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).

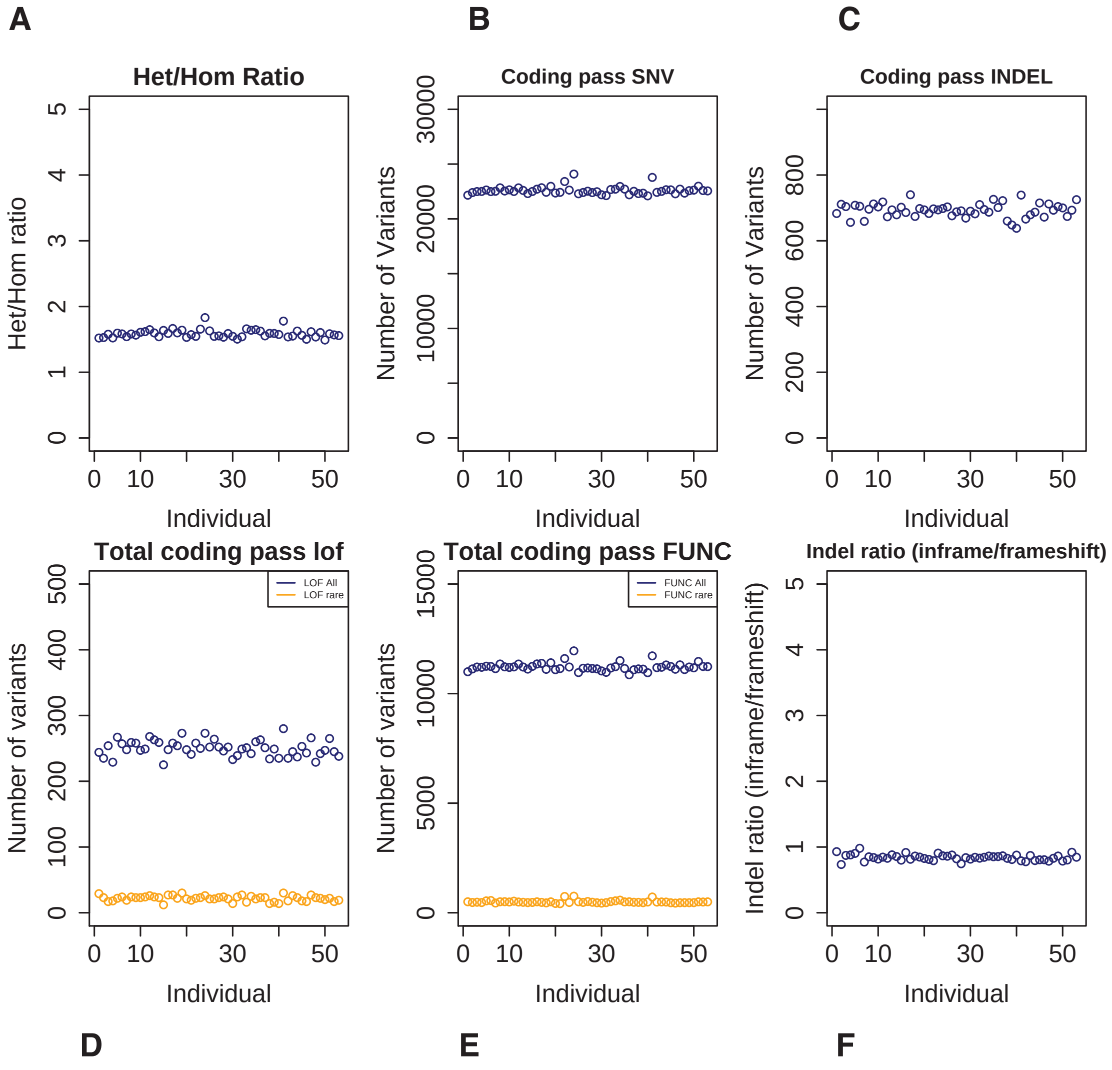

Figure 3-1: Quality control metrics for single nucleotide variants and INDELS in the WISH-WES cohort
(A) Ratio of heterozygous variants to homozygous variants. (B) Number of SNVs that lie within the coding region and have passed filters. (C) Number of INDELS that lie within the coding region and have 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).

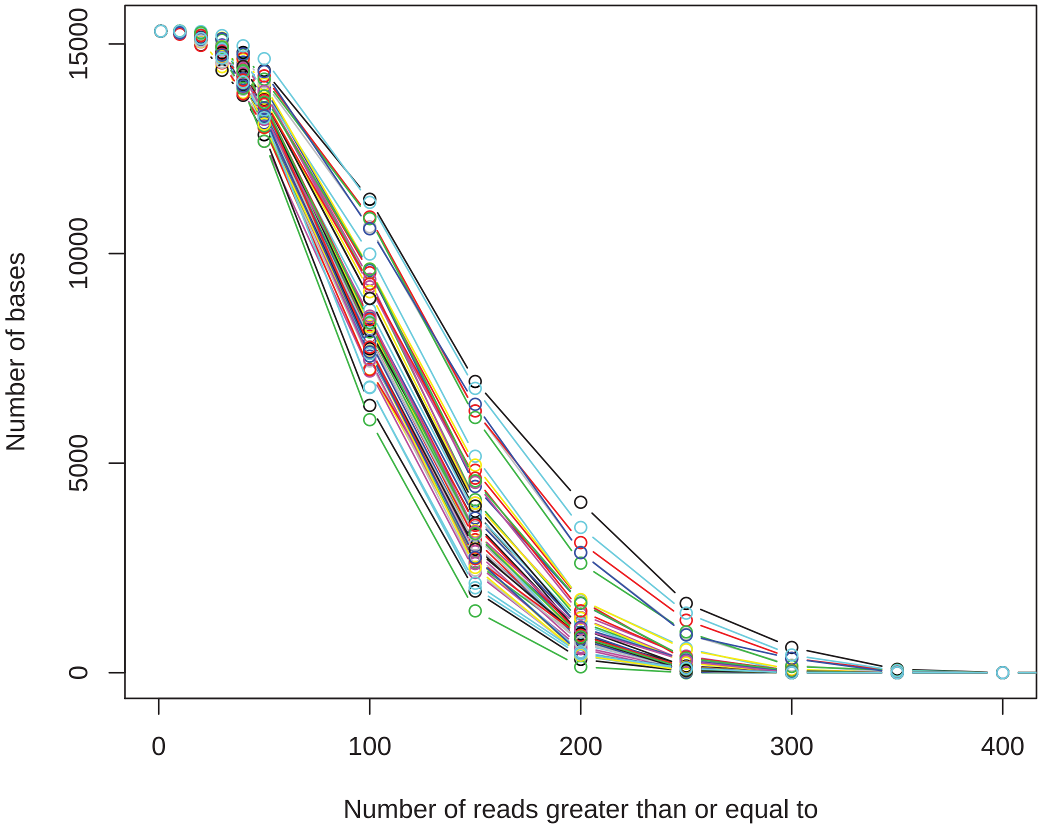

Figure 3-2: Coverage of KMT2A by exome sequencing reads by sample in the WISH-WES cohort
Number of bases covered by number of reads for parents and probands in the WISH-WES hypertrichosis cohort.

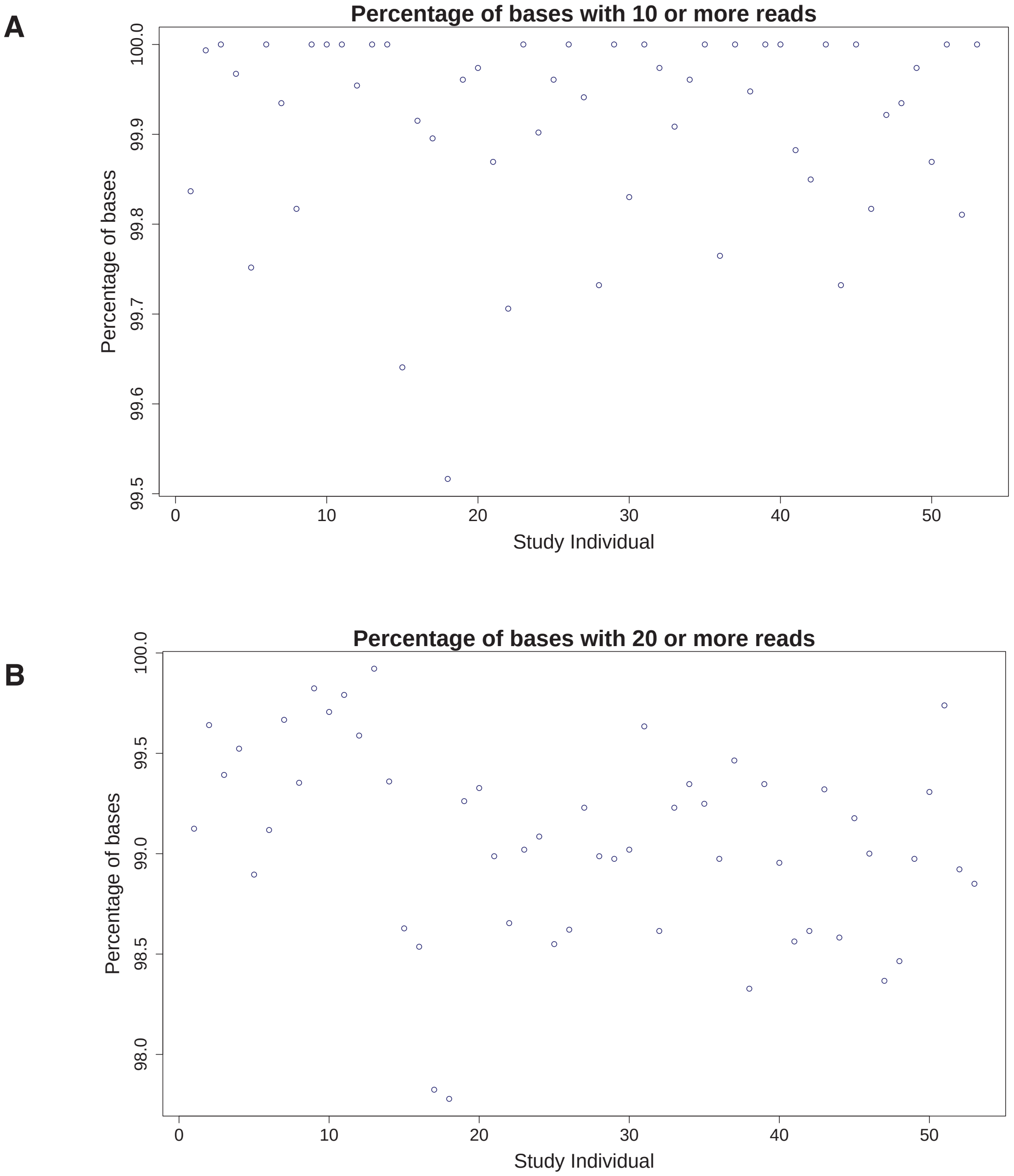

Figure 3-3: Coverage of KMT2A per study individual in the WISH-WES cohort
(A) Percentage of bases with 10 or more reads by study individual. (B) Percentage of bases with 20 or more reads by study individual.

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.