-wide search is carried out for overlapping blocks of homozygosity in affected individuals, usually from multiple different families. Then the disease causing mutation is identified through sequencing genes within overlapping regions. At first the detection of autozygous regions was carried out by genotyping individuals with panels of highly polymorphic microsatellite markers, subsequently single nucleotide polymorphism (SNP) arrays(201) were used.

In terms of limitations, despite the use of SNP arrays and computational analysis for linkage, the capillary sequencing involved to identify the disease gene is extremely time consuming, particularly in gene-rich areas or for large candidate intervals. This technique is also heavily dependent on the availability of consanguineous families. A significant proportion of individuals with recessive diseases are not the product of a consanguineous union, however gene mapping for recessive disorders in outbred populations has been much more difficult than autozyosity mapping(202). Limited linkage information from nuclear families and the heterogeneity of causative mutations in these families, are reasons why gene mapping has been so difficult in outbred populations.

The first developmental disorder solved by whole exome sequencing was the autosomal recessive condition Miller syndrome(41) (see chapter 1: Introduction). Since this time, next generation sequencing techniques have increasingly been employed as a fast alternative for sequencing genes within the overlapping blocks of homozygosity to high depth when carrying out homozygosity (autozygosity) mapping. Makrythanasis *et al* carried out autozygosity mapping and whole exome sequencing and array CGH in 50

consanguineous families with neurodevelopmental disorders and reported a diagnosis rate of 38% in 18 families for variants in known disease associated genes (1 though array CGH, 17 through whole exome sequencing)(203). However, these studies are limited by the necessity of investigating consanguineous families, small numbers and the difficulty of assigning pathogenicity to variants. Other authors have carried out recessive gene discovery using Array-comparative Genomic Hybridisation (aCGH) in combination with whole exome sequencing. Aradhya et al found 10.1% of 138 families (who had been found to have a single mutation in a bilallelic gene on sequencing) were found to have a CNV on the other allele through exonic array CGH. Array CGH has also been used in combination with a SNP array to detect a homozygous disease causing CNV in a region of autozygosity in single families, each within a large study combining SNP arrays and array CGH(203, 204).

4.2.4 Challenges to recessive gene discovery

During the last decade, the identification of *de novo* dominant copy number variants improved the diagnosis of genetically heterogeneous developmental disorders (reviewed by Mefford *et al(205))*. More recently with the advent of next generation sequencing technologies, the identification of *de novo* single nucleotide variants (SNVs) and small insertions and deletions (indels) has revolutionised the diagnosis and understanding of sporadic developmental disorders(85, 206, 207). In dominant disorders, *de novo* mutations are so rare they give a clue about causality, however everyone has some homozygous or compound heterozygous missense variants that are harder to assign pathogenicity to and understand. Also for some recessive diseases which require there to be one loss of function allele and one hypomorphic allele for pathogenicity, (as bilallelic loss of function alleles would likely not be compatible with life, and biallelic hypomorphic alleles would not cause disease), it would be impossible to detect the underlying genetic cause for these disorders using linkage in a consanguineous population using this technique.

4.2.5 Summary

The advent of next generation sequencing approaches has significantly improved discovery of the genetic cause of developmental disorders, however there are many

more disorders to discover and the contribution of the various types of autosomal disorder to developmental disorders as a whole is unknown. The DDD study is a national study to improve the diagnosis of developmental disorders that employs genome wide techniques to diagnose multiple underlying genetic mechanisms causing developmental disorders.

4.3 Methods

4.3.1 Whole exome sequencing within the DDD Study

DNA and or saliva samples were sent to the Wellcome Trust Sanger Institute (WTSI) from regional genetics centers for processing and sequencing by the WTSI core facility.

Quality control, including confirmation of family structure and gender

On arrival at the WTSI individual samples were evaluated for DNA quality, call rate and average heterozygosity using a Sequenom assay (Sequenom, San Diego, USA). For quality control, in order to detect and remove poor quality samples individual samples with a heterozygosity value below 0.195 or above 0.756 or a call rate less than 0.74 were failed. Trios were analysed for mismatches between the genotyped gender versus the stated gender versus in sequenom data. Trio samples were also analysed for the likelihood of the expected pedigree structure. This assessed for sample mix ups and non-paternity and non-maternity. All pedigrees demonstrating non-standard relatedness were evaluated manually before any further sample processing was allowed to occur.

Whole Exome Sequencing

Whole exome sequencing was carried out on DNA samples from all probands and both parents using SureSelect RNA baits: Human All Exon V3 Plus with custom ELID #C0338371 (Agilent, Wokingham, UK), and 75 base paired-end sequencing on the HiSeq™ 2000 platform (Illlumina, saffron Walden, UK). The bait design used incorporates 271,063 bait regions and includes the Agilent Sanger-Exome (Human All Exome 50mb Kit) with an additional 57,680 bait regions used to cover ultra-conserved regions, heart enhancers and additional enhancer regions. The median sequencing

depth was 90X across the whole targeted sequence with 95% of samples having an average sequencing depth in excess of 65X. The WTSI core facility carried out all of this work.

SNV and INDEL Detection (GAPI pipeline at the Wellcome Trust Sanger Institute)

The Genome Analysis Production Informatics (GAPI) pipeline at the Wellcome Trust Sanger Institute was used to process all Binary Alignment/Map (BAM) files. The reference genome (GRCh37_hs37d5) was used for read mapping. Picard (version 1.46) was used to mark duplicate fragments, GATK (version 1.1) was used to perform local realignment around INDELS and was then used to recalibrate base qualities. SNVs were called with GATK using the UnifiedGenotyper, INDELS and SNVs were called with Samtools (version 0.1.16) mpileup options -d 500 -C50 -m3 -F0.002 and variants were filtered using the vcfutils.pl utility and options –p -d 4 -D 1200 from Samtools. A dedicated INDEL caller, Dindel (version 1.01) was used to call a further set of INDELS. Individual single sample variant call formatted (VCF) files were produced by the GAPI pipeline for each caller (Samtools, GATK and Dindel). These individual files were then combined into a merged VCF file. Resolution of merging conflicts was carried out in the following caller order: Dindel, GATK, Samtools where the first caller in this list (the primary caller) was used to define the position and genotype of the variant. The Genome Analysis Production Informatics (GAPI) pipeline team at the Wellcome Trust Sanger Institute carried out this work.

4.3.2 Concepts behind my method: Transmission of disease alleles and burden analyses

There are two concepts that were important in the conception of the method I used in my investigation. These were the ‘transmission of disease alleles’ and the concept of ‘burden’. I will detail these here:

Transmission of disease alleles

If the proband has a recessive disorder it is expected that they have inherited a disease allele from each of their parents. If the proband has a new dominant disorder, then this has not been inherited from a germline variant in either of their parents. If a proband has

a dominant disorder and they have inherited this from one of their parents who is also affected (or a carrier female in the case of X-linked disorders), then the proband would also be expected to have inherited the disease allele from this parent. Therefore, the alleles the proband hasn't inherited from their parents, when put together, form the genome of a theoretical human whose phenotype would be expected to be normal. This is because for all of the above scenarios the disease alleles have been passed to the proband, or arose *de novo* in the proband in the case of a new dominant disorder. Processed whole exome sequencing data in variant call format files doesn't give the whole sequence at every allele. However, it does give all of the variants from the reference sequence. Therefore, taking all of the variants from each parent that the proband didn't inherit and putting them together gives the 'untransmitted diplotype control' for that trio. In summary, If the cause of the proband's developmental disorder is genetic then it results from a variant or variants they carry or a structural rearrangement or imprinting defect within their genome. Therefore, an individual inheriting the variants carried by both parents, that the proband did not inherit, (the 'untransmitted diplotypes') is predicted to be no different from a random individual in the population. The untransmitted diplotype control is also matched to the population of the proband and their parents and the data has been processed in the same way as that of the proband. This analysis was carried out prior to the generation of the ExAC database which contains control data from around 60,000 individuals from exome sequencing studies(120) and therefore there were less control data available at this stage.

Burden

Burden refers to the enrichment of a defined subclass of variation in cases, over null expectation. For example, Girirajan *et al* investigated children with intellectual disability and showed that children with multiple severely damaging copy number variants (a greater burden) had neurological and specific organ deficits in more domains than those with a single variant(208). The presence of a burden of a subclass of variation does not implicate any one variant as causal. Instead, it demonstrates the relevance of that class of variant, and prioritizes it for further investigation. In addition, burden analyses may help dissect the underlying architecture of genetic disorders by enabling an estimation of the proportion of variants of a particular class that are likely to be pathogenic. For example, burden analysis may show that recessive diseases contribute significantly to

undiagnosed developmental disorders, by showing an enrichment of inherited pathogenic alleles inherited from unaffected parents in affected individuals compared to controls. Alternatively, they might highlight a contribution of recessive disease to developmental disorders by demonstrating an excess of compound heterozygous or homozygous loss of function or protein altering variants in affected individuals compared to controls. Also, if there was evidence that a significant number of undiagnosed developmental disorders have recessive inheritance, it may help give parents empiric recurrence risks for future pregnancies.

4.3.3 I merged and filtered variant call format files (VCFs)

In order to generate the untransmitted diplotype control, for each trio, I merged the mother, father and proband's VCF files using VCF tools(160). From the merged VCF files, I wrote custom programs in Perl to generate the untransmitted diplotype controls. To improve variant quality and reduce the inclusion of sequencing errors I removed non 'PASS' variants. In order to remove sites where artifacts are likely, I removed variants with multiple reference alleles or multiple alternate alleles. Finally, I removed intronic and upstream variants. In addition, I removed indels, CNVs, X and Y chromosome variants. I next calculated the genotypes of the untransmitted diplotypes based on the genotypes of the mother, father and proband for each trio (table 4-3).

Mother	Father	Proband	Untransmitted diplotype
0/0	0/1	0/1	0/0
0/0	0/1	0/0	0/1
0/0	1/1	0/1	0/1
1/0	0/0	0/1	0/0
1/0	0/0	0/0	1/0
1/1	0/0	1/0	1/0
1/1	1/1	1/1	1/1
1/1	0/1	0/1	1/1
1/1	0/1	1/1	0/1
0/1	1/1	1/1	0/1
0/1	1/1	1/1	0/1
0/1	1/1	0/1	1/1
0/1	0/1	0/1	0/1
0/1	0/1	1/1	0/1
0/1	0/1	0/0	1/1

Table 4-3: Calculation of the genotype of the untransmitted diplotype for each trio. The genotype of the untransmitted diplotype was calculated based on the genotypes of the mother, father and proband for each trio. For example, if the genotype of the mother was 0/0 and the father was 0/1 and the proband was 0/0, it could be concluded that the untransmitted diplotype genotype was 0/1.

For every variant carried by the mother, father or proband I calculated the genotype of the proband at this allele. For example, if the mother and father both have the genotype 0/1 and the proband has the genotype 1/1, the genotype for the untransmitted diplotype for this variant would be 0/0. If the mother, father and proband all have the genotype 0/1, then the untransmitted diplotype would also have the genotype 0/1.

4.3.4 I removed variants that did not fit with Mendelian inheritance

When calculating the genotype of the untransmitted diplotype, I identified some genotype combinations in the mother, father and proband that were not compatible with Mendelian inheritance (Non-Mendelian variants). As it had already been confirmed that the family relationships were correct (see above), I could exclude non-paternity or non-maternity as the cause of these erroneous variant combinations.

In order to determine how to process these non-Mendelian variant combinations, I investigated their underlying cause by studying a single trio (FAMP100003). In FAMP100003, there were a total of 94,497 variants present in either the mother, father or proband or in a combination of all three. Of these 94,497 variants, 3288 variants had a trio genotype configuration not consistent with Mendelian inheritance.

I first considered why these mother, father, proband genotype combinations had occurred and if the reason was identified how this would affect the interpretation of what the untransmitted diplotype's genotype would be for that trio. For example, the mother, father, proband genotype combination 0/0, 0/0, 0/1 could be caused by: 1. A false positive variant in the proband, 2. A false negative variant in the mother or father or 3. A real de novo mutation in the proband. However, whatever the cause of this genotype combination, the resulting genotype for the untransmitted diplotype would be 0/0. I therefore removed the non-Mendelian variants that would not make a difference to the untransmitted diplotypes's genotype (table 4-4).

Mother's Genotype	Father's Genotype	Proband's Genotype
0/0	1/1	1/1
1/1	0/0	1/1
0/0	1/1	0/0
0/1	1/1	0/0
1/1	0/0	0/0
1/1	0/1	0/0
1/1	1/1	0/0
1/1	1/1	0/1

Table 4-4 Non-Mendelian variants, which will not affect the untransmitted diplotypes genotype
Non-Mendelian variants in the mother, father, proband, which whatever the cause for the abnormal genotype combination would not make a difference to the untransmitted diplotypes genotype.

I next sought to investigate the cause of the remaining 412 non-Mendelian variants for which the genotype of the untransmitted diplotype could not be calculated (table 4-5).

Mother's Genotype	Father's Genotype	Proband's Genotype	Number of variants
0/0	1/1	1/1	41
1/1	0/0	1/1	56
0/0	1/1	0/0	97
0/1	1/1	0/0	27
1/1	0/0	0/0	112
1/1	0/1	0/0	32
1/1	1/1	0/0	17
1/1	1/1	1/0	30
		Total	**412**

Table 4-5: Non-Mendelian variants in a single trio. Number and configuration of the variants not compatible with Mendelian inheritance observed in a single DDD Study trio following removal of variants on the X and Y chromosome and Intronic and upstream variants and those variants that would not affect the untransmitted diplotype's genotype whatever the reason for the erroneous genotype combination.

I investigated these 412 variants by first interrogating their read depth metrics in the merged VCF file. To determine whether there was sufficient read depth to confirm the variant and further analyse it, I assigned a cut-off value of 7 reads or greater. I selected this figure (≥ 7 reads) as this was a cut-off already used internally within the DDD study as a filter for assessment of apparent *de novo* mutations. Interrogation of the read depth metrics in the merged VCF file showed only 132 of the 412 variants had a read depth in the individuals of the trio who carry the variant of ≥ 7 reads. However, as read depth metrics are only given for a certain locus in the merged VCF file for individuals who themselves carry the variant I therefore sought to determine the read depth at the loci of all 132 variants in all three individuals by reviewing each of them manually using the Integrative Genomics Viewer (IGV)(209, 210). Manual review using the IGV of the loci of each of the 132 variants in the mother, father and proband showed that only 64 variants had a read depth of ≥ 7 reads in all three individuals. I concluded from this that low read depth results in bad quality variants which adversely affect my analysis.

I next selected these 64 variants with adequate read depth for further analysis. I reviewed each of them manually using the IGV, to estimate the most likely true genotype combination. For this analysis, I grouped these variants into those of the same non-

Mendelian genotype combination, hypothesising that the underlying cause is the same for each group of variants. Using a combination of visual inspection and the alternate and reference allele read count at each loci, I estimated the mostly likely real genotype combination using the IGV for each variant (Table 4-6). If all the reads (or the overwhelming majority of reads) showed the alternate allele, then I deduced the genotype was 1/1 (2). If none of the reads (or only one or two reads showed the alternative allele I deduced the genotype was 0/0 (0). If 50% of the reads (or by if by eye around 50% of the reads) showed the alternative allele, I deduced that the genotype was 0/1(1). I accepted that this process was unlikely to be fully accurate, however I carried this out to help determine why these non-Mendelian variants existed and to determine what to do with them. To use an illustrative example, if encountering the IGV plot shown in figure 4-1 which was called as 2.2.1 in the mother, father, child respectively. If reviewing this by eye I would note that each of the individuals (mother, father and child) all have the vast majority of reads showing the alternate allele, therefore I would conclude that the true mother, father, child genotype at this position is 2.2.2.

Figure 4-1: Example Integrative Genomics Viewer (IGV) plot to demonstrate deduction of likely true genotype combination.
This Integrative Genomics Viewer (IGV) plot shows the reads for the mother, father, child at position Chromosome 1, genomic co-ordinates: 987200. The top reads refer to the mother, the middle reads the father, and the bottom reads refer to the proband. The genotype for this trio was called as 2.2.1 in the mother, father and child respectively at this base position. However, reviewing this IGV plot gives evidence that the true genotype at this position is 2.2.2.

For some non-Mendelian genotype combinations, the non-Mendelian genotype combination appeared to be the most likely genotype. For other non-Mendelian genotype combinations, different variants appeared to have different most likely real Mendelian genotypes, suggesting that the underlying cause for the same non-Mendelian genotypes

may not be the same for each variant. For two variants it was difficult to deduce what the most likely genotype combination was, and these were labelled as unclear.

Non Mendelian Genotype combination	Number of Variants	Estimated Real Genotype combination
0.2.0	7	0.2.1(5), 0.2.0(2)
2.0.0	2	1.0.0 (1) Unclear (1)
1.2.0	1	1.2.0
0.2.0	6	1.2.1
1.2.0	4	1.2.1
0.2.2	6	1.2.2
2.0.0	12	2.0.1 (9) 2.0.0 (3)
2.0.2	2	2.0.2
2.1.0	2	2.1.1
2.0.2	6	2.1.2
2.2.1	15	2.2.2
2.2.0	1	Unclear
Total	64	

Table 4-6: Number of variants with non-Mendelian genotypes per non-Mendelian genotype combination with estimated real genotype. Variants with mother, father, proband genotype combinations that were not compatible with Mendelian inheritance were grouped by genotype combination and each manually reviewed using the Integrative Genomics Viewer (IGV) to estimate the mostly likely real genotype. Column 1 shows genotypes in the order: Mother.Father.Proband. Column 3 (Estimated real genotype) shows genotypes in the order: Mother.Father.Proband. The numbers in brackets show how many variants showed that real genotype combination.

Therefore, in total, of the 412 non-Mendelian variants in trio FAMP100003, 348 did not pass the read depth cut-off of ≥ 7 reads. Of the 64 variants with sufficient read depth, the most likely real genotype combination could not be determined in all cases. I therefore decided for ongoing analysis that variants that did not show a Mendelian pattern of inheritance within a trio would be filtered out as the non-Mendelian variants are likely erroneous and result from low read depth. As the number of variants implicated

is relatively small per trio I concluded the effect per trio on the downstream analysis would be minimal.

4.3.5 I filtered variants by QUAL score

I studied a subset of 10 trios to determine whether the number of variants carried by the inherited diplotypes seemed appropriate. I investigated the number of common and rare (MAF <0.01) SNVs in the probands and untransmitted diplotypes (Figure 4-2). A greater number of common and rare variants were observed in the untransmitted diplotypes than in the probands.

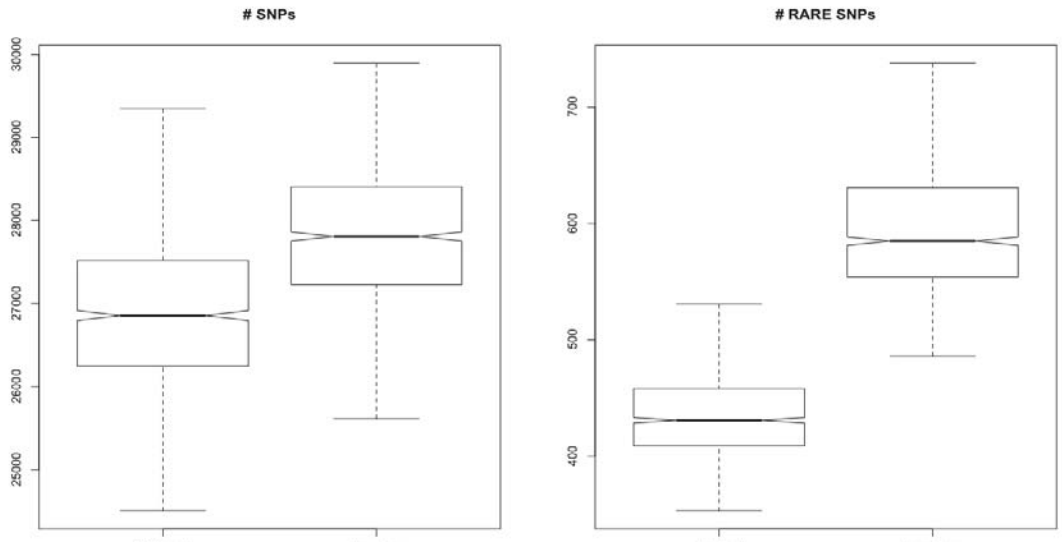


Figure 4-2: Number of single nucleotide variants (SNVs) in the probands and untransmitted diplotypes. A. Common SNVs. B. Rare SNVs MAF <0.01. These figures were plotted using data generated from a subset of 200 probands and 200 untransmitted diplotypes.

In order to determine the reason for the discrepancy in the number of common and rare variants between the probands and untransmitted diplotypes I investigated the relationship between the QUAL score and number of variants. QUAL (variant quality score) is a phred-scaled quality score generated by GATK (161). The QUAL score is an estimate of the confidence that the variant caller correctly identified that a given genome locus exhibits true variation in at least one sample, i.e. that there is a true variant and not an artefact resulting from sequencing, alignment or data processing.

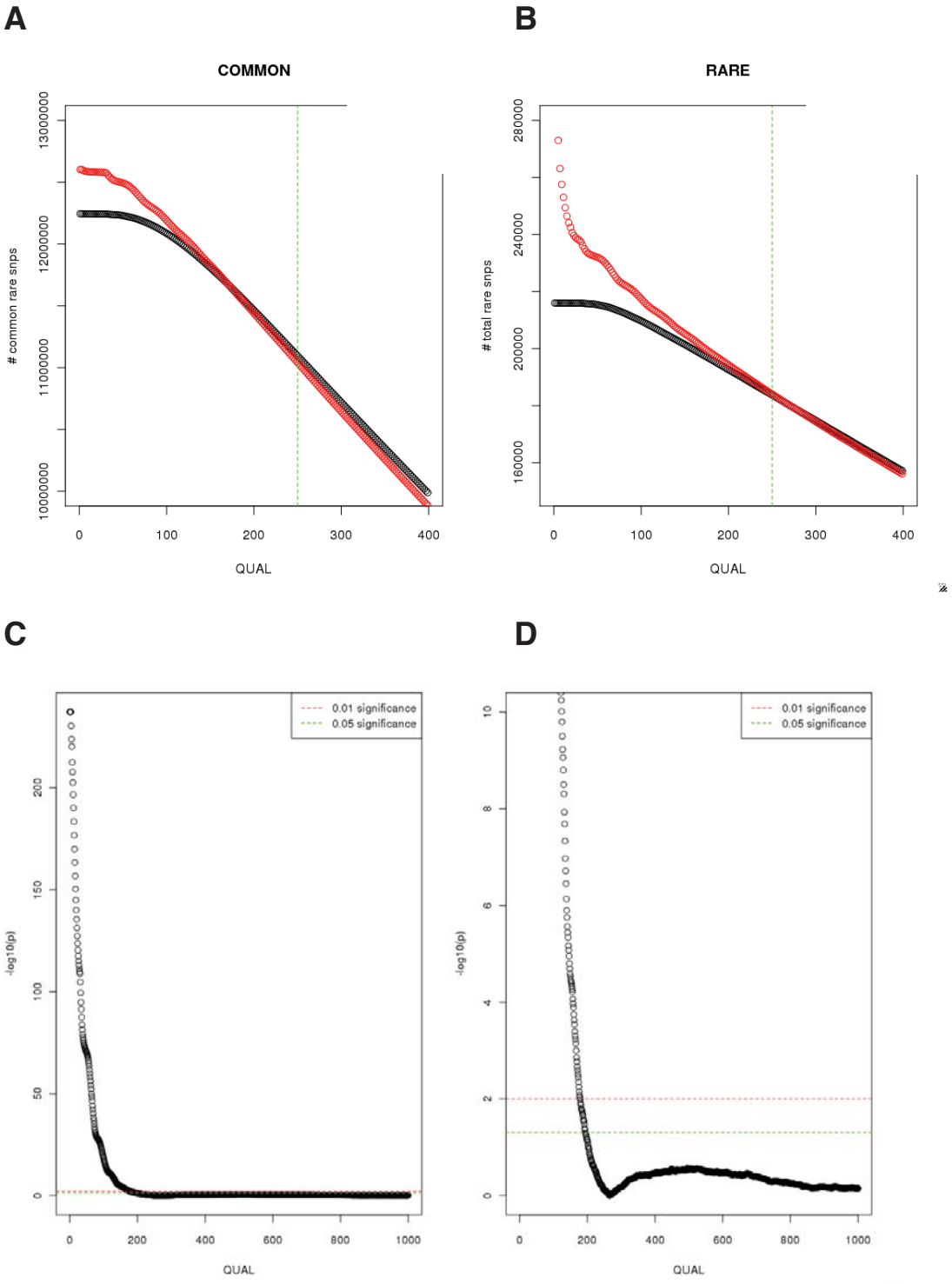


Figure 4-3: Studies of QUAL score and number of SNPs per sample for the probands and untransmitted diplotypes. **A:** Number of common variants against QUAL score for the probands and untransmitted diplotypes. Red dots represent untransmitted diplotypes, black dots represent probands. **B.** Number of rare (MAF <0.01) variants against QUAL score for the probands and untransmitted diplotypes. Red dots represent untransmitted diplotypes, black dots represent probands. A randomly selected subset of approximately 10% of 1139 probands and untransmitted diplotypes datasets were used to generate plots A and B. **C and D:** log10(p) versus QUAL threshold for a Mann-Whitney test of the numbers of rare SNPs per sample in probands and untransmitted diplotypes against QUAL score.

I first sought to identify low quality variants as the numbers of these are potentially likely to be different between the proband and the untransmitted diplotypes and they may result from low read depth. Therefore, I investigated the relationship between QUAL score threshold and statistical significance between the numbers of rare (MAF <0.01) SNPs per sample in the probands and untransmitted diplotypes. I sought to pick a QUAL score threshold that largely eliminated the difference in variant numbers between the proband and untransmitted diplotypes.

In order to do this, I carried out a Mann-Whitney test (with assistance from Tomas Fitzgerald) between the number of rare SNPs in probands compared to untransmitted diplotypes vs. QUAL(Figure 4-3). I selected nominal (uncorrected) p value cut-offs of 0.05 and 0.01 to assess what QUAL score values these corresponded to. From Figure 4-3, the 0.01 significance level was determined as being a QUAL score of 179 and the 0.05 significance level was determined to be a QUAL score of 194. Plots of number of rare variants in probands and untransmitted diplotypes using QUAL score cut-offs of 179 and 194 are shown in Figure 4-4.

I selected a QUAL score cut-off of 179 as a relatively conservative filter for the subsequent analyses. At this threshold the difference in variant numbers between the proband and untransmitted diplotypes was largely eliminated removing many of the low quality variants. I appreciated that there remained a modest difference between the number of variants in the probands and untransmitted diplotypes and that there was a balance between removing low quality variants and removing diagnoses.

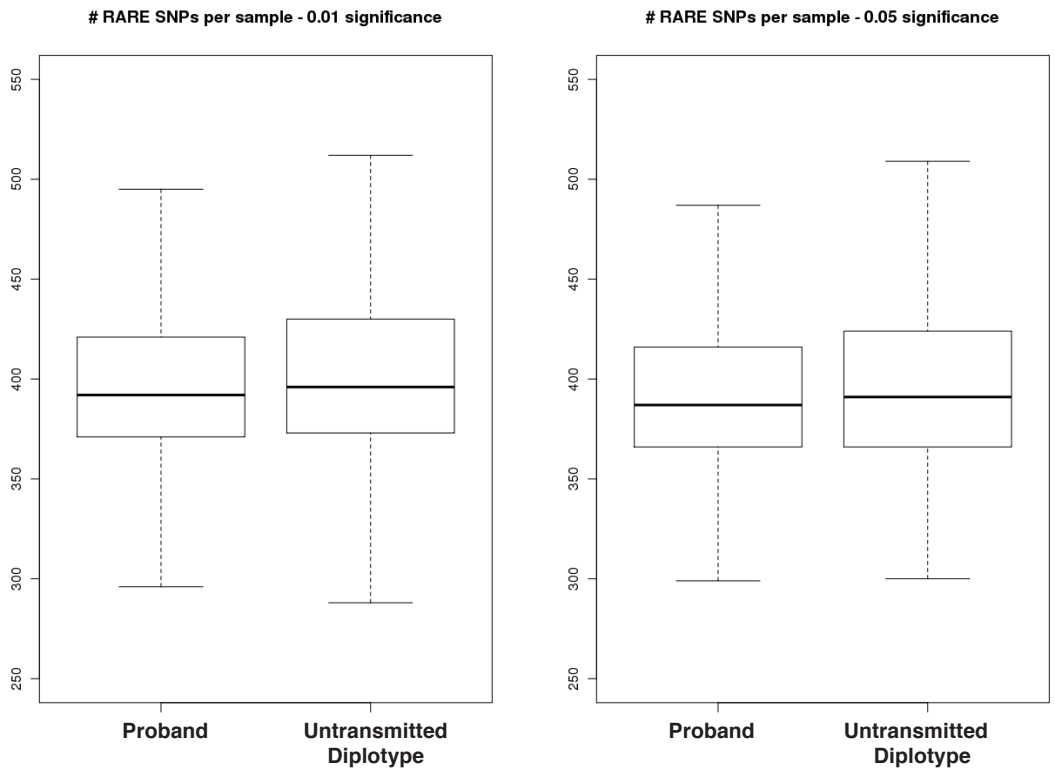


Figure 4-4: Boxplots to show number of single nucleotide variants (SNVs) in the probands and untransmitted diplotypes. A. Common SNVs. B. Rare SNVs: (MAF <0.01). These figures were plotted using data generated from a subset of 200 probands and 200 untransmitted diplotypes.

4.3.6 I removed trios with extreme variant numbers

I compared the numbers of rare variants carried by the probands and the untransmitted diplotypes. I noted that there were some outlying individuals with low or high numbers of rare variants. I first explored what these outliers represented - the hypothesis was that ancestry could be resulting in the differences between rare variant number and this was further investigated and confirmed by carrying out a principle component analysis (PCA), this was carried out by Tomas Fitzgerald (figure 4-5). For on-going analysis, I filtered out trios that had extreme rare variant numbers (less than 200 or greater than 6000) in the proband or the untransmitted diplotypes to prevent trios with extreme variant numbers from adversely affecting the investigation.

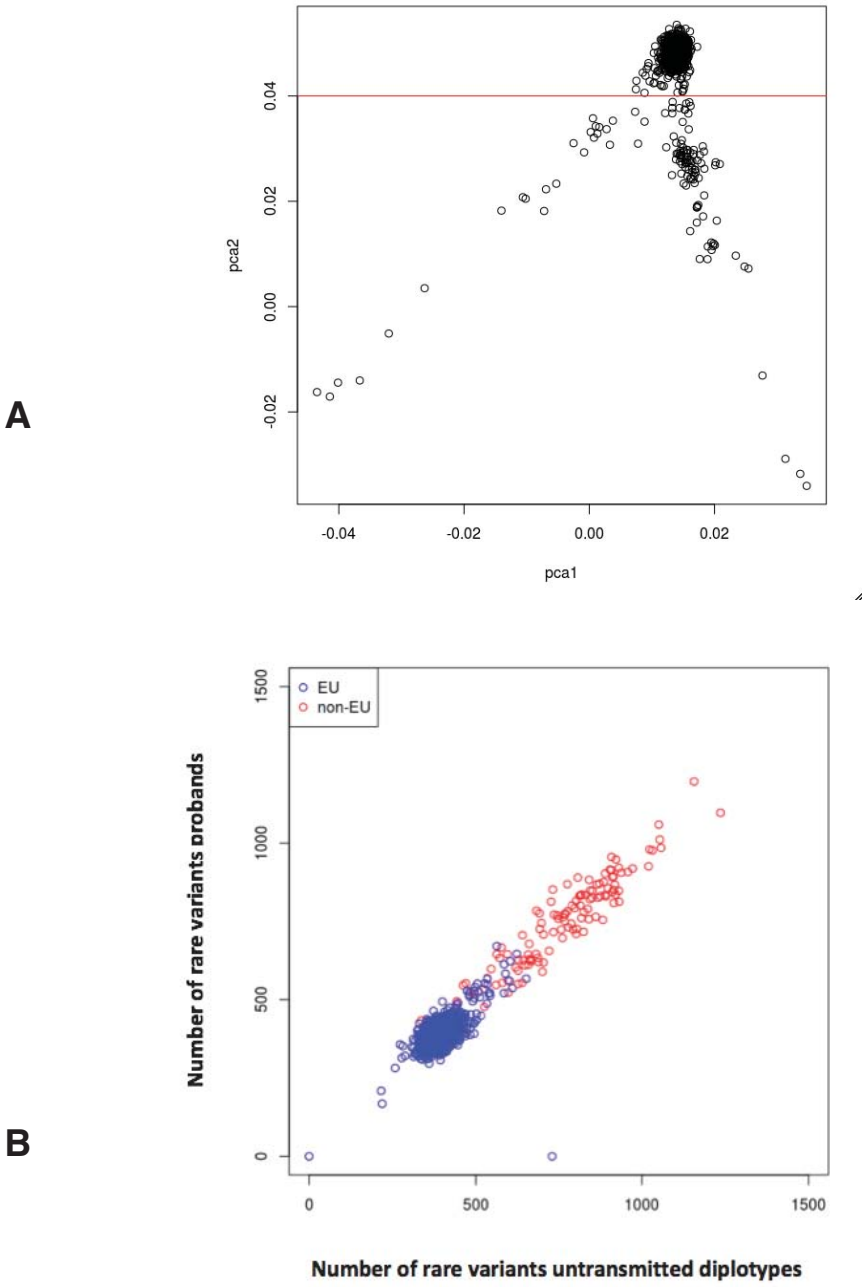


Figure 4-5: A: Principle component analysis (PCA) of individuals in the DDD.
European ancestry was defined by a PCA2 value of greater than 0.04. B: Comparison of the number of rare variants in the untransmitted diplotypes and probands. Variant numbers are labelled by ancestry, Blue circles represent European ancestry, red circles represent non-European ancestry. European ancestry was defined by a PCA2 value of greater than 0.04.

4.3.7 I generated cumulative haplotype counts of rare SNVs

Using 1127 trios, in order to perform burden analyses, I generated cumulative counts of 1. haplotypes containing rare (minor allele frequency of ≤ 5%) variants, and 2. rare bilallelic variants for the probands and untransmitted diplotype for each gene in the

genome by each variant type. I used the variant classification shown in Table 4-7, adapted from the classification devised by Purcell *et al(211).* Where there was more than one variant in the same gene on the same allele, I selected the variant with the most severe consequence (loss of function / disruptive > damaging > functional > Silent). I processed the proband's exome variant profiles used in this analysis in the same way as the untransmitted diplotypes, i.e. they had had specific variant types removed as in Figure 4-6.

Variant classification	**Type of variant**
Disruptive	Stop gained, transcript ablation, splice donor variant, splice acceptor variant, frameshift variant
Disruptive / Damaging	All of the disruptive variants plus functional variants predicted to be damaging by two algorithms: SIFT-Deleterious and PolyPhen-Probably damaging)
Functional	Missense, inframe deletion, inframe insertion, coding sequence variant stop lost (not fulfilling the above criteria for 'Damaging')
Silent	Synonymous variant

Table 4-7: Classification used for variants when generating cumulative haplotype counts. This classification was adapted from that devised by Purcell *et al(211).*

4.3.8 I compared filtered variant ratios to those observed in autism

To determine whether the number of filtered variants and filtered variant ratios were similar to those identified in other studies, I compared the number of rare (minor allele frequency of ≤ 5%) variants predicted to completely knock out the encoded protein product (homozygous or compound heterozygous loss of function variants) observed in our probands and untransmitted diplotypes to published data from children with autism(212), see Table 4-8. Our figures were significantly higher than those identified in individuals with autism and their controls used, also the individuals with autism carried more variants than the controls, whereas in our study it was vice versa. I sought to further investigate this discrepancy by investigating the effect on numbers of variants

and effect on variant ratios by excluding certain subgroups and by looking for genes that may be skewing the ratios and numbers.

	Rare (≤ 5%) Heterozygous LoF Variants in 1127 probands and 1127 untransmitted diplotypes	**Rare (≤ 5%) Homozygous or comp het LoF Variants**	**Number of complete knock out events per individual**	**Number of complete knock out events per individual (Lim *et al* 2013) in 933 probands and 869 controls**
Probands	16976	119 + 14 = 133	0.118	0.066
Controls / Unitrans. Diplo.	16965	138 + 15 = 153	0.135	0.033

Table 4-8: Comparison of the number of Rare (≤ 5%) filtered variants observed in our probands and untransmitted diplotypes compared to previously published data from children with autism(212).

4.3.9 Investigating the discrepancy of our ratios with those in autism

I took a number of approaches to investigate why there was a discrepancy between our figures and those observed in individuals with autism. I first investigated whether one or more genes harboured excessive numbers of homozygous variants with a minor allele frequency of ≤ 5% in the probands or untransmitted diplotypes and was affecting the ratios observed in our data. I next investigated the genes that harboured homozygous loss of function variants in more than one 'individual' (proband or untransmitted diplotype), see Table 4-9. However, overall the numbers of genes and 'individuals' this involved were small and I didn't think this was contributing to the discrepancy observed between ratios in my data and that of Lim *et al(212).*

A

Number of homozygous loss of function variants	Number of genes
0	19075
1	77
2	18
3	2

B

Number of homozygous loss of function variants	Number of genes
0	19068
1	81
2	15
3	6
4	1
5	1

Table 4-9 Number of genes with homozygous variants in probands and untransmitted diplotypes. (A) Number of genes with homozygous loss of function variants with a minor allele frequency of ≤ 5% in 1127 probands. (B) Number of genes with homozygous loss of function variants with a minor allele frequency of ≤ 5% in 1127 controls (untransmitted diplotypes).

4.3.10 I filtered out consanguineous trios

I next compared the number of rare (MAF ≤ 5%) loss of function and synonymous variants in consanguineous versus non-consanguineous trios. Using King Score(213). The King score is an estimation of the kinship coefficient (degree of consanguinity) between any two individuals. It is obtained by using a rapid algorithm for relationship inference that allows the presence of unknown population substructure(213). I defined consanguineous families as those having a King Score > 0. Removing the probands from consanguineous trios resulted in the relationship of the numbers of rare homozygous variants between probands and controls becoming more consistent with

the figures published in autism(212), see Table 4-10. In the study of individuals with autism there were approximately twice as many complete knock-out events in affected individuals than in controls. However, in this analysis, with the consanguineous families included there were a larger number of complete knock out events in the untransmitted diplotypes than in the probands. Removing the probands from consanguineous trios from my analysis resulted in the numbers of complete knock-out events being more equal between probands and untransmitted diplotypes. Therefore, it can be concluded, the probands and untransmitted diplotypes from consanguineous trios harbour large numbers of homozygous variants. To prevent generating untransmitted diplotypes with homozygosity by descent, I removed consanguineous families (N=47) from this analysis.

	Number of rare complete knock our events per individual		
	All trios	**Consanguineous trios**	**Non-Consanguineous trios**
Probands	0.118	0.638	0.095
Untransmitted diplotypes	0.136	0.915	0.102

Table 4-10 Number of rare complete knock our events per individual
Rare means ≤ 5%. Complete knock out = compound het and homozygous events. For consanguineous trios, N=47 individuals (47 probands and 47 untransmitted diplotypes). For Non-consanguineous, N=1080 (1080 probands and 1080 untransmitted diplotypes). For all trios, N= 1127 (1127 probands and 1127 untransmitted diplotypes)

4.3.11 QUAL 1000 filter improves ratios but likely removes diagnoses

With consanguineous trios now removed I investigated whether more stringent filtering might give variant ratios more similar to those reported in individuals with autism. I filtered the variants using a QUAL score of 1000. This resulted in a larger number of variants in the probands than in the untransmitted diplotypes. It also resulted in a number of events per proband to per control ratio, which was closer to that observed in individuals with autism(25)(see table 4-11). However, the number of loss of function homozygous and compound heterozygous variants observed at this QUAL score cut-off was substantially decreased. I therefore concluded that a number of diagnoses were likely removed as a result of this and I decided to continue the analysis with a QUAL score cut-off of 179.

	Rare (≤ 5%) Homozygous or comp het LoF Variants in 1080 probands and 1080 untransmitted diplotypes	Number of complete knock out events per individual in our data	Rare (≤ 5%) Homozygous or comp het LoF Variants in 933 probands versus 869 controls (Lim *et al* 2013)(212)	Number of complete knock out events per individual (Lim *et al* 2013)(212)
Probands	53 + 3 = 56	0.052	62	0.066
Controls	41 + 3 = 44	0.041	29	0.033

Table 4-11: Complete knock out variants at QUAL score 1000.
Number of homozygous and compound heterozygous loss of function variants in 1080 probands and 1080 untransmitted diplotypes using a QUAL score cut off of 1000, compared to the number of complete knock out events observed in individuals with autism and controls reported by Lim *et al(212)*.

4.3.12 Summary of untransmitted diplotype generation method

In summary, I generated a population based control dataset of untransmitted diplotypes using the untransmitted haplotypes from the parents of the affected probands in 1,080 non-consanguineous trios. To prevent generating untransmitted diplotypes with homozygosity by descent, consanguineous families were removed from this analysis. An exome variant profile for the untransmitted diplotypes control was generated for each trio. The trio VCF files (mother, father, child) were merged and the following variants removed: Non 'PASS' variants, INDELS, variants involving a multiallelic reference or alternate allele, CNVs, X and Y chromosome variants, intronic and upstream variants and variants with a QUAL score <179. The genotype in the untransmitted diplotypes was calculated based on the genotypes of the mother, father and proband. Variants that did not fit with Mendelian inheritance were removed. A summary of this method and all the filtering steps for the untransmitted diplotypes generation is shown in figure 4-6.

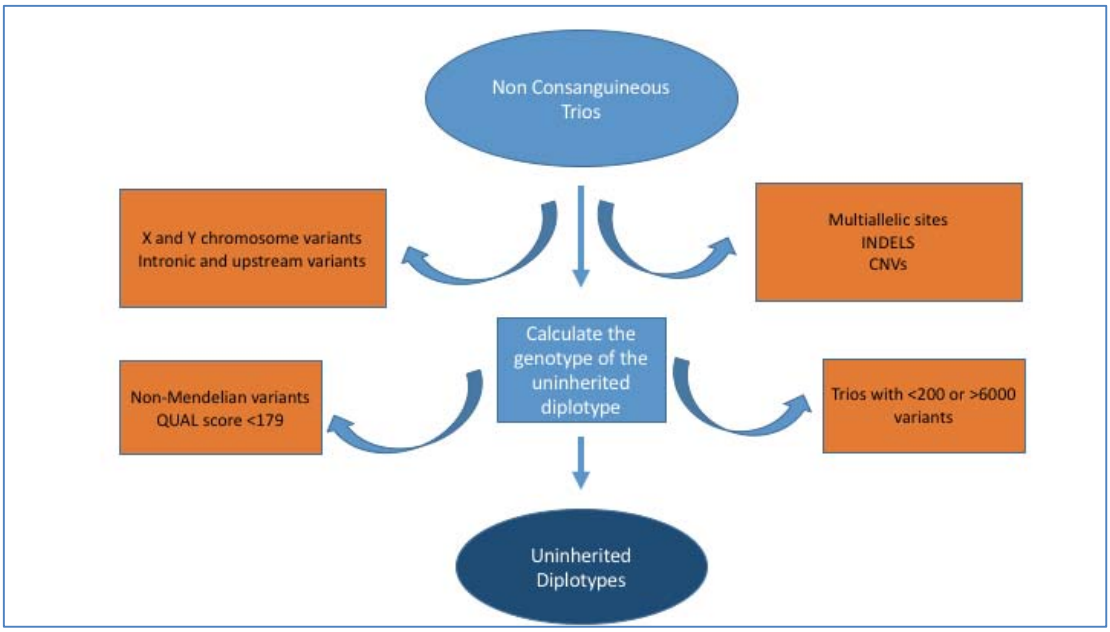


Figure 4-6: Flow diagram showing the processing steps for generating the untransmitted diplotypes.

4.3.13 Outline of burden analyses using untransmitted diplotypes

For consistency, I processed the probands' exome variant profiles used in this analysis in the same way as the untransmitted diplotypes, i.e. they had had specific variant types removed as above.

In order to perform burden analyses, I first compared cumulative counts of rare (MAF < 5%) homozygous and compound heterozygous loss of function and damaging functional variants between the probands and untransmitted diplotypes.

I next identified a specific group of probands likely to have a dominant cause of their developmental disorder, the 'dominant probands'. I hypothesised that removing these 'dominant probands' from the rest of the proband group would have the effect of enriching the remaining group of probands (the non-dominant probands) for recessive developmental disorders if they are present. I concluded I would identify this enrichment by carrying out burden analyses between the 'dominant' and 'non-dominant' probands.

4.4 Results

4.4.1 I identified over transmission of very rare inherited LoF variants to probands

I first compared cumulative counts of rare (MAF < 5%) inherited LoF variants between the probands and untransmitted diplotypes. I identified no observable genome wide trend towards over-transmission to probands for these variants (Table 4-12). I next investigated whether there was an over-transmission of very rare inherited LoF variants (MAF < 0.05%) to probands and showed a genome-wide trend towards over transmission to probands (p=0.015) (Table 4-12). I conclude that this finding gives important evidence that inherited variants are contributing to developmental disorders in this DDD study cohort. I did not observe this over transmission for very rare inherited damaging missense variants (Table 4-12).

	Probands (n = 1080)	**Untransmitted Diplotypes (n = 1080)**
Rare (MAF**<5%) Inherited LoF Variants**	15805	15749
Rare (MAF<5%) Inherited Damaging Functional Variants	98566	98455
Very Rare (MAF**<0.05%) Inherited LoF Variants**	4416	4191
Very Rare (MAF**<0.05%)** Inherited Damaging Functional **Variants**	21965	22044

Table 4-12: Total number of rare and very rare inherited variants observed in probands and untransmitted diplotype controls.

Total number of rare and very rare inherited variants in 1080 children with developmental disorders in comparison to a control dataset of 1080 untransmitted diplotypes. MAF = Minor Allele Frequency. There was an over-transmission of very rare (<0.05%) inherited Loss of Function (LoF) variants (MAF < 0.05%) to probands (p=0.015)**, using the transmission disequilibrium test (McNemar's chi-square)(214)**. There was no observable genome wide trend to over transmission of very rare (<0.05%) Damaging Functional variants to probands or of rare (MAF < 5%) LoF variants to probands.

The over transmission to probands I have identified could be consistent with individuals having a recessive disease, an inherited dominant disease, or an oligogenic disorder. The fact that only by looking at very rare inherited LoF variants (MAF < 0.05%) is there a significant difference between the probands and untransmitted diplotypes suggests that low quality variants may be affecting the results for the less rare variants (MAF < 5%

variants) or that disease resulting from inherited alleles is caused by very rare variants.

4.4.2 Stronger enrichment of bilallelic DDG2P variants than globally

I identified no genome-wide enrichment of rare (<5%) biallelic (compound heterozygous or homozygous) loss of function or missense variants in the probands versus the untransmitted diplotypes. When focusing the analysis on individual genes, there were no genes with significant differences in number of biallelic (compound heterozygous or homozygous) loss of function variants or missense variants. It is likely that low quality variants in both the probands and the untransmitted diplotypes may be preventing an observable difference being identified between probands and untransmitted diplotypes.

I next investigated specifically for enrichment of (<5%) bilallelic variants in the list of 1,142 known Developmental Disorder (DD) genes in the probands. This showed a stronger enrichment of LoF variants than in the genome-wide analysis. Of note, however, the untransmitted diplotypes contained 1 biallelic and 34 monoallelic rare LoF SNVs. This highlights the importance of when interpreting genomes of patients with developmental disorders, not to assume that any damaging variants in known developmental disorder genes are definitely pathogenic.

4.4.3 Depletion of rare biallelic LoF mutations in 'dominant probands'

I next evaluated rare (MAF < 5%) biallelic (homozygous and compound heterozygous) LoF mutations in the dominant probands compared to other probands and showed a 0.56-fold depletion of such variants (p=0.04) in dominant probands (Table 4-13). I identified no enrichment in biallelic damaging missense variants in the other probands compared to the dominant probands, consistent with the findings of Lim *et al* in individuals with autism(212). I conclude that this gives evidence of the presence of recessive disorders in the 'non-dominant' probands in the DDD study.

Biallelic Variant Types	Rate per Untransmitted Diplotype	Rate per Dominant Proband	Rate per Non-Dominant Proband
LoF/LoF (Genome-wide)	0.102	0.063	0.106
LoF/Dam (Genome-wide)	0.081	0.078	0.088
Dam/Dam (Genome-wide)	0.289	0.333	0.326
LoF/LoF (DDG2P Biallelic)	0.001	0.004	0.004
LoF/Dam (DDG2P Biallelic)	0.002	0	0.007
Dam/Dam (DDG2P Biallelic)	0.024	0.026	0.031

Table 4-13: Rate of biallelic loss of function and damaging functional variants.

Rate of rare (MAF < 5%) biallelic loss-of-function and damaging functional variants per untransmitted diplotype, dominant and non-dominant proband. 'dominant probands' refers to probands with a reported de novo mutation or affected parents, and 'other probands' to all remaining probands. 'DDG2P Biallelic' refers to confirmed and probable DDG2P genes with a biallelic mode of inheritance. For untransmitted diplotypes, N=1080, for dominant probands, N=270 and for non-dominant probands N=810.

I next investigated the properties of the dominant probands to see whether this gave any insight into differences they had from the non-dominant probands that may enable more stringent filtering of the untransmitted diplotypes. I investigated the following properties: Variant type, QUAL score, Haplotype score, Readsum score, MQ Ranksum score. However, on visual inspection, I observed no obvious difference between the plotted distributions of these properties between of the probands and the untransmitted diplotypes.

4.5 My findings in context and other contributions to the DDD study

In summary I generated a control dataset of untransmitted diplotypes with which I demonstrated evidence that inherited variants are contributing to developmental disorders in this DDD study cohort. This analysis was carried out at a time when the ExAC database(120), containing large quantities of control data from exome sequencing studies was not available. By studying the probands with likely dominant disorders (dominant probands), I showed that there was a depletion of biallelic loss of function mutations in dominant probands compared to the other probands (non-dominant

probands), this gives evidence for the presence of recessive disorders in individuals with developmental disorders in the DDD study. My findings were a key part of the analysis of the first 1133 trios in the DDD study. Other key findings from the analysis of 1133 trios, were that 12 novel genes associated with developmental disorders were discovered. Together with a multi-disciplinary team of Clinical Geneticists, scientists and bioinformaticians, I reviewed the variants in DDG2P genes flagged for clinical reporting by the bioinformatics pipeline within the DDD study in the 1133 trios in a weekly meeting. We assessed each variant for analytical and clinical validity. For each variant, we compared the patient's phenotypic features, family history and growth parameters to the known phenotype for that gene. When there was sufficient overlap we reported the variant back to the regional genetics service via the patient's local clinician. In total 31% of the 1133 probands and their families received a diagnosis for their disorder. Throughout this process we adjusted the robust bioinformatics pipeline underlying the DDD study, through identifying: problems with reporting, identifying large genes with multiple variants (such as *titin*), or genes that had multiple variants thought to be spurious or sequencing errors. I played an important role in this overall process, contributing my clinical experience and dysmorphology knowledge to help give clinical validity to new pipelines or analyses. I played a significant role in the development of the pipeline but also reporting rules. In addition, I manually reviewed in detail the first 30 *de novo* mutations we reported for clinical validity, and continued to contribute to clinical reporting thoughout my three years on the project.

Further, more recent analyses carried out as part of the DDD study included a case-control analysis looking for evidence of mosaicism in 1303 DDD trios. I played a role in reviewing the mosaic variants and clinical phenotypes in this investigation which identified 12 structural mosaic abnormalities (0.9%) that were reported back to local clinicians. 10 out of 12 of these variants were assessed as highly likely to be pathogenic in causing the individual's developmental disorder. In further analysis of analysis of 4293 trios, the DDD study identified four new genes implicated in recessive diseases and discovered 14 new dominant disease genes. Again I played a key role in reviewing the clinical phenotypes for this investigation and identified families with overlapping phenotypes. Many of the aspects of the DDD study have been incorporated into modern day clinical genetics practice, for example the DDG2P is used in Clinical Genetics

laboratories throughout the UK. Also, multi-disciplinary meetings to review whole exome sequencing findings, as pioneered by the DDD study, form an important part of the week for a number of Clinical Genetics departments.

4.6 Discussion

4.6.1 Summary

In summary I generated a control dataset of untransmitted diplotypes which I used to carry out burden analyses to look for evidence of autosomal recessive disease in individuals with developmental disorders. I carried out multiple filtering steps to generate a dataset of the untransmitted diplotypes, of the correct Mendelian pattern and minimise the number of low quality variants. This novel technique to generate a control database matched for population and sequencing technique and data processing has not to my knowledge been previously attempted. To my knowledge, my work with the untransmitted diplotypes gives the first insight into the contribution of autosomal recessive disease in individuals with developmental disorders by studying untransmitted alleles from exome sequencing data. In addition, my analyses, clinical knowledge and role in clinical reporting contributed significantly to the DDD study, which has shaped modern day clinical genetics knowledge and practice.

4.6.2 Limitations with the untransmitted diplotypes as a control dataset

The theory driving our untransmitted diplotypes control dataset is that individuals inheriting the variants the affected proband didn't inherit (the untransmitted diplotypes) would be predicted to be healthy as if the probands disorder was genetic the disease causing variant(s) would be expected to be within the variants they carry. However, the true phenotypes of this control dataset are not known and never will be. Therefore, it cannot be ruled out that the untransmitted diplotypes carry lethal variants that would result in foetal demise or a severe developmental disorder. Also our analysis doesn't account for the possibility of non-penetrance of a variant in the parents, or disorders resulting from environmental exposures or disease in the mother, or other non-genetic causes of developmental disorder in the probands.

In addition, we removed variants on the X and Y chromosomes, INDELS and variants with multiallelic ALTs or REFs from the untransmitted diplotypes. Therefore, our control dataset is not a complete representation of the exonic variation of the untransmitted alleles.

Furthermore, despite the multiple filtering steps I carried out, the untransmitted diplotypes are still likely to be enriched with false positive variants from their parents. Also filtering the probands may have removed diagnostic variants. One way I could improve this in the future would be to carry out joint variant calling on the raw sequencing data used in this investigation with that of other studies using next generation sequencing methods. 'Joint calling' methods have been shown to successfully separate out true variation from machine artifacts which are common to next generation sequencing technologies while preserving true variant sites(215)52). Implementing joint calling on my dataset, may therefore remove some of the false-positive variants in the untransmitted diplotypes.

One alternative to using the untransmitted diplotypes as controls would be to use true siblings as controls. This would overcome the problem of not knowing the untransmitted diplotypes phenotypes and also the increased number of false-positive mutations observed in the untransmitted diplotypes.

4.6.3 Our findings in context

Other studies using untransmitted alleles

Untransmitted alleles have previously been investigated in individuals with diabetes using the Transmission Test for Linkage Disequilibrium (TDT test)(214). As a test for linkage disequilibrium Spielman *et al* considered a heterozygous allele associated with disease in an affected parent and evaluated the frequency with which this allele or its alternate was passed to an affected offspring. Although theses authors also studied transmitted and untransmitted alleles, the authors only studied single alleles and didn't look at the untransmitted alleles in the context of the other untransmitted allele at the

same loci, i.e. they looked at all of the untransmitted alleles in aggregate from affected parents and they didn't pair up corresponding alleles to investigate real possible recessive combinations of alleles within families. The untransmitted diplotypes dataset is therefore to our knowledge a unique control dataset which comprises real combinations of recessive variants within families.

4.6.4 Using burden analysis to detect oligogenic inheritance

There is evidence that two hit aetiologies and oligogenic models of inheritance exist in developmental disorders and that these events are most likely to be distributed over many genes(208, 216-218). Here we have shown that burden analyses give insight into the underlying genetic architecture of developmental disorders. In the future similar methods could be used to investigate for evidence of oligogenic inheritance in individuals with developmental disorders. One way to approach this would be following assembly of a large cohort of individuals with developmental disorders to remove individuals with known monogenic diseases to leave a group that is likely to be enriched with oligogenic developmental disorders. The number and types of variants and their inheritance could then be compared between the undiagnosed group and both the diagnosed group and a control dataset. Real siblings could also play an important role in these types of analyses.

4.6.5 The future of untangling the aetiology of developmental disorders

Understanding the architecture of developmental disorders is important now as we are in an era of mass gene discovery, but will be more so in the future when we are reaching saturation of Mendelian gene discovery, as we work out how many of the remaining developmental genetic disorders have a genetic cause. Successful future dominant and recessive gene discovery requires larger datasets with international collaborations likely playing a role in this. Full and accurate sharing of standardised phenotypic data is highly likely to be needed to help facilitate gene discovery. Isolated populations / consanguineous unions may continue to help in these efforts to uncover recessive diseases. So may the use of studying real siblings and incorporating analyses of the epigenome. Further understanding of the phenotypic spectrum of genetic diseases,

reasons for disease variability and reduced penetrance will help us understand which individuals have more than genetic disorder as composite phenotypes will continue to challenge Clinical Geneticists in years to come. Clinical interpretation of variants will be crucial to the dissection of developmental disorders in the future.

Conclusions

In conclusion, I generated a control dataset of untransmitted diplotypes which I used to carry out burden analyses to look for evidence of autosomal recessive disease in individuals with developmental disorders. To my knowledge, my work with the untransmitted diplotypes gives the first insight into the contribution of autosomal recessive disease in individuals with developmental disorders by studying untransmitted alleles from exome sequencing data. In addition, my analyses, clinical knowledge and role in clinical reporting contributed significantly to the DDD study, which has shaped modern day clinical genetics knowledge and practice.

Successful future gene discovery in developmental disorders requires larger datasets with international collaborations likely playing a role in this. Full and accurate sharing of standardised phenotypic data is essential and clinical interpretation of variants identified through genome wide sequencing techniques will be crucial to the dissection of developmental disorders in the future.

Chapter 5

Discussion

In this dissertation, I have described three projects that take genetic or phenotypic approaches to understanding developmental disorders using data from next generation sequencing. In Chapter 2, I investigated the phenotype of Wiedemann-Steiner syndrome (WSS) resulting from *KMT2A* mutations, and aimed to define the phenotypic and mutational spectrum. I identified, collected and analysed standardised data from 84 individuals with WSS and *KMT2A* mutations. To my knowledge this is the largest cohort reported in the world to date. I defined the *KMT2A* mutational spectrum and provided the first detailed evaluation of the features associated with *KMT2A*-associated WSS in a cohort more than 15 times larger than the largest previous report (N=5 individuals). I defined the growth pattern and demonstrated that not all individuals with WSS have hypertrichosis. I reported new features of WSS (including genital abnormalities in females) and highlighted that epilepsy is as an important feature of WSS that is associated with poorer developmental outcomes in WSS individuals. I reported the first somatic mosaic individual with a *KMT2A* mutation with a milder clinical phenotype.

My next aim was to investigate how missense mutations affect KMT2A function. I showed that missense mutations cluster within the recognized domains of KMT2A, including the zinc-finger and zinc-binding domains and proposed disease mechanisms for these mutations, including preventing DNA binding. Finally, I aimed to investigate whether WSS caused by *KMT2A* mutations has a recognisable facial appearance. I showed that the facial appearance of individuals with WSS is distinguishable from that of other individuals with developmental disorders by carrying out a facial recognition experiment with experienced Genetics Clinicians. This investigation has significantly advanced the knowledge and understanding of WSS caused by *KMT2A* mutations. The knowledge gained will help clinicians identify and manage individuals with WSS in the future and ultimately improve the medical care of these individuals, who have multiple medical needs.

In Chapter 3, my first aim was to investigate the genetic basis of developmental disorders associated with hypertrichosis using whole exome sequencing. I identified 247 individuals with developmental disorders including hypertrichosis or with WSS or a condition similar to WSS and analysed their exome variant profiles. The *de novo* diagnostic yield from my cohort was 29%, which is higher than the *de novo* diagnostic

yield of the DDD study more generally of 23%. However, a significant proportion of this yield was *KMT2A* mutations and removing those individuals gave a diagnostic yield of 21%. My next aim was to seek evidence of a burden of variants in genes that play a role in maintaining the structure or function of chromatin (chromatin genes). I showed that the DDD hypertrichosis cohort was significantly enriched for variants in chromatin genes. My findings suggest that hypertrichosis is an important signal that an individual could have a chromatin disorder and may be more likely to harbour a diagnostic *de novo* mutation than individuals with developmental disorders more generally. This investigation also highlighted known disease genes implicated in hypertrichosis, which could enable gene panels (for phenotype-focused next generation sequencing) to be curated for individuals with hypertrichosis.

I next sought to identify new genes implicated in developmental disorders associated with hypertrichosis. I identified two individuals with identical missense mutations in *ZMYND11*, suggesting that *ZMYND11,* a well-recognized developmental disorder gene, is specifically associated with hypertrichosis. Although I didn't identify any new genes, I showed in principle this approach can successfully identify disease genes associated with hypertrichosis and will pave the way for further research into the genetic architecture of hypertrichosis using larger cohort sizes in the future.

In Chapter 4, I aimed to investigate the underlying architecture of severe developmental disorders by seeking out evidence of autosomal recessive disease using a population matched control dataset. I generated a novel control dataset of untransmitted diplotypes and analysed 1,080 non-consanguineous trios with developmental disorders in the DDD Study. By the use of the untransmitted diplotypes, I showed that there is a genome wide trend towards over transmission of very rare (MAF < 0.05%) LoF variants to DDD probands, giving evidence that inherited variants contribute to developmental disorders in the DDD study cohort. In addition, by separating out the individuals with a likely dominant cause of their disorder (dominant probands) I showed an enrichment of rare (MAF <5%) biallelic loss of function (LOF) variants in known developmental disorder genes in non-dominant probands compared to dominant probands, providing evidence of recessive disease in the non-dominant probands. To my knowledge, my work using the

untransmitted diplotypes gives the first insight into the contribution of autosomal recessive disease to developmental disorders by studying untransmitted alleles.

My final aim was to contribute to the significant improvement of the diagnosis of children with developmental disorders as a clinician researcher working as a member of the DDD study analysis team. My analyses, clinical knowledge and role in clinical reporting contributed significantly to the DDD study, which has shaped modern day clinical genetics knowledge and practice. Through analysis of 1,133 trios, 31% of probands and their families received a diagnosis for their disorder, and 12 novel genes associated with developmental disorders were discovered. A case-control analysis looking for evidence of mosaicism in 1303 DDD trios, identified 12 structural mosaic abnormalities (0.9%) that were reported back to local clinicians, 10 of which were assessed as highly likely to be pathogenic in causing the individual's developmental disorder. In further analysis of analysis of 4293 trios, the DDD study identified four new genes implicated in recessive diseases and discovered 14 new dominant disease genes. Many of the aspects of the DDD study have been incorporated into modern day clinical genetics practice; for example, the DDG2P is used in Clinical Genetics laboratories throughout the UK. Also, multi-disciplinary meetings to review whole exome sequencing findings, as pioneered by the DDD study, form an important part of the week for a number of Clinical Genetics departments.

The three projects I have presented in this dissertation have four important outcomes that increase knowledge of developmental disorders and will ultimately help individuals with rare diseases and their family members in the future. 1. I have significantly increased knowledge of the phenotypic and mutational spectrum of the rare disorder Wiedemann-Steiner syndrome. This will ultimately improve patient identification, diagnosis and medical care. 2. I have demonstrated that there is a burden of variants in chromatin genes in individuals with hypertrichosis. This suggests that hypertrichosis is an important indicator that an individual could have a chromatin disorder and may be more likely to harbour *de novo* mutation than other individuals with developmental disorders. This knowledge and knowledge of the genes implicated will help in the diagnosis of individuals with hypertrichosis in the future, and also will enable gene panels (for selective next generation sequencing) to be curated for individuals with

hypertrichosis. 3. My work using the untransmitted diplotypes has increased knowledge of the architecture of developmental disorders through studying non-transmitted alleles. I showed that inherited variants are contributing to developmental disorders in the DDD study cohort. Additionally, I found evidence for recessive disease in DDD study individuals by identifying a burden of biallelic loss of function variants in DDD non-dominant probands. 4. I have played a key role as an analyst and clinician researcher in the DDD study which has shaped modern day clinical genetics knowledge and practice.

In conclusion, I have described three investigations that take genetic or phenotypic approaches to understanding developmental disorders using data from next generation sequencing. The themes running throughout this dissertation are dominant versus recessive inheritance, loss of function versus missense variants, the use of next generation sequencing to unravel the underlying causes of developmental disorders and challenges in assigning pathogenicity to variants. Many of these themes are current key issues of Clinical Genetics more widely in the whole exome sequencing era. In the future, further understanding of the architecture, genotypes and phenotypes of developmental disorders will be driven by larger sample sizes, standardisation of phenotype terms, online portals to input and share genotype and phenotype data and large population control databases. Variant interpretation will remain a challenge for many years to come, particularly the understanding of missense variants. Ideally our mutation and control databases would together cover every variant or there would be a valid and reliable functional assay for every disease. However, until this point, a sensible and practical approach to managing patients with variants of uncertain significance is vital and assessment for dysmorphological features will continue to play an important role in variant interpretation. Though understanding more about developmental disorders, we are in a stronger position to manage patients more effectively and develop treatments, ultimately helping individuals with rare diseases and their families.

References

1. The Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519(7542):223-8.
2. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305-14.
3. Al Turki S, Manickaraj AK, Mercer CL, Gerety SS, Hitz MP, Lindsay S, et al. Rare variants in NR2F2 cause congenital heart defects in humans. American journal of human genetics. 2014;94(4):574-85.
4. Association AP. Desk Reference to the Diagnostic Criteria From DSM-IV-TR. Washington2000.
5. Topper S, Ober C, Das S. Exome sequencing and the genetics of intellectual disability. Clinical genetics. 2011;80(2):117-26.
6. Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Human mutation. 2010;31(6):631-55.
7. EURODIS. Survey of the delay in diagnosis for 8 rare diseases in Europe ('EurordisCare2'). 2007
.
8. Commission of the European Communities. Communication from the commision to the European Parliament, the council, the European economic and social committee and the committee of the regions. On Rare Diseases: Europe's challenges. 2008.
9. European Commission. Rare Diseases [Available from: https://ec.europa.eu/health/rare_diseases/policy_en.
10. Orphanet. Prevalence of rare diseases: Bibliographic data. Orphanet Reports Series 2014; Rare Diseases collection.
11. Stewart DL, Hersh JH. The impact of major congenital malformations on mortality in a neonatal intensive care unit. J Ky Med Assoc. 1995;93(8):329-32.
12. Synnes AR, Berry M, Jones H, Pendray M, Stewart S, Lee SK, et al. Infants with congenital anomalies admitted to neonatal intensive care units. Am J Perinatol. 2004;21(4):199-207.
13. Scriver CR, Neal JL, Saginur R, Clow A. The frequency of genetic disease and congenital malformation among patients in a pediatric hospital. Can Med Assoc J. 1973;108(9):1111-5.
14. Polder JJ, Meerding WJ, Bonneux L, van der Maas PJ. Healthcare costs of intellectual disability in the Netherlands: a cost-of-illness perspective. J Intellect Disabil Res. 2002;46(Pt 2):168-78.
15. Centers for Disease C, Prevention. Economic costs associated with mental retardation, cerebral palsy, hearing loss, and vision impairment--United States, 2003. MMWR Morb Mortal Wkly Rep. 2004;53(3):57-9.
16. Arvio M, Salokivi T, Tiitinen A, Haataja L. Mortality in individuals with intellectual disabilities in Finland. Brain Behav. 2016;6(2):e00431.
17. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nature genetics. 2004;36(9):949-51.
18. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
19. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. American journal of human genetics. 2009;84(2):148-61.

20. Lejeune J, Gauthier M, Turpin R. [Human chromosomes in tissue cultures]. Comptes rendus hebdomadaires des seances de l'Academie des sciences. 1959;248(4):602-3.
21. Tjio HJL, A. The chromosome numbers of man. Hereditas. 1956;42:1-6.
22. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14(5):307-20.
23. Hirschhorn K, Decker WH, Cooper HL. Human intersex with chromosome mosaicism of type XY/XO. Report of a case. N Engl J Med. 1960;263:1044-8.
24. Lamb J, Wilkie AO, Harris PC, Buckle VJ, Lindenbaum RH, Barton NJ, et al. Detection of breakpoints in submicroscopic chromosomal translocation, illustrating an important mechanism for genetic disease. Lancet. 1989;2(8667):819-24.
25. Rudd MK. Structural variation in subtelomeres. Methods Mol Biol. 2012;838:137-49.
26. Ravnan JB, Tepperberg JH, Papenhausen P, Lamb AN, Hedrick J, Eash D, et al. Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. Journal of medical genetics. 2006;43(6):478-89.
27. Knight SJ, Regan R, Nicod A, Horsley SW, Kearney L, Homfray T, et al. Subtle chromosomal rearrangements in children with unexplained mental retardation. Lancet. 1999;354(9191):1676-81.
28. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, et al. High-resolution human genome structure by single-molecule analysis. Proceedings of the National Academy of Sciences. 2010;107(24):10848-53.
29. Ray M, Goldstein S, Zhou S, Potamousis K, Sarkar D, Newton MA, et al. Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis. BMC Genomics. 2013;14(1):505.
30. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by Optical Mapping. Science. 1993;262.
31. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. The American Journal of Human Genetics. 2010;86(5):749-64.
32. Aradhya S, Lewis R, Bonaga T, Nwokekeh N, Stafford A, Boggs B, et al. Exon-level array CGH in a large clinical cohort demonstrates increased sensitivity of diagnostic testing for Mendelian disorders. Genet Med. 2012;14(6):594-603.
33. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
34. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230(4732):1350-4.
35. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in enzymology. 1987;155:335-50.
36. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991;65(5):905-14.
37. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nature genetics. 1999;23(2):185-8.
38. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.

39. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6.
40. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(45):19096-101.
41. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nature genetics. 2010;42(1):30-5.
42. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin M, Gildersleeve H, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics. 2010;42(9):790-3.
43. Clayton-Smith J, O'Sullivan J, Daly S, Bhaskar S, Day R, Anderson B, et al. Whole-exome-sequencing identifies mutations in histone acetyltransferase gene KAT6B in individuals with the Say-Barber-Biesecker variant of Ohdo syndrome. American journal of human genetics. 2011;89(5):675-81.
44. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nature genetics. 2012;44(4):376-8.
45. Santen GW, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, et al. Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nature genetics. 2012;44(4):379-80.
46. Jones WD, Dafou D, McEntagart M, Woollard WJ, Elmslie FV, Holder-Espinasse M, et al. De novo mutations in MLL cause Wiedemann-Steiner syndrome. American journal of human genetics. 2012;91(2):358-64.
47. Sawyer SL, Hartley T, Dyment DA, Beaulieu CL, Schwartzentruber J, Smith A, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clinical genetics. 2015.
48. Firth HV, Wright CF. The Deciphering Developmental Disorders (DDD) study. Developmental medicine and child neurology. 2011;53(8):702-3.
49. Winter R, Baraitser, M. London Dysmorphology Database [Available from: http://www.lmdatabases.com/.
50. POSSUM [Available from: www.possum.net.au.
51. Hennekam RCM, Cormier-Daire V, Hall JG, Méhes K, Patton M, Stevenson RE. Elements of morphology: Standard terminology for the nose and philtrum. American Journal of Medical Genetics Part A. 2009;149A(1):61-76.
52. Allanson JE, Cunniff C, Hoyme HE, McGaughran J, Muenke M, Neri G. Elements of Morphology: Standard Terminology for the Head and Face. American journal of medical genetics Part A. 2009;149A(1):6-28.
53. Biesecker LG, Aase JM, Clericuzio C, Gurrieri F, Temple IK, Toriello H. Defining Morphology: Hands and Feet. American journal of medical genetics Part A. 2009;149A(1):93-127.
54. Carey JC, Cohen MM, Curry CJR, Devriendt K, Holmes LB, Verloes A. Elements of morphology: Standard terminology for the lips, mouth, and oral region. American Journal of Medical Genetics Part A. 2009;149A(1):77-92.
55. Hall BD, Graham JM, Cassidy SB, Opitz JM. Elements of morphology: Standard terminology for the periorbital region. American Journal of Medical Genetics Part A. 2009;149A(1):29-39.
56. Hunter A, Frias JL, Gillessen-Kaesbach G, Hughes H, Jones KL, Wilson L. Elements of morphology: Standard terminology for the ear. American Journal of Medical Genetics Part A. 2009;149A(1):40-60.

57. The American Journal of Human Genetics: Instructions for authors [Available from: http://www.cell.com/ajhg/authors.
58. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research. 2014;42(Database issue):D966-74.
59. Graungaard AH, Skov L. Why do we need a diagnosis? A qualitative study of parents' experiences, coping and needs, when the newborn child is severely disabled. Child: Care, Health and Development. 2007;33(3):296-307.
60. Carmichael B, Pembrey M, Turner G, Barnicoat A. Diagnosis of fragile-X syndrome: the experiences of parents. J Intellect Disabil Res. 1999;43 (Pt 1):47-53.
61. Goker-Alpan O, Schiffmann R, LaMarca ME, Nussbaum RL, McInerney-Leo A, Sidransky E. Parkinsonism among Gaucher disease carriers. Journal of medical genetics. 2004;41(12):937-40.
62. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. New England Journal of Medicine. 2009;361(17):1651-61.
63. Siebert M, Sidransky E, Westbroek W. Glucocerebrosidase is shaking up the synucleinopathies. Brain : a journal of neurology. 2014;137(Pt 5):1304-22.
64. Mitsui J, Matsukawa T, Sasaki H, Yabe I, Matsushima M, Dürr A, et al. Variants associated with Gaucher disease in multiple system atrophy. Annals of clinical and translational neurology. 2015;2(4):417-26.
65. Hurst JA, Baraitser M, Auger E, Graham F, Norell S. An extended family with a dominantly inherited speech disorder. Developmental medicine and child neurology. 1990;32(4):352-5.
66. Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME. Localisation of a gene implicated in a severe speech and language disorder. Nature genetics. 1998;18(2):168-70.
67. Lai CS, Fisher SE, Hurst JA, Levy ER, Hodgson S, Fox M, et al. The SPCH1 region on human 7q31: genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. American journal of human genetics. 2000;67(2):357-68.
68. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 2001;413(6855):519-23.
69. Haesler S, Rochefort C, Georgi B, Licznerski P, Osten P, Scharff C. Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. PLoS biology. 2007;5(12):e321.
70. Wiedemann H-R, Kunze, J, Dibbern H. Atlas der klinischen Syndrome fur Klinik and Praxis. 3rd ed. Stuttgart: Schattauer; 1989. p. 198-9.
71. Steiner CE, Marques AP. Growth deficiency, mental retardation and unusual facies. Clin Dysmorphol. 2000;9(2):155-6.
72. Koenig R, Meinecke P, Kuechler A, Schafer D, Muller D. Wiedemann-Steiner syndrome: three further cases. American journal of medical genetics Part A. 2010;152A(9):2372-5.
73. Mendelsohn BA, Pronold M, Long R, Smaoui N, Slavotinek AM. Advanced bone age in a girl with Wiedemann-Steiner syndrome and an exonic deletion in KMT2A (MLL). American journal of medical genetics Part A. 2014.
74. Miyake N, Tsurusaki Y, Koshimizu E, Okamoto N, Kosho T, Brown NJ, et al. Delineation of clinical features in Wiedemann-Steiner syndrome caused by KMT2A mutations. Clinical genetics. 2015.

75. Stellacci E, Onesimo R, Bruselles A, Pizzi S, Battaglia D, Leoni C, et al. Congenital immunodeficiency in an individual with Wiedemann-Steiner syndrome due to a novel missense mutation in KMT2A. American journal of medical genetics Part A. 2016.
76. Strom SP, Lozano R, Lee H, Dorrani N, Mann J, PF OL, et al. De Novo variants in the KMT2A (MLL) gene causing atypical Wiedemann-Steiner syndrome in two unrelated individuals identified by clinical exome sequencing. BMC medical genetics. 2014;15(1):49.
77. Yuan B, Pehlivan D, Karaca E, Patel N, Charng W-L, Gambin T, et al. Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. The Journal of clinical investigation. 2015;125(2):636-51.
78. Dunkerton S, Field M, Cho V, Bertram E, Whittle B, Groves A, et al. A de novo mutation in KMT2A (MLL) in monozygotic twins with Wiedemann-Steiner syndrome. American journal of medical genetics Part A. 2015.
79. Bramswig NC, Lüdecke H-J, Alanay Y, Albrecht B, Barthelmie A, Boduroglu K, et al. Exome sequencing unravels unexpected differential diagnoses in individuals with the tentative diagnosis of Coffin–Siris and Nicolaides–Baraitser syndromes. Human genetics. 2015;134(6):553-68.
80. Steel D, Salpietro V, Phadke R, Pitt M, Gentile G, Massoud A, et al. Whole exome sequencing reveals a MLL de novo mutation associated with mild developmental delay and without 'hairy elbows': expanding the phenotype of Wiedemann-Steiner syndrome. J Genet. 2015;94(4):755-8.
81. Calvel P, Kusz-Zamelczyk K, Makrythanasis P, Janecki D, Borel C, Conne B, et al. A Case of Wiedemann-Steiner Syndrome Associated with a 46,XY Disorder of Sexual Development and Gonadal Dysgenesis. Sex Dev. 2015;9(5):289-95.
82. Bogershausen N, Bruford E, Wollnik B. Skirting the pitfalls: a clear-cut nomenclature for H3K4 methyltransferases. Clinical genetics. 2013;83(3):212-4.
83. Helbig KL, Farwell Hagman KD, Shinde DN, Mroske C, Powis Z, Li S, et al. Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. Genet Med. 2016.
84. Euro E-RESC, Epilepsy Phenome/Genome P, Epi KC. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. American journal of human genetics. 2014;95(4):360-70.
85. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Ercument Cicek A, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-15.
86. Lederer D, Grisart B, Digilio Maria C, Benoit V, Crespin M, Ghariani Sophie C, et al. Deletion of KDM6A, a Histone Demethylase Interacting with MLL2, in Three Patients with Kabuki Syndrome. The American Journal of Human Genetics. 2012;90(1):119-24.
87. Issaeva I, Zonis Y, Rozovskaia T, Orlovsky K, Croce CM, Nakamura T, et al. Knockdown of ALR (MLL2) Reveals ALR Target Genes and Leads to Alterations in Cell Adhesion and Growth. Molecular and cellular biology. 2007;27(5):1889-903.
88. Agger K, Cloos PAC, Christensen J, Pasini D, Rose S, Rappsilber J, et al. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature. 2007;449(7163):731-4.
89. Fahrner JA, Bjornsson HT. Mendelian Disorders of the Epigenetic Machinery: Tipping the Balance of Chromatin States. Annual review of genomics and human genetics. 2014;15:269-93.
90. Tonkin ET, Wang T-J, Lisgo S, Bamshad MJ, Strachan T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. Nature genetics. 2004;36(6):636-41.

91. Deardorff MA, Bando M, Nakato R, Watrin E, Itoh T, Minamino M, et al. HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. Nature. 2012;489(7415):313-7.
92. Deardorff MA, Kaur M, Yaeger D, Rampuria A, Korolev S, Pie J, et al. Mutations in Cohesin Complex Members SMC3 and SMC1A Cause a Mild Variant of Cornelia de Lange Syndrome with Predominant Mental Retardation. The American Journal of Human Genetics.80(3):485-94.
93. Musio A, Selicorni A, Focarelli ML, Gervasini C, Milani D, Russo S, et al. X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations. Nature genetics. 2006;38(5):528-30.
94. Deardorff MA, Wilde JJ, Albrecht M, Dickinson E, Tennstedt S, Braunholz D, et al. RAD21 mutations cause a human cohesinopathy. American journal of human genetics. 2012;90(6):1014-27.
95. Butler LH, Slany R, Cui X, Cleary ML, Mason DY. The HRX proto-oncogene product is widely expressed in human tissues and localizes to nuclear structures. Blood. 1997;89(9):3361-70.
96. Rasio D, Schichman SA, Negrini M, Canaani E, Croce CM. Complete Exon Structure of the ALL1 Gene. Cancer Research. 1996;56(8):1766-9.
97. Schuettengruber B, Martinez AM, Iovino N, Cavalli G. Trithorax group proteins: switching genes on and keeping them active. Nat Rev Mol Cell Biol. 2011;12(12):799-814.
98. Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nat Rev Cancer. 2015;15(6):334-46.
99. Milne TA, Briggs SD, Brock HW, Martin ME, Gibbs D, Allis CD, et al. MLL targets SET domain methyltransferase activity to Hox gene promoters. Molecular cell. 2002;10(5):1107-17.
100. Cosgrove MS, Patel A. Mixed lineage leukemia: a structure-function perspective of the MLL1 protein. The FEBS journal. 2010;277(8):1832-42.
101. Cosgrove MS, Patel A. Mixed lineage leukemia: a structure-function perspective of the MLL1 protein. The FEBS journal. 2010;277(8):1832-42.
102. Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, et al. Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. EMBO J. 2006;25(19):4503-12.
103. Hsieh JJ, Ernst P, Erdjument-Bromage H, Tempst P, Korsmeyer SJ. Proteolytic cleavage of MLL generates a complex of N- and C-terminal fragments that confers protein stability and subnuclear localization. Molecular and cellular biology. 2003;23(1):186-94.
104. Down JL. Observations on an ethnic classification of idiots. 1866. Ment Retard. 1995;33(1):54-6.
105. Lin HC, Le Hoang P, Hutchinson A, Chao G, Gerfen J, Loomes KM, et al. Alagille Syndrome in a Vietnamese Cohort: Mutation Analysis and Assessment of Facial Features. American journal of medical genetics Part A. 2012;158A(5):1005-13.
106. Kamath BM, Loomes KM, Oakey RJ, Emerick KE, Conversano T, Spinner NB, et al. Facial features in Alagille syndrome: specific or cholestasis facies? Am J Med Genet. 2002;112(2):163-70.
107. Rohatgi S, Clark D, Kline AD, Jackson LG, Pie J, Siu V, et al. Facial Diagnosis of Mild and Variant CdLS: Insights From a Dysmorphologist Survey. American journal of medical genetics Part A. 2010;0(7):1641-53.
108. Sokol RJ, Heubi JE, Balistreri WF. Intrahepatic "cholestasis facies": is it specific for Alagille syndrome? J Pediatr. 1983;103(2):205-8.
109. Leisti J, Hollister DW, Rimoin DL. The Floating-Harbor syndrome. Birth Defects Orig Artic Ser. 1975;11(5):305.

110. Pelletier G, Feingold, M. . Case report 1. . Syndrome Ident 1973;1:8-9.
111. Robinson PL, Shohat M, Winter RM, Conte WJ, Gordon-Nesbitt D, Feingold M, et al. A unique association of short stature, dysmorphic features, and speech impairment (Floating-Harbor syndrome). J Pediatr. 1988;113(4):703-6.
112. Basel-Vanagaite L, Wolf L, Orin M, Larizza L, Gervasini C, Krantz ID, et al. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. Clinical genetics. 2016;89(5):557-63.
113. Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. Elife. 2014;3:e02020.
114. Hammond P, Hutton TJ, Allanson JE, Buxton B, Campbell LE, Clayton-Smith J, et al. Discriminating power of localized three-dimensional facial morphology. American journal of human genetics. 2005;77(6):999-1010.
115. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human mutation. 2016;37(6):564-9.
116. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic acids research. 2016;44(D1):D710-D6.
117. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Human mutation. 2008;29(1):6-13.
118. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American journal of human genetics. 2009;84(4):524-33.
119. Epi KC, Epilepsy Phenome/Genome P, Allen AS, Berkovic SF, Cossette P, Delanty N, et al. De novo mutations in epileptic encephalopathies. Nature. 2013;501(7466):217-21.
120. (ExAC) EAC. ExAC Dataset Cambridge, MA (URL: http://exac.broadinstitute.org)2015/ [(03):[
121. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nature genetics. 2014;46(9):944-50.
122. Freeman JV, Cole TJ, Chinn S, Jones PR, White EM, Preece MA. Cross sectional stature and weight reference curves for the UK, 1990. Archives of Disease in Childhood. 1995;73(1):17-24.
123. Chu-Shore CJ, Major P, Camposano S, Muzykewicz D, Thiele EA. The natural history of epilepsy in tuberous sclerosis complex. Epilepsia. 2010;51(7):1236-41.
124. Poisson A, Nicolas A, Cochat P, Sanlaville D, Rigard C, de Leersnyder H, et al. Behavioral disturbance and treatment strategies in Smith-Magenis syndrome. Orphanet J Rare Dis. 2015;10:111.
125. Aguilar-Arnal L, Hakim O, Patel VR, Baldi P, Hager GL, Sassone-Corsi P. Cycles in spatial and temporal chromosomal organization driven by the circadian clock. Nat Struct Mol Biol. 2013;20(10):1206-13.
126. Katada S, Sassone-Corsi P. The histone methyltransferase MLL1 permits the oscillation of circadian gene expression. Nat Struct Mol Biol. 2010;17(12):1414-21.
127. Hoffman JD, Ciprero KL, Sullivan KE, Kaplan PB, McDonald-McGinn DM, Zackai EH, et al. Immune abnormalities are a frequent manifestation of Kabuki syndrome. American journal of medical genetics Part A. 2005;135(3):278-81.
128. Argentaro A, Yang J-C, Chapman L, Kowalczyk MS, Gibbons RJ, Higgs DR, et al. Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD)

domain of the chromatin-associated protein ATRX. Proceedings of the National Academy of Sciences. 2007;104(29):11939-44.
129. Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, et al. Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. The EMBO Journal. 2006;25(19):4503-12.
130. EuroEPINOMIC-RES Consortium, Epilepsy Phenome/Genome Project, Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. Am J Hum Genet. 2014;95(4):360-70.
131. Wang J, Leung JW, Gong Z, Feng L, Shi X, Chen J. PHF6 regulates cell cycle progression by suppressing ribosomal RNA synthesis. J Biol Chem. 2013;288(5):3174-83.
132. Menke LA, van Belzen MJ, Alders M, Cristofoli F, Study DDD, Ehmke N, et al. CREBBP mutations in individuals without Rubinstein-Taybi syndrome phenotype. American journal of medical genetics Part A. 2016.
133. Park U-H, Yoon SK, Park T, Kim E-J, Um S-J. Additional Sex Comb-like (ASXL) Proteins 1 and 2 Play Opposite Roles in Adipogenesis via Reciprocal Regulation of Peroxisome Proliferator-activated Receptor γ. The Journal of biological chemistry. 2011;286(2):1354-63.
134. Kaikkonen MU, Spann N, Heinz S, Romanoski CE, Allison KA, Stender JD, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular cell. 2013;51(3):310-25.
135. Guenther MG, Jenner RG, Chevalier B, Nakamura T, Croce CM, Canaani E, et al. Global and Hox-specific roles for the MLL1 methyltransferase. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(24):8603-8.
136. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nature genetics. 2012;44(8):886-9.
137. Harakalova M, van Harssel JJ, Terhal PA, van Lieshout S, Duran K, Renkens I, et al. Dominant missense mutations in ABCC9 cause Cantu syndrome. Nature genetics. 2012;44(7):793-6.
138. van Bon BW, Gilissen C, Grange DK, Hennekam RC, Kayserili H, Engels H, et al. Cantu syndrome is caused by mutations in ABCC9. American journal of human genetics. 2012;90(6):1094-101.
139. De Raeve L, Keymolen K. Congenital hypertrichosis lanuginosa in a father and son. Arch Dermatol. 2011;147(6):746-7.
140. Sun M, Li N, Dong W, Chen Z, Liu Q, Xu Y, et al. Copy-number mutations on chromosome 17q24.2-q24.3 in congenital generalized hypertrichosis terminalis with or without gingival hyperplasia. American journal of human genetics. 2009;84(6):807-13.
141. Chen W, Ring J, Happle R. Congenital generalized hypertrichosis terminalis: a proposed classification and a plea to avoid the ambiguous term "Ambras syndrome". European journal of dermatology : EJD. 2015.
142. Petrij F, Giles RH, Dauwerse HG, Saris JJ, Hennekam RC, Masuno M, et al. Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. Nature. 1995;376(6538):348-51.
143. Vidal M, Cusick Michael E, Barabási A-L. Interactome Networks and Human Disease. Cell.144(6):986-98.
144. Nakamura T, Mori T, Tada S, Krajewski W, Rozovskaia T, Wassell R, et al. ALL-1 Is a Histone Methyltransferase that Assembles a Supercomplex of Proteins Involved in Transcriptional Regulation. Molecular cell.10(5):1119-28.

145. Hughes CM, Rozenblatt-Rosen O, Milne TA, Copeland TD, Levine SS, Lee JC, et al. Menin Associates with a Trithorax Family Histone Methyltransferase Complex and with the Hoxc8 Locus. Molecular cell.13(4):587-97.
146. Cho Y-W, Hong T, Hong S, Guo H, Yu H, Kim D, et al. PTIP Associates with MLL3- and MLL4-containing Histone H3 Lysine 4 Methyltransferase Complex. The Journal of biological chemistry. 2007;282(28):20395-406.
147. Lee JH, Skalnik DG. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. The Journal of biological chemistry. 2005;280(50):41725-31.
148. Lee JH, Tate CM, You JS, Skalnik DG. Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. The Journal of biological chemistry. 2007;282(18):13419-28.
149. Dou Y, Milne TA, Ruthenburg AJ, Lee S, Lee JW, Verdine GL, et al. Regulation of MLL1 H3K4 methyltransferase activity by its core components. Nat Struct Mol Biol. 2006;13(8):713-9.
150. Ernst P, Wang J, Huang M, Goodman RH, Korsmeyer SJ. MLL and CREB bind cooperatively to the nuclear coactivator CREB-binding protein. Molecular and cellular biology. 2001;21(7):2249-58.
151. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
152. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
153. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.
154. NHLBI Exome Sequencing Project (ESP) Exome Variant Server. [Accessed on 12th January 2015]; Available at: http://snp.gs.washington.edu/popgenSNP/.
155. The Exome Aggregation Consortium (ExAC) Browser. [Accessed on 12th January 2015]; Available at: http://exac.broadinstitute.org/.
156. Sherry ST, Ward M, Sirotkin K. Use of molecular variation in the NCBI dbSNP database. Human mutation. 2000;15(1):68-75.
157. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.
158. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42(Database issue):D980-5.
159. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
160. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.
161. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491-8.
162. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
163. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2014;47:11 2 1- 2 34.

164. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nature genetics. 2014;46(9):944-50.
165. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci. 2016;19(4):571-7.
166. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
167. Huang L, Jolly LA, Willis-Owen S, Gardner A, Kumar R, Douglas E, et al. A noncoding, regulatory mutation implicates HCFC1 in nonsyndromic intellectual disability. American journal of human genetics. 2012;91(4):694-702.
168. Yu HC, Sloan JL, Scharer G, Brebner A, Quintana AM, Achilly NP, et al. An X-linked cobalamin disorder caused by mutations in transcriptional coregulator HCFC1. American journal of human genetics. 2013;93(3):506-14.
169. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9.
170. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic acids research. 2015;43(Database issue):D1049-56.
171. Meehan RR, Lewis JD, Bird AP. Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic acids research. 1992;20(19):5085-92.
172. Todd MAM, Picketts DJ. PHF6 Interacts with the Nucleosome Remodeling and Deacetylation (NuRD) Complex. Journal of Proteome Research. 2012;11(8):4326-37.
173. Wang QC, Zheng Q, Tan H, Zhang B, Li X, Yang Y, et al. TMCO1 Is an ER Ca(2+) Load-Activated Ca(2+) Channel. Cell. 2016;165(6):1454-66.
174. Akawi N, McRae J, Ansari M, Balasubramanian M, Blyth M, Brady AF, et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nature genetics. 2015;47(11):1363-9.
175. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nature genetics. 2014;46(10):1063-71.
176. Cobben JM, Weiss MM, van Dijk FS, De Reuver R, de Kruiff C, Pondaag W, et al. A de novo mutation in ZMYND11, a candidate gene for 10p15.3 deletion syndrome, is associated with syndromic intellectual disability. Eur J Med Genet. 2014;57(11-12):636-8.
177. Wang Z, Benoit G, Liu J, Prasad S, Aarnisalo P, Liu X, et al. Structure and function of Nurr1 identifies a class of ligand-independent nuclear receptors. Nature. 2003;423(6939):555-60.
178. Zetterstrom RH, Solomin L, Jansson L, Hoffer BJ, Olson L, Perlmann T. Dopamine neuron agenesis in Nurr1-deficient mice. Science. 1997;276(5310):248-50.
179. Kultgen PL, Byrd SK, Ostrowski LE, Milgram SL. Characterization of an A-kinase anchoring protein in human ciliary axonemes. Mol Biol Cell. 2002;13(12):4156-66.
180. Wise A, Tenezaca L, Fernandez RW, Schatoff E, Flores J, Ueda A, et al. Drosophila mutants of the autism candidate gene neurobeachin (rugose) exhibit neuro-developmental disorders, aberrant synaptic properties, altered locomotion, and impaired adult social behavior and activity patterns. J Neurogenet. 2015;29(2-3):135-43.
181. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic acids research. 2014;42(Database issue):D993-D1000.

182. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391-9.
183. Drury S, Williams H, Trump N, Boustred C, Gosgene, Lench N, et al. Exome sequencing for prenatal diagnosis of fetuses with sonographic abnormalities. Prenatal diagnosis. 2015;35(10):1010-7.
184. Pangalos C, Hagnefelt B, Lilakos K, Konialis C. First applications of a targeted exome sequencing approach in fetuses with ultrasound abnormalities reveals an important fraction of cases with associated gene defects. PeerJ. 2016;4:e1955.
185. Carss KJ, Hillman SC, Parthiban V, McMullan DJ, Maher ER, Kilby MD, et al. Exome sequencing improves genetic diagnosis of structural fetal abnormalities revealed by ultrasound. Human molecular genetics. 2014;23(12):3269-77.
186. Alamillo CL, Powis Z, Farwell K, Shahmirzadi L, Weltmer EC, Turocy J, et al. Exome sequencing positively identified relevant alterations in more than half of cases with an indication of prenatal ultrasound anomalies. Prenatal diagnosis. 2015;35(11):1073-8.
187. Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, et al. Clinical Diagnosis by Whole-Genome Sequencing of a Prenatal Sample. The New England journal of medicine. 2012;367(23):2226-32.
188. Westerfield LE, Stover SR, Mathur VS, Nassef SA, Carter TG, Yang Y, et al. Reproductive genetic counseling challenges associated with diagnostic exome sequencing in a large academic private reproductive genetic counseling practice. Prenatal diagnosis. 2015;35(10):1022-9.
189. Crabtree JS, Scacheri PC, Ward JM, Garrett-Beal L, Emmert-Buck MR, Edgemon KA, et al. A mouse model of multiple endocrine neoplasia, type 1, develops multiple endocrine tumors. Proc Natl Acad Sci U S A. 2001;98(3):1118-23.
190. Stoller JZ, Huang L, Tan CC, Huang F, Zhou DD, Yang J, et al. Ash2l interacts with Tbx1 and is required during early embryogenesis. Experimental biology and medicine (Maywood, NJ). 2010;235(5):569-76.
191. Bertero A, Madrigal P, Galli A, Hubner NC, Moreno I, Burks D, et al. Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. Genes & Development. 2015;29(7):702-17.
192. Minocha S, Sung T-L, Villeneuve D, Lammers F, Herr W. Compensatory embryonic response to allele-specific inactivation of the murine X-linked gene Hcfc1. Developmental Biology. 2016;412(1):1-17.
193. Sutherland HG, Newton K, Brownstein DG, Holmes MC, Kress C, Semple CA, et al. Disruption of Ledgf/Psip1 Results in Perinatal Mortality and Homeotic Skeletal Transformations. Molecular and cellular biology. 2006;26(19):7201-10.
194. Dias C, Estruch SB, Graham SA, McRae J, Sawiak SJ, Hurst JA, et al. BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. American journal of human genetics. 2016;99(2):253-74.
195. Funnell AP, Prontera P, Ottaviani V, Piccione M, Giambona A, Maggio A, et al. 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. Blood. 2015;126(1):89-93.
196. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet.385(9975):1305-14.
197. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236(4808):1567-70.
198. Mueller RF, Bishop DT. Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. Journal of medical genetics. 1993;30(9):798-9.

199. Bittles AH. Consanguinity and its relevance to clinical genetics. Clinical genetics. 2001;60(2):89-98.
200. Jaber L, Halpern GJ, Shohat M. The Impact of Consanguinity Worldwide. Public Health Genomics. 1998;1(1):12-7.
201. Carr IM, Bhaskar S, O'Sullivan J, Aldahmesh MA, Shamseldin HE, Markham AF, et al. Autozygosity mapping with exome sequence data. Human mutation. 2013;34(1):50-6.
202. Carr IM, Diggle CP, Touqan N, Anwar R, Sheridan EG, Bonthron DT, et al. Identification of autosomal recessive disease loci using out-bred nuclear families. Human mutation. 2012;33(2):338-42.
203. Makrythanasis P, Nelis M, Santoni FA, Guipponi M, Vannier A, Bena F, et al. Diagnostic exome sequencing to elucidate the genetic basis of likely recessive disorders in consanguineous families. Human mutation. 2014;35(10):1203-10.
204. Wiszniewska J, Bi W, Shaw C, Stankiewicz P, Kang S-HL, Pursley AN, et al. Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing. European journal of human genetics : EJHG. 2014;22(1):79-87.
205. Mefford HC, Batshaw ML, Hoffman EP. Genomics, Intellectual Disability, and Autism. New England Journal of Medicine. 2012;366(8):733-43.
206. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367(20):1921-9.
207. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380(9854):1674-82.
208. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med. 2012;367(14):1321-31.
209. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14(2):178-92.
210. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6.
211. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506(7487):185-90.
212. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron. 2013;77(2):235-42.
213. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
214. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American journal of human genetics. 1993;52(3):506-16.
215. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491-8.
216. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, et al. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science. 2001;293(5538):2256-9.
217. Miraoui H, Dwyer AA, Sykiotis GP, Plummer L, Chung W, Feng B, et al. Mutations in FGF17, IL17RD, DUSP6, SPRY4, and FLRT3 are identified in individuals with congenital hypogonadotropic hypogonadism. American journal of human genetics. 2013;92(5):725-43.

218. Ramos M, Trujillano D, Olivar R, Sotillo F, Ossowski S, Manzanares J, et al. Extensive sequence analysis of CFTR, SCNN1A, SCNN1B, SCNN1G and SERPINA1 suggests an oligogenic basis for cystic fibrosis-like phenotypes. Clinical genetics. 2013.

Appendix 1

Table 1: The 84 KMT2A variants observed in the individuals in my cohort with Wiedemann-Steiner syndrome.

Mutations refer to transcript: ENST00000534358.1 and Genome Reference Consortium human genome build 37 (GRCh37).

HGVSc (Transcript ENST00000534358.1)	HGVSp (ENSP00000436786.1)	Genomic co-ordinates	Consequence
c.152_186del35	p.Pro51ArgfsTer84	11:118307378-118307413	frameshift_variant
c.553C>T	p.Arg185Ter	11:118342427-118342427	stop_gained
c.1474_1493dup20	p.Pro500LeufsTer74	11:118343367-118343367	frameshift_variant
c.1660delC	p.Gln554SerfsTer13	11:118343533-118343534	frameshift_variant
c.2126_2127delCT	p.Ser709Ter	11:118343999-118344001	frameshift_variant
c.2262delC	p.Met755Ter	11:118344135-118344136	frameshift_variant
c.2318dupC	p.Ser774ValfsTer12	11:118344192-118344192	frameshift_variant
c.2318dupC	p.Ser774ValfsTer12	11:118344192-118344192	frameshift_variant
c.2318dupC	p.Ser774ValfsTer12	11:118344192-118344192	frameshift_variant
c.2452A>T	p.Lys818Ter	11:118344326-118344326	stop_gained
c.2461dupA	p.Ser821LysfsTer11	11:118344335-118344335	frameshift_variant
c.2659delG	p.Glu887SerfsTer62	11:118344532-118344533	frameshift_variant
c.2659G>T	p.Glu887Ter	11:118344533-118344533	stop_gained
c.3034C>T	p.Gln1012Ter	11:118344908-118344908	stop_gained
c.3460C>T	p.Arg1154Trp	11:118348807-118348807	missense_variant
c.3460C>T	p.Arg1154Trp	11:118348807-118348807	missense_variant
c.3482G>C	p.Cys1161Ser	11:118348829-118348829	missense_variant
c.3500G>A	p.Cys1167Tyr	11:118348847-118348847	missense_variant
c.3518_3521delGCTT	p.Cys1173Ter	11:118348864-118348868	frameshift_variant
c.3521T>G	p.Leu1174Ter	11:118348868-118348868	stop_gained
c.3556A>G	p.Lys1186Glu	11:118348903-118348903	missense_variant
c.3570-2A>G	-	11:118350887-118350887	splice_acceptor_variant
c.3613dupT	p.Tyr1205LeufsTer7	11:118350932-	frameshift_variant

		118350932	
c.3635-2A>C	-	11:118352428-118352428	splice_acceptor_variant
c.3647_3650delAAGA	p.Lys1216ArgfsTer18	11:118352441-118352445	frameshift_variant
c.3649G>T	p.Glu1217Ter	11:118352444-118352444	stop_gained
c.3651dupG	p.Lys1218GlufsTer4	11:118352446-118352446	frameshift_variant
c.3697delG	p.Val1233LeufsTer2	11:118352491-118352492	frameshift_variant
c.3697delG	p.Val1233LeufsTer2	11:118352491-118352492	frameshift_variant
c.3790C>T	p.Arg1264Ter	11:118352585-118352585	stop_gained
c.3790C>T	p.Arg1264Ter	11:118352585-118352585	stop_gained
c.3790C>T	p.Arg1264Ter	11:118352585-118352585	stop_gained
c.3790C>T	p.Arg1264Ter	11:118352585-118352585	stop_gained
c.3809delA	p.Lys1270ArgfsTer86	11:118352600-118352601	frameshift_variant
c.4012+1G>C	-	11:118352808-118352808	splice_donor_variant
c.4030C>T	p.Gln1344Ter	11:118353154-118353154	stop_gained
c.4054delA	p.Ser1352ValfsTer4	11:118353177-118353178	frameshift_variant
c.4090A>T	p.Lys1364Ter	11:118354901-118354901	stop_gained
c.4218+1delG	-	11:118355029-118355030	splice_donor_variant
c.4333-2A>C	-	11:118359327-118359327	splice_acceptor_variant
c.4503C>A	p.Cys1501Ter	11:118360530-118360530	stop_gained
c.4576-1G>A	-	11:118360843-118360843	splice_acceptor_variant
c.4599dupT	p.Lys1534Ter	11:118360867-118360867	frameshift_variant
c.4635G>A	p.Trp1545Ter	11:118360903-118360903	stop_gained
c.4635G>A	p.Trp1545Ter	11:118360903-118360903	stop_gained
c.4713_4714delCT	p.Cys1572Ter	11:118361926-118361928	frameshift_variant
c.5167delT	p.Tyr1723ThrfsTer12	11:118363933-118363934	frameshift_variant
c.5431C>T	p.Arg1811Ter	11:118366482-118366482	stop_gained
c.5646T>G	p.Tyr1882Ter	11:118367064-118367064	stop_gained
c.5664+1G>T	-	11:118367083-118367083	splice_donor_variant
c.5672G>T	p.Gly1891Val	11:118368658-118368658	missense_variant
c.5708A>G	p.His1903Arg	11:118368694-	missense_variant

		118368694	
c.5749G>T	p.Asp1917Tyr	11:118368735-118368735	missense_variant
c.5873A>G	p.His1958Arg	11:118369155-118369155	missense_variant
c.5902_5903delGT	p.Val1968LeufsTer4	11:118369183-118369185	frameshift_variant
c.5935C>T	p.Arg1979Ter	11:118369217-118369217	stop_gained
c.6002_6005delTTGT	p.Phe2001TrpfsTer8	11:118370057-118370061	frameshift_variant
c.6079+1G>A	-	11:118370136-118370136	splice_donor_variant
c.6379C>T	p.Arg2127Ter	11:118372446-118372446	stop_gained
c.6379C>T	p.Arg2127Ter	11:118372446-118372446	stop_gained
c.6571C>T	p.Arg2191Ter	11:118373178-118373178	stop_gained
c.6712delG	p.Asp2238IlefsTer8	11:118373318-118373319	frameshift_variant
c.6811delA	p.Arg2271GlyfsTer6	11:118373417-118373418	frameshift_variant
c.6913delT	p.Ser2305LeufsTer2	11:118373519-118373520	frameshift_variant
c.7144C>T	p.Arg2382Ter	11:118373751-118373751	stop_gained
c.7264G>T	p.Gly2422Ter	11:118373871-118373871	stop_gained
c.7419delT	p.Pro2474LeufsTer35	11:118374025-118374026	frameshift_variant
c.7485_7488delTTCT	p.Ser2496CysfsTer12	11:118374091-118374095	frameshift_variant
c.7567_7570delGTCA	p.Val2523LysfsTer2	11:118374173-118374177	frameshift_variant
c.7753delG	p.Asp2585IlefsTer17	11:118374359-118374360	frameshift_variant
c.8095C>T	p.Arg2699Ter	11:118374702-118374702	stop_gained
c.8099_8106delTGGCATCC	p.Leu2700ProfsTer2	11:118374705-118374713	frameshift_variant
c.8267delT	p.Leu2756Ter	11:118374873-118374874	frameshift_variant
c.8577T>A	p.Asn2859Lys	11:118375184-118375184	missense_variant
c.8806_8809delGTCT	p.Val2936Ter	11:118375412-118375416	frameshift_variant
c.8874_8875delAG	p.Lys2961GlufsTer13	11:118375480-118375482	frameshift_variant
c.9495dupA	p.His3166ThrfsTer10	11:118376102-118376102	frameshift_variant
c.9661delC	p.Leu3221SerfsTer35	11:118376267-118376268	frameshift_variant
c.9857_9858delCC	p.Pro3286GlnfsTer7	11:118376463-118376465	frameshift_variant
c.9983dupA	p.His3328GlnfsTer31	11:118376590-118376590	frameshift_variant
c.10353delA	p.Glu3451AspfsTer8	11:118376959-	frameshift_variant

		118376960	
c.10457_10458delTT	p.Phe3486TyrfsTer8	11:118377063-118377065	frameshift_variant
c.11374_11376delCCT	p.Pro3792del	11:118390723-118390726	inframe_deletion
Chr11:118354782-118362888del		11:118354782-118362888	Exonic_deletion

Figure 1. Wiedemann-Steiner syndrome and Hypertrichosis (WiSH) Study phenotype questionnaire

On the next three pages follows the phenotype questionnaire.

Wiedemann-Steiner syndrome and Hypertrichosis (WiSH) Study phenotype questionnaire

Study Number:…….. Genetics Centre…….. Clinician……..
Male / female Date of birth…….. Age at last assessment……..Yrs……..months……..

Family history

Ethnicity……. Consanguinity Yes ☐, No ☐
Family History of developmental disorders Yes ☐ No ☐ Unknown ☐
Details…….....
Mother's age at birth of child……..Yrs Father's age at birth of child……..Yrs

Pregnancy, birth and neonatal period

Conception: Natural ☐ Assisted ☐
Details.....
Were there any pregnancy or labour complications (inc. maternal illness, bleeding, abnormal scans, assisted delivery)? Yes ☐ No ☐ Unknown ☐
Details……....
Duration of pregnancy:.....weeks

Birth weight .....g (.....centile) Birth length.....cm (.....centile)
Birth OFC.....cm (.....centile)

Neonatal feeding problems? Yes ☐ No ☐ Unknown ☐
Details…….
Neonatal hypotonia? Yes ☐ No ☐ Unknown ☐
Details…….
Other neonatal problems? Yes ☐ No ☐ Unknown ☐
Details…….

Development, learning and behaviour:

Milestones (please write NYA if not yet achieved)
Sat …… ..months, Walked ……..….yrs…….. months, First words ……. Yrs…… months

Learning difficulties? No ☐ Mild ☐Moderate ☐Severe☐ Profound ☐
Behavioural problems (inc autism)? Yes ☐No ☐Unknown☐
Details…….
School type: Mainstream ☐ Special needs ☐
Details…….

Growth

Height.....cm (age .....yrs….....months) (..... centile)
Head circumference.....cm (age .....yrs.....months) (..... centile)
Weight.....kg (age….....yrs.....months) (.....centile)
Mother's height.....cm Father's height.....cm

Clinical features

Constipation? Yes ☐ No ☐ Unknown ☐
Details……...
Feeding difficulties Yes ☐ No ☐ Unknown ☐
Details…….

NG or PEG feeding Yes ☐ No ☐ Unknown ☐
Details……
Other GI problems Yes ☐ No ☐ Unknown ☐
Details……

Frequent infections? Yes ☐ No ☐ Unknown ☐
Details……...

Cardiac anomaly? Yes ☐ No ☐ if no, have they had an echo? Yes ☐ No ☐
Details of cardiac
anomaly……

Seizures? Yes ☐ No ☐ Unknown ☐
Details……
Autonomic dysfunction? Yes ☐ No ☐ Unknown ☐
Details…….....
Sleep disturbance? Yes ☐ No ☐ Unknown ☐
Details…….....
Reduced pain perception? Yes ☐ No ☐ Unknown ☐
Details……...
Other neurological abnormality? Yes ☐ No ☐ Unknown ☐
Details……...
Have they had a brain MRI? Yes ☐ No ☐ Unknown ☐
Result……...

Visual Abnormality? Yes ☐ No ☐ Unknown ☐
Details……...

Audiological Abnormality? Yes ☐ No ☐ Unknown ☐
Details……...

Premature eruption of dentition? Yes ☐ No ☐ Unknown ☐ Details……...
Other dental abnormality? Yes ☐ No ☐ Unknown ☐
Details……...

Hypertrichosis Arms Yes ☐ No ☐, Legs Yes ☐No ☐, Back Yes ☐No ☐, Face Yes ☐No ☐
Age hair growth first noted and distribution
……...
Swelling of hands or feet? Yes ☐ No ☐ Unknown ☐
Details……...
Other skin/bone/muscle abnormality Yes ☐ No ☐ Unknown ☐
Details……...

Urogenital Abnormalities? Yes ☐ No ☐ Unknown ☐
Details……...
Have they had a renal US? Yes ☐ No ☐ Unknown ☐
Result……...

Age entered puberty Normal ☐, Early ☐, Late ☐, Unknown ☐, Not reached ☐
Menstrual disturbance Yes ☐ No ☐ Unknown ☐ Details ……..
Other endocrine abnormality Yes ☐ No ☐ Unknown ☐ Details
……..

Please list any other problems / difficulties:……
……
……

On examination:

Joint hypermobility	Yes ☐	No ☐	Not Assessed ☐			
Fetal finger pads	Yes ☐	No ☐	Not Assessed ☐			
Deep palmar creases	Yes ☐	No ☐	Not Assessed ☐			
Sacral dimple	Yes ☐	No ☐	Not Assessed ☐			
Abnormal fat pads on feet	Yes ☐	No ☐	Not Assessed ☐			
Muscular Build	Yes ☐	No ☐	Not Assessed ☐			
Abnormal body fat distribution	Yes ☐	No ☐	Not Assessed ☐			

Other Examination findings:……
……

Do they take any medications? Yes ☐ No ☐ Unknown ☐, if yes please list:
……
……
……
Have they ever been admitted to hospital or attended the emergency department? Yes ☐ No ☐ Unknown ☐, if yes please list each occasion:
……
……
……
……

Please include relevant clinic letters, laboratory reports and growth data.

Do they take any medications? Yes ☐ No ☐ Unknown ☐, if yes please list:
……
……
……
Have they ever been admitted to hospital or attended the emergency department? Yes ☐ No ☐ Unknown ☐, if yes please list each occasion:
……
……
……
……

Please include relevant clinic letters, laboratory reports and growth data.

Appendix 2

Table 1: Pathogenic variants in *KMT2A*

KMT2A variants identified through Whole Exome Sequencing in 248 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features. Position refers to position on chromosome 11 and to transcript ENST00000534358. All variants are unique in the DDD dataset and WiSH dataset and not present in the eXac database [ref], The bottom two variants had previously present in ExAC but then had been subsequently removed. ** Patient is known to have an affected mother, all other individual have unaffected parents NB some trios we do not know about parents affected status. DN = de novo. PAT MOS = Paternal mosaicism.

ID	POSITION	CONSEQ	PolyPhen SIFT (Missense)	REF	ALT	HGVSc	HGVSp	INH
272567	118374704	Frameshift	-	ACTGGCATC	A	c.8098_8105delCTGGCATC	p.Leu2700ProfsTer2	*DN*
273060	118374171	Frameshift	-	ACAGT	A	c.7565_7568delCAGT	p.Val2523LysfsTer2	*DN*
267429	118353193	Stop gained	-	C	T	c.4069C>T	p.Gln1357Ter	U
259584	118375405	Frameshift	-	ATCTG	A	c.8799_8802delTCTG	p.Val2936Ter	U
260169	118367053	Frameshift	-	TG	T	c.5636delG	p.Cys1879PhefsTer2	U
260165	118352786	Stop gained	-	C	T	c.3991C>T	p.Gln1331Ter	U
258665	118374786	Frameshift	-	AC	A	c.8180delC	p.Thr2727LysfsTer30	U
265131	118372446	Stop gained	-	C	T	c.6379C>T	p.Arg2127Ter	*DN*
264841	118368658	Missense	PDamDel	G	T	c.5672G>T	p.Gly1891Val	*DN*
271611	118359327	Splice acceptor variant	-	A	C	c.4333-2A>C		*DN*
271660	118344530	Frameshift	-	CG	C	c.2657delG	p.Glu887SerfsTer62	*DN*
259227	118363932	Frameshift	-	AT	A	c.5166delT	p.Tyr1723ThrfsTer12	*DN*
273901	118375184	Missense	PDamDel	T	A	c.8577T>A	p.Asn2859Lys	*DN*
260868		Frameshift	-			c.1474_1493dup20	p.Pro500Leufs*74	*DN*
276248	118343528	Frameshift	-	GCCCCCC	GCCCCC	c.1654_1655delGCinsG	p.Gln554SerfsTer13	*DN*
270606	118390134-118392246	Multi-exonic deletion						*DN*
WW1	118348807	Missense	PDamDel	C	T	c.3460C>T	p.Arg1154Trp	*DN*
WW2	118348847	Missense	PDamDel	G	A	c.3500G>A	p.Cys1167Tyr	*DN*
WW3	118350931	Frameshift	-	C	CT	c.3612_3613insT	p.Tyr1205LeufsTer7	*DN*
WW4	118352436	Frameshift	-	AAAAG	A	c.3642_3645delAAAG	p.Lys1216ArgfsTer18	*DN*
WW5	118352491	Frameshift	-	TG	T	c.3697delG	p.Val1233LeufsTer2	PATMOS
WW6	118369155	Missense	PDamDel	A	G	c.5873A>G	p.His1958Arg	NOTMAT
WW7	118370136	Splice donor	-	G	A	c.6079+1G>A		*DN*
WW8	118374358	Frameshift	-	AG	A	c.7752delG	p.Asp2585IlefsTer17	*DN*
WW9	118376461	Frameshift	-	TCC	T	*DN*	:p.Pro3286GlnfsTer7	NOTMAT
WW10	118377062	Frameshift	-	CTT	C	c.10456_10457delTT	p.Phe3486TyrfsTer8	*DN*
WW11	118344185	Frameshift	-	A	AC	c.2311_2312insC	p.Ser774ValfsTer12	*DN*
WW12	118344185	Frameshift	-	A	AC	c.2311_2312insC	p.Ser774ValfsTer12	

Table 2: Pathogenic d*e novo* loss of function or missense mutations in DDG2P genes

De novo loss of function or missense variants assessed to be pathological in causing the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	PolyPhen/ SIFT	REF / ALT	MAF	GENO	EXAC
261629	ARID1B	6	157100465	ENST00000346085	Stop gained		C/T	***	1/0/0	NA
270826	ADNP	20	49507987	ENST00000396029	Frameshift		CA /C		1/0/0	NA
262888	CBL	11	119148874	ENST00000264033	Splice acceptor		A/T	NA	1/0/0	NA
263977	ARID1B	6	157150547	ENST00000346085	Stop gained		C/T	NA	1/0/0	NA
274225	MBD5	2	149226237	ENST00000407073	Frameshift		AC/ A	NA	1/0/0	NA
281166	SMARCA2	9	2086862	ENST00000382203	Missense	Benign DEL *	C/T	NA	1/0/0	NA
261706	SCN2A	2	166172031	ENST00000357398	Frameshift		TG/ T	NA	1/0/0	NA
273003	MED13L	12	116429567	ENST00000281928	Frameshift		G/G C	NA	1/0/0	NA
261065	MED13L	12	116452954	ENST00000281928	Stop gained		G/A	NA	1/0/0	NA
272674	CREBBP	16	3801726	ENST00000262367	Splice donor		C/A	NA	1/0/0	NA
260414	DNMT3A	2	25470582	ENST00000264709	Missense	ProbDam DEL	C/T	NA	1/0/0	NA
262471	BCL11A	2	60773352	ENST00000335712	Missense	ProbDam DEL	T/G	NA	1/0/0	NA
265865	HNRNPU	1	245022576	ENST00000283179	Splice donor		C/T	NA	1/0/0	NA
263116	ARID1B	6	157527664	ENST00000346085	Frameshift		CTG TT/ C	NA	1/0/0	NA
279294	SMAD4	18	48604664	ENST00000342988	Missense	ProbDam DEL	C/T	8.24 E-06	1/0/0	1
262215	SMARCB1	22	24175856	ENST00000263121	Inframe deletion		GA GA/ G	NA	1/0/0	NA
266748	EP300	22	41562653	ENST00000263253	Missense	PossDam DEL	A/G	NA	1/0/0	NA
257812	COL4A3BP	5	74722257	ENST00000380494	Missense	ProbDam DEL	G/A	NA	1/0/0	NA
260433	ARID1B	6	157150439	ENST00000346085	Stop gained		C/T	NA	1/0/0	NA
261034	ARID1B	6	157527539	ENST00000346085	Frameshift		CAG AA/ C	NA	1/0/0	NA
262754	SMARCA2	9	2060867	ENST00000382203	Missense	PossDam DEL	C/T	NA	1/0/0	NA
263882	ADNP	20	49508751	ENST00000396029	Frameshift		CTT TAT/ CT	NA	1/0/0	NA
267419	DYRK1A	21	38850576	ENST00000398960	Frameshift		AT/ A	NA	1/0/0	NA
272205	MED13L	12	116401227	ENST00000281928	Missense	ProbDam(0.999),DEL(0)	G/A	NA	1/0/0	NA
273042	ABCC9	12	21997785	ENST00000261200	Missense	PossDam DEL	G/T	NA	1/0/0	NA
273981	SYNGAP1	6	33411558	ENST00000418600	Frameshift		ACA GT/ A	NA	1/0/0	NA
276409	EP300	22	41543865	ENST00000263253	Frameshift		GCA TGG	NA	1/0/0	NA

							CC/G			
278805	EYA1	8	72211338	ENST00000340726	Frameshift		TG/T	NA	1/0/0	NA
279844	CTCF	16	67655480	ENST00000264010	Missense	ProbDam TOL	G/A	NA	1/0/0	NA
280914	DNMT3A	2	25467466	ENST00000264709	Missense	ProbDam DEL	C/T	NA	1/0/0	NA
265425	TUBA1A	12	49579605	ENST00000546918	stop_lost		C/A	NA	1/0/0	NA
258278	ABCC9	12	22061091	ENST00000261200	Missense	PossDam TOL	C/T	NA	1/0/0	NA
258975	ARID1B	6	157527679	ENST00000346085	Stop gained		C/T	NA	1/0/0	NA
258975	ARID1B	6	157527679	ENST00000346085	Stop gained		C/T	NA	1/0/0	NA
259221	ACTB	7	5567395	ENST00000331789	Missense	ProbDam DEL	T/C	NA	1/0/0	NA

Table 3: Pathogenic heterozygous variants in DDG2P genes inherited from an affected parent

Loss of function and functional variants in DDG2P genes inherited from an affected parent and assessed to be pathological in causing the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	Poly Phen /SIFT	REF / ALT	MAF	GENO	EXAC	PHENO
272901	RAD21	8	117866693	ENST00000297338	Frameshift		CT/C	NA	1/0/1	NA	FIT
260000	ARID1B	6	157517398	ENST00000346085	Missense	Poss Dam DEL	T/G	0.000122	1/0/1	NA	DP
263662	GRIN2A	16	10031815	ENST00000396573	Splice donor		C/A	NA	1/1/0	NA	CF
276420	ANKRD11	16	89350830	ENST00000301030	Frameshift		TC/T	NA	1/1/0	NA	FIT

Table 4: Pathogenic heterozygous variants in dominant DDG2P genes where inheritance information was not available

Loss of function and functional variants in dominant DDG2P genes where inheritance information was not available and assessed to be pathologenic in causing the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	PolyPhen/SIFT	REF/ALT	MAF	GENO	EXAC	Path
258284	ASXL3	18	31324288	ENST00000269197	Frameshift		AAGCTC/A	NA	1/NA/NA	NA	FIT
259668	HNRNPU	1	245022606	ENST00000283179	Stop gained		C/T	NA	1/NA/NA	NA	FIT
262336	NIPBL	5	37022374	ENST00000282516	Missense	ProbDamDEL	G/A	NA	1/NA/NA	NA	FIT
264183	ASXL3	18	31318772	ENST00000269197	Frameshift		A/AAATC	NA	1/NA/NA	NA	FIT
264262	EP300	22	41548351	ENST00000263253	Frameshift		AAG/A	NA	1/NA/NA	NA	FIT
264723	ARID1B	6	157527837	ENST00000346085	Frameshift		TAGAA/T	NA	1/NA/NA	NA	FIT
267485	WAC	10	28824686	ENST00000354911	Splice donor		GGTGA/GAACAGCAGTCCCCAAAGCCACTCTCAGCCCTTGCAGACGTCCCACCGCATGTGA	NA	1/NA/NA	NA	FIT
268440	ARID1B	6	157502271	ENST00000346085	Stop gained		C/T	NA	1/NA/NA	NA	FIT
270585	SETD5	3	9483320	ENST00000402198	Frameshift		TC/T	NA	1/NA/NA	NA	FIT
272241	ANKRD11	16	89351042	ENST00000301030	Frameshift		GTGTTT/G	NA	1/NA/NA	NA	FIT
259607	ANKRD11	16	89350783	ENST00000301030	Frameshift		CTT/C	NA	1/NA/NA	NA	FIT
259639	WAC	10	28878738	ENST00000354911	Frameshift		CA/C	NA	1/NA/NA	NA	FIT

Table 5: Pathogenic variants in X-linked DDG2P genes

Loss of function and functional variants in X-linked DDG2P genes assessed to be pathogenic in causing the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	PolyPhen/ SIFT	REF / ALT	MAF	GENO	EXAC
263763	PHF6	X	133527600	ENST00000332070	Missense	ProbDam DEL	C/T	NA	1/NA/ NA	NA
259485	HDAC8	X	71708853	ENST00000373561	Missense	Benign*	G/A		1/1/0	NA
260478	HDAC8	X	71715077	ENST00000373573	Missense	ProbDam DEL	A/G	NA	1/0/0	NA
258681	SMC1A	X	53432045	ENST00000322213	Missense	ProbDam DEL	G/A	0.00 012 2	1/0/0	NA
274689	DDX3X	X	41205589	ENST00000399959	Missense	PossDam DEL	C/T	NA	1/0/0	NA
269411	HDAC8	X	71571623	ENST00000373573	Missense	ProbDam DEL	A/T	NA	1/0/0	NA
267559	DDX3X	X	41205876	ENST00000399959	Splice donor		G/C	NA	1/0/0	NA
271331	HDAC8	X	71684432	ENST00000373573	Stop gained		A/T	NA	1/0/0	NA
271619	HDAC8	X	71684483	ENST00000373573	Missense	PossDam DEL	A/C	NA	1/0/0	NA
273139	DCX	X	110653435	ENST00000338081	Stop gained		G/C	NA	1/0/0	NA
277224	MECP2	X	153296516	ENST00000453960	Stop gained		G/A	NA	1/0/0	NA
278845	IQSEC2	X	53270970	ENST00000396435	Missense	ProbDam DEL	A/G	NA	2/0/0	NA
259137	DDX3X	X	41203558	ENST00000399959	Stop gained		C/T	NA	1/0/0	NA
270216	PHF8	X	54037639	ENST00000357988	Stop gained		G/A	0.00 013 8	2/1/0	NA

Table 6: Pathogenic biallelic variants in DDG2P genes

Loss of function and functional variants in biallelic DDG2P genes assessed to be pathogenic in causing the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features. *HACE1: Had recently been identified as a DD gene by the DDD analysis team. The publication documentation detailing this was under submission when this analysis was carried out.

ID	GENE	CHR	POS	TRANS	CONSEQ	REF/ ALT	MAF	GENO	EXAC
259339	TMCO1	1	165737436	ENST00000367881	Frameshift	ACT/A	0.000854	2/1/1	17
281381	HACE1* HACE1	6 6	105224626 105280997	ENST00000262903 ENST00000262903	Frameshift Stop gained	CTG/C G/A	0.000138 0.000138	1/1/0 1/0/1	NA NA

Table 7: Possible pathogenic variants in dominant DDG2P genes

Loss of function and functional variants in dominant DDG2P genes assessed to be possibly pathogenic in causing or contributing to the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	PolyPhen/ SIFT	REF/ ALT	MAF	GENO	EX AC
267975	DYNC1H1	14	102467512	ENST00000360184	Missense	Prob Dam, DEL	G/A	NA	1/NA/ NA	**NA**
279847	CHAMP1	13	115091109	ENST00000361283	Missense	Poss Dam, TOL	G/A	0.001	1/0/1	**5**
258295	SCN1A	2	166908329	ENST00000303395	Missense	Poss Dam, TOL	C/A	NA	1/0/1	**NA**
266755	ZSWIM6	5	60790136	ENST00000252744	Missense	Poss Dam, TOL	C/A	3.57E-05	1/0/1	**1**
267298	ARID1A	1	27087348	ENST00000324856	Missense	Unknown	A/C	NA	1/NA/ NA	**NA**
269266	SCN2A	2	166166908	ENST00000357398	Missense	Prob Dam, DEL	G/C	NA	1/0/0	**NA**
270556	ARID1B	6	157431623	ENST00000346085	Missense	Prob Dam, DEL	C/G	NA	1/NA/ NA	**NA**
271218	ARID1A	1	27088789	ENST00000324856	Missense	Unknown	G/C	NA	1/0/1	**NA**
272732	FANCA	16	89877126	ENST00000389301	Missense	Poss Dam, TOL	A/G	0.000276	1/1/0	**NA**
272732	FANCA	16	89882881	ENST00000567943	Missense	Unknown	G/T	0.0038	1/0/1	**2**
273187	ARID1B	6	157527742	ENST00000346085	Missense	Prob Dam, DEL	G/A	NA	1/1/0	**NA**
273379	HDAC4	2	240003811	ENST00000345617	Missense	Prob Dam, DEL	C/T	NA	1/NA/ NA	**NA**
278746	CBL	11	119149311	ENST00000264033	Missense	Poss Dam	G/T	0.000138	1/1/0	**2**
279847	IGF1R	15	99192836	ENST00000268035	Missense	Poss Dam, TOL	C/G	8.24E-06	1/0/1	**1**

Table 8: Possible pathogenic variants in biallelic DDG2P genes

Loss of function and functional variants in biallelic DDG2P genes assessed to be possibly pathogenic in causing or contributing to the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	PolyPhen/ SIFT	REF/ ALT	MAF	GENO	EX AC
275918	PC	11	66617092	ENST00000393960	Missense	Poss Dam, DEL	T/C	0.000 276	2/1/1	**5**
267275	CPS1	2	211523345	ENST00000430249	Missense	Poss Dam, DEL	T/C	0.000 244	2/1/1	**NA**
264350	COG1	17	71196139	ENST00000299886	Missense	Prob Dam, DEL	C/T	0.008 8600 9	2/1/1	**642**
271585	FAT4	4	126336669	ENST00000394329	Missense	Prob Dam, DEL	G/A	0.000 244	2/1/1	**NA**
259262	FAR1	11	13729542	ENST00000354817	Missense	Prob Dam, DEL	C/T	0.000 276	2/1/1	**NA**
259262	COG1	17	71202934	ENST00000299886	Missense	Prob Dam, DEL	C/T	0.000 276	2/1/1	1

Table 9: Possible pathogenic variants in X-linked DDG2P genes

Loss of function and functional variants in X-linked DDG2P genes assessed to be possibly pathogenic in causing or contributing to the individual's phenotype identified through whole Exome Sequencing in 228 individuals with a phenotype consistent with Wiedemann-Steiner syndrome, related phenotypes or increased body hair plus other phenotypic features.

ID	GENE	CHR	POS	TRANS	CONSEQ	Poly Phen /SIFT	REF/ ALT	MAF	GENO	EX AC
266180	HUWE1	X	53634593	ENST00000342160	Missense	Prob Dam, TOL	T/A	0.00019	1/1/0	5
267975	OPHN1	X	67333062	ENST00000355520	Missense	Poss Dam, DEL	T/C	NA	1/NA/ NA	NA
258369	OCRL	X	128721053	ENST00000371113	Missense	Prob Dam, DEL	C/G	0.00012 2	2/1/0	NA

Table 10: ZMYD11 *de novo* missense variants in the wider DDD cohort

Loss of function and functional *de novo* mutations in ZMYD11 identified from the wider DDD study of 4293 individuals. This excludes the two individuals with *de novo* mutations in ZMYD11 detailed in Chapter 3.

ID	SEX	CHR	POS	TRANS	CONSEQ	ALT/ REF	GENO (P/M/F)	ExAC FREQ
272015	F	10	294310	ENST00000397962	Missense	G/A	1/0/0	0
265790	M	10	294525	ENST00000397962	Missense	G/A	1/0/0	0
264849	M	10	298321	ENST00000397962	Missense	T/C	1/0/0	0