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.

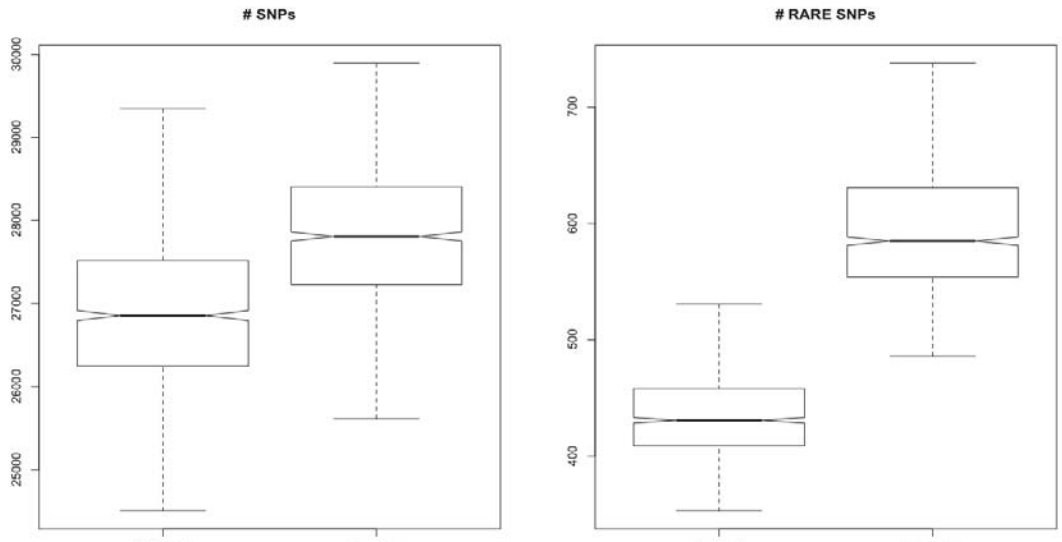

Figure 4-2: Number of single nucleotide variants (SNVs) in the probands and untransmitted diplotypes. A. Common SNVs. B. Rare SNVs MAF <0.01. These figures were plotted using data generated from a subset of 200 probands and 200 untransmitted diplotypes.

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.

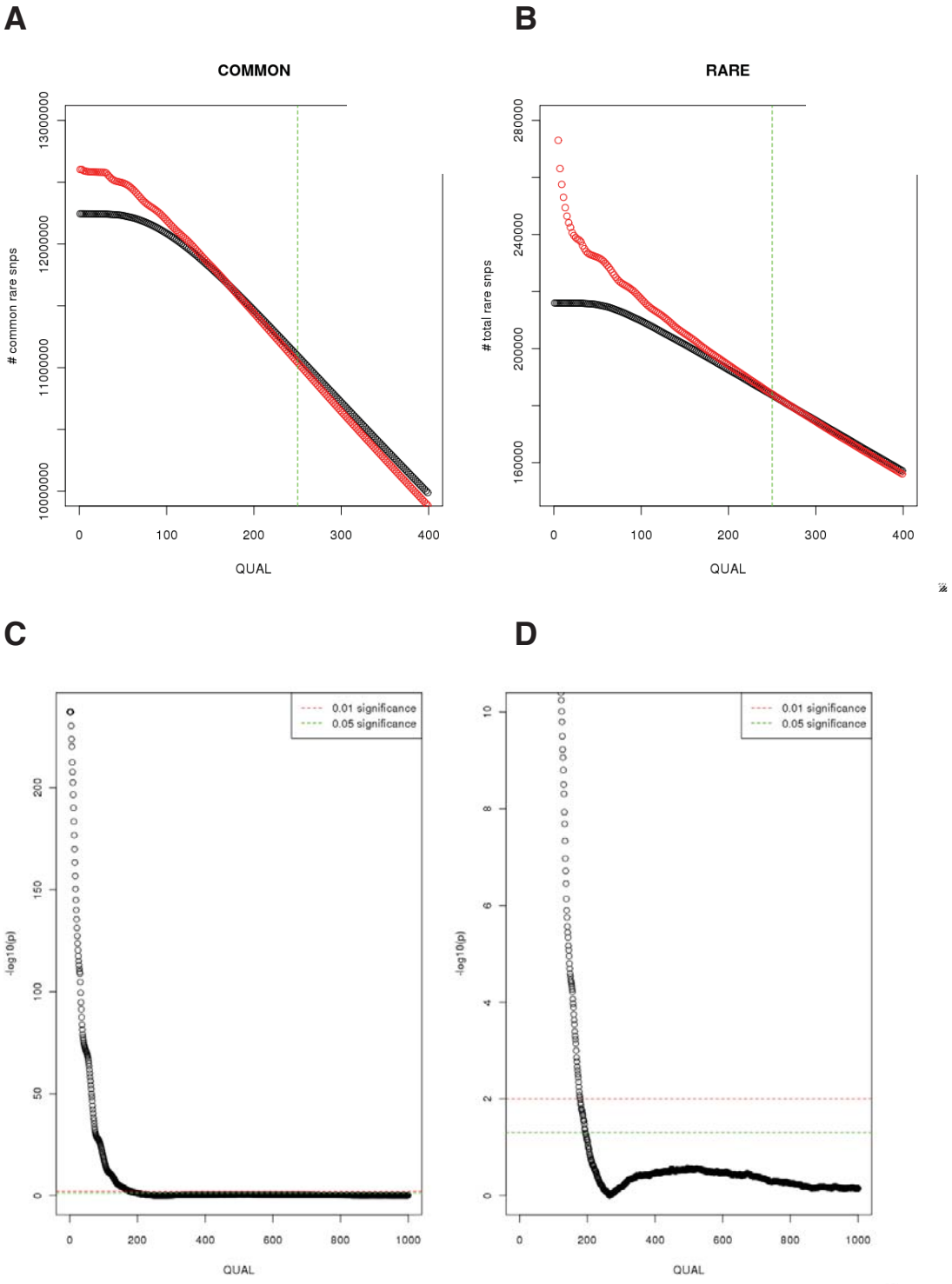

Figure 4-3: Studies of QUAL score and number of SNPs per sample for the probands and untransmitted diplotypes. **A:** Number of common variants against QUAL score for the probands and untransmitted diplotypes. Red dots represent untransmitted diplotypes, black dots represent probands. **B.** Number of 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.

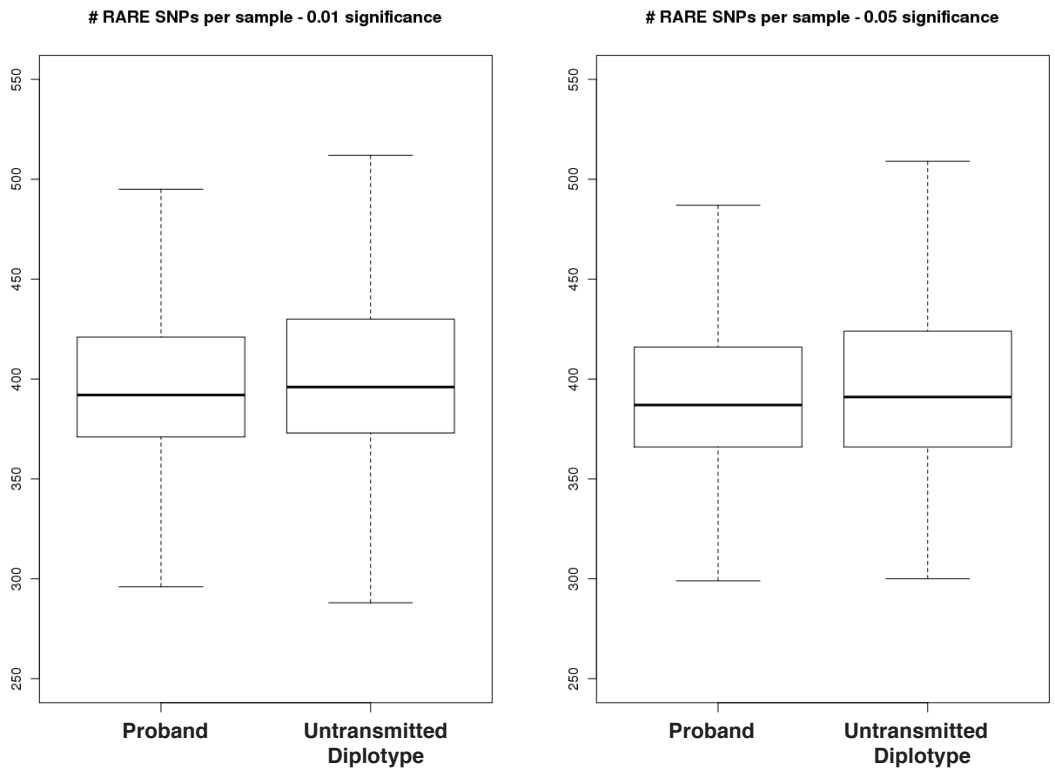

Figure 4-4: Boxplots to show number of single nucleotide variants (SNVs) in the probands and untransmitted diplotypes. A. Common SNVs. B. Rare SNVs: (MAF <0.01). These figures were plotted using data generated from a subset of 200 probands and 200 untransmitted diplotypes.

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.

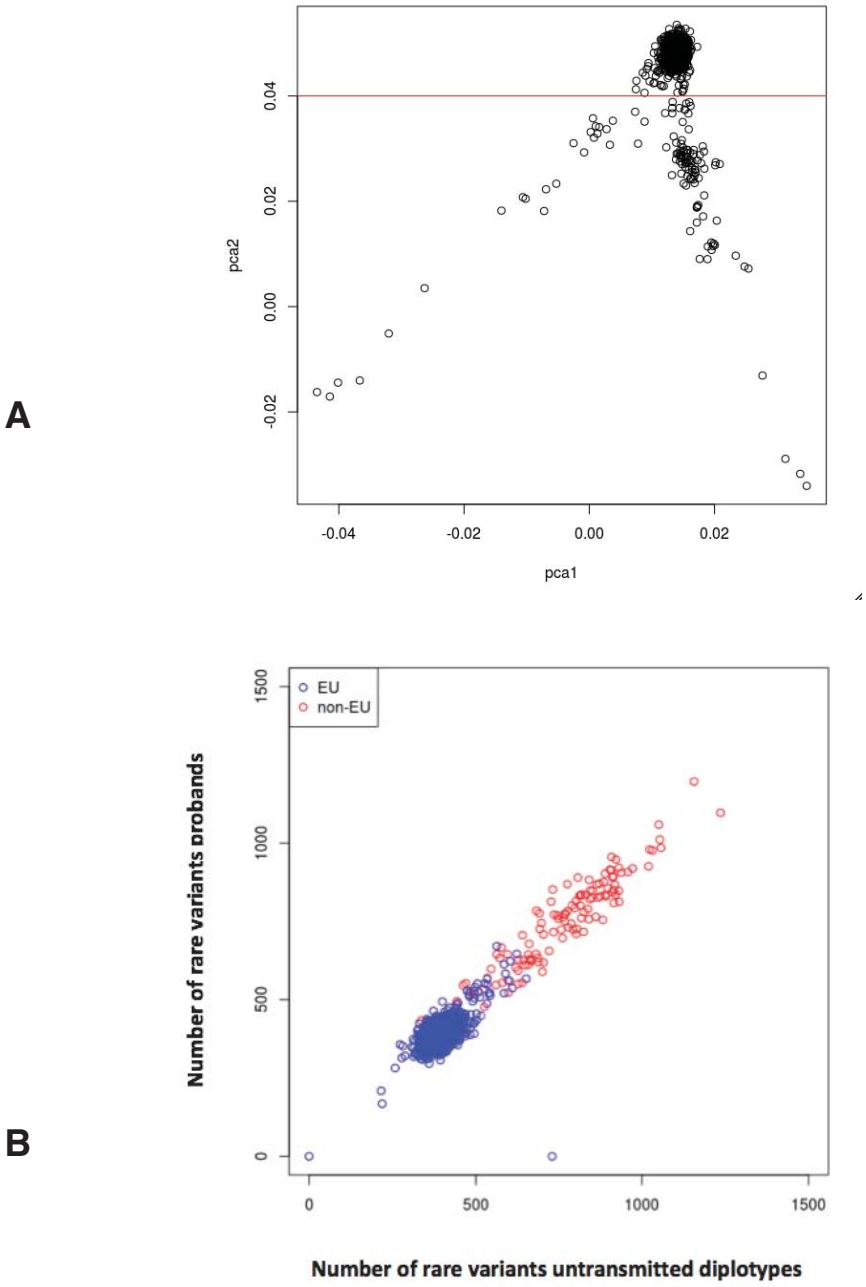

Figure 4-5: A: Principle component analysis (PCA) of individuals in the DDD.
European ancestry was defined by a PCA2 value of greater than 0.04. B: Comparison of the number of rare variants in the untransmitted diplotypes and probands. Variant numbers are labelled by ancestry, Blue circles represent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

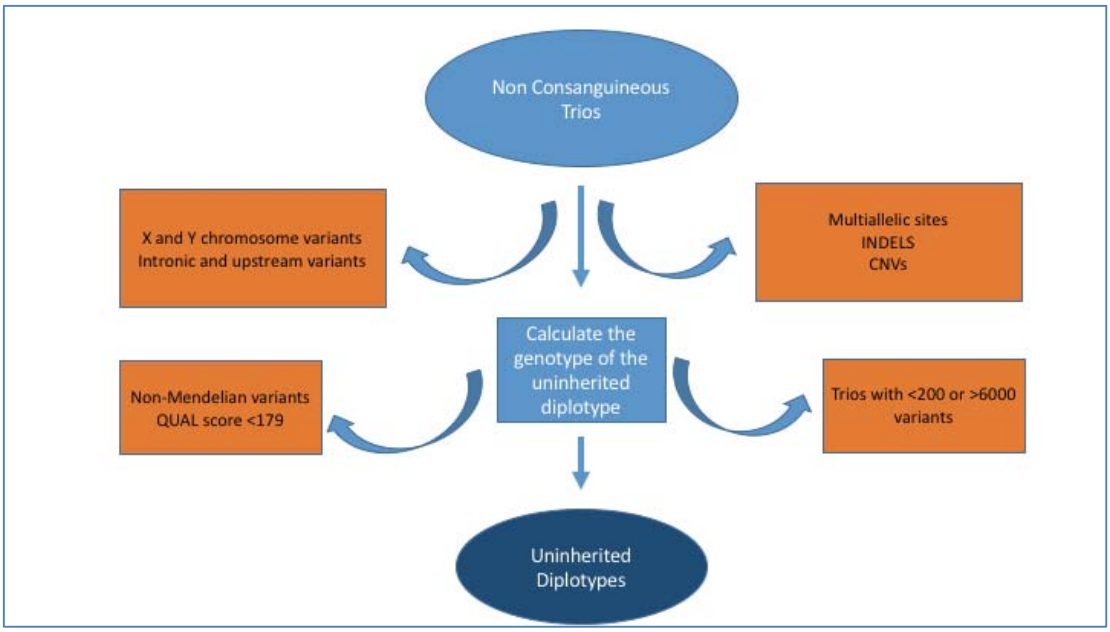

Figure 4-6: Flow diagram showing the processing steps for generating the untransmitted diplotypes.

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.