Persistent, ancient constraints shape copy number & expression variation of dosage-sensitive genes

by

Alan M. Rice

A thesis submitted to
The University of Dublin
for the degree of
Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College
The University of Dublin

September, 2017
Declaration of Authorship

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University’s open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

________________________________________  ____________________________
Signature                                      Date
Acknowledgements

The work contained in this thesis would not have been possible without the help and support of many people.

Firstly, I’d like to acknowledge the support and encouragement of my family. The countless times I’ve been chauffeured, come home to a hot meal, given a spare room to stay in, and let drone on about my work or stresses without complaint, can never be repaid in a lifetime.

Thanks to Katie, Liam, Graeme, and Craig, for all the fun, laughs, and adventures along the way. While research was usually never far from mind, it kept things in perspective. I’m looking forward to many more adventures in the future.

Past and current members of the McLysaght lab, thank you for all the helpful comments, advice, support, and making day-to-day work fly by. After four years, I’m already looking forward to another year of your company. Thanks also to extended members of our department, especially those in our Wednesday lunchtime seminars for helpful comments, and the opportunity to present my sometimes rough and ready work to a kind audience.

Finally, thank you Aoife, for all the opportunities, guidance, mentorship, and freedom. It’s been great and I’ll forever be grateful.
## Contents

1 Introduction ......................................................... 1

1.1 Copy number variation ............................................. 2
  1.1.1 Mechanisms of copy number variant generation .......... 2
  1.1.2 Contribution to disease ................................. 4
  1.1.3 Causes of pathogenicity ............................. 6
  1.1.4 Copy number variant inheritance .................... 15

1.2 Evolution through gene gain and loss ....................... 17
  1.2.1 Gene duplication ....................................... 17
  1.2.2 Whole genome duplication ............................ 19
  1.2.3 Duplicate fixation or loss? .......................... 20
  1.2.4 Biased retention of ohnologues ....................... 23

1.3 Expression evolution ............................................ 25
  1.3.1 Phenotypic evolution by expression change .......... 26
  1.3.2 Expression divergence of gene duplicates ............ 28
  1.3.3 Expression variation ................................. 31
  1.3.4 Role in disease ................................... 31

1.4 Aim .......................................................... 32

2 Materials & Methods ............................................... 33

2.1 Filtering dbVar CNV data .................................... 33

2.2 Counting mammalian gene gain and loss events ............ 34
## CONTENTS

2.2.1 Homologues and trees .................................................. 35  
2.2.2 Analysis workflow ...................................................... 36  
2.2.3 Annotation and error ................................................... 37  
2.2.4 Analysis code ........................................................... 38  

3 Dosage sensitivity is a major determinant of human copy number variant pathogenicity ............................................. 39  
3.1 Introduction ................................................................. 39  
3.2 Materials & Methods ...................................................... 42  
3.2.1 CNV Data with clinical interpretation .............................. 42  
3.2.2 CNV coverage ............................................................ 43  
3.2.3 Gene category enrichment ............................................. 43  
3.2.4 Mammalian copy number analysis ................................ 43  
3.2.5 22q11 conserved synteny analysis ................................ 44  
3.2.6 CNV data for comparative genomics ............................... 45  
3.2.7 Gene ontology (GO) enrichment analysis ........................ 45  
3.3 Results .................................................................. 45  
3.3.1 Identification of pathogenic CNV peak regions ............... 45  
3.3.2 Pathogenic CNVs are enriched for developmental genes .......................... 49  
3.3.3 Class P genes display features of dosage sensitive genes .............. 53  
3.3.4 Solitary Class P genes are enriched for neurodevelopment ...... 57  
3.3.5 Class P genes have high evolutionary copy number constraint 66  
3.3.6 Evolutionary copy number conservation in pathogenic CNVRs 67  
3.3.7 Conserved genes reveal ancient and persistent constraints ...... 70  
3.4 Discussion ................................................................. 74  

4 Prediction of pathogenic copy number variation yields insights into variant inheritance .............................................. 77
viii CONTENTS

5.3 Results ....................................................... 104
  5.3.1 eQTL enrichment of CNVR genes and dosage sensitive genes 104
  5.3.2 Ohnologues have more similar eQTL patterns in brain tissues 107
  5.3.3 eQTLs affecting dosage-sensitive genes have been shaped
    by selection ............................................ 109
  5.3.4 Variance and size of eQTL effects is reduced for ohnologues 112
  5.3.5 Ancient constraint rather than gene age predicts eQTL pat-
    terns .................................................... 113
  5.3.6 Dosage-insensitive gene eQTLs patterns are constrained
    when eQTLs are shared ............................... 113
  5.3.7 Ohnologue pairs demonstrate dosage-balance with mutually
    absent eQTLs ........................................... 116

5.4 Discussion ............................................... 116

6 Conclusions ................................................ 121

Bibliography .................................................. 128
List of Tables

2.1 dbVar studies included in CNV analysis . . . . . . . . . . . . . . 34

3.1 Summary of human CNVs used in CNV analysis . . . . . . . . . 47

3.2 Genes included in different types of CNV have different genetic
and functional characteristics . . . . . . . . . . . . . . . . . . . . 56

3.3 199 Solitary Class P genes . . . . . . . . . . . . . . . . . . . . . 57

3.4 Most enriched functional classes of solitary Class P genes . . . 65

3.5 Most enriched functional classes of genes with conserved copy number 71

4.1 Classifier features . . . . . . . . . . . . . . . . . . . . . . . . . . 84

4.2 Pathogenicity classification of CNVs by inheritance types in DECI-
PHER database . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94

5.1 Genes within CNVRs and eQTL patterns . . . . . . . . . . . . . . 104

5.2 Haploinsufficient genes and eQTL direction . . . . . . . . . . . . 107
List of Figures

1.1 Mechanism of pathogenicity - Haploinsufficiency . . . . . . . . . . 7
1.2 Mechanism of pathogenicity - Protein aggregation . . . . . . . . . 8
1.3 Mechanism of pathogenicity - Dosage balance . . . . . . . . . . . 10
1.4 Mechanism of pathogenicity - Concentration-dependent morphogen 12
1.5 Mechanism of pathogenicity - Topologically associated domains . 14

3.1 Number of genes per copy number variant (CNV) in dataset . . . 47
3.2 Number of genes per CNV region (CNVR) . . . . . . . . . . . . . 48
3.3 Percentage of CNVs containing at least one developmental gene for
   1000 randomised sets . . . . . . . . . . . . . . . . . . . . . . . . . . 50
3.4 Percentage of developmental genes overlapped by each randomised
   CNV set . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
3.5 Illustration of CNVRs and intersection with genes . . . . . . . . 53
3.6 Patterns of gene duplication and loss across mammals for ortho-
   logues of human genes in CNVs . . . . . . . . . . . . . . . . . . . . 54
3.7 Number of Class P genes per copy number gain (CNG) peak region 63
3.8 Number of Class P genes per copy number loss (CNL) peak region 63
3.9 Mammalian copy number changes for genes within known pathogenic
   CNV regions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69
3.10 Distance to nearest developmental gene for developmental genes
    grouped by number of genomes where orthologue has unchanged
    copy number . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
LIST OF FIGURES

4.1 Out-of-bag error with increasing number of trees in random forest 86
4.2 Classifier performance on test dataset 87
4.3 Feature importance and correlation in CNV classifier 89
4.4 Classified control and case CNVs 92
4.5 Classified DECIPHER case CNVs grouped by inheritance type 95
5.1 eQTL enrichment of CNVR genes and dosage sensitive genes 105
5.2 eQTL tissue specificity of dosage-sensitive genes 110
5.3 Shared constraints shaping eQTL patterns 114
5.4 Selection patterns for eQTL influencing expression of dosage-sensitive genes 117
Abbreviations

bp  base pair
CNG  copy number gain
CNL  copy number loss
CNVRs  copy number variant regions
CNVs  copy number variants
DDC  duplication-degeneration-complementation
EAC  escape from adaptive conflict
eQTLs  expression quantitative trait loci
IAD  innovation-amplification-divergence
Mb  megabase
mRNA  messenger RNA
Mya  Million years ago
SNVs  single nucleotide variants
SSD  small-scale duplication
SSDs  small-scale duplicates
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TADs</td>
<td>topologically associated domains</td>
</tr>
<tr>
<td>WGD</td>
<td>whole genome duplication</td>
</tr>
</tbody>
</table>
“I could tell you my adventures—beginning from this morning,” said Alice a little timidly; “but it’s no use going back to yesterday, because I was a different person then.”

— LEWIS CARROLL,

Alice in Wonderland
Summary

For genomes, phenotypes, and organisms to evolve, a necessary prerequisite is genomic variation. The work described in this thesis examines the effect of ancient and persistent gene product constraint on permissible copy number and expression variation affecting dosage-sensitive genes.

Copy number variants (CNVs) are regions of the genome that are duplicated or deleted in some individuals in a population. Often this variation does not produce a phenotype as CNVs, however, CNVs have previously been associated with human conditions, most notably neurodevelopmental disorders, making its study and understanding important. With the prevailing hypothesis that CNV pathogenicity is due to dosage-sensitivity of the enclosed genes, the evolutionary history of genes in benign and pathogenic CNVs were contrasted with the aim of creating a deeper understanding of how evolutionary patterns relate to CNV pathogenicity. We found through comparative genomic methods that mammalian orthologues of human genes found in pathogenic CNV regions are rarely duplicated or lost. Conversely, genes overlapped by benign CNVs displayed much higher variability in mammalian copy number. Furthermore, we found that genes with conserved copy number across mammals are depleted among CNVs in non-human healthy mammals, mirroring the pattern observed in humans.

In chapter 4, we examined the relationship between CNV pathogenicity and acquisition route of variants. Firstly, we found that our trained CNV pathogenic-
ity classifier could accurately predict pathogenic variants using simple genomic characteristics. Applying our prediction tool to independent datasets of control variants from healthy individuals and case variants from patients with rare disorders, we found a large difference in the proportion of variants being classified as pathogenic, enabling us to accurately distinguish between the groups. Finally, we observed significantly different levels of pathogenicity between *de novo* variants and inherited variants. We found evidence, consistent with previous findings, that *de novo* CNVs are more likely to be pathogenic than inherited variants.

Expression quantitative trait loci (eQTLs) are genomic regions harbouring sequence variants that influence the expression level of one or more neighbouring genes. Exploring expression variation, in the form of eQTLs, we tested the hypothesis that dosage-sensitive genes are refractory to such variants in a similar way to their depletion among benign CNVs. To test this we used ohnologues, genes with conserved copy number across mammals from chapter 3, and haploinsufficient genes as groups of genes displaying dosage-sensitivity. Contrary to our expectation, we found that dosage-sensitive genes display enrichment for eQTLs, however they show non-random biases. Notably, eQTLs affecting dosage-sensitive genes are biased towards influencing expression in fewer tissues, whereby eQTL affecting expression in many tissues are likely removed by purifying selection.

These results provide evidence of persistent, ancient dosage constraints in place for many human protein-coding genes. These constraints are a major force in shaping and defining permissible copy number and expression variation for these genes.
Chapter 1

Introduction

Variety’s the very spice of life,
That gives it all its flavour.
— WILLIAM COWPER,
‘The Task’ (1785)

For genomes, phenotypes, and organisms to evolve, a necessary prerequisite is genomic variation. Kan and colleagues began to reveal the extensive variation that pervades our genomes, discovering individuals with multiple copies of the α-globin genes (Kan et al., 1975; Goossens et al., 1980). Around the same time, they also discovered the first single nucleotide variants (SNVs) in the HpaI restriction site downstream of the gene producing β-globin (Kan and Dozy, 1978). With variant presence differing between African and non-African populations and an association with sickle cell disease, the importance of this discovery to medical and population genetics cannot be understated. In the years following, the study of SNVs grew rapidly but research of structural variation involving inversions, deletions, duplications, insertions and translocations, lagged behind until the turn of the 21st century.
1.1 Copy number variation

Copy number variants (CNVs), a form of structural variation, are regions of the genome that are duplicated or deleted in individuals of a population and account for an order of magnitude more variation than SNVs in terms of base pair length. CNVs have been widely characterised in many species (Völker et al., 2010; Liu et al., 2010; Nicholas et al., 2009; Li et al., 2012; Pezer et al., 2015; Debolt, 2010; Bai et al., 2016), but most extensively in humans, where they are found to be abundant. In 2010, it was estimated that each individual has on average 1,000 CNVs of 450 bp with respect to the human reference genome (Conrad et al., 2010). Several estimates suggest that $\sim$10% of the genome experiences recurrent CNV events (Conrad et al., 2010; Zarrei et al., 2015). These repeatedly affected regions are called CNV hotspots and their architecture predisposes them to recurrent CNVs (Lupski, 1998).

1.1.1 Mechanisms of copy number variant generation

Genomic regions that are highly similar in sequence similarity can induce non-allelic homologous recombination between each other (Pâques and Haber, 1999). Non-allelic homologous recombination occurs during meiosis when regions with high similarity align with each other and cross-over occurs. This results in the generation of two new alleles: one allele missing the sequence flanked by the aligned repeats, and another allele with the intermediate sequence duplicated. With about half of the human genome comprising repetitive elements (Treangen and Salzberg, 2012), there is ample opportunity for regions to misalign and give rise to a CNV.

Segmental duplications are one such type of genomic repeat and are relatively young, being younger than 35 million years old (Eichler, 2001). With $\sim$95%
sequence identity, these repetitive DNA regions longer than 5 kilobases are distributed non-randomly throughout the genome and have genic biases (Bailey et al., 2002). Duplications often are located within pericentromeric and subtelomeric regions (Eichler, 2001) but this varies by chromosome (Bailey et al., 2002). Additionally, gene-rich chromosomes are enriched for segmental duplications and conversely gene-poor chromosomes display fewer duplications. This distribution has caused certain regions of our genome to experience accelerated rates of evolution that played an important role during primate evolution (King and Wilson, 1975; Wilson et al., 1975; Samonte and Eichler, 2002; Bailey and Eichler, 2006; Marques-Bonet et al., 2009).

Patterns of genic biases within these regions suggest that the function of the enclosed genes may be a factor in the copy number evolution of these regions. Some genes are observed to have undergone recent rapid adaptation in the primate lineage (Johnson et al., 2001), while other genes with disadvantageous loss-of-function phenotypes appear to be depleted within duplication hotspots (Dang et al., 2008). The *morpheus* gene family within a segmental duplication, LCR16a, has experienced recent duplication and positive selection for many amino acid substitutions (Johnson et al., 2001). This two step process of gene family expansion through duplication and subsequent rapid sequence evolution may be an important reservoir for the emergence of hominoid genes. Conversely, after extensive manual curation of genes with haploinsufficient phenotypes, Dang et al. (2008) demonstrate that these genes are less likely to be situated between pairs of neighbouring segmental duplications on the same chromosome. This depletion suggests that selection has acted on the location of these genes or repeats in order to avoid the deleterious scenario arising of deletion alleles being created, reducing the gene product dosage of these haploinsufficient genes. Creating both positive and purifying selective pressure on the contained genes, CNV arising from non-
allelic homologous recombination occurring between segmental duplications has shaped the genes within these regions.

Besides these recurrent CNVs, there are additional mechanisms of CNV generation that give rise to non-recurrent CNVs. Break-induced replication that occurs at short sites of homology (microhomologies of 2-15 bp) can be error prone, joining a pair of proximal homologous DNA single strands (Bauters et al., 2008; Hastings, Ira, et al., 2009, reviewed in Hastings, Lupski, et al., 2009). This erroneous product is similar to the results of non-allelic homologous recombination with the potential to create deletions and duplications. Additionally, non-homologous end joining can generate CNVs as double-strand breaks in DNA are repaired by annealing microhomologies exposed in single-stranded overhangs (Lee et al., 2007; Lieber, 2008). This mechanism does not require pre-existing homology in the region affected, only in the exposed overhangs, hence the term non-homologous end joining.

These non-recurrent CNVs have the ability to alter the copy number of genes in any region of the genome, not just in regions flanked by repeats. Millions of years of CNV occurrence between segmental duplications have shaped the genes within these regions, but only weakly as the occurrence of CNVs within segmental duplications still have the capacity to cause disease and disorders (Emanuel and Shaikh, 2001; Shaw and Lupski, 2004; Sharp et al., 2006; Miller et al., 2009; Girirajan et al., 2013).

1.1.2 Contribution to disease

Often the occurrence of a CNV does not produce a phenotype, as CNVs are frequently small, intergenic or encompass genes that can tolerate a change in copy number. Some genes can even be completely deleted with no apparent effect (Zarrei et al., 2015). However, CNVs have previously been associated with
a number of human conditions, most notably neurodevelopmental disorders including autism spectrum disorders, schizophrenia, intellectual disability, attention deficit hyperactivity disorder, developmental delay and epilepsy (Sebat et al., 2007; Stefansson et al., 2014; Stefansson et al., 2008; Walsh et al., 2008; Mefford et al., 2008; Lesch et al., 2011; Helbig et al., 2009; Cooper et al., 2011).

Other neurological conditions have been associated with CNVs, including bipolar disorder (Green et al., 2016), Alzheimer’s disease (Rovelet-Lecrux et al., 2006; Heinzen et al., 2010; Brouwers et al., 2012), amyotrophic lateral sclerosis (Morello et al., 2017), and Parkinson’s disease (Miller et al., 2004). There is a growing number of studies implicating the role of CNVs on body weight, particularly those leading to obesity (Bochukova et al., 2010; Jarick et al., 2011; Falchi et al., 2014). In Jacquemont et al. (2011), one region on chromosome 16 is observed to give rise to extreme over- or underweight phenotypes, depending on whether the region is deleted or duplicated, respectively. Additionally, CNVs have been associated with metabolic conditions such as diabetes (Jeon et al., 2010), cardiovascular traits such as familial hypercholesterolemia (Wang et al., 2005), and autoimmune disorders including Crohn’s disease (Fellermann et al., 2006) and rheumatoid arthritis (McKinney et al., 2008). Due to these implications in disease, CNVs are subject to increasingly intense scrutiny to understand and characterise their genetic and phenotypic effects.

Determining the effect that a CNV will have at the phenotypic level is challenging. An understanding of the function of the genes affected by the CNV is required, how those genes interact with other genomic components, and how a change in copy number could perturb the function and interactions of the genes affected. These obstacles are not simple to overcome as identical CNVs can give rise to a range of phenotypes in different individuals due to different genetic backgrounds and environmental factors. For example, presence and severity of
Introduction

Phenotypes can vary for individuals with Down’s syndrome, caused by trisomy of chromosome 21 (Roper and Reeves, 2006), and individuals with CNVs in the q11 region of chromosome 22 (Hiroi et al., 2013). Therefore, there is a need to understand the genes affected by CNVs and how a copy number change can be the basis for pathogenesis.

1.1.3 Causes of pathogenicity

One of the first well-characterised cases of CNV pathogenicity was Charcot-Marie-Tooth neuropathy. Charcot-Marie-Tooth neuropathy is one of the most commonly inherited neurological disorders, affecting ~1 in every 2,500 people. The disorder has been specifically linked to changes in copy number of the dosage-sensitive gene peripheral myelin 22 (PMP22, Lupski et al., 1991). Dosage-sensitivity provides a model whereby a 50% increase or decrease in gene copy number is deleterious (Veitia, 2002; Papp et al., 2003a; Birchler et al., 2001; Birchler and Veitia, 2012). When the dosage of these genes is changed by an overlapping CNV, the function of the gene is disrupted in a way that we may observe as disease. More acutely dosage-sensitive genes may never be observed in CNVs, even in pathogenic ones, if they are so disruptive as to result in inviability (Chen et al., 2017). Thus duplication and/or loss CNVs of dosage-sensitive genes are not expected to be observed in healthy individuals (Makino and McLysaght, 2010).

Haploinsufficiency

Genes can be dosage-sensitive in a number of ways, for example some genes, termed haploinsufficient, have a minimum required concentration to achieve functionality (Figure 1.1). Haploinsufficient genes include many transcription factors and developmental genes (Fisher and Scambler, 1994). A CNV deletion will half the copy number of the genes contained within its breakpoints. This will typically
Figure 1.1 | Mechanism of pathogenicity - Haploinsufficiency
A loss-of-function mutation reduces gene dosage in half resulting in insufficient gene product concentration to effectively carry out function.
result in a reduction of messenger RNA (mRNA) expression levels produced from these genes (Henrichsen et al., 2009) and likely a reduction in final gene product concentration. If functionality cannot be achieved with this decreased concentration and absence of function decreases fitness, the copy number change will be disadvantageous. As already mentioned, haploinsufficient genes are depleted within segmental duplications (Dang et al., 2008), and it is likely the selective pressure of CNV deletion alleles being created at these loci by non-allelic homologous recombination that lead to this non-random genic bias.

Aggregation, toxicity, & overexpression phenotypes

Dosage-sensitivity can also manifest for gene products at high cellular concentrations (Figure 1.2). One such example gene is SNCA which encodes the α-synuclein protein. This protein is aggregation-prone at high concentrations and can result in rapid onset Parkinson’s disease (Singleton et al., 2003; Miller et al., 2004). Aggregation-prone proteins have been linked to other neurodegenerative diseases such as Alzheimer’s and Huntington’s disease but also systemic amyloidosis and localised diseases e.g. type II diabetes and cataracts (Chiti and Dobson, 2006).
Uncontrolled increases in the concentration of aggregation-prone proteins would likely lead to more interactions and more opportunity for potentially deleterious aggregates to form. This could explain why aggregation-prone proteins are seen to be more tightly regulated compared to non-aggregation-prone proteins, showing lower transcription rate and abundance, higher translation regulation and lower translation efficiency and also both lower protein abundance and shorter protein half-life (Gsponer and Babu, 2012). All these regulatory differences between aggregation-prone and non-aggregation-prone proteins suggests that the reduction in availability of monomers for these potentially deleterious aggregates has likely been selected for over evolutionary time.

More generally, genome-wide screens have been conducted systematically in *Saccharomyces cerevisiae* to investigate genes with overexpression phenotypes, that is where overexpression reduces fitness and growth (Gelperin et al., 2005; Sopko et al., 2006). In yeast, \( \sim 80\% \) of genes tested can be overexpressed in this way without significantly reducing growth rate. The remaining 20% are dosage-sensitive in part due to producing proteins that localise in membranes and that contain regions of structural disorder (Österberg et al., 2006; Vavouri et al., 2009). Intrinsic disorder in protein structure promotes off-target promiscuous interactions that can be deleterious and is an additional potential cause of pathogenicity.

**Dosage balance**

In some instances, dosage-sensitive genes may not give rise to deleterious consequences exclusively from under- or over-abundance but from any stoichiometric change relative to other cellular components (Figure 1.3). This is observed for protein complex members that are in stoichiometric balance with their respective complex partners (Papp et al., 2003a). Here as part of the dosage balance hypothesis (Veitia, 2002), it is hypothesised that genes that encode heteromer complex
Figure 1.3 | Mechanism of pathogenicity - Dosage balance
For heteromeric protein complexes, stoichiometric balance is important for correct complex assembly. In cases where imbalance occurs, such as the duplication of a bridging subunit, here subunit B, incomplete complex formation can occur as the overabundant subunit sequesters other complex subunits with reduced chance of forming a complete complex.
subunits must be balanced in concentration to their respective partner(s). If a trimeric complex subunit is produced in excess, a reduction in fully assembled complexes may result due to the other subunits being bound and sequestered by the abundant subunit. As many incomplete dimeric complexes result there are two reasons why this can be disadvantageous: firstly, these partial products will at best be wasteful or worse actively deleterious (e.g. toxic), and secondly, there are fewer than required complete functional complexes to carry out their cellular function(s).

A systematic screen in yeast has not only confirmed deleterious consequences of overexpressing some genes producing complex subunits, but also has demonstrated that simultaneously overexpressing partner subunits negates the reduction in fitness and rescues the phenotype (Makanae et al., 2013). This screen found that only \(\sim2\%\) (115/5,806 genes) of yeast protein-coding genes were sensitive to overexpression. In more recent work this group have attributed such a small proportion of the genome displaying sensitivity to dosage compensatory buffering at the protein-level (Ishikawa et al., 2017). They found that \(\sim10\%\) of genes tested are dosage-compensated and that these genes are enriched for subunits of multi-protein complexes. This suggests that perturbations to protein complex stoichiometry is sufficiently deleterious to warrant widespread compensation. However, it is important to note that post-translational compensation is not universal for dosage-sensitive genes as some complex members still demonstrate sensitivity to overexpression. This perhaps hints that a balance must exist between cellular robustness to variation and the ability to evolve. If a phenotype is too robust this could limit the effect of selective forces on underlying genetic variation (Kitano, 2004).
CONCENTRATION-DEPENDENT MORPHOGENS

Wild-type

Protein distribution

Overexpression

Protein distribution

Underexpression

Protein distribution

Figure 1.4 | Mechanism of pathogenicity - Concentration-dependent morphogen
Concentration-dependent morphogens pattern embryos in a highly specific manner during development to create differential gene expression at opposite embryonic poles. If increased gene dosage occurs, constitutive receptor activation can result and no transcriptional differences result at opposite poles. Conversely, underexpression will fail to active any receptors and body plan development will fail to proceed.
**Specific absolute dosages**

Dosage-sensitive genes may operate in a concentration-dependent fashion (Figure 1.4). An example of how some developmental genes are affected by a concentration change can be seen during embryonic development where morphogens establish concentration gradients. These gradients pattern the embryo in a highly specific manner to regulate transcription (Rogers and Schier, 2011). In some cases, receptors for these transcription factors have low binding affinity, requiring high concentrations of the morphogen to activate transcription of downstream elements. Alternatively, a scenario where a high number of binding sites are present and a significant proportion must be bound to activate transcription also requires high concentration. By contrast, other receptors could have high affinity and so responding to even very low levels of transcription factor. Similarly low levels of transcription factor may suffice if the number of binding sites is low and transcription can readily be activated. Concentration changes on such a precise system of regulation can greatly alter the response seen, where increased concentration would cause low binding affinity receptors to exhibit a constitutively active-like state. Conversely, a reduction in morphogen could be insufficient to activate some types of receptors at all.

Specific concentrations of gene product can also influence splicing of mRNAs as seen with pyruvate kinase M \( (PKM, \text{Chen et al., 2012}) \). PKM has two splicing forms, an adult form and an embryonic form. Which form is produced at a given time is regulated by the concentration of hnRNPA1, a heterogeneous nuclear ribonucleoprotein that functions as a splicing repressor. High hnRNPA1 concentration yields the embryonic form of PKM and low concentration results in adult PKM. In cancer cells however, hnRNPA1 is upregulated by MYC and erroneously gives rise to splicing of the embryonic form in adult tissue. Therefore, it is important to note that CNVs can not just give rise to more or less of a
Introduction

TOPOLOGICALLY ASSOCIATED DOMAINS

Wild-type

Duplication of TAD boundary

TAD interaction

loops

Figure 1.5 | Mechanism of pathogenicity - Topologically associated domains (TADs)

Duplication or deletion of TAD boundaries can result in different interactions between genomic regions. Shown here is a duplication of a TAD boundary that disrupts one TAD, TAD b, and leads to the creation of a second TAD, neo TAD. This altered genomic context means that genes are exposed to alternative regulatory elements and changes to gene expression occur.

gene product and the downstream regulatory effects that can have, but also the functional effects of alternative splice forms that can arise.

Chromosomal structure & topologically associated domains

While the prevailing hypothesis on CNV pathogenicity is that it is due to dosage sensitivity of the included genes, there is more than one possible mechanism by which a CNV can disrupt gene function and cause a phenotype, including disruption of chromosome structure and uncoupling of regulatory elements from their downstream genes (Zhang, Gu, et al., 2009). There is evidence that a CNV at a given locus can affect more than just the expression of its contained genes by also influence neighbouring genes’ expression level as well (Reymond et al., 2007). In fact this extended influence can reach as far as 2–7 Mb beyond a CNV’s boundaries, making it more difficult to elucidate causative elements of pathogenicity.

CNVs can disrupt the three-dimensional organisation of the genome in the
nucleus (Figure 1.5, Franke et al., 2016). Topologically associated domains (TADs) are megabase size regions of the genome defined as physically interacting and proximal in nuclear three-dimensional space (Dixon et al., 2012; Nora et al., 2012; Stevens et al., 2017). They operate as regulatory units where enhancers and promoters interact and their genomic boundaries appear to be conserved across species and cell types (Lupiáñez et al., 2015). Duplications within a TAD, intra-TAD duplications, appear to leave TAD structure unaffected (Franke et al., 2016). However, inter-TAD duplications that enclose a boundary and extend into the neighbouring domain create new domains (neo-TADs) that isolate that region from the rest of the genome. If only non-coding, non-functional regions are incorporated into this new domain this can result in a neutral phenotype, but if any genes or regulatory elements are included they will experience a different genomic context potentially giving rise to a deleterious phenotype.

From these examples it is clear that CNVs affecting dosage-sensitive genes can give rise to pathogenicity and how that might occur. As much of the genome remains incompletely characterised functionally, it is not easy to say with a great degree of certainty what the phenotypic affect of a CNV at a given genomic locus will be.

1.1.4 Copy number variant inheritance

When a CNV in an individual’s genome originates in one of their progenitor cells as a germline mutation, its origin is said to be de novo. About 50% of this individual’s own germline cells will also have the allele with the CNV and the potential to be inherited by his/her offspring. This means that for a CNV to be inherited from a parent already affected by it in their own genome, its resulting phenotype cannot be sufficiently deleterious to completely reduce fecundity, the ability to produce offspring. Therefore, CNVs that greatly impact fecundity and
fitness can be observed as part of de novo variation but these mutations will be absent among inherited variants. CNVs that are incompatible with life will still be unobserved in both de novo and inherited variation in individuals.

These extreme selective pressures acting on segregating CNVs should have an impact on which CNVs are permissible and differentiate them from de novo variants. Indeed, distinct patterns have been observed. Recurrent de novo CNVs are observed as occurring in a similar distribution to segregating CNVs in the genome of Plasmodium falciparum, being predominantly located in the telomeric or subtelomeric regions (Samarakoon et al., 2011). However, non-recurrent, rare de novo variants are randomly distributed throughout the genome, arising with about equal frequency across chromosomal regions. This suggests that selection has removed some variation outside of telomeric or subtelomeric regions that arose but never achieved segregation within the population.

The type of variants that are acquired by different routes, inherited or de novo, appear to play different roles in human conditions also. It has been observed that the genomes of individuals with autism spectrum disorder have a four-fold enrichment for de novo CNVs compared to their unaffected siblings (Sebat et al., 2007; Itsara et al., 2010; Levy et al., 2011). Also, patients with intellectual disability and/or multiple congenital anomalies categorised as having severe phenotypes, including organ malformations, were enriched for more de novo CNVs (Vulto-van Silfhout et al., 2013). In contrast, patients with more moderate phenotypes had a mix of inherited and de novo CNVs affecting their genomes. Additionally, de novo CNVs are observed to contribute to schizophrenia and have 8-fold enrichment in sporadic, non-familial cases compared to controls (Xu et al., 2008). Rare, inherited CNVs were only moderately enriched in sporadic schizophrenia cases. It is apparent from this evidence that de novo CNVs can be highly pathogenic and some of which likely affect fecundity which leads to a reduction in their frequency segregating
in the population.

The relationship between pathogenicity and CNV acquisition route is investigated in chapter 4 – ‘Prediction of pathogenic copy number variation yields insights into variant inheritance’.

1.2 Evolution through gene gain and loss

When CNVs arise they have two ultimate fates: loss or fixation. If a CNV negatively affects fecundity it will have reduced chances of being inherited and likely removed from the population. An advantageous CNV that increases an organism’s fitness, is more likely to be inherited and eventually reach high frequency in the population becoming fixed in all individuals of a subsequent generation. How quickly a variant reaches one of these fates is determined by a number of factors, including population size and selective pressure (Whitlock, 2003). It has been about 15 million years since humans and great apes shared a common ancestor (Moorjani et al., 2016), and this has been enough time for some CNVs to reach fixation, with a number of studies observing gene copy number differences between the genomes of human and other primates (Locke et al., 2003; Fortna et al., 2004; Newman et al., 2005; Goidts et al., 2006; Wilson et al., 2006). In fact, these fixed copy number differences appear to be extensive and common, encompassing more than 20 Mb and two hundred genes. Which variants reach fixation and which regions of the genome are affected are important questions to investigate both for understanding genome evolution and human conditions.

1.2.1 Gene duplication

In 1970, Susumu Ohno championed the importance of evolution by gene duplication, whereby new gene function originates after a gene duplication event (Ohno,
1970; Ohno, 1973). After duplication, two identical gene copies, called paralogues, exist and this genetic redundancy releases the paralogues from functional constraints (Lynch and Conery, 2000). Now, because a “backup” gene copy is present, one of the paralogues is free to acquire a mutation that would render it unable to carry out its ancestral function without deleterious effect. This paralogue can continue to accrue mutations that will either lead to its eventual loss or acquire a different, new function altogether. The “backup” gene copy will continue to carry out ancestral function but has functional constraints reimposed on it. Which of the two paralogues becomes evolutionarily restricted again is determined by which copy first acquires a mutation that prevents it carrying out ancestral function.

Gene duplications provide a basis for genomic innovation. Protein-coding genes are complex, comprised of an open reading frame, typically spread across multiple exons divided by introns, upstream and downstream regulatory elements, protein domains, etc. (He and Zhang, 2005). Gene duplication and mutation provide a mechanism to attain complex gene function without having to repeatedly evolve such complexity. Lynch and Conery (2000) provide a conservative average estimate of 0.01 gene duplicates per gene per million years, with a species range of $\sim 0.002 - 0.02$. This means that in a 35-350 million year time period, $\sim 50\%$ of protein-coding genes are expected to undergo a duplication event and reach high frequency in a population. Such abundance of raw material provides an excellent basis for genomic innovation and adaptation.

An issue arises, however, when a gene being duplicated is dosage-sensitive, specifically if it is sensitive to increased expression. Here, only one or several genes are being duplicated and this process is called small-scale duplication (SSD). In this case, a gene duplication will be deleterious and selected against. We would expect dosage-sensitive genes to have less fixed duplications over evolutionary timescales due to selection arising from these deleterious effects. This can be
tested by looking for signatures of natural selection. This hypothesis is tested in chapter 3 – ‘Dosage sensitivity is a major determinant of human copy number variant pathogenicity’.

1.2.2 Whole genome duplication

Ohno (1970) also proposed that vertebrates had experienced two whole genome duplication (WGD) events. As opposed to SSD, WGD is the complete doubling of genomic content resulting in cells that have four sets of chromosomes and are in a state of tetraploidy. Polyploidy may arise due to abnormal cell division during meiosis. As chromosomes accrue mutations, they gradually return to diploidy as they fail to retain enough sequence similarity to line up at metaphase during meiosis (Wolfe, 2001), i.e. instead of four alleles aligning at a given locus, the genome returns to two aligned alleles. How long the process of reploidisation takes likely varies between genomes but recent evidence from the salmonid-specific autotetraploidisation gives us some indication as rediploidisation is still ongoing after 80 million years (Lien et al., 2016).

The vertebrate WGD events are often called the 2R hypothesis, a name derived from two rounds of duplication, and have been extensively studied in the advent of sequence data (Holland et al., 1994; Ohno, 1999; McLysaght et al., 2002; Hokamp et al., 2003; Dehal and Boore, 2005). These events are estimated to have occurred at least 450 million years ago (Mya) and are seen as fundamental to the development and innovation of vertebrate evolution that has since taken place by providing extensive raw material upon which to select (Freeling and Thomas, 2006). Extensive genomic rearrangement and massive gene loss took place in the aftermath of the WGD events. Prior to the salmonid-specific duplication event already mentioned, a teleost-specific WGD occurred also (Jaillon et al., 2004; Meyer and Van De Peer, 2005). In fact, WGD events while rarely observed fixed
have occurred a number of times across phyla, particularly in the plant kingdom (Jiao et al., 2011; Chalhoub et al., 2014). While the majority of duplicated genes post-WGD are observed to return to single copy, this is a non-random process.

1.2.3 Duplicate fixation or loss?

Gain of a new function, termed neofunctionalisation (Force et al., 1999), is only one of a number ways gene duplicates can be selectively retained in the genome and persist (Conant and Wolfe, 2008; Innan and Kondrashov, 2010). An issue with this model of duplicate retention through gain of function proposed by Ohno (1970) is that the paralogue not carrying out the ancestral gene function has a high probability of losing all functionality due to random mutation prior to achieving a new function that ensures its retention (Bergthorsson et al., 2007). This means that it will likely become a non-functional pseudogene before the process of “mutation during non-functionality” will have time to give rise to a new function. Bergthorsson et al. reconcile Ohno’s hypothesis with the lack of a non-functionality phases through a model called innovation-amplification-divergence (IAD).

In the IAD model some pre-existing innovation is present. An example of this would be an enzyme that typically catalyses one reaction but occasionally partakes in another (Gancedo and Flores, 2008; Gancedo et al., 2016). The enzyme is optimised for its primary substrate but perhaps under certain conditions, e.g. high concentration of secondary substrate, is coerced to catalyse a reaction upon this secondary substrate. Upon gene duplication, amplification can take place in one gene copy, that is mutations arise that favour the promiscuous activity on the secondary substrate. This optimises the reaction and it takes place with higher frequency, likely at the cost of reduced affinity for the primary substrate.
Finally, divergence occurs where mutations continue to take place, the reaction on the secondary substrate becomes the main function for one gene copy. As both paralogues now undertake two different functions, they experience different selective pressures and are set upon separate, divergent evolutionary trajectories. In this model, there is no period where one of the paralogues is nonfunctional and provides an explanation for persistence of both copies until neofunctionalisation can occur.

A second reason for duplicate retention is subfunctionalisation, where the ancestral pre-duplication gene carries out several functions, perhaps via multiple protein domains (Force et al., 1999; Conant and Wolfe, 2008). Given duplication of a gene with two domains and functions, redundancy is present allowing mutations to inactivate a domain in one copy. At this point, subsequent mutations can lead to the loss of the second domain in the same copy, at which point pseudogenisation will likely occur for this paralogue. This would return this gene back to its single copy pre-duplication state. Alternatively, mutations can result in the inactivation of the second domain in the other paralogue. This would result in the two paralogues each carrying out one of the ancestral functions. This model is call "duplication-degeneration-complementation" (DDC), as neutral degeneration of function occurs after duplication and the resulting paralogues complement each as they have subfunctionalised. Both paralogues are now preferentially retained as the removal of either gene will likely be deleterious due to absence of its specific function.

A second subfunctionalisation model that explains duplication retention has been proposed called "escape from adaptive conflict" (EAC; Hittinger and Carroll, 2007). This model is very similar to the DDC model however the functions of the pre-duplication gene are not highly optimised for either function. Potential exists for adaptive mutations to refine both functions but at the cost of limiting
the other i.e. a conflict exists between adaptation of either function. Duplication provides the freedom for the two subsequent paralogues to specialise, optimising both functions independently having escaped the competitive situation prior to duplication. In the EAC model mutations that accumulate aren’t solely neutral as in the DDC model. The potential for positive selection to act on adaptive mutations is assumed to occur once duplication provides the necessary redundancy to escape conflict.

Subfunctionalisation can refer to the division of multiple functions but also of expression in multiple tissues (Li et al., 2005). The DDC model can be applied to promoters in addition to functional protein domains. Ohno (1970) proposed such a scenario, which was investigated by Ferris and Whitt (1979). After duplication, neutral mutations can accumulate in upstream or downstream regulatory motifs and inactivate expression in one or more tissues. (Papp et al., 2003b, see Figure 1). This leaves the duplicates expressed each in a distinct subset of the original ancestral tissues. Both genes are on independent evolutionary paths and will experience different selective pressures likely leading to further divergence.

Gene product dosage is another reason for duplicate retention. Gene duplication gives rise to increased gene copy number which in the absence of feedback mechanisms will increase mRNA level and protein product. If increased protein product is beneficial then gene duplication will immediately result in an advantageous scenario. Upon the creation of two paralogues, no additional processes are required to take place for both duplicates to be retained, their existence is sufficiently beneficial alone. An example of such a case occurs in *Plasmodium falciparum* (Price et al., 2004). Increased copy number of *pfmdr1* in *P. falciparum* confers resistance against mefloquine, a drug used to treat malaria. If mefloquine is a present selective pressure, duplications of *pfmdr1* will be immediately beneficial and retained. In human, another potential example has been characterised (Gon-
Introduction

A segmental duplication contains \( CCL3L1 \), a gene that encodes a suppressive chemokine of human immunodeficiency virus (HIV) and a ligand for the HIV co-receptor CCR5. Increased copy number of \( CCL3L1 \) reduces HIV susceptibility and likely is favoured in areas such as regions in Africa where HIV prevalence is high. This is supported by observations of \( CCL3L1 \) copy number where individuals in non-African populations have a median copy number of three compared to median copy number of six for individuals in African populations.

As evident in section 1.1.3, ‘Causes of pathogenicity’, genes can be dosage-sensitive meaning that for these genes a copy number change is harmful. Specifically for genes in stoichiometric balance with other genomic components, any aberration outside of a permitted expression range is deleterious. This raises the question under what conditions, if any, the process of gene duplication takes place for these genes? SSD will be deleterious as, unless the respective interacting genes are neighbouring and also included in the region duplicated, dosage imbalance will occur. The duplicates will be selected against, removed from the population and never reach fixation. However, a WGD event duplicates all genes and stoichiometric relationships together maintaining balance throughout the genome (Veitia, 2004; Veitia, 2005). Therefore, for even dosage-sensitive genes, WGD will be neutral as all genes are equally duplicated. However, after duplication, gene loss due to loss-of-function mutations or copy number losses (CNLs) will disrupt stoichiometric balance again and be deleterious. Hence, gene loss for dosage-sensitive genes should have distinct patterns after WGD events.

1.2.4 Biased retention of ohnologues

Only \(~30\%\) of human protein-coding genes are ohnologues, genes duplicated and retained following the vertebrate WGD events (Nakatani et al., 2007; Makino and McLysaght, 2010). A number of studies have noted differences between small-
scale duplicates (SSDs) and ohnologues. Davis and Petrov (2005) found that ohnologues in *S. cerevisiae* were enriched for catalytic proteins and depleted for binding proteins and enzyme regulators, while SSDs were enriched for enzyme regulators and depleted for transcriptional regulators. However there was no observed difference in codon bias or evolutionary rate. Additional studies in vertebrates have supported this, showing enrichment of binding, signalling, transcription regulation, and development functional classes amongst ohnologues (Blomme et al., 2006; Hufton et al., 2008). This functional enrichment is apparent in the genomes of teleost fish as well (Brunet et al., 2006).

Guan et al. (2007) investigated differences between duplicates arising from SSD and WGD in *S. cerevisiae*, and found that paralogues formed through WGD had more shared interaction partners and biological functions than SSDs. Furthermore, ohnologues were less likely to be essential but more likely to be synthetic lethal, that is where loss-of-function of either gene is neutral but loss-of-function of both is lethal. Hakes et al. (2007) also observed lower essentiality of ohnologues compared to SSDs.

However, the relationship with essentiality is not as simple in multicellular organisms (Makino et al., 2009). Ohnologues in mammals are as essential as singleton genes and therefore more essential than SSDs. Further, paralogues involved in development are more likely to be ohnologues than SSDs (Makino et al., 2009). This suggests preferential retention after the vertebrate WGD events, possibly due to dosage balance. In subsequent work Makino and McLysaght (2010) demonstrate that ohnologues are in fact often dosage-balanced and refractory to benign CNVs and SSD. Additionally they are found to be enriched for protein complex subunits. It seems that ohnologues have been retained in the human genome for ~400 MYs due to their loss causing stoichiometric imbalance and being deleterious. They are observed less on benign CNVs and therefore their fixation as SSDs are also
less likely. Ohnologues are seen to be enriched for genes involved in disease when perturbed and among genes found on pathogenic CNVs (McLysaght et al., 2014). This provides evidence that aberrations to their balance with respect to other genomic elements causes deleterious phenotypes and disease.

In chapter 3, ‘Dosage sensitivity is a major determinant of human copy number variant pathogenicity’, we examine the patterns of duplication and loss across mammalian genomes for human genes that are found within benign and pathogenic CNV regions. We explore the prevailing hypothesis that CNV pathogenicity is frequently due to the copy number change of one or more dosage-sensitive genes or regions found within a variant (Riggs et al., 2012). Using what we have learned of ohnologues that they are likely under an ancient, persistent expression constraint that predates the vertebrate WGD events we predict that this constraint should be evident in their copy number evolution across mammalian genomes. Specifically, we predict that copy number changes that give rise to human pathogenicity should have a reciprocal evolutionary trend and this trend should be shared across mammals.

1.3 Expression evolution

As early as the start of the 20th century, it was becoming apparent that there was a high degree of similarity between human and great ape blood proteins (Nuttall, 1904), with some human proteins having almost identical sequence to chimp or gorilla (Washburn, 1963). King and Wilson (1975) described the extraordinary and unexpected similarity of human and chimp protein sequences, comparing 44 structural proteins and establishing that they shared more than 99% sequence identity. In this seminal comparative genomics paper, the authors propose that due to such a high degree of similarity, differences in gene regulation may account
for the phenotype differences between human and chimp rather than extensive
differences in protein function. Mechanisms of controlling how and when genes
are expressed as an important factor in phenotypic evolution has gain empirical
evidence in the advent of abundant gene expression data (Romero et al., 2012).

1.3.1 Phenotypic evolution by expression change

Numerous examples of phenotypic evolution by gene regulatory changes have
now been identified. One such example is pelvis structural changes in threespine
stickleback fish (\textit{Gasterosteus aculeatus}, Shapiro et al., 2004). Freshwater stickle-
backs have lost all or almost all of the prominent pelvic skeleton found in marine
sticklebacks after less than 10,000 generations since divergence. Crosses between
marine and freshwater stickleback identified a gene \textit{Pitx1} in a region responsi-
ble for much of the pelvic size variation observed. The \textit{Pitx1} protein sequence
showed no sequence substitutions suggesting regulatory changes were responsible
for the phenotypic difference. Altered pattern of \textit{Pitx1} expression accounted for
differences in pelvic formation with no expression in the prospective pelvic region
and reduced expression in the caudal fin in freshwater sticklebacks. Other re-
gions where expression is observed were unaltered. These gene regulatory changes
suggest mutations in cis-acting regulatory elements of \textit{Pitx1}.

Rewiring of developmental gene regulatory networks has been witnessed also
in sea urchin (Hinman et al., 2003). Sea urchin (\textit{Strongylocentrotus purpuratus})
and starfish (\textit{Asterina miniata}) diverged from a common ancestor \textasciitilde 500 Mya and
share similar endomesodermal embryonic development with the exception that
starfish lacks a skeleton during embryogenesis. A three-gene feedback loop has
been independent conserved in both lineages for half a billion years yet a regulatory
gene involved in sea urchin skeletogenesis, \textit{tbr}, is used entirely differently in the
starfish embryo. \textit{Tbr} expression has newly acquired regulators that lead to its
involvement in formation of the primary digestive tube in the starfish embryo. Predicted cis-regulatory elements are likely responsible for the altered pattern of expression for a member of this important conserved gene regulatory network.

Another example of regulatory changes resulting in body plan evolution was revealed upon sequencing of several snake genomes (Kvon et al., 2016). The Zone of Polarizing Activity Regulatory Sequence is a limb-specific enhancer of the \textit{Shh} gene and is highly conserved across vertebrates. Enhancer activity is present across vertebrates except in snake species where it has progressively lost its function through the accumulation of SNVs during snake evolution. After synthetic introduction of a single ancestral transcription factor binding site lost after the snake divergence, \textit{in vivo} function was fully restored. This single enhancer’s sequence evolution is potentially the primary element responsible for snake limb loss.

A number of notable examples of expression evolution impacting human gene regulation exist. The Duffy blood group locus contains the human gene \textit{DARC} that encodes a chemokine receptor that localises in the membrane of red blood cells. This receptor acts as a point of invasion for malarial parasites. An allele responsible for the absence of the receptor on red blood cells has been positively selected for in most sub-Saharan African populations but is rare elsewhere (Tournamille et al., 1995; Hamblin and Di Rienzo, 2000). A single nucleotide substitution upstream of \textit{DARC} disrupts a binding site, consequently impairing promoter activity in red blood cells. Abolishing expression in this manner confers increased protection against malarial infection.

In most mammalian species, lactase, the enzyme used to digest lactose sugar found in milk, has reduced expression after weaning (Swallow, 2003). However, in some human populations a recent adaptation has meant that the downregulation of lactase does not occur and expression persists into adulthood (Bersaglieri et al., 2004). In Europeans, a SNV within the gene encoding lactase, \textit{LCT} accounts for
the persistence observed but in Tanzanian, Kenyan and Sudanese populations three SNVs instead enhance LCT expression (Tishkoff et al., 2007). This is a remarkable example of positive selection for expression evolution rather than functional changes to converge on the same phenotype.

In 2002, divergences in gene expression between primate species were quantified (Enard et al., 2002). A large number of tissue-specific expression changes were revealed between these closely related species, particularly so in human brain tissue. Additionally in 2005, Rockman et al. found cis-regulatory changes of the human gene prodynorphin expressed in the brain (Rockman et al., 2005). Prodynorphin is a precursor of opioids and neuropeptides, notably endorphins, and is implicated in pain perception, memory, learning, and social attachment and bonding. The peptide sequence of prodynorphin is identical between chimp and human but a 68 bp tandem repeat in the human promoter has been selected for and increases transcript inducibility.

Given these examples of substantial phenotypic changes it is therefore reasonable to investigate the interplay between gene expression evolution and evolution by duplication.

1.3.2 Expression divergence of gene duplicates

Gu et al. (2002) compared synonymous substitution divergence and correlation in expression profile between duplicates in *S. cerevisae*. Using synonymous substitutions as a measure for evolutionary time, they found that only a small fraction of ancient gene pairs have conserved expression patterns. Largely they observed that expression profile divergence between duplicate genes are correlated with their synonymous sequence divergence and thus evolutionary time. This data suggests that conservation of duplicate expression profiles is not a common or widespread phenomenon at least in yeast.
Analysis of ohnologues produced from polyploidy events in the evolutionary history of the multicellular organism, *Arabidopsis thaliana*, yielded insights into functional divergence through expression changes (Blanc and Wolfe, 2004). Immediately after duplication through a WGD event, duplicates share identical transcriptional profiles and over time they can evolve and diverge in expression patterns. In a comparison of the most recent ohnologues within *A. thaliana*, more than half of the ohnologue pairs had diverged in expression from one another. More interesting is that this expression divergence is non-random and groups of ohnologues diverge similarly to form parallel networks each containing one ohnologue from the duplicate pair. Expression is seen to be highly correlated within each network but not with its corresponding ohnologue partner. These patterns provide both a method of functional divergence and duplicate gene retention after WGD events, similar to a scenario proposed by Force et al. (1999).

Similar trends were observed by Haberer et al. (2004) in *A. thaliana*, whereby duplicate pairs had highly similar expression characteristics. Expression divergence is common however, and the authors propose that the DDC model of duplicate gene retention acting on cis-regulatory elements could explain how frequently divergence in expression occurs. Further work in *A. thaliana* suggests that divergence likely takes place either at or shortly after duplication (Ganko et al., 2007). In contrast to Gu et al. (2002) however, Ganko et al. (2007) find no evidence of relationship between expression divergence and synonymous substitutions. However, they find a strong association between nonsynonymous substitutions and expression evolution in duplicates from WGD events, suggesting correlation between purifying selection on sequence and expression.

Remarkable expression divergence has been observed in another plant, cotton (*Gossypium raimondii*, Renny-Byfield et al., 2014). Expression in petal, leaf, and seed tissues were measured and differential expression was found between 99.4%
of ohnologue pairs in at least one of the tissues. Strikingly, 85% of ohnologue pairs displayed divergence in all three tissues tested suggesting that either neofunctionalisation or subfunctionalisation of expression is responsible for their retention rather than a constraint to maintain a specific gene product dosage. This is consistent with other analyses in plants (Duarte et al., 2006; Throude et al., 2009; Guo et al., 2013), which raises questions about how stoichiometric balance relates to this trend and if dosage-sensitive genes are enriched among the remaining 15% of ohnologues pairs that have conserved expression in at least one tissue.

The orthologue conjecture puts forward the idea that orthologues are more similar functionally than paralogues (Koonin, 2005), and forms the basis for projection of function identified in one species to orthologues of unknown function in other related species. Numerous studies have rejected the hypothesis (Nehrt et al., 2011; Yanai et al., 2004) but there is more evidence that the conjecture holds true (Liao and Zhang, 2006; Brawand et al., 2011; Huerta-Cepas et al., 2011; Altenhoff et al., 2012; Chen and Zhang, 2012). Recently, a study among vertebrate species and Drosophila melanogaster confirmed strongly conserved orthologue tissue-specificity (Kryuchkova-Mostacci and Robinson-Rechavi, 2016), however could not distinguish between ohnologues and SSDs, hampering insights into dosage balance constraints.

Gout and Lynch (2015), however, specifically investigated ohnologues and expression divergence in several Paramecium species. They found a strong correlation between the most recent ohnologues in Paramecium biaurelia for absolute expression level, that is that ohnologue pairs show signs of constraint for maintaining expression level post-WGD. In the model of post-WGD evolution proposed, they posit that expression level of individual duplicates are free to vary and evolve neutrally provided that combined they maintain a summed absolute dosage that is within a tolerable threshold. This model goes some way to reconciling expression
divergence and dosage-sensitivity of retained duplicates after WGD.

1.3.3 Expression variation

Expression quantitative trait loci (eQTLs) are genomic regions harbouring sequence variants that influence the expression level of one or more genes (Albert and Kruglyak, 2015). While mapping of eQTLs affecting single genes has been conducted for decades, genome-wide eQTL mapping is about 15 years old (Brem et al., 2002; Schadt et al., 2003). Since then many mapping experiments have been undertaken in various species (Morley et al., 2004; Cheung et al., 2005; Stranger et al., 2005; Stranger et al., 2007; West et al., 2007; Dimas et al., 2009; Kelly et al., 2012; Massouras et al., 2012; GTEx Consortium, 2015). A range of positive and negative expression effect sizes, that is increased and decreased expression, are observed. These eQTL effects can occur in a tissue-specific manner or across a number of tissues, however, tissue-specific influence is more typical (Gerrits et al., 2009). In human, the expression of thousands of genes is affected by eQTLs making them a significant contribution to the genetic variation of expression and in turn phenotypic variation and complex disease.

1.3.4 Role in disease

Most human protein-coding genes are influenced by eQTLs in the general population (GTEx Consortium, 2015). Therefore, the majority of the genome must be able to tolerate some amount of mRNA level change without obvious deleterious consequences. However, in combination with genome-wide association studies eQTLs have been used to elucidate further the pathophysiology of many disease phenotypes. To date eQTLs have been associated with numerous human diseases including asthma, autoimmune disorders, diabetes, numerous cancers, Parkinson’s disease, and other brain disorders (see Table 1 in Albert and Kruglyak, 2015).
Additionally, eQTLs have been shown to undergo increased purifying selection with gene age where young, primate-specific genes are enriched for eQTLs, having higher effect size and influencing expression in more tissues (Popadin et al., 2014). Therefore, the effect of eQTLs on gene expression and association with important traits makes them worthy of study especially in the context of genes with known expression constraints.

Dosage-sensitivity genes are refractory to other types of variation, e.g. CNVs, and due to strong gene product constraints we expect that dosage-sensitivity genes are also constrained for variants within promoters and enhancer regions that affect expression level. We wished to investigate further the evolution of expression for dosage-constrained genes and test the patterns of human eQTLs affecting dosage-sensitive genes in chapter 5 – ‘Expression quantitative trait loci of dosage-sensitive genes have narrow tissue specificity bias’.

1.4 Aim

This thesis is, as far as current data allows, an examination of the evolution of copy number and expression variation of human dosage-sensitive protein-coding genes. In chapter 3, copy number evolution of human genes affected by benign and pathogenic CNVs is analysed through a comparative genomics study of mammalian genomes. In chapter 4, a CNV pathogenicity classifier is constructed and used to gain insights into CNV inheritance. Finally in chapter 5, evolution of expression variation is analysed by examining the patterns of human expression quantitative trait loci found in healthy individuals.
Chapter 2

Materials & Methods

This chapter provides an introduction and detailed explanation to some of the methods used throughout this thesis.

2.1 Filtering dbVar CNV data

dbVar is the National Center for Biotechnology Information’s (NCBI’s) database of genomic structural variation (Lappalainen et al., 2013).

A combination of different studies from dbVar were used to form a dataset of germline CNVs with clinical interpretations that are used in the analysis of chapter 3 and the training and testing of the classifier in chapter 4. CNVs and the studies from which they originate are listed in Table 2.1.

CNVs submitted for human GRCh37 genome assembly were taken as is and CNVs from earlier human genome assemblies were remapped automatically in dbVar. Both the GRCh37 submitted and remapped germline dbVar datasets released on 31st Oct 2013 were filtered for clinical assertions of benign and pathogenic. CNVs longer than a tenth of their respective chromosome were removed from the final dataset.

The python script used to implement dbVar filtering is available online at https:
### Materials & Methods

#### Table 2.1 | dbVar studies included in CNV analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of CNVs included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller et al. (2010)</td>
<td>7,586</td>
</tr>
<tr>
<td>Kaminsky et al. (2011)</td>
<td>2,507</td>
</tr>
<tr>
<td>Wapner et al. (2012)</td>
<td>1,773</td>
</tr>
<tr>
<td>Mitsui et al. (2010)</td>
<td>173</td>
</tr>
<tr>
<td>Riggs et al. (2012)</td>
<td>67</td>
</tr>
<tr>
<td>dbVar user submitted curated variants from OMIM, GeneReviews, or ClinVar (Lappalainen et al., 2013)</td>
<td>34</td>
</tr>
<tr>
<td>Sharp et al. (2008)</td>
<td>9</td>
</tr>
<tr>
<td>Zhang, Davis, et al. (2009)</td>
<td>6</td>
</tr>
<tr>
<td>Sharp et al. (2007)</td>
<td>4</td>
</tr>
<tr>
<td>Lopez-Herrera et al. (2012)</td>
<td>1</td>
</tr>
<tr>
<td>CNVs lacking study IDs in 2016 (Lappalainen et al., 2013)</td>
<td>211</td>
</tr>
</tbody>
</table>


#### 2.2 Counting mammalian gene gain and loss events

We wished to identify and count copy number changes during mammalian evolution for as many human protein-coding genes as possible. Identifying homology is complex, and although this is often achieved through comparisons of sequence similarity, homology is a state of shared common ancestry and not just similarity in sequence (Reeck et al., 1987). Adaptive and purifying selective pressures, functional convergence, fission/fusion events, and other influences can impact sequence similarity and complicate our attempts to decipher ancestry. For these reasons it can be difficult to determine homology through sequence similarity.
2.2.1 Homologues and trees

We inferred gene duplications and losses from Ensembl Compara annotations (Herrero et al., 2016). The Compara pipeline, uses automated and manually curated gene annotations from species included in the Ensembl database (Cunningham et al., 2015) to determine homologies (Vilella et al., 2009). At the start of this pipeline, a representative protein isoform is chosen for each annotated protein-coding gene. BLAST is run to identify similarity (Camacho et al., 2009) and protein sequences are clustered using Hcluster_sg from Treefam (Li et al., 2006). A multiple protein sequence alignment is constructed for each cluster using either M-Coffee (Wallace et al., 2006) or MAFFT (Katoh and Standley, 2013), depending on cluster size. The original coding DNA sequences are fit into the protein multiple alignment to create a codon alignment which is input to TreeBeST to create a phylogenetic tree. At this point, orthologue and paralogue relationships are identified through pairwise gene relationships within the phylogenetic tree. For example, for genes of different species and where an ancestral node is a speciation event the gene pairwise relationship will be orthologous. Similarly, for genes within the same species and where an ancestral node is a duplication event the gene pairwise relationship will be paralogous.

We wished to calculate the number of mammalian genomes where gene copy number differed to copy number in human. Thirteen mammalian genomes (Bos taurus, Callithrix jacchus, Canis lupus familiaris, Equus caballus, Felis catus, Gorilla gorilla, Macaca mulatta, Mus musculus, Oryctolagus cuniculus, Ovis aries, Pan troglodytes, Rattus norvegicus, Sus scrofa) were used in the analysis as they were of sufficiently high quality. Gene duplications and losses in these 13 genomes were calculated for all human gene families inferred to be present in the mammalian common ancestor. In this way, all families have an equal time to undergo gene duplication or loss events since the mammalian divergence. Presence in the
mammalian common ancestor was determined using Ensembl ‘gain/loss gene trees’. These differ from phylogenetic trees as they are estimates of gene family evolution given gene family sizes in extant species and are calculated by CAFE (Han et al., 2013). The CAFE program infers gene family sizes at internal nodes of the species tree using birth and death rates. Using these ‘gain/loss gene trees’, it was possible to estimate gene family size at the tree node of the mammalian ancestor. Only gene families with at least one gene at this node, i.e. a copy number of at least one, were included in subsequent analysis steps. This step excluded 373 protein-coding genes. Additionally, 307 human protein-coding genes were not included in an Ensembl gene tree, and a further 374 genes did not have a ‘gain/loss gene tree’. These genes were also excluded as homologous relationships could not be inferred. Therefore, mammalian copy number could be calculated for 19,260 human protein-coding genes.

2.2.2 Analysis workflow

For counting the number of species where copy number is (un)changed, most genes can be considered independently, this is because they predate the mammalian divergence and do not change in copy number on the lineage leading to human. An example would be a species-specific duplication event in mouse with all of the other genomes having one orthologue. Here, 12 species would be counted as having unchanged copy number and one species as having a duplication. In this simple case, Ensembl pairwise homologies are used where orthologues are labelled as having a one-to-one orthologous relationship or a one-to-many relationship. One-to-one orthologues are counted as having ‘unchanged’ copy number. One-to-many orthologues are counted as ‘duplicated’ and if no orthologue for that species is present it is considered as having ‘no orthologue’.
Copy number changes along human lineage

For genes that have undergone a copy number change since the mammalian divergence on the lineage leading to human, a more complex approach is required. In this scenario, duplicated genes belonging to the same duplication event are grouped as one unit and compared to the other mammalian genomes. To achieve this, the copy number status in other mammalian genomes is determined for genes unaffected by copy number changes along the human lineage first. Overlaps in orthologues of ‘affected’ human genes are then examined and ‘affected’ genes are grouped where they share at least one orthologue. These grouped human genes are assumed to be duplicates from the same duplication event. Each group is treated as having a copy number of one and the genes within the group are assigned the same copy number counts for the other mammalian genomes. These counts are determined by counting the number of grouped orthologues for each species. If one orthologue is present for a species then that species is counted as showing ‘unchanged’ copy number. If more than one orthologue is present that species is counted as a ‘duplicated’ and if no orthologue for that species is present in the group it is counted as having ‘no orthologue’.

2.2.3 Annotation and error

As this analysis is performed genome-wide, it is not possible to manually verify the results for all genes. A number of error sources are possible. Firstly, if a gene is present, but not annotated, in one of the 13 mammalian genomes used, it will be considered a gene loss event. Secondly, errors in gene trees and ‘gene gain/loss trees’ will affect the homologies determined prior to counting of genomes affected by a copy number change. Thirdly, errors in our analysis for counting, particular if a complex evolutionary history have given rise to a number of copy number changes occurring on the branch leading to human. In this scenario, our analysis
may overestimate or underestimate the number of genomes showing copy number changes. While manual curation of all human protein-coding genes is not possible, it is possible to manually assign members of some gene families a copy number count by hand and compare the performance of our analysis. Doing so we find that 86.7% (176/203) of genes are correctly assigned the correct counts for number of species with unchanged copy number, duplications, and missing orthologues. Excluding cases where it was not possible to determine copy number by hand (due to highly complex gene trees that do not follow clear species tree topology), 97.4% (185/190) of cases were between $+/−1$ of the manually assigned copy number counts. Additionally, 9 of the 27 incorrect estimates seem likely caused by annotation and homology errors rather than errors in our analysis, e.g. a gene split over over two contigs in pig, an unlabelled gorilla duplication, a duplication in dog that isn’t an annotated orthologue of the human gene, several genes being considered older than mammalian divergence but likely only annotated as such due to spurious orthologues in distant species. Therefore, we have reasonably high confidence that our analysis can show clear evolutionary signal given some noise introduced by these errors.

2.2.4 Analysis code

The python script used to count mammalian gene gain and loss events is available online at https://github.com/alanrice/paper-dosage-sensitivity-copy-number-variation/blob/master/analysis/copyNumberAnalysis.py.
Chapter 3

Dosage sensitivity is a major determinant of human copy number variant pathogenicity

The research described in this chapter has been published in *Nature Communications* (Rice and McLysaght, 2017).

3.1 Introduction

Copy number variants (CNVs) are regions of the genome that are duplicated or deleted in some individuals in a population. CNVs are most intensely studied in human but have been observed and characterised to a lesser extent in other species (Völker et al., 2010; Liu et al., 2010; Nicholas et al., 2009; Li et al., 2012; Pezer et al., 2015; Debolt, 2010; Bai et al., 2016). Accounting for more variation than single nucleotide polymorphisms in terms of base pair length, CNVs are abundant in human genomes. Each individual has on average 1000 CNVs of greater than 450bp with respect to the reference genome (Conrad et al., 2010). CNVs segregate in the population but also arise *de novo* (Conrad et al., 2010;
Dosage sensitivity underlies human CNV pathogenicity

Zarrei et al., 2015). Some regions in the genome are CNV hotspots – approximately 10% of the human genome experiences recurrent CNV events (Mefford and Eichler, 2009).

Often this variation does not produce a phenotype as CNVs are frequently small, intergenic or encompass genes that can tolerate a change in copy number. Some genes can even be completely deleted with no apparent effect (Zarrei et al., 2015). However, CNVs have previously been associated with a number of human conditions, most notably neurodevelopmental disorders including autism spectrum disorders, schizophrenia, intellectual disability, attention deficit hyperactivity disorder, developmental delay and epilepsy (Sebat et al., 2007; Stefansson et al., 2014; Stefansson et al., 2008; Walsh et al., 2008; Mefford et al., 2008; Helbig et al., 2009; Cooper et al., 2011). Due to this implication in disease, CNVs are subject to increasingly intense scrutiny to understand and characterise their genetic and phenotypic effects.

There is more than one possible mechanism by which a CNV can disrupt gene function and cause a phenotype, including disruption of chromosome structure, interference with regulatory elements, and perturbation of relative amounts of dosage sensitive genes (Zhang, Gu, et al., 2009). Several recent studies have shown a relationship between topologically associated domains (TADs) and genomic duplication effects (Xie et al., 2016; Franke et al., 2016). Still, the prevailing hypothesis on CNV pathogenicity is that it is due to dosage sensitivity of the included genes. One of the first well-characterised cases of CNV pathogenicity was Charcot-Marie-Tooth neuropathy which was specifically linked to CNV of the dosage-sensitive gene peripheral myelin 22 (PMP22) (Lupski et al., 1991). Dosage-sensitivity provides a model whereby a 50% increase or decrease in gene copy number is deleterious (Veitia, 2002; Papp et al., 2003a; Birchler et al., 2001; Birchler and Veitia, 2012). There are a number of reasons for dosage-sensitivity,
firstly, dosage-sensitive genes may be in stoichiometric balance with other genes, for example as in protein complex members (Papp et al., 2003a). Secondly, dosage-sensitive genes may operate in a concentration-dependent fashion as observed for developmental morphogens (Rogers and Schier, 2011) or some splicing cofactors (Chen et al., 2012). Thirdly, dosage-sensitive genes may produce proteins that are aggregation-prone at high concentrations as in the case of the SNCA protein (Miller et al., 2004). Additionally, dosage-sensitivity may arise due to a minimum required concentration to achieve functionality, observed for haploinsufficient genes, including many transcription factors and developmental genes (Fisher and Scambler, 1994). When the dosage of these genes is changed by an overlapping CNV the function of the gene is disrupted in a way that we may observe as disease. More acutely dosage-sensitive genes may never be observed in CNVs, even in pathogenic ones, if they are so disruptive as to result in inviability. Thus duplication and/or loss CNVs of dosage-sensitive genes are not expected to be observed in healthy individuals (Makino and McLysaght, 2010).

As dosage-sensitivity is linked to relative abundances rather than absolute amounts, whole genome duplication (WGD) is tolerable because, by definition, all genes are duplicated equally. Unlike CNVs and small-scale duplications (SSDs), WGD events preserve gene stoichiometry. Two such events occurred early in the vertebrate lineage and were followed by extensive genome rearrangement and massive gene loss (Ohno, 1999; McLysaght et al., 2002; Dehal and Boore, 2005; Nakatani et al., 2007). Duplicated genes retained from these events (ohnologues) are found to be refractory to CNVs and SSDs, that is, they evolve in a pattern that suggests ancient and persistent dosage-sensitivity (Makino and McLysaght, 2010). Ohnologues are depleted among CNVs found in healthy individuals (Makino et al., 2013), and are found to be overrepresented among genes on pathogenic CNVs (McLysaght et al., 2014) providing further supporting evidence that they
Dosage sensitivity underlies human CNV pathogenicity are under dosage constraint.

Here the evolutionary history of genes in CNVs with different clinical interpretations is examined with the aim of creating a deeper understanding of the predictive power of evolutionary patterns for understanding CNV pathogenicity. We explore the prevailing hypothesis that CNV pathogenicity is frequently due to the copy number change of one or more dosage-sensitive genes or regions found within a variant (Riggs et al., 2012) and predict that this dosage-sensitivity will similarly constrain their evolution in mammals in characteristic ways. Consistent with this hypothesis, we find that orthologues of human genes found in pathogenic CNV regions are rarely duplicated or lost in the mammalian lineage. Conversely, genes overlapped by benign variants have highly variable copy number across the tested species. Furthermore, we find that genes with conserved copy number across mammals are depleted among CNVs in non-human healthy mammals, mirroring the pattern observed in humans. These results demonstrate the role of dosage sensitivity in shaping the human genome and point to the usefulness of evolutionary metrics in refining the lists of candidate causative genes on pathogenic CNVs.

3.2 Materials & Methods

3.2.1 CNV Data with clinical interpretation

Human autosomal germline copy number variants with clinical interpretations of ‘benign’ and ‘pathogenic’ were obtained from dbVar release dated 31st October 2013 for genome assembly GRCh37 (Lappalainen et al., 2013). CNVs longer than a tenth of a chromosome were discarded and the included CNVs coordinates are summarised in Table 3.1. For more information on filtering and the specific dbVar studies included in this analysis see section 2.1, page 33.
3.2.2 CNV coverage

CNV coverage (number of CNVs overlapping a given region of genome) was calculated genome-wide using Bedtools (Quinlan and Hall, 2010). Within each CNVR, peak regions were calculated as any local maximum in CNV coverage, defined as any subregion with higher coverage than its flanking regions. Multiple peak regions were permitted within a CNVR. Where a given region has only one CNV, the entire CNV is counted as the peak region. Protein-coding gene annotations and gene ontology (GO) terms were obtained from Ensembl GRCh37 (Cunningham et al., 2015). A gene was considered to be intersecting with a CNV if the any of the gene sequence was overlapped by 1 or more bases on either strand.

3.2.3 Gene category enrichment

Median RPKM values by tissue were obtained from GTEx V6 (GTEx Consortium, 2015). The tissue with the maximum RPKM value was used for each gene. Probability of loss-of-function mutation intolerance values for genes were obtained from ExAC Release 0.3 (Lek et al., 2016) as a proxy for haploinsufficiency scores. Protein complex member genes were sourced from the Uniprot KB/Swiss-Prot database (The UniProt Consortium, 2015) by filtering for keywords in the ‘Subunit structure’ annotation field.

3.2.4 Mammalian copy number analysis

Gene duplications and losses in 13 mammalian genomes (Bos taurus, Callithrix jacchus, Canis lupus familiaris, Equus caballus, Felis catus, Gorilla gorilla, Macaca mulatta, Mus musculus, Oryctolagus cuniculus, Ovis aries, Pan troglodytes, Rattus norvegicus, Sus scrofa) were calculated for all human genes inferred to be present in the mammalian common ancestor. Gene duplications and losses were
Dosage sensitivity underlies human CNV pathogenicity inferred from Ensembl Compara annotations (Cunningham et al., 2015; Herrero et al., 2016). For each of the 13 species a given human gene was considered to be duplicated in that genome if the annotation was one-to-many. We also counted the number of instances where the Ensembl Compara annotation reports no orthologue in that genome as presumed gene loss events. Genes with a one-to-one orthologous relationship were counted as unchanged in that genome. Where a gene has undergone a duplication event since the mammalian divergence and two duplicates have persisted to present in human, both genes cannot be treated independently during analysis. Doing so could potentially confound the number of copy number change counts because all mammalian genomes tested not sharing the duplication would be counted as being changed with respect to human. Thus these recent human paralogues were grouped and their ancestral copy number of one was compared to the copy number of each other species. For a more detailed explanation see section 2.2, page 34.

### 3.2.5 22q11 conserved synteny analysis

For Figure 3.9b, grey dashed outlines show orthologue groups that are neighbouring on their respective chromosome/scaffold in each species. When no orthologue is present for a gene but orthologues exist for flanking genes and are neighbouring, one bounding outline groups all genes. This permits grouping in cases where additional genes are annotated within orthologous sequence in one of the genomes or where pseudogenisation has occurred. Neighbouring groups are broken when chromosome/scaffold changes, region inversion occurs breaking gene collinearity, or where position shifts substantially on the same chromosome/scaffold to a non-neighbouring region.
3.2.6 CNV data for comparative genomics

For comparative analysis of CNVs between species, human CNV data without clinical interpretation (presumed healthy) were obtained from the inclusive map provided in Zarrei et al. (2015). Mouse CNV data for wild-caught mice from four populations were obtained from Pezer et al. (2015). There is no pathogenicity information explicitly listed, but they are presumed to represent healthy control variation. CNVs are identified on every mouse chromosome. Mouse orthologues of human genes with one-to-one relationships in all 13 mammalian species tested were intersected with mouse CNVs. Genes overlapped by 1 or more bases were considered to be affected by CNVs.

3.2.7 Gene ontology (GO) enrichment analysis

Developmental genes were defined as those with GO term “developmental process” (GO:0032502). GO term enrichment of solitary non-passenger pathogenic genes was examined using g:Profiler (Reimand et al., 2011) with genes within full pathogenic CNVRs as a custom background gene list. The significance threshold was adjusted by Bonferroni correction. Genes that show one-to-one orthology in all 13 mammalian species were tested for GO enrichment compared to the full list of human protein-coding genes using g:Profiler.

3.3 Results

3.3.1 Identification of pathogenic CNV peak regions

We obtained human autosomal germline copy number gains (CNGs) and losses (CNLs) with clinical interpretations of ‘benign’ or ‘pathogenic’ from dbVar (Table 3.1). The operational definition of a CNV varies between studies, but in the
Dosage sensitivity underlies human CNV pathogenicity
data used here the minimum length of a CNV is 50 bp. Furthermore, we excluded CNVs that were greater than 10% of the length of the respective chromosome, as these dramatically increase the number of genes included and potentially confound the analysis. Although benign CNVs outnumber pathogenic CNVs by about 2:1, the proportion of genome covered by any pathogenic CNV (74.4%) is much larger than that covered by benign CNVs (8.3%), due to the substantially longer average length of pathogenic CNVs.

CNVs are described by their start and end points and whether they are gain or loss events (CNG and CNL respectively). A given region of genome may be overlapped by multiple CNVs with different start and end points, of different types (gain or loss), and of different clinical interpretations. Sets of partially overlapping CNVs are grouped together into CNV regions (CNVRs). By contrast other regions have no observed CNVs at all, or only have rare CNVs.

The number of genes included in pathogenic CNVs seems implausibly large for them all to be causative of disease (Figure 3.1 and 3.2). Rather, it is probable that only a subset of the genes in the pathogenic CNV regions are responsible for the associated phenotypes. We observe that 87.1% (223.8 Mb/256.9 Mb) of benign CNVR is overlapped by pathogenic CNVs. Thus, we wanted to refine the CNVs to home in on likely causative genes.

Even in well-characterised pathogenic CNV regions such as 22q11, the start and end points of the CNV region are variable between patients. However, a “critical” 1.5 Mb region has been identified which is common to most cases and it is usually inferred that the primary causative genes are present within this region (Karayiorgou et al., 2010). Similarly, Down’s Syndrome is caused by trisomy of chromosome 21, but a short segment of the chromosome has been linked to most symptoms and is considered the Down’s Syndrome Critical Region (Antonarakis et al., 2004). Mirroring this approach, we identified recurring subregions of pathogenic
### Table 3.1 | Summary of human CNVs used in CNV analysis

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>Average length (kbp)</th>
<th>Combined length (Mbp)</th>
<th>Genome coverage</th>
<th>PC genes</th>
<th>PC developmental genes</th>
<th>Contain 1+ developmental genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full</td>
<td>8,005</td>
<td>388.5</td>
<td>3,110.1</td>
<td>8.9%</td>
<td>1,802</td>
<td>28.0% (2,242)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>769</td>
<td>334.0</td>
<td>256.9</td>
<td>4.4%</td>
<td>1,108</td>
<td>18.8% (187)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>993</td>
<td>128.3</td>
<td>127.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benign CNVs</td>
<td>Full</td>
<td>4,306</td>
<td>400.1</td>
<td>1,723.0</td>
<td>6.2%</td>
<td>1,369</td>
<td>26.5% (1,139)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>494</td>
<td>362.0</td>
<td>178.9</td>
<td>3.3%</td>
<td>875</td>
<td>20.4% (126)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>619</td>
<td>82.9</td>
<td>94.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benign CNGs</td>
<td>Full</td>
<td>3,699</td>
<td>375.0</td>
<td>1,387.1</td>
<td></td>
<td></td>
<td>29.8% (1,103)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>445</td>
<td>271.6</td>
<td>120.9</td>
<td></td>
<td></td>
<td>21.1% (94)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>506</td>
<td>143.9</td>
<td>72.8</td>
<td></td>
<td></td>
<td>17.2% (87)</td>
</tr>
<tr>
<td>Benign CNLs</td>
<td>Full</td>
<td>4,366</td>
<td>3,503.0</td>
<td>15,294.2</td>
<td></td>
<td></td>
<td>95.4% (4,166)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>167</td>
<td>13,856.3</td>
<td>2,314.0</td>
<td></td>
<td></td>
<td>92.8% (155)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>923</td>
<td>545.0</td>
<td>503.1</td>
<td></td>
<td></td>
<td>58.2% (537)</td>
</tr>
<tr>
<td>Pathogenic CNVs</td>
<td>Full</td>
<td>1,097</td>
<td>3,985.1</td>
<td>4,371.6</td>
<td>48.4%</td>
<td>11,217</td>
<td>97.2% (1,066)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>178</td>
<td>7,840.5</td>
<td>1,395.6</td>
<td></td>
<td></td>
<td>92.1% (164)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>300</td>
<td>1,861.6</td>
<td>558.5</td>
<td></td>
<td></td>
<td>76.7% (230)</td>
</tr>
<tr>
<td>Pathogenic CNGs</td>
<td>Full</td>
<td>3,269</td>
<td>3,341.3</td>
<td>10,922.6</td>
<td></td>
<td></td>
<td>94.8% (3,100)</td>
</tr>
<tr>
<td></td>
<td>Regions</td>
<td>212</td>
<td>9,196.1</td>
<td>1,949.6</td>
<td></td>
<td></td>
<td>92.0% (195)</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>699</td>
<td>669.7</td>
<td>468.1</td>
<td></td>
<td></td>
<td>63.7% (445)</td>
</tr>
</tbody>
</table>

**Figure 3.1 | Number of genes per copy number variant (CNV) in dataset**

Histogram of the number of genes per CNV separated by CNV clinical interpretation (benign/pathogenic). Bin width set to 5.
Dosage sensitivity underlies human CNV pathogenicity

Figure 3.2 | Number of genes per CNV region (CNVR)
Histogram of the number of genes per CNVR separated by CNV clinical interpretation (benign/pathogenic). Bin width set to 10.

CNVRs as we consider them more likely to contain the causative genes.

We refined pathogenic CNVRs into “peak regions” defined as local maximums of CNV coverage. This approach has the advantage of promoting recurrent CNV subregions for special attention while also avoiding discriminating against rare CNVs in the dataset; in cases where there is only one CNV in a region, the entire CNV is the local “peak” (with a coverage of 1). We preferred this method to selecting an arbitrary genome-wide coverage threshold, as such an approach would exclude rare CNVs and low coverage regions, and would fail to refine high coverage regions. This is important as some rare CNVs have been implicated in disease (Kirov et al., 2014; Walsh et al., 2008), and there are likely to be more, as yet uncharacterised, rare CNVs that are causative of disease.

Using this peak region approach, 167 CNVRs (composed of 4,366 individual CNVs, and grouping duplication and deletion CNVs) which cover over 74% of the genome and encompass 16,343 protein-coding genes were broken into 923 peak
Dosage sensitivity underlies human CNV pathogenicity

regions (a CNVR can have multiple local peaks) covering 16.2% of the genome and 4,234 genes (Table 3.1). Some of these peak coverage pathogenic regions overlap benign CNVs, intersecting with 16.7% (42.9 Mb/256.9 Mb) of benign CNVR.

A similar analysis can also be applied to benign CNVs (shown in Table 3.1 for comparison), but it makes little sense to analyse the benign CNVs in this way as we presume that the entire region is benign.

3.3.2 Pathogenic CNVs are enriched for developmental genes

CNVs have been associated with diverse conditions such as heart disease, cancers, immunodeficiency, hearing loss, and obesity (Cooper et al., 2011; Jacquemont et al., 2011; Conrad et al., 2010; Craddock et al., 2010; Walters et al., 2010; Greenway et al., 2009; Zhang, Gu, et al., 2009; Orange et al., 2011; Glessner et al., 2014; Shearer et al., 2014). However, they are most often associated with developmental conditions with over 14% of developmental delay and intellectual disability cases caused by CNVs (Cooper et al., 2011). This makes intuitive sense as development is considered to be a finely-balanced, dosage sensitive process (Fisher and Scambler, 1994; Rogers and Schier, 2011). Nonetheless, one must be careful to consider the possibility of an ascertainment bias: it is not possible to know if a given individual will develop heart disease later in life so they will be noted as healthy, whereas developmental conditions are early onset by definition and so should always be observed when present. Thus, it is not currently clear if the apparent enrichment for developmental conditions reflects a detection bias or a greater inherent vulnerability in developmental processes.

We found that 95.4% of full pathogenic CNVs in the current dataset contain at least one developmental gene, compared to only 28.0% of benign CNVs. However, as pathogenic CNVs are typically longer and cover such a large proportion of the genome it is expected that they will contain more genes and in turn are more likely
Figure 3.3 | Percentage of CNVs containing at least one developmental gene for 1000 randomised sets
Pathogenic CNV positions were shuffled randomly 1000 times, after each the percentage of CNVs that contain at least one developmental gene was calculated. The observed value for dbVar pathogenic CNVs is overlaid as a black line. Bin width set to 0.3.
Dosage sensitivity underlies human CNV pathogenicity

to contain a gene involved in any given Gene Ontology (GO) category. Thus it is necessary to correct for differences in CNV length. We did this by calculating the proportion of genes on each CNV that are developmental genes. When we considered individual pathogenic CNVs that were not overlapped by benign CNVs (i.e., exclusively pathogenic regions) a mean of 37.3% of the genes were developmental genes compared to 24.2% of benign CNV genes (medians 28.4% and 0% respectively), a highly significant difference \( (P < 1.0 \times 10^{-16}, \text{Mann-Whitney U test}) \). As an alternative correction for length difference we randomised the location of the pathogenic CNVs and counted the number containing at least one developmental gene. We repeated this simulation 1000 times. Over these simulations the mean percentage of CNVs that overlapped at least one developmental gene was 74.8% and the highest percentage found in any simulation was 76.9%, significantly less than observed in the real data \( (P < 1 \times 10^{-16}; \text{Z-score: 37.2}; \text{Figure 3.3}) \).

Comparing pathogenic peak regions to benign CNVRs (full merged CNVs), pathogenic regions that have no benign overlap were significantly enriched for containing at least one developmental gene (58.5% of 684 pathogenic regions vs. 27.1% of 769 benign regions, \( P < 1.0 \times 10^{-16}, \chi^2 \text{ test}) \). Although the lengths of full benign CNVRs (mean 334.0 kb; median 151.5 kb) and pathogenic peak region CNVRs (mean 545.6 kb; median 184.0 kb) are more similar than the full regions compared with each other (full pathogenic CNVRs: mean 13.9 Mb; median 8.3 Mb), the pathogenic regions still contain more genes on average. Thus we corrected for gene number and found a mean of 33.7% developmental genes when we considered pathogenic peak CNVRs that were not overlapped by benign CNVRs compared to 24.5% of benign CNV genes (medians 20.0% and 0% respectively), a highly significant difference \( (P < 1.0 \times 10^{-16}, \text{Mann-Whitney U test}) \).

Clustering of developmental genes may contribute to their pathogenicity (Andrews et al., 2015). There is a significant enrichment of the proportion of de-
Figure 3.4 | Percentage of developmental genes overlapped by each randomised CNV set
Pathogenic CNV positions were shuffled randomly 1000 times, after each the percentage of developmental genes overlapped was calculated. The observed value for dbVar pathogenic CNVs is overlaid as a black line. Bin width set to 0.01.

Developmental genes in pathogenic peak regions compared to pathogenic regions outside of peaks (pathogenic peak regions: 24.5% of 3,452 genes exclusive to these regions; remaining pathogenic regions: 18.1% of 15,978 genes; $P < 1.0 \times 10^{-16}, \chi^2$ test). We confirmed that this is not due to clustering of developmental genes in general in the genome because upon randomising CNV location as above, the proportion of developmental genes covered by CNVs was consistently lower than observed in simulation (Supplementary Fig. 3.4). That the peak regions are enriched for developmental genes with respect to the remainder of the pathogenic CNVR is strong evidence that developmental genes are consistently implicated in CNV-related disease phenotypes across different CNVs in the genome.
3.3.3 Class P genes display features of dosage sensitive genes

A particular region of genome can be overlapped by multiple CNVs and these may differ in both their type (gain or loss) or their clinical interpretation (benign or pathogenic). We consider genes that are in CNVs with opposite clinical interpretations unlikely to be causative, particularly when the CNV is of the same type. That is, a gene found in a benign CNV gain region and a pathogenic CNV gain region is unlikely to be driving the pathogenic phenotype (Figure 3.5).

We refined the gene lists by considering all CNVs simultaneously. Fig. 3.6a shows counts of genes according to CNV type and clinical interpretation. For benign CNVs we considered the full CNV, whereas for pathogenic CNVs we considered only peak regions as above. We identify 6367 genes found only in pathogenic CNVs (shaded blue in Fig. 3.6b), we label these “Class P genes”. By contrast, 524 genes were inconsistent, reported in both benign and pathogenic CNVs, and are
Figure 3.6 | Patterns of gene duplication and loss across mammals for orthologues of human genes in CNVs.
(Continued on the following page.)
Dosage sensitivity underlies human CNV pathogenicity

Figure 3.6 | Patterns of gene duplication and loss across mammals for orthologues of human genes in CNVs.
(a) Venn diagram showing the number of protein-coding genes overlapped by different combinations of CNV types (blue, benign CNGs; yellow, benign CNLs; green, pathogenic gain peak coverage regions; red, pathogenic loss peak regions). (b) Genes that are covered exclusively by benign CNVs are labelled as ‘Class B’ (shaded red), those exclusive to pathogenic CNVs as ‘Class P’ (shaded blue) and those falling in CNVs with both clinical interpretations for gain or loss are considered as likely to be passenger genes and labelled ‘Class x’ (shaded grey). It is noteworthy that the classification refers to the CNVs that the genes fall within rather than the genes themselves. (c) Phylogenetic tree of 13 mammalian species used for gene conservation analysis and examples of human genes from each CNV overlap pattern type (Venn diagram segment) showing the orthologue distribution in the mammals. A dash indicates no change. (d) Box plot of the number of mammalian species where copy number is unchanged (black), duplication has occurred (green) and no orthologues (orange) for different categories of CNV overlap, as indicated below the boxplots. Upper and lower hinges of boxes correspond to the first and third quartiles. The median is shown within each box. Whiskers extend to values 1.5 interquartile range. These data were calculated per gene as illustrated in c. The sample size is shown below each boxplot.

deemed to be benign “passengers” (shaded grey in Fig. 3.6b; referred to as “Class X”). We refer to the 1075 genes that were consistently found in benign CNVs (shaded red in Fig. 3.6b) as Class B genes.

We tested these groups for enrichment of developmental genes, haploinsufficiency (Lek et al., 2016), protein complex members (Uniprot complex subunits (The UniProt Consortium, 2015)), ohnologues, and high gene expression (GTEx Consortium, 2015) (Table 3.2), all being features of genes previously associated with dosage sensitivity. We found Class B to be depleted for involvement in development and we found the opposite for Class P genes (14.3% vs. 25.2%, $P = 2.5 \times 10^{-11}$, $\chi^2$ test). Using the probability of loss-of-function mutation intolerance as a proxy for probability of haploinsufficiency (Lek et al., 2016), we found Class B genes to be less likely to be haploinsufficient compared to Class P genes (median probability of loss-of-function intolerance: 0.014 vs 0.028, $P = 0.005$, Mann-Whitney U test). Additionally, considering only the subset of genes with
## Dosage sensitivity underlies human CNV pathogenicity

Table 3.2 | Genes included in different types of CNV have different genetic and functional characteristics

<table>
<thead>
<tr>
<th>Class</th>
<th>Developmental genes</th>
<th>Protein complex members</th>
<th>Ohnologues</th>
<th>Haploinsufficient genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>14.3% (1,075)</td>
<td>23.4% (251)</td>
<td>26.9% (280)</td>
<td>12.0% (104)</td>
</tr>
<tr>
<td></td>
<td>25.2% (1,606)</td>
<td>33.9% (2,156)</td>
<td>37.6% (2,395)</td>
<td>19.5% (1,138)</td>
</tr>
<tr>
<td>P</td>
<td>22.4% (523)</td>
<td>28.7% (150)</td>
<td>30.4% (159)</td>
<td>13.9% (63)</td>
</tr>
<tr>
<td></td>
<td>20.2% (1,075)</td>
<td>33.0% (117)</td>
<td>29.8% (28)</td>
<td>13.4% (11)</td>
</tr>
<tr>
<td>X</td>
<td>20.2% (1,075)</td>
<td>33.0% (117)</td>
<td>29.8% (28)</td>
<td>13.4% (11)</td>
</tr>
<tr>
<td>BL/PG</td>
<td>25.5% (6,367)</td>
<td>35.5% (209)</td>
<td>41.8% (46)</td>
<td>14.9% (15)</td>
</tr>
<tr>
<td>BG/PL</td>
<td>25.5% (28)</td>
<td>33.0% (31)</td>
<td>41.8% (46)</td>
<td>14.9% (15)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P-value (x² test)</th>
<th>P-value (Mann-Whitney U test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developmental</td>
<td>2.7 × 10⁻¹¹</td>
<td></td>
</tr>
<tr>
<td>Protein complex</td>
<td>7.5 × 10⁻⁹</td>
<td></td>
</tr>
<tr>
<td>Ohnologues</td>
<td>4.1 × 10⁻¹⁰</td>
<td></td>
</tr>
<tr>
<td>Haploinsufficient genes</td>
<td>4.3 × 10⁻⁶</td>
<td></td>
</tr>
</tbody>
</table>

1 Genes exclusively observed in benign CNVRs
2 Gene exclusively observed in pathogenic CNVRs
3 Genes observed in contradictory CNV types and clinical interpretations
4 BL/PG - genes exclusively overlapped by benign loss CNVRs & pathogenic gain peak CNVRs
5 BG/PL - genes exclusively overlapped by benign gain CNVRs & pathogenic loss peak CNVRs
6 All p-values are Bonferroni corrected. Values in bold have adjusted residuals greater than ± 2 in the x² test.
7 Pairwise comparisons are indicated with dots. All p-values are Bonferroni corrected
8 Genes with probability of loss-of-function mutation intolerance >90% inferred in ref (Lek et al., 2016)
9 Probability of loss-of-function mutation intolerance inferred in ref (Lek et al., 2016)

High haploinsufficiency scores (3,230 genes with probability of loss-of-function intolerance >90%), we find Class P genes enriched (19.5%) relative to Class B and Class X genes (12.0% and 13.9%, respectively, $P = 4.3 \times 10^{-6}$, $x^2$ test).

Protein complex members are expected to have constrained relative dosages (Veitia, 2010; Veitia and Birchler, 2010). While only 23.4% of Class B genes have products functioning as subunits in protein complexes, 33.9% of Class P genes are in complexes, a significant difference ($P = 7.6 \times 10^{-9}$, $x^2$ test). Additionally, ohnologues (paralogues generated by whole genome duplication) are over-represented in Class P genes (37.6% vs. 26.9%, $P = 3.9 \times 10^{-10}$, $x^2$ test), consistent with previous observations that ohnologues are frequently associated with disease (OMIM classification) (Makino and McLysaght, 2010). There is evidence that highly expressed genes are not only strongly constrained with respect to sequence evolution (Sharp,
Dosage sensitivity underlies human CNV pathogenicity

but also have greater dosage constraint (Gout et al., 2010). Consistent with this we found that Class P genes have higher expression than Class B genes (medians: 19.6 RPKM and 9.6 RPKM respectively, \( P < 1.0 \times 10^{-16} \), Mann-Whitney U test). Furthermore, Class P genes are more highly expressed than Class X genes (medians: 19.6 RPKM vs 12.6, \( P = 4.7 \times 10^{-13} \), Mann-Whitney U test). The trends observed here are consistent with the notion that genes in pathogenic CNVs are dosage-sensitive.

3.3.4 Solitary Class P genes are enriched for neurodevelopment

When we consider the genomic distribution of the Class P genes we observe that seven of 390 CNVRs (178 pathogenic CNG regions and 212 pathogenic CNL regions) do not contain any genes exclusive to pathogenic CNVRs. In these cases, pathogenicity may be due to genes of reduced penetrance, position effects of the CNV, or a different type of dosage sensitivity (for example, if the gene is haploinsufficient, or conversely if the gene is aggregation prone at higher concentration, then these genes could be in both pathogenic loss and benign gain CNVs or vice versa, respectively, and would not be designated as exclusively pathogenic by us).

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Chromosome</th>
<th>Gene Start (bp)</th>
<th>Gene End (bp)</th>
<th>Strand</th>
<th>Associated Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000067606</td>
<td>1</td>
<td>1981909</td>
<td>2116834</td>
<td>1</td>
<td>PRKCZ</td>
</tr>
<tr>
<td>ENSG00000215912</td>
<td>1</td>
<td>2567415</td>
<td>2718286</td>
<td>-1</td>
<td>TTC34</td>
</tr>
<tr>
<td>ENSG00000142611</td>
<td>1</td>
<td>2985732</td>
<td>3355185</td>
<td>1</td>
<td>PRDM16</td>
</tr>
<tr>
<td>ENSG00000130762</td>
<td>1</td>
<td>3370990</td>
<td>3397677</td>
<td>1</td>
<td>ARHGEF16</td>
</tr>
<tr>
<td>ENSG00000116288</td>
<td>1</td>
<td>8014351</td>
<td>8045565</td>
<td>1</td>
<td>PARK7</td>
</tr>
<tr>
<td>ENSG00000116731</td>
<td>1</td>
<td>14026693</td>
<td>14151574</td>
<td>1</td>
<td>PRDM2</td>
</tr>
<tr>
<td>ENSG00000117154</td>
<td>1</td>
<td>18434240</td>
<td>18704977</td>
<td>1</td>
<td>IGSF21</td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Start Position</th>
<th>End Position</th>
<th>Chromosome</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000117425</td>
<td>45285516</td>
<td>45308735</td>
<td>-1</td>
<td>PTCH2</td>
</tr>
<tr>
<td>ENSG00000162599</td>
<td>45308735</td>
<td>45412895</td>
<td>1</td>
<td>NFIA</td>
</tr>
<tr>
<td>ENSG00000184005</td>
<td>45412895</td>
<td>45528893</td>
<td>1</td>
<td>ST6GALNAC3</td>
</tr>
<tr>
<td>ENSG0000018641</td>
<td>45528893</td>
<td>45604375</td>
<td>-1</td>
<td>DPYD</td>
</tr>
<tr>
<td>ENSG0000060718</td>
<td>45604375</td>
<td>45706952</td>
<td>-1</td>
<td>COL1A1</td>
</tr>
<tr>
<td>ENSG00000174827</td>
<td>45706952</td>
<td>45827103</td>
<td>-1</td>
<td>GPR89A</td>
</tr>
<tr>
<td>ENSG00000117262</td>
<td>45827103</td>
<td>46032647</td>
<td>-1</td>
<td>NBPFI1</td>
</tr>
<tr>
<td>ENSG00000188092</td>
<td>46032647</td>
<td>46476575</td>
<td>1</td>
<td>GPR89B</td>
</tr>
<tr>
<td>ENSG00000162687</td>
<td>46476575</td>
<td>46928355</td>
<td>-1</td>
<td>KCNT2</td>
</tr>
<tr>
<td>ENSG00000080910</td>
<td>46928355</td>
<td>47096925</td>
<td>1</td>
<td>CFHR2</td>
</tr>
<tr>
<td>ENSG00000066279</td>
<td>47096925</td>
<td>47211582</td>
<td>-1</td>
<td>ASPM</td>
</tr>
<tr>
<td>ENSG00000081237</td>
<td>47211582</td>
<td>47387654</td>
<td>1</td>
<td>PTPRC</td>
</tr>
<tr>
<td>ENSG00000092978</td>
<td>47387654</td>
<td>47584424</td>
<td>-1</td>
<td>GAPATCH2</td>
</tr>
<tr>
<td>ENSG00000143507</td>
<td>47584424</td>
<td>47742568</td>
<td>-1</td>
<td>DUSP10</td>
</tr>
<tr>
<td>ENSG00000077585</td>
<td>47742568</td>
<td>47948516</td>
<td>1</td>
<td>GPR137B</td>
</tr>
<tr>
<td>ENSG00000198626</td>
<td>47948516</td>
<td>48279288</td>
<td>1</td>
<td>RYR2</td>
</tr>
<tr>
<td>ENSG00000133019</td>
<td>48279288</td>
<td>48978750</td>
<td>1</td>
<td>CHRM3</td>
</tr>
<tr>
<td>ENSG00000182901</td>
<td>48978750</td>
<td>49250530</td>
<td>-1</td>
<td>RGS7</td>
</tr>
<tr>
<td>ENSG00000117020</td>
<td>49250530</td>
<td>49641381</td>
<td>-1</td>
<td>AKT3</td>
</tr>
<tr>
<td>ENSG00000162849</td>
<td>49641381</td>
<td>49872873</td>
<td>1</td>
<td>KIF26B</td>
</tr>
<tr>
<td>ENSG00000185420</td>
<td>49872873</td>
<td>50076614</td>
<td>-1</td>
<td>SMYD3</td>
</tr>
<tr>
<td>ENSG00000134324</td>
<td>50076614</td>
<td>50377421</td>
<td>1</td>
<td>LPIN1</td>
</tr>
<tr>
<td>ENSG00000119888</td>
<td>50377421</td>
<td>50987470</td>
<td>1</td>
<td>EPCAM</td>
</tr>
<tr>
<td>ENSG00000095002</td>
<td>50987470</td>
<td>51589450</td>
<td>1</td>
<td>MSH2</td>
</tr>
<tr>
<td>ENSG00000179915</td>
<td>51589450</td>
<td>52159674</td>
<td>-1</td>
<td>NRXN1</td>
</tr>
<tr>
<td>ENSG00000082898</td>
<td>52159674</td>
<td>53076761</td>
<td>-1</td>
<td>XPO1</td>
</tr>
<tr>
<td>ENSG00000168702</td>
<td>53076761</td>
<td>53589927</td>
<td>-1</td>
<td>LRP1B</td>
</tr>
<tr>
<td>ENSG00000169554</td>
<td>53589927</td>
<td>54421247</td>
<td>-1</td>
<td>ZEB2</td>
</tr>
<tr>
<td>ENSG00000121989</td>
<td>54421247</td>
<td>55258247</td>
<td>-1</td>
<td>ACVR2A</td>
</tr>
<tr>
<td>ENSG00000160406</td>
<td>55258247</td>
<td>56127580</td>
<td>1</td>
<td>MBD5</td>
</tr>
<tr>
<td>ENSG00000168280</td>
<td>56127580</td>
<td>56562819</td>
<td>1</td>
<td>KIF5C</td>
</tr>
<tr>
<td>ENSG00000144285</td>
<td>56562819</td>
<td>56984523</td>
<td>-1</td>
<td>SCN1A</td>
</tr>
<tr>
<td>ENSG00000168542</td>
<td>56984523</td>
<td>57587747</td>
<td>1</td>
<td>COL3A1</td>
</tr>
<tr>
<td>ENSG00000064933</td>
<td>57587747</td>
<td>59324355</td>
<td>1</td>
<td>PMS1</td>
</tr>
<tr>
<td>ENSG00000115896</td>
<td>59324355</td>
<td>60347305</td>
<td>1</td>
<td>PLC1L1</td>
</tr>
<tr>
<td>ENSG00000119042</td>
<td>60347305</td>
<td>60635989</td>
<td>-1</td>
<td>SATB2</td>
</tr>
<tr>
<td>ENSG00000116117</td>
<td>60635989</td>
<td>61648866</td>
<td>1</td>
<td>PARD3B</td>
</tr>
<tr>
<td>ENSG00000116106</td>
<td>61648866</td>
<td>62243892</td>
<td>-1</td>
<td>EPHA4</td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Chromosome</th>
<th>Start Position</th>
<th>End Position</th>
<th>Genome Position</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000153820</td>
<td>2</td>
<td>228844666</td>
<td>229046361</td>
<td>-1</td>
<td>SPHKAP</td>
</tr>
<tr>
<td>ENSG00000134121</td>
<td>3</td>
<td>238279</td>
<td>451090</td>
<td>1</td>
<td>CHL1</td>
</tr>
<tr>
<td>ENSG00000196277</td>
<td>3</td>
<td>6811688</td>
<td>7783215</td>
<td>1</td>
<td>GRM7</td>
</tr>
<tr>
<td>ENSG00000168016</td>
<td>3</td>
<td>3686311</td>
<td>36986548</td>
<td>-1</td>
<td>TRANK1</td>
</tr>
<tr>
<td>ENSG00000088538</td>
<td>3</td>
<td>50712672</td>
<td>51421629</td>
<td>1</td>
<td>DOCK3</td>
</tr>
<tr>
<td>ENSG00000183662</td>
<td>3</td>
<td>68053359</td>
<td>68594776</td>
<td>1</td>
<td>FAMIL9A1</td>
</tr>
<tr>
<td>ENSG00000114861</td>
<td>3</td>
<td>71003844</td>
<td>71633140</td>
<td>-1</td>
<td>FOXP1</td>
</tr>
<tr>
<td>ENSG00000175161</td>
<td>3</td>
<td>85008132</td>
<td>86123579</td>
<td>1</td>
<td>CADM2</td>
</tr>
<tr>
<td>ENSG00000183770</td>
<td>3</td>
<td>181429714</td>
<td>181432221</td>
<td>1</td>
<td>SOX2</td>
</tr>
<tr>
<td>ENSG00000145012</td>
<td>3</td>
<td>187871072</td>
<td>188608460</td>
<td>1</td>
<td>LPP</td>
</tr>
<tr>
<td>ENSG00000138670</td>
<td>4</td>
<td>82347547</td>
<td>82965397</td>
<td>-1</td>
<td>RASGEF1B</td>
</tr>
<tr>
<td>ENSG00000152208</td>
<td>4</td>
<td>93225550</td>
<td>94695707</td>
<td>1</td>
<td>GRID2</td>
</tr>
<tr>
<td>ENSG00000196159</td>
<td>4</td>
<td>126237554</td>
<td>126414087</td>
<td>1</td>
<td>FAT4</td>
</tr>
<tr>
<td>ENSG00000151623</td>
<td>4</td>
<td>148999913</td>
<td>149365850</td>
<td>-1</td>
<td>NR3C2</td>
</tr>
<tr>
<td>ENSG00000198589</td>
<td>4</td>
<td>151185594</td>
<td>151936879</td>
<td>-1</td>
<td>LRBA</td>
</tr>
<tr>
<td>ENSG00000171560</td>
<td>4</td>
<td>155504278</td>
<td>155511918</td>
<td>-1</td>
<td>FGA</td>
</tr>
<tr>
<td>ENSG00000174473</td>
<td>4</td>
<td>172733405</td>
<td>173962710</td>
<td>1</td>
<td>GALNTL6</td>
</tr>
<tr>
<td>ENSG00000151718</td>
<td>4</td>
<td>184020446</td>
<td>184241930</td>
<td>1</td>
<td>WWCT2</td>
</tr>
<tr>
<td>ENSG00000164342</td>
<td>4</td>
<td>186990306</td>
<td>187009223</td>
<td>1</td>
<td>TLR3</td>
</tr>
<tr>
<td>ENSG00000170561</td>
<td>5</td>
<td>2745959</td>
<td>2752969</td>
<td>-1</td>
<td>IRX2</td>
</tr>
<tr>
<td>ENSG00000170549</td>
<td>5</td>
<td>3596168</td>
<td>3601517</td>
<td>1</td>
<td>IRX1</td>
</tr>
<tr>
<td>ENSG00000154162</td>
<td>5</td>
<td>21750782</td>
<td>22853731</td>
<td>-1</td>
<td>CDH12</td>
</tr>
<tr>
<td>ENSG00000113100</td>
<td>5</td>
<td>26880709</td>
<td>27121257</td>
<td>-1</td>
<td>CDH9</td>
</tr>
<tr>
<td>ENSG00000164190</td>
<td>5</td>
<td>36876861</td>
<td>37066515</td>
<td>1</td>
<td>NIPBL</td>
</tr>
<tr>
<td>ENSG00000049860</td>
<td>5</td>
<td>73935848</td>
<td>74018472</td>
<td>1</td>
<td>HEXB</td>
</tr>
<tr>
<td>ENSG00000081189</td>
<td>5</td>
<td>88013975</td>
<td>88199922</td>
<td>-1</td>
<td>MEF2C</td>
</tr>
<tr>
<td>ENSG00000155324</td>
<td>5</td>
<td>125695824</td>
<td>125832186</td>
<td>1</td>
<td>GRAMD3</td>
</tr>
<tr>
<td>ENSG00000138829</td>
<td>5</td>
<td>127593601</td>
<td>127999878</td>
<td>-1</td>
<td>FBN2</td>
</tr>
<tr>
<td>ENSG00000186314</td>
<td>5</td>
<td>144851362</td>
<td>145214932</td>
<td>-1</td>
<td>PRELID2</td>
</tr>
<tr>
<td>ENSG00000070814</td>
<td>5</td>
<td>149737202</td>
<td>149779871</td>
<td>1</td>
<td>TCOF1</td>
</tr>
<tr>
<td>ENSG00000113327</td>
<td>5</td>
<td>161494546</td>
<td>161582542</td>
<td>1</td>
<td>GABRG2</td>
</tr>
<tr>
<td>ENSG00000120149</td>
<td>5</td>
<td>174151536</td>
<td>174157896</td>
<td>1</td>
<td>MSX2</td>
</tr>
<tr>
<td>ENSG00000165671</td>
<td>5</td>
<td>176560026</td>
<td>176727216</td>
<td>1</td>
<td>NSD1</td>
</tr>
<tr>
<td>ENSG00000054598</td>
<td>6</td>
<td>1610681</td>
<td>1614127</td>
<td>1</td>
<td>FOXC1</td>
</tr>
<tr>
<td>ENSG00000153046</td>
<td>6</td>
<td>4706393</td>
<td>4955785</td>
<td>1</td>
<td>CDYL</td>
</tr>
<tr>
<td>ENSG00000145979</td>
<td>6</td>
<td>13266774</td>
<td>13328815</td>
<td>-1</td>
<td>TBC1D7</td>
</tr>
<tr>
<td>ENSG00000124813</td>
<td>6</td>
<td>45295894</td>
<td>45632086</td>
<td>1</td>
<td>RUNX2</td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Sign</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000146085</td>
<td>6</td>
<td>49398073</td>
<td>49430904</td>
<td>-1</td>
<td>MUT</td>
</tr>
<tr>
<td>ENSG00000079841</td>
<td>6</td>
<td>72596406</td>
<td>73112845</td>
<td>1</td>
<td>RIMS1</td>
</tr>
<tr>
<td>ENSG00000185850</td>
<td>6</td>
<td>124125286</td>
<td>125146803</td>
<td>1</td>
<td>NKAIN2</td>
</tr>
<tr>
<td>ENSG00000196569</td>
<td>6</td>
<td>129204342</td>
<td>129837714</td>
<td>1</td>
<td>LAMA2</td>
</tr>
<tr>
<td>ENSG00000049618</td>
<td>6</td>
<td>157099063</td>
<td>157531913</td>
<td>1</td>
<td>ARID1B</td>
</tr>
<tr>
<td>ENSG00000185345</td>
<td>6</td>
<td>161768452</td>
<td>163148803</td>
<td>-1</td>
<td>PARK2</td>
</tr>
<tr>
<td>ENSG00000112530</td>
<td>6</td>
<td>163148164</td>
<td>163736524</td>
<td>1</td>
<td>PACRG</td>
</tr>
<tr>
<td>ENSG00000112531</td>
<td>6</td>
<td>163835032</td>
<td>163996928</td>
<td>1</td>
<td>QKI</td>
</tr>
<tr>
<td>ENSG00000182095</td>
<td>7</td>
<td>5346421</td>
<td>5465045</td>
<td>-1</td>
<td>TNRC18</td>
</tr>
<tr>
<td>ENSG00000155034</td>
<td>7</td>
<td>5470966</td>
<td>5553429</td>
<td>-1</td>
<td>FBXL18</td>
</tr>
<tr>
<td>ENSG00000175600</td>
<td>7</td>
<td>40174575</td>
<td>40900362</td>
<td>1</td>
<td>SUGCT</td>
</tr>
<tr>
<td>ENSG00000106571</td>
<td>7</td>
<td>42000548</td>
<td>42277469</td>
<td>-1</td>
<td>GLI3</td>
</tr>
<tr>
<td>ENSG00000158321</td>
<td>7</td>
<td>69063905</td>
<td>70258054</td>
<td>1</td>
<td>AUTS2</td>
</tr>
<tr>
<td>ENSG00000099550</td>
<td>7</td>
<td>73007524</td>
<td>73038873</td>
<td>-1</td>
<td>MLXIPL</td>
</tr>
<tr>
<td>ENSG00000164692</td>
<td>7</td>
<td>94023873</td>
<td>94060544</td>
<td>1</td>
<td>COL1A2</td>
</tr>
<tr>
<td>ENSG00000158528</td>
<td>7</td>
<td>94536514</td>
<td>94925727</td>
<td>1</td>
<td>PPP1R9A</td>
</tr>
<tr>
<td>ENSG00000016266</td>
<td>7</td>
<td>117105838</td>
<td>117356025</td>
<td>1</td>
<td>CFTR</td>
</tr>
<tr>
<td>ENSG00000106025</td>
<td>7</td>
<td>120427376</td>
<td>120498456</td>
<td>-1</td>
<td>TSPAN12</td>
</tr>
<tr>
<td>ENSG00000179603</td>
<td>7</td>
<td>126078652</td>
<td>126893348</td>
<td>1</td>
<td>GRM8</td>
</tr>
<tr>
<td>ENSG00000174469</td>
<td>7</td>
<td>145813453</td>
<td>148118090</td>
<td>1</td>
<td>CNTNAP2</td>
</tr>
<tr>
<td>ENSG00000130675</td>
<td>7</td>
<td>156786745</td>
<td>156803345</td>
<td>-1</td>
<td>MNX1</td>
</tr>
<tr>
<td>ENSG00000183117</td>
<td>8</td>
<td>2792875</td>
<td>4852494</td>
<td>-1</td>
<td>CSMD1</td>
</tr>
<tr>
<td>ENSG00000147316</td>
<td>8</td>
<td>6264113</td>
<td>6501144</td>
<td>1</td>
<td>MCPH1</td>
</tr>
<tr>
<td>ENSG00000175445</td>
<td>8</td>
<td>19759228</td>
<td>19824769</td>
<td>1</td>
<td>LPL</td>
</tr>
<tr>
<td>ENSG00000029534</td>
<td>8</td>
<td>41510739</td>
<td>41754280</td>
<td>-1</td>
<td>ANK1</td>
</tr>
<tr>
<td>ENSG00000104331</td>
<td>8</td>
<td>57870492</td>
<td>57906403</td>
<td>-1</td>
<td>IMPAD1</td>
</tr>
<tr>
<td>ENSG00000171316</td>
<td>8</td>
<td>61591337</td>
<td>61779465</td>
<td>1</td>
<td>CHD7</td>
</tr>
<tr>
<td>ENSG00000175073</td>
<td>8</td>
<td>67540722</td>
<td>6759452</td>
<td>-1</td>
<td>VCPIP1</td>
</tr>
<tr>
<td>ENSG00000104447</td>
<td>8</td>
<td>116420724</td>
<td>116821899</td>
<td>-1</td>
<td>TRPS1</td>
</tr>
<tr>
<td>ENSG00000178685</td>
<td>8</td>
<td>145051321</td>
<td>145086940</td>
<td>-1</td>
<td>PARP10</td>
</tr>
<tr>
<td>ENSG00000107099</td>
<td>9</td>
<td>214854</td>
<td>465259</td>
<td>1</td>
<td>DOCK8</td>
</tr>
<tr>
<td>ENSG00000107104</td>
<td>9</td>
<td>470291</td>
<td>746105</td>
<td>1</td>
<td>KANK1</td>
</tr>
<tr>
<td>ENSG00000107249</td>
<td>9</td>
<td>3824127</td>
<td>4348392</td>
<td>-1</td>
<td>GLIS3</td>
</tr>
<tr>
<td>ENSG00000153707</td>
<td>9</td>
<td>8314246</td>
<td>10612723</td>
<td>-1</td>
<td>PTPRD</td>
</tr>
<tr>
<td>ENSG00000171843</td>
<td>9</td>
<td>20341663</td>
<td>20622542</td>
<td>-1</td>
<td>MLLT3</td>
</tr>
<tr>
<td>ENSG00000169071</td>
<td>9</td>
<td>94325373</td>
<td>94712444</td>
<td>-1</td>
<td>ROR2</td>
</tr>
<tr>
<td>ENSG00000185920</td>
<td>9</td>
<td>98205262</td>
<td>98279339</td>
<td>-1</td>
<td>PTCH1</td>
</tr>
<tr>
<td>ENSG00000214645</td>
<td>9</td>
<td>110539471</td>
<td>110540419</td>
<td>-1</td>
<td>AL162389.1</td>
</tr>
<tr>
<td>ENSG00000197070</td>
<td>9</td>
<td>140500106</td>
<td>140509812</td>
<td>1</td>
<td>ARRDC1</td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Chromosome Position</th>
<th>Pathogenicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000181090</td>
<td>9 140513444 140764468 1</td>
<td>EHMT1</td>
</tr>
<tr>
<td>ENSG00000150275</td>
<td>10 55562531 57387702 -1</td>
<td>PCDH15</td>
</tr>
<tr>
<td>ENSG00000183230</td>
<td>10 67672276 69455927 -1</td>
<td>CTNNA3</td>
</tr>
<tr>
<td>ENSG00000156110</td>
<td>10 75910960 76469061 1</td>
<td>ADK</td>
</tr>
<tr>
<td>ENSG00000107779</td>
<td>10 88516407 88692595 1</td>
<td>BMPR1A</td>
</tr>
<tr>
<td>ENSG00000068383</td>
<td>10 134351324 134596979 1</td>
<td>INPP5A</td>
</tr>
<tr>
<td>ENSG00000214026</td>
<td>11 1968508 2005752 1</td>
<td>MRPL23</td>
</tr>
<tr>
<td>ENSG00000109911</td>
<td>11 31531297 31805546 1</td>
<td>ELP4</td>
</tr>
<tr>
<td>ENSG0000007372</td>
<td>11 31806340 31839509 -1</td>
<td>PAX6</td>
</tr>
<tr>
<td>ENSG00000110090</td>
<td>11 68522088 68611878 -1</td>
<td>CPT1A</td>
</tr>
<tr>
<td>ENSG00000187240</td>
<td>11 102980160 103350591 1</td>
<td>DYNC2H1</td>
</tr>
<tr>
<td>ENSG00000149571</td>
<td>11 126293254 126873355 -1</td>
<td>KIRREL3</td>
</tr>
<tr>
<td>ENSG00000060237</td>
<td>12 861759 1020618 1</td>
<td>WNK1</td>
</tr>
<tr>
<td>ENSG00000151067</td>
<td>12 2079952 2802108 1</td>
<td>CACNA1C</td>
</tr>
<tr>
<td>ENSG00000110841</td>
<td>12 27676364 27848497 1</td>
<td>PPFIBP1</td>
</tr>
<tr>
<td>ENSG00000089225</td>
<td>12 114791736 114846247 -1</td>
<td>TBX5</td>
</tr>
<tr>
<td>ENSG00000123066</td>
<td>12 116395711 116715143 -1</td>
<td>MED13L</td>
</tr>
<tr>
<td>ENSG00000102452</td>
<td>13 101706130 102068843 -1</td>
<td>NALCN</td>
</tr>
<tr>
<td>ENSG00000100888</td>
<td>14 21853353 21924285 -1</td>
<td>CHD8</td>
</tr>
<tr>
<td>ENSG00000092054</td>
<td>14 23881947 23904927 -1</td>
<td>MYH7</td>
</tr>
<tr>
<td>ENSG00000139865</td>
<td>14 38065052 38510647 1</td>
<td>TTC6</td>
</tr>
<tr>
<td>ENSG00000100592</td>
<td>14 59655364 59838123 1</td>
<td>DAAM1</td>
</tr>
<tr>
<td>ENSG00000021645</td>
<td>14 78708734 80330762 1</td>
<td>NRXN3</td>
</tr>
<tr>
<td>ENSG00000182979</td>
<td>14 105886159 105937066 1</td>
<td>MTA1</td>
</tr>
<tr>
<td>ENSG00000128739</td>
<td>15 25068794 25223870 1</td>
<td>SNRPN</td>
</tr>
<tr>
<td>ENSG00000114062</td>
<td>15 25582381 25684128 -1</td>
<td>UBE3A</td>
</tr>
<tr>
<td>ENSG00000104044</td>
<td>15 28000021 28344504 -1</td>
<td>OCA2</td>
</tr>
<tr>
<td>ENSG00000128731</td>
<td>15 28356186 28567298 -1</td>
<td>HERC2</td>
</tr>
<tr>
<td>ENSG00000166664</td>
<td>15 30653443 30686052 -1</td>
<td>CHRFAM7A</td>
</tr>
<tr>
<td>ENSG00000187951</td>
<td>15 30916697 31065196 1</td>
<td>ARHGAP11B</td>
</tr>
<tr>
<td>ENSG00000169918</td>
<td>15 31775329 32162992 -1</td>
<td>OTUD7A</td>
</tr>
<tr>
<td>ENSG00000166147</td>
<td>15 48700503 48938046 -1</td>
<td>FBN1</td>
</tr>
<tr>
<td>ENSG00000213614</td>
<td>15 72635775 72668817 -1</td>
<td>HEXA</td>
</tr>
<tr>
<td>ENSG00000169371</td>
<td>15 75890424 75918810 -1</td>
<td>SNUPN</td>
</tr>
<tr>
<td>ENSG00000169752</td>
<td>15 76228310 76352136 -1</td>
<td>NRG4</td>
</tr>
<tr>
<td>ENSG00000269360</td>
<td>15 93749295 93751277 1</td>
<td>AC112693.2</td>
</tr>
<tr>
<td>ENSG00000140443</td>
<td>15 99192200 99307759 1</td>
<td>IGFR1</td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Chromosome</th>
<th>Start Position</th>
<th>End Position</th>
<th>Strand</th>
<th>Gene Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000000168904</td>
<td>15</td>
<td>99791567</td>
<td>99930934</td>
<td>1</td>
<td>LRRC28</td>
</tr>
<tr>
<td>ENSG00000001885536</td>
<td>16</td>
<td>222846</td>
<td>223709</td>
<td>1</td>
<td>HBA2</td>
</tr>
<tr>
<td>ENSG0000000086506</td>
<td>16</td>
<td>230452</td>
<td>231180</td>
<td>1</td>
<td>HBQ1</td>
</tr>
<tr>
<td>ENSG0000000053539</td>
<td>16</td>
<td>3775055</td>
<td>3930727</td>
<td>-1</td>
<td>CREBBP</td>
</tr>
<tr>
<td>ENSG0000000078328</td>
<td>16</td>
<td>6069095</td>
<td>7763340</td>
<td>1</td>
<td>RBBFOX1</td>
</tr>
<tr>
<td>ENSG0000000140743</td>
<td>16</td>
<td>22357257</td>
<td>22448486</td>
<td>-1</td>
<td>CDR2</td>
</tr>
<tr>
<td>ENSG0000000140945</td>
<td>16</td>
<td>82660408</td>
<td>83830204</td>
<td>1</td>
<td>CDH13</td>
</tr>
<tr>
<td>ENSG0000000051523</td>
<td>16</td>
<td>88709691</td>
<td>88717560</td>
<td>-1</td>
<td>CYBA</td>
</tr>
<tr>
<td>ENSG0000000167693</td>
<td>17</td>
<td>702553</td>
<td>883010</td>
<td>-1</td>
<td>NXN</td>
</tr>
<tr>
<td>ENSG0000000108953</td>
<td>17</td>
<td>1247566</td>
<td>1303672</td>
<td>-1</td>
<td>YWHAE</td>
</tr>
<tr>
<td>ENSG0000000070366</td>
<td>17</td>
<td>1963133</td>
<td>2207065</td>
<td>-1</td>
<td>SMG6</td>
</tr>
<tr>
<td>ENSG0000000127804</td>
<td>17</td>
<td>2308856</td>
<td>2415185</td>
<td>-1</td>
<td>METTL16</td>
</tr>
<tr>
<td>ENSG000000007168</td>
<td>17</td>
<td>2496504</td>
<td>2588909</td>
<td>1</td>
<td>PAFAH1B1</td>
</tr>
<tr>
<td>ENSG0000000170425</td>
<td>17</td>
<td>15848231</td>
<td>15879060</td>
<td>1</td>
<td>ADORA2B</td>
</tr>
<tr>
<td>ENSG0000000205309</td>
<td>17</td>
<td>17206644</td>
<td>17250977</td>
<td>1</td>
<td>NT5M</td>
</tr>
<tr>
<td>ENSG0000000108557</td>
<td>17</td>
<td>17584787</td>
<td>17714767</td>
<td>1</td>
<td>RAI1</td>
</tr>
<tr>
<td>ENSG0000000196712</td>
<td>17</td>
<td>29421945</td>
<td>29709134</td>
<td>1</td>
<td>NF1</td>
</tr>
<tr>
<td>ENSG0000000214226</td>
<td>17</td>
<td>54869274</td>
<td>54916134</td>
<td>-1</td>
<td>CI7orf67</td>
</tr>
<tr>
<td>ENSG0000000154217</td>
<td>17</td>
<td>65373575</td>
<td>65693372</td>
<td>1</td>
<td>PITPNC1</td>
</tr>
<tr>
<td>ENSG0000000161533</td>
<td>17</td>
<td>73937588</td>
<td>73975515</td>
<td>-1</td>
<td>ACOX1</td>
</tr>
<tr>
<td>ENSG0000000196628</td>
<td>18</td>
<td>52889562</td>
<td>53332018</td>
<td>-1</td>
<td>TCF4</td>
</tr>
<tr>
<td>ENSG0000000130158</td>
<td>19</td>
<td>11309971</td>
<td>11373157</td>
<td>-1</td>
<td>DOCK6</td>
</tr>
<tr>
<td>ENSG0000000141837</td>
<td>19</td>
<td>13317256</td>
<td>13734804</td>
<td>-1</td>
<td>CACNA1A</td>
</tr>
<tr>
<td>ENSG0000000072071</td>
<td>19</td>
<td>14260750</td>
<td>14316999</td>
<td>-1</td>
<td>LPHN1</td>
</tr>
<tr>
<td>ENSG0000000131848</td>
<td>19</td>
<td>56732681</td>
<td>56879752</td>
<td>-1</td>
<td>ZSCAN5A</td>
</tr>
<tr>
<td>ENSG0000000125845</td>
<td>20</td>
<td>6748311</td>
<td>6760927</td>
<td>1</td>
<td>BMP2</td>
</tr>
<tr>
<td>ENSG0000000161322</td>
<td>20</td>
<td>62783144</td>
<td>62873604</td>
<td>1</td>
<td>MYT1</td>
</tr>
<tr>
<td>ENSG0000000215193</td>
<td>22</td>
<td>18560689</td>
<td>18613905</td>
<td>1</td>
<td>PEX26</td>
</tr>
<tr>
<td>ENSG0000000184979</td>
<td>22</td>
<td>18632666</td>
<td>18660164</td>
<td>1</td>
<td>USP18</td>
</tr>
<tr>
<td>ENSG0000000128185</td>
<td>22</td>
<td>20301799</td>
<td>20307603</td>
<td>-1</td>
<td>DGC6L</td>
</tr>
<tr>
<td>ENSG0000000161133</td>
<td>22</td>
<td>20704868</td>
<td>20745048</td>
<td>-1</td>
<td>USP41</td>
</tr>
<tr>
<td>ENSG0000000185651</td>
<td>22</td>
<td>21903736</td>
<td>21978323</td>
<td>1</td>
<td>UBE2L3</td>
</tr>
<tr>
<td>ENSG0000000186575</td>
<td>22</td>
<td>29999545</td>
<td>30094587</td>
<td>1</td>
<td>NF2</td>
</tr>
<tr>
<td>ENSG0000000133424</td>
<td>22</td>
<td>33558212</td>
<td>34318829</td>
<td>-1</td>
<td>LARGE</td>
</tr>
<tr>
<td>ENSG0000000100425</td>
<td>22</td>
<td>50166931</td>
<td>50221160</td>
<td>-1</td>
<td>BRD1</td>
</tr>
<tr>
<td>ENSG0000000251322</td>
<td>22</td>
<td>51112843</td>
<td>51171726</td>
<td>1</td>
<td>SHANK3</td>
</tr>
</tbody>
</table>

The remainder of CNV regions contain at least one protein-coding gene that is
Figure 3.7 | Number of Class P genes per copy number gain (CNG) peak region
Histogram of the number of Class P genes per CNG peak region. Bin width set to 1.

Figure 3.8 | Number of Class P genes per copy number loss (CNL) peak region
Histogram of the number of Class P genes per CNL peak region. Bin width set to 1.
never observed in a benign CNV. We found that 21/390 pathogenic CNVRs contain exactly one such Class P gene, and 300/999 pathogenic peak regions contain exactly one (199 unique genes out of 321 solitary genes intersecting with regions, Table 3.3, Figure 3.7 and 3.8). These latter cases have suggested pathogenicity by exclusion: they are only found in pathogenic CNVs and no other gene in the peak region is exclusively pathogenic. The observation that peak regions are enriched for solitary pathogenic genes suggests that analysing peak regions is a useful way to refine the analysis of CNVRs. Note that we consider sets of overlapping duplication CNVs separately from deletion CNVs when building these CNVRs and their peak regions. This allows for the possibility of the mechanistic basis of pathogenicity being different between duplication and deletion CNVs, though frequently it is the same gene that is the sole Class P gene: 122 out of 199 solitary Class P genes are the solitary Class P gene for a duplication and a deletion CNV peak region.

Thus, of the 4234 genes overlapped by pathogenic peak region CNVs, 199 are the solitary candidate pathogenic gene in the region. These are promising candidates for causing the pathogenicity of the CNV. With all pathogenic peak CNVR genes as a background list, we find that solitary candidate pathogenic genes are enriched for “anatomical structure development” (GO:0048856; \( P = 8.4 \times 10^{-12} \)), especially “embryonic morphogenesis” (GO:0048598; \( P = 1.0 \times 10^{-10} \)) and “neurogenesis” (GO:0022008; \( P = 3.1 \times 10^{-8} \)), “regulation of multicellular organismal process” (GO:0051239; \( P = 3.7 \times 10^{-9} \)), “adult behavior” (GO:0030534; \( P = 4.7 \times 10^{-9} \)), and “signaling” (GO:0023052; \( P = 3.4 \times 10^{-8} \)) (Table 3.4). These 199 genes are also significantly enriched for localization within axons and dendrites (cellular component term “neuron projection”, GO:0043005; \( P = 8.9 \times 10^{-6} \)) and are overrepresented for genes associated with an “abnormality of the nervous system” (HP:0000707; \( P = 4.7 \times 10^{-30} \)) in the Human Phenotype Ontology (Köhler et al.,
Table 3.4 | Most enriched functional classes of solitary Class P genes
Full table available in supplementary information of Rice and McLysaght (2017)

<table>
<thead>
<tr>
<th>Term ID</th>
<th>Term name</th>
<th>P-value (Bonferroni-adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0032501</td>
<td>multicellular organismal process</td>
<td>6.53 × 10^{-12}</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>anatomical structure development</td>
<td>8.37 × 10^{-12}</td>
</tr>
<tr>
<td>GO:0048731</td>
<td>system development</td>
<td>1.80 × 10^{-11}</td>
</tr>
<tr>
<td>GO:0007275</td>
<td>multicellular organismal development</td>
<td>9.33 × 10^{-11}</td>
</tr>
<tr>
<td>GO:0048598</td>
<td>embryonic morphogenesis</td>
<td>1.02 × 10^{-10}</td>
</tr>
<tr>
<td>GO:0044707</td>
<td>single-multicellular organism process</td>
<td>1.49 × 10^{-10}</td>
</tr>
<tr>
<td>GO:0009653</td>
<td>anatomical structure morphogenesis</td>
<td>4.77 × 10^{-10}</td>
</tr>
<tr>
<td>GO:0032502</td>
<td>developmental process</td>
<td>6.24 × 10^{-10}</td>
</tr>
<tr>
<td>GO:0044767</td>
<td>single-organism developmental process</td>
<td>1.01 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0048468</td>
<td>cell development</td>
<td>1.48 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0007399</td>
<td>nervous system development</td>
<td>1.61 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0051239</td>
<td>regulation of multicellular organismal process</td>
<td>3.69 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0009790</td>
<td>embryo development</td>
<td>3.75 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0030534</td>
<td>adult behavior</td>
<td>4.68 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0030154</td>
<td>cell differentiation</td>
<td>9.03 × 10^{-09}</td>
</tr>
<tr>
<td>GO:0009887</td>
<td>organ morphogenesis</td>
<td>1.99 × 10^{-08}</td>
</tr>
<tr>
<td>GO:0022008</td>
<td>neurogenesis</td>
<td>3.12 × 10^{-08}</td>
</tr>
<tr>
<td>GO:0023052</td>
<td>signaling</td>
<td>3.37 × 10^{-08}</td>
</tr>
<tr>
<td>GO:0044700</td>
<td>single organism signaling</td>
<td>3.37 × 10^{-08}</td>
</tr>
<tr>
<td>GO:0043005</td>
<td>neuron projection</td>
<td>8.85 × 10^{-06}</td>
</tr>
<tr>
<td>GO:0097458</td>
<td>neuron part</td>
<td>3.81 × 10^{-05}</td>
</tr>
<tr>
<td>GO:0044464</td>
<td>cell part</td>
<td>2.93 × 10^{-04}</td>
</tr>
<tr>
<td>GO:0005623</td>
<td>cell</td>
<td>2.96 × 10^{-04}</td>
</tr>
<tr>
<td>GO:0044463</td>
<td>cell projection part</td>
<td>4.61 × 10^{-04}</td>
</tr>
<tr>
<td>GO:0042995</td>
<td>cell projection</td>
<td>5.55 × 10^{-04}</td>
</tr>
<tr>
<td>GO:0030425</td>
<td>dendrite</td>
<td>1.66 × 10^{-03}</td>
</tr>
<tr>
<td>GO:0030424</td>
<td>axon</td>
<td>4.37 × 10^{-03}</td>
</tr>
<tr>
<td>HP:0000118</td>
<td>Phenotypic abnormality</td>
<td>4.58 × 10^{-31}</td>
</tr>
<tr>
<td>HP:0000707</td>
<td>Abnormality of the nervous system</td>
<td>4.73 × 10^{-30}</td>
</tr>
<tr>
<td>HP:0000005</td>
<td>Mode of inheritance</td>
<td>9.98 × 10^{-29}</td>
</tr>
<tr>
<td>HP:0011842</td>
<td>Abnormality of skeletal morphology</td>
<td>2.34 × 10^{-27}</td>
</tr>
<tr>
<td>HP:0009121</td>
<td>Abnormal axial skeleton morphology</td>
<td>5.81 × 10^{-27}</td>
</tr>
<tr>
<td>HP:0000924</td>
<td>Abnormality of the skeletal system</td>
<td>1.48 × 10^{-26}</td>
</tr>
<tr>
<td>HP:0012638</td>
<td>Abnormality of nervous system physiology</td>
<td>4.10 × 10^{-26}</td>
</tr>
<tr>
<td>HP:0000006</td>
<td>Autosomal dominant inheritance</td>
<td>1.23 × 10^{-25}</td>
</tr>
<tr>
<td>HP:0000929</td>
<td>Abnormality of the skull</td>
<td>4.26 × 10^{-25}</td>
</tr>
<tr>
<td>HP:0012759</td>
<td>Neurodevelopmental abnormality</td>
<td>1.11 × 10^{-24}</td>
</tr>
<tr>
<td>HP:0000234</td>
<td>Abnormality of the head</td>
<td>2.03 × 10^{-24}</td>
</tr>
<tr>
<td>HP:0000152</td>
<td>Abnormality of head and neck</td>
<td>5.04 × 10^{-24}</td>
</tr>
<tr>
<td>HP:0002011</td>
<td>Morphological abnormality of the central nervous system</td>
<td>1.36 × 10^{-23}</td>
</tr>
<tr>
<td>HP:0000271</td>
<td>Abnormality of the face</td>
<td>4.54 × 10^{-23}</td>
</tr>
<tr>
<td>HP:0012372</td>
<td>Abnormal eye morphology</td>
<td>1.71 × 10^{-22}</td>
</tr>
<tr>
<td>HP:0000598</td>
<td>Abnormality of the ear</td>
<td>1.86 × 10^{-22}</td>
</tr>
<tr>
<td>HP:0000478</td>
<td>Abnormality of the eye</td>
<td>2.24 × 10^{-22}</td>
</tr>
<tr>
<td>HP:0012374</td>
<td>Abnormality of the globe</td>
<td>2.95 × 10^{-22}</td>
</tr>
<tr>
<td>HP:0012639</td>
<td>Abnormality of nervous system morphology</td>
<td>6.08 × 10^{-22}</td>
</tr>
</tbody>
</table>
While representing only a small portion of all genes overlapped by pathogenic CNVs, solitary non-passengers are overrepresented for clinically relevant functional categories.

3.3.5 Class P genes have high evolutionary copy number constraint

Under the hypothesis of CNV pathogenicity being caused by the dosage-sensitivity of enclosed genes, we expect to see characteristic patterns of evolution of genes within pathogenic CNVs, namely a dearth of gene duplication and loss events. We investigated gene duplication and loss within the mammalian tree by counting the number of genomes in which there are copy number changes. For a given human gene that was inferred to have been present in the mammalian common ancestor (i.e., excluding newer genes, and genes where the orthologue is not identifiable).

We looked across 13 genomes and noted whether there was a gene duplication, absent orthologue, or no change in that genome (for examples, see Figure 3.6c).

We performed this for all human genes present in the mammalian ancestor, and grouped the results according to the presence of the human gene in benign or pathogenic, gain or loss CNVs as before. The box plots of these distributions are shown in Figure 3.6d. Panels 2, 3, 4, and 5 show that the conservation of copy number for genes in pathogenic regions is high across the genomes surveyed (the copy number is mostly unchanged, black points; median 12, with upper and lower quartiles at 13 and 11 species). By contrast for the genes in benign, presumably less dosage-sensitive regions, the copy number is more variable, with a greater proportion of genes having copy number changes in more genomes (lower quartile ranging between 8 and 4 species). These genes show more duplications and missing orthologues across the mammalian tree. Variance significantly increases from
Dosage sensitivity underlies human CNV pathogenicity

pathogenic groups to benign groups ($P < 1 \times 10^{-16}$, Fligner-Killeen test) indicative of lower copy number constraints in the latter. We also compared the counts of genes with conserved copy number in all tested species in each CNV-classification group (Figure 3.6d panels 3-11) and found a highly significant difference ($P < 1 \times 10^{-16}$, $\chi^2$ test).

This evolutionary analysis provides an independent measure of gene dosage sensitivity and is not dependent on CNV classification, yet the patterns match the expectations based on CNV clinical interpretations, namely, genes with evolutionary patterns suggestive of dosage sensitivity are more associated with pathogenic CNVs.

3.3.6 Evolutionary copy number conservation in pathogenic CNVRs

If we imagine a simplified scenario where there is a single dosage-sensitive gene in a region, the observed peak CNV region may nonetheless repeatedly contain multiple genes. This will be particularly true in the case of CNV hotspots (Mefford and Eichler, 2009) which may be located at several genes’ distance, repeatedly generating multi-gene CNVs. In this scenario neither the dosage-sensitive gene nor the closely linked non-dosage-sensitive genes will be observed in benign CNVs. Similarly, as evolutionary gene duplication events have the same mechanistic origins as CNVs, linked non-dosage-sensitive genes may have patterns of duplication and loss that somewhat track the pattern of the dosage-sensitive gene. However, if the linkage is broken by genome rearrangement events this incidental constraint on the non-dosage-sensitive gene will be broken. Thus, genes with the most consistent patterns of gene copy number conservation are the most interesting.

We applied the evolutionary constraint metric to genes within 14 well-characterized pathogenic regions associated with neurodevelopmental disorders. Fig-
Dosage sensitivity underlies human CNV pathogenicity

Figure 3.9a shows the copy number variation across mammals of genes in each of these regions along with some flanking genes. The flanking regions are included as a proxy indicator of any potential local duplication or loss biases. In some cases, such as 1q21.1 deletion and the 17q12 deletion, the pattern of gene conservation across mammals fits expectations in that the genes within the pathogenic CNV region are much more conserved than the genes in the flanking regions. Other regions have patterns of conservation that are close to expectations, and others show no obvious gross pattern. There is no significant difference between the total pathogenic CNV and the flanking region. However, when we compare the medians of number of "unchanged" gene copy number inside the critical region (shaded dark grey, when present) with the medians in the remainder plus the flanking region, we find that the genes in the critical region are significantly more evolutionarily conserved (Mann-Whitney U test, Bonferroni-corrected \( P = 0.0028 \)).

One of these is the region associated with 22q11 deletion syndrome which is shown in detail in Fig. 3.9b. Even though Fig. 3.9a shows several genes with multiple loss events, the detailed view shows that ten of the genes in this region have no duplication or loss events in any of the mammalian genomes tested (dark red shading). Of the 28 genes present in the critical region of this CNV (1.5Mb deletion that presents the same symptoms as the 3Mb deletion (Karayiorgou et al., 2010)), 16 have consistently detectable orthologues in all mammalian genomes, and three are missing an orthologue in only one of the 13 genomes. We chose these 13 genomes for analysis based on the high quality of the available data, but even so, we cannot exclude that some of these differences are due to missing data or poor annotation. Nonetheless, the subset of 22q11 genes that are completely conserved across mammals are good candidates for disease causation. Interestingly, TBX1, a candidate disease gene in this syndrome (Gao et al., 2013) is not completely conserved, not being detected in cow, sheep and pig. This is consistent with a
Figure 3.9 | Mammalian copy number changes for genes within known pathogenic CNV regions.
(Continued on the following page.)
Figure 3.9 | Mammalian copy number changes for genes within known pathogenic CNV regions.

(a) Copy number changes across mammalian species for genes within known pathogenic CNV regions associated with schizophrenia and other neurodevelopmental disorders obtained from (McLysaght et al., 2014). The minimal (min) CNV region (shaded dark grey) is typically the smallest region associated with the disease phenotype while the maximal (max) CNV region (shaded light grey) is typically observed. Ten flanking genes on each side are also plotted where possible. Each region is labelled above with the chromosomal band, and position along chromosome in megabases is shown on the x-axis. Genes are plotted by start position. Each point represents for one human gene the number of duplications (green) and losses (orange). Genes within regions are listed in a table available in supplementary information of Rice and McLysaght (2017).

(b) For each protein-coding gene within the 22q11 region, copy number changes across 13 mammalian species are shown. Green circles indicate where orthologues are duplicated, orange circles where orthologues are missing. Genes highlighted in light red are genes were at least one orthologue is present in all species and genes highlighted in dark red are genes with conserved one-to-one orthology across the mammalian species tested (completely conserved genes). Grey dashed outlines group orthologues that are neighbouring on their respective chromosome/scaffold in each species. Genes with greyed-out names were not included in copy number analysis and so no data is displayed for them.

We identified 7,014 human genes that have conserved copy number across all 13 mammalian genomes (List available in supplementary information of Rice and McLysaght, 2017). Even though this definition is independent of CNV status, this evolutionary information is suggestive of dosage constraint. Over 28% of these are involved in development, consistent with genes identified via pathogenic CNVs. Overall we found that the evolutionarily conserved genes are strongly enriched for “anatomical structure development” (GO:0048856; $P = 1.7 \times 10^{-31}$) as part
Table 3.5 | Most enriched functional classes of genes with conserved copy number
Full table available in supplementary information of Rice and McLysaght (2017)

<table>
<thead>
<tr>
<th>Term ID</th>
<th>Term name</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(Bonferroni-adjusted)</td>
</tr>
<tr>
<td><strong>Biological process GO terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0044699</td>
<td>single-organism process</td>
<td>7.18 x 10^{-65}</td>
</tr>
<tr>
<td>GO:0044763</td>
<td>single-organism cellular process</td>
<td>8.37 x 10^{-64}</td>
</tr>
<tr>
<td>GO:0009987</td>
<td>cellular process</td>
<td>3.31 x 10^{-41}</td>
</tr>
<tr>
<td>GO:0008150</td>
<td>biological_process</td>
<td>1.23 x 10^{-34}</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>anatomical structure development</td>
<td>1.65 x 10^{-31}</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>cell communication</td>
<td>3.38 x 10^{-31}</td>
</tr>
<tr>
<td>GO:0023052</td>
<td>signaling</td>
<td>4.43 x 10^{-31}</td>
</tr>
<tr>
<td>GO:0044700</td>
<td>single_organism_signaling</td>
<td>4.43 x 10^{-31}</td>
</tr>
<tr>
<td>GO:0044707</td>
<td>single_multicellular_organism_process</td>
<td>6.25 x 10^{-31}</td>
</tr>
<tr>
<td>GO:0032502</td>
<td>developmental_process</td>
<td>2.17 x 10^{-30}</td>
</tr>
<tr>
<td>GO:0044767</td>
<td>single_organismdevelopmental_process</td>
<td>7.92 x 10^{-30}</td>
</tr>
<tr>
<td>GO:0032501</td>
<td>multicellular_organismal_process</td>
<td>1.09 x 10^{-28}</td>
</tr>
<tr>
<td>GO:0009653</td>
<td>anatomical structure morphogenesis</td>
<td>2.69 x 10^{-27}</td>
</tr>
<tr>
<td>GO:0006793</td>
<td>phosphorus metabolic process</td>
<td>4.33 x 10^{-27}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cellular component GO terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005737</td>
<td>cytoplasm</td>
<td>3.82 x 10^{-40}</td>
</tr>
<tr>
<td>GO:0044444</td>
<td>cytoplasmal part</td>
<td>6.25 x 10^{-29}</td>
</tr>
<tr>
<td>GO:0044424</td>
<td>intracellular part</td>
<td>6.76 x 10^{-20}</td>
</tr>
<tr>
<td>GO:0005622</td>
<td>intracellular</td>
<td>2.10 x 10^{-18}</td>
</tr>
<tr>
<td>GO:0042995</td>
<td>cell projection</td>
<td>1.82 x 10^{-16}</td>
</tr>
<tr>
<td>GO:0044422</td>
<td>organelle part</td>
<td>8.93 x 10^{-16}</td>
</tr>
<tr>
<td>GO:0044446</td>
<td>intracellular organelle part</td>
<td>1.09 x 10^{-14}</td>
</tr>
<tr>
<td>GO:008588</td>
<td>bounding membrane of organelle</td>
<td>2.24 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0043226</td>
<td>organelle</td>
<td>2.81 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0044464</td>
<td>cell part</td>
<td>3.13 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0005623</td>
<td>cell</td>
<td>3.50 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0097458</td>
<td>neuron part</td>
<td>5.04 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0012505</td>
<td>endomembrane system</td>
<td>6.26 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0031090</td>
<td>organelle membrane</td>
<td>1.65 x 10^{-11}</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>membrane</td>
<td>1.50 x 10^{-09}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Molecular function GO terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005515</td>
<td>protein binding</td>
<td>7.61 x 10^{-36}</td>
</tr>
<tr>
<td>GO:0003824</td>
<td>catalytic activity</td>
<td>7.91 x 10^{-23}</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>binding</td>
<td>5.09 x 10^{-17}</td>
</tr>
<tr>
<td>GO:0016740</td>
<td>transferase activity</td>
<td>2.65 x 10^{-14}</td>
</tr>
<tr>
<td>GO:0043168</td>
<td>anion binding</td>
<td>4.88 x 10^{-13}</td>
</tr>
<tr>
<td>GO:0016773</td>
<td>phosphotransferase activity, alcohol group as acceptor</td>
<td>5.09 x 10^{-13}</td>
</tr>
<tr>
<td>GO:0016772</td>
<td>transferase activity, transferring phosphorus-containing groups</td>
<td>8.87 x 10^{-12}</td>
</tr>
<tr>
<td>GO:0016301</td>
<td>kinase activity</td>
<td>2.35 x 10^{-11}</td>
</tr>
<tr>
<td>GO:0004672</td>
<td>protein kinase activity</td>
<td>6.96 x 10^{-11}</td>
</tr>
<tr>
<td>GO:0036094</td>
<td>small molecule binding</td>
<td>2.44 x 10^{-10}</td>
</tr>
<tr>
<td>GO:0097367</td>
<td>carbohydrate derivative binding</td>
<td>4.76 x 10^{-08}</td>
</tr>
<tr>
<td>GO:0004674</td>
<td>protein serine/threonine kinase activity</td>
<td>1.19 x 10^{-07}</td>
</tr>
<tr>
<td>GO:0032559</td>
<td>adenyly ribonucleotide binding</td>
<td>2.26 x 10^{-07}</td>
</tr>
<tr>
<td>GO:0032553</td>
<td>ribonucleotide binding</td>
<td>4.28 x 10^{-07}</td>
</tr>
<tr>
<td>GO:0030554</td>
<td>adenyly nucleotide binding</td>
<td>8.47 x 10^{-07}</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dosage sensitivity underlies human CNV pathogenicity

Figure 3.10 | Distance to nearest developmental gene for developmental genes grouped by number of genomes where orthologue has unchanged copy number

For each developmental gene, distance to the closest upstream or downstream developmental gene in base pairs was calculated, ignoring strand. For developmental genes that overlap a distance of 1 base pair was assigned. Developmental genes are grouped by the number of genomes where orthologue copy number is unchanged (13 where a gene has a one-to-one relationship with all 13 mammalian genomes tested). Width of each box is proportional to sample size in each group and the median is shown within each box. Upper and lower hinges of boxes correspond to the first and third quartiles. Whiskers extend to values 1.5 * interquartile range.

of developmental processes (GO:0032502; \( P = 2.2 \times 10^{-30} \)), “cell communication” (GO:0007154; \( P = 3.4 \times 10^{-31} \)), “phosphorus metabolic process” (GO:0006793; \( P = 4.3 \times 10^{-27} \)), and “macromolecule modification” (GO:0043412; \( P = 5.3 \times 10^{-23} \)) specifically “protein modification process” (GO:0036211; \( P = 2.5 \times 10^{-23} \)).

Additionally, conserved genes are also enriched for “localization” (GO:0051179; \( P = 2.8 \times 10^{-23} \)), “regulation of biological process” (GO:0050789; \( P = 3.3 \times 10^{-21} \)), and “response to stimulus” (GO:0050896; \( P = 3.5 \times 10^{-21} \)), encompassing “response to organic substance” (GO:0010033; \( P = 6.7 \times 10^{-16} \)), “response to
endogenous stimulus” (GO:0009719; \( P = 7.0 \times 10^{-12} \)), and “response to oxygen-containing compound” (GO:1901700; \( P = 1.8 \times 10^{-10} \)) (Table 3.5). We confirmed that the enrichment for developmental genes is not due to clustering of those genes on the genome because we observe no effect of distance to the nearest developmental gene in the human genome and the conservation across mammalian genomes (Figure 3.10). Furthermore we observed that genes conserved across the mammalian tree are enriched for OMIM disease genes (18.2%; \( P < 1.0 \times 10^{-16} \), \( \chi^2 \) test) relative to genes with copy number changes (13.4%). We find a similar trend for candidate haploinsufficient genes (genes with probability of loss-of-function mutation intolerance >90% inferred in Lek et al. (2016)) with conserved genes enriched (22.8%) compared to genes with copy number changes (15.0%; \( P < 1.0 \times 10^{-16} \), \( \chi^2 \) test). Clearly, conserved genes are functionally distinct and involved in biologically important processes.

We tested genes with conserved copy number for representation among genes overlapped by benign CNVs and genes overlapped by an independent human CNV map (Zarrei et al., 2015). We found them to be underrepresented among benign CNV genes with 5.8% (393/6,809) of conserved genes overlapped by benign CNV compared to 10.9% (1,272/11,632) of genes not conserved in all 13 genomes tested (\( P < 1.0 \times 10^{-16} \), \( \chi^2 \) test). Similarly conserved genes are overlapped less in a control CNV map (35.6% of 7,014 conserved genes overlapped) compared to genes not conserved in the genomes tested (38.4% of 13,300 genes, \( P = 0.0001 \), \( \chi^2 \) test). This is consistent with our expectation that these genes are under copy number constraint and with previous work that has shown comparatively more duplications of genes in benign CNVRs (Nguyen et al., 2008) and of haplosufficient genes (Huang et al., 2010).

The pattern of conservation across the mammalian tree suggests an ancient and persistent dosage constraint, and as such we expect that CNVs encompassing
orthologues of these genes would also be deleterious in other mammals. We tested mouse orthologues of genes with conserved copy number and we found them to be depleted among mouse CNVs compared to other protein-coding genes (23.6% of mouse orthologues of conserved genes are overlapped by mouse CNVs compared to 27.3% of other mouse genes, $P = 3.0 \times 10^{-9}$, $\chi^2$ test). This indicates that these genes are constrained compared to other genes within mouse. These results suggest that evolutionary trends are informative in the identification of dosage sensitive genes.

3.4 Discussion

Though the phenotypes resulting from CNVs at different genomic locations can differ quite widely, there are certain commonalities that allow a deeper insight into the genetic and biological mechanisms of CNV pathogenicity. That we observe trends in function and evolutionary patterns for genes within pathogenic CNVs supports the hypothesis that gene dosage sensitivity is a predominant causative factor. In particular, CNV subregions that recur frequently in pathogenic cases, or CNV regions that are rare but associated with pathogenicity, are biased with respect to the genes they contain both in terms of function and evolution.

In particular, the observation that genes with constrained evolutionary patterns of gene duplication and loss are usually found within pathogenic CNVs strongly supports the model whereby dosage-sensitivity of individual genes enclosed by a CNV is responsible for pathogenicity. This pattern is not predicted by any other model (though does not exclude the co-existence of other mechanisms of CNV pathogenicity). Furthermore, the identification of genes with such evolutionary patterns supplies a shortened list of candidate genes for further inspection. Peak regions of pathogenic CNVs that contain only one gene that is exclusively found
in pathogenic CNVs are of particular interest. Based on an admittedly simplistic logic, these 199 genes are candidate causative disease genes. Consistent with this, these genes are rarely found to be duplicated or lost in other mammals (Fig. 3.6d, panel 2).

Importantly, this analysis of gene duplication and loss is restricted to genes where we can infer presence in the common ancestor to all 13 mammalian genomes examined. Thus we avoid any problems associated with the increased difficulty in detecting quickly-evolving genes (Elhaik et al., 2006; Wolfe, 2004). Genes which we cannot infer to be present in the common ancestor are either new genes or older genes that are difficult to detect because of gene loss or extensive sequence evolution, and it is not possible to distinguish these without more detailed inspection of the loci. However, we found that genes that were not inferred in the ancestral mammal are enriched in benign CNVRs compared to the rest of the genome (7.6% (137/1,802) vs 4.8% (852/17,628) respectively, $P = 4.7 \times 10^{-7}$, $\chi^2$ test), suggesting lower evolutionary constraint, consistent with having less phenotypic effect upon disruption.

Haploinsufficient genes are genes where there is a minimum amount of gene product required to attain a wild-type phenotype. Logically, these are distinct from dosage-balanced genes where any significant disruption in amount of product, be it increased or decreased, will induce a phenotype, however in practice the two may overlap (if, for example, one tests only for a phenotype in heterozygote knockouts). Interestingly, in their analysis of haploinsufficiency, Huang et al. (2010) observed fewer paralogues of haploinsufficient genes, even though this is not predicted by haploinsufficiency, but which would be expected of general dosage sensitivity or dosage balance. We would expect that the genes showing the pattern indicated by the yellow segment in Fig. 3.6b, that is benign gain but pathogenic loss, should naturally be haploinsufficient genes. Conversely, genes present in
Dosage sensitivity underlies human CNV pathogenicity

Pathogenic gain CNVs but benign loss CNVs (green segment in Fig. 3.6b) may be aggregation-prone at high concentration. Whereas we lack well-curated data on aggregation-prone genes to test the latter relationship, we can use the recently available haploinsufficiency data to test the former. We observed the expected enrichment for haploinsufficiency among genes found within benign gain, pathogenic loss regions. These might be considered “simple” haploinsufficient genes. However, the enrichment among class P genes described above suggests that many haploinsufficient genes are also dosage sensitive in other ways.

There is great interest in the relationship between development and dosage sensitivity and CNVs in general. As we noted, there is a potential bias in the annotation of disease CNVs due to this interest and due to the fact that developmental disorders are expected to be more reliably identified at the time of sample collection. Therefore, the enrichment for developmental genes in pathogenic CNVs must normally be interpreted in that light. However, the evolutionary measures based on conservation of copy number across mammalian species are independent of disease annotation and have no such reporter or study bias. Our finding that these evolutionarily constrained genes are indeed enriched for developmental genes confirms the view of development as an inherently dosage sensitive process.

This is the first comparison of the genome evolutionary trends of genes in benign and pathogenic CNVs. We have revealed distinct functional and evolutionary trends for the two classes of CNVs. This points to the usefulness of evolutionary metrics in the interpretation of CNVs.
Chapter 4

Prediction of pathogenic copy number variation yields insights into variant inheritance

4.1 Introduction

Human copy number variants (CNVs) are an order of magnitude more common than single-nucleotide polymorphisms (Conrad et al., 2010). Much of this variation has no phenotypic consequence, as CNVs are frequently small, intergenic or encompass genes that can tolerate a change in copy number (Zarrei et al., 2015). However, CNVs have previously been associated with a number of human conditions, most notably neurodevelopmental disorders including autism spectrum disorders, schizophrenia, intellectual disability, attention deficit hyperactivity disorder, developmental delay and epilepsy (Sebat et al., 2007; Stefansson et al., 2014; Stefansson et al., 2008; Walsh et al., 2008; Mefford et al., 2008; Lesch et al., 2011; Helbig et al., 2009; Cooper et al., 2011). Due to this implication in disease, there is a need to understand and characterise the genetic and phenotypic effects of
Pathogenic CNV prediction yields inheritance insights

Multiple mechanisms can account for and explain aberration of gene function and phenotypic changes due to the presence of a CNV at a given locus. These are detailed in 1.1.3, page 6. Dosage sensitivity of the included genes remains the prevailing hypothesis for CNV pathogenicity and here we investigate whether pathogenicity can be predicted by examination of the genes enclosed in a CNV. The ability to accurately predict pathogenicity of a given CNV would be a powerful tool to increase understanding of the impact of CNVs and to aid clinical diagnosis of patient variation. Previously, several attempts have been made to develop automated learning and interpretation of CNV pathogenicity (Hehir-Kwa et al., 2010; Engchuan et al., 2015; Erikson et al., 2015; Foong et al., 2015). However, these methods used information about overlaps with other reference data sets, features such as CNV length, or phenotype-specific variants. Overlaps with reference data sets and CNV length does not aid in elucidating a specific method of pathogenicity for a region at a given locus. Additionally, while the phenotypes from CNVs vary greatly, the underlying general mechanisms of dosage-sensitivity at the cellular level are probably more common.

A discipline of computer science is machine learning, the use of general algorithms that enable computers to learn how to achieve specific tasks without explicit programming. Machine learning is often used to solve a problem for which it is difficult to write algorithmic rules, for example recognising handwriting, filtering spam emails, identifying tumours in medical images, etc. The learning algorithms used determine useful features or patterns of features that best aid in solving a problem. A wide range of learning algorithms and protocols exist. To accurately predict pathogenicity, classification algorithms would be useful, that is where a labelled dataset of known examples is used and patterns are learned that accurately describe the difference between groups in the dataset, in this instance
benign and pathogenic variants.

Typically before attempting classification, the labelled dataset is split into a set of samples that the classifier can use to learn and a held-out set that the classifier never sees during learning. Splitting the data in this way enables testing of classifier performance on the unseen held-out data after training has taken place. Allowing the classifier to learn on all available data would not allow independent testing of accuracy as the classifier would have already witnessed every sample and adjusted its knowledge and response to each sample in the dataset. Accuracy will be overestimated and so splitting the dataset and holding back a set of samples for testing is recommended to best estimate accuracy.

Decision trees are one form of machine learning classifier where classification is determined by passing a sample through a tree-like structure of branching decisions. The various values of the sample determine its path through the tree until it reaches a leaf of the tree, at which a classification has been assigned during learning. However, decision trees suffer from high variance, that is, constructing decision trees on random subsets of the same data yields very different trees and are also prone to overfitting. Overfitting is the when the model too closely describes noise and random error in the training dataset instead of generalising on the underlying relationship in the data. When a classifier is overfit, it performs very well on the training data but has comparably worse performance on additional data.

A method of countering high variance and overfitting in decision trees is to construct a number of decision trees, each on bootstrap subsamples of the data. Bootstrap samples have the same number of samples as the total dataset but are generated randomly with replacement. During classification, samples are passed down all decisions trees and each tree votes for a class. The class with the most votes overall is the final classification. This type of classifier is called a bagged
Pathogenic CNV prediction yields inheritance insights

tree. If multiple features are correlated then variance will not be reduced in a bagged tree classifier. However, an alternative method exists called a random forest classifier. Random forests are similar to bagged trees, however, during the construction of trees only a random subset of features are considered at node splits. In a bagged tree classifier, during the decision of which feature to use at a tree split all features are considered before one feature is chosen. In random forest classifiers a limit is placed on the number of features to pick from at any given split. Employing this approach, a random forest both decorrelates trees and reduces variance and performs very well for many classification tasks.

Here, we fit a random forest classifier to a dataset of CNVs with clinical interpretations of benign or pathogenic using mostly genic features including evolutionary conservation, intolerance to mutations, and expression. We find that the trained classifier is highly accurate at predicting pathogenicity among test samples. As an independent validation, we predicted CNVs from healthy controls and a group of CNVs from individuals with rare disorders and observed significantly different proportion of CNVs being classified as pathogenic between the two groups. This difference is still evidence even when CNVs are grouped by length. Additionally, we find a relationship between the proportion of pathogenic patient CNVs and how a CNV is inherited, with de novo CNVs in patients being pathogenic more often than CNVs inherited from a healthy parent. These results demonstrate that CNV pathogenicity can be predicted by the genes contained within CNV breakpoints and that pathogenicity varies with CNV inheritance context.
4.2 Materials & Methods

4.2.1 Data

dbVar CNVs with clinical interpretations were filtered so that no two CNVs overlapped the same set of protein-coding genes (Lappalainen et al., 2013). For CNVs that overlapped the same set of genes, only the first CNV encountered was included. This filtering step left 637 benign CNG variants, 403 benign CNL variants, 709 pathogenic CNG variants and 1,756 pathogenic CNL variants. Classifier input features listed in Table 4.1 were calculated for variants using a custom Python script. The order of CNV samples was randomly sorted, and the dataset was split into training and test datasets with 80% and 20% of samples in each respectively.

4.2.2 Classifier training

Scikit-learn’s random forest classifier implementation (Pedregosa et al., 2011) was used with default parameters except for specifying a maximum tree depth of 5 and a minimum of 10 samples per leaf. A random forest classifier with these parameters was fit on the CNVs in the training set, and out-of-bag error estimates were measured with the addition of each new tree to the forest (Figure 4.1). A suitable number of trees was chosen for the final classifier with a balance between the point where the out-of-bag error stabilised and a low number of trees to optimise classifier performance. A final classifier with 200 trees, maximum tree depth of 5, a minimum of 10 samples per leaf and classes weighted inversely proportional to class frequencies was trained on the training dataset. Out-of-bag error and 10-fold cross-validation accuracy were measured.
4.2.3 Classifier testing

CNV samples in the 20% held-out test set were predicted by the classifier and output predictions were compared to true labels in a confusion matrix (Figure 4.2A). A ROC curve was plotted and area under the curve measured (Figure 4.2B).

4.2.4 Classifier validation

This study makes use of data generated by the DECIPHER community. A full list of centres who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. Funding for the DECIPHER project was provided by the Wellcome Trust. Case CNVs from DECIPHER database (Firth et al., 2009) and control CNVs from DGV (MacDonald et al., 2014) both without clinical interpretations were classified as independent validation of the CNV pathogenicity classifier. CNVs that do not intersect at least one protein-coding gene and those that match dbVar CNVs exactly were removed.

4.2.5 Inheritance types

DECIPHER CNVs were grouped by inheritance type and classified using the CNV pathogenicity classifier (Table 4.2).

4.3 Results

4.3.1 Random forest classifier accurately predicts variant pathogenicity

Identification of regions of the genome that contribute to pathogenicity when affected by a copy number variant is desirable, yet remains difficult. A subset of dbVar CNVs (Lappalainen et al., 2013) have associated clinical interpretations of
benign or pathogenic that were assigned manually, typically based on patterns of inheritance and characteristics of known syndromes (Miller et al., 2010). For example, a CNV inherited from a parent with a similar phenotype implicates that CNV as a driver of the phenotype. Similarly, a CNV containing an OMIM disease gene that when disrupted by other kinds of mutations (e.g. a heterozygous inactivating mutation) produces a similar phenotype adding further endorsement of conferring a pathogenic label.

Here, genomic features were enlisted in combination with machine learning techniques to attempt to distinguish between benign and pathogenic CNVs and enable their accurate and quick prediction. The selection of mostly genic features, listed in Table 4.1, were chosen as potential indicators of dosage-sensitivity. The features include involvement of gene products in protein interactions and complexes (Guan et al., 2007; Papp et al., 2003a; Veitia, 2004), observed frequency of deletions and duplications in exome data relative to expected frequency (Ruderfer et al., 2016), haploinsufficiency (Veitia, 2002; Dang et al., 2008), expression level in multiple tissues, percentage GC content, genic evolutionary information (Makino and McLysaght, 2010; Rice and McLysaght, 2017) and density of long interspersed nuclear elements (LINEs, Cardoso et al., 2016) on CNV.

As the dataset of dbVar CNVs with clinical interpretations has numerous overlaps, we filtered samples to avoid CNVs that overlapped the same set of protein-coding genes. For these CNVs, we removed all but the first CNV encountered. After this filtering step, the dataset contained 636 benign CNG variants, 401 benign CNL variants, 709 pathogenic copy number gain (CNG) variants and 1,756 pathogenic copy number loss (CNL) variants. Classifier input features were calculated for these remaining variants. The order of samples was randomised and the dataset was split into training and test datasets with 80% and 20% of samples in each respectively.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV type</td>
<td>Whether CNV is a loss or gain variant</td>
<td>Categorical</td>
<td>- loss / gain</td>
</tr>
<tr>
<td>Protein-coding gene density</td>
<td>Density of Ensembl GRCh37 protein-coding genes</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean gene length</td>
<td>Mean length of protein-coding genes that intersect with CNV</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean protein interactions</td>
<td>Mean number of protein interactions per gene as found in BioGRID</td>
<td>Continuous</td>
<td>BioGRID (Chatr-Aryamontri et al., 2017)</td>
</tr>
<tr>
<td>Mean gene haploinsufficiency score</td>
<td>Mean ExAC probability of intolerance to loss-of-function mutation</td>
<td>Continuous</td>
<td>ExAC (Lek et al., 2016)</td>
</tr>
<tr>
<td>Mean gene CNV deletion intolerance score</td>
<td>Mean ExAC probability of intolerance to deletion CNV</td>
<td>Continuous</td>
<td>ExAC (Ruderfer et al., 2016)</td>
</tr>
<tr>
<td>Mean gene CNV duplication intolerance score</td>
<td>Mean ExAC probability of intolerance to duplication CNV</td>
<td>Continuous</td>
<td>ExAC (Ruderfer et al., 2016)</td>
</tr>
<tr>
<td>Mean gene % GC content</td>
<td>Mean percentage GC content of protein-coding genes</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>LINE density</td>
<td>Density of long interspersed nuclear elements (LINEs)</td>
<td>Continuous</td>
<td>TranspoGene (Levy et al., 2008)</td>
</tr>
<tr>
<td>Proportion ohnologues</td>
<td>Proportion of protein-coding genes that are ohnologues</td>
<td>Continuous</td>
<td>Makino and McLysaght (2010)</td>
</tr>
<tr>
<td>Proportion dosage-balanced ohnologues</td>
<td>Proportion of protein-coding genes that are dosage-balanced ohnologues</td>
<td>Continuous</td>
<td>Makino and McLysaght (2010)</td>
</tr>
<tr>
<td>Proportion mammalian CCN genes</td>
<td>Proportion of protein-coding genes that are mammalian copy number conserved (CCN) genes</td>
<td>Continuous</td>
<td>Rice and McLysaght (2017)</td>
</tr>
<tr>
<td>Proportion protein complex subunits</td>
<td>Proportion of protein-coding genes that are protein complex subunits as found in Uniprot</td>
<td>Continuous</td>
<td>Uniprot (The UniProt Consortium, 2015)</td>
</tr>
<tr>
<td>Mean chimp % identity</td>
<td>Mean percentage identity with one-to-one chimp orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean mouse % identity</td>
<td>Mean percentage identity with one-to-one mouse orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean chimp dN</td>
<td>Mean number of nonsynonymous substitutions per nonsynonymous site (dN) for chimp one-to-one orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean mouse dN</td>
<td>Mean number of nonsynonymous substitutions per nonsynonymous site (dN) for mouse one-to-one orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean chimp dS</td>
<td>Mean number of synonymous substitutions per synonymous site (dS) for chimp one-to-one orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean mouse dS</td>
<td>Mean number of synonymous substitutions per synonymous site (dS) for mouse one-to-one orthologues</td>
<td>Continuous</td>
<td>Ensembl (Yates et al., 2016)</td>
</tr>
<tr>
<td>Mean expression</td>
<td>Mean expression of genes across 53 tissues measured in the GTEx project</td>
<td>Continuous</td>
<td>GTEx (GTEx Consortium, 2015)</td>
</tr>
<tr>
<td>Mean expression standard deviation</td>
<td>Mean standard deviation of expression across 53 tissues measured in the GTEx project</td>
<td>Continuous</td>
<td>GTEx (GTEx Consortium, 2015)</td>
</tr>
</tbody>
</table>
A random forest classifier was chosen as a suitable machine learning algorithm to distinguish between benign and pathogenic CNVs due to its high accuracy in other datasets and its robustness against overfitting. Random forest classifiers are meta estimators, consisting of multiple decisions trees. Input vectors are passed through all trees and in binary classification, each tree votes for a class. The vector is finally classified as the class with the most votes across all trees. To encourage generalisation in our classifier and reduce the potential for overfitting to the training data, we specified that decision trees could not have more than five levels of nodes and that a minimum of 10 samples was required per tree leaf. These parameters will reduce excessive tree rules being created e.g. exceptions being made for individual samples.

While many trees can be added to a random forest without penalty to accuracy or overfitting, speed of classification slows with additional trees. To estimate the number of trees necessary for a given forest, out-of-bag (OOB) error estimates can be used. OOB error is an internal estimate of classifier performance on the training data. Each tree in the forest is constructed from a different random sample of the data as a random bootstrap sample is taken during construction of a given tree. Respective omitted cases are passed through each tree for classification and this allows internal estimation of error as training takes place. Starting with two trees and increasing to 1000 trees, OOB error was measured for the training set of CNV samples (Figure 4.1). OOB error estimation was conducted for two classifiers with the previously specified parameters of tree depth and minimum samples per leaf but different limits on the number of features that can be considered when determining the best split in data at any tree node. First, a true random forest classifier where the square root of the number of input features are considered and second, a random forest classifier where all features are available for consideration, essentially a bagged tree classifier. A maximum limit on features considered mostly
Out-of-bag (OOB) error can be estimated during training with the addition of each new tree. Here, OOB error is measured at the addition a new tree for two classifiers each with two trees at the outset. The first case (orange) is for a classifier limited to considering a random subsample of features, here the square root of the number of total features, when determining the best split in data at any node. The second case (green) allows all features to be considered for comparison. This case is a bagged tree classifier rather than a random forest as all features are available for consideration instead of a random subset. It is clear that OOB error quickly drops with the addition of just a few trees and stabilises before reaching 100 trees.
A B

Figure 4.2 | Classifier performance on test dataset
A, Normalised confusion matrix showing performance of the CNV pathogenicity classifier predicting the held-out test set. A confusion matrix shows the true positive and false negative rate of the classifier for each class, and the number of CNV samples in parenthesis for each scenario. B, Receiver operating characteristic (ROC) curves allow the prediction quality of a classifier to be evaluated. With false positive rate on the x axis and true positive rate on the y axis, good predictive binary classification from a classifier will have a ROC curve further in the upper left corner. A dashed line is shown for a curve with area 0.5, where prediction in a two class dataset is no better than random chance.

affects forests with very few trees, but we observe that both classifiers converge on similar minimum error rates stabilising with less than 100 trees. Therefore, we selected 200 trees as the number of estimators for our random forest moving forward as this remained efficient computationally and gave probability resolution of 0.005.

As there are more than twice as many pathogenic samples than benign samples in our dataset, it is important to balance their effect on the classifier. To achieve balance, we adjusted class weight during training to be inversely proportional to the frequency of each class in the input data. Ten-fold cross-validation of the classifier was performed on the training set achieving a mean accuracy of 90.4% with a standard deviation of 3.1%. After training the classifier on the full training dataset, an internal OOB error rate of 9.7% was observed. Using this
Pathogenic CNV prediction yields inheritance insights

trained classifier to predict pathogenicity of samples in the 20% of samples set aside as the test dataset we achieved an accuracy of 90.4%. Examining benign and pathogenic classes separately, we saw slightly higher accuracy for pathogenic samples with a true positive rate of 91.3% compared to 88.1% for benign samples (Figure 4.2A). This prediction imbalance of over-predicting pathogenicity occurred despite balancing class weights to the frequency found of each in the training dataset and should be taken into consideration when interpreting output. A receiver operating characteristic (ROC) curve is a test of classifier quality and examines the increase in true positive rate with a corresponding increase in false positive rate. An area below a ROC curve of 0.5 for binary classification suggests that classifier performance is no better than random chance. Here for our CNV pathogenicity classifier, we found an area of 0.96, a satisfactory score indicating good classifier performance (Figure 4.2B).

4.3.2 Dosage-balanced ohnologues, conserved mammalian genes, and gene interactions are important determinants of CNV pathogenicity

Random forest classifiers provide an estimate of input feature importance, in the form of mean impurity decrease or Gini importance. Impurity decrease is a measure of how effective an input variable is at separating different classes to improve the purity of a tree node split. A random variable will not aid separation of classes and decrease impurity for a node split, however, a good predictive variable will describe differences between classes well and be useful in distinguishing between them. Mean impurity decrease was calculated for our CNV pathogenicity classifier, and this is scaled to 1 to give relative importances (Figure 4.3A). We found that the proportion of dosage-balanced ohnologues, proportion of mam-
Figure 4.3 | Feature importance and correlation in CNV classifier
(Continued on the following page.)
Figure 4.3 | Feature importance and correlation in CNV classifier

A. Feature importance is measured by the mean impurity decrease from each feature and the features are ranked according to this measure. Mean impurity decrease, or Gini importance, is a measure of how effective a variable is at separating different classes to improve the purity of tree node splits. Red bars are relative feature importances of the forest, and standard error bars show inter-trees variability. B. Pearson correlation between input features calculated for dbVar dataset. Only significant correlations are shown after Bonferroni correction.

Mammalian conserved copy number (CCN, as identified in 3.3.7, see page 70) genes and mean number of protein interactions were the three most useful features in our classifier. Involvement in protein complexes and expression level were also informative features. This is consistent with previous findings of dosage-balanced ohnologues and mammalian CCN genes being enriched on pathogenic CNVs and depleted on benign CNVs (McLysaght et al., 2014; Rice and McLysaght, 2017).

It is important to note that mean impurity decrease can be affected by correlation between features. In the case of two features that are highly correlated these features may be interchangeable when deciding best node splits. This interchangeability causes their relative feature importance to be shared and neither feature to stand out as useful. Therefore, it is important to consider feature correlations also when evaluating importance. For our input features we see that the two expression features ($r = 0.96$), chimp dN and dS ($r = 0.91$) and ohnologue proportion and its dosage-balanced subset ($r = 0.67$) are all highly correlated (Figure 4.3B). Given this, summing both expression features a combined expression importance (of 0.10) would still not match the three most important features. Chimp dN and dS values both score poorly as useful features (0.013 and 0.003, respectively) and so their correlation is inconsequential. As dosage-balanced ohnologues are a subset of ohnologues, they are naturally correlated, but it is clear that the dosage-balanced group can distinguish between benign and pathogenic samples better. It is possible that the ohnologue proportion feature may even mask the full importance of the
dosage-balanced ohnologue feature.

4.3.3 Control and case CNVs can be accurately distinguished by classifier

We expect that some of the CNVs in the genomes of patients with rare disorders are contributing to their phenotype. Therefore, when pooled a higher proportion of patient CNVs are likely pathogenic when compared to CNVs of healthy control individuals. If our classifier can correctly distinguish between benign and pathogenic CNVs, it should be able to make a distinction between two sets of CNVs: one from healthy controls and one from patients with disorders. To validate this, we used case CNVs from patients in the Database of Genomic Variation and Phenotype in Humans using Ensembl Resources (DECIPHER) (Firth et al., 2009) and control CNVs from healthy individuals in Database of Genomic Variants (DGV) (MacDonald et al., 2014). Initially, we examined all CNVs that overlap at least one protein-coding gene that did not have identical start and end breakpoints with any CNV in the dbVar dataset used in classifier construction. Our classifier labelled case CNVs as pathogenic far more often than control CNVs with 57.2% (11,203/19,581) of case CNVs labelled pathogenic compared to 1.4% (34,819/2,514,857) of control CNVs ($P < 1 \times 10^{-16}$, $\chi^2$ test; Figure 4.4A, left). We do not expect all case CNVs to be pathogenic as patients should have healthy variants in addition to one or more causative CNVs. Furthermore, due to reduced penetrance of some pathogenic variants we may see some ‘pathogenic’ variants in healthy, control individuals.

As a large difference in CNV length exists between the two groups (median length of 2,527 bp for control CNVs and 602,616 bp for case CNVs), it was important to ensure that CNV length did not play a role in pathogenicity determination. Longer CNVs are more likely to encounter a given gene of any kind and potentially
A All CNVs overlapping 1+ genes

B CNVs with unique gene overlaps

Figure 4.4 | Classified control and case CNVs
Control CNVs from DGV and case CNVs from DECIPHER were classified by our CNV pathogenicity classifier. Sample size are shown below each panel. A, Classification results of all CNVs that overlapped at least gene. Left, proportions of case and control CNVs classified as benign or pathogenic. Right, proportions of variants classified as benign or pathogenic for case and control (Con.) CNVs grouped by CNV length. Bonferroni-adjusted $\chi^2$ test P values shown above. B, Classification results of CNVs filtered to remove duplicates to give CNVs with unique gene overlaps.
then include a feature set that will lead to a pathogenic classification. To correct for this, case and control CNVs were grouped by CNV length and compared within each group (Figure 4.4A, right). While control CNVs between 125-250 kb in length have a relatively high level of pathogenicity (10.8%), it is clear that a greater proportion of case CNVs are pathogenic (26.7%). A significant difference between case and control CNVs is observed for all five groups with CNVs longer than 15 kb after Bonferroni correction. The proportion of case CNVs classified as pathogenic increases with length, reaching a maximum for case CNVs longer than a megabase with 90.6% being labeled as pathogenic.

The case and control CNV datasets each include duplicate CNVs with identical coordinates from different individuals and additionally include CNVs that have significant overlaps (i.e. different breakpoints but overlap the same set of genes). Within each dataset, we filtered case and control CNVs to remove duplicates that overlapped the same group of protein-coding genes. For example, if multiple CNVs overlap three genes, gene A, gene B, and gene C, only one CNV is included. If another CNV only overlaps gene A and gene B or gene B and gene C it is not removed. This filtering step removes 37.0% of case CNVs and 98.6% of control CNVs. After filtering, we still see a significant difference in the proportion of CNVs classified as pathogenic between case and control datasets (64.9% of 12,339 case CNVs and 13.1% of 34,666 control CNVs; $P < 1 \times 10^{-16}, \chi^2$ test; Figure 4.4B, left).

When CNVs with unique gene overlaps are grouped by length, the proportion of control CNVs classified as pathogenic increases moderately with length but plateaus at about 30% (Figure 4.4B, right). For CNVs longer than 250 kb, case CNVs are significantly more pathogenic as the proportion classified as pathogenic continues to increase with length. Comparing the full and filtered control CNV dataset in Figure 4.4, it is notable that the proportion of pathogenicity is higher
Pathogenic CNV prediction yields inheritance insights

**Table 4.2 | Pathogenicity classification of CNVs by inheritance types in DECI-PHER database**

<table>
<thead>
<tr>
<th>Inheritance type</th>
<th>% classified ‘pathogenic’</th>
<th>( \chi^2 ) test standardised residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biparental</td>
<td>5.9% (1/17)</td>
<td>-3.9</td>
</tr>
<tr>
<td><em>de novo</em> constitutive</td>
<td>83.9% (3,685/4,395)</td>
<td>45.4</td>
</tr>
<tr>
<td><em>de novo</em> mosaic</td>
<td>94.6% (35/37)</td>
<td>5.0</td>
</tr>
<tr>
<td>Imbalance arising from a balanced parental rearrangement</td>
<td>90.9% (140/154)</td>
<td>9.4</td>
</tr>
<tr>
<td>Inherited from a normal parent</td>
<td>37.1% (1,055/2,842)</td>
<td>-18.5</td>
</tr>
<tr>
<td>Inherited from parent with similar phenotype to child</td>
<td>64.9% (421/649)</td>
<td>6.0</td>
</tr>
<tr>
<td>Inherited from parent with unknown phenotype</td>
<td>44.9% (87/194)</td>
<td>-2.4</td>
</tr>
<tr>
<td>Maternally inherited, constitutive in mother</td>
<td>35.8% (656/1,835)</td>
<td>-15.7</td>
</tr>
<tr>
<td>Maternally inherited, mosaic in mother</td>
<td>58.8% (10/17)</td>
<td>0.5</td>
</tr>
<tr>
<td>Paternally inherited, constitutive in father</td>
<td>31.6% (498/1,574)</td>
<td>-17.9</td>
</tr>
<tr>
<td>Paternally inherited, mosaic in father</td>
<td>50.0% (2/4)</td>
<td>-0.1</td>
</tr>
<tr>
<td>Unknown</td>
<td>49.9% (5,165/10,349)</td>
<td>-9.4</td>
</tr>
</tbody>
</table>

For the filtered set of CNVs. In the filtered dataset, rare variants are given equal weight to common variants after filtering. For example, 99 CNVs might overlap *gene A* and one CNV might overlap *gene B*. The CNVs overlapping *gene A* are classified as benign and the CNV overlapping *gene B* is classified as pathogenic. In the full dataset, 1% of total CNVs are labelled pathogenic, however in the filtered dataset, 50% of total CNVs are classified as pathogenic. The increase in pathogenicity that we see in the actual dataset suggests that control variants that are subsequently classified as pathogenic are present at low frequency in the full dataset and may have reduced penetrance. This accounts at least in part for the increase in pathogenicity observed upon filtering control CNVs in Figure 4.4B.
Figure 4.5 | Classified DECIPHER case CNVs grouped by inheritance type
DECIPHER case CNVs from three CNV inheritance types classified by our CNV pathogenicity classifier. Increasing proportion of variants classified as pathogenic is clear from left to right.
4.3.4 De novo variants are more pathogenic than inherited variants

A subset of DECIPHER CNVs includes information about how a specific CNV was inherited e.g. inherited from a normal parent or a parent with a similar phenotype, maternally or paternally inherited, and whether a CNV is present constitutively or in mosaic. As de novo mutations have been observed to have increased pathogenicity (Vulto-van Silfhout et al., 2013), we wished to investigate if pathogenicity trends would be clear from classification results. We grouped DECIPHER case CNVs by their inheritance type and classified each group using our CNV pathogenicity classifier. We found that the proportion of variants classified as pathogenic is significantly different across groups by performing a $\chi^2$ test on Table 4.2 ($P < 1 \times 10^{-16}$). CNVs inherited from a normal, unaffected parent are less pathogenic than expected (37.1% pathogenic, standardised residual of -18.5) and that pathogenicity increases for CNVs inherited from a parent with a similar phenotype to the child (64.9%, standardised residual of 6.0; Figure 4.5). CNVs of de novo origin have amongst the highest levels of pathogenicity assigned (de novo constitutive: 83.9%, standardised residual of 45.4; de novo mosaic: 94.6%, standardised residual of 5.0).

4.4 Discussion

This study demonstrates the potential to distinguish between benign and pathogenic human CNVs and subsequently, to utilise this distinction to learn about CNV inheritance. Most dbVar CNVs with clinical interpretations were assigned their benign or pathogenic label by rules including their presence in an affected or unaffected parent and overlap of known syndromes and OMIM disease genes. The ability to accurately differentiate between these groups demonstrated by our CNV
Pathogenic CNV prediction yields inheritance insights

Pathogenicity classifier using only the genomic features listed in Table 4.1 gives insights into the biology of permissible variation. Dosage-balanced ohnologues, mammalian CCN genes and protein interactions largely contribute to informing the difference between benign and pathogenic CNVs. Dosage-balanced ohnologues and mammalian CCN genes have evolutionarily constrained copy number by definition, and both have been observed to be refractory to CNVs, being rare in healthy individuals, but enriched in disease cases. Likely due to this, both features are informative to class separation. Additionally, the number of protein interactions describes how genes are integrated into pathways or regulatory networks to which copy number changes can cause perturbations by altering expression level.

Testing our CNV pathogenicity classifier on two datasets of control and case CNVs as independent validation of performance yields confirmation that our classifier can accurately distinguish between the two groups. Additionally, we observe increasing case CNV pathogenicity with CNV length; a trend that is at least slightly if not much reduced for control CNVs. Minimising any potential bias introduced by duplicate CNVs with identical or similar breakpoints enriches the control CNV dataset for rare variants as they obtain an equal weight to common variation in the dataset. These low frequency CNVs appear to be disproportionately pathogenic, yet we still observe a significant difference in pathogenicity compared to case CNVs at lengths longer than 250 kb. The enrichment of pathogenicity among rare control CNVs might be suggestive of a biological explanation rather than simple classifier error. We would not expect classifier error to be biased in this manner and could be further evidence of rare deleterious variants present in apparent healthy control individuals (Männik et al., 2015).

In Vulto-van Silfhout et al. (2013), the authors find that while the genomes of individuals with mild phenotypes have both de novo and inherited CNVs, genomes yielding more severe phenotypes are enriched for de novo variants. We observe a
complementary trend where DECIPHER \textit{de novo} CNVs are strongly enriched for being classified as pathogenic. Furthermore, CNVs inherited from a parent with a similar phenotype to the child have an intermediate proportion of pathogenicity between \textit{de novo} CNVs and CNVs inherited from a normal parent. We expect that some portion of variants inherited by a child from a parent with the same phenotype may contribute to that phenotype in both individuals and hence are more likely to be pathogenic than variants inherited from a normal parent. Hence we see enriched pathogenicity for these CNVs inherited from a parent with a similar phenotype. Simple genomic features were used here to distinguish between benign and pathogenic variants that in part were defined as such by their inheritance patterns. That we observe clear separation of CNV inheritance types by our pathogenicity classifier, that is unaware of such information, strongly suggests genomic differences between permissible variants.
Chapter 5

Expression quantitative trait loci of dosage-sensitive genes have narrow tissue specificity bias

As part of his final year undergraduate project, Pauric Donnelly, carried out early analysis that formed the basis of this chapter of work.

5.1 Introduction

Expression quantitative trait loci (eQTLs) are genomic regions harbouring sequence variants that influence the expression level of one or more genes (Albert and Kruglyak, 2015). While mapping of eQTLs affecting single genes has been conducted for decades, genome-wide eQTL mapping is about 15 years old (Brem et al., 2002; Schadt et al., 2003). Since then many mapping experiments have been undertaken in various species (Morley et al., 2004; Cheung et al., 2005; Stranger et al., 2005; Stranger et al., 2007; West et al., 2007; Dimas et al., 2009; Kelly et al., 2012; Massouras et al., 2012; GTEx Consortium, 2015). A range of expression effect sizes, both positive and negative, are observed. These eQTL effects can
Tissue specificity bias of dosage-sensitive gene eQTLs occur in a tissue-specific manner or across a number of tissues, however, tissue-specific influence is more typical (Gerrits et al., 2009). In human, the expression of thousands of genes is affected by eQTLs making them a significant contribution to the genetic variation of expression and in turn phenotypic variation and complex disease.

The Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015) has characterised eQTLs across a diverse range of human tissues. In Release V6, 86.5% (15,757/18,208) of protein-coding genes tested had their expression influenced by at least one eQTL. As such a high proportion of the genome experiences this type of expression variation in control individuals, the majority of the genome must be able to tolerate some amount of mRNA level change without obvious deleterious consequences. However, in combination with genome-wide association studies, eQTLs have been used to elucidate further the pathophysiology of many disease phenotypes. To date eQTLs have been associated with human diseases including asthma, autoimmune disorders, diabetes, numerous cancers, Parkinson’s disease, and other brain disorders (see Table 1 in Albert and Kruglyak, 2015). Additionally, eQTLs have been shown to undergo increased purifying selection with gene age where young, primate-specific genes are enriched for eQTLs, having higher effect size and influencing expression in more tissues (Popadin et al., 2014). Therefore, the effect of eQTLs on gene expression and association with important traits makes them worthy of study especially in the context of genes with known expression constraints.

Dosage-sensitive genes have a deleterious phenotype if their dosage is perturbed (Veitia, 2002; Papp et al., 2003a; Birchler et al., 2001; Birchler and Veitia, 2012), and are often seen to be refractory to variation. Ohnologues, paralogues retained after whole genome duplication events are enriched for dosage-sensitive genes. In human, ohnologues produced from polyploidy events at the base of the
vertebrate lineage have been shown to be depleted on control and benign copy number variants (CNVs) but enriched among genes on pathogenic variants (Makino and McLysaght, 2010; Makino et al., 2013; McLysaght et al., 2014). Likewise, mammalian copy number conserved (CCN) genes and haploinsufficient genes are also depleted in benign CNVs and enriched in pathogenic CNVs (Rice and McLysaght, 2017). Constraint on copy number is likely due in part to a constraint on protein product dosage that has reciprocal effects on expression level and copy number. When a variant arises such as a CNV or eQTL that causes a deleterious aberration in expression level, the variant will experience purifying selection and be removed from the population. Therefore we expect that dosage-sensitive genes have fewer eQTLs in healthy individuals.

Ohnologues are also thought to be dosage-balanced, that is, they are constrained in expression relative to their interaction partners (Makino and McLysaght, 2010). For genes under dosage-balance, expression variation of individual genes gives rise to a deleterious stoichiometric imbalance that can lead to incomplete protein complex formation, affect function and be wasteful. The strong requirement for balance has shaped three-dimensional nuclear organisation where genomic regions containing ohnologue pairs are organised with higher spatial proximity (Xie et al., 2016). This nuclear proximity likely ensures similar regulation of gene expression for both genes in the pair. We expect to see distinct eQTLs patterns that reflect the shared constraints of stoichiometric balance between ohnologue pairs.

Here, we investigated the patterns of eQTLs affecting dosage-sensitive genes. Contrary to our expectation that ohnologues and other categories of dosage-sensitive genes should be depleted for this variation, we found that these genes are enriched for eQTLs. However, they have fewer eQTL-affected tissues than dosage-insensitive genes, as the eQTLs that affect these genes are more tissue-specific. Dosage-sensitive genes are depleted for broad tissue breadth eQTLs which are
likely removed by purifying selection as they conflict with expression constraints. We observed that ohnologue pairs have more similar eQTL-affected tissues compared to random ohnologue pairs suggesting a shared constraint between real pairs. This evidence suggests that dosage-sensitivity shapes the evolution of eQTLs influencing the expression of these genes whereby deleterious variants in conflict with constraints experience purifying selection.

5.2 Materials & Methods

5.2.1 Data

eQTLs used in this analysis were significant SNP-gene associations based on permutations obtained from The Genotype-Tissue Expression (GTEx) project V6 (GTEx Consortium, 2015). Protein-coding gene annotations were obtained from Ensembl GRCh37 (Yates et al., 2016). Copy number variant regions were obtained from the inclusive CNV map in Zarrei et al. (2015) and a gene was considered to be intersecting with a region if the any of the gene sequence was overlapped by one or more bases on either strand using Bedtools (Quinlan and Hall, 2010). Ohnologue annotations were obtained from Makino and McLysaght (2010). Haploinsufficient genes were defined as genes with a probability of loss-of-function mutation intolerance > 0.9 (Lek et al., 2016). Genes unaffected by CNVs in nearly 60,000 individuals studied in Ruderfer et al. (2016) were defined as CNV-free genes. Mammalian copy number conserved genes are genes with no copy number changes in 13 mammalian genomes (Rice and McLysaght, 2017).

5.2.2 eQTL enrichment of dosage-sensitive genes

Ohnologues, haploinsufficient genes, CNV-free genes and CCN genes were tested for eQTL enrichment. Genes considered were restricted to those tested for eQTLs
by GTEx i.e. those with at least 6 reads and \( > 0.1 \) reads per kilobase of transcript per million mapped reads (RPKM) in at least 10 individuals.

### 5.2.3 Ohnologue tissue similarity

Jaccard index was calculated between tissues for eQTL-affected ohnologues and nonohnologues using the GeneOverlap R package (Shen and Sinai, 2013).

### 5.2.4 eQTL effect size

eQTL effect sizes were quantified by the slope of the linear regression model used in identifying eQTLs in the GTEx project and represents the effect of the alternative allele on expression relative to the GRCh37/hg19 genome reference allele. The median effect size for each gene in a tissue was calculated and these medians of all ohnologues and nonohnologues were compared. This yielded 30,844 and 63,863 median negative effect sizes for ohnologues and nonohnologues, respectively and 30,622 and 61,444 positive effect sizes for ohnologues and nonohnologues, respectively. Standard deviations of eQTL effect sizes were calculated for genes affected by more than one variant in a tissue, giving 26,223 and 55,918 median negative effect size standard deviations for ohnologues and nonohnologues, respectively and 26,140 and 53,359 positive effect size standard deviations for ohnologues and nonohnologues, respectively.

### 5.2.5 Ohnologue pair analysis

Ohnologue pairs, from Makino and McLysaght (2010), considered were those on different chromosomes and those were both were affected by eQTLs, giving 5,415 pairs. For comparison, an equal number of random ohnologue pairs were generated by randomly sampling two ohnologues with replacement. The number of distinct
Table 5.1 | Genes within CNVRs and eQTL patterns

<table>
<thead>
<tr>
<th>Genes affected by both positive &amp; negative eQTLs</th>
<th>Genes affected by negative eQTLs only</th>
<th>Genes affected by positive eQTLs only</th>
<th>eQTL-free</th>
<th>% total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>std. res.</td>
<td>#</td>
<td>std. res.</td>
<td>#</td>
</tr>
<tr>
<td>CNVR genes</td>
<td>984</td>
<td>5.2</td>
<td>84</td>
<td>-3.4</td>
</tr>
<tr>
<td>CNGR genes only</td>
<td>431</td>
<td>-0.5</td>
<td>63</td>
<td>0.8</td>
</tr>
<tr>
<td>CNVL genes only</td>
<td>3790</td>
<td>12.5</td>
<td>367</td>
<td>-4.8</td>
</tr>
<tr>
<td>Genes outside CNVRs</td>
<td>7361</td>
<td>-14.1</td>
<td>1130</td>
<td>5.9</td>
</tr>
<tr>
<td>% total</td>
<td>69.0%</td>
<td>9.0%</td>
<td>8.5%</td>
<td>13.5%</td>
</tr>
</tbody>
</table>

tissues affected by eQTLs for both genes within a pair (i.e. the union of eQTL affected tissues) was calculated for both the real and random ohnologue pairs.

5.2.6 Multi-gene eQTLs

Multi-gene eQTLs were defined as variants that significantly affect the expression of two or more genes irrespective of tissue.

5.3 Results

5.3.1 eQTL enrichment of CNVR genes and dosage sensitive genes

Genes within benign copy number variant regions (CNVRs) tolerate expression variation through copy number change without strong deleterious consequences thus we expect other kinds of variants that alter mRNA level to co-occur. One such type of variation is eQTLs. We obtained human eQTL data affecting the expression of 15,757 protein-coding genes across 44 tissues from The Genotype-
Figure 5.1 | eQTL enrichment of CNVR genes and dosage sensitive genes.
A. Proportion of genes affected and not affected by eQTLs for two sets of CNVs: Zarrei et al. CNV map and ExAC CNV data; ohnologues, haploinsufficient genes and mammalian copy number conserved (CCN) genes. P-values shown above each plot are Bonferroni-adjusted. B. Proportion of ohnologues (O) and nonohnologues (N) affected by eQTLs per tissue. C. Pairwise overlap between eQTL-affected genes between tissues measured by Jaccard index. Left heatmap: Brain tissues; right heatmap: non-brain tissues. Upper triangle: Pairwise overlap of ohnologues; Lower triangle: Pairwise overlap of nonohnologues.
Tissue Expression (GTEx) project (GTEx Consortium, 2015). As expected we found that more protein-coding genes, 89.6% (6,203/6,924), within CNVRs of a human control CNV map (Zarrei et al., 2015) were affected by eQTLs compared to 84.7% (9,554/11,284) of genes outside of CNVRs ($P < 1 \times 10^{-16}$, $\chi^2$ test, Figure 5.1A). We see similar eQTL enrichment of CNV-affected genes in data for 59,898 exomes from The Exome Aggregation Consortium (ExAC) (Ruderfer et al., 2016). Here, 88.5% of genes affected by CNVs are also influenced by eQTLs compared to 85.3% of genes without CNVs ($P = 0.0004$, $\chi^2$ test; Figure 5.1A).

Genes that tolerate copy number gain regions (CNGRs) might be more likely to have positive eQTLs and those within copy number losses regions (CNLRs) to have negative eQTLs. To check if CNV direction, gain or loss, has a relationship with positive or negative eQTLs direction, we categorised genes by occurrence in control CNVR map and whether they are affected by positive or negative eQTLs. Genes in regions that experience both copy number gains (CNGs) and copy number losses (CNLs) were grouped as CNVR genes, genes only in gain regions were grouped as CNGR genes and genes only in loss regions were grouped as CNLR genes. We examined these groups, along with genes outside of CNVRs, for enrichment of being affected by positive and negative eQTLs. We found that genes outside of CNVRs are strongly deficient in having both positive (increased expression) and negative (decreased expression) effect eQTLs (standard residuals: -14.1, $P < 1 \times 10^{-16}$, $\chi^2$ test, Table 5.1). However, we did not see a trend matching CNV and eQTL direction when we considered positive only or negative only eQTLs and genes within CNGRs or CNLRs on the CNV map.

As dosage-sensitive genes such as ohnologues, mammalian (CCN) genes, and haploinsufficient genes, are refractory to benign CNVs and enriched among genes found on pathogenic CNVs, we expected them to be depleted for eQTLs as expression variation will likely be deleterious. However, we find the opposite to be the
Table 5.2 | Haploinsufficient genes and eQTL direction

<table>
<thead>
<tr>
<th>Genes affected by both positive &amp; negative eQTLs</th>
<th>Genes affected by negative eQTLs only</th>
<th>Genes affected by positive eQTLs only</th>
<th>eQTL-free</th>
<th>% total</th>
</tr>
</thead>
<tbody>
<tr>
<td># std. res. 2030</td>
<td># std. res. 329</td>
<td># std. res. 312</td>
<td># std. res. 315</td>
<td>-3.2</td>
</tr>
<tr>
<td>Haploinsufficient</td>
<td>Haplosufficient</td>
<td>% total 70.3%</td>
<td>8.9%</td>
<td>8.5%</td>
</tr>
<tr>
<td>% total 9877</td>
<td>% total 1175</td>
<td>% total 1123</td>
<td>% total 1767</td>
<td>3.2</td>
</tr>
</tbody>
</table>
| case. We observed that ohnologues are enriched for eQTLs relative to nonohnologues (89.6% affected by eQTLs vs. 84.8%, respectively; $P < 1 \times 10^{-16}$, $\chi^2$ test; Figure 5.1A). Haploinsufficient genes are also found to be slightly enriched for eQTLs (89.5% affected by eQTLs vs. 87.3% for haplosufficient genes, respectively; $P = 0.007$, $\chi^2$ test; Figure 5.1A). Similarly, for CCN genes we found that they were enriched for eQTLs relative to genes with mammalian gene duplication and loss events (89.5% affected by eQTLs vs. 76.6%, respectively; $P < 1 \times 10^{-16}$, $\chi^2$ test; Figure 5.1A). Additionally, we see that haploinsufficient genes are depleted for being affected by both positive and negative eQTLs and so are enriched for being affected by either negative or positive eQTLs only ($P = 3 \times 10^{-10}$, $\chi^2$ test, Table 5.2). However, we do not see a directional bias for haploinsufficient genes being strongly depleted for negative eQTLs only.

5.3.2 Ohnologues have more similar eQTL patterns in brain tissues

We investigated the proportion of genes affected by eQTLs for all genes expressed per tissue (Figure 5.1B). We observe that ohnologues have a lower proportion of eQTL-affected genes within each tissue. Given that the trend per tissue is the
opposite to the trend observed irrespective of tissue, we examined the possibility that different ohnologues are affected by eQTLs in different tissues. For example, if one ohnologue is affected by eQTLs in tissue A and another ohnologue is affected by eQTLs in tissue B, while two nonohnologues are affected by eQTLs in both tissue A and tissue B. If different ohnologues are influenced by eQTLs in different tissues that may explain having fewer eQTLs within a tissue but being more affected by eQTLs overall compared to nonohnologues. If cumulatively across tissues more ohnologues are affected than nonohnologues this account for the tissue-specific and genome-wide trends we observe.

The Jaccard index is a measure of similarity between sets and is the size of the intersection divided by the size of the union of the sets. If eQTL-affected ohnologues are more distinct between tissues compared to eQTL-affected nonohnologues then we expect a lower Jaccard index between sets of ohnologues (i.e. less similar). When we look at the Jaccard index for eQTL affected genes between GTEx tissues, the most striking difference initially is that the two cell lines of transformed lymphocytes and fibroblasts have the most distinct eQTL-affected genes for both ohnologues and nonohnologues (Figure 5.1C). Considering all tissues however, we do not see a difference between the sharing of eQTL-affected genes for ohnologues and nonohnologues (median Jaccard index 0.936 for both; \( P = 0.7 \), Mann-Whitney U test). Despite this, for all 45 brain tissue pairwise comparisons we see consistently higher similarity between eQTL-affected ohnologues than between eQTL-affected nonohnologues (median Jaccard index 0.973 vs. 0.965, respectively; \( P = 0.004 \), Mann-Whitney U test; Figure 5.1C, left). Excluding brain tissues, there is still no significant difference between ohnologues and nonohnologue similarity for non-brain tissues (median Jaccard index 0.934 for both; \( P = 0.8 \), Mann-Whitney U test; Figure 5.1C, right). Higher similarity among ohnologues for eQTL patterns in brain tissues might be suggestive of
more similar gene regulation wherein the same genes can or cannot evolve eQTLs influencing their expression.

5.3.3 eQTLs affecting dosage-sensitive genes have been shaped by selection

Dosage-sensitive genes are under various constraints (such as haploinsufficiency, aggregation-susceptibility, concentration-dependency, and dosage-balance) that restrict their permissible copy number. CNVs will cause the expression of genes they contain to change across tissues which can be permissible in cases where the expression change is compatible with the constraint (e.g. a copy number gain of a gene that is haploinsufficient). However, an incompatible CNV in conflict with an expression constraint can produce a deleterious phenotype and will then experience purifying selection. eQTLs, on the other hand, can influence the expression of genes across a broad range of tissues or within only a single tissue. Narrow tissue breadth eQTLs can often avoid conflicting with a constraint that exists within one or a few tissues and neutrally evolve, drifting in population frequency. eQTLs that affect a dosage-sensitive gene’s expression may arise if: first, the constraint is not present in all tissues; second, the eQTL affects expression only in a subset of unconstrained tissues; and third, the altered expression is not deleterious for reasons other than the preexisting constraint. For these reasons, we expect eQTLs that influence the expression of dosage-sensitive genes are biased towards those that affect fewer tissues. We find this to be the case.

We observe that ohnologues have a lower proportion of eQTL affected tissues where they are expressed than nonohnologues (median % tissues affected by eQTLs: 11.4% vs. 13.6%, respectively; \( P < 1 \times 10^{-16} \), Mann-Whitney U test; Figure 5.2A). This difference holds true even when genes without eQTLs are included (median % tissues affected by eQTLs: 9.3% for ohnologues vs. 10.5% for
Figure 5.2 | eQTL tissue specificity of dosage-sensitive genes.  
A, proportion of tissues where a gene is expressed that are affected by eQTLs. B, proportion of dosage-sensitive genes per number of tissues affected by eQTLs. C, number of eQTLs affecting the expression of dosage-sensitive genes. D, number of tissues affected per eQTL for dosage-sensitive genes. E, proportion of genes affected by broad tissue breadth eQTLs (influencing expression in more than 10 tissues) that are ohnologues, haploinsufficient genes or CCN genes.
nonohnologues, respectively; \( P = 0.003 \), Mann-Whitney U test). We find the same trend for haploinsufficient genes, with a bias for narrower tissue breadth (median % tissues affected by eQTLs: 9.1% vs. 13.6%, respectively; \( P < 1 \times 10^{-16} \), Mann-Whitney U test;) and a moderate difference between mammalian CCN genes and genes with duplication and loss events (median % tissues affected by eQTLs: 11.4% vs. 13.6%, respectively; \( P = 3.0 \times 10^{-6} \), Mann-Whitney U test; Figure 5.2A).

When genes are grouped by the number of tissues affected by eQTLs, we can see that the proportion of genes that are dosage-sensitive are skewed towards fewer tissues affected (Figure 5.2B).

The absolute number of eQTLs affecting a dosage-sensitive gene could also be reduced if purifying selection is acting on eQTLs removing deleterious variants that conflict with expression constraints. For genes that are affected by at least one eQTL, ohnologues have fewer eQTLs per gene compared to nonohnologues (median eQTLs: 87 vs. 96; \( P = 0.01 \), Mann-Whitney U test; Figure 5.2C). This is also strongly the case for haploinsufficient genes (median eQTLs: 68 vs. 102; \( P < 1 \times 10^{-16} \), Mann-Whitney U test; Figure 5.2C). However in contrast to ohnologues and haploinsufficient genes, we found that CCN genes have slightly more eQTLs affecting their expression per gene compared to genes with mammalian copy number gain and loss events (median eQTLs: 98 vs. 91; \( P = 0.01 \), Mann-Whitney U test).

eQTLs influencing gene expression in a broad range of tissues are more likely to come into conflict with constraints in one or more tissues and experience purifying selection. Therefore we expect to find a narrower breadth of tissues affected per eQTL influencing a dosage-sensitive gene. We find that eQTLs influencing ohnologues affect their expression in fewer tissues per eQTL than their nonohnologue counterparts (median tissues affected: 1 per ohnologue eQTL, 2 per nonohnologue eQTL; \( P < 1 \times 10^{-16} \), Mann-Whitney U test; Figure 5.2D). We find this applies to
both haploinsufficient genes and CCN genes also (median tissues affected: 1 per ohnologue eQTL, 2 per nonohnologue eQTL; \( P < 1 \times 10^{-16}, \chi^2 \text{ test}; \) Figure 5.2D). Furthermore we find that eQTLs that influence expression in more than ten tissues (broad tissue breadth eQTLs) are depleted amongst dosage-sensitive genes. For ohnologues, 9.3% (627/6,727) are affected by broad eQTLs compared to 15.4% (1,767/11,481) of nonohnologues \( (P < 1 \times 10^{-16}, \chi^2 \text{ test}; \) Figure 5.2E). We see a similarly strong depletion for haploinsufficient genes, (7.0% vs. 14.7% of haplosufficient genes; \( P < 1 \times 10^{-16}, \chi^2 \text{ test} \)) but only a moderate depletion of CCN genes (12.4% vs. 13.8% of genes with copy number change events; \( P = 0.01, \chi^2 \text{ test} \)).

5.3.4 Variance and size of eQTL effects is reduced for ohnologues

The amount of influence an eQTL has on a gene’s expression level varies; some eQTLs only moderately increase or decrease mRNA level, while others have large effects. The direction and size of eQTL effects are quantified by the slope of the linear regression model used in identifying eQTLs in the GTEx project and represents the effect of the alternative allele relative to the GRCh37/hg19 genome reference allele. The median eQTL effect size for each gene in a tissue was calculated and we see that ohnologues are affected by eQTLs with a smaller effect size than nonohnologues for both positive and negative effect eQTLs (median positive effect size for genes per tissue: 0.39 for ohnologues vs 0.41 for nonohnologues; \( P < 1 \times 10^{-16}; \) median negative effect size: -0.38 for ohnologues vs -0.41 for nonohnologues; \( P < 1 \times 10^{-16}, \text{Mann-Whitney U test} \)). Additionally, the effect size of eQTLs that influence ohnologue expression have decreased variance in comparison to eQTL effect sizes of nonohnologue eQTLs (median positive effect standard deviation: 0.035 for ohnologues vs 0.039 for nonohnologues; \( P < 1 \times 10^{-16}; \) median negative effect standard deviation: 0.033 for ohnologues vs 0.037 for nonohnologues;
\[ P < 1 \times 10^{-16}, \text{Mann-Whitney U test} \). Therefore when ohnologue expression is affected by eQTLs, it is moderately less perturbed both in size and variability.

### 5.3.5 Ancient constraint rather than gene age predicts eQTL patterns

Popadin et al. (2014) find evidence of increased purifying selection acting on older genes preferentially. They propose that integration into interaction networks, involvement in regulatory networks, and haploinsufficient are constraints on expression that increase with gene age and so increase selection forces on eQTLs. Using the same method of defining gene age (Zhang et al., 2010), we test ohnologues and nonohnologues at the oldest node, node 0, for the proportion of expression tissues affected by eQTLs. The age of this node is prior to vertebrate radiation and the majority of ohnologues (80.6%) map here. We still find that ohnologues have a lower proportion of eQTL affected tissues where they are expressed than nonohnologues (median % tissues affected by eQTLs: 9.1\% vs. 11.4\%, respectively; \[ P = 1.9 \times 10^{-12}, \text{Mann-Whitney U test} \]). This difference holds true when genes without eQTLs are excluded (median % tissues affected by eQTLs: 11.4\% for ohnologues vs. 13.6\% for nonohnologues, respectively; \[ P = 2.7 \times 10^{-12}, \text{Mann-Whitney U test} \]). While gene age may be a factor in selection acting on eQTLs, controlling for age ohnologues are still more constrained than nonohnologues.

### 5.3.6 Dosage-insensitive gene eQTLs patterns are constrained when eQTLs are shared

An eQTL can be associated with influencing the expression of more than one gene. While the majority of eQTLs identified in the GTEx project affect only one gene’s expression, many affect more than one, and a handful of variants influence
Figure 5.3 | Shared constraints shaping eQTL patterns.
A, eQTLs affecting two, three and four genes. Dark grey diamonds and the corresponding number show the mean for each group. Bonferroni-adjusted P-values for pairwise Mann Whitney U tests are displayed above each boxplot along with sample sizes for each group. B, the union of tissues affected by eQTLs was calculated for 5,415 ohnologue pairs and for an equal number of random pairs generated by randomly sampling two ohnologues with replacement.
the expression of up to 19 genes. As some eQTLs affect both dosage-sensitive and dosage-insensitive genes, we wished to investigate how the inclusion of dosage-sensitive genes shaped eQTL patterns. One possibility is that the inclusion of any dosage-sensitive gene in the group of affected genes sharply constrains the number of tissues where expression is affected for all genes influenced. An alternative to this is where with higher proportion of dosage-sensitive genes in the group of genes affected the more constrained the eQTL becomes for all genes, in a gradient-like manner. Additionally, the same eQTL can affect expression in different numbers of tissues for different genes. Given this, it is possible that eQTL constraint exclusively affects dosage-sensitive genes, leaving patterns for dosage-insensitive genes to evolve unencumbered. We used eQTLs affecting two, three and four genes to distinguish between these possibilities (Figure 5.3A).

We observe multi-gene eQTLs that exclusively influence expression of nonohnologues affect more tissues per eQTL than multi-gene eQTLs that instead exclusively affect ohnologues (median for two gene eQTLs: two tissues vs. one, respectively; \( P < 1 \times 10^{-16} \), Mann-Whitney U test; Figure 5.3A). Additionally, we see intermediate constraints for gene groups that include both dosage-sensitive and dosage-insensitive genes. The number of tissues affected per eQTL for nonohnologues decreases when the eQTL also influences the expression of at least one ohnologue. Conversely, for ohnologues the number of tissues affected per eQTL increases when the eQTL also affects at least one nonohnologue. While eQTLs influencing dosage-insensitive genes affect fewer tissues when the same eQTL also influences a dosage-sensitive gene, we also observe a difference in the number of tissues affected for ohnologues and nonohnologues influenced by the same eQTL. In this case, the number of altered tissues is higher for dosage-insensitive genes than their dosage-sensitive partner (median for two gene eQTLs: two tissues affected for nonohnologues vs. one tissue for ohnologues; \( P < 1 \times 10^{-16} \), Mann-Whitney
Tissue specificity bias of dosage-sensitive gene eQTLs

Therefore multi-gene eQTLs have a gradient-like constraint with mixed eQTLs being intermediate between eQTLs exclusively affecting the expression of dosage-sensitive genes and those affecting unconstrained genes. Furthermore, the influence of eQTLs is not equal within a mixed gene group where dosage-sensitive genes demonstrate additional expression constraint.

5.3.7 Ohnologue pairs demonstrate dosage-balance with mutually absent eQTLs

Genes within ohnologue pairs are dosage-balanced with respect to their counterpart, so this shared constraint will likely affect their eQTL patterns. Genes within pairs should have a shared absence of eQTLs detected for at least one common tissue. eQTLs can arise in other tissues and drift in frequency neutrally. Therefore if we look at all the tissues affected by eQTLs for both genes in a pair we should see signals of selection removing eQTLs for tissues with a shared constraint. We had 5,415 ohnologue pairs that met the criteria of each being on different chromosomes and both being affected by eQTLs. For comparison, we generated an equal number of random pairs by randomly sampling two ohnologues with replacement. The union of tissues affected by eQTLs for both genes in a pair was calculated for both actual and random pairs. As expected actual ohnologue pairs have fewer combined tissues affected by eQTLs (median: 9 tissues affected by eQTLs) compared to random pairs (median: 10 tissues affected by eQTLs; \( P = 0.008 \), Mann-Whitney U test; Figure 5.3B).

5.4 Discussion

This study extends our knowledge of the permissible variation of dosage-sensitive genes. Here we found that dosage-sensitive genes are enriched for eQTLs, contrary
Tissue specificity bias of dosage-sensitive gene eQTLs

Figure 5.4 | Selection patterns for eQTL influencing expression of dosage-sensitive genes.

Scenarios of eQTLs affecting the expression of dosage-sensitive genes and the selective pressure they experience. **Scenario 1**, a hypothetical ohnologue that has constrained expression in a tissue, A, and is not affected by any eQTLs. **Scenario 2**, an eQTL affecting the expression of an ohnologue in unconstrained tissues, evolving neutrally neither under positive or purifying selection. **Scenario 3**, an eQTL under purifying selection as its effect on expression comes in conflict with a tissue expression constraint (see Figure 5.2). **Scenario 4**, an eQTL that is positively selected as it does not adjust expression in a constrained tissue but allows the ohnologue to reach a different expression level that is optimum for another tissue. **Scenario 5 and 6**, adjacent ohnologue and nonohnologue both affected by the same eQTL (see Figure 5.3A). In **scenario 6** the eQTL violates expression constraints for the ohnologue in tissue B and so experiences purifying selection. The removal of this variant from the population will shaped eQTL patterns for both the affected ohnologue and the nonohnologue. **Scenario 7**, an ohnologue pair, consisting of ohnologue and ohnologue’ in different loci, that share expression constraint in tissue B (see Figure 5.3B). eQTLs that change expression level beyond the permissible range in tissue B for either ohnologue will experience purifying selection.
to our expectations. While these constrained genes have previously been shown to be refractory to other forms of variation, we show here that eQTLs influencing the expression of dosage-sensitive genes are tolerable when they influence expression in unconstrained tissues. When variants arise that alter the expression of neighbouring dosage-sensitive genes a number of outcomes can result (Figure 5.4, scenarios 2-4). If the eQTL changes expression only in tissues that are free to vary the sequence variant can neutrally evolve, without selective pressure. Narrow tissue breadth eQTLs are less likely to come in conflict with an expression constraint that exists in one or several tissues compared to broad tissue breadth eQTLs. Due to this many tissue specific eQTLs can arise and evolve neutrally. In contrast to this, eQTLs with a broad tissue breadth are more likely to give rise to deleterious expression levels that will be selected against and removed from the population. Therefore, we see dosage-sensitive genes are depleted for broad tissue breadth eQTLs and so the eQTLs remaining that influence their expression are biased towards having narrow tissue specificity. Additionally, eQTLs can allow dosage-sensitive genes to circumvent constraints and reach different expression levels in other tissues that are more optimal. It may be possible to see signatures of positive selection for some variants influencing the expression of dosage-sensitive genes by investigating high-frequency derived alleles, long haplotypes and reduced heterozygosity around variants (Sabeti et al., 2002; Nielsen, 2005; Voight et al., 2006; Sabeti et al., 2007).

We investigated two types of shared expression constraints that influence eQTL patterns of dosage-sensitive genes: multi-gene eQTLs and ohnologue pairs (Figure 5.4, scenarios 5-7). We found that multi-gene eQTLs have graduated constraint, increasing with larger proportions of dosage-sensitive genes affected. While most eQTLs identified by GTEx affect a single gene, the inclusion of dosage-sensitive genes as part the group of genes influenced by multi-gene eQTLs has conse-
quences on the variation affecting dosage-insensitive genes. Additionally, while multi-gene eQTLs that affect a mix of ohnologues and nonohnologues display intermediate constraint, ohnologues are still disproportionately more constrained than nonohnologues. This result highlights that permissible variation avoids expression constraints and often consists of random neutrally evolving variants. Here, multi-gene eQTLs that affect both ohnologues and nonohnologues do not influence ohnologue expression in constrained tissues that conflicts with restrictions. Additionally, these eQTLs can alter expression of nonohnologues randomly, neutrally evolving where expression restrictions do not exist. A natural limit on the amount of discordance for multi-gene eQTLs likely exists that places an upper bound on the number of tissues affected for nonohnologues when the number of tissues affected for ohnologues is limited to some degree.

Human ohnologues have been retained since the vertebrate whole genome duplication events approximately 500 million years ago (McLysaght et al., 2002; Dehal and Boore, 2005). It has been suggested that for ohnologues to persist for this length of time, their absence must be deleterious and indeed copy number variation of ohnologues has been association with disease (Makino and McLysaght, 2010). The requirement for relative stoichiometric balance of ohnologue expression has manifested in part as higher spatial proximity of ohnologue pairs within the nucleus to ensure similar gene regulation (Xie et al., 2016). eQTLs have the potential to disturb an ohnologue’s expression level and cause imbalance relative to its partner hence we found that ohnologue pairs show fewer combined eQTL-affected tissues. This provides further evidence that some ohnologue pairs are under strong constraint for similar expression regulation.

Similar eQTL trends presented here should be observed in other mammalian genomes as dosage-sensitive genes are likely under a persistent ancient constraint that limits permissible expression variation patterns across genomes (Makino and
Tissue specificity bias of dosage-sensitive gene eQTLs

McLysaght, 2010; Rice and McLysaght, 2017). Additionally, if orthologous genes are functionally conserved across species, similar expression variance constraints may be witnessed in the same tissues between species.

Our results are largely consistent with previous work that demonstrated that eQTLs affecting young genes were under less constraint (Popadin et al., 2014). Similarly, we found ohnologues, a subset of older genes, have more tissue specific eQTLs and their eQTL effect sizes are smaller. However, notably we show controlling for gene age, ohnologues still demonstrate fewer tissues affected by eQTLs compared to nonohnologues. Therefore we propose that constraints on gene dosage are a more important factor and that these constrains are ancient. An additional contrast with Popadin et al. (2014) is overall we see that the expression of more ohnologues are affected by eQTLs. This could be due to increased tissues and samples tested as part of the GTEx project data analysed here. It is likely that the categories of dosage-sensitive genes used here are predominantly older genes (by definition ohnologues and mammalian CCN genes must predate the vertebrate and mammalian divergences, respectively) however we show that age cannot solely explain the patterns we observe, especially for shared eQTL patterns seen within actual ohnologue pairs.
Chapter 6

Conclusions

Intricate biological processes occur within the cells of all organisms. At first glance, these processes may seem chaotic but are often stringently controlled at many levels. Evolutionary processes have refined interactions and regulatory mechanisms to perform efficiently in their respective environments. However, variants, such as SNVs and CNVs, arise and threaten to perturb these finely tuned cellular activities. Which genes can tolerate such variants, why other genes cannot, and the role these variants play in the evolution of our genome are important questions to answer. In this thesis, the three studies presented attempt to answer, at least in part, these questions.

Permissible variation can arise across much of the genome, often without any phenotype at all but for some regions, a deleterious phenotype results. In chapter 3, I explored the prevailing hypothesis that CNV pathogenicity is caused by dosage-sensitivity of the genes contained within the CNV and devised a novel test for this hypothesis using evolutionary metrics. I looked for signatures of natural selection across mammalian genomes and found that the absence of gene gain and loss events aligns well with human pathogenic variants. Conversely, genes in benign CNVRs have more variable copy number. These evolutionary constraints are
Conclusions

characteristic of genes on pathogenic CNVs and can only be explained by dosage-sensitivity of those genes. These results implicate dosage-sensitivity of individual genes as a common cause of CNV pathogenicity and suggest that they cannot tolerate variation without giving rise to a deleterious phenotype.

Variants that yield no phenotype, or a phenotype that is neutral, are at the whim of genetic drift and can randomly reach fixation or loss in the population. If a variant is slightly advantageous upon duplication, its chance of becoming fixed increases as it may experience some positive selective force. Conversely, a variant that is disadvantageous can impact fecundity, and is less likely to reach fixation and more likely to be lost.

In many instances, dosage-sensitivity is not likely to be a recent constraint. Ancestral dosage-sensitivity is one possible explanation for duplicate retention after WGD events. Persistence of that sensitivity to the present day is an explanation for the enrichment of dosage-sensitive genes for involvement in disease. In this thesis, I show that this constraint is apparent across mammalian genomes whereby constraint on dosage-sensitive genes to maintain their gene product level within a permissible threshold is pervasive and universal across mammals. Human CNV pathogenicity is predictive of copy number constraint across mammalian genomes, with the most distantly related genomes compared here sharing a common ancestor ~100 Mya (Hedges et al., 2015). Furthermore, I observe that mouse orthologues of human copy number conserved genes are depleted among mouse CNVRs, lending further support to the idea that these are conserved ancestral copy number constraints among mammalian genomes. This work provides evidence that constraint on dosage is ancient, dating back to before the vertebrate WGD events and is still persistent and common among extant mammals.

The vast majority of CNVs with clinical interpretations used here were assigned such labels due to patterns of inheritance and aspects of genic content (e.g.
Conclusions

presence of a corresponding OMIM disease gene). While this likely introduces some genic and study bias, we do not expect the same of our evolutionary metrics which are independent of disease annotation or study bias. Ideally, more CNVs in the dataset would be labeled as pathogenic due to statistical association with cases over control individuals and the results presented here would now suggest manual curation of such variants from the literature to repeat a similar analysis may be a worthwhile undertaking.

Additionally, with an improved human genome assembly and additional recently released and upcoming high quality mammalian and vertebrate genome sequences, this analysis could be repeated with increased resolution. Our rudimentary evolutionary metric of mammalian copy number conservation disregards evolutionary information, specifically the phylogenetic distance between the 13 mammalian species and human. The metric in its current basic form clearly yields valuable insights, but weighting copy number changes of the 13 genomes with respect to distance to human could highlight more relevant copy number changes. A weighted metric would make a copy number change in chimp or gorilla more significant compared to a similar change in cow or horse. While we provide evidence for constraint on dosage-sensitive genes being ancient and persistent, we also have the example of \textit{TBX1} that is essential in human and mouse but lost in even-toed ungulates. The potentially diminished importance of this essential gene by our metric due to a probable single loss event in an ungulate common ancestor could be somewhat reversed with a weighted score. We propose that a phylogenetically weighted metric is worthy of investigation in the prioritisation of candidate genes.

If a variant is sufficiently deleterious as to affect fecundity or severely impact fitness, this variant will be less inherited and more rare in segregating variation. In chapter 4, we investigated the relationship between CNV pathogenicity and
Conclusions

acquisition route. Initially, we confirmed and utilised the genic difference between benign and pathogenic CNVs to accurately predict CNV pathogenicity. Evolutionarily copy number conserved genes, specifically dosage-balanced ohnologues and mammalian CCN genes, and protein interactions are highly informative for differentiating benign and pathogenic variants. These features provide evidence, consistent with previous findings, for which kinds of genic content can tolerate CNVs without deleterious consequences and which cannot. Further, we found a strong relationship between CNV inheritance and purifying selection filtering pathogenic variants.

*De novo* variants should be relatively untouched by natural selection, only being required to be compatible with life for observation and detection. Inherited variants, on the other hand, should have at least one additional requirement of not being sufficiently deleterious to entirely reduce survival and fecundity. Therefore, inherited variants are a subset of *de novo* CNVs, after some pathogenic variation has been removed by purifying selection. Our results confirm this to be the case with lower proportion of pathogenicity found for inherited variants. This observation supports previous clinical findings of enrichment for *de novo* variation among variants of case individuals.

We expect that some portion of variants inherited by a child from a parent with the same phenotype may contribute to that phenotype in both individuals. In these cases more of these inherited CNVs are likely pathogenic than variants inherited from a normal parent as a causative variant is passed on. We do not have the same expectation for CNVs inherited from a healthy parent. Perhaps some of these variants are pathogenic in offspring but not in the parent due to increased penetrance, but this scenario is not specific to healthy parents. Our observation of a gradient of increasing pathogenicity from variants inherited from a healthy parent, those inherited from a parent with a similar disorder, to CNVs of *de novo*
Conclusions

origin, clearly implicates filtering by purifying selection and its defining role in permissible variation.

Again, genic and study biases within the initial dataset of dbVar CNVs used for training probably have an impact on the performance of our classifier, although it performs well on independent datasets regardless. Quality of training data is paramount in machine learning and any efforts to improve this, e.g. by manual curation of variants in the literature, could only be beneficial. While the phenotypes yielded from CNVs can vary greatly, the underlying general mechanisms of dosage-sensitivity at the cellular level are probably more common. As we show that genic features can predict pathogenicity that arises from CNVs, it may be possible to create either a variant- or gene-level classifier that can predict dosage-sensitivity upon copy number change in a species-independent fashion. Given universal evolutionary metrics like ohnologue status, and standardised protein interaction information (relative node degree distributions) and other features, mechanisms of dosage-sensitivity could be predicted.

Finally in chapter 5, we tested the expectation that dosage-sensitive genes are also refractory to dosage changes by evolution of gene expression. We explored expression evolution of dosage-sensitive genes by comparing expression variation, in the form of eQTLs, of dosage-constrained genes with other genes in healthy individuals. Here, we found dosage-constrained genes are biased towards having eQTLs of narrow tissue specificity, suggesting that broad tissue breadth eQTLs are likely removed by purifying selection due to conflicts with expression constraints. Dosage-sensitivity shapes the evolution of these genes by restricting them to a narrow route of evolution through expression changes in unconstrained tissues. Additionally, we show that this restriction is likely ancient and is consistent with constraints present at the vertebrate WGD events.

The effects on dosage by gene duplication and expression variation are not
necessarily the same, and thus the constraints on permissible variation can be different. While a CNV has a global effect, altering dosage in every tissue, eQTLs can have more precise influence allowing avoidance of specific constraints for dosage-sensitive genes and consequently a tolerable path to vary and evolve. In this way, dosage-sensitive genes are not static and incapable of evolving, however, any change must be compatible with dosage constraints to avoid yielding a deleterious phenotype. Indeed this can impact variation on genes without constraints, where eQTLs in a region affect multiple neighbouring genes, one of which has a dosage constraint. Purifying selection will act on the multi-gene eQTL to remove it from the population so in effect, a neighbouring gene without constraints will have similarities in its variation profile to the dosage-sensitive gene. While the majority of eQTLs in the dataset here are gene-specific (~two-thirds), constraints radiating from neighbouring dosage-sensitive genes muddle expectations of permissible variation without considering wider genomic context.

As eQTLs provide a potential avenue for dosage-sensitive genes to circumvent constraints and optimise expression levels in other tissues, it may be possible to see signatures of positive selection on some variants. We propose searching for high-frequency derived alleles, long haplotypes and reduced heterozygosity around variants as a test for this hypothesis. Additionally, as the constraint in place on some human genes is ancient, we expect that similar eQTL trends presented here should be observed in other mammalian genomes. Further, if orthologous genes are functionally conserved across species, similar dosage constraints may be witnessed in the same tissues between species elucidating their functional constraints. Examination of mouse eQTL data may confirm or reject these expectations, although the data available may not be as extensive as human. A limitation of the work here is that our analysis only consists of healthy control expression variation. To fully confirm purifying selection acting on broad tissue breath eQTLs that conflict
Conclusions

with gene dosage constraints examination of pathogenic variants will be necessary.

Here, we performed the first comparison of genome evolutionary trends of genes in benign and pathogenic CNVs and provide evidence for the usefulness of evolutionary metrics in the identification of candidate disease genes. Further, we show their use in the interpretation of CNVs both as a clinical tool and a method of gaining insights into genomic trends of permissible variation. Additionally, we provide evidence that expression evolution by tissue-specific variation is almost the only route available to evolve gene product levels for dosage-sensitive genes. For these genes we expect at least one tissue with reduced expression variation potentially revealing which tissues have constrained gene expression, and thus providing greater insight into the basis of pathogenicity when these genes are affected by a CNV.
Bibliography


Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz (2012). “Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs”. In: *PLoS Computational Biology* 8.5, e1002514. doi: 10.1371/journal.pcbi.1002514.


Bersaglieri, Todd, Pardis C. Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F. Schaffner, Jared A. Drake, Matthew Rhodes, David E. Reich, and Joel N. Hirschhorn (2004). “Genetic signatures of strong recent positive selection at the lactase gene”. In: *American Journal of Human Genetics* 74.6, pp. 1111–1120. DOI: 10.1086/421051.


Chen, Xiaoshu and Jianzhi Zhang (2012). “The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data”. In: PLoS Computational Biology 8.11, e1002784. DOI: 10.1371/journal.pcbi.1002784.

of chromosomal abnormalities in pregnancy losses reveals critical genes and loci for human early development”. In: *Human Mutation* 38.6, pp. 669–677. DOI: 10.1002/humu.23207.


Rehder, Blake C. Ballif, Lisa G. Shaffer, and Evan E. Eichler (2011). “A copy number variation morbidity map of developmental delay”. In: Nature Genetics 43.9, pp. 838–846. doi: 10.1038/ng.909.

Craddock, Nick et al. (2010). “Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls”. In: Nature 464.7289, pp. 713–720. doi: 10.1038/nature08979.


Falchi, Mario, Julia Sarah El-Sayed Moustafa, Petros Takousis, Francesco Pesce, Amélie Bonnefond, Johanna C Andersson-Assarsson, Peter H Sudmant, Rajku-


American Journal of Human Genetics 66.5, pp. 1669–1679. DOI: 10.1086/302879.


years of echinoderm evolution.” In: Proceedings of the National Academy of Sciences of the United States of America 100.23, pp. 13356–13361. DOI: 10.1073/pnas.2235868100.


Kaminsky, Erin B., Vineith Kaul, Justin Paschall, Deanna M. Church, Brian Bunke, Dawn Kunig, Daniel Moreno-De-Luca, Andres Moreno-De-Luca, Jennifer G. Mulle, Stephen T. Warren, Gabriele Richard, John G. Compton,


Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O’Donnell-Luria, James S Ware, Andrew J


Massouras, Andreas, Sebastian M Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, Eric A


Fichera, Santina Reitano, Mariangela Lo Giudice, Kelly E. Li, Iris Casuga, 
Adam Broomer, Bernard Conrad, Markus Schwerzmann, Lorenz Räber, Sabina 
Gallati, Pasquale Striano, Antonietta Coppola, John L. Tolmie, Edward S. To-
bias, Chris Lilley, Lluis Armengol, Yves Spysschaert, Patrick Verloo, Anja 
De Coene, Linde Goossens, Geert Mortier, Frank Speleman, Ellen van Bins-
bergen, Marcel R. Nelen, Ron Hochstenbach, Martin Poot, Louise Gallagher, 
Michael Gill, Jon McClellan, Mary-Claire King, Regina Regan, Cindy Skinner, 
Roger E. Stevenson, Stylianos E. Antonarakis, Caifu Chen, Xavier Estivill, 
Björn Menten, Giorgio Gimelli, Susan Gribble, Stuart Schwartz, James S. Sut-
cliffe, Tom Walsh, Samantha J.L. Knight, Jonathan Sebat, Corrado Romano, 
Charles E. Schwartz, Joris A. Veltman, Bert B.A. de Vries, Joris R. Vermeesch, 
“Recurrent Rearrangements of Chromosome 1q21.1 and Variable Pediatric Phe-
notypes”. In: New England Journal of Medicine 359.16, pp. 1685–1699. DOI: 
10.1056/NEJMo0805384.

Meyer, Axel and Yves Van De Peer (2005). “From 2R to 3R: Evidence for a fish-
specific genome duplication (FSGD)”. In: BioEssays 27.9, pp. 937–945. DOI: 
10.1002/bies.20293.

Miller, D T, Y Shen, L A Weiss, J Korn, I Anselm, C Bridgemohan, G F Cox, H 
Dickinson, J Gentile, D J Harris, V Hegde, R Hundley, O Khwaja, S Kothare, 
C Luedke, R Nasir, A Poduri, K Prasad, P Raffalli, A Reinhard, S E Smith, 
M M Sobeih, J S Soul, J Stoler, M Takeoka, W-H Tan, J Thakuria, R Wolff, R 
Yusupov, J F Gusella, M J Daly, and B-L Wu (2009). “Microdeletion/duplication 
at 15q13.2q13.3 among individuals with features of autism and other neuropsy-
chiatric disorders.” In: Journal of Medical Genetics 46.4, pp. 242–248. DOI: 
10.1136/jmg.2008.059907.


Nicholas, Thomas J., Ze Cheng, Mario Ventura, Katrina Mealey, Evan E. Eichler, and Joshua M. Akey (2009). “The genomic architecture of segmental dupli-


Nuttall, George Henry Falkiner (1904). “Blood Immunity and Blood Relationship”. In:

Ohno, Susumu (1973). “Ancient linkage groups and frozen accidents.” In: Nature 244.5414, pp. 259–262. DOI: 10.1038/244259a0.


ton) exhibit near-complete expression divergence”. In: *Genome Biology and Evolution* 6.3, pp. 559–571. DOI: 10.1093/gbe/evu037.


Rice, Alan M. and Aoife McLysaght (2017). “Dosage sensitivity is a major determinant of human copy number variant pathogenicity”. In: *Nature Communications* 8, p. 14366. DOI: 10.1038/ncomms14366.


Romero, Irene, Ilya Ruvinsky, and Yoav Gilad (2012). “Comparative studies of
gene expression and the evolution of gene regulation”. In: *Nature Reviews
Genetics* 13.7, pp. 505–516. DOI: 10.1038/nrg3229.

Roper, Randall J. and Roger H. Reeves (2006). “Understanding the basis for
Down syndrome phenotypes”. In: *PLoS Genetics* 2.3, pp. 0231–0236. DOI:
10.1371/journal.pgen.0020050.

Rovelet-Lecrux, Anne, Didier Hannequin, Gregory Raux, Nathalie Le Meur, Annie
Laquerrière, Anne Vital, Cécile Dumanchin, Sébastien Feuillette, Alexis Brice,
Martine Vercelletto, Frédéric Dubas, Thierry Frebourg, and Dominique Campion
Alzheimer disease with cerebral amyloid angiopathy.” In: *Nature Genetics* 38.1,
pp. 24–26. DOI: 10.1038/ng1718.

Ruderfer, Douglas M, Tymor Hamamsy, Monkol Lek, Konrad J Karczewski, David
Kavanagh, Kaitlin E Samocha, Exome Aggregation Consortium, Mark J Daly,
Daniel G MacArthur, Menachem Fromer, and Shaun M Purcell (2016). “Patterns
of genic intolerance of rare copy number variation in 59,898 human
exomes.” In: *Nature Genetics* 48.10, pp. 1107–11. DOI: 10.1038/ng.3638.

Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel
J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, Jill V. Platko, Nick J.
Patterson, Gavin J. McDonald, Hans C. Ackerman, Sarah J. Campbell, David
Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S. Lander
(2002). “Detecting recent positive selection in the human genome from
haplotype structure”. In: *Nature* 419.October, pp. 832–837. DOI: 10.1038/
nature01140.

Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter,
Chris Cotsapas, Xiaohui Xie, Elizabeth H. Byrne, Steven A. McCarroll,
Rachelle Gaudet, Stephen F. Schaffner, Eric S. Lander, and International


Sharp, Andrew J., Sierra Hansen, Rebecca R. Selzer, Ze Cheng, Regina Regan, Jane A. Hurst, Helen Stewart, Sue M. Price, Edward Blair, Raoul C. Hennekam, Carrie A. Fitzpatrick, Rick Segraves, Todd A. Richmond, Cheryl Guiver, Donna G. Albertson, Daniel Pinkel, Peggy S. Eis, Stuart Schwartz, Samantha J. L. Knight, and Evan E. Eichler (2006). “Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome”. In: *Nature Genetics* 38.9, pp. 1038–1042. DOI: 10.1038/ng1862.


Stefansson, Hreinn, Andreas Meyer-Lindenberg, Stacy Steinberg, Brynja Magnusdottir, Katrin Morgen, Sunna Arnarisdottir, Gyda Bjornsdottir, G Bragi Walters, Gudrun A. Jonsdottir, Orla M. Doyle, Heike Tost, Oliver Grimm, Solveig Kristjansdottir, Heimir Snorrason, Solveig R. Davidsdottir, Larus J. Gudmundsson, Gudbjorn F. Jonsson, Berglind Stefansdottir, Isafold Helgadottir, Magnus Haraldsson, Birna Jonsdottir, Johan H. Thygesen, Adam J. Schwarz,


Xie, Ting, Qing Yong Yang, Xiao Tao Wang, Aoife McLysaght, and Hong Yu Zhang (2016). “Spatial colocalization of human ohnolog pairs acts to maintain dosage-balance”. In: Molecular Biology and Evolution 33.9, pp. 2368–2375. DOI: 10.1093/molbev/msw108.


Yates, Andrew, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah


