Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Identifying causative elements within structural variants associated with developmental disorders

Hannah Boulding

HERTFORD COLLEGE
OXFORD

Michaelmas Term 2012

Abstract

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Hannah Boulding, Hertford College, Michaelmas 2012

It has been well established that copy number variation contributes substantially to genetic variation within human populations. However, the extent to which *de novo* and inherited copy number variants (CNVs) underlie human disease is not well known. In this thesis, I investigate the role of *de novo* and inherited CNVs in a wide range of developmental abnormalities.

First, I compare disease associated and apparently benign CNVs for structural differences, with the aim of identifying distinguishing features of disease causing CNVs. I identified significant enrichments of protein-coding genes, protein-coding genes associated with disease in OMIM and miRNAs amongst disease associated disease. Conversely, inherited CNVs observed in healthy individuals show depletions of these features.

Following this, I employ functional enrichment approaches to identify the copy number variable genes within these *de novo* CNVs that contribute to the patient's developmental abnormalities. I predict candidate genes for 143 different developmental abnormalities, with 65% of the candidate genes not having been previously associated with disease in OMIM. Through examining the distribution of these candidate genes within the patient's CNVs, I found evidence of extensive pleiotropy and epistasis as well as a small number of simple additive effects.

Finally, I extend my analyses to examine the role of inherited CNVs as the underlying cause of human developmental disorders. I implicate inherited CNVs and their overlapping copy number variable genes in the underlying causes of 45 human developmental abnormalities. Additionally, I re-examine the patients possessing both *de novo* CNVs and inherited CNVs using functional enrichment analyses. I reveal significant enrichments for a greater number of human developmental abnormalities when combining both the *de novo* and inherited CNVs, suggesting it is *de novo* mutations in combination with the inherited genomic background that are responsible for many instances of human developmental abnormalities.

Acknowledgements

I would like to sincerely thank my supervisors Dr Caleb Webber and Prof. Chris Ponting for their guidance and support in supervising this project and for their helpful feedback on this thesis.

I would also like to thank all the members of the Webber and Ponting groups for their help and advice during my project and for making my graduate experience so enjoyable. I would particularly like to thank Steve Meader, Avi Agam, Charlotte Tibbit and Rob Young for their friendship and support.

Outside of the lab, I would like to thank my parents and my sister for their encouragement and support. I also thank Emma Robinson, Jen White, Emma Brookman and Kanika Dharmayat for their valued friendship.

Finally, I would like to thank the UK Medical Research Council and Hertford College for financial support.

Table of Contents

Chapter 1: Introduction

1.1 Structural variation in the human genome

Any two individuals' DNA sequences are estimated to be 99.9% identical (Przeworski *et al.* 2000); Reich *et al.* (2002); (Consortium 2010). The remaining 0.1% of these two individuals' genomes comprises single nucleotide polymorphism (SNPs) which were, until recently, held responsible for much of the phenotypic diversity between two individuals. The development of genome scanning technologies revealed microscopic and submicroscopic structural variation, which include deletions, duplications, inversions and translocations.

The first instances of human structural variation were observed as karyotypic changes through a microscope. They were consequently large, were often spontaneous changes and had large phenotypic effects. Although rare, both aneuploidies (the deletion or duplication of a whole chromosome) and other large chromosomal rearrangements were quickly associated with disease, for example, trisomy 21 with Down's syndrome and trisomy 18 with Edwards syndrome (Edwards *et al.* 1960; Patau *et al.* 1960; Coco and Penchaszadeh 1982; Warburton 1991).

The identification of smaller submicroscopic structural variation was enabled from the invention of DNA sequencing methods. The original DNA sequencing technology was developed in 1975 by Fred Sanger. This dideoxy chain termination method enabled the sequencing of the first genome, that of the bacteriophage phi X 174 (Sanger *et al.* 1977), which was only possible due to the small size of this bacteriophage's genome (~ 5000 bp). It wasn't until the 1980's that new technologies arrived enabling sequencing to be scaled up to cope with the large size and highly repetitive sequence content of eukaryotic

genomes. In 2000 the leaders of the public and private genome sequencing efforts announced the completion of two draft reference sequences of the human genome (Lander *et al.* 2001; Venter *et al.* 2001). The completion of the human reference genome followed by more recent developments in next generation sequencing technologies have enabled much work to be undertaken to assess the extent of human genetic diversity (Tuzun *et al.* 2005). This work has rapidly added to our understanding of both the types and prevalence of different forms of structural variation in the human genome (**Table 1.1**).

Types of structural variation

Genomic structural variation can take many forms, several of which are described below (**Table 1.1**).

Structural Variant	Description
Single nucleotide polymorphism	• Deletion, duplication or substitution of a single nucleotide within the genome. • May result in a premature stop codon or amino acid substitution.
Segmental Duplication	• 1-200kb blocks of duplicated DNA that occur at multiple sites in the genome and share a high level of sequence identity.
Variable number tandem repeat	• Nucleotide sequence that occurs in tandem blocks of repeat length 1-60bp.
Uniparental disomy	• When an individual inherits both copies of a chromosome from the same parent
Small Indels	• Small insertions and deletions ranging from 1bp-1kb
Copy number variant	• Duplication or deletion of DNA greater than 1Kb.

Table 1.1: Summary table describing the different types of structural variation identified in the genome.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the substitution of one nucleotide for another. Coding SNPs are either synonymous (when the change in nucleotide does not affect the amino acid sequence) or non-synonymous (when the nucleotide change results in either a change in amino acid sequence or creates a premature stop codon). Recent

estimates have predicted that there is 1 SNP for every 300bp of an individual's genome (Ke *et al.* 2008). Due to the widespread abundance of SNPs within the human genome, they have until recently been held responsible as the main contributor of human phenotypic diversity.

Variable number tandem repeats

Variable number tandem repeats (VNTRs) are locations in the genome where the nucleotide sequence occurs in repeat blocks of length 1-60 bp. VNTRs are very diverse within the human population, with variation in the size of the repeat block, in the number of repeat blocks and in the chromosomal location of the repeat block. VNTRs can be divided into two main sub-groups, microsatellites with a repeat block length of ~5bp and minisatellites with a repeat block of ~60bp (Breen 2010).

Segmental duplications

The completion of the human genome reference sequence revealed the prevalence of segmental duplications within the human genome (Lander *et al.* 2001; Venter *et al.* 2001). Segmental duplications comprise ~5% of the human genome, and are formed of 1-200 kb blocks of genomic sequence that possess ~90% sequence identity. Their high prevalence within the human genome, coupled with the previous estimates that segmental duplications appeared more frequently in primates than non-primate vertebrates, led to the hypothesis that they were responsible for a large proportion of human diversity (Stankiewicz *et al.* 2004). More recently, the completion of the mouse genome sequence revealed that mouse also possesses a high number of segmental duplications and the previous underestimates within non-human primate vertebrates are the result of draft genome assemblies capturing fewer segmental duplications than the finished assemblies

(Church *et al.* 2009). Consequently, segmental duplications are not thought to underlie as much human phenotypic diversity as was previously thought.

Uniparental disomy

Uniparental disomy (UPD) describes the occasion when a pair of homologous chromosomes within a diploid organism are derived from a single parent. Segmental uniparental disomy describes the instance when a portion of a pair of chromosomes is inherited from a single parent. UPD can be either heterodisomic or isodisomic depending on whether the UPD is a result of a meiosis I or a meiosis II error. The majority of cases of UPD do not result in an observable phenotype. However, cases have been observed of isodisomic UPD where the offspring inherited two copies of a recessive disease-causing allele from a carrier parent (Altug-Teber *et al.* 2005; Raghavan *et al.* 2005). Uniparental inheritance of imprinted genes can also result in an abnormal human phenotype (Cassidy *et al.* 2011). Imprinted genes show parent of origin specific levels of expression and thus possessing two maternal or two paternal chromosomes will drastically affect the expression level of an imprinted gene compared to individuals with one maternally and one paternally inherited chromosome.

Small Insertions and Deletions

Small insertions and deletions (INDELs) are insertions or deletions of genomic DNA <1Kb in the genome (Mullaney *et al.* 2010). Genome-wide studies have revealed over 2000 INDELs that range from 1-153bp in size within the human population (Dawson *et al.* 2001). More recent studies using sequencing data have estimated the number of INDELs within the human population to be over 200,000 (Mills *et al.* 2006). INDELs are known to contribute to much human phenotypic diversity and human disease. For example, a single base pair deletion within the *CFTR* gene is known to cause cystic fibrosis.

Copy number variants

Copy number variants (CNVs) are the deletion or duplication of genomic DNA >1kb in length (Redon *et al.* 2006). Before 2004, there had been little research into the role of CNVs (excluding aneuploidy) in human phenotypic diversity, indeed until this time it was widely believed that SNPs were the most important structural variant underlying human diversity. The first CNVs that were identified were thought to be rare and to be mainly implicated in cancer. During 2004, the development of new genome scanning technologies (see **Section 1.3**), revealed several hundred different copy number variable regions in healthy human individuals (see **Section 1.4**). Sebat *et al.* and Iafrate *et al.* identified several hundred CNVs in 20 and 55 individuals respectively (Iafrate *et al.* 2004; Sebat *et al.* 2004). Both studies revealed that between any two unrelated individuals, 11-12 different copy number variable regions can be observed. The two studies also highlighted the high prevalence of CNVs across the human population. Some CNVs were identified in ~90% of individuals, suggesting that they are very common within the human population. However, across the two studies only a small number of CNVs were replicated, suggesting that many CNVs that will be identified in the human genome are rare. Both studies revealed that CNVs overlap many genes and thus affect coding sequence, strongly implicating their role in human phenotypic diversity (see **Section 1.4**). After 2004, many large scale studies have been undertaken to identify copy number variable regions in both healthy individuals and patients with a wide range of different disorders (Pinto *et al.* 2007). Recently, efforts have been made to catalogue structural variants, for example in the Database of Genomic Variants, which catalogues CNVs from genome wide scans, and the databases DECIPHER and ECARUCA that catalogue CNVs observed in patients with developmental abnormalities (Feenstra *et al.* 2006; Zhang *et al.* 2006; Firth *et al.* 2009) (see **Section 1.6**).

1.2 CNV formation

Much research has been undertaken to identify the mechanisms underlying formation of copy number variants in the human genome. At present there are four mechanisms that are generally accepted as underlying CNV formation (**Table 1.2**).

Mechanism	Description
Non-allelic homologous recombination	• Recombination of two non-homologous genomic regions • Occurs in genomic hotspots • These hotspots are conserved across several primate species
Non-homologous end joining	• Initiated by genomic double strand breaks • Exploits regions of microhomology
Fork stalling template switching	• Initiated by stalling of the replication fork during DNA replication • The lagging strand disengages and switches templates via microhomology
Micro-homology mediated break induced replication	• Expansion of the fork stalling template switching model • Allows for more complex copy number variation than the first three models

Table 1.2: Summary table describing the different mechanisms of CNV formation.

Non-allelic homologous recombination

Non-allelic homologous recombination (NAHR) results in the duplication or deletion of a genomic sequence through the recombination of two non-allelic but sequence similar regions of the genome. NAHR can lead to a wide range of different structural outcomes depending on both the location and orientation of the two genomic regions that are to recombine.

NAHR events can be defined as interchromosomal events, intrachromosomal events or intrachromatid events (**Figure 1.1**). Each of these events leads to the duplication of genomic sequence on one chromosome/chromosome arm and the deletion of genomic sequence on the other. In the case of intrachromatid NAHR, an acentric circle of genomic DNA is produced. Often these circles, due to the absence of any double strand breaks, are stable and can persist in the genome.

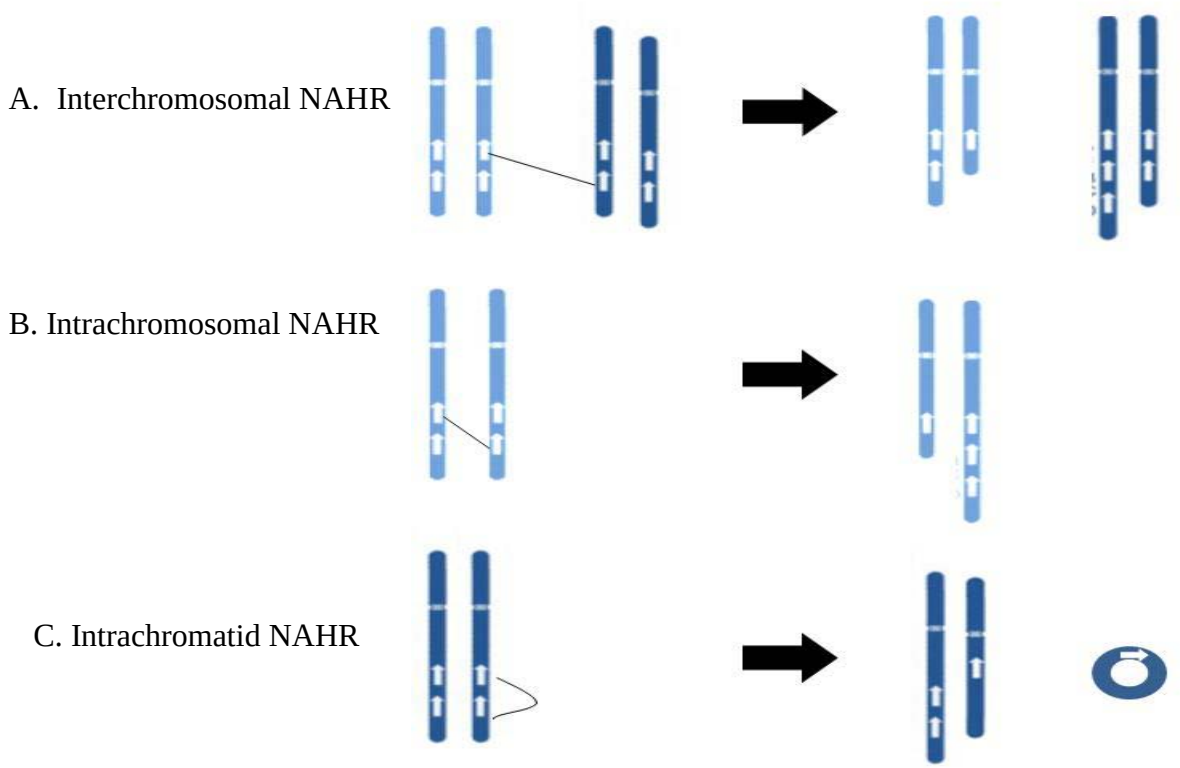


Figure 1.1: Mechanisms of genomic deletions and duplications through non-allelic homologous recombination (NAHR). A. NAHR resulting from the recombination of sequence similar regions on a pair of chromosomes that are not alleles. B. NAHR as a result of non-allelic sequence similar regions from different arms of the same chromosome recombining. C. NAHR due to non-allelic regions from the same chromosome arm recombining.

NAHR does not occur uniformly in the genome and has instead been shown to occur in genomic hotspots. These hotspots have been observed to be conserved across closely related primate species (Perry *et al.* 2008; Gazave *et al.* 2011). NAHR hotspots have been found to occur near DNA structures known to induce double-strand breaks, such as

minisatellites, transposons and palindromes. Identifying genomic regions that are more likely to undergo NAHR will enable the identification of potential disease causing deletions and duplications in the genome. Indeed, Uddin, M *et al.* searched for and identified 1,963 rearrangement hotspots in the human genome assembly. These hotspots were enriched in both NAHR events and flanking pathogenic deletions and duplications (Uddin *et al.* 2011).

Non-homologous End Joining

Non-homologous end joining (NHEJ) is a biological process initiated by double strand breaks in genomic DNA. It is defined as "non-homologous" as the process involves the direct ligation of two strands of genomic DNA without the need for an extensive homologous DNA template. NHEJ uses regions of microhomology to initiate the repair. These regions are found in the short single strand overhangs that are generated in a double strand break. Deletions and duplications can occur via NHEJ if there are multiple regions of microhomology close to the break region, which can therefore align incorrectly (**Figure 1.2**).

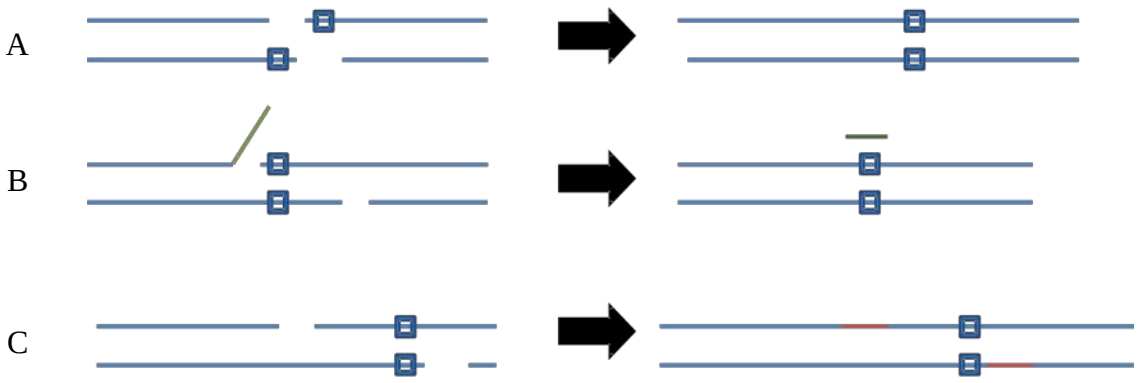


Figure 1.2: CNV formation via non-homologous end joining. Double strand breaks are ligated using regions of micro-homology (blue squares). (A): NHEJ that does not result in any duplication or deletion of DNA. (B): NHEJ that results in a deletion (green line). (C): NHEJ that results in a duplication (red lines).

Fork Stalling and Template Switching

Fork stalling and template switching (FOSTES) was proposed as a method of copy number formation in 2007 (Lee *et al.* 2007). The model proposes that the replication fork stalls during DNA replication as it approaches regions of DNA instability. This enables the lagging strand to disengage and move to another replication fork that can be moving in either the 3' or 5' direction. The switching of the templates requires micro-homology and as the lagging strand moves along its new template, the DNA is copied into the original sequence until the lagging strand disengages and returns to its original template. For FOSTES to occur, the two replication forks need to be in close proximity to each other. However, due to the folding of DNA, these two forks can be separated by large linear distances.

Micro-homology Mediated Break Induced Replication

Micro-homology mediated break induced replication (MMBIR) is an extension of the FOSTES model to include additional mechanistic molecular detail. MMBIR is a model for restarting collapsed replication forks independent of the RecA/Rad51 heterodimer required for homology induced repair. The difference between NHEJ and MMBIR, is that during MMBIR the regions of microhomology required for the model are followed by longer stretches of DNA inserted at the junction. Both FOSTES and MMBIR are models that allow for more complex copy number variations than simple duplications and deletions caused by NAHR and NHEJ, for example, discontinuous duplications mixed with deletions, inversions and or triplications (Zhang *et al.* 2009).

1.3 CNV detection

Recent advancements in both experimental and computational techniques have enabled scientists to call copy number variants from genomic data. Presently used methods of CNV detection involve either a genome wide or targeted analysis. Each of these methods have different strengths and weaknesses, however none of the current CNV detection methods has high quantitative accuracy (**Table 1.3**). As the CNV data used in my thesis are derived from comparative genomic hybridisation (CGH) methods, I will be concentrating my discussion on these methods.

Method	**Description**	**Pros (+)/Cons (-)**
Array CGH-BACs	Competitive hybridisation of a test and reference genome to bacterial artificial chromosome (BAC) targets.	+ Extensive genome coverage
aCGH-oligonucleotides	Competitive hybridisation of a test and reference genome to oligonucleotide targets.	+ Good resolution due to smaller probe spacing distances in comparison to other methods
SNP genotyping	Microarray containing short oligonucleotide targets	+ Can identify cases of loss of heterozygosity.
PCR based methods	Targeted approach using oligonucleotide probes or fluorescent reporters	—Unreliable for calling the precise number of observed duplications
Sequencing methods	Copy number variants are identified through paired end mapping or using short read depth	+Able to detect smaller CNVs in comparison to the previous methods

Table 1.3: Descriptions and pros and cons of CNV detection methods.

Comparative Genomic Hybridisation

Comparative genomic hybridisation detects copy number differences between test and reference genomes (**Table 1.3**). Chromosomal CGH was the first implementation of the CGH method, and involves hybridising both test and reference DNA sequences to metaphase chromosomes. The two DNAs are labelled with different fluorescent dyes, which when activated emit different wavelengths of light. Consequently, the differing

abundances between test and reference sequences can be quantified to reveal locations of deleted and duplicated genomic regions. Due to the low resolution of this method, the smallest identifiable CNV using this method is 5Mb (Molinaro 2002).

Currently the most frequently used CGH method for CNV identification is genome wide array-CGH methods (**Figure 1.3**). To identify differences in copy number these techniques competitively hybridise labelled fragments of both a reference genome and the genome of interest to an array containing DNA targets. The type of target found on the array can vary; possibilities (with different advantages and disadvantages) include BACs, oligonucleotides and SNPs (**Table 1.3**). BAC aCGH was the first implementation of aCGH and is popular due to its extensive coverage of the genome in comparison to other targets. BAC arrays consist of ~3000 probes of approximately 100-200kb in size, which can limit the detection of smaller CNVs. Conversely, oligonucleotides have an improved resolution over that achieved from employing BACs due to the former employing ~2 million probes with smaller probe spacing distances. To further improve the signal-to-noise ratio observed when using oligonucleotides, a process called representative oligonucleotide microarray analysis (ROMA) can be used (Lucito *et al.* 2003). This is achieved by amplifying the regions of interest through linker-mediated PCR amplification and removing DNA not of interest via restriction enzyme digests.

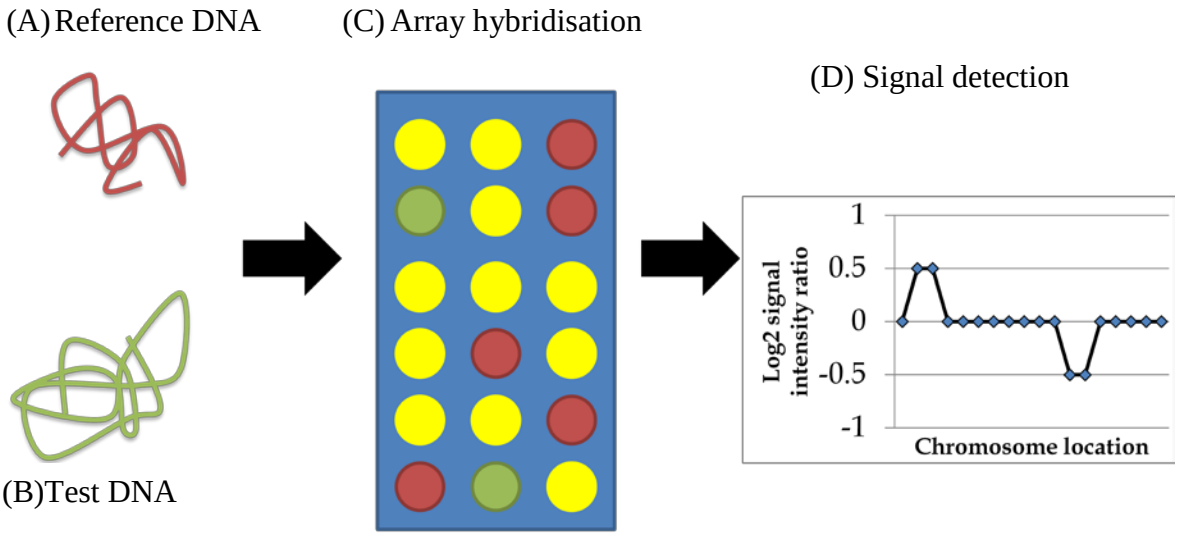


Figure 1.3: Array comparative genomic hybridisation protocol. (A) Reference DNA labelled with a red fluorescent tag. (B) Test DNA labelled with a green fluorescent tag. (C) The test and reference DNA are hybridised to targets spotted on an array. Example targets include BACs, oligonucleotides or PCR fragments. (D) The red to green fluorescence ratio is calculated to reveal copy number differences between the two samples.

SNP genotyping arrays can also be used to perform genome-wide CNV detection. These arrays contain short oligonucleotides and have the advantage of being able to genotype genomic data alongside the identification of copy number variation. This approach can therefore also identify cases of loss of heterozygosity (Mei *et al.* 2000).

CNV Detection using Computational Approaches

Once CNV data have been generated by CGH methods, various alternative computational algorithms can be employed to quantify each CNV call. These can be classified into four types of algorithm: smoothing, segmentation, threshold and hidden Markov model (HMM). Smoothing algorithms supply a visual aid to interpreting CNV data but do not automatically classify regions of copy number variation. They work by fitting a curve to

aCGH intensity plots across genomic regions with observed abrupt copy number jumps (Eilers and de Menezes 2005). Segmentation algorithms use the same specifications as smoothing algorithms (constant regions of copy number change with jumps in between) to identify CNV breakpoints in the region. The user can change parameters within this algorithm to balance the trade-off between calling small CNVs and partitioning the data too finely (Wang *et al.* 2005). Thresholding models classify genomic regions as copy number variable or normal (diploid) through creating log2 ratio thresholds. This calculation can be imposed on both individual probes or groups of probes (Pollack *et al.* 2002). Hidden Markov models are an advance on the previous three methods as they enable the incorporation of the relationship between different probes when making their CNV predictions. The previous three algorithms assume that each signal intensity is independent of those that neighbour it. HMMs however incorporate the dependence of neighbouring probes. This is particularly important when the aCGH protocol uses closely placed probes for example, oligonucleotides (DeSantis *et al.* 2009).

PCR Based Methods

In contrast to genome-wide array CGH detection methods, PCR based methods employ a targeted approach in identifying CNVs in the human genome (**Table 1.3**). The most common of these is quantitative-PCR (qPCR). The procedure follows the same method of the PCR reaction which is used to amplify a target section of DNA. During the qPCR reaction amplified DNA is detected as the reaction progresses, using one of two methods, either sequence-specific DNA oligonucleotide probes labelled with a fluorescent reporter or by using fluorescent dyes that bind non-discriminately to double stranded DNA. These methods work well for identifying genomic regions of duplications and deletions. However, like the other CNV detection methods described above, they are also not suitable for

accurately calling high numbers of observed duplications (i.e. differing between 5 or 6 copies) due to the variability in signal intensities (Ponchel *et al.* 2003).

Sequencing methods

CNVs can also be detected using data generated from next generation sequencing experiments. Next generation sequencing methods work by sequencing DNA in parallel, which in turn lowers the cost and time required in comparison to chain termination methods. There are two main methods for calling CNVs from sequencing data: paired end read mapping or using read depth (Chen *et al.* 2009; Yoon *et al.* 2009). In silico methods of CNV identification have revealed larger numbers of structural variants <1kb than was originally estimated. Paired end read mapping methods, for example Breakdancer, can detect deletions, duplications, inversions and translocations (Chen *et al.* 2009). The initial step is to map the paired end reads generated from a sequencing run to the reference genome. The number of anomalous read pairs can be counted and subsequently clusters of anomalous read pairs can be identified (**Figure 1.4**).

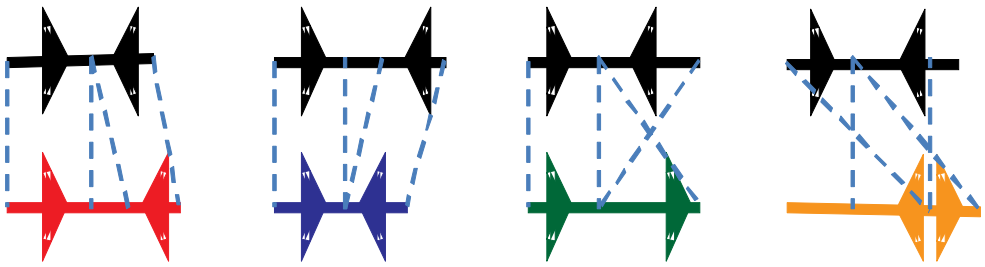


Figure 1.4: Structural variant classification is based on the distance and orientation of the paired end reads. The black line represents the test genome. The coloured arrows illustrate where the test genome maps to the reference genome (Red) Deletion. (Blue) Insertion. (Green) Inversion. (Yellow) Translocation.

Read depth methods (e.g. event wise testing) can detect deletions and duplications. In order to estimate read depth, the start position of sequencing reads within 100bp sliding windows are counted.

1.4 Copy Number Variation in Control Population Cohorts

Since the creation of the human genome reference assembly, many studies have been undertaken to identify the prevalence of CNVs in both healthy and disease state human populations (**Table 1.4**). Much effort has been undertaken to catalogue these structural variants, perhaps the most well known being the Database of Genomic Variants comprising CNVs overlapping 538Mb of genomic DNA obtained from ~1000 healthy individuals (Zhang *et al.* 2006).

Study	Description
Sebat *et al.* 2004	• Oligonucleotide array • Low coverage, detecting 76 CNVs
Iafrate *et al.* 2004	• BAC array • Higher coverage than the Sebat paper, identifying a greater number of CNVs • Lower resolution than the Sebat method, therefore identifies larger CNVs
Hinds *et al.* 2005	• Olignucleotide array identifying 215 CNVs • Identified many CNVs that were observed in multiple individuals • Able to identify small CNVs (<10Kb)
Sharp *et al.* 2005	• BAC array identifying 119 CNVs
Tuzun *et al.* 2005	• Identified 297 CNVs using sequence mapping • Identified complex rearrangements e.g. inversions
Locke *et al.* 2006	• BAC array identifying 384 CNVs
Redon 2006	• Multiple complementary BAC arrays identifying >1000 CNVs
McCarroll *et al.* 2006/ Conrad *et al.* 2006	• Hapmap trios and SNP genotypes to identify deletions • Unable to identify complex structural variants as those identified in the Tuzun *et al.* paper.
Wong *et al.* 2007	• BAC array identifying 3,654 CNVs. • Low replication rate suggesting a high false discovery rate
Shaikh *et al.* 2009	• Illumina hapmap 550 chip identifying 54462 CNVs
1000 genomes project 2010	• Sequenced the genomes of 1000 individuals drawn from different ethnic groups • Identified 1,500 copy number variable regions

Table 1.4: Summary of recent studies of CNVs in control populations

Early studies of CNVs in healthy individuals employed a combination of BAC arrays, oligonucleotide arrays, sequence comparison and HapMap trios (genomic maps for a proband and their parents formed from known observations of genomic variants that are inherited together). The first two papers describing global CNVs were Sebat *et al.* and Iafrate *et al.* who employed oligonucleotides and BACs, respectively (Iafrate *et al.* 2004; Sebat *et al.* 2004). The Sebat *et al.* paper identified 76 CNVs, however the design provided very low genomic coverage and the technology was not widely adopted (Scherer *et al.* 2007). Iafrate *et al.*'s method identified a greater number of CNVs (255) due to the increased genomic coverage. However, it had a lower resolution and was limited to identifying large CNVs. Hinds *et al.'s* later work improved on Sebat *et al.*'s work using oligonucleotides to identify 215 CNVs (Hinds *et al.* 2005). This method identified a CNV set of which a greater proportion was observed in multiple individuals (67% compared to Sebat's 41%). This study was also capable of identifying small CNVs (<10Kb).

Following the Iafrate *et al.* global BAC array study, Sharp *et al.*, Locke *et al.*, Redon *et al.* and Wong *et al.* used BACs to identify the extent of CNVs in the human genome (Sharp *et al.* 2005; Locke *et al.* 2006; Redon *et al.* 2006; Wong *et al.* 2007). Sharp *et al.* and Locke *et al.* identified similar numbers of CNVs to those seen in the Iafrate *et al.* study (119 and 384 respectively). The Redon *et al.* publication used multiple complementary arrays, and generated what is thought of as the first comprehensive genome map, identifying >1000 CNVs. This paper was also the first to discuss copy number variant regions (CNVRs; a contiguous region of DNA formed from individual overlapping copy number variants). The Wong *et al.* BAC paper identified 3,654 CNVs, which in comparison to other BAC studies appears to be a high rate of detection. This, in addition to the low replication rate (22%) implies a high false discovery rate in this particular experiment.

Tuzun *et al.* was the first study to employ paired-end sequence mapping to identify CNVs across the whole genome (Tuzun *et al.* 2005). The study identified a total of 297 CNVs and was able to identify more complex rearrangements such as inversions. Conrad *et al.* and McCarroll *et al.* were the first two studies to use HapMap trios to identify CNVs in healthy individuals (Conrad *et al.* 2006; McCarroll *et al.* 2006). By using HapMap trios, the parent of origin of each inherited CNV can be identified. These were the first two papers to use SNP genotypes to identify deletions. However, they were not able to describe complex structural variants such as those observed in the Tuzun *et al.* analysis.

Most recently, in 2010 the 1000 genomes consortium completed its pilot project. Genomes from 1000 individuals of various ethnicity were examined and 1,500 copy number variable regions covering 12% of the genome were identified (Consortium 2010).

As described above, the first instances of identifying benign CNVs (CNVs observed in healthy individuals and therefore thought to not underlie human disease phenotypes) arose from studies that employ both different detection platforms and processing algorithms. Consequently, data from different studies will vary in their accuracy, with datasets containing diverse false negative and false positive rates thereby making the integration of CNV datasets challenging. One issue in comparing CNVs from various datasets is the variation in terminology used. For example, in my thesis, I define a CNV as a deletion or duplication of genomic DNA greater than 1Kb that differs from the reference genome. The 1Kb definition stems from the limits of resolution in the earlier CNV detection technologies. As detection techniques improve, the 1kb to sub-microscopic indels will be included in the copy number variant definition, indeed recent work identified 415,436 indels (1-10,000bp) across 79 diverse human genomes (Mills *et al.* 2011). Additionally, many structural variants were initially defined cytogenetically, such as chromosomal trisomies and monosomies and consequentially are often described separately

from copy number variants, when in fact if judged by the "greater than 1Kb" definition they can be grouped together with CNVs.

To identify a CNV, one requires a reference genome to compare the test sample against. Currently, there is no standard reference DNA source, with individual studies using different reference genomes or even pooled genomic data (Scherer *et al.* 2007). This discrepancy can lead to different CNV calls in the same region, however using a pooled reference genome may help resolve some of the reference biases (Scherer *et al.* 2007). Indeed, comparing 6 studies that call CNVs using the sample test sample (NA15510) but using different reference sample and platforms leads to different numbers of CNV calls (**Table 1.5**). The first five platforms identify similar numbers of CNVs; however the number of CNVs detected by multiple platforms is small. The sequencing method identifies a much larger number of CNVs, validating the CNVs identified with the first five methods as well as identifying many smaller CNVs.

Platform	**Method**	**Reference**	**CNVs detected**	**Average CNV size**
BAC (WTSI)	Clone aCGH	NA10851	74	237Kb
Nimblegen 385K	Oligo CGH	NA10851	63	343Kb
Agilent 244K array	Oligo CGH	NA10851	42	74Kb
Affymetrix 500K	Comparative Intensity	Pool	24	316Kb
Illumina 650Y	Comparative Intensity	Pool	9	236Kb
Sequencing	Fosmid Ends	Build 35	241	29Kb

Table 1.5: Numbers and sizes of copy number variants identified using the same test sample (NA15510) and altering the detection method and the reference genome. (Adapted from Scherer *et al.* 2007). Different detection resolutions and genomic backgrounds alter the size and number of CNVs that can be accurately identified.

It is clear that a standardised approach to identifying CNVs is needed in order to integrate CNVs from different studies. Nonetheless, no one method can identify all sizes

and occurrences of CNVs accurately, hence it is important to collect these data from different sources into central databases while providing detailed annotations as to which discovery platform, method of detection and reference genome were used.

1.5 CNVs and Human Disease

Identifying the contributions of CNVs to human disorders is still ongoing. However, there are several well known examples of copy number variants that give rise to human disease, disease susceptibility and variation in drug response. For example, copy number variants on the X chromosome have been identified as the cause of *de novo* X-linked intellectual disability (Roeleveld *et al.* 1997). Copy number variants affecting cytokine gene dosage have been shown to alter an individual's HIV susceptibility (Gonzalez *et al.* 2005). Also, dosage variation in cytochrome P450 genes is known to alter an individual's ability to metabolise drugs, including nicotine (Oscarson *et al.* 1999). The mechanisms of how a CNV results in human disease are numerous, with some easier to detect and interpret than others (**Table 1.6**).

Mechanism	Description
Gene dosage	• CNVs overlapping protein coding genomic sequence can alter the number of transcribed copies of a gene • Responsible for Potocksi-Lupski syndrome and Smith-Magenis syndrome
Uncovering a recessive mutation	• A "benign" CNV deleting a protein coding gene in combination with a recessive mutation on the other allele may cause disease • Responsible for some cases of Cohen syndrome
Disruption of allelic exclusion	• A deletion or duplication overlapping genes that undergo allelic exclusion • Consequently a CNV will have different phenotypic affects depending on which allele it disrupts. • Responsible for Prader Willi and Angelman syndrome
Fusion genes	• Deletions and duplications can fuse genes or genes and promoters together resulting in abnormal gene products and gene expression

Table 1.6: Descriptions of how copy number variants can result in human disease.

The simplest way to understand how a CNV can result in an undesirable phenotype is through affecting gene dosage. CNVs result in the deletion or duplication of a genomic region of DNA greater than 1Kb, and thus if this region completely overlaps a gene and its regulatory elements the individual will possess one copy more or one copy less of the gene. For many genes in the genome, this does not result in a drastic phenotype change, for example healthy individuals have been observed to possess between 1 and 11 copies of the *CCL3L1* gene (Colobran *et al.* 2010). However, for other genes, their protein products need to be kept within strict limits in order to produce a "normal" human phenotype. In the case of a deletion of a gene whose protein product comprises a rate limiting step within a metabolic pathway, normal cellular function can be drastically disrupted (**Figure 1.5 A**). For example, patients with deletion CNVs encompassing the tyrosine hydroxylase gene, which encodes a rate limiting enzyme in dopamine synthesis, develop Parkinson's like symptoms (Bademci *et al.* 2010). Normal cellular function can also be disrupted through CNVs duplicating genes whose protein products aggregate with other proteins to regulate cell metabolism (**Figure 1.5 B**). For example, *SNCA* gene duplications result in increased alpha-synuclein aggregation, which is thought to promote degredation of nigrostriatal neurons in Parkinson's disease (Singleton 2005). Furthermore, some genes have been identified to be sensitive to dosage changes in both directions (**Figure 1.5 C**), for example, Smith-Magenis syndrome and Potocksi-Lupski syndrome are caused by a deletion and a duplication of genes (always including the *RAI1* gene) respectively on the p arm of chromosome 17 (Yatsenko *et al.* 2005).

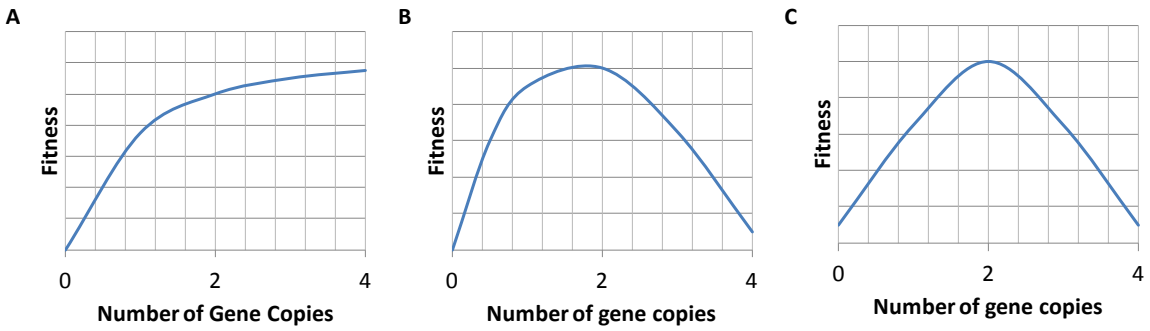


Figure 1.5: Illustrations of gene dosage against various fitness outcomes. (A) In the case of enzymes involved in a rate limiting step, a diminishing marginal return is observed. Cells with less than a diploid gene copy see a drastic reduction in their fitness. However as gene copies increase above diploid the increase in fitness per gene dosage decreases. (B) The result of fitness in regards to protein aggregation. As protein levels increase above the optimal threshold, the cell's fitness decreases. (C) The result of fitness in relation to gene dosage of genes sensitive to copy number change in both directions.

An abnormal phenotype can also occur when an individual harbours a copy number variant that deletes a gene while the other copy of that gene in the same individual has a recessive mutation. Ordinarily, a patient with a CNV overlapping this gene would present as healthy, as would a patient harbouring the single recessive mutation. However, the combination of the two can result in disease. For example, Rivera-Brugues *et al.* identified a group of patients presenting with the recessive disorder Cohen syndrome that had both a deletion CNV overlapping one allele of the *COH1* gene and a point mutation in the other allele, however the mutations were not observed together in controls (Rivera-Brugues *et al.* 2011).

CNVs can also cause disease through disrupting allelic exclusion. Allelic exclusion is the process where only one of a pair of alleles is expressed and the other is either transcriptionally or translationally repressed and consequently a CNV would have different phenotypic affects depending on which allele it resides on. The most well known examples of this phenomenon are imprinted genes such as *IGF2*, where only the allele from the father is expressed and the *UBE3A* gene where only the maternally inherited

allele is expressed. CNV deletions in the long arm of chromosome 15 are known to cause Prader Willi/Angelman syndrome depending on whether the deletion is observed on the paternally or maternally inherited chromosome (Cassidy *et al.* 2000). Recently, allelic exclusion has been observed in a tissue specific manner, suggesting CNVs have the potential to be benign and pathogenic, depending on the tissue under examination (Sun *et al.* 2010).

Fusion genes can result from both CNV deletion and duplication events. A deletion can lead to two genes being fused or a gene being fused to a new promoter thereby altering its expression level. A duplication may be inserted into a different region of the genome, again being fused to another gene and/or promoter, or the gene could be inserted into a gene causing that gene to be disrupted, essentially acting as a deletion. Recently a cohort of autism and control patients' CNVs were examined using RT-PCR to identify fusion gene transcripts. Fusion genes were found in both sets, suggesting that this type of duplication should be considered when identifying how a CNV causes a disease phenotype (Holt *et al.* 2012).

Partial gene duplications and deletions can also affect gene function. For example, a gene that is partially duplicated and inserted within its original gene may disrupt function and in effect act as a loss. Partial gene deletions can result in fusion genes, as discussed above or produce truncated proteins that cannot function normally.

1.6 The genetics of developmental disorders

On average, developmental disorders occur in 3% of births. They include a wide range of anatomical, metabolic and behavioural abnormalities, including facial clefts (0.12%), autism (1%) and intellectual disability (1-3%) (Chelly *et al.* 2006; Kogan *et al.* 2009; Bister *et al.* 2010). Often, patients presenting with a developmental disorder are observed to have multiple large (>100kb) *de novo* CNVs which are not often observed in the healthy population (median ~1 large *de novo* per person) (Itsara *et al.* 2010). Much work has been, and is currently being, undertaken to determine whether the CNV observed within the patient is the source of their developmental abnormality. Often, this is achieved by identifying patients who share the same developmental abnormality and an overlapping copy number variable region. This approach has identified many syndromes, for example, Di George syndrome (prevalence 1 in 4000 births), where patients present with learning difficulties, cardiovascular and palate abnormalities which is the result of a deletion at 22q11.2 (Yamagishi 2002). An additional example is Cri du Chat syndrome which has a prevalence of 1 in 50,000 births. Patients with this disorder present with severe intellectual disability, behavioural and feeding problems, and all are observed to have a deletion on the p arm of chromosome 5 (Rodriguez-Caballero *et al.* 2010).

Many developmental abnormalities occur much less frequently, and consequently it can be difficult for a single clinician to identify multiple patients with the same phenotypic presentation and an overlapping CNV. To remedy this situation and facilitate the identification of new pathogenic CNVs, databases such as DECIPHER and ECARUCA have been created. DECIPHER (DatabasE of Chromosomal Imbalance and Phenotypes in Humans using Ensembl Resources) and ECARUCA (European Cytogenetics Association Register of Unbalanced Chromosomal Aberrations) collate cytogenetic and clinical data across a large number of patients (Feenstra *et al.* 2006; Firth *et al.* 2009).

Currently, DECIPHER and ECARUCA hold over 2000 cases each within their respective databases. Within each database, cases have been obtained from several different clinical centres. Consequently, the CNVs have been called using a variety of different platforms and CNV detection algorithms. However, the patient phenotypes within DECIPHER and ECARUCA are described consistently using terms drawn from a strict medical ontology, the London Medical database (LMD) (Fryns and de Ravel 2002). Through using a strict medical ontology, patients obtained from different clinical centres can be grouped by their phenotypic presentations. Thus these resources enable the identification of recurrently copy number variable regions amongst patients presenting with the same developmental abnormality as well as enabling the application of functional enrichment analysis approaches to sets of non-overlapping CNVs observed in patients with the same developmental abnormality (see **Chapter 4** and **Section 1.7**).

CNVs observed in patients that are suspected to be pathogenic can be both *de novo* or inherited. There are arguments supporting how both of these types of CNVs can underlie the pathophysiology of developmental disorders. However, it is possible that it is a combination of both types, relative to the patient's genomic background that results in the disease phenotype. Two lines of evidence suggest *de novo* CNVs underlie many cases of developmental disorders. Firstly, due to the severity of the phenotype, patients with developmental disorders are less likely to reproduce, consequently not propagating their inherited structural variants to the next generation. However, from examining twin and adoption studies, many developmental abnormalities are observed to be highly heritable. For example, autism has a fertility ratio of 0.05 and yet a heritability of 0.9, suggesting mutations of recent origin are likely to contribute to the disorder (Uher 2009). The second piece of evidence supporting the role of *de novo* CNVs in developmental disorders is the observation of a paternal age effect (a positive correlation between the disorder's prevalence and the age of the father) in patients with developmental disorders,

particularly autism and intellectual disability (Saha *et al.* 2009; Uher 2009). The evidence that inherited CNVs underlie some developmental abnormalities is derived from the observation of an increased burden of CNVs amongst patients with certain developmental abnormalities, as is the case with patients with short stature and epilepsy who have been observed to have an increased burden of inherited CNVs (Dauber *et al.* 2011; Striano *et al.* 2012).

Despite the paternal age effect and the observation of an increased CNV burden amongst patients with developmental disorders, it can be difficult to predict whether a rare CNV is a major contributor to the developmental phenotype. Indeed Vermeesch *et al.* provides evidence that the causality of *de novo* CNVs is overestimated. The article states that CNV-patient phenotype analysis assumes the occurrence of both a rare disorder and a genomic mutation is statistically unlikely and hence the two must be linked causally (Vermeesch *et al.* 2011). However, the paper shows that the *de novo* CNV mutation rate is ~2.5 CNVs per 100 live births and therefore many *de novo* CNVs observed in patients are not unexpected. In my thesis, I make use of functional enrichment analysis approaches which are critically dependent on the expectation that a proportion of the CNVs being analysing are pathogenic, and consequently I remove CNVs deemed to be benign in order to analyse CNVs that are most likely to be pathogenic (see **Chapters 4** and **5**).

1.7 Functional Enrichment Analysis

Disease causing genes within CNVs can be identified by grouping patients presenting with the same disease phenotype, and finding overlapping structural variants within the patient sample. However, often patients are observed with a common phenotype but possess structural variants in different, non-overlapping genomic regions. In order to assess

whether a CNV that has not been observed in any other patient is involved in disease pathology, functional enrichment analyses (FEAs) can be utilised. The hypothesis behind these approaches is that distinct mutations in the human genome can give rise to the same human phenotype by affecting genes and/or regulatory elements belonging to the same biological pathway or process. Therefore, by disrupting any of the genes within this pathway the same human phenotype presentation may result. These genes are likely to share common attributes more often than is expected from a group of randomly selected genes, and this can be assessed by examining sets of genes for enrichments of features including protein-protein interactions (PPIs), phenotype data, cellular locations, genetic interactions, gene expression and literature co-citations.

There are many existing bioinformatics tools that implement FEA. These vary widely in their approach to testing functional enrichments. Enrichment tools can be classified into three categories: singular enrichment analysis, gene set enrichment analysis and modular enrichment analysis. Singular enrichment analyses calculate a P-value for each functional term of interest from a pre-selected gene list. The significantly enriched terms can then be ranked from lowest to highest P-value. These tests can be used on any list of genes, however singular enrichment analyses does not include data about the relationships between the terms being tested during its analysis. Well known tools that employ singular enrichment analyses include GoStat and DAVID (Beissbarth and Speed 2004; Huang da *et al.* 2009). Gene set enrichment analysis uses both a gene list and their experimental values during the functional enrichment analysis, allowing experimental values to be integrated into the P-value calculation. Tools that use this method include FatiGO and FatiScan (Al-Shahrour *et al.* 2006). Modular enrichment analysis, as used by Ontologizer, includes term-term and gene-gene interactions when calculating the P-value for the observed functional enrichment (Bauer *et al.* 2008).

To determine whether a functional annotation is enriched amongst a given gene set, one needs to identify a suitable control gene set against which to compare the test set (Webber 2011) (**Table 1.7**). The simplest of these approaches is to compare against genes from the whole genome (see **Chapter 4**). However, as the test set genes are being compared to the whole genome, overlapping CNV regions will need to be merged so each test gene is only counted once, thereby reducing the power obtained from recurrently occurring CNV regions. A second approach is to use a genomic background selected from the whole genome that represents the biases observed in the test CNV set, which could be used for example to account for the observed biases of GC content within CNVs (see **Chapter 3**) (Nguyen *et al.* 2008). For inherited CNVs it may be possible to compare gene content to a set of suitably matched inherited control CNVs (see **Chapter 5**). When matching case and control CNV sets, it is important to account for potential biases due to different methods of CNV detection between the two sets. Cases and control "hits" can be measured across the gene set as a whole, or through counting the number of individuals with a gene of interest. This second approach enables recurrently copy number variable regions to be counted more than once in the analysis.

Control	Description
Whole genome	• Merge CNV test regions • Compare the number of test genes with an annotation against all the genes in the genome
Whole genome accounting for GC bias	• Merge CNV test regions • Compare against sections of the whole genome with comparable GC content to the test set regions.
Case-control counting genes	• Merge test and control CNV regions • Count and compare the number of genes with and without genomic annotations in both sets
Case-control counting patients	• Count and compare the number of patients in both sets that have at least one copy number variable gene with a genomic annotation

Table 1.7: Possible control sets for FEA.

To perform FEAs on sets of CNVs, a list of protein-coding genes and genomic elements that overlap the set of CNVs needs to be identified. This process itself is subject to a number of biases (see **Chapter 2**). First, different median gene lengths have been observed amongst genes with different tissue specific expression (Webber 2011). Genes observed to be expressed specifically in the brain (defined as 4 times higher expression than the median expression of all other tissues) have been discovered to be an average of 44% larger than the median of all non brain specifically expressed genes. In comparison, genes specifically expressed in the heart are observed to be the smallest of all of the tissues specific expressed genes. Due to their larger size, brain expressed genes are more likely to be overlapped by CNVs than other categories of tissue specific genes, potentially leading to false positive enrichments. However, if it is only genes completely overlapped by CNVs that are examined (an approach often used when CNVs are called on older platforms which frequently overestimate CNV boundaries) this creates biases against longer genes, potentially generating false negatives when examining the roles of CNVs in neurological disorders.

Once a suitable test and control gene list has been generated, the most appropriate functional genomic resource or combination of resources to use to test for evidence of functional enrichment must be decided upon. Data within functional genomic resources are generated through direct experimental analysis, computational analysis, literature reviews or a combination of the three (**Table 1.8**). An example of a functional genomic resource generated from experimental evidence is the Mouse Genome Informatics resource which describes phenotypes resulting from the disruption of genes in the mouse (Smith and Eppig 2009). Due to the cost and amount of time needed to generate a knock-out mouse, this online resource contains data for only approximately 1/3 of all human:mouse 1:1 orthologues, fewer than that seen in other online resources. In addition, the genes with

annotations were likely to have been chosen for analysis due to a hypothesis that their disruption would produce an interesting phenotype result. This creates biases if the control set of genes is randomly chosen from the genome. The Gene Ontology (GO) is an example of an online resource where data are identified from a combination of experimental, computational and literature based searches (Ashburner *et al.* 2000). Exploiting a resource where data are generated using different methods leads to results with differing reliabilities. Only 20% of GO terms have been assigned experimentally, the remainder have been assigned through literature searches (and are therefore not expertly curated) and through computational predictions (prone to assigning GO terms to paralogues) (see **Chapter 5**). The MirTarBase is an example of a functional genomic resource that provides information for non-protein coding elements, in this case microRNAs (Hsu *et al.* 2011). This resource is formed from data obtained from literature. In addition, each annotation is annotated according to whether the original evidence was experimental or computational, enabling the user to decide the trade-off between ascertainment biases (experimental) or false positive rate (computational).

Resource	Data origin	Description
Mouse Genome Informatics	Experimental evidence	Mouse phenotypes resulting from the disruption of genes in the mouse
Gene Ontology	Experimental and computational evidence	Three ontologies labelling genes with cellular location, biological processes and molecular function annotations
KEGG	Experimental	Descriptions of metabolic pathways and their associations with protein coding genes
DAPPLE	Literature review	Protein-protein interaction data obtained from the literature
HMDD	Literature review	Associations between microRNAs and human diseases
MirTarBase	Experimental and computational	A list of microRNAs and their predicted protein coding gene targets

Table 1.8: Description of commonly used functional genomic resources

In summary, functional enrichment analyses have great potential to not only identify candidate genes, but to also identify disrupted biological processes/pathways that underlie disease pathology. However, to be used effectively, the biases associated with each potential method need to be accounted for (see **Chapter 5**).

1.8 Aims and structure of thesis

In this thesis I describe how I sought to identify disease causing variants that underlie human developmental abnormalities. Through exploiting a wide range of functional genomic resources I aimed to identify candidate genes and regulatory elements whose disruption underlies the disease pathology. Through analysing the candidate disease genes I sought to identify biological pathways and processes that are associated with human disease.

In **Chapter 3** I examine the genomic properties of disease associated and "benign" CNVs. I aimed to identify genomic properties that differentiate disease associated or benign CNVs. Within **Chapter 4** I exploit mouse phenotype data to identify candidate genes from two large sets of *de novo* CNVs observed in patients with developmental abnormalities. Following this, I examined the usefulness of using recurrent copy number regions to identify candidate genes. I also analysed the resultant candidate disease genes for evidence of epistasis (the interaction of genes that are not alleles), pleiotropy (the production of two or more unrelated effects by a single gene) and additive effects (the phenotypic effects of a combination of genes are equal to the sum of their individual effects). During **Chapter 5**, I consider the value of additional functional genomic resources in identifying candidate genes within CNVs, as well as those that enable the identification of non-protein coding elements that underlie disease. Following this, in **Chapter 6**, I analyse a set of CNVs obtained from patients not described using a medical

ontology, describing both the challenges and benefits associated with this type of clinical data. Finally in **Chapter 7,** I make use of GO and protein-protein interaction data to identify the biological pathways and processes underling individual developmental disorders. I analyse a set of inherited CNVs observed in patients and compare the results to those seen in the *de novo* CNV sets.

Chapter 2: Materials and Methods

2.1 CNV Datasets

During my DPhil project I made use of three datasets of CNVs observed in patients with developmental disorders and three sets of CNVs obtained from apparently healthy individuals (**Table 2.1**).

CNV set	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (Median)
DECIPHER	4749	0.3Mb	15532	4
ECARUCA	2798	16Mb	19918	105
Guys + St Thomas's	1843	0.6Mb	13240	5
SHAIKH	54462	8.1kb	7022	0
AGP	1843	0.6Mb	13240	5
NIJMEGEN	361	0.74Mb	1567	6

Table 2.1: The number, size and gene coverage of CNVs within the 6 CNV sets. DECIPHER, ECARUCA and Guys + St Thomas's set consist of CNVs observed in patients presenting with a range of developmental disorders. The SHAIKH, AGP and NIJMEGEN set consist of CNVs observed in healthy individuals. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

DECIPHER

DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Human using Ensembl Resources) is a web-based database that catalogues copy number variants observed in patients with developmental abnormalities (Firth *et al.* 2009). CNV data obtained through array comparative genome hybridisation (aCGH) from over 150 different centres are present in the database, together with a list of developmental abnormalities recorded in each patient. Currently, DECIPHER holds over 2000 cases. For each case,

patient and CNV data are entered by a clinical geneticist and molecular cytogeneticist, respectively. In order to ensure consistent human phenotypic annotation amongst the different centres, patient disorders are described using terms from the London Medical Database (Fryns and de Ravel 2002) (see **Chapter 2.2**).

I obtained 4749 CNVs from 1564 patients from the DECIPHER database. Patients present with multiple developmental disorders and congenital malformations (total = 861, median = 5, range = 1-32). CNV coordinates were mapped to the NCBI 36 genome assembly. The CNVs can be subdivided into those that are *de novo*, inherited or of unknown origin (**Table 2.2**).

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All CNVs	4749	0.3Mb	15532	4
(Gain)	2412	0.2Mb	9423	3
(Loss)	2337	0.3 Mb	11801	5
(1) Affected	41	1.2Mb	658	9
(Gain)	17	1.2Mb	303	8
(Loss)	24	1.3Mb	360	9
(2) Clinical	3	0.2Mb	2	1
(Gain)	0	0	0	0
(Loss)	3	0.2Mb	2	1
(3) De novo	626	2.2Mb	10487	18
(Gain)	162	1.9Mb	4717	22
(Loss)	464	2.3Mb	7411	17
(4) Inherited	2480	0.2Mb	3512	2
(Gain)	732	0.2Mb	2532	2
(Loss)	1016	0.2Mb	1556	2
(5) Unknown	1599	0.3Mb	11169	5
(Gain)	768	0.2Mb	5852	4
(Loss)	831	0.5Mb	8270	8

Table 2.2: The number, size and gene coverage of CNVs within DECIPHER. CNVs are split by their inheritance status. (1) Affected: Both the CNV and the abnormal human phenotype are observed in the child and the parent. (2) Clinical: CNVs deemed to be of clinical significance by DECIPHER. (3) *De* novo: CNVs observed in the child but not the parent. (4) Inherited: CNVs observed in the child and the healthy parent. (5) Unknown: CNVs of unknown inheritance. Each CNV set is further split into gain and loss CNV sets. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

ECARUCA

ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosomal Aberrations) is an online database for rare chromosomal abnormalities (Feenstra *et al.* 2006). At present, the database contains approximately 1500 chromosomal aberrations observed within ~4000 patients. Several of the cases within ECARUCA were obtained from the Zurich Cytogenetic Database (Schinzel 1994). ECARUCA excludes cases of autosomal trisomies, and X or Y chromosomal aberrations unless observed with another autosomal aberration. The CNVs within ECARUCA have been called using many different array platforms and cytogenetic techniques due to the CNV data originating from multiple diagnostic centres. As in DECIPHER, the patient's phenotypes are reported using terms from the London Medical Database (Fryns and de Ravel 2002).

I obtained 2798 CNVs observed in 2509 patients from ECARUCA. Again, patients present with multiple developmental disorders and congenital malformations (total = 1187, range of phenotypes per patient = 1-42, median number of phenotypes per patient = 8). CNV coordinates are mapped to the NCBI 35 genome assembly. However, during my analyses I map the ECARUCA CNV coordinates to the NCBI 36 genome assembly in order to match the DECIPHER CNV coordinates. Once again, CNVs were annotated as *de novo*, inherited or of unknown origin (**Table 2.3**).

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	2798	16Mb	19918	105
(Gain)	833	24.3Mb	17585	145
(Loss)	1965	14.1Mb	17301	99
(1) De novo	1143	17.3Mb	17791	112
(Gain)	159	32.2Mb	12330	208
(Loss)	984	16.4Mb	16026	107
(2) Inherited	568	16.3Mb	13181	127
(Gain)	263	24.9Mb	12430	170
(Loss)	305	12.7Mb	7426	99
(3) Unknown	1088	12.7Mb	16094	84
(Gain)	411	23.2Mb	12705	89
(Loss)	677	8Mb	12872	80

Table 2.3: The number, size and gene coverage of CNVs within ECARUCA. CNVs are split by their inheritance. (1) *De* novo: CNVs observed in the child but not the parent. (2) Inherited: CNVs observed in the child and the healthy parent. (3) Unknown: CNVs of unknown inheritance. Each CNV set is further split into gain and loss CNV sets. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Comparison between DECIPHER and ECARUCA

The DECIPHER and ECARUCA sets consist of CNVs observed in patients presenting with a wide range of developmental abnormalities, described using terms from the London Medical database (LMD) (Fryns and de Ravel 2002). In total, the DECIPHER set consists of a greater number of CNVs than the ECARUCA set. In addition, when the CNVs are segregated by their mode of inheritance, I observe that the proportions of *de novo,* inherited and unknown CNVs that make up each set differ substantially. The inherited CNVs make up the greatest proportion of the DECIPHER set (2480/4749 (52%), compared to 568/2798 (20%) in ECARUCA), whereas in the ECARUCA set it is *de novo* CNVs that form the greatest proportion of the sample (1143/2798 (41%), compared to 626/4749 (13%) in DECIPHER). Assuming *de novo* CNVs are more likely to produce severe developmental phenotypes than CNVs inherited from healthy parents, I hypothesised that this may be due to the differences in the number and severity of the patient's developmental abnormalities between the two sets. Indeed, I found that the data

support this hypothesis and on average, the DECIPHER patients present with a median of 5 abnormalities (range = 1-32), whereas the ECARUCA patients present with a median of 8 (range = 2-102).

The median size of the CNVs between the two sets varies greatly. Within DECIPHER the median CNV size is 0.3Mb, whereas in ECARUCA the median CNV size is 16Mb. This may be due to the differences of CNV calling techniques used in the two CNV datasets. Within DECIPHER, the CNVs are called using various array CGH platforms, whereas within ECARUCA a combination of array CGH and cytogenetic techniques are used, the latter often used to identify larger genomic rearrangements. Indeed, several of the CNVs within the ECARUCA dataset are from the Zurich cytogenetics database, and are substantially older than the DECIPHER CNVs. Consequently these CNVs would be large due to the CNV calling methods available at the time. Within DECIPHER, the *de novo* and inherited CNVs vary vastly in size (2.2 and 0.2Mb, respectively), however this is not observed with the ECARUCA set (17 and 16 Mb). Again, this may be due to the newer CNV detection methods employed within the DECIPHER set, which may have better resolution to detect small, as well as large, CNVs. Examining the DECIPHER and ECARUCA CNV set reveals an overlap of 193 Mb.

Guys and St Thomas's Dataset

I obtained 1843 CNVs from the Guys and St Thomas's hospital cytogenetics laboratory. Patients present with an average of 3 developmental abnormalities. The CNVs are mapped to Ensmart 54 and are described as either *de novo*, maternally inherited, paternally inherited, inherited from both parents, inherited from an affected parent or of unknown inheritance (**Table 2.4**). As with the other two datasets, CNVs were called

using various array CGH platforms. Unlike the DECIPHER and ECARUCA databases, the patient symptoms are not described using the London Medical database. Instead, no ontology is used, and the symptoms are described using terms deemed most suitable by each individual clinician.

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	1843	0.6Mb	13240	5
(Gain)	903	0.6Mb	10743	6
(Loss)	940	0.5Mb	7778	4
(1) De novo	218	1.4Mb	5621	18
(Gain)	52	1.4Mb	2261	17
(*Loss)*	166	1.4Mb	3635	18
(2) Inherited	643	0.4Mb	2609	4
(Gain)	347	0.4Mb	1969	5
(*Loss)*	296	0.2Mb	1072	2
(3) Unknown	985	0.7Mb	10936	6
(Gain)	507	0.7Mb	9242	7
(*Loss)*	478	0.6Mb	4957	8

Table 2.4: The number, size and gene coverage of CNVs within the Guys and St Thomas's dataset. CNVs are split by inheritance. (1) *De novo*: CNVs observed in the child but not the parent. (2) Inherited: CNV observed in the patient and one of their parents. (3) Unknown: CNVs of unknown inheritance. Each CNV set is split by copy number direction into gain and loss sets. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Control CNVs

In order to examine a set of CNVs for their involvement in a human disorder, a suitable genomic control needs to be identified. If that control is to be a set of CNVs, it is important to ensure that the two sets have similar attributes, for example size, gene overlap, and G+C content much of which can be controlled for by ensuring the test and control CNVs have the same mode of inheritance (i.e. *de novo* or inherited) and were called on the same detection platform and using the same calling algorithm.

Clearly, identifying a suitable set of CNV controls for the *de novo* CNVs in DECIPHER, ECARUCA and Guys and St Thomas's sets is difficult due to the combination of their unusually large size and *de novo* inheritance. Consequently, when I examine the *de novo* CNVs for their role underlying patient's disorders (**Chapters 3**, **4, 5** and **6**) I compared either randomly generated CNVs (randomly generated CNV sets with a similar size and GC content to the test set of CNVs) or the whole genome (all protein coding genes within the genome) to the DECIPHER and ECARUCA test sets. During **Chapter 7** I examine the inherited DECIPHER CNVs against a set of inherited "benign" CNVs from apparently healthy individuals.

Shaikh *et al.* Dataset

In **Chapters 3** and **7** I make use of a set of 54462 CNVs mapped to Ensmart 54 observed in healthy individuals, herein referred to as the Shaikh *et al.* dataset (Shaikh *et al.* 2009) (**Table 2.5**). The CNVs were detected using the Illumina humanhap 550 beadchip from whole blood obtained from healthy Caucasians, African-Americans and Asian-Americans. The Shaikh *et al.* dataset can be further split to create a list of CNVs observed in more than one individual (common CNVs).

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	54462	8.1kb	7022	0
(Gain)	8544	39.1Kb	3495	1
(Loss)	45918	6.3Kb	5059	0
Common	7641	0.17Kb	2603	1
(Gain)	2204	0.6Kb	982	1
(Loss)	5623	0.11Kb	1921	1

Table 2.5: The number, size and gene coverage of CNVs within the Shaikh *et al.* dataset. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Nijmegen *et al.* Dataset

I also make use of a further set of 361 inherited CNV controls, herein referred to as the Nijmegen *et al.* dataset (Nguyen *et al.* 2008) (**Table 2.6**) (see **Chapter 3**). The set consists of 361 CNVs observed in healthy trios. The CNVs were detected using a 32-k tiling resolution genomic microarray consisting of 32,447 BAC clones.

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	361	0.74Mb	1567	6
(Gain)	197	0.76Mb	1280	8
(Loss)	164	0.63Mb	773	6

Table 2.6: The number, size and gene coverage of CNVs within the Nijmegen control CNV dataset. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

AGP Dataset

In **Chapter 3** I also make use of the Autism Genome Project (AGP) "benign" set consisting of 2537 rare CNVs observed in controls (Pinto *et al.* 2010) (**Table 2.7**). CNVs were called using two CNV prediction algorithms, QuantiSNP (Colella *et al.* 2007) and iPattern (Gai *et al.* 2010), with 40% of the CNVs undergoing additional qPCR experimental validation. CNVs less than 30Kb in size in the AGP total sample (both cases and controls) were not included in the dataset.

CNV	Number of CNVs	CNV size (Median)	Overlapping Genes (total)	Overlapping genes* (median)
All	2537	96.5Kb	2789	1
(Gain)	1310	119.2Kb	2133	1
(Loss)	1227	77.3Kb	933	0

Table 2.7: The number, size and gene coverage of CNVs within the AGP control CNV dataset. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Comparison of Control CNVs

The three control CNV datasets vary in their size and gene overlap. The Shaikh *et al.* and AGP set consist of CNVs of a similar size, whereas the Nijmegen *et al.* dataset contains CNVs that tend to be larger (**Table 2.6**). Although the Shaikh *et al.* and AGP CNVs have a very different median size, the number of genes overlapped per CNV is similar. The AGP CNVs overlap a median of one gene per 96.5Kb whereas the common Shaikh *et al.* CNVs overlap one gene per 0.17Kb. These differences between the three datasets can be attributed to the different CNV calling methods. The Shaikh *et al*, Nijmegen *et al.* and AGP datasets were genotyped on an Illumina Hapmap 550 beadchip, BAC array and Illumina 1M platform, respectively. Although BAC arrays are popular due to their extensive coverage of the genome, their large probe size can limit the detection of smaller CNVs, which may explain why the Nijmegen set contains larger CNVs in comparison to the other two sets. The Nijmegen CNVs were also identified against a reference pool, rather than a single individual ensuring a low false positive rate. The Illumina 1M platform contains almost double the number of probes that the Illumina 550K platform has, giving better genomic coverage, which may explain the differences in gene overlap between the Shaikh *et al.* dataset and the AGP set.

2.2 Patient Phenotypes

Approximately 90% of patients within the DECIPHER, ECARUCA and Guys and St Thomas's dataset present with multiple developmental abnormalities. Within DECIPHER and ECARUCA the human developmental abnormalities are consistently described using terms from a medical ontology, the London Medical Database (Fryns and de Ravel 2002). Conversely, the patients within the Guys and St Thomas's dataset are described

inconsistently, as no ontology is used by the multiple clinicians entering patient data into the database.

London Medical Database

Patient phenotypes within DECIPHER and ECARUCA are described using terms from the London Medical Database (LMD) (Fryns and de Ravel 2002). The LMD is a three layer ontology consisting of 1682 terms describing human phenotypic abnormalities, arranged into hierarchical relationships with three levels of specificity. At its highest level the LMD consists of 34 overarching human symptom terms (e.g. neurology, skeleton). Each of the parent terms has multiple child (e.g. mental cognitive function, general abnormalities) and grand-child terms (e.g. mental retardation) listed beneath them (**Figure 2.1**). Each term in the LMD is listed beneath a single parent term, and there is no overlap between the branches of the ontology. Therefore, the symptoms observed in patients held within DECIPHER and ECARUCA are frequently described by three terms, one from each level of the ontology. The advantages of the London Medical database are twofold. Firstly, by annotating patients using a strict vocabulary it facilitates the grouping of patients that present with the same abnormality. Secondly, by exploiting the hierarchical structure of the ontology, I can group patients with similar phenotypes together, using an overarching phenotype term. As the LMD contains such a large number of terms, it is unlikely that each individual patient has been tested for all of them. Consequently, a lack of an LMD annotation for any particular patient does not confirm that the patient does not have that particular abnormality.

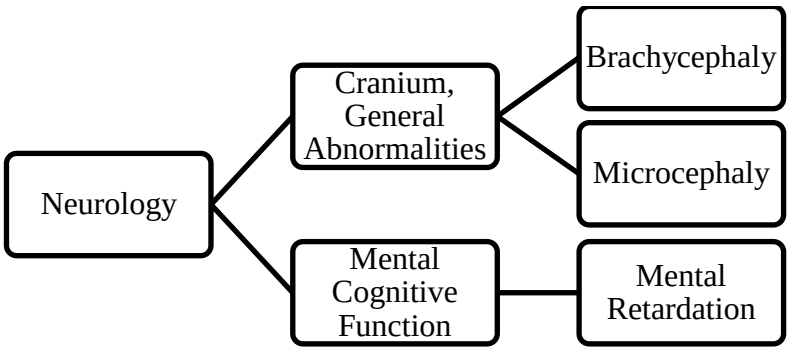


Figure 2.1: The London Medical Database Ontology. Patients presenting with a wide range of developmental disorders are consistently annotated with terms from the London Medical Database ontology. The ontology describes human phenotypes within a three level hierarchy. Each second and third level term is described beneath one single parent term. Therefore an abnormality observed in a patient can be described by up to three terms. For example, "Neurology" is the parent term of "Mental cognitive function, general abnormalities" which in turn is the parent term of "mental retardation/developmental delay".

The DECIPHER and ECARUCA patients present with 685 and 892 different phenotype terms respectively, with 489 different LMD terms observed in both the DECIPHER and ECARUCA sets. On average the patients within DECIPHER present with 5 LMD annotations and the ECARUCA patients present with 8.

Guys and St Thomas's

The patients' developmental abnormalities described within the Guy's and St Thomas's dataset are derived from records obtained from the clinicians who diagnosed the patients. As the patients within the dataset were described by different clinicians, patients with the same human phenotype are described using different terms e.g. Autism is described as "autism", "autistic", "ASD", and "autism-spectrum". Therefore, to group patients with a similar developmental abnormality I employed a pattern matching procedure using terms from different medical ontologies (see **Chapter 5**).

2.3 Protein coding gene annotations

Protein coding gene annotations were obtained from Ensembl (Flicek *et al.* 2010). When identifying genes disrupted by a CNV, I selected genes where at least one exon from every known transcript of the gene was overlapped by the CNV. This ensured that protein coding sequence is affected within the genes I took forward for analysis. The decision to use exon overlap as the selective criteria for putatively affected genes is a result of previous observations of gene length biases in the genome. Previously, it has been observed that genes with tissue specific expression levels have different median lengths. For example, brain specific genes have been shown to have the longest median length (37Kb), while heart specific genes have been shown to have the shortest median length (10Kb) (Raychaudhuri *et al.* 2010). Therefore, CNVs are more likely to overlap brain specific genes than other tissue specific genes, which when employing a functional enrichment analysis (see **Section 2.6**) against the whole genome will result in false positives. One possibility to remedy this is to only consider genes completely overlapped by a CNV. This however biases towards smaller genes (potentially resulting in false positives) and against longer genes (potentially resulting in false negatives). The difference in length of different tissue specific genes is reduced by calculating gene length using only the protein-coding (exonic) regions of the gene; indeed extra intronic sequence can explain the majority of the length differences between heart (median length of protein coding region = 425 residues) and brain (median length of protein coding region = 375 residues) specific genes (Webber 2011). Consequently, submitting genes for analysis only when the CNV overlaps at least one exon will counteract many of the gene length biases.

2.4 Non-protein coding gene annotations

In addition to protein coding genes, I identified other genomic elements affected by CNVs, namely micro-RNAs, long intergenic non-coding RNAs, genomic repeats and conserved non-coding elements.

MicroRNAs

MicroRNAs (miRNAs) are short ribonucleic acids that act as post-translational regulators. They bind to complementary sequences found in mRNA, targeting the mRNA for repression or degradation. A list of 718 miRNAs mapped to the human reference genome hg18, was downloaded from the UCSC genome browser (Kent *et al.* 2002; Griffiths-Jones 2006; Meyer *et al.* 2012). The given genomic locations were obtained from miRBase, and calculated using wublastn requiring 100% sequence identity (Griffiths-Jones *et al.* 2008).

Conserved Non-Coding Elements

Conserved non-coding elements (CNEs) are DNA sequences that have no protein coding function and yet are evolutionarily conserved. CNEs are interesting due to their potential to regulate gene transcription and translation. I obtained 7025 conserved non-coding elements mapped to the human reference hg19, from Greg Elgar's lab at the MIMR (personal communication; http://www.nimr.mrc.ac.uk/research/greg-elgar/). The set of CNEs was generated by comparing the *Fugu rubripes* and the human genome using megaBLAST, while masking both genomes for repeats, protein coding and RNA coding regions (Edwards *et al.* 2006). Overlapping CNEs in the original data set were merged to form the final set of 6839 CNEs.

Long intergenic non-coding RNAs

Long intergenic non-coding RNAs (LincRNAs) are non-coding, intergenic sequences of DNA greater than 200bp in length. LincRNAs mapped to hg19 were obtained from the UCSC genome browser (Meyer *et al.* 2012). The transcripts originate from the human body map data (Cabili *et al.* 2011), and were generated through the integration of pre-existing annotation data and *de novo* assembled transcripts from 4 billion RNA seq reads.

RNA Genes

RNA genes are functional RNA molecules that are not translated into a protein. RNA genes mapped to hg18 were downloaded from the UCSC genome browser (Mourelatos *et al.* 2002). This list comprises transfer RNAs, ribosomal RNAs, small cytoplasmic RNAs, small nuclear RNAs, small nucleolar RNAs, miRNAs and mitochondrial tRNA-derived pseudogenes.

Genomic repeats

Genomic repeats mapped to hg18 comprising short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeat elements (LTRs), DNA repeats, micro-satellites, low complexity repeats, satellite repeats, and RNA repeats were downloaded from the UCSC genome browser (Meyer *et al.* 2012). All repeats were determined using the RepeatMasker program (Jurka *et al.* 2005).

2.5 Functional Genomic Resources

Chapters 4, 5, 6 and **7** describe work associating copy number variants with a patient's developmental abnormality. Within these CNVs I attempt to identify candidate genes/candidate genomic elements whose copy number change contributes to the patient's phenotype. To achieve this, I make use of multiple functional genomics resources (**Table**

2.8). Using these functional resources to annotate protein coding and non-coding elements overlapping a patient's CNV, I investigate sets of CNVs for significant enrichments of annotations associated with the genes or non-coding elements.

Functional Genomics Resource	Description
OMIM	• A database containing human genes and their associated human phenotypes • Contains genic annotations for all currently known mendelian disorders
Mouse genome informatics database	• Mouse phenotypes resulting from the determined disruption of genes in the mouse. • The mouse phenotypes are described using terms from the Mouse Phenotype Ontology
Gene ontology	• A controlled vocabulary describing gene products and functions • Annotations are derived using computational and experimental approaches
KEGG	• Describes genes and their associated biological pathways
Gene expression data	• mRNA tissue expression levels obtained using Affymetrix chips
HMDD	• Human miRNA disease database associating mRNAs and human disease
MiRTarBase	• Database of miRNAs and their protein coding targets

Table 2.8: Descriptions of functional genomics resources employed in this thesis

Online Mendelian Inheritance in Man

The Online Mendelian Inheritance in Man (OMIM) is a database of human genes and their associated human disorders developed and maintained at John Hopkins University (McKusick 1998). OMIM contains phenotype information for all known mendelian disorders and their associated human genes. Each disease and gene is assigned a MIM code, where the first number in the sequence determines the mode of inheritance for example, dominant or recessive. Each of the entries within OMIM also contains a written summary of the currently known information about the gene and human phenotype and

links to the biomedical literature from where the information was obtained. I downloaded human disease phenotype data associated with 2154 Ensembl gene IDs from OMIM.

The Mouse Genome Informatics Resource

The Mouse Genome Informatics resource (MGI) provides mouse phenotype descriptions obtained from published experiments that examine the determined disruption of genes in the mouse (Eppig JT 2007; Smith and Eppig 2009). The phenotypes resulting from the experimental disruptions of mouse genes are recorded in the MGI using annotations derived from the Mammalian Phenotype Ontology (MPO) (Smith and Eppig 2009). In total, the MPO consists of 5,283 phenotype terms, organised under one or more of the 33 overarching mouse phenotype terms, each with multiple levels of finer phenotypic terms beneath them. Within the database, each gene is annotated with the most specific observed phenotype term and the overarching mouse phenotype term under which the specific term is described. For each analysis that made use of the MGI, I assigned all intermediate phenotype terms between the overarching and finest mouse phenotype to a gene. I only considered mouse phenotype terms associated with at least 1% of all genes associated with their overarching phenotype category. During the enrichment analysis this reduces uninformative results from the mouse phenotypes associated with a very low number of genes, as well as the number of tests (see **Chapters 4** and **5**). Using 1:1 gene orthology relationships between mouse and human genes, I mapped 5,283 MGI phenotype terms to 5,671 human genes.

Using the MGI resource I defined a set of 1226 haploinsufficient genes and 78 haplosufficient genes. Genes were deemed haploinsufficient if any mouse phenotype data were recorded after a heterozygous deletion of the gene. Genes were termed haplosufficient

if the disrupted gene is annotated with the overarching mouse category "normal". These data are likely to be incomplete as the majority of genes within the MGI have been subjected to homozygous knockout experiments. Also, when a "normal" phenotype is entered into the database, it is unlikely the mouse has been analysed for all 5,283 phenotype terms, and is instead relative to the individual mouse phenotype tests of interest within the original experiment.

Gene Ontology

The gene ontology (GO) standardises genes and gene product attributes across different species and databases using a controlled vocabulary of terms to describe gene characteristics. GO terms were obtained from the Gene Ontology website (http://www.geneontology.org) (Ashburner *et al.* 2000). GO terms are organized into three sub ontologies: biological process; cellular location; and molecular function. GO terms are assigned to genes using various methods including, but not limited to: experimental evidence (e.g. expression pattern); computational analysis (e.g. sequence similarity); and evidence from the literature (e.g. traceable author statement). The GO is acyclic and hierarchical; however it is possible for a child term to have multiple parent terms. I obtained 8245 GO terms that map to 18532 human genes for my analyses.

Kyoto Encylopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains descriptions for 350 biological pathways associated with 5264 human genes (http://www.genome.jp/kegg/) (Kanehisa *et al.* 2008). KEGG pathways consist of both disease pathways and metabolic pathways, for example, Alzheimer's disease and tricarboxylic acid cycle, respectively, described in a three layer hierarchy where child terms only belong to one parent term. Genes are annotated to each pathway using evidence from the literature.

Reactome

The Reactome database contains descriptions for 1326 biological pathways associated with 6436 human genes (Croft *et al.* 2011). Reactome pathways and gene annotations are manually curated by "biological experts" and cross-referenced to several other bioinformatics resources, for example, NCBI, KEGG and GO.

Gene expression data

mRNA tissue expression levels were obtained from The Genomics Institute of the Novartis Research Institute's online tissue expression database (http://symatlas.gnf.org) (Su *et al.* 2004). Expression levels for 11595 human genes were determined using Affymetrix microarray chips. To determine foetal to adult expression ratio for each gene I used the expression levels of the 4 different foetal tissues and 31 different adult tissues. During the analyses that used this data I excluded gene expression levels that were derived from cancer tissues due to the high mutational burden compared to healthy tissues.

Human MicroRNA Disease Database (HMDD)

I obtained a list of 396 microRNAs and associations to 277 different human diseases from the HMDD (http://202.38.126.151/hmdd/mirna/md/) (Lu *et al.* 2008). Within the database microRNAs and disease associations are manually curated using evidence from the literature.

miRTarBase

miRTarBase is an online genomic resource that describes the downstream protein-coding gene targets of microRNAs (http://mirtarbase.mbc.nctu.edu.tw/) (Hsu *et al.* 2011). Interactions between microRNAs and protein-coding genes are collected using evidence from the literature. Each interaction in the database is annotated with the experimental

method used to identify the interaction. I obtained a set of 283 microRNAs and 868 downstream targets identified through reporter assay and western blot experiments (see **Chapter 5**).

Evolutionary rate

In **Chapter 3** I examine the differences between 'healthy' and disease associated CNVs by annotating overlapped genes with their evolutionary rate. The evolutionary rate of a protein coding gene can be estimated by examining the nucleotide substitution rate between 1:1 orthologues from two species. Nucleotide substitutions can be either synonymous (ds; no amino acid change) or non-synonymous (dn; resulting in an amino acid change). The ratio of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site of a protein coding gene is calculated by dividing dn/ds. A result of less than one suggests that purifying selection has acted on that gene. A result greater than one indicates positive selection of non-synonymous mutations within that gene since the two species diverged. I obtained dn/ds values from the Ensmart54 database. I downloaded the dn/ds values for human to mouse, human to macaque, and human to dog 1:1 gene orthologues.

Mapping between different human genome builds

The numbering and position of bases within each the human genome sequence alters as new genome assemblies describing a more complete version are released. LiftOver, a program provided by UCSC Genome Informatics (http://hgdownload.cse.ucsc.edu/downloads.html) enables the user to convert sets of genome coordinates, such as those describing the boundaries of the copy number variants, between different assembly versions of the human genome. LiftOver uses chain data files, which describe the alignment of different genome assembly versions to map coordinates

from one genome assembly to another. In my research CNVs and functional annotations are mapped to a mixture of hg17, hg18 and hg19. Consequently, I employed LiftOver to ensure the coordinates of each dataset used in the same analysis are consistently mapped to the hg19 version of the human genome assembly.

2.6 Functional enrichment analysis

To associate functional genomic annotations with human developmental disorders, I test whether the copy number variable genes observed in a group of patients with a particular developmental disorder are annotated with a functional annotation more frequently than is expected by chance. To calculate this enrichment, the number of annotations amongst the test gene set needs to be compared against the number of annotations that would be expected by chance. The significant difference between these two values can then be assessed using the hypergeometric distribution. Additionally, if multiple functional genomic annotations are examined, a multiple testing correction needs to also be applied.

When choosing functional annotations it is important to identify potential biases within the data. For instance, the annotations within the Mouse Genome Informatics database are experimentally derived and therefore the genomic elements may have been studied due to prior evidence that they are disease relevant. Consequently, large *de novo* CNVs may possess significant enrichments of functional annotations as the genomic elements with annotations are not randomly drawn from the genome. Conversely, functional annotations derived computationally from sequence similarity may result in runs of paralogues within the genome with the same functional annotations. If a CNV overlaps these genes, a significant enrichment of that functional annotation may be produced even though the annotation is not representative of all the CNVs within the dataset that is being tested.

In order for an accurate enrichment to be identified, an appropriate number of expected annotations for each test needs to be derived. There are several different methods of calculating the number of expected annotations. The first is to compare the test set to the genomic background as a whole, comparing the number genes in the test set and the number of genes within the genome that have a specific functional annotation. This is appropriate when the test CNVs cannot be matched against a suitable set of control CNVs. For example, in **Chapters 4**, **5** and **6** I examine sets of *de novo* CNVs for enrichments of functional genomic annotations. These CNVs are both large and rare, consequently making it difficult to identify a control CNV dataset from healthy individuals. A disadvantage of using the whole genome to determine significant functional enrichments is that overlapping copy number variable regions and their disrupted genes in the test set can only be counted once and therefore power from recurrently copy number variable regions is lost.

An alternative to comparing the test set to the whole genome is to compare the test set to a set of control CNVs. These can be obtained from a healthy population, or randomly generated by randomly sampling the genome. Through using control CNVs or randomly generated CNVs, recurrently copy number variable genes in the test set can be examined. In **Chapter 7** I examine sets of inherited CNVs obtained from patients with developmental abnormalities, which I compare to a set of inherited control CNVs identified in healthy patients. To remove artefacts from the results, I need to ensure that the test and control CNVs are obtained from matched patients and matched CNV detection methods. Other potential sources of bias centre around the test set. Often, the inherited CNVs in the DECIPHER and ECARUCA test sets are present in the database because a clinician has predicted their role in disease. Consequently, there will be biases when comparing against control CNVs where there has been no initial CNV selection step. It may also be that patients have more inherited CNVs than control individuals, resulting

in the observation of significant enrichments caused by an increased burden of CNVs rather than the patient's CNVs specifically disrupting genes with a specific functional annotation.

If no matched control set can be identified, a set of randomly generated controls can also be used. Through creating random CNVs, the CNVs can be matched to the test set CNVs for size and other additional biases, such as G+C content. I used the method of creating randomly generated CNVs for my analysis in **Chapter 3**. This ensured that the different CNV sets obtained from different patients, and obtained using different CNV detection methods, were compared against CNVs that shared similar genomic biases.

In order to ensure no spurious results are generated, irrespective of which background is used, the observed and expected results under the null hypothesis needs to be examined. Within functional enrichment approaches, the null hypothesis is that there should be an even distribution of P values of each functional annotation across a range of randomly generated CNV sets. An observation of a large number of high P-values may be the result of a functional annotation assigned to a very small number of genes in the genome. Observing a large number of small P-values suggests a high number of false positive results will be observed.

Hypergeometric test

The hypergeometric test (Sokal 1995) describes the probability of obtaining x successes over n samples from a total of m successes in a total of N items. (**Equation 2.1**).

$$P = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

Equation 2.1: The hypergeometric distribution. The probability of obtaining x successes over n samples from a total of m successes in a total of N items.

In this thesis I apply the hypergeometric test to identify whether a particular type of genomic annotation is over or under-represented amongst the CNVs that I observe in my patient samples. I use the hypergeometric test for each of the gene enrichment analyses in **Chapters 4**, **5** and **6** by comparing the genes within the CNV test sets to all genes within the human genome. For example, in the Gene Ontology enrichment analysis, N is all the genes in the genome with a GO annotation, and m = the number of genes within the genome with the GO term annotation of interest. X is the number of genes with the GO annotations from sample size n (the genes overlapped by the CNVs of interest).

Fisher's Exact Test

The Fisher's exact test is used to determine whether there are significant non-random associations between two categorical samples, for example patients with disease, and healthy individuals. Thus, if we have two groups of people (healthy individuals and patients with mental retardation) and the number of copy number variable genes in these patients associated with a specific GO term, the contingency table may look like that depicted in **Table 2.9**. The probability of obtaining this distribution of numbers can then be determined (**Equation 2.2**).

	Healthy Patients	**MR Patients**	**Total**
Copy number variable genes associated with GO term	A	B	A+B
Copy number variable genes not associated with GO term	C	D	C+D
Total	A+C	B+D	A+B+C+D (N)

Table 2.9: Distribution of copy number variable genes and their association with a specific GO term observed in healthy and mental retardation patients.

$$P = \frac{\binom{A+B}{A}\binom{C+D}{C}}{\binom{N}{A+C}}$$

Figure 2.2: Fisher's exact test. The probability of observing non-random associations between two categorical variables.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (ks test) determines whether there is a difference between the distributions of two data sets. This test requires two datasets comprising continuous data and does not make any assumptions as to whether the data are normally distributed. The ks test measures the distance between the cumulative distributions of the two datasets in order to test the probability that the two datasets were drawn from the same distribution. In **Chapter 3** I use the ks test to compare the distributions of nucleotide substitution rates between several different datasets.

Multiple Testing Correction

Applying multiple tests on the same set of data simultaneously can lead to the identification of false positives. The greater the number of tests applied to a set of data, the more likely the null hypothesis will be rejected at a given confidence value just by chance. Thus, it is important to account for the number of tests being performed on a data set when determining whether a result is statistically significant.

When employing a multiple testing correction during the functional enrichment analysis (**Chapters 4, 5, 6** and **7**) and when using the Genomic Association Tester (GAT; see below) (**Chapters 3** and **5**) I use the false discovery rate (FDR) (Benjamini and Hocjberg 1995) (Storey, 2003). The FDR calculation describes the predicted number of false positives that are occurring in a set of results found to be statistically significant for example, when using the hypergeometric test. For instance, if 100 results are found to be significant at $P<0.05$, and the FDR is set to 5%, you would expect 5 of these results to be false positives.

Genomic Association Tester

The Genomic Association Tester (GAT) tests whether two sets of intervals overlap more than is expected by chance (http://code.google.com/p/genomic-association-tester/source/list). GAT accepts three main input files: (i) segments – a set of intervals whose association to annotations is to be tested, for example CNVs; (ii) annotations – various annotations within the genome for example miRNAs; (iii) workspace – the genomic region available to sample segments and annotations from for example, the whole genome.

First, GAT uses these three inputs listed in the previous paragraph to determine the proportion of segments that overlap with the annotations. Overlap can be measured using either a count of the number of annotations that overlap a segment, or a count of the number of nucleotides that are intersected by both the segment and annotation. Throughout my thesis I measure overlap by counting the number of annotations. Secondly, GAT creates sets of control segments randomly sampled from the genomes. Segments are randomly placed on the workspace using the size distribution of the original set of segments. Segments are generated until the same number of nucleotides overlaps the workspace as in the original segment sample. The proportion of sampled segments that overlap the annotations is then calculated and compared to the original test segments. The random sampling procedure is repeated for several hundred iterations, and for each sample the amount of overlap with the annotations is compared to that in the original segment sample. This enables the calculation of a P-value and false discovery rate associated with the observed over or under-representation of overlap in the original sample (Storey 2002).

In addition to the three input files discussed in the previous paragraphs, additional genomic information, for example G+C content, can be incorporated into the analysis. These genomic features are important to include in the GAT analysis if the annotations or segments have known biases. In **Chapter 3** I used GAT to examine CNVs for enrichments of several different genomic annotations. CNVs have been observed to have biases in their G+C content, and therefore are not randomly distributed throughout the genome. This needs to be accounted for when generating the randomly sampled segments within the workspace. In order to account for G+C content during the analysis, GAT can be provided with isochore data. This enables the genome workspace to be segmented into smaller regions based on the G+C distribution in the genome. GAT performs simulations for each GC bin separately. At the end, results for each bin are aggregated. Using isochores is required in my analysis of genomic features as CNVs are correlated with G+C content within the genome.

Chapter 3: Evolutionary analysis of disease associated CNVs

3.1 Abstract

Copy number variants (CNVs) contribute to a large amount of the genomic variation observed within the human population, and are therefore expected to contribute to many human phenotypic differences. Several CNVs have been strongly associated with human disease, for example, the 22q11.2 deletion and Di George syndrome. However, for many other CNVs observed in the human population it has not been possible to determine whether they contribute to a disease phenotype. In this chapter, I analysed two sets of disease associated and three sets of "benign" CNVs for evidence of genomic biases. Through these experiments I aimed to provide evidence of where CNVs are likely to arise, identify the selection pressures acting upon CNVs and reveal differences between benign and disease associated CNVs. I identify significant biases in G+C content in each of the 5 CNV sets that I examined. In addition, I identify significant enrichments of protein-coding genes, miRNAs and protein-coding genes with OMIM disease annotations amongst the disease associated CNV sets and significant depletions of these features amongst the "benign" CNV sets. I also identify genomic features that do not differ between disease associated and "benign" CNVs, namely RNA genes and repeat elements, which contradicts previous work in this area. Finally, this chapter highlights the variation of genomic properties amongst different disease associated and "benign" CNV sets that have been called with different CNV detection methods. These results indicate that different CNV detection methods are biased as to which kind of CNVs they are able to detect. This

observation needs to be taken into account when comparing sets of CNVs or generalising results from one subset of CNVs to another.

3.2 Introduction

Copy number variants (CNVs) are a large contributor to the differences observed between two individuals' genomes (Redon *et al.* 2006). Consequently, CNVs may contribute to much of the healthy and disease phenotypic differences observed within a population (see **Chapter 1**). Through the analysis of the genomic biases that occur within CNVs, it will be possible to further our knowledge of where CNVs are likely to arise and the selective pressures that act upon them. By examining and identifying differences between benign and disease associated CNVs it may be possible to use this information to predict the pathogenicity of an unknown CNV.

It has previously been shown that copy number variation does not occur uniformly within the genome (Iafrate *et al.* 2004). Several publications have noted further biases within benign CNVs. Nguyen *et al.*, 2006 examined a set of 627 CNVs and concluded that positive selection has acted upon CNVs in the human population. They observed an increased gene density and an elevated rate of protein evolution within the CNV regions that were analysed (Nguyen *et al.* 2006). In 2009, Nguyen *et al.* analysed additional sets of CNVs, reconsidering their previous conclusions, and suggested that CNVs are instead retained in the human population due to reduced purifying selection. This conclusion was reached from the observation that benign CNVs are enriched in non-essential genes, which in turn allows for the fixation of CNVs with a high gene density in the human genome.

Much research has also concentrated on identifying biases within disease associated CNVs. Webber *et al.,* 2009 examined *de novo* CNVs identified in patients with intellectual

disability, identifying an enrichment of genes associated with KEGG neurodegenerative pathways and the mouse model phenotypes *abnormal axon* and *dopaminergic neuron morphologies* (Webber *et al.* 2009). These observed biases were consequently exploited to develop a classifier that accurately distinguishes benign CNVs from those underlying mental retardation (Hehir-Kwa *et al.* 2010).

In this chapter, I examine the genomic biases observed between two sets of disease associated CNVs and three sets of benign CNVs in order to identify distinguishing features of disease associated and benign CNVs. Testing the null hypothesis that CNVs are randomly distributed across the genome, I examine each of the five sets of CNVs for an enrichment over that expected by chance of GC content, protein-coding genes, OMIM annotations, haplosufficient genes, haploinsufficient genes, RNA genes, miRNA genes, miRNA target sites, repeat elements, long interspersed nuclear elements (LINEs) or short interspersed nuclear elements (SINEs). In addition, I examine whether disease or benign associated CNVs are observed more frequently in centromeric or telomeric regions than expected by chance. Lastly, I examine the nucleotide substitution rates between human and mouse, human and dog and human and macaque of the genes overlapped by the CNVs.

My results reveal that disease associated CNVs are enriched in protein-coding and OMIM annotated genes, in comparison to benign CNVs where I observe significant depletions. I observe a significant enrichment of miRNA loci amongst disease associated CNVs and a significant depletion of miRNA target sites amongst the benign CNV sets. Additionally, I observe a depletion of disease associated CNVs and an enrichment of benign CNVs in centromeric regions within the genome. I also observe (when employing a 5Mb size cut-off) a significant reduction in dn/ds (the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per

synonymous sites) values amongst genes overlapping disease associated CNVs in comparison to the whole genome and three benign CNV sets.

The work in this chapter also investigates how G+C biases within the genome can affect the observed enrichments of genomic elements within CNVs through examining random CNV sets that account for the G+C content bias in the test set and those that do not. Furthermore, the control CNV sets used in this chapter have been identified using different array platforms, allowing the biases between different platforms to be analysed through a comparison of the results between the three sets. In addition, the disease CNV sets are large and consist of patients presenting with a wide range of developmental disorders, expanding on previous studies that have concentrated on CNVs associated with more specific disorders.

3.3 Methods

CNVs and patient phenotypes

In this chapter, I make use of two sets of *de novo* CNVs associated with developmental disorders and three sets of "control" CNVs observed in seemingly healthy individuals. The two sets of *de novo* disease associated CNVs come from the DECIPHER (Firth *et al.* 2009) and ECARUCA databases (Feenstra *et al.* 2006) and are formed of 626 (median size = 2.2Mb) and 1143 (median size = 17.3Mb) CNVs, respectively. The patients from the DECIPHER and ECARUCA databases present with one or more developmental abnormalities described using terms from the London Medical Database ontology (Fryns and de Ravel 2002). The control CNV sets are from Shaikh *et al.* (Shaikh *et al.* 2009), AGP (Pinto *et al.* 2010) and Nijmegen *et al.* (Nguyen *et al.* 2008) and comprise 54462 (median size = 8.1Kb), 2537 (median size = 96.5Kb) and 361 (median size = 0.74Mb)

CNVs, respectively. The Shaikh *et al.* set was also filtered to form a set of common control CNVs (CNVs observed in more than one individual). Each CNV set was further subdivided by the direction of copy number, i.e. whether a loss or a gain.

Enrichments of genomic elements within CNVs

To examine whether disease associated and benign CNV sets are enriched in various genomic elements within the genome I obtained a set of GC rich regions, protein coding genes, OMIM annotations, miRNA genes, miRNA target sites, RNA genes and repeat sequences each mapped to the hg18 genome assembly from the UCSC genome browser (see **Chapter 2**).

For each CNV dataset I examined whether the CNVs were enriched in these genomic elements using the Genomic Association Tester (GAT) (see **Chapter 2**). To prevent bias during the random CNV sampling procedure, the genomic workspace was split based on the GC content distribution of the genome. Enrichments were deemed significant if the reported P-value was <0.05.

Haplosufficient and haploinsufficient genes

I obtained a set of 1226 haploinsufficient and 77 haplosufficient genes from the MGI database (Smith and Eppig 2009). Haploinsufficient genes were defined as genes described within the MGI resource annotated with one or more mouse phenotypes (apart from *normal*) when they are hemizygously disrupted. Haplosufficient genes are defined as those genes within the MGI database annotated with a *normal* mouse phenotype when hemizygously disrupted. I examined each set of CNVs for an enrichment of these two sets of genes using the whole genome as a background. Statistical significance was achieved when $P<0.05$.

Genomic location of CNVs

I downloaded the hg18 coordinates of the centromeres and telomeres for each human chromosome from the UCSC genome browser. I used GAT to examine whether the centromeric and telomeric regions are enriched in CNVs. I defined these regions as within 10Mb of the hg18 coordinates obtained from the UCSC genome browser.

dn/ds of genes within CNVs

In order to examine the selective pressures acting on copy number variable genes in patients and healthy individuals I compared the evolutionary rates of genes overlapping *de novo* CNVs observed in patients with developmental disorders, CNVs observed in healthy individuals and the genome as a whole. I obtained the dn and ds values of 1:1 orthologues of human and mouse, human and macaque and human and dog from the ensmart 54 database. I compared the median synonymous nucleotide substitution rate (ds) and the median evolutionary rate (dn/ds) of protein coding genes overlapping CNVs in each of the five CNV sets. I compared three different human:vertebrate ds and dn/ds values in order to identify if different selective pressures are acting on the 5 CNV sets. Macaque has the smallest evolutionary distance to human and is therefore the better comparison for identifying human specific selective pressures. However, due to the recent common origin of the two species, the number of synonymous mutations will often be low which in turn will make it difficult to calculate the dn/ds ratio accurately.

3.4 Results

I sought to identify genomic biases that differed between disease associated CNVs and benign CNVs. To achieve this I examined 2 sets of disease associated CNVs and three sets of benign CNVs for enrichments of a range of different genomic features.

Disease associated and benign CNVs have differing enrichments of genomic elements

Two sets of *de novo* CNVs obtained from patients with developmental disorders (DECIPHER and ECARUCA; (Feenstra *et al.* 2006; Firth *et al.* 2009)) and three sets of control CNVs (Shaikh *et al.*, AGP and Nijmegen *et al*; (Nguyen *et al.* 2008; Shaikh *et al.* 2009; Pinto *et al.* 2010)) from healthy individuals were examined for enrichments of G+C content, protein coding genes, protein coding genes associated with disease in OMIM, RNA genes, miRNAs, miRNA target sites and repeat elements (split by all, and LINEs and SINEs separately) (**Figure 3.1A-E**). The Shaikh *et al.* set was further split to create a set of "common" control CNVs, consisting of CNVs observed in more than one individual. Additionally, for each of the CNV sets the duplication and deletion CNVs were separated into "gain" and "loss" sets, respectively. In the following figures, the CNV sets Shaikh *et al.*, Shaikh common, AGP and Nijmegen *et al.* are referred to as Ctl-S, Ctl-SC, Ctl-A, Ctl-N, respectively.

The two disease associated CNV sets and the Shaikh *et al.* and Nijmegen *et al.* control CNV sets are significantly enriched in genomic regions with a high GC content (**Figure 3.1A**). The AGP loss set is significantly depleted in GC rich regions, whereas the Shaikh-common set appears to contain a comparable amount of GC rich regions observed to the genome sampled at random. The differences in GC content compared to what is expected by chance need to be taken into account when examining CNVs for enrichments of genomic elements that also correlate with percentage GC content in the genome (see **Discussion**). Consequently, for all of the following analyses I compare each CNV set to randomly generated CNVs drawn from the genome with a comparable GC content.

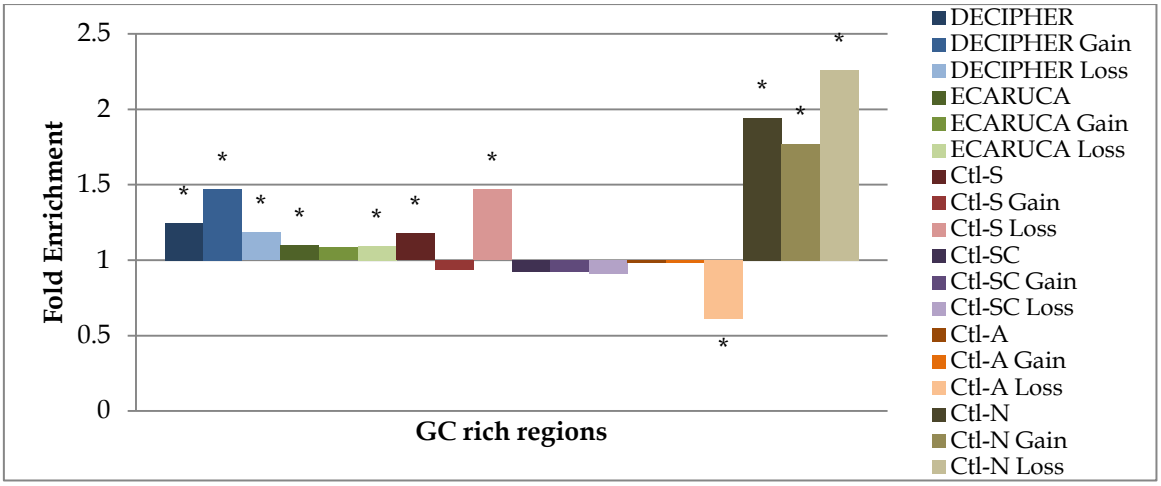


Figure 3.1A: Fold enrichments over that expected by chance of GC rich regions amongst the 5 CNV datasets. Enrichments, compared to randomly generated CNV regions, are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

Several of the CNV datasets show a small yet significant enrichment of protein-coding genes and/or OMIM-annotated genes over that expected by chance (**Figure 3.1B**). Indeed, within the disease CNV sets, the fold enrichments, although significant, are very close to one. Within the CNVs observed in healthy individuals, I observe a depletion in both the amount of protein coding genes and OMIM annotated genes over that expected by chance.

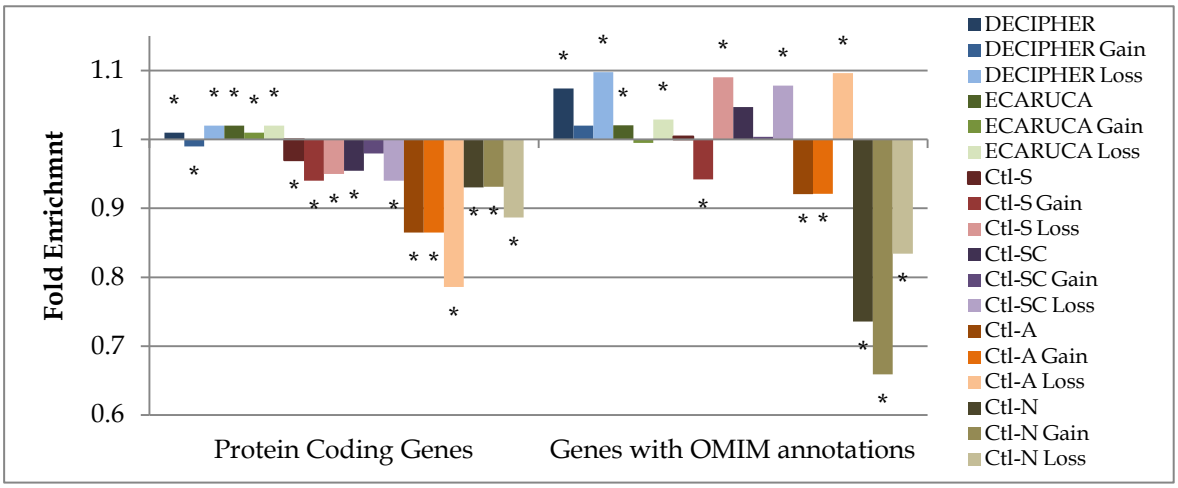


Figure 3.1B: Fold enrichments over that expected by chance of genes and genes with OMIM annotations amongst the 5 CNV datasets. Enrichments, compared to randomly generated CNV regions, are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

The above approach measures gene density by assessing the number of genes that are overlapped by each of the CNV sets. An alternative approach is to investigate the number of CNVs that overlap protein coding genes. These two complementary methods reveal whether CNVs contain more genes than expected and whether the number of CNVs that overlap genes is more or less than expected. I examined each of the 5 datasets for an enrichment of CNVs that overlap a protein-coding gene (**Figure 3.1:C**). I observe depletions of CNVs overlapping protein coding genes over the number expected by chance (see **Discussion**).

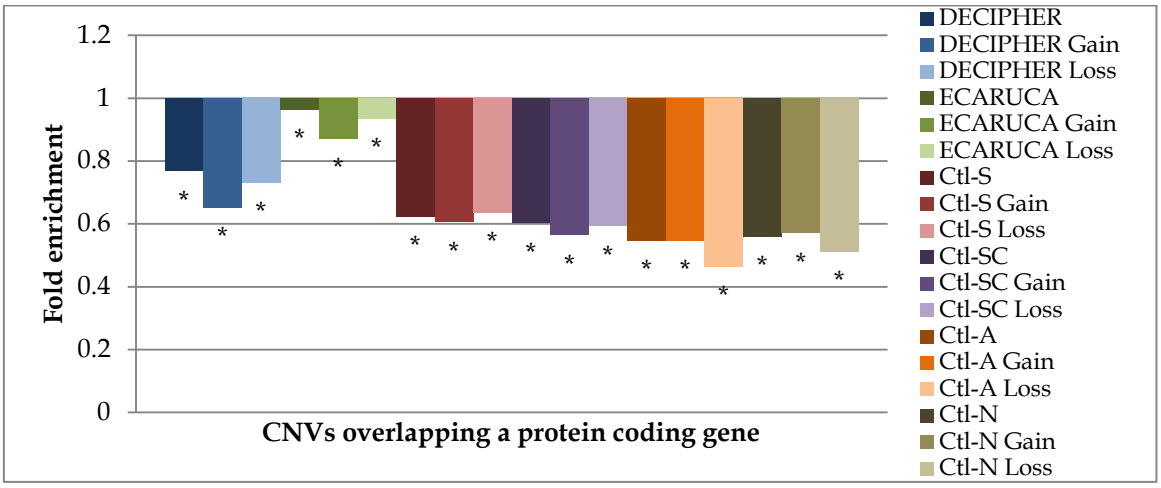


Figure 3.1C: Fold enrichments over that expected by chance of CNVs overlapping protein-coding genes. Enrichments, compared to randomly generated CNV regions, are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

Of the 5 CNV datasets, 4 show either a significant enrichment or depletion of RNA genes (**Figure 3.1D**). The DECIPHER – gain set shows a very small significant enrichment of RNA genes, as does the Shaikh *et al.* – gain CNV set. Conversely, the AGP and Nijmegen *et al.* sets are depleted in RNA genes compared to the randomisations. However, this observation is only significant in the AGP-loss set. RNA genes are known to be clustered in the genome and indeed, the significant enrichments and depletions are due to an "all or nothing" overlap of the RNA clusters by the CNVs.

The DECIPHER CNVs are enriched in miRNAs, with the all and gain results reaching significance ($P<0.05$) (**Figure 3.1D**). Depletions of miRNAs are observed amongst the majority of the control sets, with the Shaikh *et al.*-gain, AGP-loss and Nijmegen *et al.*-all sets reaching significance. However, within the AGP-all and AGP-gain sets, an enrichment of miRNAs is observed, although these enrichments were not found to be statistically significant.

Within the disease associated CNV sets there are small yet significant enrichments of miRNA target sites (**Figure 3.1D**). The Shaikh *et al.* and AGP controls show a depletion of miRNA target sites over what is expected by chance. This result may be dependent on the depletion of protein coding genes within these sets (see **Discussion**). The Nijmegen *et al.* sets vary considerably in comparison to the other control set, with large significant enrichments observed amongst the Nijmegen *et al.*-all and Nijmegen *et al.*-loss CNVs. I also examined each CNV set for enrichments of miRNA target sites within genes known to be highly expressed in the brain. Each of the CNV sets showed depletions of miRNA target sites of brain expressed genes, however the CNVs from control individuals had larger depletions than the CNVs observed in patients from the DECIPHER and ECARUCA databases.

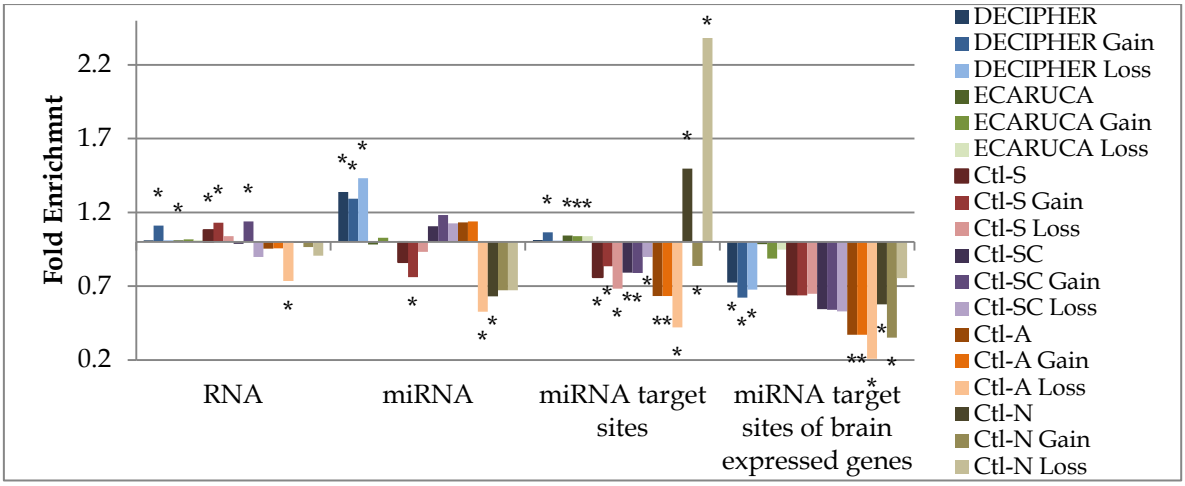


Figure 3.1D: Fold enrichments over that expected by chance of RNAs, miRNAs and miRNA target sites within the 5 CNV datasets. Enrichments, compared to randomly generated CNV regions, are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

Within the 5 sets of CNVs I observe very small enrichments and depletions of repeat elements (**Figure 3.1E**). The DECIPHER, ECARUCA, Shaikh *et al.* and AGP sets all show enrichments, however these are all less than 1.05 fold. The Nijmegen *et al.* set is depleted in repeat elements, although again by only a very small amount (>0.9 fold). The enrichments of LINEs amongst the 5 CNV datasets mirror that of the enrichments of all repeat elements amongst the CNVs (**Figure 3.1E**). The Nijmegen *et al.* set has a significant depletion of LINEs, whereas within the other 4 sets small enrichments are observed. Apart from the ECARUCA CNV sets, the CNVs show a depletion of SINEs over that expected by chance (**Figure 3.1E**). Many of these results reach significance ($P<0.05$), and again it is the Nijmegen *et al.* set where the largest depletions are observed.

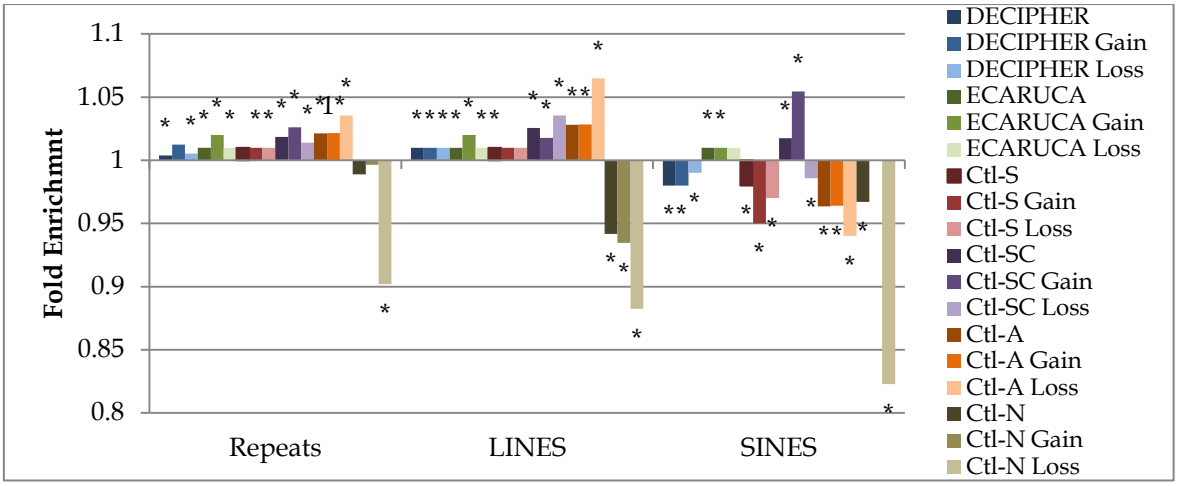


Figure 3.1E: Fold enrichments over that expected by chance of repeat elements, LINEs and SINEs within the 5 CNV datasets. Enrichments are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

Correcting for GC biases within the genome alters the observed enrichments of genomic elements within disease and benign CNVs.

Previously I have shown that disease associated and benign CNVs have significant enrichments and depletions in their GC content. Consequently, when examining CNVs for enrichments of genomic elements, I decided to account for GC bias when generating the random CNV sets. To examine the consequences of not accounting for GC, I repeated the GAT analysis without splitting the genomic workspace by GC content (**Table 3.1**). When not accounting for GC the observed enrichments increase for those genomic elements whose prevalence is closely associated with GC (e.g. protein-coding genes).

Dataset	Initial Result	Without accounting for GC bias
Protein Coding Genes - DECIPHER	1.01	1.08*
ECARUCA	1.02	1.06*
Ctl-S	0.97	0.97
Ctl-SC	0.95	1.00
Ctl-A	0.86	0.88*
Ctl-N	0.93	1.03
OMIM Genes – DECIPHER	1.07*	1.14*
ECARUCA	1.02*	1.05*

Ctl-S	1.00	0.97
Ctl-SC	1.04*	1.09
Ctl-A	0.92*	0.95
Ctl-N	0.73*	0.82
RNA genes - DECIPHER	1.01*	1.07
ECARUCA	1.01	1.05*
Ctl-S	1.08*	1.14
Ctl-SC	1.09*	1.02
Ctl-A	0.95	0.94
Ctl-N	0.99	1.05
miRNA genes-DECIPHER	1.33*	1.57*
ECARUCA	0.98	1.05
Ctl-S	0.86	0.76
Ctl-SC	1.10	1.15
Ctl-A	1.13	1.15
Ctl-N	0.63*	0.86
Target genes-DECIPHER	1.01*	1.08
ECARUCA	1.04*	1.09*
Ctl-S	0.76*	0.88
Ctl-SC	0.79*	0.83
Ctl-A	0.63*	0.65*
Ctl-N	1.49*	1.58*
Repeats-DECIPHER	1.00*	1.03
ECARUCA	1.01*	1.03*
Ctl-S	1.01	1.04*
Ctl-SC	1.01*	1.06*
Ctl-A	1.02	1.05
Ctl-N	0.98*	1.02
LINEs-DECIPHER	1.01*	1.04*
ECARUCA	1.01*	1.03*
Ctl-S	1.01*	1.03*
Ctl-SC	1.02*	1.07*
Ctl-A	1.02*	1.07*
Ctl-N	0.94*	0.94
SINEs-DECIPHER	0.98*	1.04*
ECARUCA	1.01*	1.04*
Ctl-S	0.98*	0.98
Ctl-SC	1.01*	1.05*
Ctl-A	0.96*	0.96
Ctl-N	0.96*	1.05

Table 3.1: Fold enrichments of genomic elements amongst CNVs in comparison to randomly generated CNV sets. Significant results are indicated by asterisks.

Disease and control CNVs are depleted in haplosufficient and haploinsufficient genes

The five CNV datasets are significantly depleted in both haploinsufficient and haplosufficient genes (**Figure 3.2**). The difference in the results between the disease and control CNVs is small, which is surprising given the expectation that for a CNV to cause disease it must disrupt a gene(s) that requires a diploid copy for normal healthy function.

This may be due to the large size of the *de novo* disease associated CNVs which may overlap several haplosufficient genes as well as those genes that are disease causing (**see Discussion**).

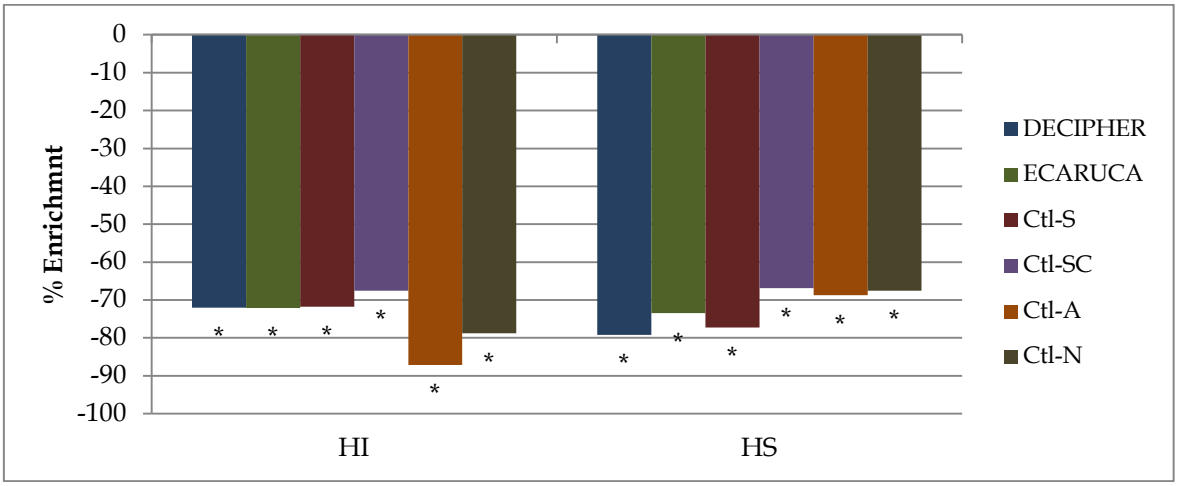


Figure 3.2: Percentage enrichments over that expected by chance of haploinsufficient (HI) and haplosufficient (HS) protein coding genes within the 5 CNV datasets. Percentage enrichments are given for each dataset and an asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

CNV distribution across the genome

The chromosomal distribution of CNVs was examined to determine whether CNVs are uniformly distributed across the genome and whether CNVs associated with disease and CNVs associated with healthy individuals have differences in their genomic distributions. Using GAT, I examined whether the subtelomeric and pericentromeric regions (10Mb) of chromosomes are significantly enriched in CNVs from each of the 5 CNV datasets (**Figure 3.3**). I obtained the genomic coordinates of the centromeres and telomeres from the hg18 genome assembly build and defined the subtelomeric and pericentromeric regions as the surrounding 10Mb of these coordinates (see **Methods**).

The two disease associated CNV sets are significantly depleted in pericentromeric regions ($P<0.05$), whereas the three control CNV sets are significantly enriched in the pericentromeric regions (**Figure 3.3**). However, when examining subtelomeric regions all of the sets, apart from the Shaikh-common set have significant enrichments.

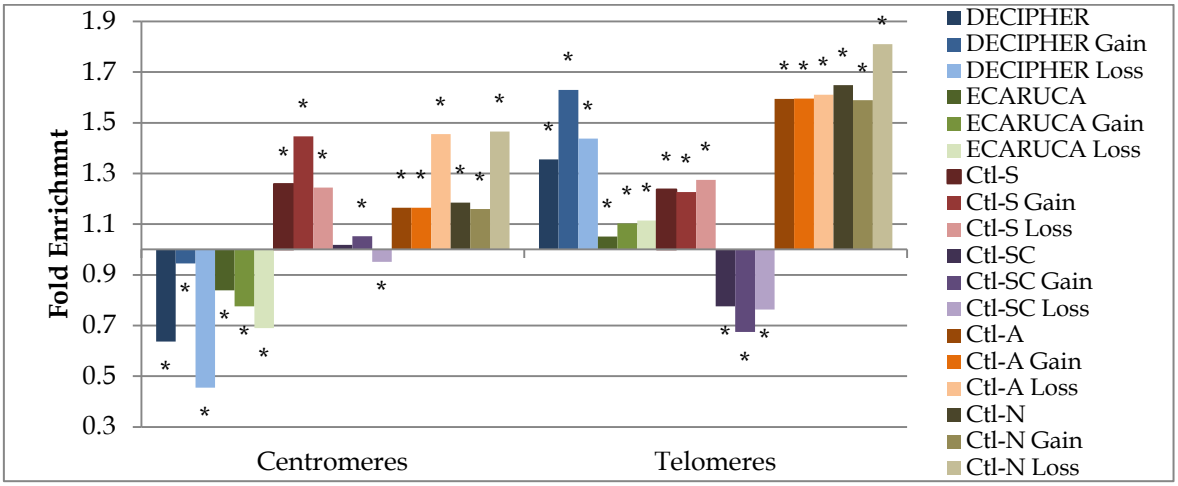


Figure 3.3: Fold enrichments over that expected by chance of subtelomeric and pericentromeric regions within the 5 CNV datasets. Enrichments are given for each dataset and further subdivisions of each dataset by copy number directions (gain/loss). An asterisk indicates the fold enrichment is significant at a P value of less than 0.05.

Nucleotide substitution rates of CNV genes

Previous work identified nucleotide substitution rates as an important feature in classifying CNVs as disease causing or benign (Hehir-Kwa *et al.* 2010). To examine differences of nucleotide substitution rates of protein coding genes amongst my two disease associated CNV sets and three benign CNV sets, I obtained the ds values of all 1:1 orthologues between human and mouse, human and macaque and human and dog from the Ensmart 54 database. For each of the 5 CNV sets (and a subset of the two disease CNV sets formed using a 5Mb cut-off) I calculated the median ds values of the overlapping protein coding genes (**Table 3.1A-C**).

Human: mouse	All	DECIPHER	(<5Mb) *	ECARUCA *	(<5Mb) *	Ctl-S	Ctl-A *	Ctl-N *
Min	0.01	0.01	0.05	0.01	0.15	0.09	0.08	0.15
Q1	0.46	0.46	0.47	0.46	0.51	0.47	0.48	0.57
Med	0.58	0.59	0.61	0.57	0.65	0.58	0.59	0.78
Q3	0.75	0.76	0.81	0.73	0.83	0.75	0.77	0.99
Max	1.49	1.49	1.49	1.49	1.40	1.49	1.49	1.49

Table 3.1A: The minimum, quartile, median and maximum ds values for all 1:1 human:mouse protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference (using a two sample Kolmogorov-Sminov test) from the all set is indicated with an asterisk.

Human:dog	All	DECIPHER *	(<5Mb) *	ECARUCA *	(<5Mb) *	Ctl-S *	Ctl-A *	Ctl-N *
Min	0.02	0.03	0.03	0.02	0.08	0.02	0.04	0.08
Q1	0.28	0.29	0.31	0.28	0.35	0.29	0.30	0.40
Med	0.39	0.40	0.43	0.38	0.47	0.40	0.40	0.61
Q3	0.55	0.57	0.63	0.54	0.63	0.57	0.59	0.85
Max	1.49	1.49	11.4	1.49	1.49	1.49	1.49	1.49

Table 3.1B: The minimum, quartile, median and maximum ds values for all 1:1 human:dog protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference (using a two sample Kolmogorov-Sminov test) from the all set is indicated with an asterisk.

Human: macaque	All	DECIPHER	(<5Mb) *	ECAURCA *	(<5Mb) *	Ctl-S *	Ctl-A *	Ctl-N *
Min	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Q1	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.07
Med	0.07	0.07	0.08	0.07	0.10	0.08	0.08	0.13
Q3	0.12	0.12	0.14	0.11	0.15	0.12	0.13	0.19
Max	1.42	1.39	1.50	1.42	1.13	1.42	1.31	1.42

Table 3.1C: The minimum, quartile, median and maximum ds values for all 1:1 human:macque protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference (using a two sample Kolmogorov-Sminov test) ($P<0.05$) from the all set is indicated with an asterisk.

In each of the three tests the ECARUCA set has the lowest median ds value of the 5 sets. The ECARUCA ds value is lower than the median ds for all 1:1 protein coding orthologues in the genome. When comparing human:macaque and human:dog, the other 4 CNV sets have a higher median ds value than the median observed across all 1:1 orthologues. However, the results differ when the ds of human:mouse is examined, as the Shaikh set has a lower ds than the median ds of all 1:1 orthologues.

When comparing the ds values between disease-associated and benign CNVs, the AGP and Nijmegen sets have a higher median ds than the DECIPHER and ECARUCA sets in all three tests, and the Shaikh set has a higher median ds in the human to macaque comparison. However, the difference in median ds between the 4 sets, excluding the Nijmegen set, is very small, which contradicts the findings in the Hehir-Kwa *et al.* paper that ds is a suitable feature for a CNV classifier.

Evolutionary rates of CNV genes

I wished to examine whether the evolutionary rates of copy number variable genes differed substantially between disease associated and benign CNVs. I analysed the dn/ds (ratio of non-synonymous to synonymous substitution rate) of the 5 CNV sets and a subset of the two disease CNV sets formed using a 5Mb cut-off (**Table 3.2A-C**). Again, to achieve this I used the values observed between 1:1 orthologues of human and mouse, human and macaque and human and dog.

Human: mouse	All	DECIPHER	(<5Mb) *	ECARUCA	(<5Mb) *	Clt-S *	Ctl-A *	Ctl-N *
Min	0	0	0	0	0	0	0	0
Q1	0.04	0.04	0.04	0.04	0.03	0.04	0.05	0.04
Med	0.09	0.09	0.09	0.09	0.07	0.09	0.10	0.09
Q3	0.17	0.17	0.17	0.17	0.14	0.17	0.18	0.16
Max	0.40	0.40	0.37	0.40	0.53	0.40	0.40	0.40

Table 3.2A: The minimum, quartile, median and maximum dn/ds values for all 1:1 human:mouse protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference from the all set is indicated with an asterisk.

Human: dog	All	DECIPHER *	(<5Mb) *	ECARUCA *	(<5Mb) *	Ctl-S *	Ctl-A *	Ctl-N *
Min	0	0	0	0	0	0	0	0
Q1	0.05	0.05	0.05	0.05	0.04	0.05	0.06	0.05
Med	0.11	0.11	0.11	0.11	0.09	0.11	0.12	0.10
Q3	0.21	0.20	0.19	0.21	0.17	0.20	0.21	0.18
Max	0.47	0.47	0.43	0.47	0.36	0.47	0.47	0.47

Table 3.2B: The minimum, quartile, median and maximum dn/ds values for all 1:1 human:dog protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference from the all set is indicated with an asterisk.

Human: macaque	All	DECIPHER	(<5Mb) *	ECARUCA	(<5Mb) *	Ctl-S	Ctl-A *	Ctl-N *
Min	0	0	0	0	0	0	0	0
Q1	0.07	0.07	0.07	0.07	0.05	0.07	0.08	0.07
Med	0.18	0.18	0.17	0.18	0.14	0.18	0.2	0.17
Q3	0.35	0.34	0.32	0.34	0.29	0.34	0.35	0.32
Max	0.76	0.76	0.73	0.76	0.64	0.76	0.76	0.75

Table 3.2C: The minimum, quartile, median and maximum dn/ds values for all 1:1 human:macque protein coding genes, and the genes overlapping the 5 different CNV datasets. Significant difference from the all set is indicated with an asterisk.

When comparing mouse and human dn/ds values, the two disease sets have a higher dn/ds than all human:mouse 1:1 orthologues (**Table 3.2A**). The Shaikh *et al.* and Nijmegen *et al.* set have a lower dn/ds value than that observed in the disease sets and the all set, however, the AGP set has a higher dn/ds. When examining human:macaque dn/ds values the results differ (**Table 3.2B**). Here the DECIPHER, Shaikh *et al.* and Nijmegen *et al.* sets have a lower dn/ds value than the median dn/ds value of all 1:1 orthologues. The ECARUCA and Nijmegen *et al.* sets have a higher dn/ds value than the 1:1 orthologues. Comparing human and dog orthologues reveals similar results to comparing the human and mouse dn/ds values. The disease sets have a higher dn/ds value than the all set and the Shaikh *et al.* and Nijmegen *et al.* set to have a lower dn/ds value than all 1:1 orthologues. Again here, the AGP set has the highest dn/ds value of the 5 sets. When employing a 5Mb CNV size cut-off in the two disease associated CNV sets, the median dn/ds values decrease. In each case the median dn/ds values for the 5Mb size cut-off sets are below that of the all set and the majority of the control sets (see **Discussion**).

3.5 Discussion

This chapter examined the genomic properties of five CNV sets: two sets observed in patients with developmental disorders and three sets observed in healthy individuals.

The 5 CNV sets were found to have biases in their GC content. The two disease sets and the Shaikh *et al.* and Nijmegen *et al.* controls have significant enrichments of GC rich regions within their respective CNVs. Conversely, the AGP and Shaikh-common sets are depleted in GC rich regions. Consequently, GC will become a confounding factor when examining CNV regions for significant enrichments of genomic elements, for example LINEs and SINEs, that segregate with GC rich regions in the genome, and therefore, in all further analyses discussed below, I accounted for GC when generating the random CNV sets for each test CNV set.

My analysis reveals a small significant enrichment of protein-coding genes amongst disease associated CNVs and significant depletions within the three benign CNV sets. Assuming the CNVs observed in patients with developmental disorders are indeed responsible for the patients' phenotypes, it seems logical that these CNVs would disrupt more protein coding genes than those CNVs that do not cause disease. However, when examining CNVs that intersect a protein-coding gene I identify significant depletions, suggesting that either not all the disease associated CNVs are the underlying cause of the patient's developmental abnormality or some patient's disorders are the result of CNV disrupting non-protein coding genomic elements. The CNVs associated with disease are enriched in genes known to be associated with disease in OMIM, whereas the benign CNVs are depleted in known disease associated genes, indicating that the amount of disrupted protein coding gene product and amount associated with OMIM annotations are potential predictors of the pathogenicity of a CNV. Past research has revealed differing conclusions on the amount of

enrichments or depletions of protein coding genes within CNVs. In 2006, Nguyen examined CNVs observed in healthy individuals and observed an increased gene density (Nguyen *et al.* 2006). This analysis did not account for GC biases within the CNVs due to previous observations that the CNVs did not have an elevated GC content. In 2008, Nguyen compared 4 additional "control" CNV sets while accounting for GC bias and found the enrichments in three sets can be explained by their elevated GC content (Nguyen *et al.* 2008). In this analysis I account for GC by sampling CNVs for GC bins at the same observed frequency in the test sample. It has been observed that larger CNVs are more prone to having a higher GC content and therefore overlap more protein coding genes (which are also more prone to be found in GC rich regions) (Antequera and Bird 1993). As discussed in **Chapter 1**, different CNV detection platforms tend to identify CNVs of different sizes and indeed the five CNV datasets were identified using different methods and are of different sizes. Of the three control sets the Shaikh *et al.* CNVs are the smallest at 8.1Kb in median length, the AGP are 10 fold bigger at 96.5 Kb and the Nijmegen *et al.* a further ~10 fold bigger at 740kb. However, by individually accounting for GC content for each set, I observed a similar amount of protein coding gene depletion in each set.

RNA genes are not largely under or over-represented amongst any of the five CNV sets, however when miRNAs are examined separately a significant enrichment is observed amongst the disease associated DECIPHER set and significant depletions are observed amongst the benign Shaikh *et al.*, AGP and Nijmegen *et al.* sets. Individual miRNAs are known to regulate multiple (>100) protein coding genes and approximately 50% of all known miRNAs are highly expressed in the brain (Serafini *et al.* 2012). Patients within DECIPHER have multiple developmental abnormalities, and >90% of the sample present with mental retardation, and therefore the enrichment of miRNAs may explain the patients' phenotype. The enrichments of miRNA target sites amongst the disease CNVs and the depletion of the target sites amongst the control CNVs may indicate that disease

associated CNVs overlap genes that are under tighter regulatory control than benign CNVs. However the enrichments and depletions may be confounded by the enrichment and depletions of all protein coding genes within these CNVs. As miRNA targets sites often reside in the 3' UTRs of protein coding genes, the increased numbers of genes within the disease sets automatically increases the likelihood of a miRNA target enrichment. In addition, brain expressed genes have longer 3' UTRs than genes with other tissue specificity. Consequently, brain expressed genes will possess more miRNA target sites. To account for this, I would need to repeat the randomisations accounting for both GC content and the number of overlapping protein coding genes.

The LINE and SINE repeats are not largely over or under-represented within the five CNV sets, apart from the Nijmegen *et al.* set where a small significant depletion of LINEs or SINEs is observed. Previous research identified that benign CNVs are enriched in LINE repeats, indeed this observation is used in the GECCO classifier to sort disease from benign CNVs (Hehir-Kwa *et al.* 2010). It is when examining the LINE and SINE enrichments that I observe the biggest difference between the genomic features amongst the Nijmegen *et al.* controls and the AGP/Shaikh *et al.* controls, with the Nijmegen *et al.* controls showing a much larger depletion in LINEs and SINEs in comparison to the other two control sets. The Nijmegen *et al.* CNV set was called using a 32K BAC microarray, which does not have good signal-to-noise ratio within repeat regions of the genome and thus may be predisposed to call CNVs in non-repeat regions resulting in the apparent depletion of SINEs and LINEs in these control CNVs.

As some of my results differed from those observed in previous experiments, particularly the depletion of LINEs and SINEs, I repeated the enrichment analyses without accounting for GC distribution biases within the 5 sets. By not accounting for GC biases the observed

enrichments of several genomic elements associated with high GC increase, including those of protein-coding genes and RNA genes.

The five CNV sets show significant depletions in both haploinsufficient and haplosufficient protein coding genes, as defined by the MGI database. One would expect disease associated CNVs to be enriched in haploinsufficient genes, but this is not observed in my analysis. However, in subsequent chapters (see **Chapter 4**) I repeat this analysis on a subset of genes within the DECIPHER and ECARUCA set that I predict to be responsible for the patients' disease phenotypes and observe a significant enrichment of haploinsufficient genes, supporting the association of these CNVs with disease. The depletion I describe in this chapter may be due to the large size of the CNVs which may encompass many genes that are not haploinsufficient as well as the disease associated ones that are. To examine this hypothesis I tested DECIPHER and ECARUCA CNVs under 5Mb for an enrichment of haploinsufficient genes and observed a far smaller depletion (-30% and -38% respectively). The overall depletion of haploinsufficient genes within the disease CNVs may therefore be due to the small proportion of overlapping protein coding genes that are underlying the disease. The depletion of haplosufficient genes is expected, as one would expect benign CNVs to disrupt genes that do not require a tightly regulated copy number for normal function. The observed results may be caused by the difficulty of accurately defining a set of haplosufficient genes. The majority of the mouse phenotype annotations in the MGI resource are the result of homozygous deletions and in addition the mouse phenotype "normal" is rarely reported due to the relative infrequency of a 'complete' mouse phenotype analysis being performed; indeed several gene disruptions are annotated with both a normal and an abnormal phenotype.

I also repeated the analysis of CNV distribution across the genome as carried out in the Nguyen 2006 paper (Nguyen *et al.* 2006). Their analysis examined the frequency of CNVs

observed in subtelomeric and pericentromeric regions compared to CNVs that were randomly sampled from the genome. The 2006 paper identifies a significant enrichment of benign CNVs amongst subtelomeric and pericentromeric regions. My results replicate their analysis, revealing enrichments of pericentromeric and subtelomeric regions for each of the three sets of benign CNVs. I also observe a depletion of centromeric regions within the two disease associated sets. Due to the low GC and gene content of centromeric regions, this may be the result of disease CNVs being more prone to encompass both gene dense and GC rich regions.

The nucleotide substitution rates within the 5 CNV sets are quite similar. However, the benign CNVs have very slightly increased ds values in comparison to the disease CNV sets, and all sets have a higher value than the genome average. This may be explained by the observation that ds values are known to be tightly correlated with G+C content (Stoletzki and Eyre-Walker 2011). The two disease sets have a higher median dn/ds value than the three benign sets. This contradicts what I expected, that the benign sets would have a higher dn/ds as their overlapped genes are free to be disrupted without resulting in an abnormal phenotype. Through forming a 5Mb size cut-off set for the two sets of disease associated CNVs, I observe a decrease in the median dn/ds values. This observation may be the result of the smaller CNVs having a higher disease/benign overlapping gene ratio than the bigger CNVs in the dataset.

In summary, I identify enrichments of several different genomic elements amongst disease associated and benign CNVs. This work shows that CNVs, both large and *de novo* as well as smaller inherited CNVs, are not distributed randomly in the genome. Indeed, it is already known that CNVs are correlated with regions of high

GC content in the genome and that CNVs resulting from NAHR are frequently observed in repetitive regions of the genome (Gazave *et al.* 2011).

In this Chapter, I also attempted to identify several genomic features that differ between CNVs associated with disease and those that are observed in healthy individuals. Larger enrichments of protein coding genes, disease associated genes and miRNA loci are observed in disease CNVs compared to controls. Benign CNVs are depleted in miRNA target sites and disease CNVs are depleted in pericentromeric regions compared to the disease and benign sets respectively. I also identify genomic features that do not differ between disease and benign CNV sets, namely RNA genes, repeat elements, haploinsufficient genes and haplosufficient genes. This chapter also describes the identification of genomic properties that differ between CNVs with the same 'disease' or 'benign' status, suggesting that different CNV detection methods have a bias in what kind of CNVs they are able to detect. This needs to be taken into account when comparing sets of CNVs or generalising results from a subset of CNVs to all CNVs within the human population.

Chapter 4: Identifying mouse phenotype enrichments amongst CNVs observed in human developmental disorders

4.1 Abstract

De novo CNVs are thought to underlie many human developmental abnormalities. However, it is not clear how these *de novo* CNVs exert their effects, or indeed how CNVs in different regions in the genome can cause the same developmental abnormality. I hypothesised that copy number variants that give rise to the same human developmental abnormality overlap genes whose 1:1 disrupted mouse orthologues give rise to the same mouse phenotype. Therefore, by collecting large numbers of disparate human CNVs, I hoped to identify over-represented mouse phenotype associations amongst the copy number variable genes indentified in patients with the same human phenotype. I obtained 1,624 *de novo* CNVs identified in patients with developmental abnormalities from the DECIPHER and ECARUCA databases. Of the 1,088 human developmental abnormalities observed within these two datasets, I was able to associate 143 human developmental abnormalities with mouse model phenotypes. Many of the significantly associated mouse phenotypes are directly comparable to the human developmental abnormality, however others are less so, generating novel biological hypotheses as to which biological processes are disrupted in each individual human developmental abnormality. I propose the genes

contributing these associations as candidate genes for the patients' developmental abnormalities. Of the 2,086 candidate genes, 65% have not previously been associated with disease in OMIM. The distribution of the candidate genes amongst the patients provides evidence of extensive pleiotropy and epistasis in the underlying causes of human developmental disorders. Significant GO associations and significant direct protein-protein interactions amongst the candidate genes are observed for 67 and 37 of the human developmental abnormalities, providing further supporting evidence of their role in human disease.

4.2 Introduction

Developmental abnormalities occur in approximately 3% of births, and encompass a wide range of different human presentations including cleft palate, autism and intellectual disability (Chelly *et al.* 2006; Kogan *et al.* 2009; Bister *et al.* 2010). Frequently, it is revealed that patients who present with one or more of these disorders possess large *de novo* copy number variants (CNVs; genomic deletions and duplications >1Kb) which are put forward as the underlying cause of the patient's phenotype (Morrow 2010). These CNVs are further suspected as being the likely cause of disease due to their different attributes in comparison to apparently benign inherited CNVs (see **Chapter 3**), and because they are *de novo* and consequently not observed in the patients' healthy parents. For several developmental disorders a particular genomic locus has been identified as the cause through the observation of overlapping CNVs in multiple patients; for example, the 22q11.2 deletion is commonly observed in patients with di-George syndrome (see **Chapter 1**). For many other developmental abnormalities it has not been possible to identify a single recurrently copy number variable region across multiple patients.

In order to identify further recurrently copy number variable regions that underlie a patient's developmental disorder, databases such as the DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) and the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA) have been formed (Feenstra *et al.* 2006; Firth *et al.* 2009). The databases collate cytogenetic and clinical data obtained from a large number of patients acquired from numerous different clinical centres. Through collecting CNVs from patients presenting with the same abnormality, it may be possible to associate a recurrently copy number variable region with a disorder.

For patients that present with the same disorder but have pathogenic CNVs in different regions in the genome it is possible to employ functional enrichment analysis (FEA) approaches (see **Chapter 1**). FEA approaches seek to identify disease causal pathways and hypothesise that dispersed CNVs observed in patients presenting with the same developmental abnormality overlap genes that participate in a shared biological process. It is the disruption of this common process within each of the patients that results in their common disease presentation. As a result, FEA approaches can identify whether there is anything unusually common about the function of genes overlapping CNVs observed in patients presenting with the same developmental abnormalities. In comparison to approaches that identify candidate genes at a single genomic locus, for example genome wide and linkage association studies, FEA methods have the potential to gain sizeable power increases to identify the commonly affected biological process(es) through the simultaneous examination of the contribution of numerous structural variants across many patients' genomes. The null hypothesis employed in FEA approaches proposes that CNVs observed in patients with the same developmental abnormality are randomly sampling the human genome. Through the rejection of the null hypothesis, the observed functional

enrichment is associated with the patients' disorder and proposes genes that contribute to the enrichment as candidate genes for the patients' disorder.

When employing FEA approaches, "function" can be defined through exploiting different datasets of functional genomic annotations. In this chapter I exploit a set of mouse phenotypes arising from the intentional disruption of genes in the mouse (Blake *et al.* 2009). Mouse models have contributed substantially to our understanding of gene function and also provide mechanistic insights into human developmental disorders (Wilson 2000; Smith and Eppig 2009; Webber *et al.* 2009; Miller *et al.* 2010; NHGRI 2011). Within this chapter, I propose that an overrepresentation of one or more mouse phenotypes amongst the human:mouse 1:1 orthologues disrupted by CNVs will identify shared features amongst genes underlying individual developmental abnormalities. Consequently, an overrepresentation of a mouse phenotype would provide three conclusions: firstly, the association of the mouse phenotype with the human phenotype; secondly, the genes that contribute to the mouse phenotype enrichments are candidate genes whose disruption (deletion or duplication) underlies the patient's developmental abnormality; and, thirdly that the CNVs in which the candidate genes reside underlie the patients' phenotype.

In order to identify associations between mouse models and human phenotypes, both of the descriptors need to be defined in a consistent manner. With DECIPHER and ECARUCA the human phenotypes are described using terms from the London Medical Database (LMD) (Fryns and de Ravel 2002). The LMD is a three level hierarchical ontology, enabling patients to be grouped by specific phenotype terms or in larger groups using broader phenotype descriptors. The mouse model phenotypes are held by the Mouse Genome Informatics resource (MGI) and described using terms from the mammalian

phenotype ontology (Smith and Eppig 2009). As with other functional genomic resources, mouse phenotype annotations are not available for all genes, nor are the genes within the database comprehensively tested for each available annotation. Consequently, the absence of a significant mouse phenotype enrichment amongst a gene set can not lead to the conclusion that the functional annotation is not of interest to the disorder under investigation (see **Discussion**).

In this chapter I associate over 100 mouse model phenotypes with human developmental abnormalities through identifying significant enrichments of copy number variable genes whose disrupted mouse orthologues result in a specific mouse model phenotype. I propose the genes contributing these enrichments as being candidate genes for the patients' abnormal phenotypes. Further analyses of these genes suggest that extensive pleiotropy and epistasis play an important role in the causes of developmental disorders.

The work within this chapter has been published as a research article in Human Mutation (Boulding and Webber 2012).

4.3 Methods

CNVs and Patient Phenotypes

I obtained two sets of *de novo* CNVs for analysis. The largest set, from the ECARUCA database (Feenstra *et al.* 2006), consists of 988 CNVs observed in 958 patients. The second set, consisting of 636 CNVs from 525 patients, was obtained from the DECIPHER database (Firth *et al.* 2009) (**Table 4.1**).

CNV set	Number of CNVs	Median CNV size (range) (Mb)	Median number of symptoms per patient (range)	Total number of symptoms in the dataset	Genes covered* (median/ CNV)
DECIPHER	636	2.3 (0.003-53.7)	5 (1-32)	685	10487 (18)
ECARUCA	988	18.1 (1.3-146)	8 (2-102)	892	17791 (112)
Combined	1421	11.7 (0.003-146)	7 (1-102)	489	19156 (78)
Overlap	1386	4.3 (0.001-8.6)	7 (1-102)	385	11644 (31)

Table 4.1: Summary statistics of the DECIPHER, ECARUCA, combined and overlap datasets. CNV size range, human phenotype data and genomic coverage of CNVs. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

In the DECIPHER and ECARUCA set, contributing patients are annotated with terms from the London Medical Database (LMD) (Fryns and de Ravel 2002) (see **Chapter 2**). Patients in both sets are described by multiple LMD terms (median = 5 for DECIPHER and median = 8 for ECARUCA). For each of the different developmental abnormalities defined by the LMD I formed a non-exclusive group of *de novo* CNVs drawn from those patients annotated with that LMD term (herein termed Symptom-CNV sets). This was performed separately for each of the DECIPHER and ECARUCA sets as well as for their combined set of CNVs. As >90% of patients are annotated with multiple LMD terms, the majority of the CNVs were non-exclusively assigned to a Symptom-CNV set (**Figure 4.1**). I also considered that the direction of copy number change (deletion/duplication) could affect the underlying pathoetiology of a particular disorder, and thus each CNV set formed was further subdivided into "loss" (deletions) and "gain" (duplication) CNV sets associated with a particular LMD phenotype. I assigned genes to CNVs using the method outlined in **Chapter 2**.

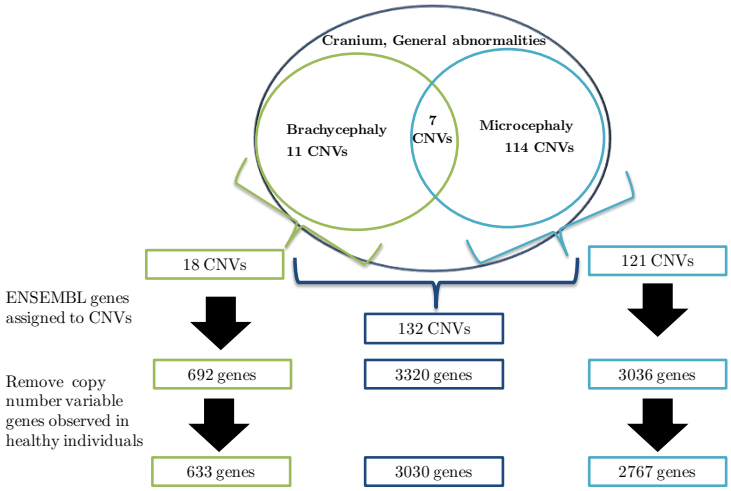


Figure 4.1: The formation of Symptom-CNV sets from the DECIPHER and ECARUCA data. The CNVs were grouped non – exclusively into Symptom-CNV sets. **Stage 1:** For each symptom-CNV set, genes overlapped by these CNVs were obtained from ENSEMBL (see **Chapter 2**). **Stage 2**: Genes observed to be copy number variable in the same direction (deletion or duplication) within apparently healthy individuals were removed.

Mouse model phenotypes

Mouse phenotype descriptions resulting from the disruption of mouse orthologues of human genes were obtained from the Mouse Genome Informatics online resource (MGI) (see **Chapter 2**) (Eppig JT 2007). Using 1:1 gene orthology relationships between mouse and humans defined by the MGI, I mapped mouse phenotype terms to 5,671 ENSEMBL genes. For each CNV set in the DECIPHER, ECARUCA, and the COMBINED DECIPHER and ECARUCA sets I tested for an enrichment of genes whose disrupted mouse orthologues are associated with specific mouse phenotypes (**Figure 4.2**). Previously in our lab a two step procedure was employed, firstly testing for enrichments within each of the 30 overarching mouse phenotype categories, and then secondly examining each mouse finer phenotype term within any overarching category found to be significant (Webber *et al.* 2009). However as >90% of the patients in both data sets present with multiple developmental abnormalities, the vast majority of the CNVs are non-exclusively assigned to a Symptom-CNV set and consequently straddle several phenotype categories. Therefore, to concentrate my analysis of each CNV set toward the

developmental abnormality of interest, I tested the mouse finer phenotypic terms within the mouse overarching phenotypes most relevant to the human symptom being investigated (Shaikh *et al.* 2011).

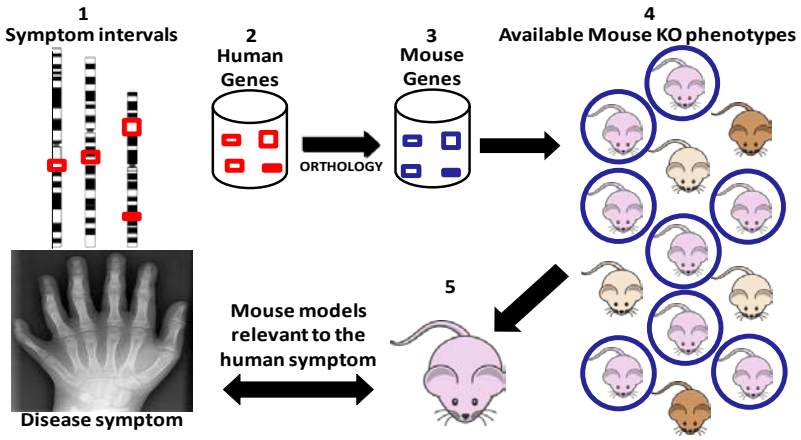


Figure 4.2: Examining human Symptom-CNV sets for enrichments of genes associated with mouse phenotypes. (**1**) Formation of a set of CNVs associated with a particular human symptom. (**2**) Identification of human genes affected (duplicated or deleted) by the CNVs. (**3**) Identification of unique mouse orthologues of affected human genes. (**4**) Identification of the phenotypes reported for the knockout models of these mouse orthologues. (**5**) Identification of mouse phenotype enrichments significantly above that expected by chance.

Given the absence of a suitable control (see **Chapter 1** and **Chapter 2**), I compared the gene content of these *de novo* CNV sets to the genomic background (Raychaudhuri *et al.* 2010), noting that the gene assignment procedure does not appear to incur any concerning tissue-specific gene bias (Webber 2011). When assigning genes to a Symptom-CNV set I take genes that have at least one exon overlapped by a CNV. This prevents an ascertainment bias of selecting only longer genes or shorter genes, which are known to have tissue specific expression levels (see **Chapter 2**). I employed a hyper-geometric test to test the null hypothesis that *de novo* CNVs identified in patients with a specific developmental disorder randomly sample all genes. As many mouse phenotypes were tested, a multiple testing correction, False Discovery Rate (FDR) <5% was applied (Benjamini and Hocjberg 1995). The application of this significance threshold is

conservative, given that the FDR correction assumes each test to be independent. In reality, many phenotype terms within the MPO are directly related, for example the mouse phenotype term *abnormal learning and memory* and the term *abnormal learning/memory and conditioning* are associated with the same genes.

Overlapping CNV region analysis

I hypothesised that dispersed CNVs contributing to the same developmental phenotype will possess common functional genomic annotations, specifically one or more shared mouse model phenotype annotations amongst affected genes' mouse orthologues. The CNVs that I observe in the two patient samples are large and therefore overlap several genes, not all of which are likely to be involved in the patient's disorder. To concentrate the signal I decided to perform a further analysis using the smallest CNV regions present in multiple patients with a shared developmental disorder (**Figure 4.3**)

Figure 4.3: For each LMD term I identified the smallest CNV regions observed in multiple patients. **(1)** On this chromosome I observe two overlapping CNVs that result in one overlapping CNV region for analysis. **(2)** The four overlapping CNVs result in two regions of maximal overlap to be taken forward for the overlapping CNV region analysis.

Patient resampling method

To examine the role of CNV sample size on obtaining statistically significant results I repeated the mouse phenotype analysis of the DECIPHER set using 25%, 50% and 75% of the sample. At each percentage, that proportion of patients from the DECIPHER set were randomly sampled 100 times.

Analysis of candidate genes

For each Symptom-CNV set I considered the genes contributing each significant mouse phenotype enrichment as candidate genes for that particular developmental abnormality. I wished to examine each set of genes for evidence of shared molecular features and protein-protein interactions. To achieve this I examined the candidate genes identified for each Symptom-CNV set for significant enrichments of one or more Gene Ontology (GO) terms (see **Chapter 2**) and protein-protein interactions (using DAPPLE – a protein-protein interaction network hosted at the Broad Institute). Significance for the GO enrichments was determined by comparing the candidate gene list to the whole genome using a hypergeometric test and a multiple testing correction (FDR<5%). For the protein-protein interactions the direct interactions between the candidate genes were compared to 10,000 randomisations drawn from the entire genome.

4.4 Results

I sought to objectively associate mouse model phenotypes with individual human developmental abnormalities. To achieve this I examined Symptom-CNV sets formed from *de novo* CNVs observed in human patients that share a common abnormality for enrichments of genes whose disrupted mouse orthologues result in the same mouse phenotype.

I obtained two sets of *de novo* CNVs identified in patients with multiple symptoms. The first set from DECIPHER comprises 636 CNVs (Firth *et al.* 2009). The second set obtained from the ECARUCA database comprises 988 CNVs (Feenstra *et al.* 2006). The median size of the DECIPHER CNVs is 2.2Mb, mean = 3.7Mb, S.D = 4.4Mb. The

ECARUCA CNVs have a median size of 18.1Mb, mean = 21.2 Mb S.D = 14Mb. In total, 685 (median per patient = 5) different LMD terms are annotated to patients whose CNVs form the DECIPHER set, while 892 LMD terms (median per patient = 8) are observed for the ECARUCA set's patients.

I wished to identify mouse model phenotypes associated with individual human symptoms, separately for the DECIPHER and ECARUCA datasets. Thus, for each of the human symptoms observed, I formed a non-exclusive set of *de novo* CNVs drawn from the subset of patients with that symptom (herein termed Symptom-CNV sets; **Figure 4.1**). As 97% of patients present with more than one symptom, >97% of the CNVs belong to more than one Symptom-CNV set. For symptoms observed among both the DECIPHER and ECARUCA patients, additional Symptom-CNV sets were created from the patients presenting with each human symptom observed across both CNV datasets (herein termed combined-Symptom-CNV sets). Considering that the direction of copy number change (gain/loss) of a gene could affect the underlying pathoetiology of a symptom, the Symptom-CNV sets were further subdivided to form groups of *Gain* and *Loss* CNVs associated with each of the human symptoms. For each of the Symptom-CNV sets, the protein-coding genes affected by CNVs were identified using ENSEMBL (see **Methods**).

Mouse model phenotype data resulting from the disruption of 5671 1:1 human:mouse orthologues were obtained from the Mouse Genome Informatics Resource (MGI) (see **Methods**). I examined each Symptom-CNV sets for an enrichment of mouse model phenotypes listed under the overarching category or categories I deemed most relevant to the human symptom under investigation (**Supplementary Table 4.1**).

Associating mouse model phenotypes with individual developmental abnormalities

Of the 685 DECIPHER and 892 ECARUCA symptom-CNV sets examined, I identified 46 (7%) and 101 (11%) sets respectively that are significantly enriched (FDR<5%) in genes whose disrupted mouse orthologues result in a specific mouse phenotype (**Tables 4.2 and 4.3; Supplementary Table 4.2 and Supplementary Table 4.3**). Within both CNV data sets many mouse phenotypes are readily comparable to the human developmental abnormality under investigation, while others are less, thereby generating novel pathoetiological hypotheses (**Figure 4.4**). Across the DECIPHER and ECARUCA datasets I identify two human symptoms with the same significant mouse phenotype enrichment, "mental retardation" and "syndactyly of toes", that are associated with the mouse model phenotypes *abnormal brain morphology* and *abnormal skeleton extremities morphology",* respectively. Of the 489 combined Symptom-CNV sets (DECIPHER and ECARUCA) I identify mouse model phenotype enrichments for 74 (15%) human developmental abnormalities of which 41/74 were not observed when considering each dataset separately (**Table 4.4 and Supplementary Table 4.4**).

Human Symptom	Mouse phenotypic enrichment	% Enriched	Patients hit/ total	Gene count
Build	Decreased Birth Body Size	182	9/17	12
Thin or slender build, general abnormalities	Decreased Fetal Size	179	11/26	20
Low Birthweight	Decreased Fetal Size	213	9/24	16
Short Stature, general abnormalities	Increased Lean Body Mass	259	9/72	8
Short Stature, prenatal onset	Abnormal fetal growth/weight/body size	138	12/19	21
Tall stature, proportionate	Abnormal chest morphology	4245	1/1	2
Prominant forehead/frontal bossing	Abnormal soft palate	1262	3/27	4
Broad base to nose	Malocclusion	2117	2/4	4
Large nose	Absent palatine shelf	3172	2/5	3
Cupid bow shape of mouth	Abnormal secondary palate development	1491	2/2	4
Open mouth appearance	Branchial arch hypoplasia	3681	2/7	3
Thin lower lip	Abnormal tooth mineralisation	8428	1/1	2
Short Philtrum	Malocclusion	1281	3/7	4
Malocclusion of teeth	Small branchial arch	2542	2/2	3
Nasal Speech	Abnormal prepulse inhibition	2468	1/5	5
Speech defect/dysarthria	Abnormal thermal nocioception	1059	3/5	6
Loose skin in neck	Premature hair loss	6989	1/1	2
Scapulae, general abnormalities	Abnormal cartilage morphology	772	2/3	5

Broad Hands	Short femur	1377	2/4	4
Camptodactyly	Abnormal metacarpal bone morphology	825	2/8	6
Clinodactyly	Decreased long bone epiphyseal plate size	560	5/28	6
Thin brittle nails	Abnormal limb/digit/tail morphology	464	2/2	6
Syndactyly 2-3 of toes	Abnormal metacarpal bone morphology	823	1/9	5
Syndactyly of toes (not 2-3)	Abnormal metacarpal bone morphology	1702	6/7	6
Haematology/Immunology, general abnormalities	Abnormal T cell activation	180	9/13	20
Neurology	Abnormal brain morphology	15	275/384	401
Mental retardation/developmental delay	Abnormal brain morphology	15	253/349	383
Hyperactivity	Abnormal conditioning behaviour	589	5/24	9
Psychotic behaviour	Abnormal prepulse inhibition	3182	1/1	5
Seizures, general abnormalities	Increased sensory neuron number	252	16/67	13
Complex partial seizures	Abnormal circadian rhythm	4070	2/2	3
Febrile Convulsion	Abnormal ammon gyrus morphology	2864	1/1	3
Spinal myoclonus	Abnormal prepulse inhibition	1377	1/1	4
Paroxymal disorders general abnormalities	Abnormal prepulse inhibition	1263	2/4	6
Intermittent tremor at rest	Abnormal prepulse inhibition	1377	1/1	4
Brachycephaly	Abnormal Purkinje cell dendrite morphology	747	5/13	6
Dolichocephaly/sachocephaly	Abnormal cued conditioning behaviour	1156	2/3	6
Ataxia, general abnormalities	Abnormal limbic system morphology	264	6/8	12
Spasticity/brisk reflexes/Babinski	Abnormal neurotransmitter secretion	740	3/7	6
Neuroradiology general abnormalities	Abnormal cerebellum development	290	9/13	11
Basal ganglia lesion*	Abnormal prepulse inhibition	1377	1/1	4
Thalamic lesion*	Abnormal prepulse inhibition	1377	1/1	4
Nevi or lentigines	Thick dermal layer	8240	1/1	2

Table 4.2: The most significant mouse phenotype enrichment observed amongst genes in each DECIPHER symptom-CNV set. Enrichments are given as the percentage change over that expected by chance. The number of patients with at least one gene contributing a mouse phenotype enrichment is given as a fraction of the total number of observed human patients presenting with that symptom. The full listing of all associated mouse models phenotypes for DECIPHER Symptom-CNV sets is given in Supplementary Table 4.2. *These enrichments are comprised of the same genes and come from the same patient.

Human Symptom	Mouse Phenotype Enriched	% Enriched	Patients hit	Gene count
Brachycephaly	abnormal malleus morphology	90	32/110	24
Dolichocephaly/Scaphocephaly	abnormal malleus morphology	166	17/60	17
Microcephaly	small branchial arch	58	82/218	37
Plagiocephaly/Asymmetrical skull	abnormal malleus morphology	197	14/35	23
Cerebral atrophy/heterotopias	abnormal basisphenoid bone morphology	148	17/36	18
Cerebellar abn./hypopl.(Structural)	abnormal nasal bone morphology	252	11/15	14
Hydroceph/Large ventricles non-spec	abnormal external auditory canal	138	19/100	13
Flat occiput	abnormal malleus morphology	129	15/53	18
Delayed closure of/Large fontanelle	abnormal maxilla morphology	72	53/74	60
Ridged cranial sutures	abnormal hyoid bone morphology	550	3/4	7
Wide cranial sutures	anencephaly	385	9/20	7
Prominent forehead/frontal bossing	abnormal palate morphology	28	93/151	125
Hyperplastic supra-orbital ridges	abnormal basioccipital bone morphology	797	2/2	5
Metopic ridge*	abnormal occipital bone morphology	798	2/2	7
Narrow forehead/Temporal narrowing	small branchial arch	103	17/60	22

Sloping forehead*	abnormal occipital bone morphology	840	2/2	7
Hypoplastic supra-orbital ridges	branchial arch hypoplasia	458	5/14	7
Dysplastic ears	abnormal malleus morphology	60	79/224	31
Posteriorly rotated ears	abnormal endolymphatic duct morphology	115	34/116	19
Pre-auricular pits/fistulas	abnormal organ of Corti supporting cell morphology	168	25/36	18
Prominent ears	abnormal malleal manubrium morphology	147	21/81	11
Simple ears	abnormal malleus morphology	219	22/35	18
Prominent anti-helix	abnormal malleus morphology	149	15/38	18
Corneal abnormalities	abnormal cornea morphology	352	3/4	9
Coloboma of iris	abnormal eye development	75	29/35	40
Iris atrophy/dysplasia	anophthalmia	556	4/4	6
Absent eyelids	abnormal basioccipital bone morphology	974	2/3	5
Blepharophimosis	abnormal hard palate	78	28/79	28
Palpebral fissures slant down	abnormal craniofacial development	40	121/150	139
Epicanthic folds	mandible hypoplasia	57	80/267	21
Broad base to nose	abnormal malleus morphology	131	19/45	18
Large nose	abnormal head morphology	48	24/28	73
Flat nose	small nasal bone	421	7/21	7
Small/short nose	abnormal hard palate	67	62/153	35
Depressed/flat nasal bridge	abnormal craniofacial development	24	72/222	173
High/prominent nasal bridge	abnormal maxillary shelf	127	21/82	16
Wide nasal bridge	abnormal hard palate	65	76/170	33
Anteverted nares	abnormal craniofacial development	26	103/143	136
Flared nares	abnormal malleus morphology	650	3/4	6
Asymmetric face	abnormal middle ear ossicle morphology	158	27/41	23
Mid-face hypopl.(excl.flat malar)	abnormal first branchial arch morphology	129	48/85	21
Flat malar region	abnormal mandibular angle morphology	891	8/17	5
Small mandible/micrognathia	abnormal craniofacial development	18	316/395	202
Prominent maxilla	abnormal second branchial arch morphology	441	5/7	8
Down-turned corners of the mouth	abnormal molar morphology	73	54/148	27
Long philtrum	abnormal malleal manubrium morphology	98	24/138	12
Short philtrum	abnormal maxilla morphology	63	46/71	43
Wide philtrum	abnormal hyoid bone morphology	234	4/10	11
Cleft upper lip (non-midline)	abnormal palatal shelf fusion at midline	219	31/42	12
Prominent upper lip	abnormal malleus morphology	154	10/33	16
High palate	small branchial arch	40	92/223	41
Prominent lateral palatine ridges	abnormal third branchial arch morphology	788	1/1	5
Delayed tooth eruption/development	abnormal skull morphology	63	14/14	71
Irregular or crowded teeth	abnormal craniofacial development	48	30/37	80
Neonatal teeth	abnormal occipital bone morphology	475	3/4	7
Abnormally shaped teeth	abnormal nasal capsule morphology	557	6/13	7
Lordosis	abnormal calvaria morphology	262	2/2	12
Meningocele/Meningo-myelocele	abnormal cranial base morphology	459	4/4	8
Scoliosis	abnormal metacarpal bone morphology	115	12/63	19
Sacral dimple/sinus	abnormal rib morphology	45	55/65	83
Vertebrae,general abnormalities	abnormal malleus morphology	207	10/21	18

Asymmetric thorax	decreased birth body size	138	8/12	16
Broad/Barrel thorax	abnormal chest morphology	453	2/2	6
Pulmonary incompetence	abnormal cardiovascular development	475	1/1	6
Cardiac situs inversus/dextrocardia	abnormal heart ventricular pressure	1429	1/1	4
Tricuspid incompetence	abnormal tricuspid valve morphology	374	6/11	8
Respiratory difficulties,general	abnormal olfactory placode morphology	607	3/4	5
Abdomen,general abnormalities^	abnormal small intestine morphology	452	2/2	7
Small bowel atresia/absence/obstr. ^	abnormal small intestine morphology	410	2/2	9
Feeding problems in infants	abnormal palate development	63	54/98	32
Inguinal hernia	cleft palate	42	46/67	82
Megacolon or Hirschprungs syndrome	decreased pancreatic beta cell number	720	2/6	5
Abnormal liver (inc.function)	abnormal pancreatic alpha cell morphology	1096	3/4	5
Stomach tumours	abnormal intestinal goblet cells	1475	1/1	4
Pelvis,general abnormalities	abnormal thoracic cage	261	2/2	20
Pubic ossification defect	abnormal occipital bone morphology	1311	1/1	5
Fused labia	abnormal prostate gland morphology	603	2/2	6
Nephritis or nephropathy	renal fibrosis	656	1/2	6
Renal tumours (inc.Wilms)	renal fibrosis	627	19/19	6
Webbing at elbow °	abnormal long bone morphology	891	1/1	5
Fingers,general abnormalities	abnormal forelimb morphology	579	1/1	7
Adducted thumbs	decreased long bone epiphyseal plate size	260	9/16	10
Proximal placement of thumb	absent radius	262	10/53	8
Absent or hypoplastic patella °	abnormal long bone morphology	891	1/1	5
Genu valgum	abnormal forelimb morphology	215	3/3	15
Genu varum	abnormal caudal vertebrae morphology	669	2/2	6
Hypoplastic or absent tibia	abnormal long bone morphology	396	2/2	7
Ankle,general abnormalities °	abnormal long bone morphology	891	1/1	5
Short hallux	abnormal skeleton extremities morphology	123	9/9	29
Syndactyly 2-3 of toes	short radius	106	22/70	20
Syndactyly of toes (not 2-3)	abnormal skeleton extremities morphology	58	24/26	53
Bleeding diatheses	decreased spleen weight	2306	1/1	4
Recurrent infections	abnormal T cell apoptosis	56	59/106	51
Hemiplegia/Tetraplegia	abnormal prepulse inhibition	1058	1/2	5
Hypotonia	abnormal brain morphology	12	291/298	601
Lethargy	abnormal sensory neuron morphology	269	2/2	13
Mental retardation	abnormal forebrain morphology	8	78/711	560
Seizures/Abnormal EEG	abnormal brain morphology	15	72/175	507
Cartilagineous exostoses	abnormal skeleton development	160	13/13	35
Osteoporosis	cervical vertebral transformation	385	3/6	9
Stippled or fragmented epiphyses	abnormal calvaria morphology	504	2/2	10

Table 4.3: The most significant mouse phenotype enrichment observed amongst genes in each ECARUCA symptom-CNV set. Enrichments are given as the percentage change over that expected by chance. The number of patients with at least one gene contributing to a mouse phenotype enrichment is given as a fraction of the total number of observed human patients presenting with that symptom. The full listing of all associated mouse models phenotypes for ECARUCA Symptom-CNV sets is given in Supplementary Table 4.3. *^°These enrichments are comprised of the same genes and come from the same patient.

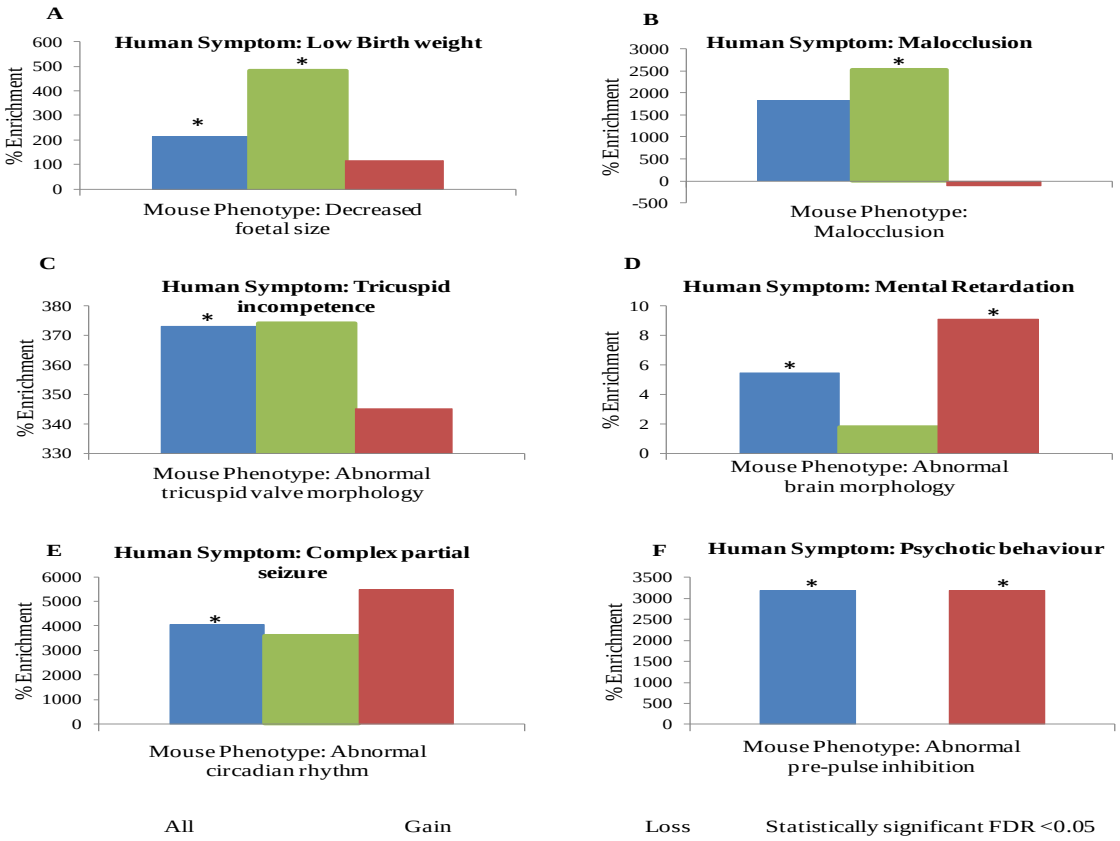


Figure 4.4: Example enrichments of mouse phenotypes amongst CNVs observed in patients with a common symptom. Enrichments are shown as the percentage increase over that expected by chance. Those marked with an asterisk are significant (FDR<5%). The enrichments shown in Panels A, B, E and F are those identified in DECIPHER Symptom-CNV sets (**Table 4.2**, **Supplementary Table 4.2**) while those shown in Panels C and D were identified in ECARUCA Symptom-CNV sets (**Table 4.3**, **Supplementary Table 4.3**). In panel F, the Symptom-CNV set contains only one loss CNV which is why there are only percentage enrichments for the "all" and "loss" groups.

Human disorder	Mouse phenotype	% Enrichment	Patients hit	Gene Count
Hydrocephaly/Large ventricles non-specific	abnormal external auditory canal	128.17	20/106	13
High frontal hairline	long incisors	637.84%	4/10	5
Metopic ridge	abnormal occipital bone morphology	797.72%	2/2	7
Narrow forehead/Temporal narrowing	small branchial arch	109.40%	19/74	24
Hypoplastic supra-orbital ridges	branchial arch hypoplasia	453.65%	5/15	7
Dysplastic ears	abnormal malleus morphology	57.63%	80/237	31
Posteriorly rotated ears	abnormal endolymphatic duct morphology	110.18%	36/129	19
Pre-auricular pits/fistulas	abnormal organ of Corti supporting cell morphology	151.91%	25/42	18
Pre-auricular tags	head tilt	220.21%	9/24	12
Prominent ears	abnormal malleal manubrium morphology	141.45%	21/90	11
Simple ears	abnormal malleus morphology	193.23%	23/44	18
Prominent anti-helix	abnormal malleus morphology	149.38%	15/39	18

Deafness,non-specific	absent saccule	352.39%	11/43	7
Corneal abnormalities	abnormal cornea morphology	336.90%	3/5	9
Coloboma of iris	abnormal eye morphology	33.44%	38/39	135
Synophrys	absent premaxilla	362.94%	10/60	6
Blepharophimosis	abnormal occipital bone morphology	63.90%	35/84	39
Flat nose	small nasal bone	389.36%	7/24	7
Small/short nose	abnormal hard palate	62.16%	62/168	35
High/prominent nasal bridge	abnormal maxillary shelf	115.93%	22/94	16
Wide nasal bridge	abnormal hard palate	59.16%	79/184	33
Flat malar region	abnormal mandibular angle morphology	832.12%	8/21	5
Long philtrum	abnormal malleus morphology	80.53%	34/130	27
High palate	small branchial arch	40.40%	96/251	42
Prominent lateral palatine ridges	abnormal third branchial arch morphology	723.32%	1/2	5
Irregular or crowded teeth	abnormal craniofacial development	44.08%	31/41	81
Lordosis	abnormal calvaria morphology	262.13%	2/2	12
Vertebrae,general abnormalities	abnormal malleus morphology	189.04%	11/26	18
Broad/Barrel thorax	abnormal chest morphology	453.45%	2/2	6
Mitral incompetence	thin myocardial wall	214.95%	7/11	13
Cardiac situs inversus/dextrocardia	abnormal heart ventricular pressure	1428.57%	1/1	4
Small bowel atresia/absence/obstr.	abnormal intestine morphology	184.59%	3/3	15
Feeding problems in infants	abnormal palate development	57.65%	65/154	35
Inguinal hernia	cleft palate	45.02%	50/79	87
Restriction,supernation/pronation	abnormal long bone morphology	254.08%	3/3	10
Adducted thumbs	decreased long bone epiphyseal plate size	241.50%	9/18	10
Proximal placement of thumb	absent radius	254.74%	10/58	8
Narrow feet	abnormal skeleton extremities morphology	310.76%	2/2	10
Short hallux	abnormal skeleton extremities morphology	105.38%	10/11	30
Blood vessels,general abnormalities	interrupted aortic arch	740.15%	4/6	6
Recurrent infections	abnormal T cell apoptosis	56.46%	63/116	53
Seizures/Abnormal EEG	abnormal brain morphology	14.17%	226/241	543

Table 4.4: The most significantly enriched mouse phenotypes observed among genes in human symptom-CNV set unique to the combined data set analysis. Enrichments are given as the percentage change over that expected from the genome by chance. The number of patients with at least one copy number variable gene contributing a mouse phenotype enrichment is given as a fraction of the total number of observed human patients presenting with that symptom**.**

I examined the results from the DECIPHER and ECARUCA datasets to identify whether any of the significantly enriched ($P<0.05$, FDR<5%) mouse model phenotypes are nominally enriched ($P<0.05$) in the other dataset. Using this approach I identify a further 15 human symptoms whose mouse phenotype enrichments are replicated across the two datasets (**Supplementary Table 4.7**).

Examining the contribution of CNV size to mouse phenotype enrichments

In both the DECIPHER and ECARUCA sets the CNVs are large, encompassing many genes that may not be responsible for the patients' disorder, in comparison to smaller CNVs where a higher proportion of the overlapping genes may be responsible for the patients' symptom. I hypothesised that smaller disease associated CNVs would overlap fewer non-disease causing genes and thereby would increase functional enrichments and therefore I would observe a greater number of significant enrichments when concentrating the analysis on the smaller CNVs in the two CNV datasets. To investigate this I examined CNVs in DECIPHER (median = 2.2Mb) less than 2.5Mb in size. In the ECARUCA database (median = 21.2Mb) I examined CNVs less than 25Mb in size. However, the observed results were not improved by considering only smaller CNVs. In the DECIPHER analysis I identify enrichments for 35/516 (6.7%) human symptoms and in the ECARUCA set I identify enrichments for 61/831 (7.3%). The reduced number of significant results may be due to the small size of the dataset not having the necessary power to detect enrichments. However, it has been previously shown that patients with developmental disorders have a greater enrichment of larger CNVs than smaller ones (Vermeesch *et al.* 2011). This provides evidence that epistasis has an underlying role in human developmental disorders. This is further supported by my identification of multiple candidate genes per patient (**Tables 4.2, 4.3 and 4.4**).

Larger CNV datasets identify candidate genes for a greater number of human developmental abnormalities and a greater proportion of patients analysed

I wished to examine how the growth of these CNV databases will affect the power of FEA analysis and our ability to identify causal genes within these CNVs and to identify biological processes underlying individual developmental disorders. I randomly re-sampled the DECIPHER dataset 100 times at 25%, 50% and 75% of the total number of patients. For each of these new datasets I repeated the mouse phenotype enrichment analysis. As the percentage of patients sampled grows, I identified enrichments for an increasing number of LMD terms. At the 25% sampling I identify enrichments for 2.6% of the available LMD terms, at 50% I identify enrichments for 4.1% of LMD terms, at 75% I observe enrichments for 5.3% and at 100% I observe enrichments for 6.7% (**Figure 4.5**). As the patient sample size increases I also identify enrichments for a greater proportion of the patient sample analysed (**Figure 4.6**). These results demonstrate the increased power obtained by pooling medical genomics data.

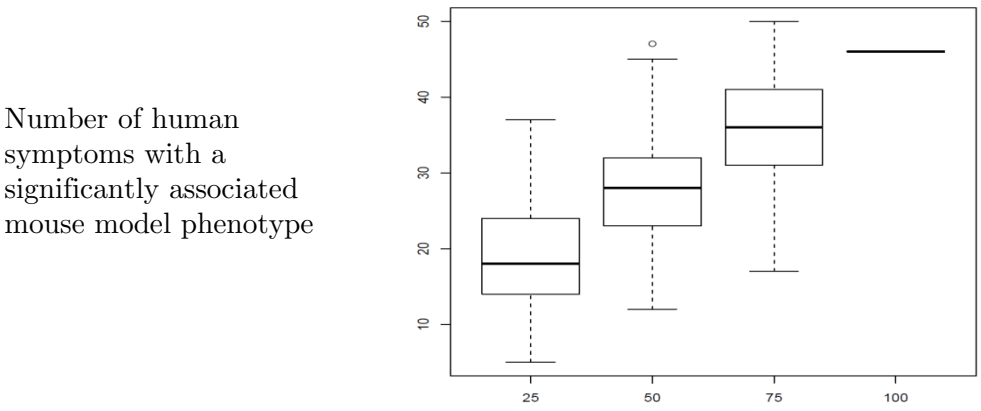


Figure 4.5: The number of human symptoms with significant mouse phenotype enrichments at 25%, 50%, 75% and 100% of the original DECIPHER patient sample. An increase in patient sample size enables the identification of significant mouse phenotype enrichments and candidate genes for a greater number of human symptoms

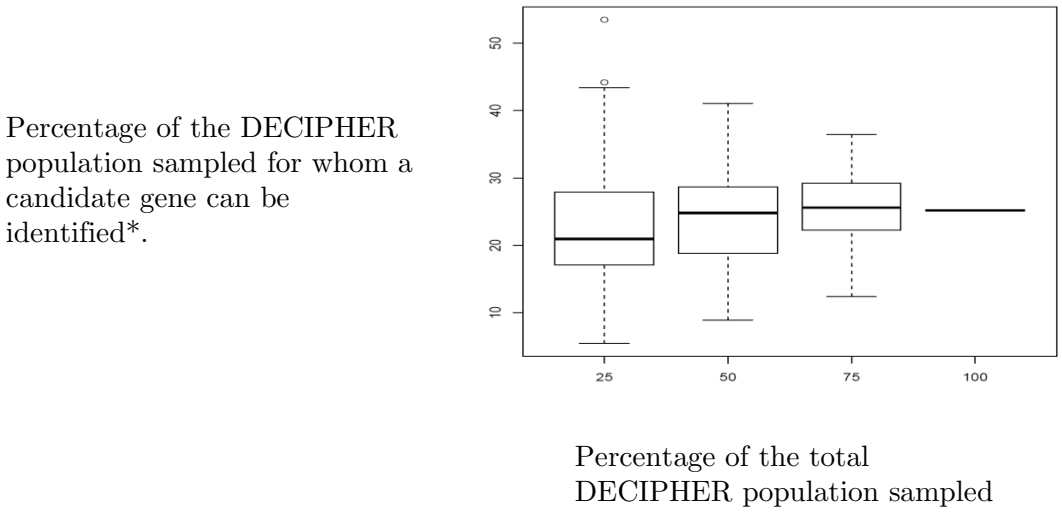


Figure 4.6: The proportion of patients for whom I identify at least one candidate gene at 25%, 50%, 75% and 100% of the original DECIPHER patient sample. An increase in patient sample size increases the proportion of patients for whom we can provide a causal hypothesis. *For this analysis I exclude the candidate genes identified for the human symptom mental retardation. This is due to the enrichment revealing 383 candidate genes, causing a large increase in the number of patients with a candidate gene.

Recurrently copy number variable regions observed in patients with a shared symptom identify additional mouse phenotype enrichments

Databases such as DECIPHER and ECARUCA are often exploited by clinicians to robustly associate a recurrently copy number variable locus observed in multiple patients with a developmental abnormality shared across those patients (Feenstra *et al.* 2006; Firth *et al.* 2009). Therefore, I re-examined only those regions observed to be copy number variable in multiple patients sharing the same developmental abnormality drawn from the combined DECIPHER and ECARUCA datasets for enrichments of genes associated with particular mouse phenotypes (**Figure 4.3**). By only examining CNV regions that occur multiple times in patients presenting with a particular human phenotype I aimed to reduce "noise" contributed by the large number of copy number

variable genes within the datasets that may not be causally-related to the patients' disorder.

Amongst the 385 combined-Symptom-CNV recurrent region sets, I identified 45 (12%) human developmental abnormalities with significant enrichments of genes associated with a mouse model phenotype, of which 28 (62%) developmental abnormalities had not been associated in the previous analyses (**Table 4.5**). For 16 of the 17 human phenotypes that had mouse model associations in the previous analyses, the enrichment is higher in the recurrent region set (**Table 4.5 and Supplementary Tables 4.6**). Although these findings show that recurrency is a powerful aid to identifying disease associations, the proportion of patients considered in each analysis for whom we can identify a candidate gene is higher (80%) in our original analysis than in the recurrent region analysis (58%) illustrating that dispersed loci provide significant, but also complementary, additional power.

Human Symptom	**Mouse Phenotype Enrichment**	**% Enrichment**	**CNVs hit**	**Genes hit**
Microcephaly	*abnormal maxilla morphology*	107.43	19/94	26
Plagiocephaly/Asymmetrical skull	*abnormal calvaria morphology**	161.23	7/10	21
Agenesis/hypopl. of corpus callosum	*abnormal occipital bone morphology*	341.26	5/15	11
Craniosynostosis	*abnormal basisphenoid bone morphology*	580.25	6/7	9
Prominent forehead/frontal bossing	*abnormal craniofacial development*	85.64	25/46	49
Wide forehead	*abnormal calvaria morphology**	150.71	6/12	19
Dysplastic ears	*abnormal malleus morphology*	213.38	9/39	15
Low-set ears	*abnormal incus morphology*	170.05	14/85	15
Posteriorly rotated ears	*absent semicircular canals*	489.50	6/42	8
Deafness,sensorineural	*saccular degeneration*	3639.12	2/6	3
Over-folded ear helix, lop ear	*saccular degeneration*	998.50	3/14	4
Microphthalmia	*iris hypoplasia*	1056.17	3/9	4
Photophobia	*iris hypoplasia*	7776.39	1/1	2
Arched eyebrows	*abnormal occipital bone morphology*	515.58	3/6	8
Ptosis of eyelids	*absent palatine shelf*	707.83	4/27	5
Blocked/absent nasolacrimal duct	*abnormal calvaria morphology**	937.02	1/1	7
Palpebral fissures slant down	*abnormal cranial base morphology*	100.68	18/45	36
Epicanthic folds	*small branchial arch*	160.26	15/55	19
Broad base to nose	*abnormal parietal bone morphology*	222.75	8/16	13
Large nose	*abnormal secondary palate development*	529.91	5/8	7

High/prominent nasal bridge	*abnormal maxillary shelf*	420.45	6/18	8
Bulbous nasal tip	*abnormal occipital bone morphology*	1310.70	1/1	5
Asymmetric face	*abnormal tooth morphology*	130.73	6/10	21
Thin/Long face	*abnormal ear distance/ position*	1882.87	3/8	4
Cleft upper lip (non-midline)	*abnormal malleus morphology*	488.28	6/8	9
Palate,general abnormalities	*abnormal orbital bone morphology*	236.86	7/16	12
Cleft palate	*abnormal skull morphology*	58.68	26/37	56
Malocclusion of teeth	*malocclusion*	2052.18	2/3	4
Voice,general abnormalities	*malocclusion*	1846.12	3/6	5
Atrial septum defect	*abnormal lung vasculature*	300.25	8/37	14
Ventricular septal defect	*intracranial hemorrhage*	180.95	17/41	23
Split hands	*ectrodactyly*	6201.11	1/1	2
Camptodactyly	*abnormal phalanx morphology*	179.08	9/23	16
Perthes/dysplastic hip	*abnormal long bone morphology*	891.43	1/1	4
Cleft foot	*ectrodactyly*	6201.11	1/1	2
Over-riding toes (inc.clinodactyly)	*abnormal skeleton extremities morphology*	201.55	2/2	16
Syndactyly 2-3 of toes	*abnormal metacarpal bone morphology*	331.28	6/22	11
Syndactyly of toes (not 2-3)	*abnormal metacarpal bone morphology*	796.05	4/8	9
Neuro,general abnormalities	*increased vertical activity*	639.37	4/30	6
Hypotonia	*abnormal brain morphology*	24.44	52/57	233
Mental retardation	*cochlear ganglion degeneration*	135.82	24/285	24
Seizures/Abnormal EEG	*abnormal telencephalon morphology*	46.18	36/63	86
Autistic behaviour	*abnormal hippocampus region morphology*	2764.14	3/4	3
Hyperactivity	*abnormal sensory neuron innervation*	1015.46	3/6	5
Cartilagineous exostoses	*abnormal cartilage morphology*	1295.08	2/2	4

Table 4.5: The most significant mouse phenotype enrichments amongst recurrently copy number variable regions observed in multiple patients with a shared symptom. % Enrichment: The enrichment is given as the percentage change over that expected by chance from randomly sampling the genome. **Genes hit**: The number of genes contributing to the mouse phenotype enrichment. **CNVs hit** gives the number of CNVs that harbour enrichment-contributing genes out of the total number of CNVs within that Symptom-CNV set. *These enrichments originate from largely the same CNVs.

Mouse model phenotype associations identify candidate genes

I propose that the genes contributing the significant mouse model phenotype associations are candidate genes, whose copy number change causally underlies the patient's developmental abnormality. The 46 Symptom-CNV set associations from DECIPHER, and the 101 from ECARUCA identify 595 and 1896 candidate genes respectively (**Supplementary Table 4.5**). Of these genes, 463 are identified in both of the two datasets, and of these 399 (86%) are identified from the enrichments observed in the

human phenotypes "syndactyly of toes" and "mental retardation" that are replicated across the two datasets. The 2028 total candidate genes provide a causal hypothesis for one or more developmental abnormalities for 54% of the DECIPHER patients and 96% of the ECARUCA patients. Of the candidate genes identified within DECIPHER and ECARUCA, 199/595 (33%) and 671/1896 (35%) respectively, are described as associated with human disease within OMIM (McKusick 1998). Of the 463 candidate genes identified by both of the two datasets, the proportion annotated by OMIM is similar (164/463 (35%)).

Although the combined DECIPHER and ECARUCA Symptom-CNV sets contained fewer developmental abnormalities, the increased power achieved by analysing more CNVs per human phenotype led to the identification of 2086 candidate genes drawn from 86% of the 1433 patients analysed (**Supplementary Table 4.5**). Of the 2086 candidate genes, 167 (8%) are not identified by either of the DECIPHER or ECARUCA Symptom-CNV sets individually and thus are novel. Again the number of genes previously associated with human disease within OMIM is comparable (718 candidate genes (34%)). Of the 167 new candidate genes, 101 contribute to 41 human developmental abnormalities that only have enrichments in the combined analyses.

The DECIPHER and ECARUCA patients present with an average of 7 (range = 1-102) developmental abnormalities each, and of the patients for whom I identify a candidate gene, 1,156 (78%) have more than one candidate gene. Among the patients with multiple candidate genes I asked whether the genes were associated with a single human phenotype, or distinct human phenotypes. Of the 1,258 individuals with a candidate gene, 1,147 (91%) have multiple genes associated with a single developmental abnormality (median = 2). The large number of patients for whom I observe this phenomenon, indicates that this is a general feature across the entire patients' CNVs, and not solely

occurring amongst the larger CNVs. I also examined the dataset to find instances where an individual patient's candidate genes are associated with multiple distinct LMD terms for that patient (**Figure 4.7**). I identified 18 DECIPHER patients and 25 ECARUCA patients possessing multiple candidate genes where different genes were associated with distinct symptoms. These patients represent only 4% of those patients with multiple candidate genes while the combined DECIPHER and ECARUCA analyses identified only a further 4 patients.

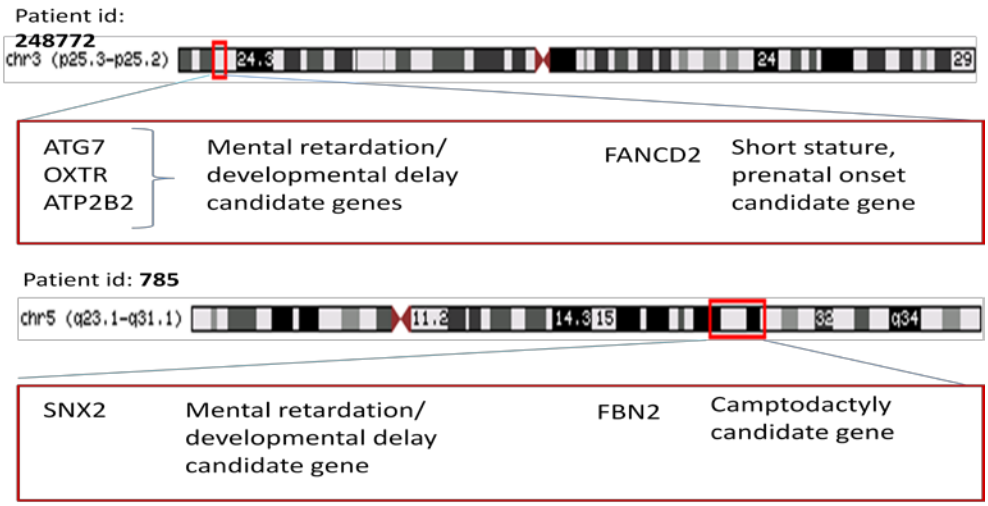


Figure 4.7: The molecular dissection of individual CNVs by human symptom. For a subset of patients, I identify distinct non-overlapping associations among the affected genes that suggest an additive pathoetiology. Two examples from a total of 47 identified are presented. For each patient, the candidate gene(s) associated with each symptom are shown.

I then examined the diversity of developmental abnormalities that each candidate gene's copy number change is associated with. Clearly this is inevitably dependent upon the symptomatic detail within the LMD definitions used to describe patients' clinical presentations (see **Figure 4.1**), and thus I examined how many human phenotypes each candidate gene is responsible for at each level of the LMD ontology. A large proportion of the candidate genes contribute to multiple developmental abnormalities whatever level of symptom specificity within the LMD ontology was considered: Of 2,086 candidate genes,

1,721 (83%) each contribute to multiple distinct human phenotypes at the most specific level of the LMD, while 1,407 (67%) each contribute to multiple distinct phenotypes at the intermediate level and 934 (45%) each contribute to multiple distinct phenotypes at the most general level. This reveals that most of the candidate genes contribute to multiple aspects of the patient's disorder, with just under half contributing to very different/distinct human phenotypes.

Analysis of candidate genes reveals insights into the biological processes that are disrupted in developmental disorders

The above analyses seek to identify copy number variable genes underlying the pathology of numerous developmental abnormalities. I identify, through exploiting the Mouse Genome Informatics database, significant biases in the protein coding gene content of CNVs observed in patients presenting with 1088 different developmental disorders (685 in DECIPHER and 892 in ECARUCA). Several mouse phenotypes suggest biological processes underlying the human phenotype, for example the *abnormal prepulse inhibition* phenotype enrichment observed amongst CNVs identified in patients with psychotic behaviour. However many of the mouse phenotypes describe structural abnormalities, such as the *decreased fetal size* mouse phenotype enrichment associated with the human phenotype low birth weight, and therefore do not provide additional information as to the biological process(es) disrupted in these disorders. To remedy this, I examined the candidate genes obtained for the 74/489 human phenotypes identified in the DECIPHER and ECARUCA combined analysis (**Supplementary Table 4.4**) for enrichments of Gene Ontology (GO) terms and direct protein-protein interactions.

Candidate genes identified for several human developmental disorders are significantly associated with gene ontology terms

I identify significant GO enrichments for 67 of the 74 candidate gene sets (**Supplementary Table 4.8**). For many of the human phenotypes the GO enrichments complement the mouse phenotype enrichment and provide further insight into the molecular processes that underlie the pathology of the disorder. For example, patients presenting with short hallux have CNVs enriched in genes associated with *abnormal skeleton extremities morphology*. The candidate genes that compromise the enrichment are associated with the GO term *fibroblast growth factor receptor activity* (see **Discussion**). I also wished to identify whether duplicated and deleted candidate genes had different roles in the underlying disease pathology of each individual phenotype. Therefore, I separated the candidate genes by the direction of copy number change (gain/loss) and repeated the GO analysis. Only one human phenotype (Cartilagenous exostoses) had significant GO enrichments amongst both the gain and loss genes. The gain genes are enriched in the GO term *fibroblast growth receptor activity*. The loss genes are enriched in the GO terms *skeletal system development, heparan sulfate N-acetylglucosaminyltransferase activity, N-acetylglucosaminyl-proteoglycan 4-beta-glucuronosyltransferase activity, glucuronosyl-N-acetylglucosaminyl-proteoglycan 4-alpha-N-acetylglucosaminyltransferase activity, heparan sulfate proteoglycan biosynthetic process* and *polysaccharide chain biosynthetic process*.

For many of the human phenotypes I observe a large number of significant GO enrichments, particularly for the human neurological developmental phenotypes. This makes it challenging to fully interpret which biological processes are disrupted in the human disorder. I aimed to reduce redundancy in the GO results by using Revigo, an online software tool that applies semantic clustering using the simrel measure to each set of GO results. By reducing redundancy within the GO terms, potential candidate

biological processes underlying human developmental disorders are revealed. For example, the candidate genes identified in patients with hyperactivity are enriched in GO terms under the broader GO category *positive regulation of axon extension* and *neuron apoptotic processes*. For patients presenting with mental retardation I identify an association with the GO terms related to *growth factor activity*. The candidate genes identified for patients presenting with seizures are enriched in GO terms listed within the broader GO term categories *neuron migration* and *circadian rhythm*.

Candidate genes identified for human developmental disorders have a greater number of direct protein-protein interactions than expected by chance

I examined whether the candidate genes identified for each of the 74 human phenotypes with significant mouse phenotype associations are clustered in a protein–protein interaction network. I used DAPPLE (see **Methods**) to examine the direct interactions of each of the 74 gene sets. The interactions were compared against 10,000 randomisations, derived from the genome as a whole, to determine whether more interactions are occurring than expected by chance.

Of the 74 genes sets 37 (50%) possess significantly more direct protein-protein interactions than expected by chance (**Table 4.6**).

Human Phenotype	**Direct PPI Observed/ Expected (P Value)**	**Human Phenotype**	**Direct PPI Observed/ Expected (P Value)**
Abnormally shaped teeth	1/0.007 (0.007)	Abducted Thumbs	1/0.1018 (0.096)
Anteverted nares	164/52.3 (0.0002)	Asymmetric Face	55/21.9 (0.0001)
Blepharophimosis	22/5.8 (0.00009)	Blood Vessels, general abnormalities	0/0.09 (1)
Brachycephaly	4/0.55 (0.003)	Broad Barrel thorax	0/0.06 (1)
Broad base to nose	2/0.4 (0.075)	Build, thin/slender	0/0.44 (1)
Cardiac situs inversus dextrocardia	0/0.01 (1)	Cartilaginous exostoses	3/1.5 (0.19)

Cerebral atrophy heterotopias	92/33.7 (0.00009)	Cleft upper lip (non-midline)	30/10.6 (0.0001)
Coloboma of iris	46/15.3 (0.0001)	Corneal abnormalities	0/0.1 (1)
Deafness,non-specific	0/0.006 (1)	Delayed closure of Large fontanelle	241/97.1 (0.00009)
Delayed tooth eruption development	40/15.4 (0.0009)	Depressed flat nasal bridge	642/268.4 (0.001)
Dolichocephaly/Scaphocephaly	1/0.4 (0.34)	Dysplastic ears	20/2.5 (0.0009)
Feeding problems in infants	10/2.3 (0.0001)	Flat malar region	1/0.42 (0.34)
Flat nose	4/0.66 (0.004)	Genu valgum	1/0.1 (0.11)
Haemat Immunology general abn.	44/19.93 (0.0001)	High frontal hairline	0/0.006 (1)
High palate	744/286.8 (0.001)	High prominent nasal bridge	1/0.1 (0.003)
Hydroceph Large ventricles non-spec	191/68.4 (0.00009)	Hyperactivity	12/7 (0.08)
Hypoplastic supra-orbital ridges	0/0.008 (1)	Hypotonia	
Inguinal hernia	12/7 (0.08)	Irregular or crowded teeth	31/7 (0.00009)
Large nose	0/0.1 (1)	Long philtrum	2/0.58 (0.12)
Lordosis	0/0.03 (1)	Mental retardation	*
Metopic ridge	0/0.03 (1)	Microcephaly	480/201.3 (0.0009)
Mid-face hypopl. (excl.flat malar)	5/0.9 (0.003)	Mitral incompetence	6/1.4 (0.0024)
Narrow feet	0/0.09 (1)	Narrow forehead/ Temporal narrowing	2/0.3 (0.037)
Nevi or lentigenes	0/0 (1)	Open mouth appearance	4/0.56 (0.01)
Palpebral fissures slant down	359/160 (0.02)	Plagiocephaly/Asymmetrical skull	81/32 (0.02)
Posteriorly rotated ears	127/63.18	Pre-auricular pits fistulas	17/6.3 (0.02)
Pre-auricular tags	2/0.54 (0.14)	Prominent anti-helix	2/0.36 (0.03)
Prominent ears	3/0.16 (0.02)	Prominent forehead frontal bossing	464/168.68 (0.02)
Prominent lateral palatine ridges	0/0 (1)	Prominent upper lip	2/0.36 (0.059)
Proximal placement of thumb	7/1.66 (0.02)	Pulmonary incompetence	0/0.06 (1)
Recurrent infections	44/12.16 (0.02)	Restriction, supernation pronation	0/0 (1)
Small bowel atresia absence obstr.	1/0.68 (0.47)	Short hallux	1/0.5 (0.41)
Short philtrum	0/0 (1)	Simple ears	1/0.76 (0.47)
Small short nose	122/34.9 (0.02)	Syndactyly 2-3 of toes	2/0.48 (0.098)
Syndactyly of toes (not_2-3)	5/2.12 (0.039)	Synophrys	1/0 (0.02)
Vertebrae, general abnormalities	14/1.86 (0.02)	Wide nasal bridge.	13/2.92 (0.02)

Table 4.6: The number of protein-protein interactions observed amongst the candidate genes identified for each human developmental abnormality. The numbers of interactions are given as a fraction of those expected to be observed by chance. Results highlighted in green are statistically significant. *The number of genes in this test set is above the maximum limit accepted by DAPPLE.

4.5 Discusion

Results within this chapter reveal that *de novo* CNVs observed in patients with numerous developmental abnormalities exhibit significant biases in their protein coding gene products. These biases have enabled the association of mouse model phenotypes with several human phenotypes observed in patients with developmental disorders (**Figure 4.4**, **Tables 4.2, 4.3, 4.4, 4.5, and Supplementary Tables 4.2, 4.3, 4.4, 4.6**). The mouse model phenotype enrichments associated with a particular symptom-CNV set suggest a causal role for both the genes contributing to these enrichments and for the CNVs harbouring the genes in the patients' developmental disorders.

For each Symptom-CNV set, I only tested for enrichments of the mouse phenotypes described beneath the overarching mouse category deemed most relevant to the human symptom being tested. The reasoning for this was that each Symptom-CNV set is non-exclusive and each set would therefore contain disease causing genes for other symptoms not being tested, and in this experiment I wished to identify the candidate genes relevant to each individual human symptom under test. The limitations of this approach are two fold. Firstly, statistically significant associated mouse phenotypes may appear relevant to the human phenotype under investigation as I only tested the terms in the MGI deemed most similar to the human symptom. Secondly, by only testing "relevant" mouse phenotype terms there is no opportunity to identify potential associations between mouse and human phenotypes that are currently thought to be unrelated but whose associations may reveal information about the pathoetiology of the human disorder. However, several of the mouse model phenotypes are very comparable to the human developmental abnormality they are associated with. This is particularly true for

the human anatomical malformations, which offers additional confidence in these results beyond that of the statistical significance. For example, CNVs observed in patients diagnosed with "malocclusion" are enriched in genes associated with *malocclusion* in the mouse, while CNVs identified in patients with "low birth weight" are enriched in genes associated with *decreased fetal size* phenotypes in mouse (**Table 4.2**). Other mouse phenotypes are less obviously comparable to the human phenotype under investigation. For instance, CNVs observed in patients diagnosed with "complex partial seizures" are enriched in genes associated with the mouse model phenotype *abnormal circadian rhythm*, which fits well with previous research showing that patients with complex partial seizures frequently suffer seizures at a set time point during their wake-sleep cycle (**Figure 4.4**) (Yalyn *et al.* 2006). Similarly, significant enrichments are observed between patients with "psychotic behaviour" and the mouse phenotype *decreased pre-pulse inhibition*. Previous research has observed that patients with psychosis exhibit decreased pre-pulse inhibition (**Figure 4.4**) (Kumari *et al.* 2008).

The mouse phenotype enrichments are directly comparable to the human phenotypes, perhaps more so than functional annotations from other functional genomic resources (see **Chapter 5**). For example, the GO often describes annotations at a molecular level whereas the mouse phenotype abnormalities are described at the organismal level. Indeed many of the mouse phenotype descriptions within the mouse phenotype ontology are the same as those within the London Medical Database ontology, for example, the term *malocclusion* is observed in both the LMD and MPO. The annotations within the MPO can also be compared to human disorders for which we do not know the underlying cause. For example, the underlying causes of autism in humans are not fully understood, and therefore comparing this human phenotype with a significantly enriched molecular functional annotation is challenging. However, comparing a human behavioural abnormality to behavioural phenotypes in the mouse may be more comparable.

While many of the comparisons between the mouse and human phenotypes are apparent, there are challenges in their comparisons due to the differences in the mouse and human genotypes underlying the two phenotypes. The mouse phenotypes are the result of homozygous knockouts, whereas the human phenotypes are caused by copy number variable genes that are either heterozygous deletions or duplications. However, if there was no correlation between the phenotypes resulting from differing abnormal copy number of the same gene, then one would not expect to identify significant associations bar those between human patients and mouse models with an identical copy number change. Evidently, the numerous and directly comparable associations identified in this chapter suggest that the differing abnormal copy number variations of a gene most likely affect the same biological process. Corroboratively, several microduplication syndromes have been revealed to have a reciprocal microdeletion syndrome at the same loci. These syndromes are shown to affect the same organ system, e.g. Smith-Magenis syndrome (deletion) and Potocki-Lupski syndrome (duplication) (Girirajan *et al.* 2010), or the mirrored body mass index phenotypes associated with 16p11.2 gene dosage (Jacquemont *et al.* 2011). Unfortunately, although heterozygous mouse models are often generated, the resultant phenotypes are often not recorded. Of the mouse models analysed in this chapter, 23% have phenotype information for the heterozygous deletions. The candidate genes I identify in my analysis are significantly enriched in genes whose mouse orthologue's heterozygous deletions are annotated as haploinsufficient (+38% enrichment; $P<10^{-16}$) (see **Chapter 2**).

Several of the symptom-CNV sets with significant mouse model phenotype enrichments consist of CNVs observed in a single patient (10 from DECIPHER and 17 from ECARUCA). A significant enrichment obtained through a single patient challenges the generalisation of these associations to the human developmental abnormality. Nonetheless,

the mouse phenotype associations are only detectable as these CNVs affect multiple genes associated with a specific mouse phenotype (median number of candidate genes = 3.5), and therefore remain of clinical interest to the patient and their disorder.

When employing functional enrichment analyses as a general method, it needs to be considered whether significant enrichments are found because the genes considered have received more experimental consideration than those that have not. Due to the cost of generating a knock out mouse model, the gene chosen for disruption is often selected due to a hypothesis that it will be scientifically interesting. Hence, it could be the case that through the analysis of the mouse phenotypes, I am only considering copy number variable genes within the patients already thought to be involved in human disease, and thus not identifying novel disease genes. However, when examining the candidate genes for OMIM annotations, I find that only approximately 35% have been previously associated with human disease, and thus consequently this appears not to be the case in this study.

The mouse model phenotype enrichments identified in this chapter identify candidate genes for a minimum of one individual human developmental abnormality in 54% of the DECIPHER patients and 96% of the ECARUCA patients. As more knockout data are added to the MGI database/ mouse phenotype ontology, I would expect this candidate gene list to increase, and the associated mouse phenotypes to become more specific. Of the 2,086 candidate genes identified in this chapter, 17% contribute to one human developmental abnormality presented within the patient dataset, whereas 83% of the candidate genes contribute to multiple human developmental abnormalities. This implies that for the majority of the patients analysed, their overall developmental disorder is caused by the pleiotropic effects of copy number variable gene(s). As well as pleiotropy, the results in this chapter suggests that epistasis is a large factor in the patient's

phenotypes. This is demonstrated by the observation that the median number of candidate genes per patient per symptom set is 2. However, it is important to consider that the candidate genes are determined by identifying enrichments of genes with a shared annotation over what is expected by chance. Consequently, several genes within these candidate gene lists would be expected by chance and therefore are unlikely to underlie the patients' disorder, particularly when the candidate genes are obtained from significant results with small percentage enrichment values. Therefore, for some of the patients with multiple candidate genes per symptom, not all of their candidate genes will actually be causative. Over 50% of the candidate genes identified in this chapter are for the human developmental abnormality "mental retardation", which is the most common human phenotype in both the DECIPHER and ECARUCA datasets, with more than 90% of patients presenting with this abnormality. Mental retardation is one of two of the symptom-CNV sets whose mouse phenotype associations were replicated across the two CNV datasets (although replications were determined through obtaining an FDR <5% and not just through nominal significance). Using nominal validations, I am able to replicate mouse phenotype enrichments for an additional 15 human symptoms. There are many possible reasons as to the lack of overlap of mouse phenotype enrichments between DECIPHER and ECARUCA. The results may reflect that the underlying pathology of a human symptom can be caused by disrupting more than one biological pathway or process, or the results may be caused by the differences in human clinical annotations and/or the variation in the CNV calling methodology represented within the two datasets. Alternatively, it may be that the significant enrichments that I identify are not biologically relevant to the human symptom and are instead the result of uncontrolled for biases that occur when applying functional enrichment approaches to gene lists obtained from *de novo*

CNVs (See **Chapter 8**). Nevertheless, the observed increases of number, percentage enrichment, and/or specificity of the mouse phenotype enrichment when examining the "combined" and "recurrent" symptom-CNV sets suggest that many of the substantial sources of noise will be reduced as the number of patients/CNVs increases.

The significant enrichments of gene ontology terms amongst the candidate gene lists provides additional evidence as to which biological processes are disrupted in individual human developmental abnormalities. For example, the candidate genes identified in patients with hyperactivity are enriched with the GO term *positive regulation of axon extension*. The enrichments of direct protein-protein interactions amongst several of the candidate genes suggest that the genes are interacting in a shared biological pathway or process. For those where significance is not reached, this may be due to a lack of power or an incomplete network within DAPPLE. As further candidate genes and protein-protein interactions are identified more significant results may be observed.

In summary, the FEA analysis in this chapter objectively links multiple human developmental disorders with mouse model phenotypes. These associations identify over 2000 candidate genes that contribute to the human disorders. Finally, this chapter illustrates the importance of the ongoing collection of centralised clinical data that is consistently annotated and freely available, in facilitating large-scale genomic approaches of human disease analysis.

Chapter 5: Comparing the utility of functional genomic resources to identify candidate genes underlying developmental abnormalities

5.1 Abstract

In addition to mouse model phenotypes, there is a wide range of different functional genomic resources available to utilise in functional enrichment analyses. Within this chapter, I examined the *de novo* DECIPHER and ECARUCA CNVs for enrichments of genes associated with several different functional annotations, namely Gene Ontology terms, KEGG pathways or genes with a high level of tissue specific expression. I identify significant enrichments amongst 547 of the human phenotypes, implicating several of the CNVs and their overlapping copy number variable genes in the pathology of the human developmental abnormalities. By comparing the results obtained from each of the different functional genomics resources, I identify and discuss potential biases amongst different functional resources; including ascertainment biases in functional resources formed from laboratory experiments and biases from paralogues in functional resources formed from computational experiments. Following this, I examined the role of non-protein-coding

elements in human developmental abnormalities identifying significant enrichments of miRNAs and conserved non-coding elements amongst *de novo* CNVs identified in patients. Although more sparsely populated than the protein-coding functional resources, I identified significant associations between miRNAs with functional genomic annotations and specific human developmental abnormalities.

5.2 Introduction

In **Chapter 4** I exploit mouse phenotype data resulting from the disruption of 1:1 mouse:human orthologues, to identify candidate genes within sets of *de novo* CNV genes associated with individual human developmental disorders. The mouse genome informatics database is one of many resources available that contain functional genomic annotations suitable for functional enrichment analysis (FEA) approaches. Other functional resources include protein-protein interaction data, gene expression data, literature based evidence, cellular location data and genetic interaction data.

The Gene Ontology (GO) is the most commonly exploited resource used in FEA (Ashburner *et al.* 2000). GO consists of three ontologies that describe the molecular function, the biological process or the cellular location associated with a gene. These annotations arise from a variety of different sources, including direct experimental evidence, computational analysis and literature reviews. GO has been used to identify functional enrichments amongst genes overlapped by CNVs in studies of both idiopathic and inherited disorders. This is demonstrated by Kariminejad *et al.,* who examined rare CNVs observed in patients with various brain malformations for enrichments of GO terms (Kariminejad *et al.* 2011). Comparing against a control set of randomly generated CNVs, they observed an enrichment of genes associated with axonal transport within the disease associated CNV set. Grond-Ginsbach *et al.* examined inherited CNVs amongst patients

with cervical artery dissection for enrichments of GO terms amongst the overlapping genes (Grond-Ginsbach *et al.* 2012). For this analysis the pre-existing tool GO term mapper (Schroeder 2011) was employed, and several identified enrichments including collagen fibril organisation.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is another well used example of a functional genomics resource containing protein coding gene annotations (Kanehisa *et al.* 2008). KEGG annotates protein coding genes with metabolic and disease pathways using data obtained from the literature. Webber *et al.* exploited the KEGG resource to examine whether a set of *de novo* CNVs observed in patients with mental retardation are enriched in genes associated with one or more KEGG pathway (Webber *et al.* 2009). Webber *et al.* identified an enrichment of genes associated with *Parkinson's disease* (one of 6 of KEGG's neurodegenerative pathways) and an enrichment of genes associated with a combined set of all the KEGG neurodegenerative pathways amongst the mental retardation associated CNV set. More recently, all of the KEGG neurodegenerative pathways have been used (in combination with other genomic annotations) to create a classifier capable of separating benign and pathogenic CNVs observed in patients with mental retardation (Hehir-Kwa *et al.* 2010).

In addition, gene expression data can be employed to examine whether CNVs overlap genes that have a high expression in a tissue related to the developmental abnormality under investigation. This approach can be used to implicate the CNV in the developmental disorder; however this will not provide evidence of the disrupted biological pathways underlying each disorder.

In addition to functional genomic resources containing information associated with protein-coding genes, there are several resources documenting functional annotations for

non-protein coding regulatory elements. Recently, the proportion of non-coding regions within the genome thought to be functional has increased, suggesting that copy number variable non-coding regions may underlie developmental abnormalities (Stamatoyannopoulos *et al.* 2012). At present, resources focussed on non-protein coding genomic elements are less common and are less well populated with annotations than those that concentrate on protein coding genes, however they are still proving valuable for identifying disease causing elements. The Human Micro RNA disease database (HMDD) is one such example, and annotates miRNAs with disease annotations through literature based evidence (Lu *et al.* 2008). Tacutu *et al.* exploited this database to examine the role of cellular senescence and age related diseases (Tacutu *et al.* 2011), reporting that the observed connectivity between genes associated with cellular senescence increases greatly when microRNA with an age related disease annotation are included in the network.

In this chapter, I associate GO and KEGG terms with a wide range of human developmental abnormalities catalogued in DECIPHER and ECARUCA. I also identify an enrichment of genes with a high foetal/adult expression ratio within both sets. Through exploiting functional genomic resources I provide evidence that implicates conserved non-coding elements and microRNAs in the underlying pathology of a range of different developmental disorders. In addition, I examine and discuss the biases encountered with the different functional annotation resources.

5.3 Methods

CNVs and patient phenotypes

Within this chapter I exploit two CNV datasets obtained from patients presenting with a wide range of developmental abnormalities. The two disease sets are obtained from the

DECIPHER database and the ECARUCA database and consist of 626 (median size = 2.2Mb) and 1143 (median size = 17.3Mb) CNVs respectively (Feenstra *et al.* 2006; Firth *et al.* 2009). The CNVs within the DECIPHER and ECARUCA sets are of *de novo* inheritance (see **Chapter 2**). The patients within DECIPHER and ECARUCA present with multiple developmental abnormalities (median = 5 and 8 respectively) described using terms from the London Medical Database ontology (LMD) (Fryns and de Ravel 2002) (see **Chapter 2**). Using the LMD ontology, the CNVs within the DECIPHER and ECARUCA set were grouped into non-exclusive CNV associated with each term within the LMD that is observed within the patient sample.

I also make use of three control CNV datasets (see **Chapter 2**). The Shaikh *et al.* set consists of 54462 CNVs (median size = 8.1Kb), the AGP set consists of 2537 CNVs (median size = 96.5Kb) and the Nijmegen *et al.* set consists of 361 CNVs (median size = 0.74Mb) (Nguyen *et al.* 2008; Shaikh *et al.* 2009; Pinto *et al.* 2010). CNVs from all three sets were observed in healthy individuals. For each of the three datasets, the CNVs were further subdivided by copy number direction (gain/duplication and loss/deletion).

Assigning ENSEMBL genes to CNVs

Genes were defined by ENSEMBL Ensmart 54 (Flicek *et al.* 2010). I assigned protein-coding genes to a CNV if the given CNV overlapped a minimum of one protein-coding exon from every known transcript of that gene. This ensures that the CNV affects protein-coding sequence whichever transcript is expressed. This gene-CNV annotation method reduces the effect of length biases associated with genes that show tissue-specific expression patterns (Webber, 2011) (see **Chapter 2**).

Functional Enrichment Analyses

Gene ontology (GO) terms, KEGG pathway terms and gene expression data were obtained from their respective online resources (Harris *et al.* 2004; Su *et al.* 2004; Kanehisa *et al.* 2008) (see **Chapter 2**). For each Symptom-CNV set within DECIPHER and ECARUCA databases, I tested for an enrichment of genes associated with one or more gene ontology terms, KEGG pathway terms or genes with a high foetal to adult expression ratio. As in **Chapter 4**, I employed a hypergeometric test to examine the null hypothesis that *de novo* CNVs observed in patients with developmental disorders randomly sample all genes. To account for multiple testing, I applied a false discovery rate (FDR) <5% (Benjamini and Hocjberg 1995).

Non-protein coding gene annotations

I obtained a set of 718 miRNA and 20940 lincRNAs from the UCSC genome browser and a set of 7025 conserved non-coding elements from Greg Elgar's lab (see **Chapter 2**) (Edwards *et al.* 2006). I examined whether the DECIPHER and ECARUCA disease associated CNVs and the Shaikh *et al.*, AGP and Nijmegen *et al.* benign CNVs were enriched in any of the three non-coding elements using the Genomic Association Tester (GAT) (see **Chapter 2**).

To further examine the role of miRNAs in developmental disorders I obtained a list of 396 microRNAs and their associations to 277 human diseases from the human microRNA disease database (HMDD) (Lu *et al.* 2008). I also downloaded a set of 283 microRNAs and their 868 experimentally identified downstream targets from the online resource mirTarBase (see **Chapter 2**) (Hsu *et al.* 2011). These data enabled me to examine the miRNAs overlapped by CNVs, and their downstream targets for functional enrichments.

5.4 Results

I sought to objectively associate functional genetic annotations from online genomic resources with human developmental disorders, in order to identify biological processes underlying the developmental abnormalities and to replicate or identify additional candidate genes identified in **Chapter 4** that are responsible for the patient's disorder. To achieve this I examined the *de novo* CNVs within the DECIPHER and ECARUCA sets for an enrichment of genes associated with a specific gene ontology term, KEGG pathway or a high ratio of foetal to adult tissue expression level.

For each of the human abnormalities within DECIPHER and ECARUCA I created a non-exclusive set of *de novo* CNVs drawn from the subset of patients with that symptom (Symptom-CNV sets). As in **Chapter 4**, I further subdivided the CNVs to form groups of gain and loss CNVs for each human symptom. For each set, the protein coding genes affected by CNVs were identified using ENSEMBL.

Associating protein-coding gene annotations with human developmental abnormalities

Of the 685 DECIPHER and 892 ECARUCA symptom-CNV sets examined, I identified 368/685 and 498/892 sets respectively that are significantly enriched (FDR <5%) in genes annotated with a specific GO term (**Table 5.1, Supplementary Tables 5.1 and 5.2**). Again, within both CNV datasets some of the enrichments are comparable to the human developmental abnormality. However, most are less directly comparable than the mouse and human phenotypes observed in **Chapter 4** (see **Discussion**) (**Table 5.2**). Within the GO analysis I observe many human symptoms with the same GO enrichment (see **Discussion**). This is most true for the GO term *keratin filament* and *receptor activity*. The GO term *keratin filament* is significantly enriched within CNVs associated with 28

developmental abnormalities in DECIPHER and the GO term *receptor activity* is associated with 143 human developmental disorders in ECARUCA.

Human Symptom	**Gene Ontology Enrichment** (% Enriched) [Gene Count] |CNVs hit|	**Mouse Phenotype Enrichment** (% Enriched) [Gene Count] |CNVs hit|
Low birthweight (< 3rd centile)	keratin filament (535.78) [22] |2/25|	decreased fetal size (190.10) [17] |10/25|
HAEMATOL/IMMUNOLOGY	keratin filament (669.25) [22] |2/17|	abnormal T cell activation (180.27) [20] |11/17|
SEIZURES, general abnormalities	cornified envelope (386.11) [14] |4/89|	increased sensory neuron number (252.49) [13] |16/89|

Table 5.2: Comparison of Gene Ontology and mouse phenotype enrichments amongst CNVs associated with developmental disorders in DECIPHER. Enrichments are given as the percentage change over that expected by chance. The number of CNVs with at least one gene contributing an enrichment is given as a fraction of the total number of CNVs associated with that symptom.

When examining the distribution of candidate genes across the CNVs I observe that many candidate genes reside within a small proportion of the CNVs tested. This differs from the mouse phenotype enrichment analysis (**Chapter 4**) which observes fewer candidate genes per patient, however the genes are spread across a greater percentage of the CNVs tested (**Table 5.2**) (see **Discussion**). I hypothesised that this may be due to the method of assigning GO terms within the GO. Many GO terms are assigned using sequence similarity, and therefore CNVs overlapping paralagous genes resulting from tandem duplications would contain multiple genes with the same annotation, potentially resulting in this term reaching statistical significance even though it is only observed in the minority of the patients' CNVs in the sample. To test whether the enrichments observed were in fact representative of the entire patient sample and of the human disorder I repeated the analysis using only experimentally derived GO terms. This analysis identifies fewer human developmental abnormalities with significant GO enrichments (62 human disorders in DECIPHER and 157 within ECARUCA) (**Supplementary Tables 5.3 and 5.4**). However, on average the candidate genes have a wider distribution across the CNVs

with a median of 50% of the CNVs in DECIPHER overlapping a candidate gene (increased from 25%) and 98.08% in ECARUCA (increased from 40%) (see **Discussion**).

I examined the 685 DECIPHER and 892 ECARUCA symptom-CNV sets for an enrichments of genes associated with one or more KEGG pathways. After correcting for multiple testing, 288 and 547 CNV sets were found to be enriched in one or more KEGG pathways (**Supplementary Tables 5.5 and 5.6**). The KEGG pathway terms are less directly relatable to the human phenotype in comparison to the results I obtained in **Chapter 4** using the mouse phenotype data. Indeed, for many human phenotypes I observe enrichments of multiple, seemingly unrelated KEGG pathways. In addition the majority of the KEGG pathways are enriched amongst several human disorders and a specific link between a particular human disorder and biological pathway cannot be made (**Supplementary Tables 5.5 and 5.6**).

I hypothesised that genes underlying developmental disorders are more likely to be expressed within developing foetal tissue than human adult tissue. To test this I examined both the entire set of DECIPHER and ECARUCA *de novo* CNVs for an enrichment of genes with a high (top 5%) foetal/adult tissue expression ratio. I identify a significant enrichment (8.71 %) of genes over that expected by chance of genes with a high foetal to adult expression ratio within DECIPHER. This enrichment is formed of 401 candidate genes overlapping CNVs observed in 56 patients. The ECARUCA set has a 7.64% enrichment of foetal to adult gene tissue expression ratio. The enrichment consists of 185 genes found to be copy number variable in 69% of patients.

Examining the contribution of non-coding elements to human developmental disorders

The mouse phenotype analysis in **Chapter 4** and the GO, KEGG and gene expression analysis described above concentrates on finding biological processes underlying a patient's developmental disorder through examining protein coding genes overlapping a CNV. To examine the role of non-protein coding genomic elements in the underlying pathology of human developmental disorders, I examined the DECIPHER and ECARUCA *de novo* disease associated CNVs and three sets of CNVs observed in healthy individuals (Shaikh *et al.*, AGP and Nijmegen *et al.*) for an enrichment of conserved non-coding elements (CNEs), long intergenic non-coding RNAs (lincRNAs), and microRNAs using GAT (**Figure 5.1**).

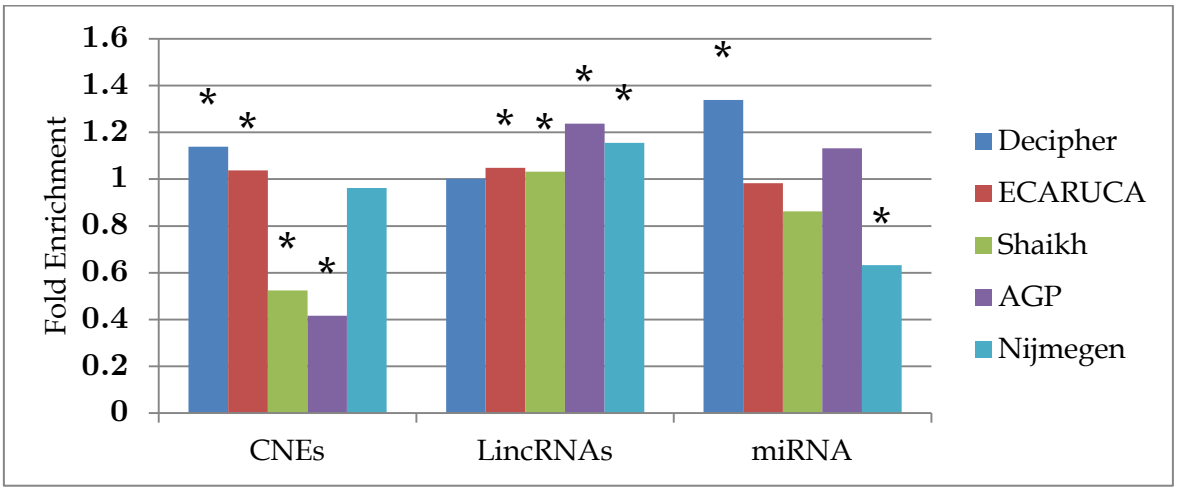


Figure 5.1: Fold enrichments over that expected by chance of conserved non-coding elements (CNEs), LincRNAs and miRNAs amongst the 5 CNV datasets. Enrichments are given for each dataset and an asterisk indicates that the fold enrichment or depletion is significant at a P value of less than 0.05.

No large differences of enrichments and depletions of lincRNAs are observed between the disease and control CNV sets. There is a small significant enrichment of CNEs and miRNAs in the DECIPHER set, however this is not replicated by the ECARUCA dataset. Two of the benign CNV sets are significantly depleted in CNEs, however the third set (Nijmegan) does not follow this trend. There is a similar for the miRNAs, with only the

Nijmegen set showing a significant depletion and the other two benign sets showing smaller but not statistically significant depletions.

I decided to explore the significant enrichment of microRNAs amongst the DECIPHER set further. To achieve this I exploited two functional genomic resources, the HMDD (human microRNA disease database) (Lu *et al.* 2008) and MirTarBase, an online resource of miRNAs and their experimentally derived downstream targets (Hsu *et al.* 2011). I examined the 685 DECIPHER symptom-CNV sets for an enrichment of miRNAs associated with one of the 18 overarching human disease categories relevant to the human developmental disorder under investigation. Of the 685 CNV sets 143 were enriched in miRNAs associated with a human disease (**Supplementary Table 5.7**). Some of the disease enrichments are relevant to the human disorder (e.g. the association between the human phenotype "vision, general abnormalities" and miRNAs annotated with *eye disease*), however many associations are spurious (e.g. the association between the human phenotype "deafness" and miRNAs with an *abdominal phenotype* annotation) (see **Discussion**).

To examine whether the downstream targets of the disrupted miRNAs are involved in the underlying pathology of the patients developmental abnormality, I examined the experimentally validated downstream targets (obtained from mirTarBase) of miRNAs overlapping each symptom-CNV set for an enrichment of genes associated with one or more mouse phenotypes (see **Chapter 4**). Of the 685 miRNA-Symptom-CNV sets, 155 were enriched in genes associated with mouse phenotypes, with each human symptom possessing a median of 24 mouse phenotype enrichments (**Supplementary Table 5.8**). Surprisingly, this number was much higher than the number of enrichments seen in the protein-coding gene analysis described in **Chapter 4**, but can be explained by the biases

within the mirTarBase. The MGI database contains mouse phenotype data for 1/3rd of all human:mouse orthologues but 2/3 of all genes within mirTarBase (see **Discussion**).

5.5 Discussion

Results within this chapter provide additional evidence that *de novo* CNVs observed in patients with the same developmental disorders house significant biases in their protein coding gene content as well as their non-coding element content. These biases enable the association of gene ontology terms and KEGG pathway data with individual human phenotypes as well as implicating miRNAs and conserved non-coding elements (CNEs) in human developmental disorders.

This chapter enables the comparison of the utility of different functional annotations in varying circumstances. At the beginning of the chapter I examined CNVs associated with each developmental disorder in DECIPHER and ECARUCA for an enrichment of copy number variable genes associated with one or more GO terms. GO is one of the most widely used resource in functional enrichment studies (Webber 2011) and consists of three different acyclic ontologies describing cellular location, molecular functions and biological processes associated with protein-coding genes. Gene ontology terms are assigned to genes using a variety of methods (e.g. experimental evidence, inferred from literature, computational assignment through sequence similarity) and thus some annotations may be more reliable than others. The GO contains annotations for many more genes than the MGI resource (see **Chapter 4**), and this feature of the GO enabled the identification of significant enrichments for more human abnormalities within DECIPHER and ECARUCA than I could identify using the MGI database (**Chapter 4**). The enrichments however are not specific to the human abnormality, with many of the same GO terms reaching significance across multiple disorders. However, when I examined the genes

contributing the enrichments they are clustered within a small number of patients (**Table 5.2**), whereas the genes from the mouse phenotype enrichments are distributed across a larger proportion of the patients tested. Many of the significant GO enrichments are therefore not representative of the biological pathways involved in the human disorder and are instead a result of the way that the GO annotations have been assigned to genes. Many of the annotations are assigned computationally using sequence similarity. This will result in runs of paralogous genes formed from duplication events being annotated with the same GO term, which if overlapped with a CNV will elevate the GO terms towards significance. This trend needs to be kept in mind particularly when using FEA approaches on structural variants that overlap multiple genes. To identify GO term enrichments representative of all patients sharing a developmental abnormality I repeated the analysis using the GO terms annotated to genes using experimental evidence. With only ~20% of the GO term's annotations resulting from experimental evidence, this drastically lowered the number of significant enrichments, however it increased the proportion of patients per symptom for whom I can identify a candidate gene. In summary, when implementing an FEA using GO, the trade-off between power (the number of GO annotations) and reliability (the method of functional annotation) must be considered.

The majority of functional enrichment resources contain annotations for protein coding genes. However, with many patients within the DECIPHER dataset possessing CNVs that do not overlap genes I wished to examine the role of non-coding elements in developmental abnormalities. Significant enrichments of both copy number variable miRNAs and conserved non-coding elements were identified amongst the DECIPHER patients, suggesting their disruption may underlie some patients symptoms. The potential role of miRNAs underlying disorders observed in the DECIPHER patients is supported by the observation that miRNAs are highly expressed in the brain and affect synaptic plasticity (Serafini *et al.* 2012), and of all DECIPHER patients, 98% have mental

retardation. Also, single miRNAs are known to regulate multiple genes (Lewis *et al.* 2005), therefore, their disruption may underlie the numerous developmental abnormalities observed in some patients. I examined the role of miRNAs in developmental abnormalities further by examining each symptom-CNV set for enrichments of miRNAs associated with known human diseases, or whose downstream gene targets are associated with a specific mouse phenotype. Although 143 symptom-CNV sets are enriched in miRNAs associated with an overarching human disease category, the enrichments are not very comparable to the human phenotype under investigation. This is likely due to the lack of power in the dataset, for example even when testing the broad overarching categories in the HMDD the number of miRNAs with these annotations ranges from 6 to 72. As freely available online resources for non-coding elements become as well populated as those that currently exist for protein coding genes there will be more opportunity to identify specific significant enrichments relatable to the human disorder under investigation. I also intended to examine the downstream targets of miRNAs for significant enrichments of mouse phenotype annotations and subsequently compare them to the enrichments found in **Chapter 4**. Of the 685 symptom-CNV sets within DECIPHER I identify significant mouse phenotype enrichments for 23%. The reason for the large number of enrichments is due to the biases within the mirTarBase database. The MGI contains phenotype annotations for ~30% of all human genes however there are annotations for >80% of the target genes listed in mirTarBase, therefore comparing the number of annotations within downstream targets while using the entire human genome as the background is not appropriate. The downstream targets within mirTarBase are identified using experimental evidence, and usually the miRNA and/or the target gene is chosen for analysis as it is hypothesised to be biologically interesting and therefore the selection of miRNAs and annotated genes within mirTarBase are not representative of miRNAs across the whole human genome.

In summary, this chapter has shown how freely available online resources can be exploited to implicate both coding and non-coding genomic elements in human developmental disorders, as well as identifying possible biological processes and pathways underlying these disorders. I also show the importance of recognising potential biases within both the functional genomic data and structural variant data, and how these will potentially affect results.

Chapter 6: Analysis of CNVs observed in patients whose developmental abnormalities are not described using a medical ontology

6.1 Abstract

In order to perform functional genomics analyses it is necessary to obtain sets of CNVs identified in patients with the same developmental abnormality. In **Chapters 4** and **5** this process was facilitated by the patient's phenotypes being described by the London Medical Ontology. However, there are many readily available CNV datasets where the patient's phenotypes are not described using a medical ontology which, in turn, makes the grouping of patients with the same phenotype challenging. In this chapter, I exploited one such dataset from the Guys and St Thomas's cytogenetics laboratory. I employed pattern matching and MeSH terms to assign each human phenotype in the dataset to one of the terms in the London Medical database. I examined sets of CNVs identified from patients with each individual human developmental abnormality for an enrichment of genes associated with one or more specific mouse phenotypes. I identified 13 human phenotypes with a specific mouse phenotype enrichment, with 4 of the human phenotypes replicating the mouse phenotype enrichments observed in the DECIPHER and

ECARUCA datasets. Following this, I examined an additional set of patients presenting with autism spectrum disorder. I identified several significant enrichments of mouse model phenotypes and Gene Ontology terms amongst the copy number variable genes in these patients. I identify many functional enrichments relating to the synapse, suggesting disruption at the synapse may underlie some instances of autism spectrum disorder.

6.2 Introduction

The work discussed in **Chapters 4** and **5** examines sets of CNVs observed in patients whose abnormal phenotype is described using terms described by the London Medical Ontology (LMD) (Fryns and de Ravel 2002). The patient's phenotypes within **Chapter 4** were annotated with a strict ontology enabling me to group patients with the same specific phenotype together. Consequently, this provided the necessary power to identify functional enrichments amongst the patient's copy number variable genes. There are however many CNV datasets in existence that have been obtained from patients whose abnormal phenotypes are described using multiple different ontologies or even without the use of a medical ontology.

The LMD is one of many existing medical ontologies, for example, the human phenotype ontology (HPO), the disease ontology (DO), Online Mendelian Inheritance in Man (OMIM) and the MeSH (Medical Subject Headings) (McKusick 1998; Lipscomb 2000; Osborne *et al.* 2009; Robinson and Mundlos 2010). Medical ontologies are useful as they enable patients with the same human phenotype to be grouped together. Due to their often hierarchical structure they also enable patients with comparable phenotypes to be grouped together under a broader phenotypic term, although which patients are grouped together may vary depending on the structure of each ontology. For example, an ontology that groups abnormal human phenotypes via closely related anatomical structures will

form different groupings to those that use disrupted physiological processes. An anatomically driven ontology may group temporal lobe and occipital lobe epilepsy in different sub-sections due to the seizures originating in different regions of the brain. However, a different ontology using groupings based on similar physiological disturbances may group these two disorders closer together due to their shared underlying cause of abnormal synchronous neuronal activity. An ontology based on the outward patient presentation may place these two disorders under entirely different sections due to the very different symptoms caused by these two disorders (temporal – déjà vu; occipital – visual disturbances). Grouping different sets of patients together will affect the ability of FEA analysis to identify significant enrichments. As well as the grouping together of patients, ontologies can also be used to group together pre-existing studies for further analysis (Beck *et al.* 2012).

Much readily available patient data are not phenotyped using a strict medical ontology. Clearly this may provide problems when attempting to group together patients, particularly when the analysis involves large numbers of human phenotypes. There are, however, potential advantages to analysing patient data that have not been recorded in an ontological format. Ontologies use strict definitions which limit the detail to which a clinician can describe a patient's phenotype. For example, the London Medical Database has only one entry for mental retardation, which doesn't allow the clinician to describe the severity of the patient's phenotype if this is apparent. Secondly, most ontologies do not allow clinicians to annotate the absence of a patient's phenotype. However, when using ontologies it cannot be presumed that a lack of an annotation is equivalent to the patient not presenting with that phenotype, as ontologies often contain hundreds of descriptors and it is unlikely a patient has been tested for all of them (see **Chapter 2**). Finally, through using a strict ontology, data regarding how the patient was tested for each of the phenotypes that they present with are lost. Although many human phenotypes

are subject to specific tests before diagnosis can be reached (e.g. autism), many phenotypes are measured more subjectively (e.g tics, sleeping disorders and feeding problems) where information regarding the method of diagnosis may be useful.

In this chapter, I analyse a set of CNVs observed in patients obtained from the Guys and St Thomas's cytogenetics lab. These patients are diagnosed with a wide range of developmental abnormalities, but are, however, not described using a structured medical ontology. Using both the London Medical database and MeSH (Medical Subject Headings) I used pattern matching to group CNVs identified in patients with a shared phenotype. I then examined each group of CNVs for an enrichment of overlapping genes associated with a mouse model phenotype, Gene Ontology term or KEGG pathway. I identify significant enrichments of copy number variable genes associated with all three resources amongst the patients, implicating the CNVs in the patient's disorders. Finally, I examine a larger set of CNVs observed in patients with autism also obtained from the Guys and St Thomas's lab. I associated these CNVs with several mouse phenotypes that are comparable to the patient's autism presentation, for example *abnormal learning, memory and conditioning*. The lack of a medical ontology enabled me to segregate the autism patients by the severity of their disorder for additional analysis. Further investigation of the autism candidate genes suggests that disruptions in synaptic pathways may underlie the same patient's phenotype.

6.3 Methods

CNVs and Patient Phenotypes

Patients presenting with a wide range of developmental disorders (median = 3) were obtained from the Guys and St Thomas's hospital cytogenetics lab. Amongst these patients a total of 1843 CNVs were identified using various microarray platforms (**Table**

6.1). Unlike the DECIPHER and ECARUCA CNV sets, the patient phenotypes are not described using the London Medical Database, or indeed any other medical ontology. Instead, each patient phenotype is described using terms deemed most suitable by each individual clinician who originally assessed the patient. Consequently, patients with the same disorder can be annotated with variety of terms. For example, patients with autism can be labelled with "autism", "ASD" and "autism-spectrum".

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	1843	0.6Mb	13240	5
(Gain)	903	0.6Mb	10743	6
(Loss)	940	0.5Mb	7778	4
(1) De novo	218	1.4Mb	5621	18
(Gain)	52	1.4Mb	2261	17
(*Loss)*	166	1.4Mb	3635	18
(2) Inherited	643	0.4Mb	2609	4
(Gain)	347	0.4Mb	1969	5
(*Loss)*	296	0.2Mb	1072	2
(3) Unknown	985	0.7Mb	10936	6
(Gain)	507	0.7Mb	9242	7
(*Loss)*	478	0.6Mb	4957	8

Table 6.1: The number, size and gene coverage of CNVs within the Guys and St Thomas's dataset. CNVs are split by inheritance. (1) *De novo*: CNVs observed in the child but not the parent. (2) Inherited: CNV observed in the patient and one of their parents. (3) Unknown: CNVs of unknown inheritance. Each CNV set is split by copy number direction into gain and loss sets. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Grouping of the Guys and St Thomas's CNVs by patient phenotype

In order to identify functional enrichments of genes amongst *de novo* CNVs observed in patients presenting with a shared human developmental abnormality I needed to group CNVs by patient phenotype. To group patients with a shared disorder I employed a pattern matching procedure to assign each of the patient phenotypes to one of the medical annotations described by the London Medical Database. By assigning MeSH term

descriptors to each of the LMD terms, I increased the chance of each Guys and St Thomas's phenotypes being successfully matched to a LMD term.

For each different developmental abnormality defined by the LMD, I formed a group of *de novo* CNVs (symptom-CNV sets) drawn from those patients annotated with that LMD term. The median number of CNVs assigned to an LMD term is 3 (range 1-113). As in the DECIPHER and ECARUCA CNV sets (see **Chapter 4**), ~90% of patients are annotated with more than one LMD term, and consequently CNVs are not exclusively assigned to a CNV set. For each Symptom-CNV set I generated a list of overlapping copy number variable genes (see **Chapter 2**) and genes observed to be copy number variable in apparently healthy individuals were removed from each list.

Mouse model phenotypes

I obtained mouse model phenotype resulting from the disruption of 1:1 mouse orthologues of human genes from the Mouse Genome Informatics database. Using 1:1 human:mouse orthology, I assigned the mouse phenotype terms to ~5000 human ENSEMBL genes. For each CNV set associated with a London Medical database term, I examined the copy number variable genes for an enrichment of 1:1 mouse orthologues associated with a specific mouse phenotype (see **Chapter 4: Methods** for additional details). As ~90% of the patients present with multiple developmental abnormalities, each of their CNVs are assigned to multiple Symptom-CNV sets. To concentrate the analysis of each set toward the LMD term of interest, I tested only the mouse phenotype terms in the ontology listed under the overarching mouse phenotype terms deemed most relevant to the human symptom being investigated (**Supplementary Table 4.1**). I compared the gene content of each Symptom-CNV set to the genomic background. I used a hypergeometric test to test the null hypothesis that each Symptom-CNV set randomly samples genes from the

genome. As I tested multiple mouse phenotype enrichments for each Symptom-CNV set I employed a multiple testing correction (False Discovery Rate <5%).

Gene Ontology, KEGG pathway and gene expression data

I obtained a list of gene ontology terms, KEGG pathway terms and gene expression data from three distinct online resources (see **Chapter 2**). For each of the different CNV sets annotated with a London Medical database term, I examined the copy number variable genes for an enrichment of GO terms, KEGG pathway terms or a high foetal to adult gene expression ratio. As with the mouse phenotype analysis, I tested the null hypothesis using the hypergeometric test and a multiple testing correction.

Analysis of the Guys and St Thomas's autism dataset

I obtained an additional 322 CNVs from the Guys and St Thomas's dataset, all of which were observed in patients diagnosed with autism, plus a wide range of additional human phenotypes. The wide variety of clinical descriptors used within the dataset enabled me to group the patients into subgroups based on the severity of their autism and their developmental delay/mental retardation (**Table 6.2**).

Overarching category	**Patient Phenotype**
Autism **mild/unconfirmed**	?ASD | ? ASD | Atypical ASD | atypical autism | Atypical autism | ?autism | ? autism | autism? | ?Autism | ? Autism | Autism/?LD | ?Autism spectrum disoder | Autism spectrum disoder? | Autism spectrum disoder (being assessed) | Autism spectrum disoder - features of | Autism spectrum disoder (under assessment) | Autism spectrum disoder - under consideration | Autism spectrum disoder (under investigation) | ? Autism spectrum disorder | Autism spectrum disorder? | autism spectrum disorder-high functioning | "Autism spectrum disorder - possible, under investigation" | ?Autistic | autistic features | Autistic

	features | autistic features (awaiting formal assessment) | ?autistic spectrum disorder | autistic traits | features of ASD | features of autism | likely autism | mild autistic spectrum LD | obesity with autistic features | possible ASD | possible autism | possible Autism spectrum disorder | possible mild features of autistic spectrum disorder | some features of autism | undergoing assessment for autism spectrum disorder
Autism	ASD | autism | Autism | Autism likely | Autism spectrum | Autism spectrum disoder | autism spectrum disorder | Autism spectrum disorder | Autism Spectrum Disorder | autistic | Autistic | autistic spectrum | Autistic spectrum condition | autistic spectrum disorder | Autistic spectrum disorder| Autistic Spectrum Disorder | childhood autism | Clinically autistic | developmental delay suggestive of ASD | diagnosis of autism | disintegration disorder/autism | ? Early onset of autism | History of LD/ASD | on the autistic spectrum | "some difficulties, on autism spectrum"
Autism severe	Severe ASD | severe autism
Development normal	cognitive develoment typical | No delay | Typical cognitive development
Developmental delay mild	cognitive development mild | Mild cognitive delay (IQ 50-69; for adults mental age 9-12 yrs) | mild delay | Mild delay (atypical) (IQ 35-49;for adults mental age 6-9yrs) | mild & variable learning problems
Developmental delay	?atypical cognitive development | atypical delay | atypical / moderate dev delay | cognitive development delay (Atypical) | delay atypical | Delayed (atypical) cognitive development | Delayed (atypical) cognitive development – global | Delayed (atypical) cognitive development - spastic diplegia | dev delay | dev delay | Dev delay | developmental delay | early dev delay | general dev delay | generalised learning difficulties | global delay | global delay including motor skills | global dev delay | Global dev delay | Global dev. Delay | global dev delay mild/atypical | Global developmental delay | "global learning difficulties (function at around 6-7 years of age, he is 10 years old)" | intellectual disability | learning diff | learning difficulties | Learning difficulties | learning difficulty | learning disability | Learning disability | mild learning difficulties | mild to moderate learning difficulties | mod dev delay | Moderate cognitive delay (IQ 35-49; for adults mental age 6-9 yrs) | moderate delay | moderate delay (atypical) | moderate delay (IQ 35-49) | moderate dev delay | moderate global dev delay | moderate learning difficulties | moderate learning disability | MR | Some degree of learning difficulties

Developmental delay **severe**	severe learning difficulty | severe learning disability | cognitive delay profound delay | cognitive development delay severe | moderate to severe learning difficulties | Profound cognitive delay (IQ <20; for adults mental age <3 yrs) | Severe cognitive delay (IQ 20-34; for adults mental age 3-6 yrs) | severe dev delay | severe global dev delay | severe global developmental delay | severe learning difficulties | Severe learning difficulties | severe learning disabilities | significant degree of learning difficulties | significant dev delay | significant learning difficulties |

Table 6.2: Patient phenotypes grouped by mild, moderate and severe autism, and grouped by normal development, mild, moderate and severe developmental delay.

To examine whether there is an unusual CNV burden or *de novo* CNV burden amongst patients with severe autism or autism with severe developmental delay, I examined the number of CNVs, amount of CNV sequence and the number of genes within the test set and randomly generated "cases" sampled from the whole of the autism CNV dataset.

For each set of CNVs I examined whether there was an enrichment of genes associated with a specific mouse phenotype from the mouse phenotype ontology (see methods of **Chapter 4**) (Eppig JT 2007). Significance was determined using the hypergeometric test and a multiple testing correction (FDR) <5% (Benjamini and Hocjberg 1995). I examined the candidate genes obtained from the MGI analysis for significant enrichments of gene ontology terms (using the same hypergeometric test and multiple testing correction as in the DECIPHER and ECARUCA analysis). I also examined the candidate genes for enrichments of protein-protein interactions using DAPPLE, an online resource hosted at the Broad Institute (Rossin *et al.* 2011).

6.4 Results

The functional enrichment analysis using the DECIPHER and ECARUCA datasets relies on the patient's developmental disorders being described consistently using terms from the London Medical Ontology. Many additional copy number variant datasets exist where the patient descriptors are derived from multiple sources, and thus the patients are not consistently annotated. This provides additional challenges and opportunities when grouping patients with a similar human phenotype (see **Discussion**). I obtained 218 *de novo* CNVs from the Guy's and St Thomas's cytogenetics laboratory. Patients present with a wide range of different developmental phenotypes (median = 3). Patients are not annotated using terms from a medical ontology and consequently patients with the same developmental abnormality are annotated with different terms. I wished to identify whether CNVs drawn from patients with a shared developmental abnormality are enriched in overlapping genes with a common functional annotation, namely mouse phenotype terms, gene ontology terms, KEGG pathway annotations and gene expression levels.

Grouping patients presenting with the same human phenotype

I used partial word pattern matching and MeSH terms to assign each human phenotype in the Guys and St Thomas's dataset to a term from the London Medical Database (LMD). I assigned the human phenotypes to 211 LMD terms. For human phenotypes that were assigned to multiple LMD terms I picked the most relevant of the assigned LMD terms by hand. If this was not possible I assigned the human phenotype to the nearest parent term of all LMD terms assigned to that phenotype term. I was left with 16 human phenotype terms that were not matched to any LMD term (**Supplementary Table 6.1**).

Mouse model phenotype analysis

Mouse phenotypes described by terms from the Mouse Phenotype Ontology were obtained from the Mouse Genome Informatics Database. I examined each Symptom-CNV set for an enrichment of genes whose disrupted mouse orthologues are associated with a specific mouse phenotype. Of the 211 Symptom-CNV sets I identified 13 with one or more significant mouse phenotype enrichments (FDR <5%) (**Table 6.3**).

Human Symptom	Mouse Phenotype	Enrichment (%)	Genes / Total	CNVs	Patients
Abnormal MRI	abnormal prepulse inhibition	743.90	7/331	3/8	3/7
Abnormal teeth	abnormal coronal suture morphology	10805.77	2/33	1/3	1/3
Bifid uvula	abnormal coronal suture morphology	14441.03	2/23	1/1	1/1
Cerebral palsy	abnormal prepulse inhibition	2735.50	6/77	2/2	2/2
Clinodactyly	abnormal long bone morphology	891.43	4/22	1/1	1/1
Dysmorphic	abnormal outer ear morphology	247.49	13/717	8/29	8/26
Floppy episodes	abnormal prepulse inhibition	2087.89	5/84	1/2	1/2
Long philtrum	abnormal coronal suture morphology	14441.03	2/23	1/1	1/1
mental retardation	abnormal cued conditioning behavior	757.29	6/234	3/9	3/8
Paralysis	abnormal prepulse inhibition	3181.83	5/56	1/1	1/1
Philtrum	abnormal craniofacial development	566.78	6/75	3/3	2/2
Polymicrogyria	abnormal prepulse inhibition	3181.83	5/56	1/1	1/1
Regression of skills	decreased prepulse inhibition	9351.67	3/12	1/1	1/1

Table 6.3: The most significantly enriched mouse phenotypes observed amongst copy number variable genes in each Guys and St Thomas's Symptom-CNV set. Enrichments are given as the percentage change over that expected by chance. The number of patients and CNVs with at least one gene contributing a mouse phenotype enrichment are given as a fraction of the total number of patients and CNVs labelled with that symptom. The full listing of all significantly associated mouse model phenotypes for each CNV set is given in **Supplementary Table 6.2**.

Several of the mouse phenotype enrichments are comparable to the human abnormalities, for example, patients with a clinodactyly phenotype have CNVs enriched in genes associated with an *abnormal skeleton extremities* morphology in the mouse. Others are less comparable, perhaps indicating biological processes underlying the human disorder, for example, the association of the mouse phenotype *abnormal sensory ganglion morphology* and the human phenotype cerebral palsy. I observed different mouse

phenotype enrichments amongst the cerebral palsy gain and loss Symptom-CNV sets. Within the loss set I observe enrichments relating to prepulse inhibition mouse phenotypes. Within the gain set I observed enrichments relating to sensory neurons. This may indicate that genomic deletions and duplications that underlie cerebral palsy do so by affecting different biological processes.

I propose the genes contributing each enrichment as candidate genes for each human developmental abnormality. For each disorder with a significant enrichment, each patient that contributes to the enrichment has more than one candidate gene. Unfortunately, the majority of the enrichments come from a very small number of patients and therefore the mouse phenotype enrichments cannot be assumed to be a general feature of the human abnormality (see **Discussion**).

Gene ontology, KEGG pathway and gene expression analysis

I examined whether the 211 Symptom-CNV sets are enriched in copy number variable genes associated with one or more gene ontology terms or KEGG pathways. I observe significant GO enrichments amongst 85 of the 211 Symptom-CNV sets (**Supplementary Table 6.3**). Although not as directly comparable as the mouse phenotypes and human symptoms, many GO term associations are comparable to the human symptom, for example, patients with "dysmorphia" have CNVs enriched in genes associated with the GO term *organ morphogenesis* and patients with the human phenotype "sandal gap" have CNVs enriched in genes associated with *abnormal structure morphogenesis*. Some GO enrichments complement the MGI enrichments observed in the Symptom-CNV sets. For example, patients with cerebral palsy have CNVs enriched in genes associated with the mouse phenotype *abnormal innervations* and the GO term *neurotransmitter activity*. For

several of the human phenotypes, the candidate genes that make up the GO and MGI enrichments do not overlap substantially (**Table 6.4**)

Human Phenotype	MGI Candidate genes	GO Candidate genes	Overlap
Abnormal MRI	7	183	3
Abnormal teeth	2	3	0
Bifid uvula	2	3	0
Cerebral palsy	6	3	0
Clinodactyly	4	0	0
Dysmorphic	13	20	6
Floppy episodes	5	6	0
Long philtrum	2	3	0
mental retardation	6	39	1
Paralysis	5	0	0
Philtrum	6	4	1
Polymicrogyria	5	0	0
Regression of skills	3	8	0

Table 6.4: The number of candidate genes identified for each human phenotype via the MGI and GO analysis and the overlap between these two sets of genes.

I examined each of the 211 Symptom-CNV sets for enrichments of one or more KEGG pathways. As I examined multiple KEGG pathways I employed a multiple testing correction for each Symptom-CNV set. I identified significant KEGG pathway enrichments for 101 Symptom-CNV sets ($P<0.05$, FDR$<5\%$) (**Supplementary Table 6.4**). Only a few of the KEGG enrichments are comparable to the human symptom under investigation, for example, patients with deafness have CNVs enriched in genes associated with the KEGG pathway *sensory system*.

Finally, I examined all the copy number variable genes within the Guys and St Thomas's dataset for an enrichment of genes with a high foetal to adult gene expression ratio. I used a hypergeometric test ($P<0.05$) to test for significance. The gene set contained a comparable number of genes with a high foetal to adult gene expression ratio (-0.44% enrichment) than what is observed across the whole genome.

Comparison of the Guys and St Thomas's results to those observed in the DECIPHER and ECARUCA CNV sets

As with the analysis of the DECIPHER and ECARUCA CNV sets (see **Chapters 4** and **5**), I observed that the mouse phenotype analysis gives more readily interpretable results than those identified with the GO and KEGG analysis. The genes contributing the mouse phenotype enrichment results are drawn from a greater proportion of the patients tested in comparison to the KEGG and GO analysis, suggesting that the mouse phenotype enrichments are able to identify a larger proportion of candidate causal elements underlying the pathology of these patient's symptoms. There are two human symptoms (clinodactyly and mental retardation) with mouse phenotype enrichments within both the DECIPHER and Guys and St Thomas's set, and three human symptoms (abnormal teeth, long philtrum and mental retardation) with mouse phenotype enrichments in ECARUCA and Guys and St Thomas's set. However, the human symptoms have different mouse phenotype enrichments across the different CNV sets, with none of the mouse phenotypes in the Guys and St Thomas's replicated in the DECIPHER and ECARUCA set.

As observed in the DECIPHER and ECARUCA analysis, I identify multiple candidate genes per patients for each of the mouse phenotype, GO annotation and KEGG pathway analyses. The large percentage enrichments within these analyses suggest that many of the candidate genes are causal. This provides further evidence that epistasis underlies many different human developmental disorders. Some candidate genes are observed amongst enrichments in more than one human symptom, suggesting that pleiotropy is involved in many patients' overall disease presentation.

Functional enrichment analysis of autism associated CNVs

I obtained an additional set of CNVs from the Guys and St Thomas's cytogenetics database whose patients' phenotypes have not been annotated consistently using a medical ontology but were all annotated as presenting with an autistic or autism relevant symptom. I examined the patients presenting with autism and a range of other developmental abnormalities (1-13 per patient) for an increased burden of CNVs and *de novo* CNVs, thus implicating their CNVs in the human disorder. I also examined the patients' copy number variable genes for a significant association with a mouse model phenotype or Gene Ontology term, thus associating biological pathways/processes with autism and identifying candidate genes for these patients and their disorder.

The autism dataset consists of 303 CNVs (see **Methods**). The median size of the CNVs within the dataset is 0.36Mb (range = 2.3kb-18.3Mb). The patients with autism present with multiple additional human developmental disorders (range 1-13). I grouped CNVs using several different methods. The first is by the type of inheritance of that CNV. The second is by the severity of the patient's autism. The third is by the severity of the patient's accompanying developmental delay (**Table 6.2**). Lastly I grouped patients by the number of additional disorders that they present with.

I examined whether patients with severe autism (5 patients), severe developmental delay (23 patients) or many additional developmental disorders (33 patients) are enriched with CNVs or *de novo* CNVs. To obtain the list of patients with multiple developmental disorders I ranked the patients by the number of additional phenotypes co-occurring with their autism, and then took forward the top 10% of these patients (patients with 7 or more accompanying disorders). Burden was calculated using a variety of methods: (1) the number of CNVs, (2) the total amount of copy number variable DNA sequence and (3)

the number of protein-coding genes overlapping a CNV. These values were compared to 10,000 random CNV datasets (consisting of the same number and sizes of CNVs as in the test set) generated from the total autism dataset (**Figure 6.1A-C**).

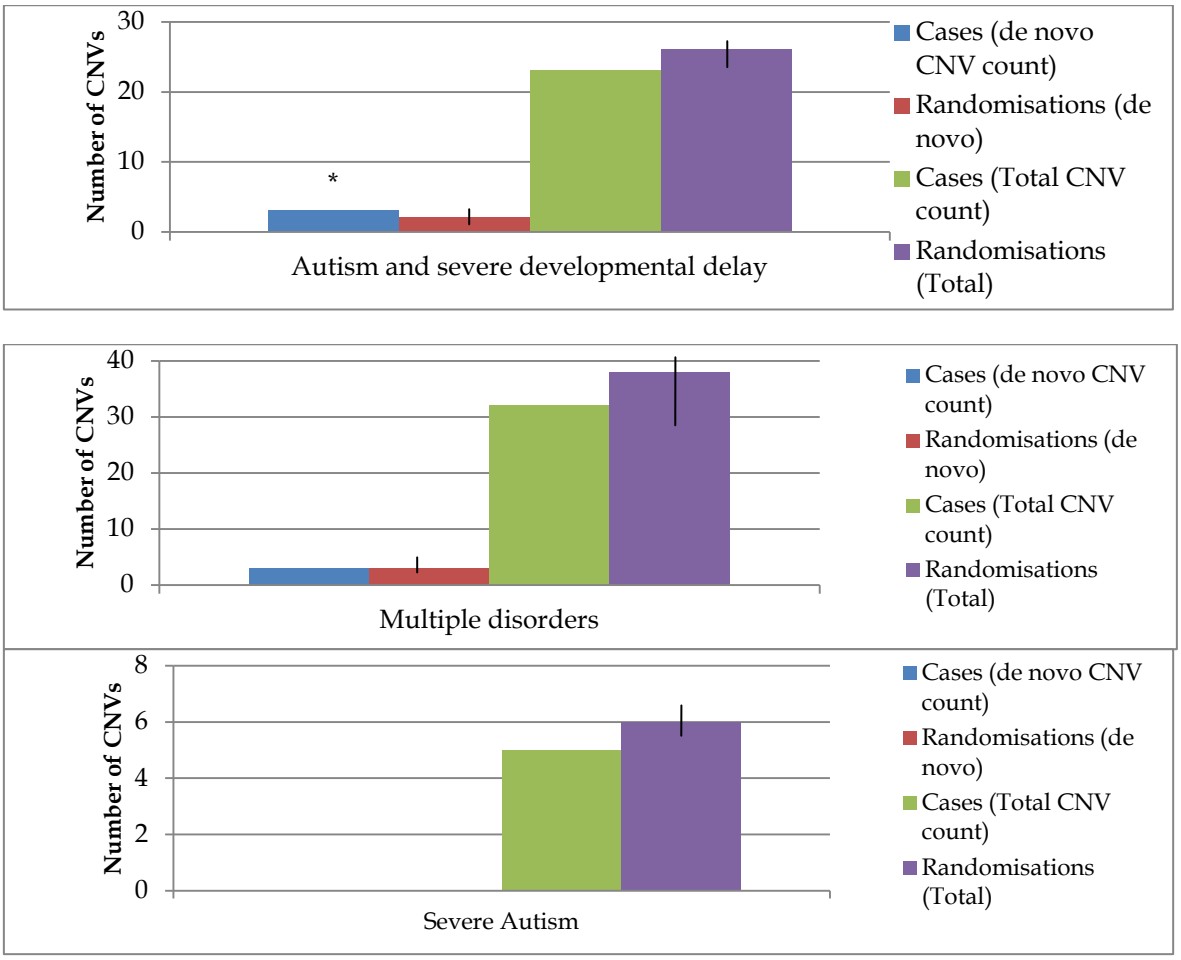


**Figure 6.1A: CNV burden calculated by examining the number of CNVs per dataset. ***The increased or decreased burden in the test cases is significant. The black line indicates the variance of the randomisations.

All three data sets have a smaller number of total CNVs compared to that observed for the randomizations (**Figure 6.1A**). Patients with severe developmental delay have a significantly increased burden of *de novo* CNVs in comparison to the median value of 10,000 randomizations.

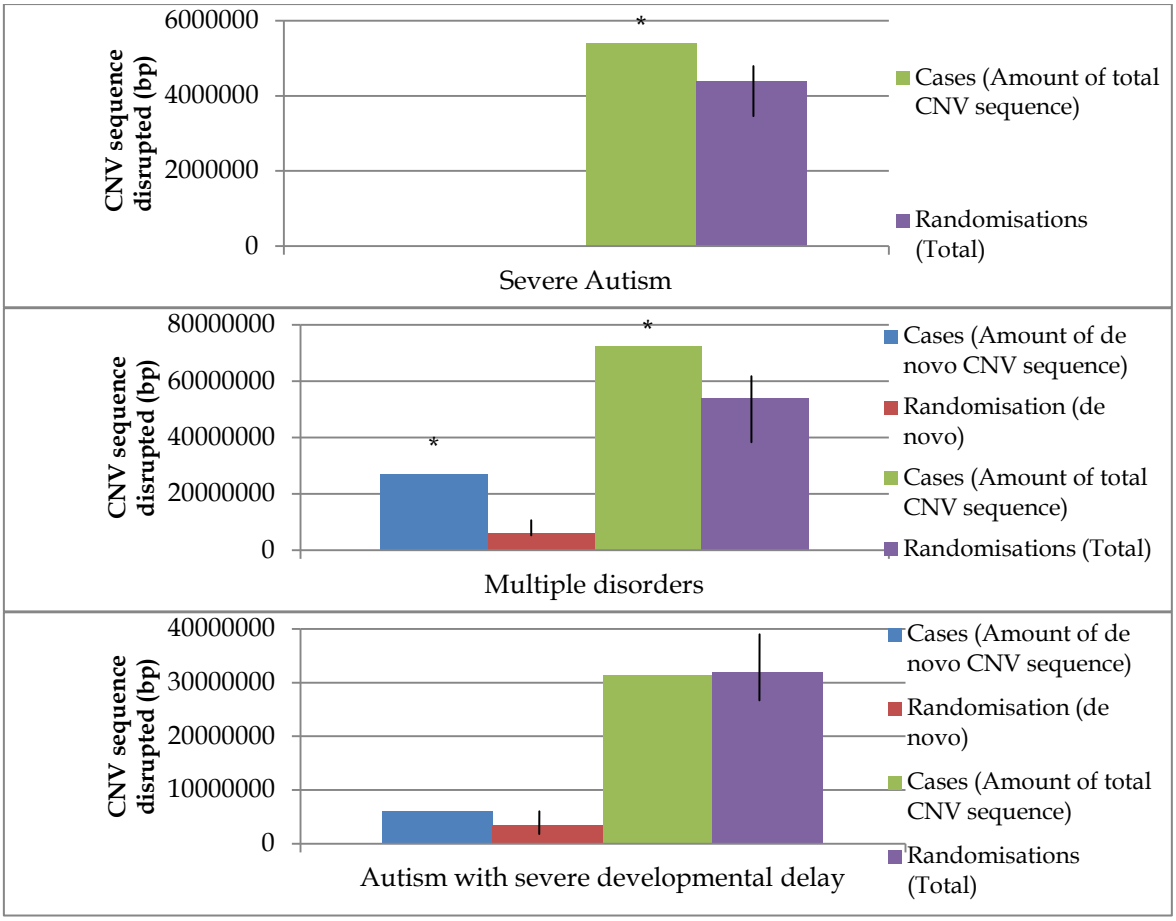


**Figure 6.1B: CNV burden calculated by examining the amount of copy number variable DNA sequence per dataset. ***The increased or decreased burden in the test cases is significant. The black line indicates the variance of the randomisations.

Patients with severe autism have a significantly increased burden of disrupted CNV sequence in comparison to the randomisations (**Figure 6.1B**). There is an increased, but not significant, burden of *de novo* sequence and of total CNV sequence in patients with severe developmental delay. Patients with multiple disorders have a significantly increased burden of both *de novo* and total CNV sequence.

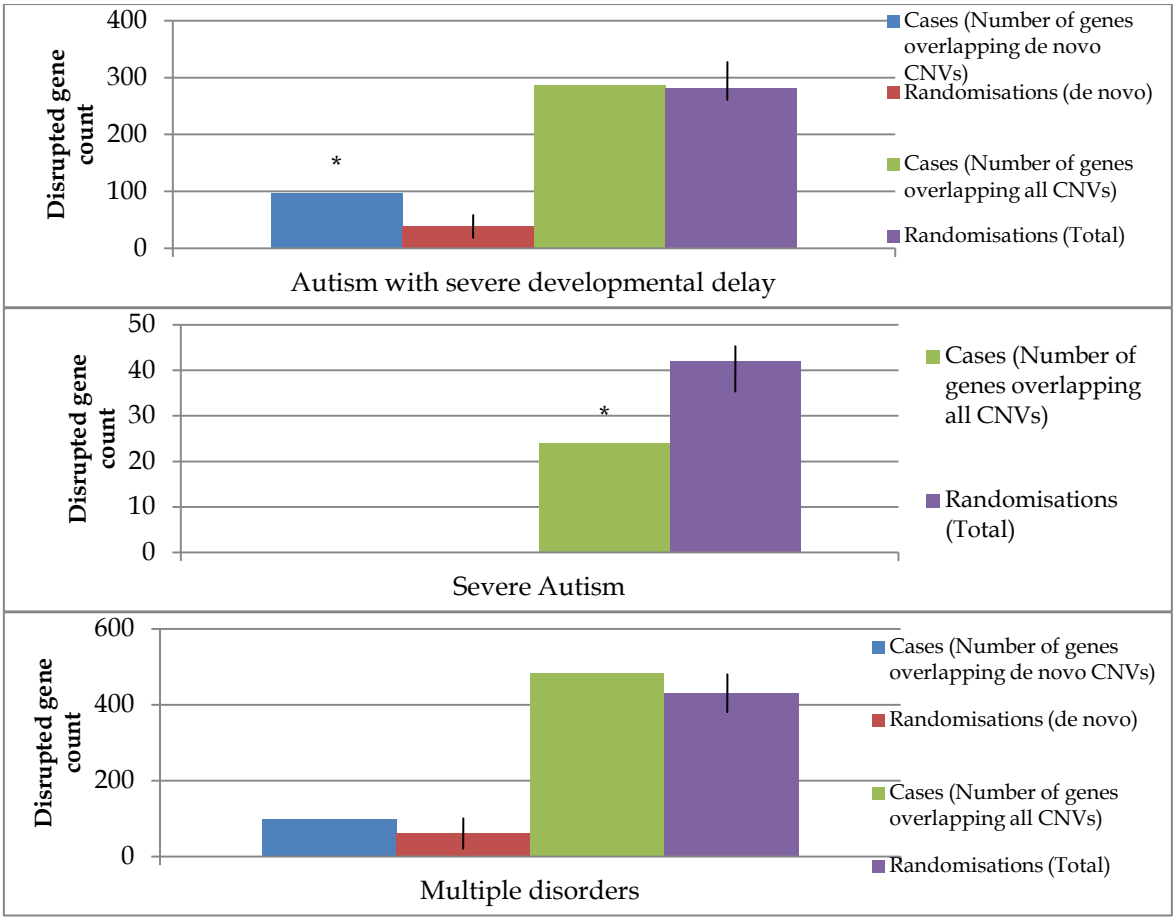


Figure 6.1C: CNV burden calculated by examining the number of copy number variable genes per dataset. *The increased or decreased burden in the test cases is significant. The black line indicates the variance of the randomisations.

Patients with severe autism have a significant depletion of genes disrupted by CNVs compared to what is observed across the 10,000 randomised CNV datasets (**Figure 6.1C**). Patients with severe developmental delay and patients with multiple additional phenotypes have an increased burden of disrupted genes observed in both *de novo* and all CNVs.

Mouse model phenotypes were obtained from the Mouse Genome Informatics Resource (MGI). The mouse phenotypes, defined using terms from the mouse phenotype ontology (MPO), are the result of the targeted disruption of 5,671 unique 1:1 human: mouse orthologues. I examined the autism CNVs for an enrichment of genes associated with one

or more mouse phenotype terms described under the overarching category *nervous system* and *behaviour/neurological*. Due to the large number of mouse phenotype terms tested I employed a multiple testing correction to ensure a FDR of less than 5%. I observe 15 different mouse phenotype enrichments across the "Inherited from an affected mother", de novo and unknown inheritance CNV sets (**Table 6.5**).

CNV set	Mouse Phenotype	% Enriched	Genes		CNVs hit
Affected mother (all)/(gain)	absence seizures	9054.29	2/5	UBE3A GABRB3	1/1
	nonconvulsive seizures	7667.27	2/5	UBE3A GABRB3	1/1
	abnormal motor capabilities/coordination/movement	357.39	5/5	GABRA5 UBE3A GABRB3 SNRPN HERC2	1/1
De novo (Gain)	abnormal cued conditioning behavior	726.84	6/62	GABRA5 ARC LYNX1 VDAC3 PLAT CREM	5/10
	abnormal associative learning	322.82	9/62	GABRA5 UBE3A SLC18A2 CREM ARC VDAC3 PRLHR LYNX1 PLAT	6/10
	abnormal contextual conditioning behaviour	496.28	6/62	GABRA5 ARC LYNX1 VDAC3 PLAT CREM	5/10
	abnormal temporal memory	412.50	6/62	GABRA5 ARC LYNX1 VDAC3 PLAT CREM	5/10
Unknown (all)	abnormal learning/memory/conditioning	69.79	44/374	NR3C2 OPA1 DOC2A ARHGEF9 UBE3A CYLN2 GDI1 MAOB MT1A ARC DSCR1 IGBP1 OTC DLG3 LIMK1 APOD PTPRA RPS6KA3 LYNX1 AFF2 ZDHHC8 NLGN3 MECP2 OPHN1 ARX DGCR8 FMR1 PSEN2 HTR2C FZD9 GTF2IRD1 MAPK3 STX1A AGTR2 NTAN1 HPRT1 L1CAM GABRA5 DCX FGF12 PAK3 DISC1 SEPT5 MAOA	28/168
	abnormal spatial learning	112.96	22/374	NLGN3 NR3C2 OPHN1 ARHGEF9 UBE3A ARX HTR2C FMR1 FZD9 NTAN1 ARC IGBP1 L1CAM OTC DCX DLG3 GABRA5 APOD PTPRA DISC1 SEPT5 AFF2	18/168
	abnormal prepulse inhibition	194.49	11/374	GABRA5 MECP2 DGCR8 TBX1 DISC1 GABRQ FMR1 GABRA3 SEPT5 GNB1L ZDHHC8	12/168
Unkno	abnormal	86.63	45/358	OPA1 DOC2A ARHGEF9	20/82

wn (gain)	learning/memory/conditioning			UBE3A CYLN2 APBA2 GDI1 MAOB ARC DSCR1 IGBP1 OTC DLG3 LIMK1 APOD RTN4R PTPRA RPS6KA3 LYNX1 AFF2 ZDHHC8 NLGN3 MECP2 OPHN1 ARX DGCR8 KL FMR1 PSEN2 HTR2C FZD9 GTF2IRD1 MAPK3 STX1A AGTR2 NTAN1 HPRT1 L1CAM GABRA5 DCX FGF12 PAK3 DISC1 SEPT5 MAOA	
	abnormal cued conditioning behaviour	219.17	13/348	NLGN3 GABRA5 MECP2 LIMK1 ARX KL GDI1 GTF2IRD1 STX1A ARC LYNX1 AFF2 IGBP1	7/82
	abnormal spatial learning	118.47	21/348	NLGN3 OPHN1 ARHGEF9 UBE3A ARX HTR2C FMR1 FZD9 NTAN1 ARC IGBP1 L1CAM OTC DCX DLG3 GABRA5 APOD PTPRA DISC1 SEPT5 AFF2	12/82
	abnormal CNS synaptic transmission	71.58	41/348	DOC2A ARHGEF9 UBE3A IL1RAPL1 APBA2 GDI1 ARC GNB1L DLG3 LIMK1 NRXN1 LYNX1 SEZ6L2 AFF2 ZDHHC8 NLGN3 MECP2 OPHN1 SYP DGCR8 FMR1 HTR2C GABRQ SYN1 GLRA2 MAPK3 STX1A L1CAM GABRA5 DCX SST FGF12 TBX1 PAK3 TRPC5 DISC1 GABRA3 SEPT5 GABRB3 CNGA2 GRIA3	16/82
	abnormal prepulse inhibition	216.49	11/348	GABRA5 MECP2 DGCR8 TBX1 DISC1 GABRQ FMR1 GABRA3 SEPT5 GNB1L ZDHHC8	6/82

Table 6.5: Significantly enriched mouse phenotypes amongst CNV observed in patients with autism. Mouse phenotype enrichments are described as the percentage enriched over that expected by chance. The number of genes contributing the enrichment and the total number of genes tested are listed. The number of CNVs overlapping a gene contributing to the mouse phenotype enrichment is described as a fraction of the total CNV tested.

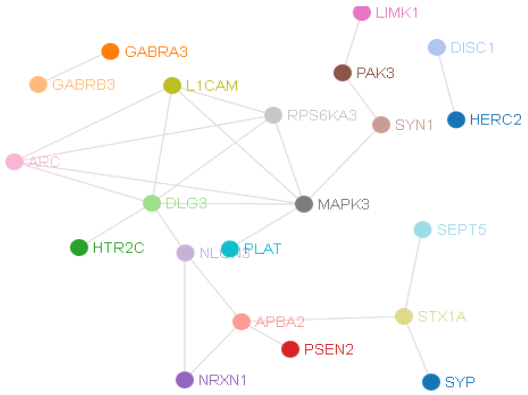
Following this, I examined the subsets of patients with severe autism, severe mental retardation and those with numerous developmental abnormalities for significant mouse phenotype enrichments to see how they would compare to the set of autism patients as a whole. The sets of CNVs obtained from patients with severe autism and severe mental

retardation were not significantly enriched in genes associated with a specific mouse phenotype after multiple testing correction. This may be due to the small number of CNVs (5, 23 and 38 respectively) within these sets. Although not passing the multiple testing correction, there are several mouse phenotypes that are significant at a single test level. The patients with severe autism have CNVs enriched in genes associated with *stereotypic behaviour, abnormal parental behaviour* and *abnormal CNS synaptic transmission.* The patients with severe mental retardation have mouse phenotype enrichments related to the structure of the nervous system, e.g. *abnormal brain morphology* (observed in the mental retardation Symptom-CNV set in DECIPHER and ECARUCA) and *abnormal neural tube morphology/development.*

I propose the genes contributing the significant (FDR <5%) mouse phenotype enrichments in **Table 6.5** as candidate genes for autism, whose copy number change underlies the patient's phenotype. The MGI analysis identifies a total of 69 candidate genes. In total this analysis provides a candidate gene for 53/322 (16%) of CNVs. Of the 69 candidate genes 33 are described as associated with human disease in OMIM, and of the OMIM annotations, 5 are for mental retardation, 1 is for learning disability and 1 is for Asperger's disorder. The median number of candidate genes per CNV is 2, suggesting it may be the combinatorial effects of disrupting multiple genes that underlies these patients autism.

To look for evidence of epistasis within the set of candidate genes I examined their direct protein-protein interaction properties using DAPPLE (Disease Association protein-protein link evaluator). Of the 69 candidate genes 21 candidate genes have direct edges with another candidate gene over what is expected by chance (Expected value = 4.59 ($P = 1x10^{-4}$)) (**Figure 6.2**). The 22 genes in the network are observed in 27 different patients and 8 patients have CNVs that overlap more than one of these genes. The genes

GABRA3, *SYP*, *NLGN3*, *HTR2C*, *DLG3*, *RPS6KA3*, *L1CAM*, *PAK3* and *SYN1* are observed in 4 patients. The genes *GABRB3* and *HERC2* are copy number variable in two



patients. *STX1A* and *LIMK1* are observed in one patient as are *PSEN2* and *DISC1*.

Figure 6.2: Autism candidate genes with a direct protein-protein interaction in the DAPPLE network.

I examined whether the set of 69 candidate genes is involved in a common biological process. To do so, I considered whether the genes are enriched for an association with a particular GO term over that expected by chance. The 69 candidate genes are significantly enriched (FDR<0.05) for 17 GO terms, many of which are involved in synaptic transmission (**Table 6.6;** ordered from most significant to lowest). The number of genes contributing an enrichment ranges from 2 – 11, and the total number of candidate genes that contribute to the GO enrichments is 30.

GO term	% Enriched	No. Genes	Genes
synapse	1078.78	11	SYN1 GABRA3 GLRA2 SYP GRIA3 GABRQ DOC2A GABRB3 GABRA5 NLGN3 ARC
GABA-A receptor activity	3825.72	5	GABRA3 GLRA2 GABRQ GABRB3 GABRA5
gamma-aminobutyric acid signaling pathway	3402.53	5	GABRA3 GLRA2 GABRQ GABRB3 GABRA5
postsynaptic membrane	1254.89	7	GABRA3 GLRA2 GRIA3 GABRQ GABRB3 GABRA5 ARC
neurotransmitter receptor activity	2463.74	5	GABRA3 GLRA2 GABRQ GABRB3 GABRA5
extracellular ligand-	2463.74	5	GABRA3 GLRA2 GABRQ GABRB3

gated ion channel activity			GABRA5
cell junction	585.95	10	SYN1 GABRA3 GLRA2 SYP GRIA3 GABRQ DOC2A GABRB3 GABRA5 ARC
ion channel activity	694.45	8	GABRA3 TRPC5 GLRA2 GRIA3 GABRQ GABRB3 CNGA2 GABRA5
Behaviour	2982.22	4	ZDHHC8 HPRT1 AGTR2 MAOA
synaptic transmission	840.79	7	SYN1 APBA2 STX1A HTR2C DOC2A SST NRXN1
synaptic vesicle membrane	6142.16	3	SYN1 SYP STX1A
chloride transport	1483.33	5	GABRA3 GLRA2 GABRQ GABRB3 GABRA5
chloride ion binding	1470.29	5	GABRA3 GLRA2 GABRQ GABRB3 GABRA5
positive regulation of exocytosis	23016.67	2	STX1A SEPT5
social behavior	4853.57	3	TBX1 GNB1L NLGN3
chloride channel activity	1993.72	4	GABRA3 GABRQ GABRB3 GABRA5
nervous system development	458.71	8	APBA2 TRPC5 OPHN1 LIMK1 FGF12 DOC2A FZD9 L1CAM

Table 6.6: Significantly enriched gene ontology terms amongst autism candidate genes identified through the mouse phenotype analysis. GO term enrichments are described as the percentage enriched over that expected by chance. The number of genes contributing the enrichment and their names listed. GO enrichments formed by the same candidate genes are highlighted in the same colour.

6.5 Discussion

In **Chapters 4** and **5** I analysed CNVs observed in patients whose phenotypic abnormalities are described using terms from the London Medical Database. Currently, there are many other medical ontologies that are used by the medical community for example, HPO and MeSH (Lipscomb 2000; Robinson and Mundlos 2010) which enable the patients presenting with the same developmental abnormality to be grouped together easily. The CNV data I analysed in this chapter were obtained from patients whose developmental abnormalities that were not annotated according to any medical ontology.

To analyse CNVs observed in patients with the same human phenotype, I grouped the patients using terms from the London Medical Database as this would enable me to compare the results against those observed in the DECIPER and ECARUCA datasets

(also annotated using the LMD). There are, however, many different medical ontologies and using a different ontology may have led to more of the human phenotypes being assigned to a category. There has previously been work to evaluate the usefulness of various medical ontologies. Yao *at al.* ranked 4 medical ontologies by measuring conciseness (e.g. whether concepts have multiple names), consistency (e.g. circularity errors) and completeness, identifying the Canonical Clinical Problem Statement System (CCPSS) as the best. At present, the ability to map terms from one medical ontology to another is limited.

The ability to group patients whose developmental abnormalities have not been annotated using terms from an ontology, and consequently to enable the identification relevant of functional enrichments amongst their CNVs, reveals the untapped resources of CNVs currently being held in cytogenetics databases. However, I was unable to assign all the patient phenotypes to an LMD annotation, with 16 not being assigned, often due to ambiguous phenotype abbreviations. Another downside is that patients grouped without the use of a medical ontology may be quite variable in their phenotypic presentation. A medical ontology entry is often followed by criteria that must be fulfilled before the patient can be assigned with the term resulting in a "minimum standard" a patient must meet before they can be assigned to a term. The patients within the Guys and St Thomas's set would consequently not have this, indeed many patients were annotated with a phenotypic descriptor followed by a question mark. However, my results suggest that many of the diagnoses may be supported by the identification of molecular commonalities in these patient's genotypes thereby proposing that they do share a common pathology.

As observed in the DECIPHER and ECARUCA analysis, I observe that the mouse phenotype analysis gives results drawn from genes from a larger proportion of patients

than the results obtained from the GO and KEGG analysis. This is due to the method of annotation of each functional genomic ontology term to each gene (see **Chapter 5**). Mouse phenotypes are annotated to genes via experimental evidence whereas GO and KEGG use a combination of experimental and computational evidence which can lead to biases if the CNVs overlap runs of paralogues, thereby resulting in a larger number of enrichments whose contributing genes are drawn from a smaller proportion of the patients tested. Each symptom-CNV set contained fewer patients than those in the DECIPHER and ECARUCA sets. This results in a lack of power to observe significant enrichments. Also many results that are observed are drawn from only a small number of patients and can therefore not be generalised to the disorder. When comparing the results from the Guys and St Thomas' analysis with the DECIPHER and ECARUCA analysis, I was unable to replicate any of the significant mouse phenotype enrichments. This may due to the small sample size of the Guys and St Thomas' data set not providing enough power to identify significant enrichments for many of the human symptoms. Alternatively, this may suggest that many of the enrichments are instead false positives and not biologically relevant to the human symptoms under investigation (See **Chapter 8**).

The autism symptom-CNV set analysed in the second half of the results section was substantially larger than any of the other Symptom-CNV sets providing the power to identify multiple significant mouse phenotype enrichments. As the clinicians were not restricted to using a single term to describe the patient's autism and additional developmental abnormalities, I was able to group patients by the severity of both their autism and mental retardation phenotype, a feature which is not possible when using the London Medical Database Ontology. I identified an increased significant burden of CNVs amongst patients with severe autism and patients with severe mental retardation in

comparison to a random randomised patient samples drawn from the total autism dataset. The mouse phenotype enrichment analysis reveals significant enrichments of copy number variable genes associated with a mouse model phenotype amongst the CNVs observed in patients with an affected mother, *de novo* CNVs and unknown-inheritance CNVs. Several enrichments are directly comparable to the human autism phenotype, for example *abnormal learning memory and conditioning*. Others are comparable to disorders known to be co-morbid with autism such as non-convulsive seizures (Munoz-Yunta *et al.* 2008). The candidate genes that I identify from this analysis are observed to share significantly more protein-protein interactions than you would expect by chance. This supports the hypothesis that dispersed CNVs within the genome underlie the same human abnormality by disrupting genes that belong to the same biological pathway or process. To identify the biological processes the candidate genes are involved in I examined them for GO term enrichments. The 69 candidate genes are enriched in 17 GO terms, with a large number of genes contributing each enrichment (median = 5). The enrichments indicate that the synapse plays an important role in the underlying pathoeitology of autism, which is supported by other recent studies (Peca and Feng 2012).

Chapter 7: Analysing the role of inherited CNVs in developmental abnormalities

7.1 Abstract

Many large CNVs of *de novo* inheritance have previously been implicated in human developmental disorders (Mefford and Eichler 2009). However, it is less well understood how inherited CNVs, particularly those inherited from a healthy parent, contribute (if at all) to a patient's developmental abnormality. In this chapter, I examine a set of inherited CNVs and a set of CNVs of unknown inheritance for a role in human developmental abnormalities. I identify significant functional annotation biases amongst genes affected by inherited and unknown CNVs identified in patients with developmental abnormalities implicating these CNVs in the underlying pathoetiology of several human developmental abnormalities. In addition, several of the significantly enriched terms observed amongst the inherited and unknown CNVs are also observed amongst *de novo* CNVs, analysed in **Chapter 4,** thought to underlie the same human developmental abnormalities.

I also examined the subset of patients with developmental abnormalities from the DECIPHER database who possess both inherited and *de novo* CNVs or who possess both *de novo* CNVs and CNVs of unknown inheritance. I identify more significant mouse phenotype enrichments amongst the three types of CNVs when analysed together than when analysed by their mode of inheritance separately. This suggests that the combined effects of inherited and *de novo* CNVs may underlie many patients' developmental

abnormalities. Across all of the 265 human developmental abnormalities considered, 14% had candidate genes that are enriched with protein-protein interactions. In addition, 28% of the human developmental abnormalities that I considered have candidate genes enriched in Reactome pathways, providing additional evidence as to which biological pathways are disrupted in specific human developmental disorders. Within the set of patients tested, I identify 4 patients who have both *de novo* and inherited copy number variable candidate genes for their developmental abnormality. In summary, inherited CNVs are an important contributor to the causes of human developmental abnormalities. It is therefore important, when analysing large *de novo* CNVs to also consider the inherited genomic background.

7.2 Introduction

Thus far my thesis has concentrated on the role of *de novo* CNVs in developmental disorders. In **Chapters 4, 5** and **6** I identify significant enrichments of mouse model phenotypes, Gene Ontology terms and KEGG pathway annotations amongst genes affected by CNVs associated with several different human developmental abnormalities, in turn implicating these CNVs as the underlying causes of these developmental abnormalities. There is, however, substantial evidence that rare inherited CNVs as well as *de novo* CNVs underlie, or contribute to, a patient's abnormal phenotype. For example, several studies have identified an increased burden of rare inherited CNVs amongst patient's with bipolar disorder and schizophrenia suggesting inherited CNVs predispose an individual to several neurological phenotypes (Van Den Bossche *et al.* 2012). In this chapter, I wished to examine whether the rare inherited CNVs within DECIPHER contribute to the patient's developmental abnormalities. To achieve this I examined the inherited CNVs identified in patients with a shared developmental abnormality for an enrichment of genes associated with a specific mouse model phenotype. Following this, I

compared significant enrichments identified amongst inherited CNVs to those identified in the *de novo* analysis **(Chapter 4**) for comparable enrichments or shared candidate genes. Comparing the affects of *de novo* and inherited CNVs can be challenging. CNVs within DECIPHER are added to the database by clinicians due to their suspected involvement in a disorder. *De novo* CNVs tend to be large, overlap many genes and are not seen in the patient's parents and consequently are usually added to the database. In contrast, both patients and healthy individuals possess inherited CNVs. If clinicians only add inherited CNVs suspected of being pathogenic to the database, not all of the patients' inherited CNVs will be available for analysis.

It is possible that a combination of *de novo* and inherited structural variants result in a patient's developmental disorder. There are two potential mechanisms as to how this could occur. The first is due to allelic exclusion and the second is due to epistasis. Allelic exclusion occurs when an individual experiences a *de novo* structural variant that disrupts one copy of a gene, while the other copy of that gene in the same individual has an inherited recessive structural variant. Ordinarily, the patient harbouring just the single recessive mutation or just the single *de novo* mutation would be healthy, however the combination of the inherited and the *de novo* mutations results in a detrimental human phenotype. For example, several patients with Cohen syndrome have been observed to possess both a *de novo* CNV and an inherited point mutation (see **Chapter 1**) (Rivera-Brugues *et al.* 2011). The second mechanism is through epistatic interactions between the inherited CNVs and the *de novo* CNVs. In this instance, the *de novo* mutation itself is sufficient to cause an abnormal human phenotype, but the additional inherited structural variant alters or increases the severity of the patient's developmental abnormality through epistatic interactions. For example, the severity of a lung disorder caused by *SFTPC* mutations has been shown to be altered by mutations in the *ABCA3* gene (Bullard and Nogee 2007).

In this chapter, I examine the DECIPHER rare inherited CNVs and CNVs of unknown inheritance for evidence that they contribute to abnormal human phenotypes. I identify significant mouse phenotype enrichments amongst inherited CNVs associated with several different human developmental disorders. Several of the observed mouse phenotype enrichments are the same amongst both the *de novo* and inherited CNV supporting a causative role for both inherited and *de novo* CNVs in human developmental abnormalities. I also examine patients who each carry both *de novo* and inherited CNVs to look for evidence that both of these types of CNVs simultaneously contribute to these patient's developmental abnormalities. I find that I identify more significant mouse phenotype enrichments when I combine the *de novo* and inherited CNVs than when I examine them separately.

7.3 Methods

Inherited CNVs, CNVs of unknown inheritance and Patient Phenotypes

In this chapter I make use of both inherited CNVs and CNVs of unknown inheritance from the DECIPHER database (Firth *et al.* 2009) (**Table 7.1**). These CNVs are identified in patients presenting with a wide range of developmental abnormalities. The dataset consists of 2480 inherited CNVs (with a median size of 0.2Mb, with each CNV overlapping a median number of 2 genes) and 1599 CNVs of unknown inheritance (with a median size of 0.3Mb, with each CNV overlapping a median of 5 genes). The unknown CNVs share more similar characteristics to the inherited DECIPHER CNVs (small, only overlapping a few genes) than they do to the *de novo* CNVs analysed in **Chapter 4** (large, overlapping many genes).

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
Inherited	2480	0.2Mb	3512	2
(Gain)	732	0.2Mb	2532	2
(Loss)	1016	0.2Mb	1556	2
Unknown	1599	0.3Mb	11169	5
(Gain)	768	0.2Mb	5852	4
(Loss)	831	0.5Mb	8270	8

Table 7.1: Summary statistics of the inherited CNVs and CNVs of unknown inheritance from the DECIPHER set. CNV size range and genomic coverage of CNVs are shown. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Of the 2480 inherited CNVs within the DECIPHER set, 480 have an annotation describing whether they were inherited from the patient's mother or father. For the remaining 2000 inherited CNVs, there is no information regarding which parent the CNV originates from.

The patients within the DECIPHER database present with a range of developmental abnormalities (median number of abnormalities = 5). The patients' phenotypes are described using terms from the London Medical Database (LMD) (see **Chapter 2**) (Fryns and de Ravel 2002). Patients present with multiple terms from the LMD, and consequently for each LMD term I formed a non-exclusive group of CNVs drawn from those patients annotated with that LMD term (herein termed Symptom-CNV sets) (see **Chapter 4, Figure 4.1**). As >90% of patients are annotated with numerous LMD terms, the majority of the CNVs are assigned to multiple Symptom-CNV sets. I assigned genes to each Symptom-CNV set using the method described in **Chapter 2.**

Control CNVs

I exploit a set of 54462 CNVs mapped to NCBI 36 that have been observed in healthy individuals as a control set (Shaikh *et al.* 2009) (**Table 7.2**). The CNVs were detected using the Illumina hapmap 550 beadchip from whole blood obtained from healthy Caucasian, African-Americans and Asian-Americans.

CNV	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	54462	8.1kb	7022	0
(Gain)	8544	39.1Kb	3495	1
(Loss)	45918	6.3Kb	5059	0

Table 7.2: Summary statistics of the inherited CNVs from healthy individuals from the Shaikh set. CNV size range and genomic coverage of CNVs are shown. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Patients with multiple CNVs of differing origins

To look for evidence of epistatic interactions between *de novo,* inherited and copy number variable genes of unknown inheritance that underly human developmental abnormalities, I created a subset of patients from the DECIPHER database who possess more than one CNV with more than one mode of inheritance i.e. patients with a *de novo* CNV and an inherited CNV (57 patients) or patients with a *de novo* CNV and a CNV of unknown inheritance (34 patients) or patient with an inherited CNV and a CNV of unknown inheritance (66 patients) or a patient with all three (20 patients) (**Table 7.3**).

CNV	Patients	Number of CNVs	CNV size (Median)	Overlapping genes (total)	Overlapping genes* (median)
All	121	3029	170.2Kb	4243	2
De novo	72	131	203.5Kb	1500	3
Inherited	104	2166	175.6Kb	2126	2
Unknown	81	726	162.2Kb	1876	1

Table 7.3: Summary statistics of the *de novo*, Unknown and inherited CNVs from DECIPHER patients with more than one CNV. CNV size range and genomic coverage of CNVs are shown. *A CNV is deemed to overlap a gene if it completely overlaps at least one protein coding exon from every possible transcript of that gene.

Mouse Model Phenotypes

I obtained mouse model phenotype descriptions from the Mouse Genome Informatics Resource (MGI) (see **Chapter 2**) (Blake *et al.* 2009). The mouse phenotypes described within this resource are the result of the experimental determined disruption of mouse genes. The mouse phenotypes are described using 5,283 terms organised under 33 overarching mouse phenotypes (Smith and Eppig 2009). Using 1:1 gene orthology relationships between mouse and human (defined by the MGI) I mapped the mouse phenotype terms to 5,671 human ENSEMBL genes.

For each Symptom-CNV set, I wished to examine whether the copy number variable genes are enriched in a mouse model phenotype annotation. For each Symptom-CNV set, I examined whether the CNVs were enriched in genes associated with one or more mouse phenotypes from within the most relevant MPO overarching category (**Supplementary Table 4.1; Figure 4.2;** see **Chapter 4.3**).

In **Chapters 4, 5** and **6** I used functional enrichment approaches and mouse model phenotypes to examine the role of *de novo* CNVs in human developmental disorders. The large size and *de novo* inheritance of these CNV sets makes acquiring a set of suitable CNV control sets from healthy individuals challenging. Consequently, in these previous analyses I compared each *de novo* CNV set against the whole genome background. Unlike *de novo* CNVs that are not frequently observed in healthy individuals, the inherited CNVs observed in patients from the DECIPHER set can potentially be compared against inherited CNVs identified in "control" patients, thereby presenting multiple different possible methods for identifying mouse phenotype enrichments amongst disease associated CNVs (**Figure 7.1**).

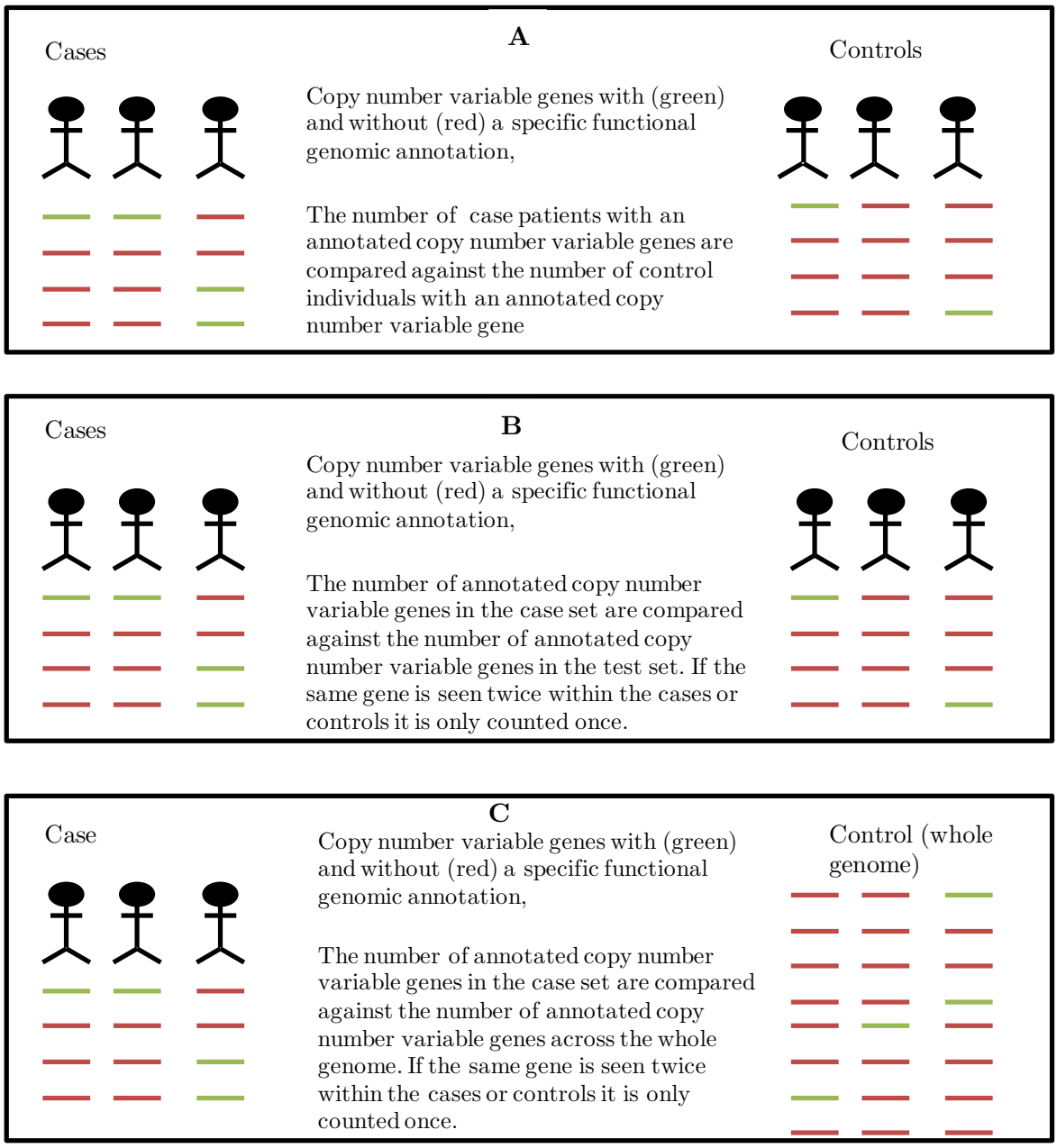


Figure 7.1: Methods for identifying functional enrichments amongst disease associated CNVs. (A): A patient based case-control frequency method. The number of patients with and the number of patients without a copy number variable gene with a given functional annotation is compared against the number of control individuals with and without a copy number variable gene with that functional annotation. **(B):** A gene centric case-control frequency method. The number of copy number variable genes with and without a functional annotation in the case and control set are compared. If the same copy number variable gene is observed more than once in the case or the control set it is only counted once. **(C):** Comparison of patients to the whole genome. The number of copy number variable genes with a functional annotation is compared to the number of genes within the whole genome with the same functional annotation. If the same copy number variable gene is observed more than once in the case set it is only counted once.

The first method I implemented is a patient based case-control frequency method (**Figure 7.1A**). For each Symptom-CNV set I counted the number of patients with a copy number variable gene associated with a specific mouse phenotype and the number of patients without. These values were compared against the number of control individuals from the Shaikh *et al.* set that do and do not have a copy number variable gene associated with the same specific mouse model phenotype. For this experiment I did not consider the total number of genes with the annotation of interest per patient in my analysis, only whether a patient or control had a CNV that affected a gene with that annotation. Significance is determined using the Fisher's exact test and a multiple testing correction ($P<0.05$, FDR$<5\%$).

The second method I employed to analyse the inherited CNVs is a gene based case-control frequency based approach (**Figure 7.1B**). For each Symptom-CNV set and separately for the Shaikh *et al.* control dataset, I identified a non-redundant list of genes disrupted by the CNVs in these datasets (see **Chapter 2**). Within each Symptom-CNV set I counted the number of copy number variable genes that were and were not associated with a specific mouse model phenotype of interest, as above (**Supplementary Table 4.1**). Following this, I compared the frequency of genes with a given annotation to that observed in the Shaikh *et al.* Control CNV gene sets. Significance was determined using a Fisher's exact test and a multiple testing correction ($P<0.05$, FDR$<5\%$).

The third method consists of comparing the Symptom-CNVs to the whole genome (**Figure 7.1C**). Here, the number of copy number variable genes associated with a specific mouse phenotype is compared to that expected by chance given the frequency of genes in the genome associated with that mouse phenotype (see **Chapter 4**). Statistical significance is determined using a hypergeometric test and multiple testing correction

($p<0.05$, FDR $<5\%$). There are several different arguments as to which of these three methods is the most appropriate to use, which I address in the **Discussion** section.

Protein-protein interactions

I examined the CNV genes contributing to each significant mouse phenotype enrichment for evidence of protein-protein interactions. I implemented this using DAPPLE (see **Chapter 2**), comparing the direct protein-protein interactions against 10,000 random gene sets drawn from the genome.

Reactome

Reactome contains descriptions for 1326 biological pathways associated with 6436 human genes (Croft *et al.* 2011). Reactome pathways and gene annotations are manually curated by "biological experts" and cross-referenced to several other bioinformatics resources, for example NCBI, KEGG and GO. For each of the candidate gene sets identified in the combined *de novo,* inherited and unknown CNV sets from the mouse phenotype analysis, I examined the copy number variable genes for an enrichment of genes associated with one or more Reactome pathways.

7.4 Results

I obtained 2480 inherited CNVs and 1599 CNVs of unknown inheritance from the DECIPHER database (Firth *et al.* 2009). CNVs were obtained from patients with a wide range (median = 5) of developmental abnormalities described by the London Medical Database Ontology (LMD) (Fryns and de Ravel 2002). To examine whether CNVs observed in patients with the same developmental abnormality share common overlapping genomic features, I created sets, for both the inherited and unknown datasets, of CNVs (Symptom-CNV sets) associated with each term within the LMD that was represented

among these patients (see **Methods**). I examined each Symptom-CNV set for an enrichment of overlapping 1:1 human:mouse orthologues associated with one or more mouse model phenotypes using three different functional enrichment approaches (see **Methods**).

Case-control analysis of patients with inherited CNVs

Unlike *de novo* CNVs, inherited CNVs observed in patients with developmental abnormalities can readily be compared against inherited CNV sets observed in healthy individuals (see **Discussion**). I wished to examine whether each inherited Symptom-CNV set from the DECIPHER database was enriched in genes associated with one or more mouse phenotypes when compared to the Shaikh *et al.* controls (see **Methods**). Using a Fishers Exact Test to determine significance, I compared the number of patients within the test and control set that did and did not possess copy number variable genes associated with a specific mouse phenotype. As I tested multiple mouse phenotypes for each Symptom-CNV set I employed a multiple testing correction (FDR<5%). As >90% of the patients present with more than one LMD annotation, the majority of the Symptom-CNV sets are non exclusive. Consequently, for each Symptom-CNV set I only tested mouse phenotypes described beneath the overarching mouse phenotype ontology category deemed to be most relevant to the human developmental abnormality under investigation (**Supplementary Table 4.1**).

Of the 571 symptom-CNV gene sets considered, 50 (8.8%) exhibit significant mouse phenotype enrichments (**Supplementary Table 7.1**). Many of the mouse phenotype enrichments are directly relevant to the human developmental abnormality of interest. For example patients presenting with "Obesity, general abnormalities" have an enrichment of copy number variable genes associated with *increased susceptibility to weight gain*.

Other enrichments provide evidence as to which biological processes are disrupted in individual human developmental disorders. For example, patients with "autism" have, in comparison to healthy patients, an enrichment of copy number variable genes associated with *abnormal synaptic depression* in the mouse. In addition, a significant enrichment of copy number variable genes associated with the mouse phenotype *abnormal cerebellum morphology* is identified in patients presenting with "short attention span".

I examined the significantly enriched mouse phenotypes from the inherited CNV analysis to those significantly enriched mouse phenotypes identified among genes within the DECIPHER *de novo* CNVs (**Chapter 4**). Only 4 human developmental abnormalities have enrichments in both experiments, namely: "Build", "Mental retardation", "Seizures" and "Brachycephaly", while only two of these four symptoms, namely "Mental retardation" and "Brachycephaly", have *de novo* and inherited CNVs enriched in the same mouse phenotypes. I examined the genes that contribute each significant mouse phenotype enrichment for each of the two human phenotypes (mental retardation and brachycephaly) to see whether the same genes are contributing to the inherited and *de novo* CNV's enrichments or whether the results in the inherited CNV sets are independent of the *de novo* CNV sets (**Table 7.4**). For both "Mental retardation" and "Brachycephaly" the genes contributing the significant mouse phenotype enrichments have a very small overlap. Thus, these replicated mouse phenotype enrichments are independent of each other and show that functionally similar genes are being hit by both *de novo* and inherited CNVs in patients with the same developmental abnormality.

Human Phenotype	De novo genes	Inherited genes	Overlap
Mental retardation	383	200	36
Brachycephaly	6	16	1

Table 7.4: Number of genes contributing the significant mouse phenotype enrichments observed in the Mental retardation and Brachycephaly Symptom-CNV sets from the inherited and *de novo* sets.

Case-control analysis of genes overlapping inherited CNVs

The enrichments seen using the patient based case-control method may have been the result of an increased burden of CNVs within the DECIPHER dataset in comparison to the controls. This increased burden may be a feature of the underlying pathology of the patients' developmental abnormalities or may be caused due to potential biases in CNV discovery between the different platforms. The CNVs within the DECIPHER database have been identified using a range of different array platforms. Conversely, the CNVs within the Shaikh *et al.* dataset are called using the Illumina hapmap 550 beadchip and therefore comparing the patients within two datasets is challenging due to the potential biases in CNV discovery between the different platforms (see **Chapter 3**). To address this, I instead compared the inherited copy number variable genes with a mouse phenotype annotation (in comparison to comparing patients in the above analysis) from the DECIPHER set to the Shaikh *et al.* Set (**Figure 7.1B**). For each Symptom-CNV set, I created a list of copy number variable genes observed in each patient. For each Symptom-CNV set I only counted each copy number variable gene once, consequently losing the power obtained from recurrently copy number variable regions that could be exploited in the previous analysis. For each mouse phenotype term described beneath the overarching mouse phenotype term deemed most relevant to the human symptom under investigation, I examined whether the DECIPHER Symptom-CNV sets were enriched in genes associated with that mouse phenotype in comparison to the Shaikh *et al.* set. Significance was determined using a Fishers Exact test ($P<0.05$) and a multiple testing correction (FDR $<5\%$).

Of the 571 Symptom-CNV sets tested, 205 have significant mouse phenotype enrichments (**Supplementary Table 7.2**). Many of the mouse phenotype enrichments are

comparable to the human disorder under investigation. For example, CNVs obtained from patients with "Generalised Obesity" are enriched in genes associated with *obesity* in the mouse. As in the previous analyses, some of the significantly enriched mouse phenotypes provide an insight into the disrupted biological processes that underlie human developmental abnormalities. For example, the CNVs identified in patients presenting with "Dyspraxia" have an enrichment of copy number variable genes associated with *impaired coordination* and the inherited CNVs observed in patients with "self mutilation" are enriched in genes associated with *hyperalgesia* in mice. Of the 205 human symptoms with significant enrichments in the gene case-control analysis, 16 (8%) share the same significant enrichments using the patient case-control method described in the preceding section, and I observe 45 (22%) symptoms with significant but different mouse phenotype enrichments across both sets.

Again, I compared the significantly enriched mouse phenotype results from the inherited case-control analysis to those mouse phenotypes observed in **Chapter 4** from the *de novo* DECIPHER CNVs. I identify 11 human symptoms with an enrichment in both the gene based case-control and *de novo* analysis ("Prominent forehead/frontal bossing", "Large nose", "Nasal speech", "Back and spine", "Haematology/Immunology", "Psychotic behaviour", "Parxymal disorders, general abnormalities", "Ataxia, general abnormalities", "Pyramidal signs, general abnormalities", "Spasticity/brisk reflexes/Babinski", "Neuroradiology, general abnormalities"). "Paroxymal disorders, general abnormalities" and "Ataxia general abnormalities" have the same significant mouse phenotype enrichment in the gene-case-control and *de novo* analysis. I examined the genes that contribute the enrichments across the two experiments and found that they did not overlap, making the two results independent of each other (**Table 7.5**). Both of these enrichments however, are formed from a small number of genes obtained from a small number of patients.

Consequently, the mouse phenotype enrichments cannot be generalised to the human abnormality.

Human Phenotype	De novo CNV genes	Inherited CNV genes	Overlap
Paroxymal disorders	6	1	0
Ataxia	5	1	0

Table 7.5: Number of genes contributing the significant mouse phenotype enrichments observed in the paroxysmal disorders and ataxia Symptom-CNV sets from the inherited and *de novo* sets.

Analysis of inherited CNVs against the genome

The inherited CNVs within the DECIPHER set are added to the database as they are suspected to be pathogenic (most often due to their rarity within human populations). Consequently, not all inherited CNVs that a patient possesses will be listed within the DECIPHER database. Therefore, comparing the inherited CNVs in DECIPHER patients to those observed in controls may not be appropriate. Given this, I also compared the DECIPHER inherited Symptom-CNV gene sets to genes across the whole genome (**Figure 7.1C**). I examined each Symptom-CNV set for enrichments of genes associated with one or more mouse phenotypes described under the MPO overarching category deemed most relevant (**Supplementary Table 4.1**).

Of the 571 human symptoms under investigation, I identify 10 (2%) with significant mouse phenotype enrichments (**Table 7.6, Supplementary Table 7.3**). As before, many of the enrichments are comparable to the human phenotypes under investigation, for example, patients presenting with "Simple ears" have CNVs enriched in genes associated with *abnormal spiral ligament morphology* in the mouse. Also, other mouse phenotype enrichments provide insights into the disrupted biological processes that may underlie human developmental abnormalities. For example, patients with "self mutilation"

have an enrichment of copy number variable genes associated with *abnormal sensory neuron innovation* in the mouse.

I examined the genes that contribute to the significant mouse phenotype enrichments, and I identify that the median number of candidate genes per patient is 2. This is observed for each of the three methods (patient case-control, gene case-control and gene case-genome) used. This suggests that epistatic interactions may play an important role in the underlying causes of human developmental abnormalities. In addition, several of the candidate genes contribute enrichments for multiple different human developmental abnormalities, providing evidence of pleiotropy within human developmental abnormalities.

Human Phenotype	**Human/Mouse Phenotype**	**Enrichment**	**Genes**	**CNVs hit**	**Patients hit**
Simple ears	type IV spiral ligament fibrocyte degeneration	8444.00%	2	1/5 20%	1/5 20%
Cleft mandible	arrest of tooth development	19925.00%	2	1/2 50.00%	1/1 100.00%
Thick/Broad neck	decreased embryo size	1129.94%	4	1/33%	1/3 33%
Asymmetric thorax	decreased embryo size	406.45%	7	1/2 50.00%	1/2 50.00%
Neurology	decreased vertical activity	138.69	25	40/1676 2.39%	34/287 11.85%
MENTAL,COGNITIVE FUNCTION, general abnormalities	decreased vertical activity	132.63	23	37/1613 2.29%	31/255 12.16%
Mental retardation/developmental delay	decreased vertical activity	138.28	23	37/1597 2.32%	31/249 12.45%
BEHAVIOURAL PROBLEMS, general abnormalities	abnormal cerebrum morphology	146.65	22	19/89 21.35%	19/60 31.67%
Aggressive behaviour	abnormal spatial learning	364.23%	6	3/6 50%	3/6 50%
Self-mutilation	impaired coordination	891.95	5	2/2 100%	2/2 100%

Table 7.6: The most significant mouse phenotype enrichments observed amongst the genes in each inherited DECIPHER Symptom-CNV set. Enrichments are given as the percentage change over that expected by chance. The number of patients and CNVs with at least one gene contributing a mouse phenotype enrichment are given as a fraction of the total number of observed patients/CNVs presenting with that human developmental abnormality.

Of the 10 human symptoms with significant mouse phenotype enrichments, 3 also had significant mouse phenotype enrichments in the **Chapter 4** *de novo* CNV analysis. These human symptoms were "Neurology", "mental cognitive function, general abnormalities" and "mental retardation/developmental delay". These three human phenotypes are grandparent, parent and child terms within the London Medical database and share a high proportion of patients. Indeed, the significant mouse phenotype enrichments are the same for all three phenotypes. However, between the inherited and *de novo* sets the mouse phenotype enrichments differ. Within the inherited set I observe an enrichment of *decreased vertical activity* whereas in the *de novo* set I observe an enrichment of *abnormal brain morphology*. In addition, the candidate genes for the *de novo* and the inherited CNVs do not overlap. This may suggest that these *de novo* and inherited CNVs contribute to the underlying causes of mental retardation by disrupting different biological pathways and that there are many biological pathways that, when disrupted, result in mental retardation. If more human phenotypic data were available, it may be possible to elucidate different sub-phenotypes of mental retardation caused by the disruption of different biological processes.

The inherited CNVs tend to be much smaller than the *de novo* CNVs within the DECIPHER set, and consequently are less likely to be considered disease causing. In addition, the inherited CNVs are observed in healthy parents and are therefore unlikely to cause disease on their own. I hypothesised that the patients may have inherited multiple CNVs, possibly from each of their parents. I wished to examine whether patients that contribute to the significant mouse phenotype enrichments do so through multiple inherited CNVs from both parents. However, the median number of CNVs contributing to an enrichment per patient is one, with only two patients that possess candidate genes originating from multiple CNVs.

Analysis of Patients with both *de novo*, inherited and unknown CNVs

Previous analyses within my thesis have examined the role of *de novo* and inherited CNVs in developmental disorders separately. I wished to examine whether it is a combination of *de novo* and inherited CNVs that contribute to a patient's developmental abnormalities. Within the DECIPHER set I observe 121 patients that possess both *de novo* and inherited CNVs or *de novo* CNVs and CNVs of unknown inheritance. The unknown CNVs share similar characteristics to the inherited CNV sets, namely their similar size and gene content. I created Symptom-CNV sets for each of the developmental abnormalities observed in the dataset. For each Symptom-CNV set, I created subsets based on the direction of copy number change (loss/gain) and also on the mode of CNV inheritance (*de novo*/inherited/unknown). I examined each Symptom-CNV set for an enrichment of genes associated with one or more mouse phenotypes over that expected by chance ($P<0.05$, FDR $<5\%$). I compared the genes within each Symptom-CNV set to the whole genome as a background (**Figure 7.1C**).

When examining patients with CNVs of different inheritance, I identify significant mouse phenotype enrichments ($P< 0.05$, FDR $<5\%$) for 7 of the 265 Symptom-CNV sets tested (**Supplementary Table 7.4**). To ensure that the enrichment signals were not all coming from the *de novo* set, I examined the contribution of candidate genes coming from CNVs of different inheritance, finding that candidate genes from each type of CNV were represented in the results.

Following this, I counted the number of human phenotypes with significant mouse phenotypes when the CNVs with different inheritance are analysed separately. I observe

enrichments for more human phenotypes when the CNVs are combined (7) than when I split the CNVs into *de novo* (5), inherited (4) and unknown (2), showing that all three types of CNVs contribute to the enrichments and therefore should be considered together when analysing an individual's developmental abnormalities. The 7 human phenotypes with significant enrichments in the combined set consist of 4 human phenotypes that are observed to have significant enrichments when analysed separately in one or more of the *de novo*, inherited and unknown sets, and 3 human symptoms that do not have significant enrichments when the *de novo*, inherited and unknown CNVs are analysed separately.

Many of the mouse phenotypes are directly comparable to the human phenotype under investigation. For example, patients with "Seizures, general abnormalities" have CNVs significantly enriched in genes associated with *abnormal neuron physiology* in the mouse and patients presenting with "mental retardation" have CNV enriched in genes associated with *abnormal amon gyrus morphology* in the mouse. By combining the *de novo*, inherited and unknown CNVs I can significantly associate human phenotypes with mouse phenotypes that are not significantly associated when examining the de novo CNVs (**Chapter 4**) and inherited CNVs (above) separately. (**Supplementary Table 7.4**). The median number of candidate genes per patient per human abnormality is 2.

To look for further evidence that it is the combinatorial effects of *de novo,* inherited and unknown CNVs that contribute to the underlying causes of human developmental abnormalities, I examined the candidate genes that I identified for the 7 human developmental disorders for a significant enrichment of direct 1:1 protein-protein interactions (**Table 7.8**). Due to the low number of candidate genes, only one human developmental abnormality (speech delay) was observed to be significantly enriched in

genes with direct protein-protein interactions. The candidate genes for each human developmental disorder are drawn from both the *de novo,* inherited and unknown CNVs (**Table 7.8**).

Human disorder	Candidate Genes	PPI (P value)
Speech delay	162 (67xD,20xI, 78xU)	60/35.1 (0.0196)
Hypopituitism	2 D,I,U (2xD)	0/0.02 (1)
Isolated growth hormone deficiency	2 D,I,U (2XD)	0/0 (1)
Autism	26 21xD,3xI,2xU	1/0.44 (0.35)
Arnold Chiari malformation	2 2xD,I,U	0/0 (1)
Microcephaly	7 3xD,I,4xU	0/0 (1)
Down turned corners of mouth	2 2xD,I,U	0/0 (1)

Table 7.8: The number of direct protein-protein interactions observed amongst candidate genes identified for each human developmental abnormality. The numbers of interactions are given as a fraction of those expected to be observed by chance. *D=*de novo*, I-inherited, U=unknown.

I examined each of the 7 sets of candidate genes for enrichments of genes associated with Reactome pathways ($P<0.05$, FDR 5%). Both the "speech delay" and "autism" candidate genes were enriched in Reactome pathways, namely *axon guidance pathways* and *potassium channel pathways,* respectively. The total number of candidate gene that contribute to the Reactome enrichments are 88 (speech delay) and 13 (autism).

I wished to identify whether the different types of CNVs (*de novo*, inherited and unknown) contribute to a patient's individual developmental abnormality. For four patients I find that the candidate genes for their individual developmental disorders come from a combination of *de novo* CNVs, inherited CNVs and CNVs of unknown inheritance (**Table 7.9**).

Patient	Human Symptom	Mouse Phenotype Enrichment	Genes from *de novo* CNVs	Genes from inherited CNVs	Genes from unknown CNV
1617	Speech Delay	abnormal synaptic transmission | abnormal CNS synaptic transmission | impaired coordination | abnormal pain threshold | abnormal behaviour	ENSG00000163273 ENSG00000078053	ENSG00000143858 ENSG00000106123 ENSG00000165125	ENSG00000065000 ENSG00000175344 ENSG00000125740 ENSG00000101180 ENSG00000101188 ENSG00000174469 ENSG00000166922
966	Speech Delay	Impaired coordination | abnormal pain threshold | abnormal learning /memory	ENSG00000081803 ENSG00000105976 ENSG00000106278 ENSG00000128573 ENSG00000170775 ENSG00000174697 ENSG00000184408	ENSG00000006704 ENSG00000106665	
	Autism	Impaired coordination | convulsive seizures | abnormal response to new environment	ENSG00000081803 ENSG00000105976 ENSG00000106278 ENSG00000128573 ENSG00000170775	ENSG00000006704 ENSG00000106665	
248657	Speech Delay	Abnormal behaviour | abnormal learning /memory	ENSG00000131791	ENSG00000204174	
248755	Autism	Impaired coordination | convulsive seizures | abnormal learning /memory | abnormal response to new environment	ENSG00000117411 ENSG00000198198		ENSG00000117411

Table 7.9 Patients for whom I identify candidate genes from a combination of *de novo* CNVs, inherited CNVs and CNVs of unknown inheritance.

For each patient I examined whether their candidate genes for each of their symptoms interacted within a protein-protein interaction network and whether these direct protein-protein interactions occurred more often than is expected by chance. To achieve this I examined where each patient's candidate genes lie within the DAPPLE network and compared the number of interactions against 10,000 randomisations (**Table 7.10**). One patient's candidate genes possess direct protein-protein interactions, although the number

of interactions is not significantly more than expected by chance. The PPIs for this patient are between a *de novo* and inherited candidate gene.

Patient	Human Symptom	Number of Direct PPIs (P-Value	Type of interactions
1617	Speech Delay	1/0.14 (0.15)	De novo - inherited
966	Speech Delay	0/0	N/A
966	Autism	0/0	N/A
248657	Speech Delay	0/0	N/A
248755	Autism	0/0	N/A

Table 7.10: The number of direct protein-protein interactions observed amongst candidate genes identified for each human developmental abnormality. The numbers of interactions are given as a fraction of those expected to be observed by chance.

7.5 Discussion

Results within this chapter indicate that inherited CNVs contribute to the underlying pathology of patients' developmental disorders. Previous work (**Chapters 4**, **5** and **6**) associated large *de novo* CNVs with multiple human developmental disorders. In this chapter, I analysed smaller inherited CNVs, finding similar biases in the protein coding genes they affect. These biases enabled me to associate mouse model phenotypes with human phenotypes observed in patients with developmental disorders, as well as implicating the copy number variable protein-coding genes and CNVs in a causal role with the disorders.

When examining *de novo* CNVs for protein coding biases, the lack of a suitable control set of CNVs led me to compare the gene content of the *de novo* CNVs against the whole genome. With inherited CNVs it is possible to identify a set of inherited CNVs in healthy individuals as a suitable control. I initially examined the DECIPHER inherited CNVs by comparing them to the Shaikh *et al.* inherited CNV controls. I examined whether the DECIPHER CNVs were enriched in genes associated with one or more mouse phenotypes,

by comparing the proportion of patients in the case set and individuals in the control set with at least one copy number variable genes associated with a specific mouse phenotype. The benefits of this analysis are that it enables recurrently copy number variable regions to be considered in the analysis, in comparison to experiments that only count copy number variable genes once even if they are observed many times in each individual dataset. Using this method, I identify 50 different human developmental abnormalities where the CNVs observed in the patients are enriched in genes associated with a specific mouse phenotype. A significant enrichment in this analysis is caused by a larger proportion of the patients with a shared developmental abnormality possessing at least one copy number variable gene associated with a specific annotation than the control individuals. Consequently, this result may be a consequence of the disease associated CNVs overlapping more functionally annotated genes. Alternatively, it may be that the patients have an increased CNV burden in comparison to the controls, which in turn would make it more likely for their CNVs to overlap a functionally annotated gene. Finally, it may be that biases between the CNV calling detection platforms between the disease and control patients may be responsible for the results. Indeed, the Shaikh *et al.* control set CNVs are called using an Illumina hapmap 550 beadchip, whereas the DECIPHER CNVs are called using a range of different detection platforms. As described in **Chapter 3**, different platforms can lead to different amounts of protein-coding overlap amongst CNVs.

As the CNVs in DECIPHER are larger than those in the Shaikh *et al.* dataset the patients within DECIPHER are more likely to possess a copy number variable gene associated with a specific mouse phenotype. To address the biases caused by the difference in the CNV size distributions between the DECIPHER and the Shaikh *et al.* set, I next examined the DECIPHER set for mouse phenotype enrichments by comparing, for each

Symptom-CNV set, the proportion of copy number variable genes associated with a mouse phenotype in the disease associated set to the control set. Although this method removes some of the biases, it also loses the power obtained from copy number variable regions. The results from this analysis reveals many more significant mouse phenotype enrichments than that observed in the patient-case-control analysis. This may be due to biases between the DECIPHER and Shaikh *et al.* sets as to which CNVs are included in the dataset. Within the Shaikh *et al.* set all inherited CNV observed in the individuals are added to the dataset. However, CNVs within the DECIPHER set are only added if they are suspected of being involved in the human developmental abnormality. As the DECIPHER set does not contain all inherited CNVs observed in each individual, the CNVs contained within the set may be more likely to overlap a functionally annotated gene as the clinician may have selected the CNV as a candidate CNV due to it disrupting a gene with an interesting known function. Consequently it is not suitable to compare the DECIPHER inherited CNVs to the Shaikh *et al.* inherited CNVs.

As the Shaikh *et al.* control CNV set may not be a suitable control for the DECIPHER CNVs I instead examined each Symptom-CNV set for significant mouse phenotype enrichments using the whole genome as a background. The mouse phenotype enrichment analysis of this dataset reveals significant biases amongst the gene content of the DECIPHER CNVs, suggesting that smaller inherited CNVs, and not just the larger *de novo* CNVs, contribute to the patients' developmental disorders. Some of the significant mouse phenotype enrichments amongst the inherited/unknown CNVs replicate those observed for *de novo* CNVs in **Chapter 4**, which provides confidence in the results. Again, as with the *de novo* CNV analysis the median number of candidate genes per patient is more than one, suggesting epistasis may play an important role in the underlying causes of human developmental disorders

The candidate genes identified amongst the combined *de novo,* inherited and unknown CNV sets are enriched in direct protein-protein interactions supporting the hypothesis that many of these CNVs underlie the same human developmental abnormality by disrupting a common biological processes. The Reactome enrichments provide evidence as to which biological processes are disrupted in "speech delay" and "autism". The candidate genes identified from patients with "speech delay" are enriched in genes associated with axon guidance pathways. The candidate genes identified in patients with "autism" are enriched in genes associated with potassium channel pathways and processes. This is supported by previous evidence that disrupted potassium channels may underlie autism (Lee and Jan 2012).

As the results from this chapter and **Chapter 4** suggest that both *de novo* and inherited CNVs may underlie human developmental abnormalities, I wished to examine instances where both of these types of CNVs are involved in the same patient. I identified a set of 121 patients who possessed either a *de novo* CNV and an inherited CNV or a CNV of unknown inheritance and a *de novo* CNV. I analysed each of these three types of CNV against the genome background separately and indentified significant enrichments, providing further evidence that all three play a role in the underlying cause of human developmental disorders. However, when combing the three types of CNVs I identify more enrichments than when the different types of CNVs are analysed separately, suggesting that it is a combination of these CNVs that cause human developmental abnormalities. It is therefore important to consider the impact of a large *de novo* CNV in the context of its inherited genomic surroundings. In addition, I find 4 patients whose candidate genes come from CNVs that have different sources of inheritance. For one of these patients I identify a protein-protein interaction between one of their *de novo* copy number variable genes and one of their inherited copy number variable genes.

In summary, I reveal that inherited CNV possess biases in their protein-coding gene content, in turn implicating these CNVs in the underlying causes of human developmental disorders. By combining inherited CNVs and de novo CNVs, I identify more significant enrichments than by analysing them separately suggesting both inherited and *de novo* mutations may act together to cause human developmental abnormalities.

Chapter 8: Conclusions and Future Perspectives

8.1 Summary

The work discussed within this thesis uses functional enrichment analysis approaches to elucidate the role of CNVs in human developmental abnormalities. Through these approaches I propose candidate CNVs, candidate genes and disrupted biological processes that underlie several human developmental abnormalities. Work within the field of disease genomics often identifies copy number variable candidate genes for human developmental disorders through the identification of overlapping copy number variable regions in multiple patients with the same developmental abnormality. Through work in this thesis I brought together sets of disparate CNVs identified in patients with the same developmental abnormalities and examined the copy number variable genes within these CNVs for enrichments of one or more terms from a range of functional genomic resources. Using these approaches, I associate CNVs and their overlapping genes with a range of different human developmental disorders. In addition, I identified possible biological processes that are disrupted in these human disorders, improving our understanding of the pathoeitiology of a several human developmental abnormalities.

In **Chapter 3**, I analysed two sets of disease associated CNVs and three sets of "benign" CNVs. I aimed to identify genomic features that differed between disease associated and benign CNVs in order to identify features that could be used to assess the pathogenicity of newly identified CNV sets. I identified significant G+C biases amongst each of the 5 sets, with differences both between and within the 2 disease-associated and 3 "benign"

CNV sets, suggesting that different CNV detection platforms are biased in the types of CNVs they are able to detect. I identified genomic features whose prevalence differ between disease-associated and "benign" CNVs, namely an enrichment of protein-coding genes, protein-coding genes associated with disease in OMIM and miRNA genes amongst disease-associated CNVs that are not observed in "benign" CNVs. I also identified genomic features that do not differ between disease-associated and "benign" CNVs, namely RNA genes and genomic repeats. Some of the results contradict previous findings; however this can be explained by previous experiments not accounting for G+C bias (Lander *et al.* 2001).

In **Chapter 4** I examined two sets of *de novo* CNVs for an enrichment of overlapping genes whose disrupted 1:1 mouse orthologues result in one or more shared mouse model phenotypes. For each of the developmental abnormalities observed in the two datasets, I created a non-exclusive CNV set. Of the 1088 developmental abnormalities considered within the DECIPHER and ECARUCA datasets, I identified 147 that have significant enrichments of overlapping genes whose disrupted mouse ortholgues result in the same mouse model phenotype. I proposed the 2086 genes that contributed these enrichments as candidate genes, whose distribution amongst patients may suggest an extensive role of pleiotropy and epistasis. In addition, many of the candidate genes sets are significantly enriched in Gene Ontology terms and direct protein-protein interactions, providing further supporting evidence for their roles in disease. For several human phenotypes, the significantly enriched mouse phenotypes propose a disrupted biological process that underlies the human abnormality, for example, the association of the human disorder complex partial seizures with the mouse model phenotype *abnormal circadian rhythm*. The most widely observed human developmental abnormality within the DECIPHER and ECARUCA databases in "mental retardation". I observe an enrichment of copy number

variable genes associated with *abnormal brain morphology* amongst patients presenting with this phenotype, identifying >300 candidate genes.

In **Chapter 5** I examined the utility of additional functional genomic resources in identifying candidate genes within *de novo* CNVs identified in patients with developmental abnormalities. Through using Gene Ontology and the KEGG pathway resource, I identified significant enrichments of functional genomic annotations amongst CNVs associated with 547 different human developmental abnormalities. Many of the genes that contribute these enrichments are also identified as candidate genes in the mouse phenotype enrichment analysis in **Chapter 4**. I also identified biases amongst different functional genomic resources, for example in the Gene Ontology resource some annotations are assigned using sequence similarity, resulting in runs of paralogues within CNVs pushing GO terms towards statistical significance. My work in this chapter indicates that the mouse phenotype ontology is the most informative resource for functional genomic analyses. In **Chapter 5**, I also examined the role of non-coding elements in the underlying pathology of human developmental abnormalities, identifying significant enrichments of conserved non-coding elements and miRNAs amongst disease associated CNVs.

In **Chapter 6** I obtained a third set of CNVs associated with human developmental abnormalities. This set differed from the CNVs analysed in **Chapters 3** and **4** as the patients' developmental abnormalities were not annotated using a medical ontology. Consequently, to carry out the functional enrichment analysis I created my own medical ontology using the London Medical Ontology and MeSH terms. In **Chapter 6**, I also analysed a larger set of CNVs observed in patients presenting with autism spectrum

disorders. The significant functional enrichments that I identified implicate that the disruption of the synapse plays an important causal role in the underlying pathology of autism spectrum disorders.

In **Chapter 7** I examined the role of inherited CNVs in human developmental disorders. In the first part of this chapter I explored the different available approaches of examining inherited CNVs for significant enrichments of functional genomic annotations, addressing the challenges of finding a suitable control CNV set and the ascertainment bias of inherited disease associated CNVs. Using functional enrichment analyses, I identify significant enrichments amongst the inherited CNVs for 45 human developmental abnormalities suggesting that inherited, in addition to *de novo,* CNVs that underlie human developmental abnormalities. By examining patients who each possess both *de novo* and inherited CNVs, I implicated both types of CNVs in many of those individual patients developmental abnormalities, providing further possible evidence of epistasis. This demonstrates the importance of examining the role of *de novo* CNVs in the context of the inherited genomic background.

Throughout my thesis I aimed to apply functional enrichment approaches to CNVs in order to identify the underlying genomic causes of a wide range of developmental abnormalities. As described above, I successfully identified numerous statistically significant enrichments of genomic functional annotations amongst the CNV sets. The large numbers of significant enrichments may indicate that the CNVs observed in these patients do indeed share commonalities supporting the hypothesis that they underlie the human developmental disorders. Alternatively, these significant enrichments may be false positives arising from uncontrolled biases within either the functional annotation

resources, biases within the test CNV set or may be the result of an inappropriate choice of control CNV dataset that I used to compare the test CNVs against.

How genes are annotated within a functional resource may introduce false positives into FEA experiments. For many functional resources, functional annotations are derived using sequence similarity approaches which can result in clusters of paralogues within the genome being annotated with the same annotation. Then, if a CNVs overlaps this genomic cluster there may be a "jack pot" effect pushing this term towards statistical significance. Indeed, I observe this affect in **Chapter 5** when exploiting the gene ontology resource, observing many of the genes that formed each significant enrichment came from a single CNV and therefore the association was not biologically relevant to the human symptom under investigation.

Alternatively, biases may have been introduced into my approach by the way the CNVs are added to the DECIPHER, ECARUCA and Guys and St Thomas' databases. Within the DECIPHER and ECARUCA databases, the CNVs are added as a clinical geneticist hypothesises that they may underlie the patients' developmental abnormality. Often a clinical geneticist will make this decision as the CNV is observed to overlap genes known or predicted to have interesting biological functions. Therefore the CNVs I performed FEA on may be pre-enriched with functional annotations compared to what is observed in all human CNVs/the whole human genome and therefore when comparing the genes within these CNVs to my control set ("benign" CNVs/all the genes in the genome) a statistical enrichment would be observed. I found this in **Chapter 7** when examining inherited CNVs in patients with developmental abnormalities. Each patient had, on average, fewer inherited CNVs than the healthy individuals in the control set but their CNVs were more likely to overlap a gene with a functional annotation than individuals in the control group.

False positives from FEA approaches can also occur from using an inappropriate control dataset. When performing FEA it is important to employ a control CNV set whose CNV distribution reflects as closely as possible the biases observed in the test set. When examining the *de novo* CNVs in **Chapters 4**, **5** and **6** I used the whole genome as the test set for two main reasons. Firstly, this was due to the unavailability of readily available sets of *de novo* CNVs that have been observed in healthy individuals. Secondly, although there are many benign inherited CNV sets I concluded they would not be an appropriate control due to their differences in size and the differences in selective pressures that act on inherited and *de novo* CNVs. However, by comparing the de novo DECIPHER and ECARUCA CNVs against the whole genome, my null hypothesis assumes that *de novo* CNVs are equally likely to occur throughout the whole genome. However, this is known not to be the case. CNV coverage across the genome is uneven, with CNVs more likely to occur in regions of the genome containing exons, segmental duplications and mobile elements, particularly Alu repeats (Cooper *et al.* 2007). Consequently, my results may contain false positives as the genomic coverage of my control set is not representative of the genomic coverage of the de novo CNVs. In **Chapter 7**, I examined inherited CNVs and consequently tried to use a control set of benign inherited CNVs in order to better match the genomic biases of the test set. Later in the chapter, I repeated the experiment using the whole genome as a control set and found that the significant results varied greatly from the results obtained using the inherited CNV control set. This demonstrates the large effect a control set has on functional enrichment analyses and the importance of identifying controls that share the same genomic biases as the test set.

In **Chapter 4** and **6** I perform FEA using the MGI on three different sets of *de novo* CNVs (DECIPHER, ECARUCA and Guys and St Thomas'). For many of the same

human developmental abnormalities I Identify significantly mouse phenotypes for each of the datasets, however there are only two human phenotypes (Mental retardation and syndactyly) where the mouse terms are replicated. This suggests that many of the significant mouse phenotype enrichments are false positives for the reasons described above. However, for the two human phenotypes where I do observe the replication of the mouse phenotype result they are two CNV datasets with a large number of patients within them. The lack of replication in some of other symptom-CNV sets may be due to their small size and consequently their lack of power. As the DECIPHER, ECARUCA and Guy's and St Thomas's datasets grow in size, it may be possible to observe an increase in replicable results using this FEA method.

I use the genes contributing the significant MGI functional enrichments as candidate genes for the patients' developmental abnormality. However it is important to note that the significant association of the mouse phenotype term does not mean that every gene contributing definitely underlies the patients' symptoms. The enrichments are calculated by comparing the number of annotated genes in the test set to the number of genes expected by chance with that annotation. Therefore, for a mouse phenotype term observed to be 200% enriched, only one in every two of those genes are not expected by chance. Consequently, this approach identifies the genes within the CNVs most likely to underlie the patients' disorder but further analysis of the candidate gene sets are needed to differentiate the true candidate genes from the false positives. Within **Chapters 4** and **6** I identify multiple candidate genes per patient which I put forward as evidence of epistasis. However, due to the false positive results it is impossible to know how prevalent the cases of multiple candidate genes per patient per symptom actually is.

8.2 Future prospects

The studies contained within this thesis analysed the role of copy number variants (CNVs) in human developmental abnormalities by utilizing functional genomics resources. This thesis demonstrates the usefulness of functional enrichment approaches when analysing the role of CNVs in human developmental abnormalities and also the importance of accounting for biases when employing for FEA approaches; including how the CNVs are identified in the test set, how the functional resources are populated and which control set should be used. As the ability to more accurately detect CNVs and the number of genes with functional genomic annotations increases, more candidate genes and other genomic elements that underlie human developmental abnormalities will be revealed.

Within this thesis I exploited three sets of CNVs identified in patients with a range of developmental abnormalities. The CNVs within these sets were called using a range of different array CGH technologies. As the cost of next generation sequencing decreases and new methods of sequencing are further developed e.g. nanopore and mass spectrometry methods, additional methods of identifying structural variants in patients will become available. Indeed, there are currently many studies that have identified causative mutations within patients using sequencing methods (Audo *et al.* 2012; Buxbaum *et al.* 2012; Shanks *et al.* 2012).

In **Chapter 3,** I observe different enrichments/depletions of genomic elements amongst three different disease associated CNV sets. These three sets were called using different platforms and the results suggest that different CNV detection methods are biased towards detecting CNVs in different regions of the genome. Although sequencing methods

possess some coverage biases, by detecting CNVs through sequencing methods this may remove some of the biases in CNV detection seen with varying CNV array CGH detection platforms. Additionally, work in **Chapters 3**, **4** and **5** show that the CNVs from the ECARUCA database are substantially larger than those in either the DECIPHER and Guys and St Thomas's dataset. This is due to the CNVs in ECARUCA being called by older detection methods; however the advent of new sequencing methods may enable the detection of much smaller CNVs.

In **Chapters 4**, **5** and **6**, I create subsets of CNVs for analysis based on whether the CNVs are losses (deletions) or gains (duplications). Although aCGH methods can accurately determine how many copies of a gene an individual has if the number is low, they cannot accurately predict higher copy numbers (i.e. above 4 copies) or determine whether a gene duplication is a cis or trans duplication or accurately call the specific breakpoints of a deletion. For example, a gene that is partially duplicated and inserted within its original gene may disrupt function and in effect act as a loss. In the case of a deletion, the result may be to create a fusion gene or to place a promoter next to a gene that it is not normally adjacent to resulting in increased expression of that gene. The prevalence and effects of the hypothetical scenarios are difficult to assess. The further development of DNA sequencing technologies will enable us to view the novel genomic structures created by copy number variation and improvements in RNA sequencing may reveal the effects of CNVs on the surrounding genes expression levels.

In **Chapter 7**, I examine the role of inherited CNVs in developmental disorders revealing that both *de novo* and inherited CNVs within a patient contribute to the developmental abnormalities. This indicates that the effects of large *de novo* CNVs need to be considered

in the context of their inherited genomic background. One area of research in this field is that of combining SNPs and CNVs in the analysis of the underlying cause of developmental abnormalities. For example, some instances of Cohen syndrome have been revealed to have been caused by the combination of a deletion CNV and a point mutation. Sequencing would identify both CNVs and single nucleotide variants simultaneously, allowing the analysis of their potential combinatorial effects.

During the section of the thesis that examined the *de novo* CNVs, I used the whole genome as a control set for my functional enrichment analyses due to the lack of an appropriate control CNV set. As our knowledge of the biases and selection pressures acting upon *de novo* CNVs increases it may be possible to generate random CNV control sets that more accurately mirror the biases within the test set, creating a more suitable control. If the experiments in this thesis were repeated using such a control many, I would expect to observe fewer but more biologically relevant statistically significant enrichments.

The ability to further investigate the role of CNVs in developmental abnormalities will also improve as functional genomics resources become more detailed. In **Chapters 4**, **5** and **6** I indentify enrichments of functional annotations amongst several CNV sets indentified in patients with a shared developmental abnormality. As further annotations are provided for a greater number of human genes, we will be able to identify significant functional enrichments for additional CNV sets. For example, the Mouse Genome Informatics resources contains phenotype data for ~1/3 of all 1:1 human:mouse orthologues. As this resource becomes more populated e.g. through the International Mouse Phenotyping Consortium, we will have the opportunity to examine many more copy number variable genes for their potential role in disease (Mallon *et al.* 2012). One of the limitations of my thesis was that the list of proposed candidate genes are not all

responsible for the human disorder. However, as more and more functional resources such as expression atlases and protein-protein interaction databases become more populated I beleice will help identify common features in these sets to identify the disease causing genes.

In **Chapter 5**, I examine the role of non-protein-coding elements in the underlying cause of human developmental abnormalities. I identified significant enrichments of miRNAs and conserved non-coding elements amongst disease associated CNVs. However, functional genomics resources for non-coding elements are far more incomplete than protein-coding resources making the identification of significant associations difficult. Recent evidence is supporting the importance of non-coding elements in human disease. For example, the ENCODE project estimates that 80% of the genome has an associated biochemical function (Dunham *et al.* 2012). In addition, recent research implicates lincRNAs as markers of disease progression, for example, PCAT-1 as a marker of disease progression in prostate cancer (Prensner *et al.* 2011).

In conclusion, analysing the role of CNVs in developmental disorders is providing insight into the biological processes disturbed in developmental abnormalities and the effects of copy number variable genes and non-coding elements. As CNV calling methods improve and functional genomics resources become more detailed this will further enhance our ability to identify the underlying causes of human developmental abnormalities.

References

Al-Shahrour, Minguez, Tarraga, Montaner, Alloza, Vaquerizas, Conde, Blaschke, Vera and Dopazo (2006). "BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments." Nucleic Acids Res **34**(Web Server issue): W472-476.

Altug-Teber, Dufke, Poths, Mau-Holzmann, Bastepe, Colleaux, Cormier-Daire, Eggermann, Gillessen-Kaesbach, Bonin and Riess (2005). "A rapid microarray based whole genome analysis for detection of uniparental disomy." Hum Mutat **26**(2): 153-159.

Antequera and Bird (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-11999.

Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, Eppig, Harris, Hill, Issel-Tarver, Kasarskis, Lewis, Matese, Richardson, Ringwald, Rubin and Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Audo, Bujakowska, Leveillard, Mohand-Said, Lancelot, Germain, Antonio, Michiels, Saraiva, Letexier, Sahel, Bhattacharya and Zeitz (2012). "Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases." Orphanet J Rare Dis **7**: 8.

Bademci, Edwards, Torres, Scott, Zuchner, Martin, Vance and Wang (2010). "A rare novel deletion of the tyrosine hydroxylase gene in Parkinson disease." Hum Mutat **31**(10): E1767-1771.

Bauer, Grossmann, Vingron and Robinson (2008). "Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration." Bioinformatics **24**(14): 1650-1651.

Beck, Free, Thorisson and Brookes (2012). "Semantically enabling a genome-wide association study database." J Biomed Semantics **3**(1): 9.

Beissbarth and Speed (2004). "GOstat: find statistically overrepresented Gene Ontologies within a group of genes." Bioinformatics **20**(9): 1464-1465.

Benjamini and Hocjberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. **57**(1): 289-300.

Bister, Set, Cash, Coleman and Fanshawe (2010). "Incidence of facial clefts in Cambridge, United Kingdom." Eur J Orthod.

Blake, Bult, Eppig, Kadin and Richardson (2009). "The Mouse Genome Database genotypes::phenotypes." Nucleic Acids Res **37**(Database issue): D712-719.

Boulding and Webber (2012). "Large-scale objective association of mouse phenotypes with human symptoms through structural variation identified in patients with developmental disorders." Hum Mutat **33**(5): 874-883.

Breen (2010). "Practical informatics approaches to microsatellite and variable number tandem repeat analysis." Methods Mol Biol **628**: 181-194.

Bullard and Nogee (2007). "Heterozygosity for ABCA3 mutations modifies the severity of lung disease associated with a surfactant protein C gene (SFTPC) mutation." Pediatr Res **62**(2): 176-179.

Buxbaum, Daly, Devlin, Lehner, Roeder and State (2012). "The Autism Sequencing Consortium: Large-Scale, High-Throughput Sequencing in Autism Spectrum Disorders." Neuron **76**(6): 1052-1056.

Cabili, Trapnell, Goff, Koziol, Tazon-Vega, Regev and Rinn (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes Dev **25**(18): 1915-1927.

Cassidy, Dykens and Williams (2000). "Prader-Willi and Angelman syndromes: sister imprinted disorders." Am J Med Genet **97**(2): 136-146.

Cassidy, Schwartz, Miller and Driscoll (2011). "Prader-Willi syndrome." Genet Med.

Chelly, Khelfaoui, Francis, Cherif and Bienvenu (2006). "Genetics and pathophysiology of mental retardation." Eur J Hum Genet **14**(6): 701-713.

Chen, Wallis, McLellan, Larson, Kalicki, Pohl, McGrath, Wendl, Zhang, Locke, Shi, Fulton, Ley, Wilson, Ding and Mardis (2009). "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." Nat Methods **6**(9): 677-681.

Church, Goodstadt, Hillier, Zody, Goldstein, She, Bult, Agarwala, Cherry, DiCuccio, Hlavina, Kapustin, Meric, Maglott, Birtle, Marques, Graves, Zhou, Teague, Potamousis, *et al.* (2009). "Lineage-specific biology revealed by a finished genome assembly of the mouse." PLoS Biol **7**(5): e1000112.

Coco and Penchaszadeh (1982). "Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause." Am J Med Genet **12**(2): 155-173.

Colella, Yau, Taylor, Mirza, Butler, Clouston, Bassett, Seller, Holmes and Ragoussis (2007). "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data." Nucleic Acids Res **35**(6): 2013-2025.

Colobran, Pedrosa, Carretero-Iglesia and Juan (2010). "Copy number variation in chemokine superfamily: the complex scene of CCL3L-CCL4L genes in health and disease." Clin Exp Immunol **162**(1): 41-52.

Conrad, Andrews, Carter, Hurles and Pritchard (2006). "A high-resolution survey of deletion polymorphism in the human genome." Nat Genet **38**(1): 75-81.

Consortium (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Croft, O'Kelly, Wu, Haw, Gillespie, Matthews, Caudy, Garapati, Gopinath, Jassal, Jupe, Kalatskaya, Mahajan, May, Ndegwa, Schmidt, Shamovsky, Yung, Birney, Hermjakob, *et al.* (2011). "Reactome: a database of reactions, pathways and biological processes." Nucleic Acids Res **39**(Database issue): D691-697.

Dauber, Yu, Turchin, Chiang, Meng, Demerath, Patel, Rich, Rotter, Schreiner, Wilson, Shen, Wu and Hirschhorn (2011). "Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions." Am J Hum Genet **89**(6): 751-759.

Dawson, Chen, Hunt, Smink, Hunt, Rice, Livingston, Bumpstead, Bruskiewich, Sham, Ganske, Adams, Kawasaki, Shimizu, Minoshima, Roe, Bentley and Dunham (2001). "A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence." Genome Res **11**(1): 170-178.

DeSantis, Houseman, Coull, Louis, Mohapatra and Betensky (2009). "A latent class model with hidden Markov dependence for array CGH data." Biometrics **65**(4): 1296-1305.

Dunham, Kundaje, Aldred, Collins, Davis, Doyle, Epstein, Frietze, Harrow, Kaul, Khatun, Lajoie, Landt, Lee, Pauli, Rosenbloom, Sabo, Safi, Sanyal, Shoresh, *et al.* (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Edwards, Harnden, Cameron, Crosse and Wolff (1960). "A new trisomic syndrome." Lancet **1**(7128): 787-790.

Edwards, Walter, McEwen, Vavouri, Kelly, Abnizova, Woolfe, Goode, Goodson, North, Snell, Callaway, Smith, Gilks, Cooke and Elgar (2006). "Characterisation of conserved non-coding sequences in vertebrate genomes using bioinformatics, statistics and functional studies." Comp Biochem Physiol Part D Genomics Proteomics **1**(1): 46-58.

Eilers and de Menezes (2005). "Quantile smoothing of array CGH data." Bioinformatics **21**(7): 1146-1153.

Eppig JT (2007). "Mouse genome informatics (MGI) resources for pathology and toxicology." Toxicol Pathol. **35**(3): 456-457.

Feenstra, Fang, Koolen, Siezen, Evans, Winter, Lees, Riegel, de Vries, Van Ravenswaaij and Schinzel (2006). "European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities." Eur J Med Genet **49**(4): 279-291.

Firth, Richards, Bevan, Clayton, Corpas, Rajan, Van Vooren, Moreau, Pettett and Carter (2009). "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." Am J Hum Genet **84**(4): 524-533.

Firth, Richards, Bevan, Clayton, Corpas, Rajan, Van Vooren, Moreau, Pettett and Carter (2009). "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." Am J Hum Genet **84**(4): 524-533.

Flicek, Aken, Ballester, Beal, Bragin, Brent, Chen, Clapham, Coates, Fairley, Fitzgerald, Fernandez-Banet, Gordon, Graf, Haider, Hammond, Howe, Jenkinson, Johnson, Kahari, *et al.* (2010). "Ensembl's 10th year." Nucleic Acids Res **38**(Database issue): D557-562.

Fryns and de Ravel (2002). "London Dysmorphology Database, London Neurogenetics Database and Dysmorphology Photo Library on CD-ROM [Version 3] 2001R. M. Winter, M. Baraitser, Oxford University Press, ISBN 019851-780, pound sterling 1595." Hum Genet **111**(1).

Gai, Perin, Murphy, O'Hara, D'Arcy, Wenocur, Xie, Rappaport, Shaikh and White (2010). "CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics." BMC Bioinformatics **11**: 74.

Gazave, Darre, Morcillo-Suarez, Petit-Marty, Carreno, Marigorta, Ryder, Blancher, Rocchi, Bosch, Baker, Marques-Bonet, Eichler and Navarro (2011). "Copy number variation analysis in the great apes reveals species-specific patterns of structural variation." Genome Res **21**(10): 1626-1639.

Girirajan, Rosenfeld, Cooper, Antonacci, Siswara, Itsara, Vives, Walsh, McCarthy, Baker, Mefford, Kidd, Browning, Browning, Dickel, Levy, Ballif, Platky, Farber, Gowans, *et al.* (2010). "A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay." Nat Genet **42**(3): 203-209.

Gonzalez, Kulkarni, Bolivar, Mangano, Sanchez, Catano, Nibbs, Freedman, Quinones, Bamshad, Murthy, Rovin, Bradley, Clark, Anderson, O'Connell R, Agan, Ahuja, Bologna, Sen, *et al.* (2005). "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility." Science **307**(5714): 1434-1440.

Griffiths-Jones (2006). "miRBase: the microRNA sequence database." Methods Mol Biol **342**: 129-138.

Griffiths-Jones, Saini, van Dongen and Enright (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-158.

Grond-Ginsbach, Chen, Pjontek, Wiest, Jiang, Burwinkel, Tchatchou, Krawczak, Schreiber, Brandt, Kloss, Arnold, Hemminki, Lichy, Lyrer, Hausser and Engelter

(2012). "Copy number variation in patients with cervical artery dissection." Eur J Hum Genet.

Harris, Clark, Ireland, Lomax, Ashburner, Foulger, Eilbeck, Lewis, Marshall, Mungall, Richter, Rubin, Blake, Bult, Dolan, Drabkin, Eppig, Hill, Ni, Ringwald, *et al.* (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32**(Database issue): D258-261.

Hehir-Kwa, Wieskamp, Webber, Pfundt, Brunner, Gilissen, de Vries, Ponting and Veltman (2010). "Accurate distinction of pathogenic from benign CNVs in mental retardation." PLoS Comput Biol **6**(4): e1000752.

Hinds, Stuve, Nilsen, Halperin, Eskin, Ballinger, Frazer and Cox (2005). "Whole-genome patterns of common DNA variation in three human populations." Science **307**(5712): 1072-1079.

Holt, Sykes, Conceicao, Cazier, Anney, Oliveira, Gallagher, Vicente, Monaco and Pagnamenta (2012). "CNVs leading to fusion transcripts in individuals with autism spectrum disorder." Eur J Hum Genet.

Hsu, Lin, Wu, Liang, Huang, Chan, Tsai, Chen, Lee, Chiu, Chien, Wu, Huang, Tsou and Huang (2011). "miRTarBase: a database curates experimentally validated microRNA-target interactions." Nucleic Acids Res **39**(Database issue): D163-169.

Huang da, Sherman and Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Iafrate, Feuk, Rivera, Listewnik, Donahoe, Qi, Scherer and Lee (2004). "Detection of large-scale variation in the human genome." Nat Genet **36**(9): 949-951.

Itsara, Wu, Smith, Nickerson, Romieu, London and Eichler (2010). "De novo rates and selection of large copy number variation." Genome Res **20**(11): 1469-1481.

Jacquemont, Reymond, Zufferey, Harewood, Walters, Kutalik, Martinet, Shen, Valsesia, Beckmann, Thorleifsson, Belfiore, Bouquillon, Campion, de Leeuw, de Vries, Esko, Fernandez, Fernandez-Aranda, Fernandez-Real, *et al.* (2011). "Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus." Nature **478**(7367): 97-102.

Jurka, Kapitonov, Pavlicek, Klonowski, Kohany and Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenet Genome Res **110**(1-4): 462-467.

Kanehisa, Araki, Goto, Hattori, Hirakawa, Itoh, Katayama, Kawashima, Okuda, Tokimatsu and Yamanishi (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-484.

Kariminejad, Lind-Thomsen, Tumer, Erdogan, Ropers, Tommerup, Ullmann and Moller (2011). "High frequency of rare copy number variants affecting functionally related genes in patients with structural brain malformations." Hum Mutat **32**(12): 1427-1435.

Ke, Taylor and Cardon (2008). "Singleton SNPs in the human genome and implications for genome-wide association studies." Eur J Hum Genet **16**(4): 506-515.

Kent, Sugnet, Furey, Roskin, Pringle, Zahler and Haussler (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kogan, Blumberg, Schieve, Boyle, Perrin, Ghandour, Singh, Strickland, Trevathan and van Dyck (2009). "Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007." Pediatrics **124**(5): 1395-1403.

Kumari, Antonova and Geyer (2008). "Prepulse inhibition and "psychosis-proneness" in healthy individuals: an fMRI study." Eur Psychiatry **23**(4): 274-280.

Lander, Linton, Birren, Nusbaum, Zody, Baldwin, Devon, Dewar, Doyle, FitzHugh, Funke, Gage, Harris, Heaford, Howland, Kann, Lehoczky, LeVine, McEwan,

McKernan, *et al.* (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lee, Carvalho and Lupski (2007). "A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders." Cell **131**(7): 1235-1247.

Lee and Jan (2012). "Fragile X syndrome: mechanistic insights and therapeutic avenues regarding the role of potassium channels." Curr Opin Neurobiol **22**(5): 887-894.

Lewis, Burge and Bartel (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lipscomb (2000). "Medical Subject Headings (MeSH)." Bull Med Libr Assoc **88**(3): 265-266.

Locke, Sharp, McCarroll, McGrath, Newman, Cheng, Schwartz, Albertson, Pinkel, Altshuler and Eichler (2006). "Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome." Am J Hum Genet **79**(2): 275-290.

Lu, Zhang, Deng, Miao, Guo, Gao and Cui (2008). "An analysis of Human MicroRNA and DIsease Associations " Plos one **3**(10).

Lucito, Healy, Alexander, Reiner, Esposito, Chi, Rodgers, Brady, Sebat, Troge, West, Rostan, Nguyen, Powers, Ye, Olshen, Venkatraman, Norton and Wigler (2003). "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation." Genome Res **13**(10): 2291-2305.

Mallon, Iyer, Melvin, Morgan, Parkinson, Brown, Flicek and Skarnes (2012). "Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans." Mamm Genome **23**(9-10): 641-652.

McCarroll, Hadnott, Perry, Sabeti, Zody, Barrett, Dallaire, Gabriel, Lee, Daly and Altshuler (2006). "Common deletion polymorphisms in the human genome." Nat Genet **38**(1): 86-92.

McKusick (1998). "Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders." Johns Hopkins University Press.

Mefford and Eichler (2009). "Duplication hotspots, rare genomic disorders, and common disease." Curr Opin Genet Dev **19**(3): 196-204.

Mei, Galipeau, Prass, Berno, Ghandour, Patil, Wolff, Chee, Reid and Lockhart (2000). "Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays." Genome Res **10**(8): 1126-1137.

Meyer, Zweig, Hinrichs, Karolchik, Kuhn, Wong, Sloan, Rosenbloom, Roe, Rhead, Raney, Pohl, Malladi, Li, Lee, Learned, Kirkup, Hsu, Heitner, Harte, *et al.* (2012). "The UCSC Genome Browser database: extensions and updates 2013." Nucleic Acids Res.

Miller, Horvath and Geschwind (2010). "Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways." Proceedings of the National Academy of Sciences **107**(28): 12698-12703.

Mills, Luttig, Larkins, Beauchamp, Tsui, Pittard and Devine (2006). "An initial map of insertion and deletion (INDEL) variation in the human genome." Genome Res **16**(9): 1182-1190.

Mills, Pittard, Mullaney, Farooq, Creasy, Mahurkar, Kemeza, Strassler, Ponting, Webber and Devine (2011). "Natural genetic variation caused by small insertions and deletions in the human genome." Genome Res **21**(6): 830-839.

Molinaro (2002). Comparative Genomic Hybridization Analysis. U.C. Berkeley Division of Biostatistics Working Paper Series, U.C Berkley.

Morrow (2010). "Genomic copy number variation in disorders of cognitive development." J Am Acad Child Adolesc Psychiatry **49**(11): 1091-1104.

Mourelatos, Dostie, Paushkin, Sharma, Charroux, Abel, Rappsilber, Mann and Dreyfuss (2002). "miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs." Genes Dev **16**(6): 720-728.

Mullaney, Mills, Pittard and Devine (2010). "Small insertions and deletions (INDELs) in human genomes." Hum Mol Genet **19**(R2): R131-136.

Munoz-Yunta, Palau-Baduell, Salvado-Salvado, Valls-Santasusana, Rosendo-Moreno, Clofent-Torrento and Manchado (2008). "[Autism, epilepsy and genetics]." Rev Neurol **46 Suppl 1**: S71-77.

Nguyen, Webber, Hehir-Kwa, Pfundt, Veltman and Ponting (2008). "Reduced purifying selection prevails over positive selection in human copy number variant evolution." Genome Res **18**(11): 1711-1723.

Nguyen, Webber and Ponting (2006). "Bias of selection on human copy-number variants." PLoS Genet **2**(2): e20.

NHGRI. (2011, 6/10/2011). "NIH to make a mightier mouse resource for understanding disease." from http://www.genome.gov/27545656.

Osborne, Flatow, Holko, Lin, Kibbe, Zhu, Danila, Feng and Chisholm (2009). "Annotating the human genome with Disease Ontology." BMC Genomics **10 Suppl 1**: S6.

Oscarson, McLellan, Gullsten, Agundez, Benitez, Rautio, Raunio, Pelkonen and Ingelman-Sundberg (1999). "Identification and characterisation of novel polymorphisms in the CYP2A locus: implications for nicotine metabolism." FEBS Lett **460**(2): 321-327.

Patau, Smith, Therman, Inhorn and Wagner (1960). "Multiple congenital anomaly caused by an extra autosome." Lancet **1**(7128): 790-793.

Peca and Feng (2012). "Cellular and synaptic network defects in autism." Curr Opin Neurobiol.

Perry, Yang, Marques-Bonet, Murphy, Fitzgerald, Lee, Hyland, Stone, Hurles, Tyler-Smith, Eichler, Carter, Lee and Redon (2008). "Copy number variation and evolution in humans and chimpanzees." Genome Res **18**(11): 1698-1710.

Pinto, Marshall, Feuk and Scherer (2007). "Copy-number variation in control population cohorts." Hum Mol Genet **16 Spec No. 2**: R168-173.

Pinto, Pagnamenta, Klei, Anney, Merico, Regan, Conroy, Magalhaes, Correia, Abrahams, Almeida, Bacchelli, Bader, Bailey, Baird, Battaglia, Berney, Bolshakova, Bolte, Bolton, *et al.* (2010). "Functional impact of global rare copy number variation in autism spectrum disorders." Nature **466**(7304): 368-372.

Pollack, Sorlie, Perou, Rees, Jeffrey, Lonning, Tibshirani, Botstein, Borresen-Dale and Brown (2002). "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors." Proc Natl Acad Sci U S A **99**(20): 12963-12968.

Ponchel, Toomes, Bransfield, Leong, Douglas, Field, Bell, Combaret, Puisieux, Mighell, Robinson, Inglehearn, Isaacs and Markham (2003). "Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions." BMC Biotechnol **3**: 18.

Prensner, Iyer, Balbin, Dhanasekaran, Cao, Brenner, Laxman, Asangani, Grasso, Kominsky, Cao, Jing, Wang, Siddiqui, Wei, Robinson, Iyer, Palanisamy, Maher and Chinnaiyan (2011). "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression." Nat Biotechnol **29**(8): 742-749.

Przeworski, Hudson and Di Rienzo (2000). "Adjusting the focus on human variation." Trends Genet **16**(7): 296-302.

Raghavan, Lillington, Skoulakis, Debernardi, Chaplin, Foot, Lister and Young (2005). "Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias." Cancer Res **65**(2): 375-378.

Raychaudhuri, Korn, McCarroll, Altshuler, Sklar, Purcell and Daly (2010). "Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function." PLoS Genet **6**(9).

Redon, Ishikawa, Fitch, Feuk, Perry, Andrews, Fiegler, Shapero, Carson, Chen, Cho, Dallaire, Freeman, Gonzalez, Gratacos, Huang, Kalaitzopoulos, Komura, MacDonald, Marshall, *et al.* (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-454.

Reich, Schaffner, Daly, McVean, Mullikin, Higgins, Richter, Lander and Altshuler (2002). "Human genome sequence variation and the influence of gene history, mutation and recombination." Nat Genet **32**(1): 135-142.

Rivera-Brugues, Albrecht, Wieczorek, Schmidt, Keller, Gohring, Ekici, Tzschach, Garshasbi, Franke, Klopp, Wichmann, Meitinger, Strom and Hempel (2011). "Cohen syndrome diagnosis using whole genome arrays." J Med Genet **48**(2): 136-140.

Robinson and Mundlos (2010). "The human phenotype ontology." Clin Genet **77**(6): 525-534.

Rodriguez-Caballero, Torres-Lagares, Rodriguez-Perez, Serrera-Figallo, Hernandez-Guisado and Machuca-Portillo (2010). "Cri du chat syndrome: a critical review." Med Oral Patol Oral Cir Bucal **15**(3): e473-478.

Roeleveld, Zielhuis and Gabreels (1997). "The prevalence of mental retardation: a critical review of recent literature." Dev Med Child Neurol **39**(2): 125-132.

Rossin, Lage, Raychaudhuri, Xavier, Tatar, Benita, Cotsapas and Daly (2011). "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology." PLoS Genet **7**(1): e1001273.

Saha, Barnett, Foldi, Burne, Eyles, Buka and McGrath (2009). "Advanced paternal age is associated with impaired neurocognitive outcomes during infancy and childhood." PLoS Med **6**(3): e40.

Sanger, Air, Barrell, Brown, Coulson, Fiddes, Hutchison, Slocombe and Smith (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-695.

Scherer, Lee, Birney, Altshuler, Eichler, Carter, Hurles and Feuk (2007). "Challenges and standards in integrating surveys of structural variation." Nat Genet **39**(7 Suppl): S7-15.

Schinzel (1994). Zurich Cytogenetics Database. O. M. Databases.

Schroeder (2011). GO Term Mapper, Lewis-Sigler Institute for Integrative Genomics.

Sebat, Lakshmi, Troge, Alexander, Young, Lundin, Maner, Massa, Walker, Chi, Navin, Lucito, Healy, Hicks, Ye, Reiner, Gilliam, Trask, Patterson, Zetterberg, *et al.* (2004). "Large-scale copy number polymorphism in the human genome." Science **305**(5683): 525-528.

Serafini, Pompili, Innamorati, Giordano, Montebovi, Sher, Dwivedi and Girardi (2012). "The role of microRNAs in synaptic plasticity, major affective disorders and suicidal behavior." Neurosci Res **73**(3): 179-190.

Shaikh, Gai, Perin, Glessner, Xie, Murphy, O'Hara, Casalunovo, Conlin, D'Arcy, Frackelton, Geiger, Haldeman-Englert, Imielinski, Kim, Medne, Annaiah,

Bradfield, Dabaghyan, Eckert, *et al.* (2009). "High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications." Genome Res **19**(9): 1682-1690.

Shaikh, Haldeman-Englert, Geiger, Ponting and Webber (2011). "Genes and biological processes commonly disrupted in rare and heterogeneous developmental delay syndromes." Hum Mol Genet **20**(5): 880-893.

Shanks, Downes, Copley, Lise, Broxholme, Hudspith, Kwasniewska, Davies, Hankins, Packham, Clouston, Seller, Wilkie, Taylor, Ragoussis and Nemeth (2012). "Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease." Eur J Hum Genet.

Sharp, Locke, McGrath, Cheng, Bailey, Vallente, Pertz, Clark, Schwartz, Segraves, Oseroff, Albertson, Pinkel and Eichler (2005). "Segmental duplications and copy-number variation in the human genome." Am J Hum Genet **77**(1): 78-88.

Singleton (2005). "Altered alpha-synuclein homeostasis causing Parkinson's disease: the potential roles of dardarin." Trends Neurosci **28**(8): 416-421.

Smith and Eppig (2009). "The mammalian phenotype ontology: enabling robust annotation and comparative analysis." Wiley Interdiscip Rev Syst Biol Med **1**(3): 390-399.

Sokal (1995). Biometry : the principles and practive of statistics in biological research. New York, W. H. Freeman and Co.

Stamatoyannopoulos, Snyder, Hardison, Ren, Gingeras, Gilbert, Groudine, Bender, Kaul, Canfield, Giste, Johnson, Zhang, Balasundaram, Byron, Roach, Sabo, Sandstrom, Stehling, Thurman, *et al.* (2012). "An encyclopedia of mouse DNA elements (Mouse ENCODE)." Genome Biol **13**(8): 418.

Stankiewicz, Shaw, Withers, Inoue and Lupski (2004). "Serial segmental duplications during primate evolution result in complex human genome architecture." Genome Res **14**(11): 2209-2220.

Stoletzki and Eyre-Walker (2011). "The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions." Mol Biol Evol **28**(4): 1371-1380.

Storey (2002). "A direct approach to false discovery rates." J.R. Statist. Soc. B **64**(3): 479-498.

Striano, Coppola, Paravidino, Malacarne, Gimelli, Robbiano, Traverso, Pezzella, Belcastro, Bianchi, Elia, Falace, Gazzerro, Ferlazzo, Freri, Galasso, Gobbi, Molinatto, Cavani, Zuffardi, *et al.* (2012). "Clinical significance of rare copy number variations in epilepsy: a case-control survey using microarray-based comparative genomic hybridization." Arch Neurol **69**(3): 322-330.

Su, Wiltshire, Batalov, Lapp, Ching, Block, Zhang, Soden, Hayakawa, Kreiman, Cooke, Walker and Hogenesch (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-6067.

Sun, Southard, Witonsky, Olopade and Di Rienzo (2010). "Allelic imbalance (AI) identifies novel tissue-specific cis-regulatory variation for human UGT2B15." Hum Mutat **31**(1): 99-107.

Tacutu, Budovsky, Yanai and Fraifeld (2011). "Molecular links between cellular senescence, longevity and age-related diseases - a systems biology perspective." Aging (Albany NY) **3**(12): 1178-1191.

Tuzun, Sharp, Bailey, Kaul, Morrison, Pertz, Haugen, Hayden, Albertson, Pinkel, Olson and Eichler (2005). "Fine-scale structural variation of the human genome." Nat Genet **37**(7): 727-732.

Uddin, Sturge, Peddle, O'Rielly and Rahman (2011). "Genome-wide signatures of 'rearrangement hotspots' within segmental duplications in humans." PLoS One **6**(12): e28853.

Uher (2009). "The role of genetic variation in the causation of mental illness: an evolution-informed framework." Mol Psychiatry. **14**(2): 1072-1082.

Van Den Bossche, Johnstone, Strazisar, Pickard, Goossens, Lenaerts, De Zutter, Nordin, Norrback, Mendlewicz, Souery, De Rijk, Sabbe, Adolfsson, Blackwood and Del-Favero (2012). "Rare copy number variants in neuropsychiatric disorders: Specific phenotype or not?" Am J Med Genet B Neuropsychiatr Genet.

Venter, Adams, Myers, Li, Mural, Sutton, Smith, Yandell, Evans, Holt, Gocayne, Amanatides, Ballew, Huson, Wortman, Zhang, Kodira, Zheng, Chen, Skupski, *et al.* (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vermeesch, Balikova, Schrander-Stumpel, Fryns and Devriendt (2011). "The causality of de novo copy number variants is overestimated." Eur J Hum Genet **19**(11): 1112-1113.

Wang, Kim, Pollack, Narasimhan and Tibshirani (2005). "A method for calling gains and losses in array CGH data." Biostatistics **6**(1): 45-58.

Warburton (1991). "De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints." Am J Hum Genet **49**(5): 995-1013.

Webber (2011). "Functional enrichment analysis with structural variants: pitfalls and strategies." Cytogenet Genome Res **135**(3-4): 277-285.

Webber (2011). "Functional Enrichment Analysis with Structural Variants: Pitfalls and Strategies." Cytogenet Genome Res.

Webber, Hehir-Kwa, Nguyen, de Vries, Veltman and Ponting (2009). "Forging links between human mental retardation-associated CNVs and mouse gene knockout models." PLoS Genet **5**(6): e1000531.

Wilson (2000). "Coloboma mouse mutant as an animal model of hyperkinesis and attention deficit hyperactivity disorder." Neurosci Biobehav Rev **24**(1): 51-57.

Wong, deLeeuw, Dosanjh, Kimm, Cheng, Horsman, MacAulay, Ng, Brown, Eichler and Lam (2007). "A comprehensive analysis of common copy-number variations in the human genome." Am J Hum Genet **80**(1): 91-104.

Yalyn, Arman, Erdogan and Kula (2006). "A comparison of the circadian rhythms and the levels of melatonin in patients with diurnal and nocturnal complex partial seizures." Epilepsy Behav **8**(3): 542-546.

Yamagishi (2002). "The 22q11.2 deletion syndrome." Keio J Med **51**(2): 77-88.

Yatsenko, Treadwell-Deering, Krull, Lewis, Glaze, Stankiewicz, Lupski and Potocki (2005). "Trisomy 17p10-p12 due to mosaic supernumerary marker chromosome: delineation of molecular breakpoints and clinical phenotype, and comparison to other proximal 17p segmental duplications." Am J Med Genet A **138A**(2): 175-180.

Yoon, Xuan, Makarov, Ye and Sebat (2009). "Sensitive and accurate detection of copy number variants using read depth of coverage." Genome Res **19**(9): 1586-1592.

Zhang, Feuk, Duggan, Khaja and Scherer (2006). "Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome." Cytogenet Genome Res **115**(3-4): 205-214.

Zhang, Khajavi, Connolly, Towne, Batish and Lupski (2009). "The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans." Nat Genet **41**(7): 849-853.