Pathogenicity and selective constraint in the non-coding genome

Patrick J. Short
St. Edmund's College

Submitted: September, 2018

Supervisor: Dr. Matthew Hurles

This dissertation is submitted in accordance with the requirements of the University of Cambridge for the degree of Doctor of Philosophy.

Acknowledgements

I would like to thank all of the members of the Hurles team and the DDD project for four wonderful years of research and for being such great scientists and friends. To Jeremy, Sebastian, and Jeff especially, thank you for all of your mentorship.

To my supervisor Matt, I am incredibly thankful to have a supervisor that is both a rigorous, visionary scientist, and such a down-to-earth person. Thank you so much for all of your support.

To the Wellcome Trust and the Mathematical Genomics and Medicine PhD programme, thank you for providing the structure and financial support to make all of this work possible.

To my parents, Jack and Denice, my brother John, and my fiancée Timarie, thank you so much for supporting me, even as I moved ‘across the pond’ to pursue this dream. I consider myself so lucky to have a family like you.

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the School of Biological Sciences.

Table of Contents

Abbreviations

3C – Chromosome Conformation Capture
ASD – autism spectrum disorder
ASE – allele specific expression
BWA – Burrows-Wheeler Aligner
CADD – combined annotation dependent depletion
CDTS – context dependent tolerance score
CNEs – conserved non-coding elements
CNN – convolutional neural network
CNVs – copy number variations
CRISPRa – CRISPR activation
CRISPRi – CRISPR inactivation
DDD – Deciphering Developmental Disorders
DD – Developmental Disorders
DHS – DNase I hypersensitive site
DNA – deoxy ribonucleic acid
DNM – de novo mutation
DSB – double strand break
ENCODE – Encyclopedia of DNA Elements
eQTL – expression quantitative trait loci
ESCs – embryonic stem cells
ExAC – exome aggregation consortium
FDR – false discovery rate
GAM - Genome Architecture Mapping
GATK – Genome Analysis Toolkit
GFP – green fluorescent protein
gnomAD – genome aggregation database
HARs – human accelerated regions
HGEs – human gained enhancers
Indels – insertions/deletions
iPSCs – induced pluripotent stem cells
LCLs – Lymphoblastoid Cell Lines
lncRNAs – long non-coding RNAs
MAPS – mutability adjusted proportion of singletons
MAVEs – multiplexed assays of variant effect
MPRAs – massively parallel reporter assays
NHEJ – non-homologous end-joining
NHS – National Health Services
PAH – pulmonary arterial hypertension
pLI – probability of loss-of-function intolerance
REP – Roadmap Epigenome Project
RFR – random forest regression
RNA – ribonucleic acid
SHH – Sonic Hedgehog
SNPs – single nucleotide polymorphisms

SNV – single nucleotide variant
SVM – support vector machine
TBP - TATA-box binding protein
TCGA – The Cancer Genome Atlas
TFBS – transcription factor binding site
TFs – Transcription Factors
TSS – Transcription Start Site
UMIs – unique molecular identifiers
UTR – untranslated region
VEP – Variant Effect Predictor
VFS – variant frequency spectrum
VUS – variant of unknown significance
WGS – whole genome sequencing

Chapter 1: Introduction

Section 1.1: The function of non-coding DNA in the human genome

Section 1.1.1. Functional roles of non-coding elements in the human genome

The vast majority of DNA sequence in the human genome does not code for protein[1]. Non-coding DNA can serve many functional roles including transcriptional activation, repression, and three-dimensional genome organisation. Based on comparative evolutionary analyses, between 3% to 15% of the non-coding genome is predicted to be under purifying selection[2-4], while protein-coding sequence comprises less than 2% of the genome[1]. Thus, the amount of functional non-coding sequence likely outnumbers functional coding sequence, yet the function of most non-coding sequence is poorly understood.

The regulation of gene expression in time and space is an essential part of organismal development. The 'core promoter' includes a transcription start site (TSS), a binding site for RNA polymerase II, and binding sites for more general transcription factors[5]. The proximal promoter is less rigidly defined and includes regulatory sequence to the 5' end of the transcribed strand which may harbour tissue-specific transcription factors (TFs). For many genes, particularly those with complex tissue-specific regulation, enhancers play an important role in regulating transcription[6,7]. Enhancers contain numerous transcription factor binding sites (TFBS), which can be identified by distinctive sequence patterns called motifs. These sequence patterns are often very flexible, maintaining binding efficiency despite changes to the underlying sequence. Furthermore, enhancers and TFBS show a high degree of evolutionary turnover[8,9]. Villar et. al identified putative regulatory regions in the liver of twenty different mammals and showed that nearly half of all enhancers show rapid lineage specific evolution[9]. While some principles of the enhancer 'grammar' have emerged, including cooperation and antagonism between different TFs[10], the nucleotide-level logic in most enhancers are poorly understood. Enhancers can be proximal (tens of kilobases from the TSS) as well as distal (in some cases, more than one million bases from the TSS).

While distal enhancers may be hundreds of kilobases from the genes they regulate in genomic space, they often colocalise in physical space. Topologically associated domains (TADs) are large genomic segments (median size 880kb) that interact more frequently than expected under a model of random interactions[11]. The prevailing model by which looping is thought to occur is by loop extrusion by the cohesin complex, until the complex encounters a 'boundary element' containing the transcription factor CTCF[12]. TAD boundaries have been shown to be highly conserved across different cell types and tissues[11]. Interactions between regulatory elements and genes are far more likely occur within a TAD than between TADs, although there are examples of interactions between enhancers and promoters in different TADs[13]. While physical looping has been proposed as the primary mechanism for long-range enhancer-promoter interactions (Figure 1A), there is evidence that looping may not occur in all cases[14]. In particular, other mechanisms have been proposed including induced phase-separation, chromatin decompaction, and subsequent diffusion of transcription factors (Figure 1B)[14,15]. Via looping or diffusion, general transcription factors and tissue specific transcription factors complex with RNA polymerase II to promote transcription[16]. While general principles of gene regulation are beginning to emerge, there is still no broad consensus around a unified model of gene regulation in all cases.

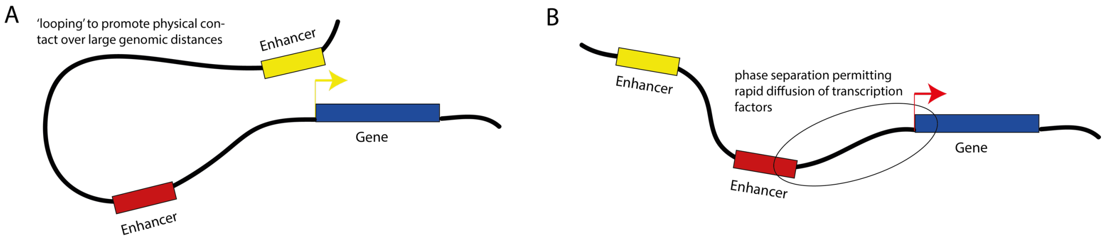

A
'looping' to promote physical contact over large genomic distances
Enhancer
Gene
Enhancer
B
Enhancer
phase separation permitting rapid diffusion of transcription factors
Gene
Enhancer

Figure 1 Schematic of transcriptional activation via enhancer looping and diffusion facilitated by chromatin decompaction and phase separation (a) Looping brings enhancers into physical proximity of the gene promoter. (b) Phase transition and chromatin decompaction allows rapid 1D diffusion of transcription factors to the target promoter.

Many genes encoding transcription factors are only transcribed in a subset of cells, tissues, or developmental time-points. As a result, enhancers and promoters harbouring binding sites for these TFs can drive tissue-specific transcriptional patterns. For example, the GLI3 gene is expressed primarily in the developing brain and limb. Antagonistic interactions between GLI3 and Sonic Hedgehog (SHH) in the developing limb results in SHH expression in the Zone of Polarising activity, but nowhere else in the developing limb bud. Together, interactions between GLI3 and SHH impact the expression of more than 1,000 downstream genes in the developing limb[17]. Detailed annotations of enhancer activity across mouse development suggest that the majority of enhancers are expressed during temporally-restricted windows, which were associated with different stages of organogenesis[18].

Multiple enhancers can act in a particular tissue or time-point in a coordinated manner. For example, the *Wap* gene is upregulated more than 1,000 fold in the mouse mammary gland[19] by the coordinated interaction of three different enhancers. A series of experiments deleting or perturbing individual enhancers and pairs of enhancers in the locus showed that all three were necessary to achieve the 1,000-fold upregulation. Clusters of coordinated enhancers, bound by the mediator complex, with non-additive contributions to gene expression have been termed super-enhancers[20]. However, the appropriateness and utility of this new category has been disputed, and it is not clear whether super-enhancers are simply collections of 'normal' enhancers with acting with varying degrees of strength or represent a novel functional category where constituent enhancers display synergistic properties[21].

Enhancers can also exhibit functional redundancy[22]. Osterwalder et. al created ten mouse lines with homozygous deletions of individual enhancers shown to regulate genes critical for normal limb development. Single deletions showed no discernible limb malformations. However, combinatorial deletion of multiple enhancers in a single locus did result in limb malformations. For example, deletion of *mm1179* and *hs1586*, two enhancers shown to regulate GLI3 expression, results in a duplication of the first digit[23].

Section 1.1.2. Characterising non-coding elements in the human genome

In light of the evidence for an important functional role for non-coding elements in the human genome, there have been a number of efforts to annotate non-coding function. These efforts have introduced novel methods to annotate non-coding function and collaborative efforts have been formed to apply new and established methods to a diverse range of organisms, tissues, and time points.

The earliest large-scale annotations of putative functional non-coding DNA leveraged DNA sequence from different mammalian and vertebrate species to identify non-coding sequence with low levels of sequence divergence[24]. Early work on ultra-conserved non-coding elements (defined as having >97% sequence identity between human, mouse, and rat) showed that a large fraction of these elements were likely acting as enhancers[24-27]. These enhancers were enriched near genes previously known to be involved in development, suggesting that their evolutionary conservation was due to a highly conserved role in gene expression during development[27]. Beyond the most highly ultra-conserved elements, which comprised a very small fraction of the genome, multi-species alignments indicated that 3-15% of the genome was conserved[2-4]. As only 1-2% of genome encodes protein-coding, large-scale efforts are underway to annotate non-coding functional elements genome-wide.

The Encyclopedia of DNA Elements (ENCODE) project launched in 2007 to systematically map functional elements across the human genome using a variety of different biochemical methods[28,29]. Nucleosome occupancy and epigenetic modification of the histone proteins comprising the core of the nucleosome have a profound impact on the accessibility of a segment of DNA. DNase-seq identifies regions of DNA that are sensitive to cleavage by DNase I[30]. These regions are often denoted as 'open' as they are fully or partially free of nucleosomes. More recently, ATAC-seq has emerged as the method of choice for identifying open chromatin due to its low requirement for input material compared to DNase-seq, and comparative speed and ease of use[31].

In addition to nucleosome occupancy, epigenetic modifications to histone proteins can shed light on the underlying function of a piece of DNA. Promoter regions and transcription start sites are often marked by trimethylation of K4 (H3K4me3)[32] and the bodies of transcribed genes by trimethylation of histone H3 at lysine 36 (H3K36me3)[33]. Like promoters, active enhancers can be identified by distinctive chromatin signatures, including open chromatin, monomethylation of histone H3 at lysine 4 (H3K4me1) and acetylation of histone H3 at lysine 27 (H3K27ac)[34]. H3K27ac by p300/CBP has been shown to destabilise nucleosomes, promoting accessibility by transcriptional machinery and TFs. In contrast, repressed or inactive chromatin is marked by trimethylation at histone H3 lysine 27 (H3K27me3) or trimethylation at histone H3 lysine 9 (H3K9me3)[35,36]. These marks are assayed using a technique called chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq)[37-39]. In the ChIP-seq method, DNA is first cross-linked to preserve any DNA-protein interactions. An antibody to the protein of interest in then used to extract the short DNA fragments bound to the protein. These short fragments are sequenced and mapped back to the reference genome, providing indication of where the protein was bound[37-39].

While there are many other potential histone modifications, six core marks (H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3) have been shown to sufficient to define a number of different 'chromatin states'. Assaying these modifications has been the focus of large-scale efforts such as the ENCODE project[29] and the Roadmap Epigenome Project[40] to annotate non-coding function across different tissues and developmental time-points. A number of methods have been developed to integrate data from open chromatin assays and histone modifications to define the chromatin state. One of the most widely used methods, chromHMM, integrates aligned reads from each input feature and uses a hidden markov model to partition the genome into predicted chromatin states[41]. The simplest model has 15-states, which include active states such as strong and weak transcription, active/weak/poised promoters, and active/weak/poised enhancers as well as inactive or repressed states such as heterochromatin, or polycomb repression.

Beyond histone modifications, ChIP-seq can be used to identify DNA-binding events for any protein for which an antibody is available. To date, the ENCODE project has assayed more than 167 human transcription factors in 127 different tissue/cell types. These data sets have been used to identify canonical and non-canonical transcription factor binding motifs[42]. Due to the difficulties associated with testing thousands of transcription factors across hundreds of different spatial and temporal contexts, machine learning techniques have been developed to detect TF-binding motifs in the absence of direct measurement by ChIP-seq[43].

Regulatory elements can also be identified using variation in gene expression or chromatin features resulting from natural genetic variation in a population. Associations between genetic variation and gene expression, termed expression quantitative trait loci[44-46] (eQTLs), were first established using Lymphoblastoid Cell Lines (LCLs) and have expanded substantially to include a wide range of cell types and tissues[47-49]. Analysis of the underlying genetic architecture of eQTLs suggests that a large fraction of associations fall within open chromatin peaks and directly impact transcription factor binding[50]. The majority of SNPs implicated in genome-wide association studies of common and complex disease risk fall outside coding regions, and there is a substantial overlap between these risk-associated loci and eQTLs in disease-relevant cell types, suggesting gene mis-regulation as a primary contribution to common and complex disease[49].

As discussed in section 1.1, genomic elements that are far apart in sequence space may interact by forming loops in three-dimensional space. The spatial organisation of chromatin can be detected by a number of different methods. The most widely used techniques are innovations around the chromosome conformation capture (3C) technique[51]. The original 3C technique, which used PCR to check for interaction between known fragments, has been modified to produce genome-wide interaction maps (Hi-C[52]), and detailed interaction maps centred on promoters or other elements of interest (Capture-C[53-55]). Hi-C has been used to define TADs in a number of different tissues and cell types[11]. However, Hi-C and Capture-C can also reveal more fine-grained regulatory interactions within TADs, including promoter-promoter and enhancer-promoter interactions. While Hi-C and Capture-C have proved valuable tools for linking enhancers to putative target genes, the sheer complexity of potential tissues, cell types, and time points to assay, combined with the incomplete sensitivity of the tools means that gene target prediction remains a challenge.

Other methods for interrogating three-dimensional organisation that are not based on ligation have recently emerged. Genome architecture mapping (GAM) assays three-dimensional organisation by freezing the cells of interest, cryosectioning the frozen cells, and isolating nuclei using laser capture microdissection. The DNA in each of these thin slices is sequenced and the frequency of interaction is quantified based on the proportion of cellular slices in which two genomic segments are sequenced together[13]. This method has the advantage of improved detection of interactions between three different segments – in ligation based methods where only two interacting segments can be identified at a time, two separate pair-wise interactions cannot be easily disambiguated from a single three-way interaction. Furthermore, GAM requires a small amount of cellular material and makes use of laser microdissection, which enables this method to be used on rare cell types or directly on patient tissue[13].

The applications of open chromatin assays, ChIP-seq, and 3C techniques in the ENCODE and Roadmap Epigenome Projects have produced fundamental insights into human biology, but have focused primarily on natural variation present in human tissues, cell lines, induced pluripotent stem cells (iPSCs) and iPSC-derived tissues[29,40]. Improvements in oligo synthesis and genome-editing technologies have given rise to new high-throughput methods for generating programmed variation in the underlying DNA sequence.

Reporter assays have been long been used to test putative promoters and enhancers for their ability to drive gene expression[56]. Reporter assays use a bacterial plasmid with a putative regulatory element upstream of a reporter gene, often GFP or luciferase. These plasmids can then be transfected into a cell type of interest and the output of the reporter gene quantified, either by measuring fluorescence intensity or by quantitative PCR. The original reporter assays, designed to test a small number of putative regulatory elements, have been adapted to allow testing of tens of thousands of elements at once. For example, massively parallel reporter assays (MPRAs) can quantify the regulatory activity of tens of thousands of enhancers in a single experiment[57,58]. MPRAs work by adding a DNA barcode to each putative regulatory element. The reporter gene is then inserted between the enhancer element and the barcode, resulting in the barcode being transcribed on the 3' end of the reporter gene. These barcodes can be quantified with RNA-sequencing and regulatory elements driving stronger reporter gene expression will have a greater number of barcodes relative to the amount of input DNA. Different MPRA variations have emerged including STARR-seq, which places the regulatory element downstream of the reporter gene, allowing a direct readout of the element rather than a linked barcode.

Saturation mutagenesis experiments using this technology have recovered known disease-associated variants, for example in the TERT promoter (unpublished work from Nadav Ahituv) and to test for causal variants in eQTL studies from a large number of SNPs in linkage disequilibrium[59]. Beyond testing variations in wild-type sequence, these assays can be used to understand the 'grammar' of non-coding elements at a more fundamental level, by building functional sequences from scratch, and systematically altering or destroying synthetic sequences[10,57,60]. These assays have a number of drawbacks – in particular, the enhancer or promoter is being tested outside of its native context, either in an episome or integrated into the genome by a lentivirus. Furthermore, not all tissues or cellular models are amenable to transfection. Greater detail on the rapid evolution of the experimental techniques and applications of MPRA to interrogate enhancer and promoter function is included in the Introduction to Chapter 4.

In order to test the impact of genetic changes in their native genetic context, several methods have been pioneered using programmed guide-RNAs and the genome-editing enzyme Cas9. *Gasperini et. al* use pairs of programmed guide-RNAs to delete thousands of kilo-base sized genomic regions[61]. Their first application of the method was applied to a housekeeping gene, *HPRT1*, which has little distal regulatory sequence, but in principal this method has the potential to test the impacts on cellular fitness or gene expression from deletion of coding or non-coding sequence.

Cas9 cutting can be paired with a repair template, allowing more precise edits such as single base pair changes via homology directed repair (HDR). However, the double-strand break (DSB) induced by the CRISPR enzyme most often resolves by non-homologous end-joining (NHEJ) usually resulting in small deletions at the cut site. Recent work by *Findlay et. al* has shown that performing the edits in a cell-line with *LIG4*, an important gene in the NHEJ pathway, knocked out results in lowered efficiency in NHEJ and a greater proportion of DSBs proceeding via HDR. *Findlay et. al* created more than 96.5% of all possible SNVs in two exons of the BRCA1 gene, paving the way for CRISPR-based alternatives to the massively parallel reporter assays discussed above[62].

Cellular models may not reveal the full spatial and temporal complexity of enhancer activity on an organismal level. Mouse transgenesis assays use a GFP reporter construct driven by the putative enhancer of interest to test for tissue-specific enhancer activity. To date, more than 2,800 putative enhancers have been tested and results made available through the VISTA browser[63], with more than 1,500 testing consistently positive in at least one tissue. Beyond transgenesis assays, mouse knockouts can reveal the function of evolutionarily conserved enhancers *in vivo*. *ARX* is an essential neuronal transcription factor in the human and mouse brain with at least two enhancers (*hs119* and *hs121*) that control expression in the ventral forebrain. These enhancers are ultraconserved, exhibiting >97% identity between human and mouse.[24] Homozygous deletion of either *hs119* or *hs121* show subtle phenotypic changes including subtle changes in body weight and density of cholinergic neurons. However, deletion of both ultraconserved enhancers shows a dramatic reduction in both body weight and cholinergic neurons[64].

Taken together, these technologies provide a detailed overview of the functional non-coding sequence across the genome for hundreds of different cellular contexts and time-points. An overview of each of these techniques is described in Table 1 which includes the type of elements/interactions tested, the approximate throughput (number of genomic elements tested in a typical experiment). Projects such as ENCODE[29], The Roadmap

Epigenome Project[40], and FANTOM[65] have focused on building an atlas of reference epigenomes and functional non-coding elements in primarily healthy tissues and cellular contexts. As the role of non-coding elements in common and Mendelian disease becomes increasingly apparent, these technologies can be applied to understand the mechanism of gene mis-regulation in these diseases.

Name of technique	Element/interaction assayed	Number of elements tested in a typical experiment
DNase-seq	Open chromatin	Genome-wide
ATAC-seq	Open chromatin	Genome-wide
ChIP-seq	Protein-occupancy (commonly histone marks, or transcription factors)	Genome-wide
Hi-C/Capture-C	Three-dimensional interaction between DNA elements	Genome-wide
Reporter Assays	Ability of putative regulatory element to drive gene expression	100s of loci
Mouse Transgenesis Assays	Ability of putative regulatory element to drive gene expression, with tissue specificity.	10s of loci
Massively Parallel Reporter Assays (MPRAs)	Activity of putative enhancer/promoter to drive gene expression	>10,000 loci
CRISPR-inactivation	Assessing the impact repressing a genomic element	>10,000 loci
CRISPR-editing	Assessing the impact of deletion/alteration of a genomic element.	>10,000 loci

Table 1 Overview of established and emerging techniques in non-coding genome annotation.

Section 1.2: The role of non-coding variation in Mendelian disease

Section 1.2.1. The contribution of protein-coding variation to Mendelian disease

The overwhelming majority of established genetic causes of Mendelian disease are caused by protein-altering single-nucleotide variation, small insertions/deletions (indels), or larger copy number variations (CNVs). Clinical microarrays and gene panels have been used extensively to test for diagnostic protein-altering variation in Mendelian disease, particularly in developmental disorders. As the cost of genome-sequencing has continued to decline, whole exome sequencing, which allows for targeted sequencing of protein-coding regions, has proven to deliver a higher diagnostic yield[66,67]. Across a broad category of Mendelian disorders, diagnostic yield is approximately 25-30%, but the actual diagnostic yield differs substantially between different disease groups[68]. For example, RASopathies, a collection of disorders resulting from mutations in the RAS-MAP kinase pathway, have diagnostic rates of over 60% whereas developmental disorders, which suffer from the issue of phenotypic similarity across a wide range of potential causal genes, have a diagnostic yield of 25-30%[68-70].

Study design can also influence diagnostic yield. For example, sequencing of parent-offspring trios has been shown to greatly improve the diagnostic yield compared to sequencing the proband-only[71,72]. In the past few years, hundreds of novel Mendelian disease genes have been identified using exome sequencing, improving the diagnostic yield even further[73]. In the case of severe developmental disorders, *Wright et. al* estimate that implementing parent–offspring whole-exome sequencing as a first-line diagnostic test would diagnose approximately 50% of patients[74].

McRae et. al estimated that 42% of undiagnosed severe developmental disorder cases harboured damaging *de novo* mutations in protein-coding genes[70]. Of this estimated 42%, ~25% could be robustly linked to a known or novel developmental disorder gene while the remaining ~17% were found in genes not yet robustly linked to developmental disorders. Beyond the contribution from *de novo* mutations in genes with a monoallelic disease mechanism, *Martin et. al* estimated the contribution of protein-altering variation in biallelic genes to be only 4% in British-European cases, and up to 20% in British-Pakistani[75].

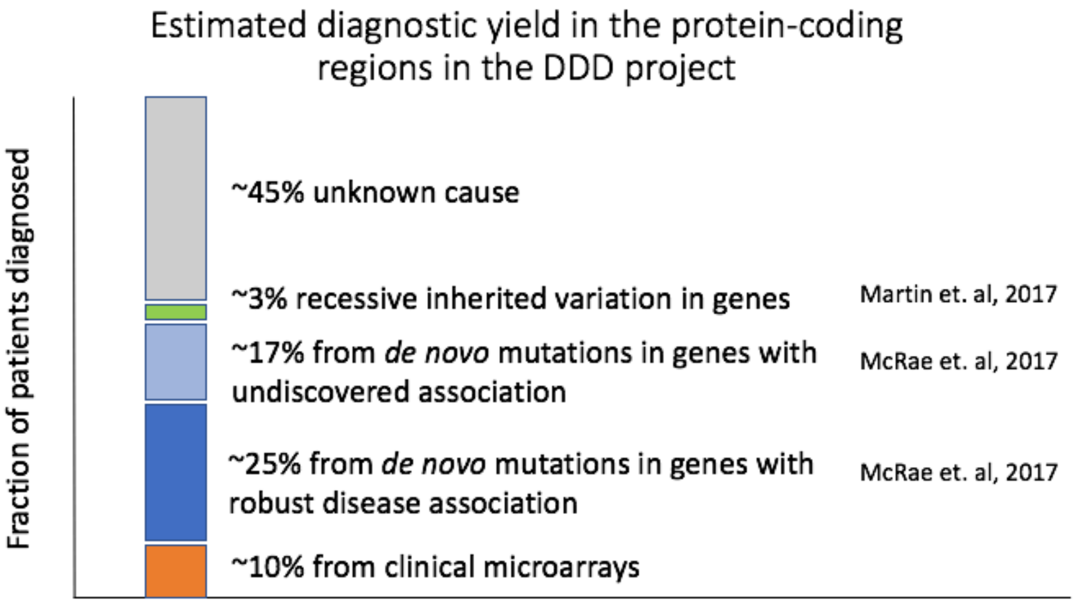


Figure 2 Diagnoses in the DDD project broken down by variant class. *De novo* mutations in protein-coding genes make up the largest fraction of diagnoses (25%), and analyses suggest that a substantial fraction of the missing diagnoses (~17%) will come from genes not yet robustly associated to DD. As of this writing, ~45% of the DDD cases likely lack a highly penetrant protein-coding variant contributing to their disorder.

Beyond developmental disorders, other severe Mendelian diseases have a high rate of unsolved cases, despite large-scale exome sequencing projects. In pulmonary arterial hypertension (PAH) heterozygous mutations in the coding sequence BMPR2 are found in an estimated 80% of familial cases and 20% of sporadic cases[76]. However, sporadic cases greatly outnumber familial cases; in a study of 1,048 individuals affected with PAH, 5.5% reported a family history[77]. Mutations in BMPR2 and other recently reported novel genes were found in just 23.5% of cases[77]. Thus, a substantial fraction of individuals remain without a cause in the protein-coding regions, even in large and well-powered studies of Mendelian Diseases (Figure 2). These results have motivated the search for causal variation outside of protein-coding genes.

Section 1.2.2. Regulatory variation in Mendelian phenotypes

The importance of the non-coding genome in complex disease is well established - the vast majority of disease-associated single nucleotide polymorphisms (SNPs) lie in intergenic or intronic regions[78,79]. Fine-mapping studies have shown that in most cases associations with a non-coding SNP cannot simply be accounted for by linkage with a coding variant on the same haplotype. Analysis of forty coding SNPs associated with Type-II Diabetes showed that a large fraction of associations (13/40) are actually 'false leads' that are likely driven by a nearby non-coding SNP[80]. In Mendelian disorders, the role of enhancers or other non-coding elements is less clear. There have been a number of regulatory elements linked to Mendelian disorders through targeted re-sequencing and pedigree analyses. *Lettice et. al* identified a set of single nucleotide variants in evolutionarily conserved sites in a regulatory element located 1Mb from the target gene, *Shh*, responsible for polydactyl[81]. Inherited single nucleotide variants as well as whole-element deletions in *SOX10* enhancers have also been shown to reduce *SOX10* expression, contributing to isolated Hirschprung disease[82]. *Weedon et. al* describe a set of six different recessive variants in an enhancer located 25kb from *PFT1A* that disrupt transcription factor binding sites for *FOXA2* and *PDX1*, abolishing enhancer activity and causing pancreatic agenesis[83]. A single nucleotide variant in an ultraconserved element regulating the expression of *PAX6*, a critical gene in eye development, has also been shown to cause Aniridia. Other examples of point mutations and small insertions/deletion in non-coding elements that have been causally linked to human disease are described in a review by *Mathelier et. al*[84].

In addition to point-mutations and small insertions/deletions, larger-scale copy-number variations and rearrangements of regulatory sequence have been shown to cause Mendelian disorders. Loss of function can be caused by deletion of an enhancer, as in the *SOX10* case described above resulting in isolated Hirschsprung disease[82], as well as in Pierre-Robin syndrome[85,86] , a condition marked by malformations of the cranial skeleton. Genomic rearrangements or disruptions of topologically associated domains (for example, by deletion of a *CTCF* binding site) can cause enhancers to regulate a gene they do not normally regulate. This phenomenon is termed 'enhancer adoption' and has been implicated in brachydactyl type A2[87] (a shortening of the digits) as well as sex-reversal due to a gonad-specific gain of function of *SOX9*[88]. A number of studies in which copy number variation and rearrangements of non-coding cis-regulatory elements contribute to Mendelian disorders by disrupting or altering gene regulation are reviewed by Malte Spielmann and Stefan Mundlos[89].

Nearly all of these well-established examples of regulatory causes of Mendelian disease share the common characteristic of being non-syndromic. In each of the cases, a single organ or organ system is affected, reflecting the tissue-specificity of the enhancer element whose function is being perturbed. It is unclear whether this is a common characteristic of Mendelian disease caused by variants in regulatory regions, or a by-product of ascertainment based on recognisable phenotypes. Furthermore, unlike loss-of-function mutations in protein-coding genes that may only impact the amount of functional protein produced, mutations in regulatory elements have the potential to cause more complex mis-regulation. Thus, the impact of variation in regulatory elements may not be directly comparable to changes in the protein-coding sequence.

Section 1.2.3 The role of regulatory variation in cancer and neurodevelopmental disorders

There have been reports of recurrent somatic mutations in regulatory regions in several cancer types using whole genome sequence data from The Cancer Genome Atlas (TCGA)[90,91]. These studies included nearly 500 whole genome sequences of tumours and matched normal tissue and primarily identified mutations in promoter sequences as well as 3’ UTRs and 5’ UTRs. The strongest signal, in the *TERT* promoter, comprises mutations that show substantial overexpression of a luciferase reporter as well as RNA levels of the *TERT* transcripts in the cancerous tissue[90,91]. Further work using a combination of whole genome sequencing and chromatin immunoprecipitation sequencing (ChIP-exo) highlighted disruption of *CTCF* and cohesin binding sites in colorectal cancer which contribute to genomic stability and establishment/maintenance of TAD boundaries[92].

Expanding these analyses to nearly 1,000 tumour whole genomes with matching transcriptomes, *Zhang et. al* recapitulated the variants resulting in overexpression in the *TERT* promoter and discover novel regulatory elements regulating the expression of *DAAM1*, *MTG2*, and *HYI*[93]*.* Their approach leverages the matched transcriptomes to identify ‘somatic eQTLs’ linking regulatory elements to putative target genes. This is a promising new approach for disorders where RNA transcript levels can be measured in the relevant tissue, or where accurate cellular models can be derived, for example, by reprogramming induced pluripotent stem cells (iPSCs).

There is substantial evidence for protein-coding DNMs contributing to autism spectrum disorder (ASD), albeit at a lower contribution than in severe developmental disorders[94]. Thus, there has been a similar motivation to assess the contribution of DNMs in the non-coding genome to ASD. For example, *Turner et. al* reported a nominally significant enrichment for *de novo* mutations in autism cases compared to unaffected siblings in 3' UTRs, promoters, and conserved transcription factor binding sites[95] using data from whole genome sequenced trios. However, independent analysis of the same cohort from *Werling et. al* find no significant enrichment and show that without appropriate correction for multiple statistical tests, plausible associations can be observed in both cases and controls[96].

Thus, the contribution of variation in the non-coding genome to developmental disorders and other rare disorders is not clear. The Deciphering Developmental Disorders study has nearly an order of magnitude more cases than the ASD studies discussed above, and included approximately 4 megabases of regulatory sequence in the exome capture. This study design may afford greater power to test for DNM burden, albeit in a more limited set of non-coding annotations. This study will be covered in detail in Chapter 2. The dramatic increase in the number of whole genome sequenced individuals affords an opportunity to identify non-coding elements under purifying selection across the genome in a disease-agnostic manner. Analyses based on more than 25,000 deep whole genomes using population genetics methods to identify selectively constrained non-coding elements will be covered in detail in Chapter 3. Finally, methods for assaying the function of non-coding elements and the impact of genetic variation in these elements can be used to delineate benign from damaging variation in these non-coding elements and to understand their function. Chapter 4 includes results from a series of MPRA experiments and mouse transgenesis assays including non-coding mutations identified in developmental disorder cases.

1 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

2 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

3 Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769-1776, doi:10.1101/gr.116814.110 (2011).

4 Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).

5 Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annual Review of Biochemistry* **72**, 449-479, doi:10.1146/annurev.biochem.72.121801.161520 (2003).

6 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286, doi:10.1038/nrg3682 (2014).

7 Banerji, R., Schaffner. Expression of a Beta-Globin Gene is Enhanced by REmote SV40 DNA Sequences. *Cell* (1981).

8 Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**, 1335-1343, doi:10.1101/gr.108795.110 (2010).

9 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).

10 Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* **45**, 1021-1028, doi:10.1038/ng.2713 (2013).

11 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

12 Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102-105, doi:10.1126/science.aar7831 (2018).

13 Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519-524, doi:10.1038/nature21411 (2017).

14 Benabdallah, N. S. & Bickmore, W. A. Regulatory Domains and Their Mechanisms. *Cold Spring Harb Symp Quant Biol* **80**, 45-51, doi:10.1101/sqb.2015.80.027268 (2015).

15 Benabdallah, N. S. *et al.* PARP mediated chromatin unfolding is coupled to long-range enhancer activation. *bioRxiv*, doi:10.1101/155325 (2017).

16 Hardison, R. C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**, 469-483, doi:10.1038/nrg3242 (2012).

17 Tickle, C. & Towers, M. Sonic Hedgehog Signaling in Limb Development. *Front Cell Dev Biol* **5**, 14, doi:10.3389/fcell.2017.00014 (2017).

18 Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531, doi:10.1016/j.cell.2013.11.033 (2013).

19 Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics*, 1-10, doi:10.1038/ng.3606 (2016).

20 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

21 Hay, D. *et al.* Genetic dissection of the alpha-globin super-enhancer in vivo. *Nature Genetics* **48**, 895-903, doi:10.1038/ng.3605 (2016).

22 Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490-493, doi:10.1038/nature09158 (2010).

23 Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243, doi:10.1038/nature25461 (2018).

24 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).

25 Kleinjan, D. A. & Van Heyningen, V. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. *Am. J. Hum. Genet* **76**, 8-32, doi:10.1086/426833 (2005).

26 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

27 Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99, doi:10.1186/1471-2164-5-99 (2004).

28 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816, doi:10.1038/nature05874 (2007).

29 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

30 Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**, 123-131, doi:10.1101/gr.4074106 (2006).

31 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

32 Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in S-cerevisiae. *Plos Biol* **3**, 1753-1769, doi:10.1371/journal.pbio.0030328 (2005).

33 Bannister, A. J. *et al.* Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem* **280**, 17732-17736, doi:10.1074/jbc.M500796200 (2005).

34 Creygthon, C., Welstead, Kooistra, Carey, Steine, Hanna, Lodato, Frampton, Sharp, Boyer, Young, Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS* (2010).

35 Young, M. D. *et al.* ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res* **39**, 7415-7427, doi:10.1093/nar/gkr416 (2011).

36 Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128**, 735-745, doi:10.1016/j.cell.2007.02.009 (2007).

37 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

38 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).

39 Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**, 1813-1831, doi:10.1101/gr.136184.111 (2012).

40 Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

41 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28**, 817-825, doi:10.1038/nbt.1662 (2010).

42 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).

43 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**, doi:10.1038/nmeth.3547 (2015).

44 Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-1207, doi:10.1038/ng2109 (2007).

45 Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-1224, doi:10.1038/ng2142 (2007).

46 Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747, doi:10.1038/nature02797 (2004).

47 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

48 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

49 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089, doi:10.1038/ng.2394 (2012).

50 Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* **13**, R7, doi:10.1186/gb-2012-13-1-r7 (2012).

51 Dekker, R., Dekker, Kleckner. Capturing Chromosome Conformation. *Science* (2002).

52 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

53 Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178, doi:10.1038/ncomms7178 (2015).

54 Davies, J. O. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* **13**, 74-80, doi:10.1038/nmeth.3664 (2016).

55 Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* **46**, 205-212, doi:10.1038/ng.2871 (2014).

56 Dewet, J. R., Wood, K. V., Deluca, M., Helinski, D. R. & Subramani, S. Firefly Luciferase Gene - Structure and Expression in Mammalian-Cells. *Mol Cell Biol* **7**, 725-737, doi:Doi 10.1128/Mcb.7.2.725 (1987).

57 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

58 Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).

59 Tewhey, R. *et al.* Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Resource Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529, doi:10.1016/j.cell.2016.04.027 (2016).

60 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 800-811, doi:10.1101/gr.144899.112 (2013).

61 Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet* **101**, 192-205, doi:10.1016/j.ajhg.2017.06.010 (2017).

62 Findlay, G. M. *et al.* Accurate functional classification of thousands of BRCA1 variants with saturation genome editing. *bioRxiv*, doi:10.1101/294520 (2018).

63 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, 88-92, doi:10.1093/nar/gkl822 (2007).

64 Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491, doi:10.1016/j.cell.2017.12.017 (2018).

65 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, doi:10.1038/nature12787 (2014).

66 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-U153, doi:10.1038/nature08250 (2009).

67 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502-1511, doi:10.1056/NEJMoa1306555 (2013).

68 Rehm, H. L. Evolving health care through personal genomics. *Nat Rev Genet* **18**, 259-267, doi:10.1038/nrg.2016.162 (2017).

69 Cizmarova, M. *et al.* New Mutations Associated with Rasopathies in a Central European Population and Genotype-Phenotype Correlations. *Ann Hum Genet* **80**, 50-62, doi:10.1111/ahg.12140 (2016).

70 Mcrae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438, doi:10.1038/nature21062 (2017).

71 Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genet Med* **18**, 696-704, doi:10.1038/gim.2015.148 (2016).

72 Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *Journal of the American Medical Association* **312**, 1880-1887, doi:10.1001/jama.2014.14604 (2014).

73 Chong, Jessica X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* **97**, 199-215, doi:10.1016/j.ajhg.2015.06.009 (2015).

74 Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*, doi:10.1038/gim.2017.246 (2018).

75 Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *bioRxiv*, doi:10.1101/201533 (2017).

76 Lane, K. B. *et al.* Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension. *Nature Genetics* **26**, 81-84 (2000).

77 Graf, S. *et al.* Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nature Communications* **9**, doi:ARTN 1416
10.1038/s41467-018-03672-4 (2018).

78 Hindorff, L. a. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

79 Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

80 Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* **50**, 559-571, doi:10.1038/s41588-018-0084-1 (2018).

81 Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725-1735, doi:10.1093/hmg/ddg180 (2003).

82 Lecerf, L. *et al.* An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease. *Hum Mutat* **35**, 303-307, doi:10.1002/humu.22499 (2014).

83 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Genetics* **46**, 61-64, doi:10.1038/ng.2826 (2014).

84 Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics* **31**, 67-76, doi:10.1016/j.tig.2014.12.003 (2015).

85 Gordon, C. T. *et al.* Long-range regulation at the SOX9 locus in development and disease. *J Med Genet* **46**, 649-656, doi:10.1136/jmg.2009.068361 (2009).

86 Gordon, C. T. *et al.* Identification of Novel Craniofacial Regulatory Domains Located far Upstream of SOX9 and Disrupted in Pierre Robin Sequence. *Hum Mutat* **35**, 1011-1020, doi:10.1002/humu.22606 (2014).

87 Dathe, K. *et al.* Duplications Involving a Conserved Regulatory Element Downstream of BMP2 Are Associated with Brachydactyly Type A2. *Am J Hum Genet* **84**, 483-492, doi:10.1016/j.ajhg.2009.03.001 (2009).

88 Amiel, J., Benko, S., Gordon, C. T. & Lyonnet, S. Disruption of long-distance highly conserved noncoding elements in neurocristopathies. *Ann N Y Acad Sci* **1214**, 34-46, doi:10.1111/j.1749-6632.2010.05878.x (2010).

89 Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics* **25**, 157-165, doi:10.1093/hmg/ddw205 (2016).

90 Melton, C., Reuter, J. a., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, doi:10.1038/ng.3332 (2015).

91 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).

92 Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, 818-821, doi:10.1038/ng.3335 (2015).

93 Zhang, W. *et al.* A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nature Genetics* **50**, 613-+, doi:10.1038/s41588-018-0091-2 (2018).

94 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **13**, 216-221, doi:10.15154/1149697 (2014).

95 Turner, Tychele N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *The American Journal of Human Genetics* **98**, 58-74, doi:10.1016/j.ajhg.2015.11.023 (2016).

96 Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature Genetics* **50**, doi:10.1038/s41588-018-0107-y (2018).

Chapter 2: *De novo* mutations in regulatory elements contribute to severe neurodevelopmental disorders

Introduction

The Deciphering Developmental Disorders Study

The Deciphering Developmental Disorders (DDD) study is a collaboration between the National Health Services (NHS) Regional Genetics Centres in the United Kingdom and Ireland and the Wellcome Trust Sanger Institute which has enrolled nearly 14,000 patients with severe developmental disorders (DD)[1]. Patients enrolled in the DDD study do not have a molecular diagnosis for their disorder after testing by clinical microarray, single-gene tests, or gene panels by a senior clinical geneticist in the NHS. The sheer number of genes implicated in DD (2,045 as of this writing https://decipher.sanger.ac.uk/ddd#ddgenes), and the often-overlapping phenotypes from different genes (for example, intellectual disability or developmental delay) have made diagnosing this class of rare disorders very challenging. The rate of molecular diagnosis for these disorders has been estimated to be between 20% and 60% dependent on a number of factors, including the specificity of the clinical presentation and the technology used to detect mutations, underscoring the need for greater understanding in the underlying genetic architecture and improvements in diagnosis[2].

Successful molecular diagnosis may provide opportunities for treatment. A number of metabolic disorders, provided they are correctly diagnosed, are amenable to treatment by vitamin or mineral supplementation. In Duchenne Muscular Dystrophy, exon-skipping therapies have been used in clinical trials to restore the open-reading frame of the *DMD* gene. Approaches such as these may be amenable to other developmental disorders, and successful diagnosis is a critical first step to understanding the mechanisms underlying these disorders, and conducting clinical trials. Furthermore, knowing the likely genetic cause of the disorder is essential for counselling families considering having more children. For example, the recurrence risk for a family whose child carries a *de novo* mutation (DNM) in a dominant developmental disorder gene is far lower than for a family where the child has inherited a rare and damaging genetic variant from each parent in a recessive developmental disorder gene.

The DDD study has performed exome sequencing on approximately 14,000 affected individuals. Nearly 10,000 of the affected individuals were recruited as full trios (mother, father, and child). The trio study design allows for detection of both rare recessive variants and *de novo* mutations, which are more challenging to detect from singleton affected patients. A subset of patients were also tested using array CGH, which allows for detection of large copy number variations (CNVs). The affected individuals and parents enrolled in the study have undergone systematic phenotyping using the Human Phenotype Ontology[3] as well as growth parameters including height, weight, and head circumference, and developmental milestones.

Large copy-number variations, rare genetic variants, and *de novo* mutations disrupting protein-coding genes have been identified as likely causal in nearly one-third of patients. The majority of these diagnosis are from *de novo* mutations in protein-coding genes. In the first 4,000 DD trios analysed, 25% of patients carried a likely causal *de novo* mutation in a known developmental disorder gene, and an additional 17% were predicted to carry pathogenic mutations in genes not yet robustly associated to developmental disorder[4]. However, after accounting for pathogenic de novo mutations and inherited variants, the majority of the individuals in the DDD study remain undiagnosed. For this reason, I sought to explore the contribution of *de novo* mutations outside of the protein-coding sequence.

Assaying non-coding sequence in the Deciphering Developmental Disorders Study

The majority of disease-associated common SNPs lie in intergenic or intronic regions, albeit with low effect sizes. In severe Mendelian disorders, rare sequence and structural variants in relatively few regulatory elements have been causally linked to Mendelian disorders[5-7] (reviewed in Chapter 1, and *Mathelier et. al, 2015, Zhang and Lupski, 2015,* and *Spielmann and Mundlos, 2016*). These pathogenic regulatory variants can act by loss-of-function[8-11] or gain-of-function[12,13] and most act dominantly, with a few exceptions[14]. These regulatory elements can lie far from the gene they regulate. For example, sequence variants in an evolutionarily conserved regulatory element located 1Mb from its target gene, *SHH*, can cause polydactyly[12]. As a consequence, it can be challenging to identify the gene whose regulation is being perturbed by an associated regulatory variant[15-17]. Moreover, the contribution of highly penetrant mutations in regulatory elements to genetically heterogeneous rare diseases, such as neurodevelopmental disorders, has not been firmly established.

To assess the contribution of variation outside the coding regions to severe DD, more than 6,000 putative regulatory elements were included in the exome capture. These elements were derived from three major categories: 4,307 highly evolutionarily conserved non-coding elements[18], 595 experimentally-validated enhancers[19], and 1,237 putative heart enhancers[20], together covering 4.2Mb of sequence (see Methods).

Variant effect prediction in the non-coding genome

A number of computational tools have been developed that combine evolutionary conservation, chromatin modifications in different tissues, transcription factor binding sites, and other genomic features to predict the impact of variation genome-wide. Combined Annotation Dependent Depletion uses a support vector machine (SVM) to predict pathogenicity of coding and non-coding variation and has proved to be a powerful tool in the coding regions, particularly for improved stratification of missense variation[21]. FATHMM-MKL also uses an SVM to predict the pathogenicity of variation genome-wide, but instead of combining all features into one model, constructs a kernel for each different feature group (Histone modifications, TF binding sites, Open Chromatin, evolutionary conservation, GC content, amongst others)[22].

In contrast to CADD and FATHMM-MKL, which attempt to model pathogenicity genome-wide, Genomiser addresses the more precise use-case of variant prioritisation in Mendelian disease[23]. Genomiser relies on a set of 453 non-coding variants previously associated with Mendelian disorders, collected through literature review and in some cases, computational prediction. Genomiser employs a similar set of features to CADD and FATHMM-MKL, including evolutionary conservation, histone modifications, and open chromatin data and uses a Random Forest model to make pathogenicity predictions.

Other methods have been developed to predict disruptions to TFBS, and integrate these disruptions with evolutionary conservation data to infer pathogenicity. One such method, DeepSEA uses a convolutional neural network (CNN) to predict TF binding intensity based on sequence context alone, using peak intensity from the ENCODE project to train the model[24]. While this method has been shown to outperform motif-based predictions of TF-binding, neural networks and other machine learning models that predict on sequence data alone are more challenging to interpret than models with structured feature sets.

CADD, Fathmm-MKL, Genomiser, and other non-coding variant effect predictors have purported to improve variant effect prediction for non-coding variation, but the utility of these scores has not been robustly verified in the non-coding genome to the same degree as tools such as Variant Effect Predictor[25] (VEP), SIFT[26], PolyPhen[27], or CADD[21] which are used in interpretation of coding variation. This difference is in large part due to the paucity of disease associated regulatory variation to verify the models. In the case of Genomiser, which was trained using curated disease variants, testing on an independent set of disease-linked variants will not be possible until more variants independent of the training set have been reported. Furthermore, effect sizes in the non-coding genome may be smaller in general, as suggested by the greater burden of non-coding variation in common disease, making the detection of damaging variation more difficult.

An alternative approach to validate the utility of these models is to use selective constraint, which can be inferred from the allele frequency of variation in modern humans. Sites under negative selection will on average harbour fewer variants, and variants at lower allele frequencies, than sites evolving neutrally[28-30] This approach has been successfully applied in the coding regions of the genome—CADD, SIFT, and PolyPhen scores show a strong correlation with pathogenicity predictions and inferred strength of purifying selection[30]. In this chapter, the mutability adjusted proportion of singletons (MAPS[30]) method is used to assess the relationship between non-coding variant effect predictors and selective constraint as a proxy for their potential in non-coding variant prioritisation. The topic of variant effect prediction and selective constraint in the non-coding genome is explored in greater detail in Chapter 3 using whole genome-sequence data from >28,000 individuals.

Methods

Recruitment in the Deciphering Developmental Disorders (DDD) project

At the time of this analysis, 7,930 trios (mother, father, affected child) were recruited through NHS Clinical Genetics Centres in the U.K. and Ireland and had been sequenced. The families gave informed consent for participation and the study has research ethics approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). DECIPHER was used to collect and store clinical data including family history, growth measurements, developmental milestones, and structured phenotypic descriptions (using the Human Phenotype Ontology[3]). Saliva was collected from all family members, and blood was collected from affected probands and DNA was extracted for sequencing.

Sequencing in the DDD Cohort

Genomic DNA was extracted from patient saliva and sheared into 150bp fragments. Libraries were created following standard Illumina paired-end protocols in the Wellcome Trust Sanger Institute sequencing facility. SureSelect RNA baits were used to do the exome pulldown. Enriched libraries were sequenced on Illumina HiSeq at the Wellcome Trust Sanger Institute using 75-base paired-end sequencing.

Alignment, Variant Calling, and Quality Control

Mapping of short-read sequences was carried out using the Burrows-Wheeler Aligner[31] (BWA; version 0.59) algorithm with the GRCh37 1000 Genomes Project phase 2 reference. The Genome Analysis Toolkit[32] (GATK; version 3.1.1) and SAMtools[33] (version 0.1.19) were used for sample-level BAM improvement. Ensembl Variant Effect Predictor (VEP[25]) based on Ensembl gene build 76 was used to annotate variants and, in coding regions, the transcript with the most severe consequence was selected. I identified a trinucleotide specific error mode (GTN->GGN) that introduced false positives which was corrected by strict strand filtering (FS < 20). I determined the number of variants called per individual, and excluded unaffected parents with variant counts on the extremes of the distribution (top 1% and bottom 1%). Across the 7,080 unaffected parents that passed quality control filters, I identified 1,520,250 unique variants in the targeted non-coding elements and coding regions.

De novo mutation calling

De novo mutations were called as described in McRae et al, 2017, excluding SNVs and indels with posterior probability <0.00781 as annotated by DeNovoGear[34].

Defining targeted non-coding elements

The placental mammal 28-way phastCons score[35] was used to select the top 4,432 conserved non-coding elements with no overlap with RefSeq genes (downloaded from UCSC on August 4th, 2010). Using the VISTA enhancer browser[19], all 622 putative enhancers with evidence of *in vivo* activity in developing mouse embryos were downloaded on August 3rd, 2010. At the time the capture was designed, the observation had been made that heart enhancers are depleted among ultra-conserved elements[20]. As heart defects are the largest group of non-CNS abnormalities in the DDD cohort ultra-conserved elements were supplemented with an early annotation of heart enhancers. These putative heart enhancers were kindly provided by Axel Visel based on ChIP-seq of p300 in human fetal heart described in May et. al, 2012 in GRCh36 coordinates, mapped over to GRCh37[36]. Collectively, these elements cover approximately 4.6 megabases of total sequence. First, elements were filtered to exclude any targeted sequences with less than 10x coverage across the DDD data set. Second, any elements previously annotated to be non-coding, but classified as protein-coding in Gencode v19[37] were removed. Finally, any elements less than 50bp in length were excluded. After filtering, 4,307 conserved elements, 595 enhancer elements and 1,237 putative heart enhancers remain.

Defining intronic control sequences

The exome baits designed to capture the coding regions frequently have considerable overlap with non-coding intronic regions. To define a set of putative well-covered introns, a 10bp buffer was added upstream and downstream of all gencode v19 coding sequence (to avoid classifying any critical splice sites in the control introns) and this coding sequence was subtracted from the exome probes. Furthermore, any introns within known developmental disorder genes (the DDG2P gene set[38]) were excluded. This set of control introns filtered to include only elements 30bp in length or larger with >30x coverage.

Evolutionary conservation using phastcons and phylop

The degree of evolutionary conservation across vertebrates at the element level was calculated using the phastcons vertebrate 100-way score. Scores were retrieved in R using the Bioconductor[39] package phastCons100way.UCSC.hg19[35] (Siepel et. al, 2005).

PhyloP scores represent the –log10 p-value that a given nucleotide is evolving neutrally[40] (Pollard et. al, 2010). I used a tabix file of pre-computed PhyloP vertebrate 100-way scores for every site in the genome in order to annotate the DNMs observed in exome-negative probands, exome-positive probands and the simulated null model.

Functional genomic annotation using the Roadmap Epigenome Project data

Data from DNase hypersensitivity assays (broadPeak set, FDR 1%) were downloaded from the Roadmap Epigenome Project[41] ftp site (http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/) in order to predict regulatory function and tissue-specificity in the enhancers and conserved noncoding elements. The GenomicRanges Bioconductor[39,42] package was used to intersect DHS peaks with the elements sequenced in this analysis. All code used in this analysis can be found at https://github.com/pjshort/DDDNonCoding2017.

Chromatin state predictions (chromHMM 15-state model[43]) for 111 different tissue types were downloaded from the Roadmap Epigenome Project[41] (REP) ftp site (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/indivModels/default_init/). I considered a CNE to be inactive in a given tissue if it was completely contained within a chromHMM segment described as Quiescent, Heterochromatin, or Polycomb Repressed ("9_Het", "13_ReprPC", "14_ReprPCWk", and "15_Quies") in the 15-state model. Using the GenomicRanges[42] Bioconductor package and coding sequence from gencode v19, I calculated the distance of each active and broadly inactive element to the nearest exon or transcription start site. All code used in this analysis can be found at https://github.com/pjshort/DDDNonCoding2017.

Trinucleotide mutation rate model with CpG-methylation status

The germline mutation rate model described in *Samocha et. al, 2014* based on tri-nucleotide context[44] was adapted to include a correction at CpG sites for methylation status. This method models the null mutation rate at a given site as a Poisson rate parameter that is dependent on the tri-nucleotide context, where the second base is mutated. I fit a linear model to the ratio of observed/expected variants at MAF <0.1% in CpG sites based on their methylation status in embryonic stem cells. For all CpG sites, I corrected the tri-nucleotide mutation rate based on the methylation status to produce a methylation-aware mutation rate model. As the sum of Poisson random variables is Poisson, the rate parameter for a given element, or set of elements, can be determined by summing the mutation rate for each individual site. Simulated mutations were based on the same tri-nucleotide mutation framework and implemented in an R software package: https://github.com/pjshort/DenovoSim.

Statistical testing for mutational burden

The p-value for the number of observed *de novo* mutations compared to expected is calculated in R as:

```
ppois(n_obs - 1, lambda = mu, lower.tail = FALSE)
```

where n_obs is the number of observed mutations within an element and mu is the mutability of the element(s) being tested (under the null model described above) multiplied by the number of probands. The burden testing I performed across subsets of elements and phenotypes included multiple nested hypotheses that were accounted for with a conservative Bonferroni-adjusted p-value threshold based on the number of explicit and implicit tests.

Gene-target prediction

I used four different methods of gene target prediction to link CNEs and enhancers to putative target genes.

The first method, Genomicus, predicts gene targets based on evolutionary conservation with nearby genes. Genomicus determines the extent to which each CNE is within the same syntenic block with nearby genes across a number of vertebrate species and predicts one or more targets[16]. The Genomicus method produces at least one prediction for 90% of CNEs (approximately 1/3 of these are the closest gene).

The second method, described in *Shooshtari et. al, 2016*, compares DNase hypersensivity at each CNE to expression of nearby genes in 56 different tissues (using RNA-seq) to search for CNE-gene pairs that show a correlation between DNase signal and gene expression[45]. This method produces statistically significant predictions for only 28% of CNEs in our set and is likely underpowered to detect elements that are active in specific tissues or timepoints.

The third method is to link CNEs to putative target genes using chromatin interaction data (Hi-C) in two different regions of the fetal brain derived from Won et. al, 2016[46]. Using Hi-C data is the most direct and tissue specific of all of the prediction methods used, but the prediction is sparse (26% of CNEs with evidence for fetal brain activity have a predicted target).

The fourth method used is a simple heuristic to choose the gene with the closest TSS (for intergenic elements) or the gene containing the element (for introns). Choosing the closest gene allows us to make a prediction for 100% of elements, but comparison with chromatin conformation and DHS-based methods has shown that the closest gene is likely the target 7% and 12% of cases[45,47].

I used the Genomicus, DHS, and Hi-C predictions to generate aggregated predictions which I considered 'high confidence' if predicted by at least two of the three methods.

To assess the pair-wise concordance, I took the set of CNEs for which at least one gene target was reported in both methods and tested how frequently both methods identified the same gene within the set of predicted targets.

Transcription factor binding site analyses

The JASPAR2016 and TFBSTools Bioconductor packages[48] were used to retrieve position weight matrices for 454 human transcription factors (TFs). Analyses in this chapter focus on the 202 TFs predicted to be expressed in the brain (cortex-expressed from GTEx dataset[49]).

A custom R package called 'denovoTF' (https://github.com/pjshort/denovoTF) was written to predict any change in TF binding at sites where DNMs were observed or simulated. This analysis works by scanning the reference and alternative sequence for all 202 PWMs and comparing predicted binding events on both sequences. By comparing the potential binding affinity for ref and alt sequences, I can predict loss of binding (alt binding < ref binding), gain of binding (alt binding > ref binding), and silent (no difference). 'Silent' DNMs fall into two classes: those for which binding is predicted on both reference and alternate, but strength of binding is unchanged, and those which do not lie in a predicted TF binding site.

The analysis of motif enrichment (AME) tool from the meme suite was used to identify a subset of PWMs that was significantly enriched in the fetal brain active elements[50]. Comparing the fetal brain active CNEs to the fetal brain inactive CNEs returned a set of 90 transcription factors, of which 45 were expressed in the brain and had PWMs available in JASPAR2016[48]. This analysis was performed on the meme-suite web server using the following command:

```
ame --verbose 1 --oc . --control meme_chromHMM_fb_inactive_all.fasta --bgformat 1 --scoring avg --method ranksum --pvalue-report-threshold 0.05 meme_chromHMM_fb_active_all.fasta db/JASPAR/JASPAR_CORE_2016.meme
```

In order to test for enrichment of loss of binding or gain of binding events in the observed DNMs, I compared predicted impact on TF binding in observed DNMs to 1,000 simulations of mutations across the 2,613 fetal brain active elements for 6,147 probands.

Power calculations at different study sizes

I used the tri-nucleotide null model described previously in order to estimate our power to detect disease-associated elements. Parameters impacting power include the fold enrichment for disease-causing mutations in the DDD cohort (proportional to the incidence of severe developmental disorders with a genetic basis in the population), the proportion of mutations within a true disease-associated element expected to be pathogenic, the penetrance of such mutations, the size/mutability of the elements tested, and the number of trios analysed. In order to estimate the power across different study sizes, I fixed the remaining parameters as follows: 120-fold enrichment for disease-causing mutations, proportion of mutations expected to be pathogenic at 8% (lower bound estimate for coding regions), penetrance at 100%, and the elements tested were the 2,613 fetal brain active conserved non-coding elements. Code for power analysis can be found in the R script: https://github.com/pjshort/DDDNonCoding2017/analysis_notebooks/ Figure4_maximum_likelihood_and_genome_estimate.Rmd.

Estimating the genome-wide burden of DNMs in fetal brain active elements contributing to severe DD

First, I intersected all targeted non-coding sequence, irrespective of original class, with fetal brain DHS peaks. I used the phastcons100 score (scores retrieved in R using the Bioconductor package phastCons100way.UCSC.hg19[35] (Siepel et. al, 2005) to rank these elements by evolutionary conservation. The ratio of observed/expected DNMs was computed with a sliding window across the elements (window size of 1000 elements, shift of 100 elements). This approach resulted in a median of 62 DNMs expected in each bin (minimum 51, maximum 68) which was compared to the observed number of DNMs. I fit a logistic regression to the excess observed/expected in each window, setting any window with observed less than expected to have excess of zero. I used the logistic regression fit on the CNEs sequenced in our analysis to predict the burden of DNMs in this genome-wide set.

Modelling the likelihood of different proportions of elements and sites with monoallelic disease mechanism given observed data

In order to test the likelihood of different models of dominant disease mechanism within the non-coding space I adapted the power calculation framework described above to test the probability of observing our data across two different parameters: the number of elements (out of 2,613) with a dominant disease mechanism and the proportion of mutations expected to be pathogenic. I tested the likelihood of observing 286 DNMs, 25 recurrently mutated elements, and zero elements at genome-wide significance while systematically varying two parameters: the proportion of mutations expected to be pathogenic parameter from 0.01% to 10.0% in increments of 0.01% and the proportion of elements with true disease-associations (from 0 to 2,613 in increments of 5). In this analysis, the remaining parameters were held constant: 120x enrichment for pathogenic mutations, penetrance at 100%, testing 2,613 fetal brain active conserved non-coding elements, and number of trios at 6,147. Code for power analysis can be found in the R notebook: https://github.com/pjshort/DDDNonCoding2017/analysis_notebooks/Figure4_maximum_likelihood_and_genome_estimate.Rmd.

Results Section 2.1: Assessing the role of *de novo* mutations in regulatory elements in severe DD

Results Section 2.1.1. Modelling the germline mutation rate in the non-coding genome

Assessing the contribution of *de novo* mutations (DNMs) to developmental disorders relies on an accurate null model of the germline mutation rate in the absence of disease association. Local sequence context has been shown to have a significant influence on the per-base mutation rate[44]. For example, CpG dinucleotides have an approximately 10x greater mutation rate than other dinucleotides due to spontaneous deamination when the C is methylated. This spontaneous deamination results in a high rate C to T transitions, and is reflected in an overrepresentation of polymorphisms at CpG sites, and a higher rate of DNMs[30,44,51,52].

A well-established mutation rate model incorporating the sequence one base upstream and downstream of a nucleotide of interest, referred to as the 'triplet context', has been shown to significantly outperform sequence-agnostic models in predicting the number of DNMs or level of rare variation in a protein-coding gene. This model has been used extensively in assessing the burden of DNMs in the autism spectrum disorder (ASD) patients and DD patients[53], and to predict the level of expected rare variation in protein-coding genes in order to detect genes under purifying selection[30].

However, the widely used tri-nucleotide mutation rate model does not account for methylation status. Unmethylated CpGs have a substantially lower mutation rate. As a result, loci with high levels of methylation relative to the genome-wide average for CpG dinucleotides would have a higher than expected mutation rate. Conversely, loci with low levels of methylation at CpG dinucleotides (for example, promoters) would have a much lower mutation rate than predicted by the trinucleotide model. Thus, I adapted this tri-nucleotide model to include CpG methylation genome-wide. There is a strong correlation between the observed number of rare variants at CpG sites and their methylation levels in either embryonic stem cells (ESCs) or sperm (Figure 1A). This methylation-aware model better accounts for levels of rare variation observed in the set of non-coding elements sequenced in DDD (**Figure 1B**).

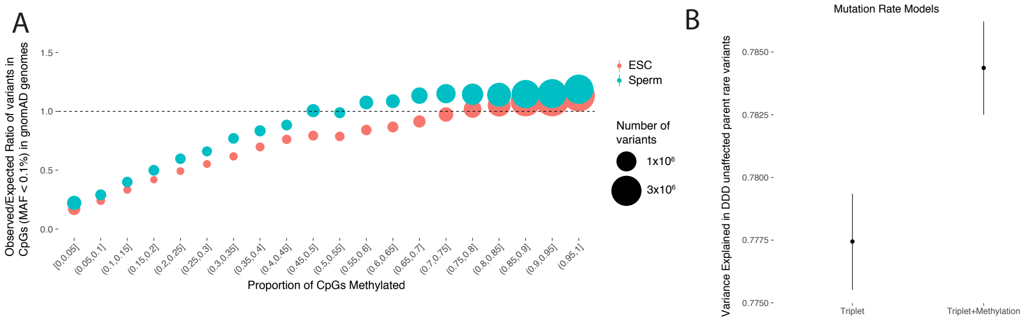


Figure 1 **Methylation-aware mutation rate model.** (a) Methylation proportion at a given CpG site in Embryonic Stem Cells (ESC) and sperm correlates with observed rare variation in gnomAD. (b) Incorporating CpG methylation proportion into mutation rate model improves model fit to rare variant data.

Results Section 2.1.2. Selective constraint acting on non-coding elements

I reasoned that if the non-coding elements sequenced in DDD were contributing to severe developmental disorders, there should be evidence for negative selection in these elements. I first assessed how much purifying selection had skewed allele frequencies in non-coding elements using the mutability-adjusted proportion of singletons (MAPS) metric[30] in 7,080 unrelated, unaffected DDD parents. I tested six different element classes: introns, heart enhancers, validated enhancers, conserved non-coding elements (CNEs), protein-coding genes, and known DD-associated genes. The validated enhancers from the VISTA enhancer browser vary across the spectrum of evolutionary conservation, while the heart enhancers are poorly conserved, consistent with previous reports[20], and the CNEs show high levels of evolutionary conservation (Figure 2A). The introns and heart enhancers show little evidence of purifying selection, while the experimentally-validated enhancers and CNEs are constrained to a similar degree to protein-coding genes, but less than known DD-associated genes (Figure 2B), consistent with evolutionary conservation maintained by purifying selection as has been previously reported[54,55].

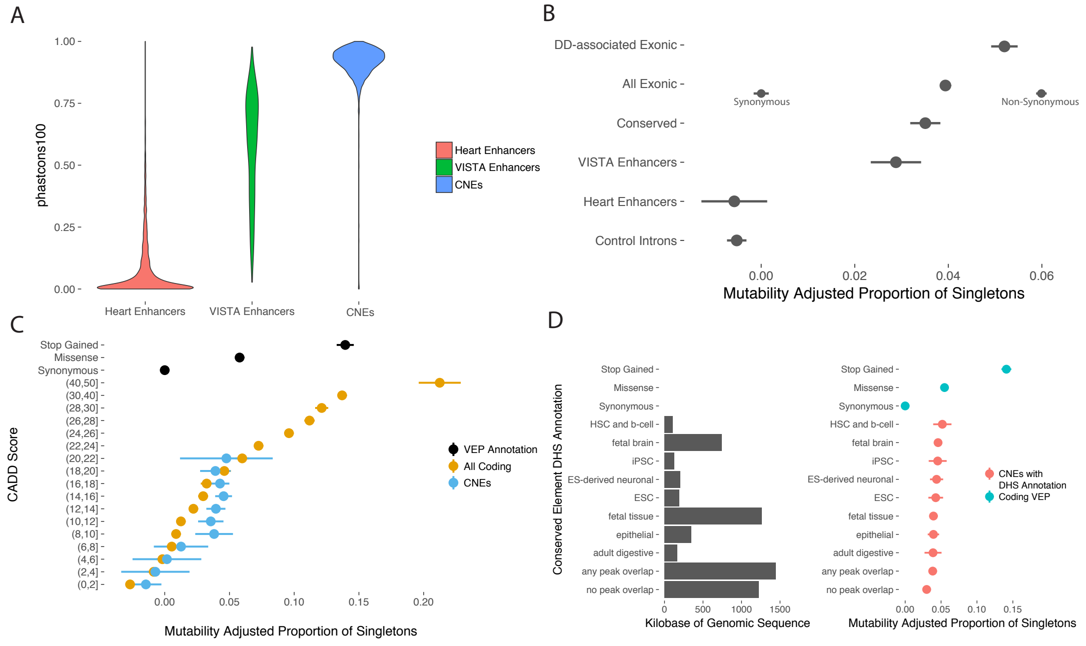


Figure 2 Selective constraint in targeted non-coding elements. (a) Evolutionary conservation score (phastcons100[35]) for CNEs (n=4,307), experimentally validated enhancers (n=595), and putative heart enhancers (n=1,237). (b) Strength of selection (MAPS metric, mean and 95% CI represented by dot and bars) in targeted non-coding elements compared to protein-coding regions, where 'Exonic' refers to all variation within protein coding exons. Stratification based on synonymous/non-synonymous consequence displayed on the same row to illustrate power of even a simple discriminator. Introns and putative heart enhancers show little evidence of purifying selection while CNEs show selection on par with all genes, but less than known DD genes. (c) Using CADD to stratify coding and non-coding variants observed in unaffected parents differentiates neutral variation from weakly and strongly constrained sites in coding regions, but fails to identify non-coding variation with selection pressure on par with protein-truncating variants (stop gained). (d) Sites overlapping a DNase I hypersensitive site (DHS) in at least one tissue are under stronger purifying selection than sites not overlapping a DHS.

Statistical power to detect functionally relevant variants in protein-coding genes is strengthened considerably by stratification of variants by their likely impact on the encoded protein and variant deleteriousness metrics such as CADD[21]. I computed the MAPS within bins of CADD scores encompassing 1,520,250 variants in unaffected DDD parents to assess whether CADD was predictive of selective constraint. In protein-coding genes, the strong correlation between CADD score and strength of purifying selection enabled differentiation between variants that are neutral (synonymous sites), weakly constrained (missense variants), and highly constrained (protein-truncating variants). In CNEs, CADD differentiates neutral variation from variation under weak constraint (comparable to missense variants), but failed to identify highly deleterious variants with selective constraint on a par with protein-truncating variants (Figure 2C). Other deleteriousness metrics were assessed, but none were more informative than CADD (Figure 3).

I used DNase I hypersensitivity sites (DHS) in 39 tissues and chromHMM genome segmentation predictions in 111 tissues[41] to predict tissue activity for the targeted non-coding elements. Of the 4,307 CNEs sequenced, 4,046 (93.9%) were active in at least one of the 111 surveyed tissues while 261 (6.1%) were inactive or repressed in all tissues. Variants within a DHS peak in at least one tissue were under stronger purifying selection than variants not overlapping a DHS peak ($p = 0.019$), but I did not identify significant differences in selective constraint between different tissues (Figure 2D).

The non-coding elements sequenced in this study are not representative of regulatory elements genome-wide, and the MAPS metric used here is not the only method for detecting selective constraint. Evidence of selective constraint in the non-coding genome is explored in greater detail in Chapter 3 using genome-wide assays of putative regulatory function and deep whole genome sequences to test for evidence of selective constraint.

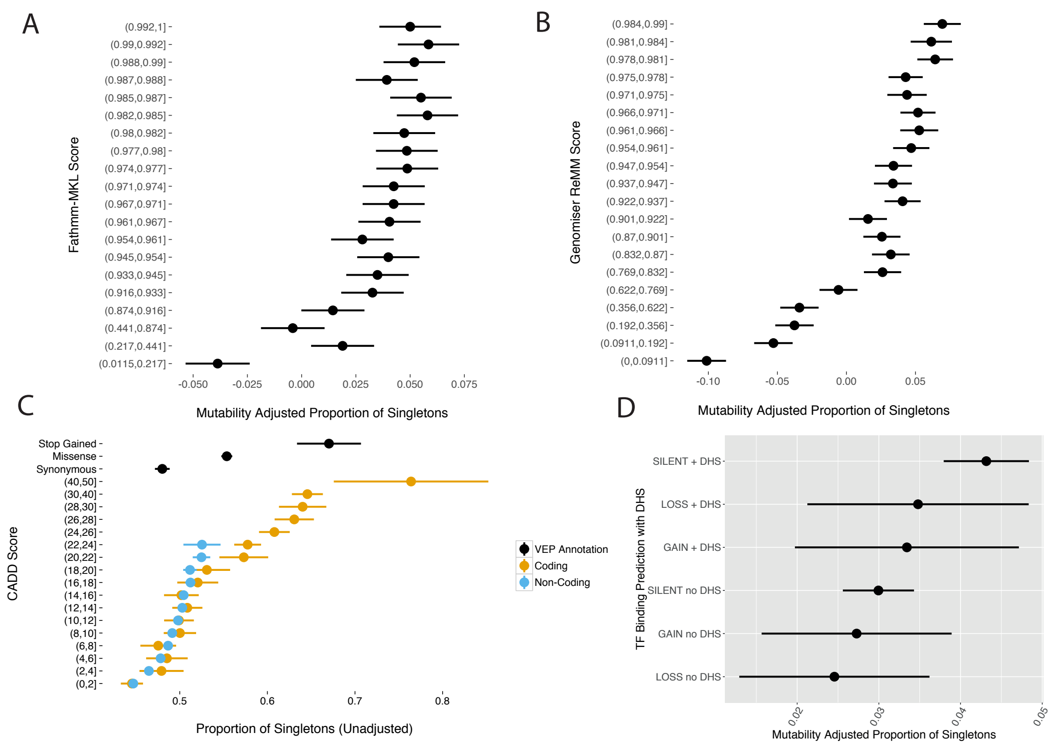


Figure 3 Assessment of variant deleteriousness metrics and selective pressure in conserved non-coding elements. Dots and bars represent the point estimate and 95% confidence interval, respectively for MAPS and proportion singletons. (a) Fathmm-MKL and (b) Genomiser separate benign variation (low MAPS score) from likely damaging variation (high MAPS score), but do not identify any classes of variation under strong selective constraint. (c) Validation of Figure 2C in main text using whole genome data from UK10K project. While CADD can identify coding variation under strong selective constraint (as measured by the proportion of singletons), CADD is unable to identify strongly constrained non-coding variants. (d) There was no significant difference in strength of purifying selection measured by MAPS between sites predicted to result in loss, gain, or no change of transcription factor binding.

Results Section 2.1.2. Fetal brain active ultra-conserved non-coding elements are enriched for mutations in exome-negative probands with neurodevelopmental phenotypes

To assess the contribution of DNMs in regulatory elements sequenced in the DDD families, I identified candidate *de novo* single nucleotide mutations in 7,930 trios using DeNovoGear (see Methods). I identified 1,691 'exome-positive' individuals with a likely pathogenic protein-altering DNM or inherited variant in a known DD-associated gene, with the remaining 6,239 being ‘exome-negative’. Using the methylation-aware mutation model, I compared the numbers of observed and expected DNMs in the targeted non-coding elements in these individuals. No significant DNM enrichment was observed in exome-positive probands in the targeted non-coding elements, demonstrating that the mutation model is reasonably well-calibrated and that a large proportion of exome-positive cases likely represent Mendelian syndromes caused by high-penetrance protein-coding mutations (Figure 4A). I note that the number of exome-positive individuals affords only limited power to reject modest mutation enrichment in the non-coding elements. Based on these results, I chose to focus on the 6,239 exome-negative individuals for subsequent analyses.

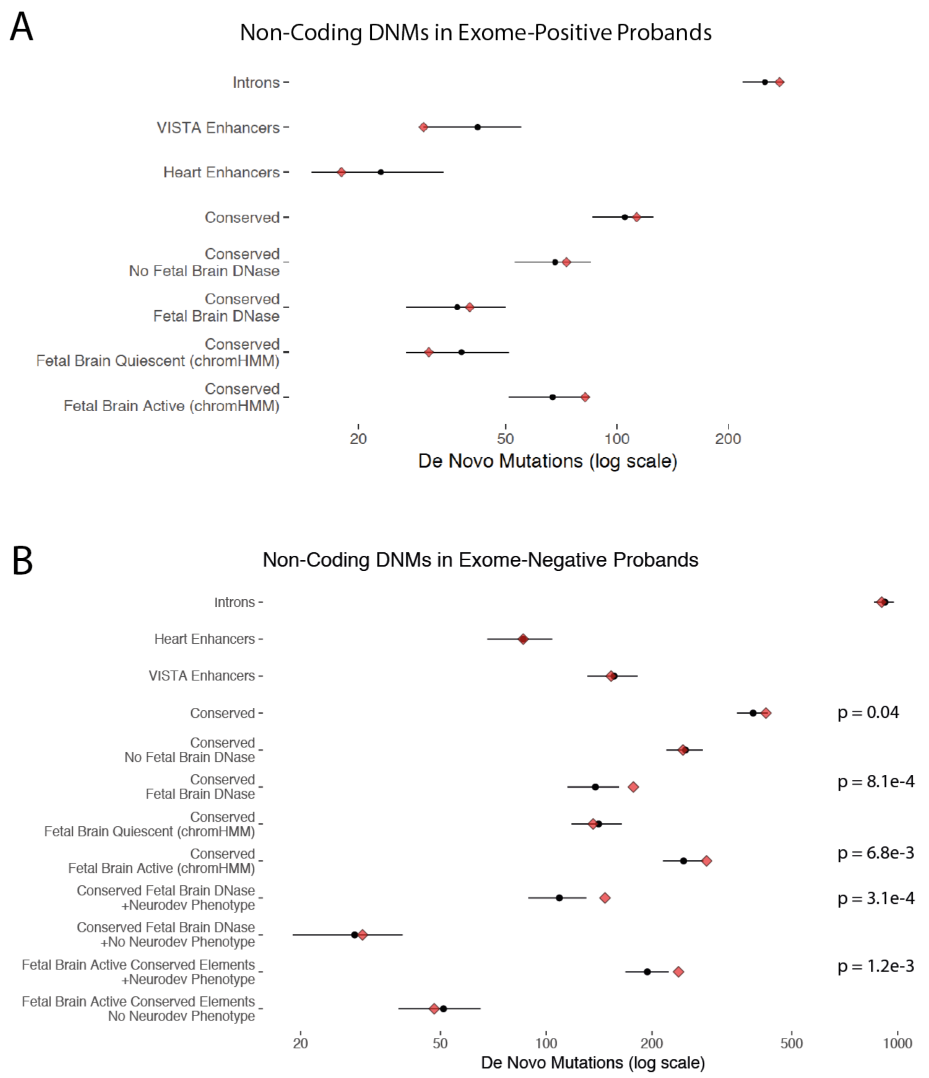


Figure 4 Non-coding mutations in exome-positive and exome-negative probands. Red diamonds indicate observed counts, while black circles and bars indicate expected count and 95% CI. (a) In

the 1,691 'exome-positive' probands, there is no evidence for a burden of *de novo* mutations in any of the non-coding element classes tested. (b) Enrichment of DNMs across element classes and functional annotations in exome-negative probands (n=6,239). Targeted CNEs showed a modest enrichment for DNMs (422 observed, 388 expected, p = 0.04) while heart enhancers, experimentally-validated enhancers, and control introns matched the null model. Observed enrichment is specific to CNEs predicted to be active in the fetal brain and to patients with neurodevelopmental disorders (238 observed, 194 expected, p = 1.2e-3). Confidence intervals and p-values derived from a Poisson distribution.

I found that the CNEs are nominally significantly enriched for DNMs (422 observed, 388 expected, p = 0.04), whereas experimentally-validated enhancers (153 observed, 156 expected, p = 0.605), heart enhancers (86 observed, 86 expected, p = 0.514), and intronic controls (901 observed, 919 expected, p = 0.728) were not enriched (Figure 4B).

Given the preponderance of individuals with neurodevelopmental disorders in our cohort but broad range of the tissue activity of the targeted CNEs, I focused on CNEs active in fetal brain. I observed a strong significant enrichment of DNMs within 2,077 fetal brain DHS peaks in CNEs (177 observed, 138 expected, p = 8.1e-4) but no enrichment in sites in CNEs falling outside of fetal brain DHSs (245 observed, 249 expected, p = 0.608) (Figure 4B). I also used chromHMM[43] predictions of fetal brain activity and again identified a significant enrichment of DNMs in the 2,613 fetal brain active CNEs (Figure 4B). Moreover, the DNMs observed in fetal brain active CNEs in exome-negative probands were at more highly conserved sites (Wilcoxon rank sum test on PhyloP 100-way score[40], p = 7.5e-4) compared to DNMs observed in exome-positive probands (Figure 5).

The excess of DNMs observed in fetal brain active CNEs is concentrated exclusively within the 79% of exome-negative probands with neurodevelopmental phenotypes (fetal brain DHS peaks: 147 observed, 109 expected, p = 3.1e-4, fetal brain active by chromHMM: 238 observed, 194 expected, p = 1.2e-3), with no significant enrichment observed in those without neurodevelopmental phenotypes (fetal brain DHS: p=0.413; fetal brain active by chromHMM: p=0.681) (Figure 4B). The highly significant and specific enrichment of DNMs in fetal brain active CNEs in exome-negative probands with neurodevelopmental disorders is robust to

Bonferroni correction for thirteen explicitly and implicitly tested hypotheses (Figure 6A). The fold-enrichment of DNMs is consistent with DNMs in fetal brain active CNEs comprising a mixture of 70-80% non-pathogenic DNMs and 20-30% of pathogenic DNMs.

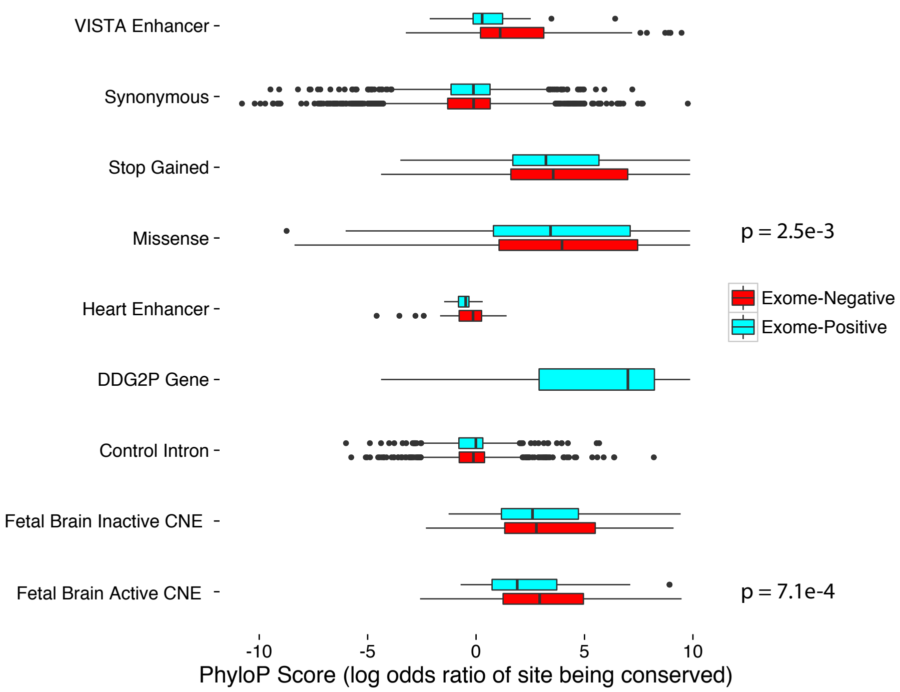


Figure 5 Evolutionary conservation of mutated base in conserved non-coding elements. DNMs in exome-negative probands show a greater degree of evolutionary conservation (measured by PhyloP score) compared to DNMs observed in exome-positive probands in two classes: fetal brain active CNEs (median 1.57 exome-positive, 2.85 exome-negative, n=368 mutations) and missense changes (median 3.43 exome-positive, 3.98 exome-negative, n = 6244 mutations).

Having identified a signal of DNM enrichment only within CNEs active in fetal brain, I re-evaluated the experimentally-validated enhancers with functional evidence for activity in fetal brain (N=383, 64%) and observed a nominally significant enrichment for DNMs only within the top quartile of evolutionary conservation (18 observed, 9 expected, p = 0.01) (Figure 6B). This result is suggestive that even for experimentally validated fetal brain enhancers, DNM enrichment is concentrated within elements with strong evolutionary conservation.

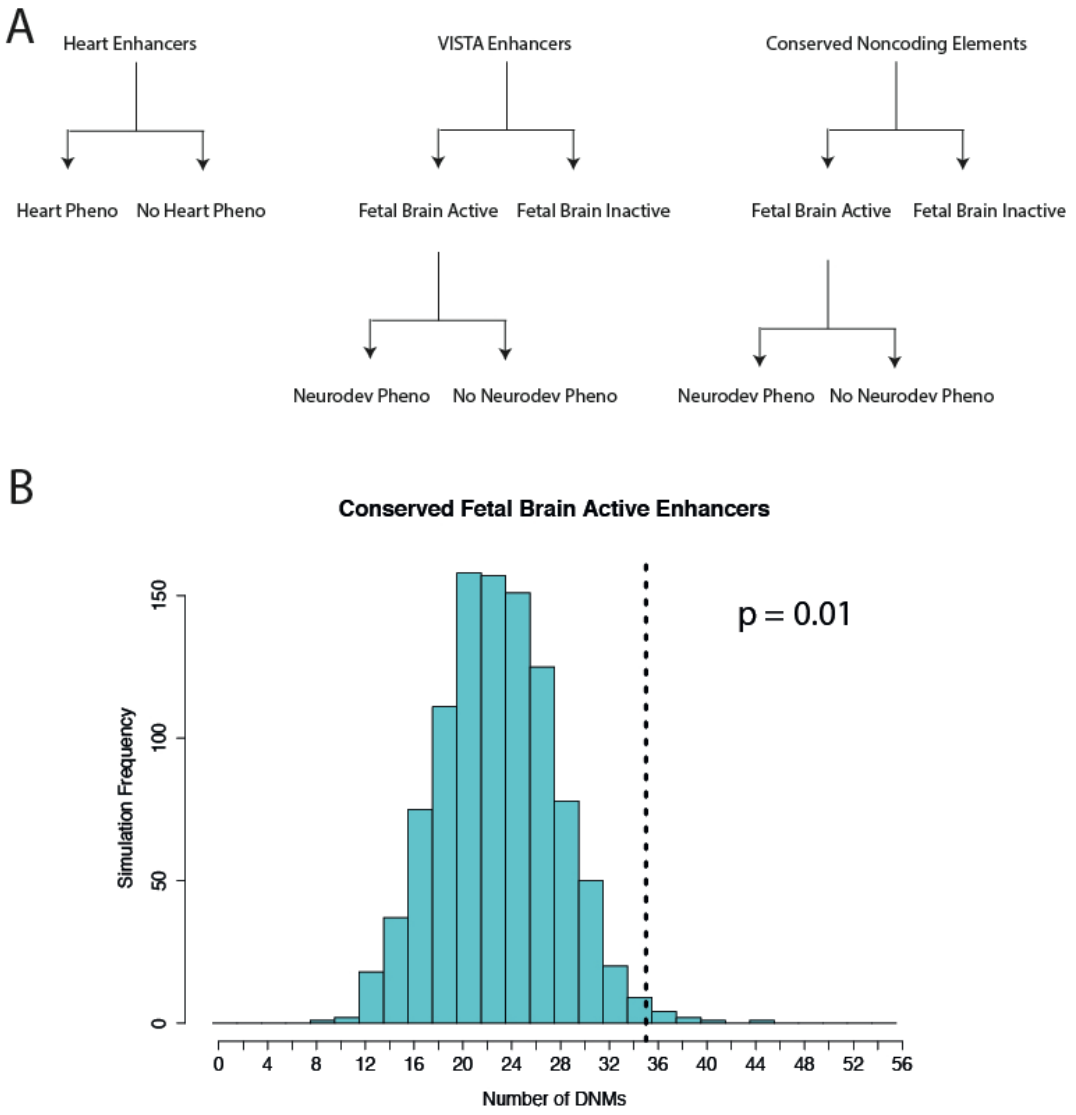


Figure 6 Hypothesis test enumeration and enrichment for mutations in highly conserved fetal brain active enhancers. *(a) We corrected for thirteen tests in order to account for the nested hypotheses based on element class and phenotype in this analysis. (b) Evolutionarily conserved fetal brain active enhancers (n=106) are enriched for DNMs in exome-negative probands.*

In addition to methylation at CpG sites, other genomic features have been previously associated with mutagenicity. To test whether the enrichment for mutations I observe is a result of hypermutability rather than disease association, I tested four genomic features previously associated with mutagenicity[56] for enrichment in non-coding elements with DNMs. I found no evidence that these genomic features were enriched in non-coding elements with DNMs (H3K27me3 χ^2-test p=0.4809, H3K9me3 χ^2-test p=0.1966, replication timing[57] Figure 7A, recombination rate[58] Figure 7B). To test for any as yet unknown factors causing differential mutability, I compared the levels of rare variation in fetal brain active and inactive CNEs in 7,509 deep whole genomes from the gnomAD consortium and found no evidence for a higher germline mutation rate in fetal brain active elements (Figure 7C, D). These results indicate that the enrichment of DNMs in regulatory elements in exome-negative probands is not likely to be the result of hypermutability not captured by the mutation rate model.

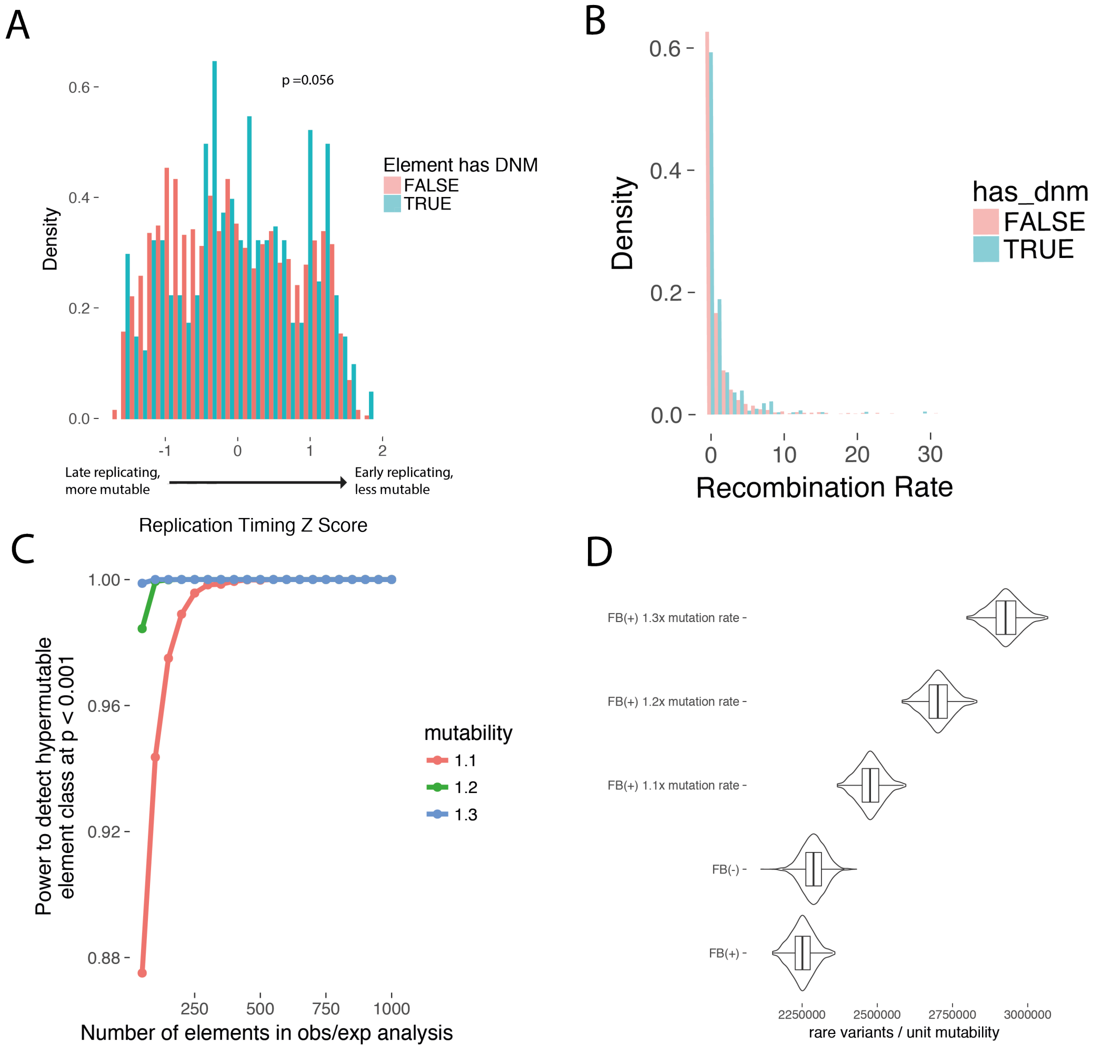


Figure 7 Genomic factors influencing mutation rate in non-coding elements. (a) Elements with *de novo* mutations observed in our study are not enriched in late-replicating regions or (b) in regions with higher recombination rate. (c) Levels of rare variation in deep whole genomes (n=7,509 non-Finnish Europeans) were used to estimate power to detect a hypermutability of 1.1X, 1.2X, or 1.3X. (d) The level of rare variation in the fetal brain active elements (n=2613, labelled FB(+)) is slightly lower than the fetal brain inactive elements (n=1694, labelled FB(-)), consistent with similar mutability between the two element sets with slightly stronger purifying selection in the fetal brain

active elements.

I next sought to estimate what proportion of the fetal brain active CNEs sequenced in the DDD patients act as enhancers. I compared a set of 617 high confidence brain enhancers from the VISTA enhancer database to two orthogonal enhancer annotations, FANTOM5[59] and EnhancerAtlas[60]. The high confidence enhancers from the VISTA database were also identified as enhancers in the FANTOM5 or EnhancerAtlas datasets, 12% and 36% respectively, providing an estimate for the sensivitity of these two datasets. Applying the same test to the CNEs, 6% of the fetal brain active CNEs were identified as enhancers by FANTOM5 and 28% were identified as enhancers by EnhancerAtlas. Taken together, these results suggest that at least half and possibly up to three quarters of CNEs are acting as enhancers.

In principle, linking regulatory elements to the gene(s) they regulate may improve power (by grouping distinct elements, and by combining burden of regulatory and coding mutations). Furthermore, predicting the target gene(s) for a given regulatory element may provide greater insight into mechanisms of pathogenic mutations in regulatory elements. I assessed four different methods for gene target prediction: Genomicus[35] (based on evolutionary synteny), correlation between DNase accessibility and gene expression[36], Hi-C in fetal brain[33], and choosing the closest gene. Genome annotations are rapidly evolving and the sensitivity and specificity of gene target prediction methods is not yet known. However, independent eQTL, eRNA and Hi-C data all suggest that the closest gene is often not the target of non-coding regulatory variation[32-34]. Across the four methods tested, the proportion of fetal brain active CNEs for which a target gene was predicted was 28% (fetal brain Hi-C), 48% (DHS-RNA correlation), 91% (evolutionary synteny), and 100% (closest gene). The pairwise-concordance between any two methods (given that both methods make a prediction) was between 17% and 35% (Figure 8A). Intersecting multiple independent methods may provide higher confidence predictions, but comes at a cost of sensitivity and therefore power. I did not identify any enrichment for DNMs in elements predicted to target known DD genes, likely dosage sensitive genes (pLI metric[22]), or genes differentially expressed in the brain (Figure 8B). Gene target predictions using fetal brain Hi-C were performed by Hyejung Won, a collaborator in the Geschwind Lab at UCLA.

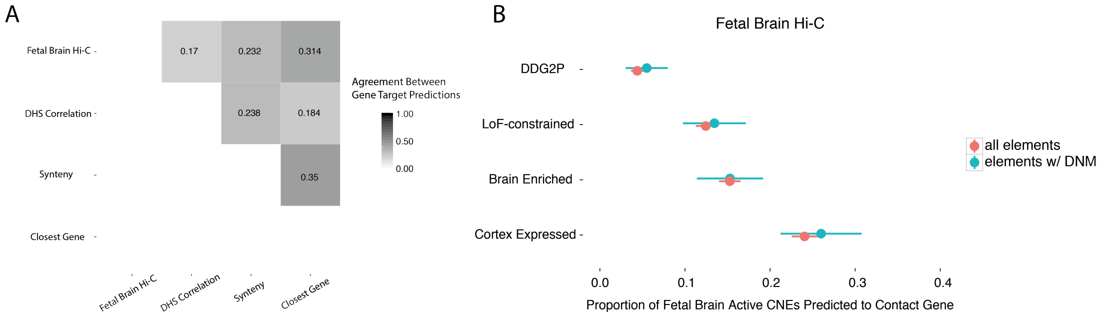


Figure 8 Gene target prediction for targeted non-coding elements. (a) Pairwise concordance between four different gene target prediction methods is low. (b) Using predicted targets from fetal brain Hi-C data, elements with an observed DNM in an exome-negative probands (n=286) do not show any bias toward any of the gene sets consistently implicated in neurodevelopmental disorders. Dots and bars represent the point estimate and 95% confidence interval, respectively.

I hypothesized that the mutations observed in DDD patients may be altering transcription factor binding. To this end, I assessed the impact of DNMs on a set of 45 transcription factor binding motifs enriched in fetal brain active CNEs (see Methods), and observed a nominally significant enrichment for DNMs predicted to increase binding affinity which did not survive multiple hypothesis correction (Figure 9). Given the number of DNMs identified, and the relative immaturity of *in silico* predictions of the impact of non-coding variation, it is not currently possible to determine precise mechanisms by which these DNMs contribute to DDs.

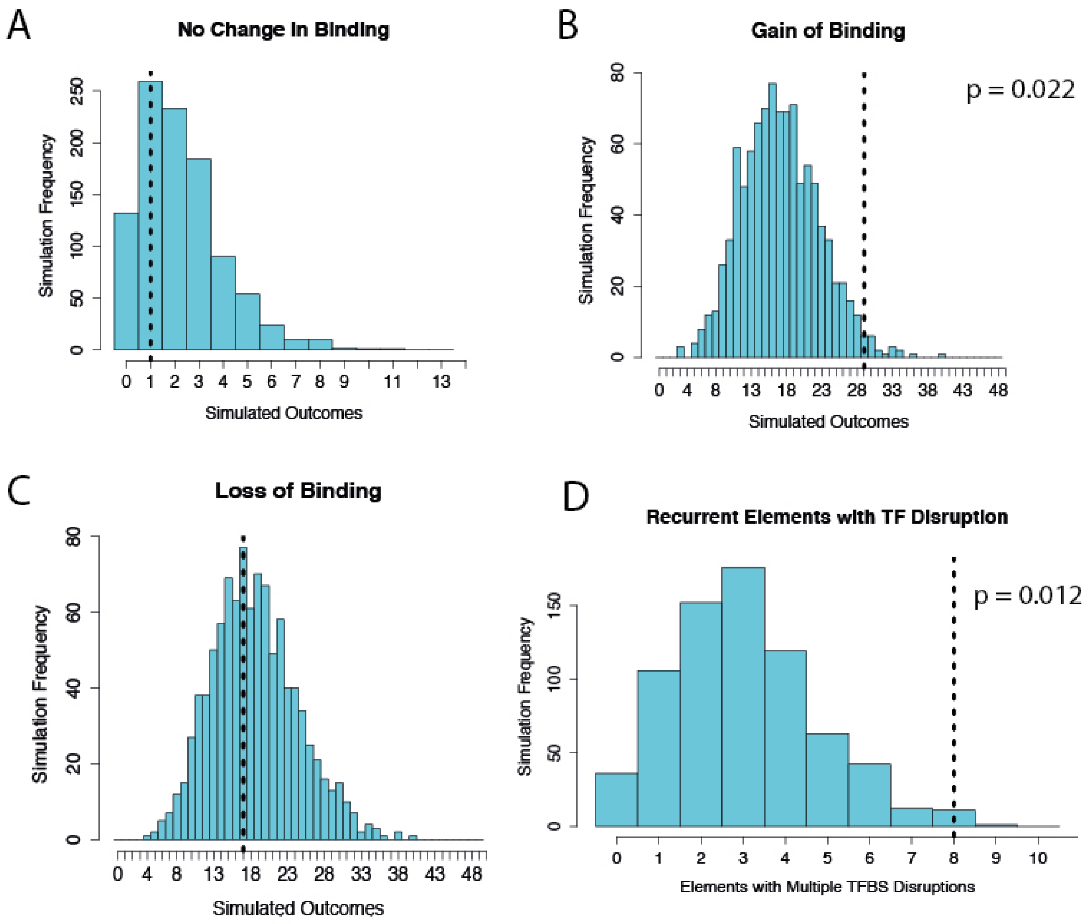


Figure 9 Predicted transcription factor binding site disruption. (a-d) Comparing predicted change in transcription factor binding for observed DNMs compared to null mutation model. Empirical p-values derived from comparison with mutations simulated from the null mutation model.

Results Section 2.1.3. Recurrently mutated regulatory elements

Observing mutations independently in multiple different families has led to the discovery of dozens of novel developmental disorder-associated protein-coding genes[4] . I applied the same approach to the set of targeted non-coding elements to test for recurrently mutated non-coding elements (two or more DNMs in unrelated individuals). I observed a significant excess of recurrently mutated elements in the fetal brain active CNEs and evolutionarily conserved enhancers compared to the expectation under the null mutation model (31 observed, 15 expected, p = 9.3e-5) (Figure 10A). However, no individual element exceeds a conservative genome-wide significance threshold of p<1.91e-5 (Bonferroni-correction for independent tests on 2,613 fetal brain active elements) (Figure 10B). Nonetheless, the set of thirty-one recurrently mutated elements provides a source of elements for medium and high-throughput functional assays, albeit with a high false-discovery rate (FDR of approximately 50%). I tested these thirty-one recurrently mutated using a saturation mutagenesis massively parallel reporter assay in order to gain a better understanding of the nucleotide-level patterns underlying these elements. Furthermore, a subset of these elements with robust evidence for fetal brain enhancer activities were tested in mouse knockout studies in collaboration with Evgeny Kvon, Diane Dickel, Len Pennachio, and Axel Visel. These results are discussed in detail in Chapter 4, and these assays and other potential functional follow-ups are detailed in the Discussion of this chapter.

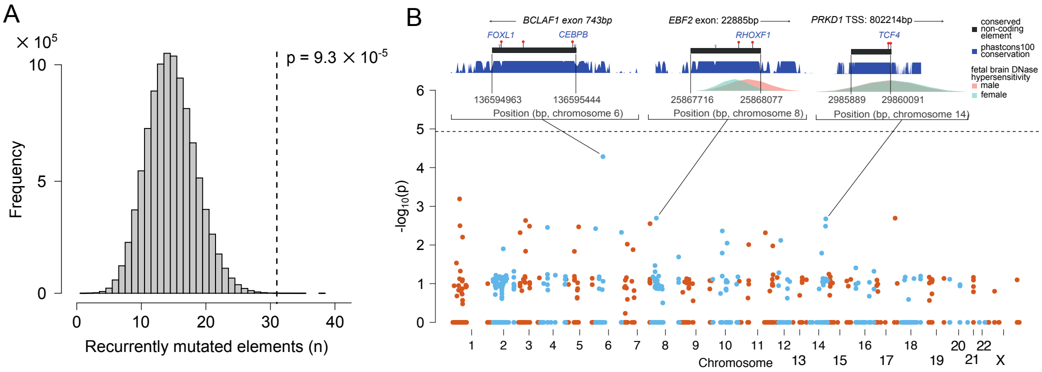


Figure 10 Recurrently mutated elements. (a) Approximately two-fold enrichment of recurrently mutated non-coding elements. Grey histogram shows distribution of expected number of recurrently mutated fetal brain active non-coding elements under the null model and vertical line indicates observed number. (b) Enrichment test of individual non-coding elements. No element was significant at a genome-wide threshold of $P < 1.9 \times 10^{-5}$ (Bonferroni correction for testing 2,613 fetal brain active elements). Inset plots for three elements show the nearest exon or transcription start site, location of DNMs (red markers) with any predicted transcription factor binding site disruptions (gain of binding in blue, loss of binding in red), location of rare variants in unaffected parents (grey markers), evolutionary conservation (blue, higher indicates more conserved), and fetal brain DNase I hypersensitivity (male in pink, female in blue).

I used chromHMM[43] to assign the recurrently mutated CNEs to a predicted chromatin state. I observed the greatest excess of DNMs in CNEs predicted to be enhancers (N=9) or strongly/weakly transcribed (N=8) (Figure 11). Five of the eight transcribed recurrently mutated elements fall in close proximity to exons, but are not in protein-coding transcripts and show evidence for involvement in alternative splicing (*BCLAF1*, *SRRT*, *SLC10A7*, and MKNK1) or as a 3' UTR (*CELF1*). The full set of recurrently mutated elements is described in Supplementary Table 1 (Appendix 1) and the location of DNMs relative to population variation and additional annotations is shown in Supplementary Figure 1 (Appendix 1).

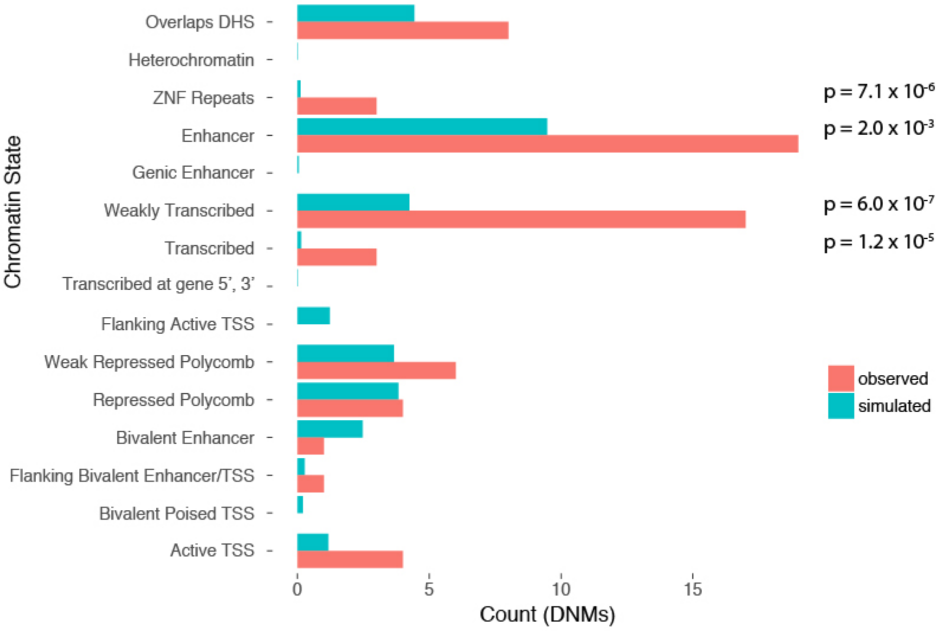


Figure 11 chromatin state of the recurrently mutated elements. chromHMM was used to annotate each of the recurrently mutated elements with a predicted chromatin state. The recurrently mutated elements showed an enrichment for enhancers and transcribed elements.

Increased power to detect locus-specific enrichments of DNMs could be gained from aggregating DNMs across elements regulating the same target gene(s). However, as described above, gene target prediction is lacking in coverage and accuracy. CNEs have been shown to cluster together within the genome[61] and are enriched around developmentally important genes[61]. An alternate approach, analogous to aggregating distinct exons into a single gene, is to cluster regulatory elements based on their location in genomic space. Therefore I applied hierarchical clustering on the 2,613 fetal brain active CNEs to identify 356 clusters (Methods). I found an excess of recurrently mutated clusters, defined as two or more elements with at least one DNM in each element (11 observed, 6 expected, p = 0.016), but did not find any element clusters with a significant excess of DNMs at a genome-wide significance threshold.

Results Section 2.2: Extrapolating results from targeted regulatory elements to genome-wide estimates

Results Section 2.2.1. Estimate of proportion of probands carrying a pathogenic mutation in a regulatory element genome-wide.

While only 4.2Mb of non-coding sequence was analysed in this cohort, there is more than 80Mb of non-coding sequence overlapping DNase I hypersensitive sites in the fetal brain. Unlike the set of elements I targeted, which is biased toward highly evolutionarily conserved elements, the vast majority of open chromatin regions in the fetal brain are poorly evolutionarily conserved (Figure 12A). Thus, the results from the DDD project is biased towards highly evolutionarily conserved elements, but also includes elements with lower levels of evolutionary conservation. In the targeted elements in this study, all of the observed DNM enrichment is concentrated in the highly conserved elements and I see no evidence for enrichment in poorly conserved regions, even though a large fraction of the poorly conserved regions have strong evidence for enhancer activity in mouse transgenesis assays.

To extrapolate the excess of DNMs I observed in the targeted non-coding elements to a genome-wide estimate, I modelled the enrichment of DNMs in the targeted non-coding elements as a function of evolutionary conservation and extrapolated to non-coding elements genome-wide (Methods). Factoring in the distribution of evolutionary conservation of fetal brain DHS peaks genome-wide, I predicted a genome-wide excess of 88 DNMs (95% CI: 48-140), corresponding to 1.0% - 2.8% of exome-negative cases carrying pathogenic mutations in regulatory elements (Figure 12A) in contrast to 13.4% and 28.4% carrying protein-truncating variants and missense variants estimated by McRae et. al, 2017 (Figure 12B). This estimate does not include small insertions or deletions due to a low number of observations in this study and the lack of a well-calibrated mutation model for small insertions and deletions. Furthermore, the lack of signal in poorly conserved elements should not be considered definitively negative – the targeting strategy in this analysis focused on highly evolutionarily conserved elements and comparable numbers of affected trios with deep whole genome sequencing will provide a more unbiased assessment of the contribution of mutations in non-coding elements genome-wide.

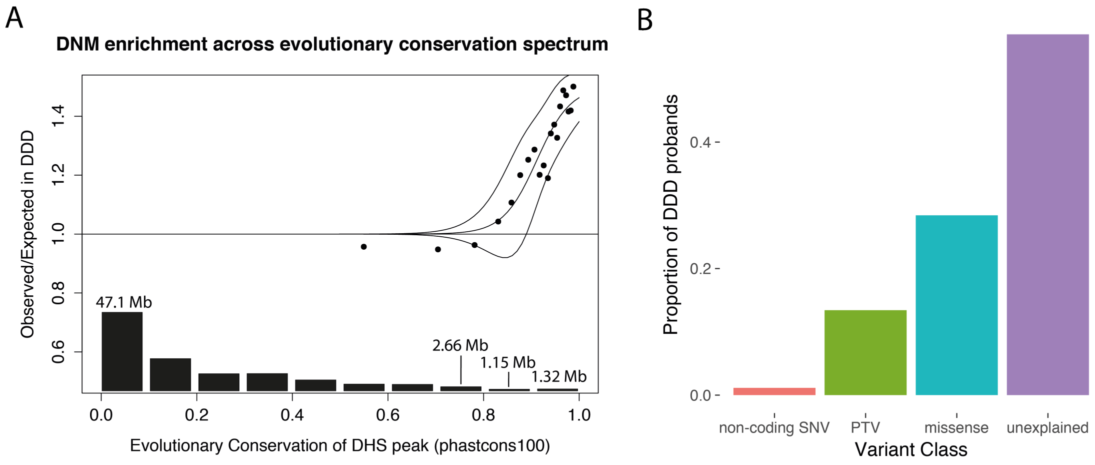


Figure 12 Genome-wide estimate of DNM burden. (a) Logistic regression used to model the genome-wide contribution of dominant-acting DNMs in fetal brain DNase hypersensitive sites in non-coding elements as a function of level of evolutionary conservation using a sliding window approach including 1,000 elements in each bin (see Methods). Dashed lines indicate the upper and lower 95% CI. The barplot shows the fetal brain active DHS peaks genome-wide (in megabase of total sequence) at a given level of evolutionary conservation. (b) The proportion of probands carrying a pathogenic *de novo* SNV in a fetal brain active regulatory element (1-2.8%) is far lower than the proportion carrying a pathogenic *de novo* protein-truncating variant (PTV) (~13.4%) or *de novo* missense variant (~28.4%).

Results Section 2.2.2. Power calculations and estimation of fraction of bases contributing to severe DD when mutated

The significant excess of recurrently mutated elements, but absence of individual non-coding elements with a genome-wide significant enrichment of DNMs is indicative of low power. This was initially surprising, as in a study comprising 4,293 trios, roughly half the sample size of this analysis, McRae et. al discovered 94 robustly associated protein-coding genes[4]. However, a large fraction of the difference in power can be attributed to the smaller size of CNEs compared to protein-coding genes and the lack of nucleotide-level variant effect prediction in the non-coding genome. The CNEs sequenced in this analysis are a median of 600bp in length, while protein-coding genes are median of 1800bp. Furthermore, variant effect predictor (VEP) reliably stratifies likely damaging from benign variation in protein-coding genes, while no such tool exists in non-coding elements. Down-sampling gene length to 600bp and masking protein consequence annotation results in an 80% drop in empirical power for the 94 genes passing the genome-wide significance threshold in McRae et. al, 2017 (Figure 13A). However, even after down-sampling genes in size and masking the predicted consequence of individual mutations, I still discover more than 20 genes at a genome-wide significance threshold (Figure 13A).

Beyond element length and consequence annotation, the proportion of sites that, when mutated, result in a severe developmental disorder with a dominant mechanism may differ between non-coding elements and protein-coding genes. At least 8% of mutations in protein-coding genes are predicted to cause loss of function due to protein-truncation[44,62], and many variants may result in missense changes causing full or partial loss of function. The fact that I do not discover any genome-wide significant CNEs at this sample size suggests that the proportion of DNMs in CNEs that are pathogenic and highly penetrant must be substantially lower than 8%.

I modelled the likelihood of observing 286 DNMs, 25 recurrently mutated CNEs, and zero CNEs at genome-wide significance across different values for the number of fetal brain active CNEs (out of 2,613) and the proportion of mutations in those elements that are pathogenic with a dominant mechanism for neurodevelopmental disorders (see Methods). The maximum likelihood model is one in which 3.5% of mutations within approximately 100 elements are pathogenic with a dominant mechanism. However, there is considerable uncertainty around this point estimate (Figure 13B), with the credible interval including scenarios whereby tens of elements have ~5-7% of mutations being pathogenic or thousands of elements have <1% of mutations being pathogenic. Our results support a model in which most sites in highly conserved non-coding elements are reasonably tolerant to heterozygosity. This result implies that the extreme sequence conservation in these elements may be maintained by selection against lower-effect size heterozygotes, compound heterozygotes, or oligogenic selection reflecting full or partial redundancy between multiple regulatory elements.

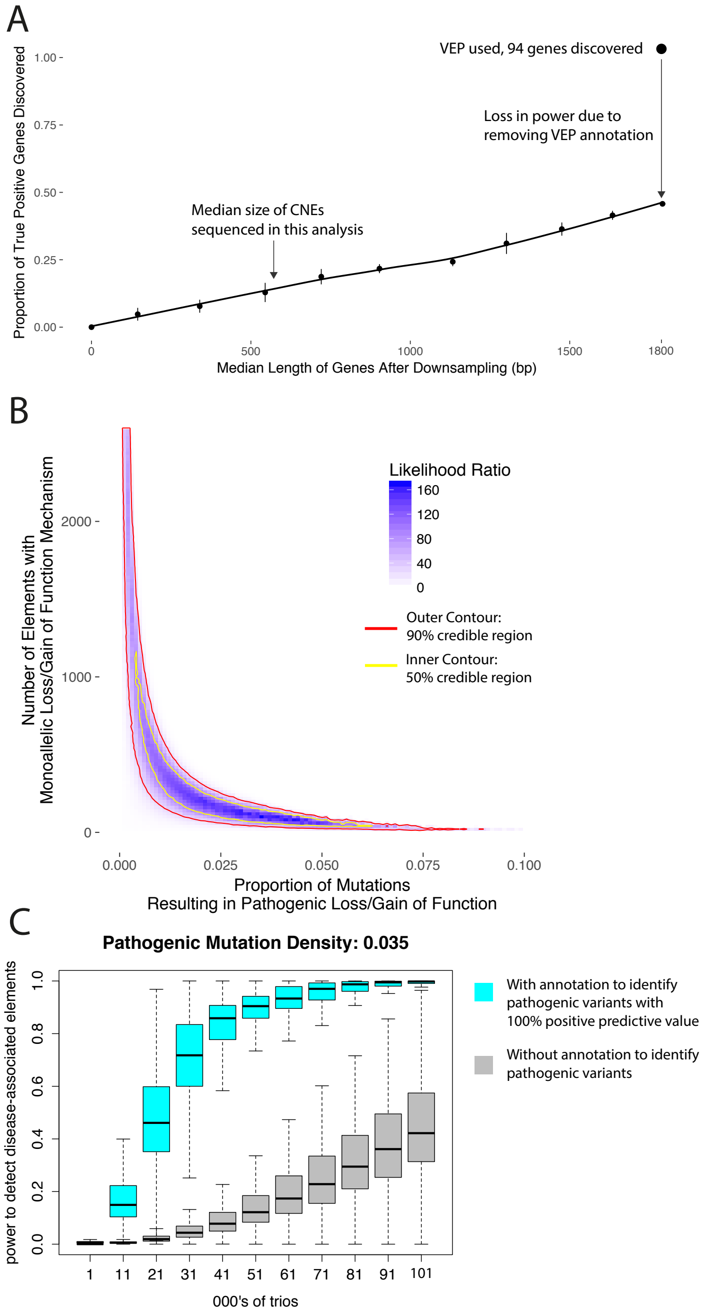


Figure 13 Power calculations estimates of genetic architecture in non-coding elements. (a) Estimating the reduction in power due to size differences between non-coding elements and genes (median 600bp vs. 1800bp) and ignoring VEP annotations used to stratify benign from likely damaging variants. Dots and bars represent the point estimate and 95% confidence interval, respectively. (b) Credible intervals for the proportion of fetal brain active conserved elements and proportion of sites within those elements with a dominant mechanism for developmental disorders. Based on our observation of zero non-coding elements at genome-wide significance in 6,239 exome-negative probands, very few sites within these elements (<5%) are likely to

contribute to developmental disorders through a highly penetrant dominant mechanism. (c) Power calculations for disease-associated non-coding element discovery. Without annotation or tools to discriminate pathogenic from benign variants in non-coding elements (grey), more than 100,000 trios are required to achieve 40% power. With annotation or tools to fully discriminate likely pathogenic from benign variants (blue), 40% power is achieved with only 21,000 trios.

Discussion

In summary, I have demonstrated that *de novo* mutations in regulatory elements contribute to severe neurodevelopmental disorders. This significant excess of DNMs is only observed in highly evolutionary conserved elements that are active in the fetal brain. I observed a 1.3-fold excess of DNMs within DHS peaks in these regulatory elements, suggesting that a minority of such DNMs are pathogenic. Moreover, our modelling suggests that there are few, if any, regulatory elements in which >4% of mutations cause neurodevelopmental disorders with a dominant mechanism. Our data are consistent with only 0.15% of mutations within fetal brain active CNEs being highly penetrant for neurodevelopmental disorders (Appendix 1, Figure S7A), likely considerably lower than the proportion of dominant pathogenic mutations in protein-coding regions. As a consequence, this class of pathogenic non-coding DNMs is only likely to account for a small proportion (<5%) of 'exome-negative' individuals and robustly identifying disease-associated regulatory elements will present a greater challenge than protein-coding genes.

This estimate does not include the potential contributions from indels. While the indel mutation rate is approximately 10-fold lower than the SNV mutation rate, indels have been suggested by evolutionary studies to be more deleterious than single nucleotide changes in non-coding regions[63]. I also did not quantify the contribution of large copy number variations in regulatory elements. Deletion of enhancer elements as well as enhancer adoption due to genomic rearrangements have been previously associated with severe disorders. Whole genome sequencing of affected trios will allow us to better detect this class of variation in order to accurately quantify the contribution of structural variants in non-coding elements to severe DD.

Our study design focuses on highly conserved elements and fetal-brain active elements and is relatively uninformative with respect to pathogenic 'gain-of-function' DNMs within elements that show no wildtype activity in fetal brain, and are not highly evolutionarily conserved. While our findings have focused on the highly conserved elements, I do not consider our observations to be definitively negative about the role of less highly conserved fetal brain enhancers in neurodevelopmental disorders, or the role of heart enhancers in CHD (due to the low proportion of subjects with CHD). The field of regulatory element annotation has progressed tremendously over the past six years since this study design was initially conceived. Therefore, a comprehensive analysis of the contribution of variation within all classes of non-coding elements to neurodevelopmental disorders is likely to require whole genome sequencing (WGS) of many tens of thousands, if not hundreds of thousands of parent-proband trios (Figure 13C).

A few recently published studies using whole genome sequencing in trios with Autism Spectrum Disorder have produced mixed results. One challenge of interpreting WGS data is the vast universe of hypotheses that could be tested, and thus how to account appropriately for multiple hypothesis testing. *Turner et al.* recently reported a nominally significant enrichment (p = 0.03) of *de novo* SNVs and private copy number variants in fetal brain DHS or at sites with PhyloP conservation score of >4, within 50kb of known autism-associated genes in WGS from 53 individuals with autism[64]. Caution should be exercised in interpreting findings based on: small sample sizes relative to that required for well-powered analyses (as discussed above) and analyses requiring multiple, arbitrary, levels of variant stratification (e.g. gene set, genomic proximity threshold, and conservation score). WGS-based analyses need to account for all explicit and implicit hypothesis testing. *Sanders et. al,* analysed a larger data set using an unbiased method to test for enrichment of mutations in more than 50,000 different annotations and found little evidence for enrichment of DNMs in non-coding annotations.

The disease-associated elements identified in this analysis primarily act either as enhancers or to regulate alternative splicing, but establishing the precise mechanism for each element has proved challenging. Our analyses highlight an urgent need for improved tools to stratify benign and damaging variants within non-coding elements and annotate gene targets for regulatory elements. Improvements in annotation of functional non-coding elements and nucleotides within these elements will dramatically increase power to detect highly-penetrant disease-associated non-coding variation, for example, increasing power more than ten-fold from 8% to 83% in 40,000 trios (Figure 13C).

Functional characterisation of increasing numbers of robustly-associated, highly-penetrant, regulatory variants in cellular and animal models will be critical in moving from a descriptive to a more predictive understanding of non-coding variation in the human genome. There are a number of experimental tools available to better understand the impact of mutations in regulatory elements. Massively parallel reporter assays (MPRAs) can be used to assess the function of tens of thousands of putative regulatory elements in parallel. Whereas mouse transgenesis assays or mouse knock-in models can assay particular variants in great detail, MPRAs provide far greater throughput allowing every base within an element to be assessed for allele specific expression. Results from a set of MPRA experiments in elements harbouring DNMs in the DDD patients are detailed in Chapter 4. While MPRA allows for very high throughput characterisation of regulatory element function, mouse or zebrafish assays can reveal spatiotemporal mis-regulation that may not be apparent in reporter assays. Results from a collaborative study contrasting expression patterns between wildtype and mutated alleles in mouse embryos is also discussed in detail in Chapter 4.

1 Study, D. D. D. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228, doi:10.1038/nature14135 (2015).

2 Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*, doi:10.1038/gim.2017.246 (2018).

3 Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic acids research* **45**, D865-D876, doi:10.1093/nar/gkw1039 (2017).

4 Mcrae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438, doi:10.1038/nature21062 (2017).

5 Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics* **31**, 67-76, doi:10.1016/j.tig.2014.12.003 (2015).

6 Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110, doi:10.1093/hmg/ddv259 (2015).

7 Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics* **25**, 157-165, doi:10.1093/hmg/ddw205 (2016).

8 Jeong, Y. *et al.* Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nature Genetics* **40**, 1348-1353, doi:10.1038/ng.230 (2008).

9 Benko, S. *et al.* Disruption of a long distance regulatory region upstream of SOX9 in isolated disorders of sex development. *J Med Genet* **48**, 825-830, doi:10.1136/jmedgenet-2011-100255 (2011).

10 Bhatia, S. *et al.* Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am J Hum Genet* **93**, 1126-1134, doi:10.1016/j.ajhg.2013.10.028 (2013).

11 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Genetics* **46**, 61-64, doi:10.1038/ng.2826 (2014).

12 Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725-1735, doi:10.1093/hmg/ddg180 (2003).

13 Hill, R. E., Lettice, L. A. & Hill, R. E. Alterations to the remote control of Shh gene expression cause congenital abnormalities. (2013).

14 Sellick, G. S. *et al.* Mutations in PTF1A cause pancreatic and cerebellar agenesis. *Nature Genetics* **36**, 1301-1305, doi:10.1038/ng1475 (2004).

15 Noonan, J. P. & McCallion, A. S. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics* **11**, 1-23, doi:10.1146/annurev-genom-082509-141651 (2010).

16 Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nature Communications* **6**, 6904, doi:10.1038/ncomms7904 (2015).

17 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48**, 488-496, doi:10.1038/ng.3539 (2016).

18 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).

19 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, 88-92, doi:10.1093/nar/gkl822 (2007).

20 Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics* **42**, 806-810, doi:10.1038/ng.650 (2010).

21 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

22 Shihab, H. a. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)* **31**, 1536-1543, doi:10.1093/bioinformatics/btv009 (2015).

23 Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *The American Journal of Human Genetics* **99**, 595-606, doi:10.1016/j.ajhg.2016.07.005 (2016).

24 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**, doi:10.1038/nmeth.3547 (2015).

25 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26**, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).

26 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nature Protocols* **11**, 1, doi:10.1038/nprot.2015.123 (2015).

27 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

28 Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).

29 Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90, doi:10.1038/nature14962 (2015).

30 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

31 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

32 Aaron McKenna, M. H., 1 Eric Banks,1 Andrey Sivachenko,1 Kristian Cibulskis,1 Andrew Kernytsky,1 Kiran Garimella,1 David Altshuler,1,2 Stacey Gabriel,1 Mark Daly,1, 2 & and Mark A. DePristo1. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 254-260, doi:10.1101/gr.107524.110.20 (2009).

33 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

34 Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature Methods* **10**, 3-7, doi:10.1038/nmeth.2611 (2013).

35 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

36 May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nature Genetics* **44**, 89-93, doi:10.1038/ng.1006 (2012).

37 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

38 Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet* **385**, 1305-1314, doi:10.1016/S0140-6736(14)61705-0 (2015).

39 Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

40 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).

41 Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

42 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

43 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).

44 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944-950, doi:10.1038/ng.3050 (2014).

45 Shooshtari, P., Huang, H. & Cotsapas, C. Integrative genetic and epigenetic analysis uncovers regulatory mechanisms of autoimmune disease. *bioRxiv*, doi:10.1101/054361 (2016).

46 Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527, doi:10.1038/nature19847 (2016).

47 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

48 Mathelier, A. *et al.* JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **44**, D110-D115, doi:10.1093/nar/gkv1176 (2016).

49 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

50 McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC bioinformatics* **11**, 165, doi:10.1186/1471-2105-11-165 (2010).

51 Duncan, B. K. & Miller, J. H. Mutagenic Deamination of Cytosine Residues in DNA. *Nature* **287**, 560-561, doi:DOI 10.1038/287560a0 (1980).

52 McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

53 Kosmicki, J. A., Samocha, K. E., Howrigan, D. P. & Sanders, S. J. Refining the role of de novo protein truncating variants in neurodevelopmental disorders using population reference samples. *Nature Genetics*, 1-18 (2016).

54 Chen, C. T. L., Wang, J. C. & Cohen, B. A. The Strength of Selection on Ultraconserved Elements in the Human Genome. *Am J Hum Genet.* **80**, 692-704, doi:10.1086/513149 (2007).

55 Derti, A., Roth, F. P., Church, G. M. & Wu, C.-t. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nature Genetics* **38**, 1216-1220, doi:10.1038/ng1888 (2006).

56 Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *bioRxiv*, doi:https://doi.org/10.1101/108290 (2017).

57 Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015-1026, doi:10.1016/j.cell.2014.10.025 (2014).

58 Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103, doi:10.1038/nature09525 (2010).

59 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, doi:10.1038/nature12787 (2014).

60 Gao, T. *et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**, 3543-3551, doi:10.1093/bioinformatics/btw495 (2016).

61 Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99, doi:10.1186/1471-2164-5-99 (2004).

62 Kryukov, G. V., Pennacchio, L. a. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics* **80**, 727-739, doi:10.1086/513473 (2007).

63 Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).

64 Turner, Tychele N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *The American Journal of Human Genetics* **98**, 58-74, doi:10.1016/j.ajhg.2015.11.023 (2016).

Chapter 3: Mutation rate and selective constraint in the non-coding genome

Introduction

Heterogeneity in the human germline mutation rate

Identifying associations between *de novo* mutations and disorders such as Autism Spectrum Disorder (ASD) or Developmental Disorders (DD) relies on a well-calibrated model of the underlying mutation rate, particularly if a large number of unaffected siblings or healthy trios are unavailable. Many methods for detecting positive and negative selection also rely on well-calibrated models of the germline mutation rate[1,2]. Assuming a homogeneous mutation rate across the whole genome can confound tests of purifying selection, increasing the false positive rate in hypomutable regions, and the false negative rate in hypermutable regions.

Historically, the mutation rate in humans has been inferred by comparing divergence between closely related species, for example other great apes[3]. These evolutionary methods have a number of drawbacks, including the difficulty in deconvoluting changes in mutational processes, generation times, and selective pressure. Furthermore, evolutionary measures have difficulty in distinguishing changes due to mutations from changes due to recombination-associated events such as biased gene conversion[2,4].

The rapid decline in sequencing cost has resulted in the sequencing of thousands of human exomes and whole genomes in recent years and rare variation in these individuals has also been used to model the germline mutation rate. These studies have focused on rare and putatively neutral sites (e.g. synonymous sites with minor allele frequency < 0.1%) where the mutational origin is likely recent and selection has had little time to act[5,6]. Under these assumptions, rare synonymous variation will closely reflect the underlying mutational processes[6,7].

Furthermore, increasing numbers of whole exome and whole genome sequenced trios have allowed more direct estimates of the germline mutation rate. Trio sequencing data has the advantage of measuring *de novo* variation directly, but at current sample sizes, the number of mutations observed is still low, approximately 100,000 *de novo* mutations in the largest study as of writing.

Trio sequencing studies also allow for the parent-of-origin to be determined for a large fraction of mutations. Using these data, the contribution of maternal and paternal age to the SNV mutation rate has been quantified[8]. Subtly different mutational spectra in the maternal and paternal germline have also been characterised, including an enrichment for C>T mutations with increasing paternal age, and increasing C>G mutations with maternal age with particularly striking maternal-biased hotspots in a small number of sub-chromosomal segments[8,9]. These maternal-biased hotspots coincide with large hypermutable regions in the genome, most strikingly apparent on chromosomes 2,7,8,9, and 16. Thus, trio sequencing data can be used to estimate the total human mutation rate, but may not be sufficient to detect heterogeneity, except on a megabase scale.

Studies using *de novo* mutations, rare variants, as well as evolutionary comparisons to detect heterogeneity in the underlying mutation rate have identified a number of genomic features associated with mutation rate and spectra. Local sequence context has been shown to have a significant influence on the per-base mutation rate. SNV mutation models incorporating triplet context (the base before and after the mutating base) capture variation in the mutation rate and have been used extensively in rare disease studies and models of selective constraint[7]. Extending beyond the triplet context, pentamer (5-bases) or heptamer (7-bases) context centred on the nucleotide of interest further improve modelling of DNMs and rare variants[5].

Beyond sequence context, other features related to cellular processes and chromatin modifications have been linked to mutation rate heterogeneity. Later replicating regions of the genome have been shown to accumulate a greater number of mutations[10]. A number of mechanisms have been proposed for this phenomenon, including exhaustion of the pool of dNTPs, causing polymerase stalling and that later replicating regions lack sufficient time to undergo mismatch repair. Increased recombination rate has also been linked to increase SNP density and mutation rate[11,12]. The proposed mechanism for this relationship is the mutagenic effect of double strand breaks, which are required for recombination. Two chromatin marks characteristic of repression, H3K27me3 and H3K9me3, have also been suggested to influence the germline mutation rate[5], perhaps by reducing accessibility of the DNA repair machinery. Transcription factor binding has also been suggested to influence mutation rate[13].

Despite analyses implicating different genomic features and sequence features in germline mutation rate heterogeneity, there are no widely used models incorporating these features together. Furthermore, the heterogeneity in the germline mutation rate has not been fully explained by the features studied to date, motivating continued work in characterising factors contributing to variation in the germline mutation rate.

Measuring selective constraint in the human genome

Putative functional DNA can be identified by using alignments between multiple species to detect orthologous segments and testing the divergence of these segments relative to a neutral model of evolution. Based on this approach, the majority of putative functional sequence is thought to be non-coding[14]. A small set of these elements show extreme sequence conservation and were originally dubbed 'ultra-conserved'[15]. These evolutionary methods have also suggested that non-coding sequence may be more intolerant to small insertions and deletions than single-nucleotide changes[16].

More recently, different population genetics methods have been applied to large sets of exome and whole genomes sequences to detect selective constraint in humans. In sites or regions undergoing purifying selection, there will be a greater proportion of observed rare alleles compared to sites or regions evolving neutrally. These shifts in the variant frequency spectrum (VFS) have been quantified using different methods including the proportion of singletons (the number of polymorphic sites observed once in a population sample of unrelated individuals divided by the total number of observed polymorphic sites) as well as methods such as RVIS that quantify the difference between two allele frequency distributions[6,17,18]. As sample sizes have increased beyond tens of thousands of individuals, the assumptions of the infinite sites model have been broken for some classes of variation with high mutation rates (e.g. CpG sites) and measures of selection based on the variant frequency spectrum have required revision[6]. The mutability adjusted proportion of singletons (MAPS) is an extension of the proportion of singletons measure that corrects for biases due to hypermutability[6].

Yet another way to quantify purifying selection is to compare the number of observed rare variants to the number of expected variants under a neutral model. This approach has been applied to protein-coding genes to identify genes intolerant of protein truncating variation[6]. Using a germline mutation rate model, the expected number of protein-truncating variants can be determined for each gene in the absence of selection. Genes with few observed variants relative to the number expected are annotated as loss of function intolerant. The degree of intolerance can be quantified as a ratio of observed divided by expected with an associated Z-score, or using a mixture model to identify genes with a high 'probability of loss of function intolerance' (pLI)[6]. The pLI score has been utilised in a number of different contexts to prioritise disease causing variants. For example, *Kosmicki et. al* show that the enrichment for damaging DNMs in ASD and DD patients can be explained almost entirely by enrichment in genes with pLI > 0.9, which represent only ~17% of all genes[19]. This approach has also been applied to identify protein-coding genes or gene sub-regions that are intolerant to missense variation[20].

Using data from approximately 10,000 deep whole genome sequences, *Di Iulio et. al* developed the 'context dependent tolerance score' (CDTS), to measure selective constraint genome-wide. Like the pLI and MAPS scores, which use the triplet sequence context to correct for heterogeneity in the germline mutation rate, the CDTS model uses the heptamer sequence context. However, the CDTS method suffers from a few potential drawbacks, most notably failure to model germline mutation rate heterogeneity and background selection[21] (the loss of heterogeneity or depression of allele frequencies at a site due to selection on nearby sites due to linkage). The pLI and MAPS scores also do not account for mutation rate heterogeneity beyond sequence context, but failing to account for mutation rate heterogeneity in coding regions is likely less impactful due to greater similarity in features such as replication timing and recombination rate compared to non-coding regions[22].

Both the evolutionary methods and the population genetics methods described above rely on a well-calibrated model of neutral evolution or the germline mutation rate. The risk of a mis-specified mutation rate causing false positive or false negative rates can be mitigated in protein-coding genes by testing for enrichment or depletion of variation in synonymous variants alongside missense or protein-truncating variants[6]. In the event that a gene has a higher mutation rate than predicted under the mutation rate model, this will be apparent as an increased number of synonymous sites. While these mutation rate differences are not modelled explicitly, genes with great deviations in synonymous rates can be flagged for removal or further analysis. In non-coding elements, it is more challenging to identify likely benign variation that can be assumed to be shaped almost entirely by mutation and not selection. As a result, negative selection will be confounded by hypomutability. In the case of the CDTS score discussed above, non-coding regions in the highest percentile CDTS score are highly enriched for promoter marks, which *Di Iulio et. al* interpret as evidence of strong purifying selection on promoters. However, promoters are enriched for CpG sites which are highly mutable when methylated, but CpG sites are often hypomethylated in promoters. As a result, a model based only on sequence and not incorporating methylation status will greatly overestimate the mutation rate, leading to a false positive prediction of constraint.

Detecting selectively constrained non-coding elements has the potential to greatly improve the signal-to-noise ratio in disease association studies in non-coding elements. However, identification of constrained elements relies critically on a robust germline mutation rate model, and a sufficient number of deep whole genome sequences to ensure power to detect modest selective constraint. The analyses I describe in this chapter make use of over 28,000 deep whole genome sequences and more than 1,500 whole genome sequenced trios to construct an improved model of germline mutation rate and apply this model to detect selective constraint genome-wide.

Methods

De novo mutations from 1,548 healthy trios

De novo mutations (DNMs) from 1,548 healthy trios described in Jonsson et. al 2017 were downloaded from the online supplementary information: https://www.nature.com/articles/nature24018#supplementary-information. These DNMs were lifted over from GRCh38 to GRCh37 using the UCSC LiftOver tool (https://genome-store.ucsc.edu/) and filtered to include only single nucleotide changes, leaving a total of 100,714 *de novo* SNVs. The DNMs were intersected with the regions used to build the model in *Samocha et. al, 2014*. These observed number of DNMs per triplet were compared to the fraction of triplets in the reference genome to determine the per triplet mutation rate.

Quality control of allele frequency data from 15,000 whole genome sequenced individuals from the genome aggregation database (gnomAD) and 13,000 individuals from the BRIDGE consortium

The gnomAD consortium (http://gnomad.broadinstitute.org/about) runs a unified bioinformatics pipeline on a large number of exome and whole genome sequencing samples and provides access to allele frequency data and other meta-data including depth of coverage across samples. Depth of coverage was included as a covariate for in the mutation rate model and all analyses related to selective constraint. All sites with low-quality variants (defined by the gnomAD random forest model) were flagged as low-quality, and the proportion of variants called as low quality in a given element was also included as a covariate.

The BRIDGE project also runs a unified bioinformatics pipeline for whole genome sequence data in a collection of rare disease cohorts. Coverage for every base as well as variant quality metrics provided were available, and as with the gnomAD sites, depth of coverage and the proportion of low quality variants was included as a covariate in the mutation rate model and all analyses related to selective constraint.

All analyses in the selective constraint section were run on the full set of 28,000 deep whole genome sequences from unrelated individuals as well as a subset of 15,000 whole genome sequenced from non-Finnish Europeans defined by principal components analysis. The number of rare variants in 50kb bins genome-wide was highly correlated between the full set of individuals compared to using only the non-Finnish Europeans ($r^2 = 0.93$, $p < 2.2e-16$). As a result, I performed all analyses with the full set of 28,000 individuals as the greater number of individuals improves power to detect depletion or enrichment for rare variation.

Tri-nucleotide mutation rate table based on 100,714 observed mutations

A custom R script was used to determine the trinucleotide sequence context for 100,714 *de novo* SNVs identified in 1,548 healthy trios. These *de novo* SNVs were used to construct a table describing the empirical mutation rate for each of 96 possible trinucleotide changes (e.g CCG -> CTG is one such change where CCG is the reference sequence and a C to T DNM is observed).

Genomic features included in the germline mutation rate random forest regression model

To model the variation in mutation rate due to genomic features, annotations in different functional categories were assembled:

- Recombination rate

  The 1000 genomes Phase 3 recombination map[23]

  Recombination rate in the male germline[24].

  Recombination rate in the female germline[24].

- Replication timing

  Replication timing in lymphoblastoid cell lines[25].

  Replication timing in embryonic stem cells[26].

- Chromatin features

  ATAC-seq data in human spermatogonial stem cells[27]

  ATAC-seq data in human embryonic stem cells[27]

  Embryonic stem cell H3K9me3, H3K27me3, H3K4me3, H3K4me1, H3K36me3, and H3K9ac from the Roadmap Epigenome Project[28]

  Ovary H3K9me3, H3K27me3, H3K4me3, H3K4me1, H3K36me3, and H3K9ac from the Roadmap Epigenome Project[28]

  CTCF binding sites from the ENCODE project[29]

Modelling genome-wide regional variation in mutation rate using random forest regression

The expected variation in sites predicted to be evolving neutrally based on the PhyloP score (PhyloP < 1 and PhyloP > -1) was calculated in 50kb bins genome-wide using the trinucleotide mutation rate model from *Samocha et. al* with the correction for CpG methylation status described in Chapter 2. Dividing the observed variation in these 50kb bins by the expected variation yielded the observed/expected ratio based on sequence context alone.

In order to determine the contribution to variation in observed/expected ratio from genomic and technical features, a random forest regression model was used. Seventy-percent of the data was used for hyperparameter tuning and model selection using 10-fold cross validation. The remaining thirty percent of data was held out to evaluate the model. The model was also evaluated on a completely independent set of *de novo* mutations 1,548 trios described in *Jonsson et. al, 2017.*

Measuring selective constraint using rare single nucleotide variants

A linear model was trained to predict number of observed variants given the mutation rate of a genomic segment using only sites with PhyloP between -1 and 1 in the ENCODE Ancestral Repeat sequences with >25x coverage in BRIDGE and gnomAD and >80% high quality variant calls in BRIDGE and gnomAD. Given a new genomic element, set of elements, or set of sites, the mutation rate was determined from the mutation rate model discussed above and the number of expected variants was generated using this linear model. Dividing the observed number of variants by the number of expected variants yields the observed/expected ratio.

Measuring selective constraint using rare indels

The number of rare indel variants per megabase (indels per megabase, or IPM) was first calculated for the ENCODE Ancestral Repeat sequences with >25x coverage in BRIDGE and gnomAD and >80% high quality variant calls in BRIDGE and gnomAD. The number of *de novo* indel calls per megabase (dnIPM) in 1,548 DECODE trios was also determined.

Given a new set of genomic elements of interest, for example DHSs, the $\mathrm{IPM_{DHS}}$ can be determined as above. The $\mathrm{IPM_{DHS}}$ is then multiplied by $\mathrm{dnIPM_{AR}}$ /$\mathrm{dnIPM_{DHS}}$ and the observed/expected value is:

$$\frac{\mathrm{dnIPM_{AR}}}{\mathrm{dnIPM_{DHS}}} * \frac{\mathrm{IPM_{DHS}}}{\mathrm{IPM_{AR}}}$$

Including the *de novo* indel rates is critical, as elements with $\mathrm{dnIPM_{DHS}} < \mathrm{dnIPM_{AR}}$ will appear under selective constraint, when in fact the paucity of rare variants is due to differences in mutation rate. Due to the relatively small number of observed *de novo* indels, Indel constraint is only feasible for element sets with a sufficient number of observations. For this work, only element sets with at least 100 expected mutations under the neutral model were included, limiting the noise introduced by the mutation term to approximately 20%, assuming the rate of indels follows a Poisson distribution.

Evolutionary conservation using PhyloP and PhastCons

PhastCons scores describing degree of evolutionary conservation at the element level in vertebrates and primates were retrieved in R using the Bioconductor[30] package GenomicScores (http://bioconductor.org/packages/devel/bioc/vignettes/GenomicScores/inst/doc/GenomicScores.html).

PhyloP scores represent the –log10 p-value that a given nucleotide is evolving neutrally[31] (Pollard et. al, 2010). A tabix file of pre-computed PhyloP vertebrate 100-way scores and primate 46-way scores were used to annotate rare variants and *de novo* mutations.

ENCODE open chromatin clusters

All of the ENCODE V3 DNase I hypersensitive peaks were overlapped to generate singly-linked clusters. These clusters were downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz.

Simulations to determine the power to detect selective constraint in non-coding elements

Random segments of the required size (10bp, 20bp, 100bp, 200bp) were selected from the ENCODE V3 DNase I hypersensitivity sites and the mutation rate and expected amount of rare variation under a neutral model was calculated. To simulate selective constraint, the amount of observed variation was sampled from a Poisson distribution with lambda = 0.8*expected. This simulated constraint is approximately equal to the deficit of variation observed in the average protein-coding exon, thus is a conservative estimate if constrained non-coding elements are less constrained, on average, than protein-coding exons. The observed/expected ratio and Z score was calculated from the simulated observed count and the expected count. The proportion of elements with $Z < -2.58$ was recorded as the power to detect a true association. This Z-score results in a false discovery rate of approximately 10%, under the assumption of ~5% of the genome under selective constraint, as a Z-score of less than or equal to -2.58 would be achieved under the null model in ~1/200 tests (false positive rate of 0.005).

Conserved transcription factor binding sites

Conserved transcription factor binding sites based on the Transfac database were downloaded from the UCSC genome browser (http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgTables?db=hg19&hgta_table=tfbsConsSites). This set includes any binding sites that successfully align in human, mouse, and rat and are predicted to bind based on position weight matrices in the Transfac Matrix Database (v7.0).

Analysing selective constraint stratified by number of active tissue groups

Non-coding elements defined by DNase hypersensitive sites were annotated with a predicted chromatin state in each of the following ten tissue groups from the Roadmap Epigenome Project data:

- Embryonic stem cells (Roadmap IDs: "E002", "E008", "E001", "E015", "E014", "E016", "E003", "E024")
- Blood (Roadmap IDs: "E062", "E034", "E045", "E033", "E044", "E043", "E039", "E041", "E042", "E040", "E037", "E048", "E038", "E047")
- Hematopoietic Stem Cell and B Cell (Roadmap IDs: "E029", "E031", "E035", "E051", "E050", "E036", "E032", "E046", "E030")
- Mesenchymal Cells (Roadmap IDs: "E026", "E049", "E025", "E023")
- Epithelial Cells (Roadmap IDs: "E055","E056","E059","E061","E057","E058","E028","E027")
- Adult Brain (Roadmap IDs: "E071", "E074", "E068", "E069", "E072", "E067", "E073", "E070")
- Adult Muscle (Roadmap IDs: "E100", "E108","E107")
- Adult Heart (Roadmap IDs: "E104", "E095", "E105", "E065")
- Fetal Tissues (Roadmap IDs: E082", "E081", "E080", "E083", "E084", "E085", "E086", "E088", "E089", "E090", "E092", "E093", "E017")
- Smooth Muscle (Roadmap IDs: "E078", "E076", "E103", "E111")

An element was considered active in a tissue group if it was annotated as an enhancer in any of the constituent tissues (Roadmap IDs above). The obs/exp ratio was calculated for elements grouped by the number of tissue groups they were active in (from 0 to 10).

Results Section 3.1: Modelling the human germline mutation rate

Results Section 3.1.1. Improved modelling of the human germline mutation rate.

The reduction in cost of whole genome sequencing has led to adoption of this technology in a number of different research efforts worldwide. Many of these efforts are focused around generating diagnoses in rare or common disease, but whole genome sequence data is also of general utility to address questions in population genomics. With this goal in mind, researchers at the genome aggregation database (gnomAD) have collected data from dozens of such studies and processed the raw data through a unified quality control and variant calling pipeline. To date, data from more than 15,000 deep whole genome sequences have been released by the gnomAD consortium. The latest gnomAD release (r2.0.2) was downloaded for use in these analyses. This release includes allele counts at each polymorphic site and metadata including depth of coverage. The BRIDGE consortium has sequenced more than 13,000 deep whole genomes from individuals with different rare diseases and in some cases, unaffected family members. Allele counts from the BRIDGE and gnomAD data were combined and annotated with depth of coverage and variant call quality (see Methods). I also downloaded high-quality *de novo* mutations from 1,548 whole genome sequenced healthy trios[8] in order to validate analyses on the germline mutation rate derived from rare variant data (see Methods).

The widely used germline mutation rate model from *Samocha, 2014* described previously relies on a 96-row table that describes the mutation rate from one triplet sequence to another[7]. This table was derived using polymorphism data in orthologous chimp and human sequence. As it has been shown that there have been changes in triplet-specific mutation rates between humans and other great apes, as well as within human populations[32,33], I reasoned that building this table directly from high-quality DNMs would be more accurate. To build this table, I intersected the DNMs with the same set of regions used to build the original mutation rate table (see Methods).

The triplet mutation rates derived from *de novo* mutations differed slightly from those derived from evolutionary estimates, notably in the rate of mutations at CpG sites. The DNM-based triplet model modestly but significantly outperforms the triplet model derived from polymorphisms in orthologous sequence in predicting the observed number of variants with MAF < 0.1% in 2kb bins genome-wide ($r^2 = 0.61$ compared to $r^2 = 0.64$). As described in Chapter 2, the addition of CpG methylation status also improved the model fit.

A number of additional genomic features have been previously associated with differences in rare variant density and divergence over evolutionary time, specifically replication timing, recombination rates, and H3K9me3, with conflicting evidence for H3K27me3[4,5,10]. As a previous study by *Carlson et. al,* used rare variant data from just 3,000 individuals to test potential mutation-associated germline features, I reasoned that with over 28,000 deep whole genomes as well as mutations from whole genome sequenced trios, an analysis to test association across a wide range of potential associated features would have much greater power. Furthermore, *Carlson et. al* was descriptive, but did not provide a model to predict the mutation rate given a sequence and set of genomic features[5]. Such a model integrating known sequence-associated variation in mutation rate with genomic features would be a useful tool in disease-association studies, population genomic models of selection, and for identifying 'mutational outliers' that are not well-explained by known mutation-associated futures which may provide insight into novel mutational mechanisms.

I reasoned that rare variants in deep whole genome sequences could be used to assess variation in the germline mutation rate independent of selection by focusing only on sites that are likely to be evolving neutrally. While our understanding of the 'regulatory code' does not allow for identification of benign sites to the same degree of certainty as in protein-coding genes, I reasoned that nucleotide level evolutionary conservation could be a useful proxy. I annotated every base genome-wide with the PhyloP score (see Methods). Sites with PhyloP greater than one (referred to as 'conserved sites' going forward) had on average 25% fewer variants than sites with PhyloP less than 1 and greater than -1 (referred to as 'neutrally evolving sites' going forward). There was no significant difference in the mutation rate between conserved sites and neutral sites, indicating that the deficit of variation observed in the conserved sites is likely the result of purifying selection. Taken together, these results imply that, in the aggregate, sites predicted to be evolving neutrally by PhyloP are under minimal purifying selection and can be used to quantify the background mutation rate.

I compiled data from twenty-six different sources representing established or potential mutation-associated genomic features including replication timing, recombination rate, and chromatin marks from the male and female germline (see Methods). I also included technical covariates including depth of coverage, presence of low complexity repeats[34], and variant quality to quantify the amount of variation that is attributable to technical sources rather than variation in mutation rate. Finally, I included the proportion of polymorphisms observed split across the six different 1mer possibilities (C>T, C>A, C>G, T>A, T>C, T>G). Including this feature was motivated by the observation that a subset of highly hypermutable regions of the genomes are enriched for C>G mutations which are high bias toward maternal-origin[8,9]. Unlike other genomic features, associations with this feature may not necessarily have a clear mechanistic interpretation, but are nonetheless useful to model mutation rate heterogeneity, or generate hypotheses for the source of mutational heterogeneity based on known mutational signatures, for example those found in the Cosmic database (https://cancer.sanger.ac.uk/cosmic).

I then split the genome into non-overlapping segments and annotated these segments with the number of observed rare putatively neutral variants, and the number expected given the mutation rate based on sequence context alone. The ratio of observed to expected variation was fit using a random forest regression (RFR) on a randomly selected subset of the two-kilobase segments (the 'training' set). The performance of the model was then tested on a held-out subset (the 'test' set) (see Methods). I hypothesized that different genomic features may influence the mutation rate at different length scales. Thus, a separate model was trained for input segments of size 2kb, 10kb, 50kb, 200kb, and 1Mb.

Incorporating genomic features substantially improved prediction of rare variation over sequence context and technical covariates alone across all of the different length scales, with the largest length scales showing the greatest improvement (Figure 1A), as variation in sequence context plays a larger role for smaller regions, while genomic features dominate over larger length scales where sequence context becomes more homogeneous (Figure 1B).

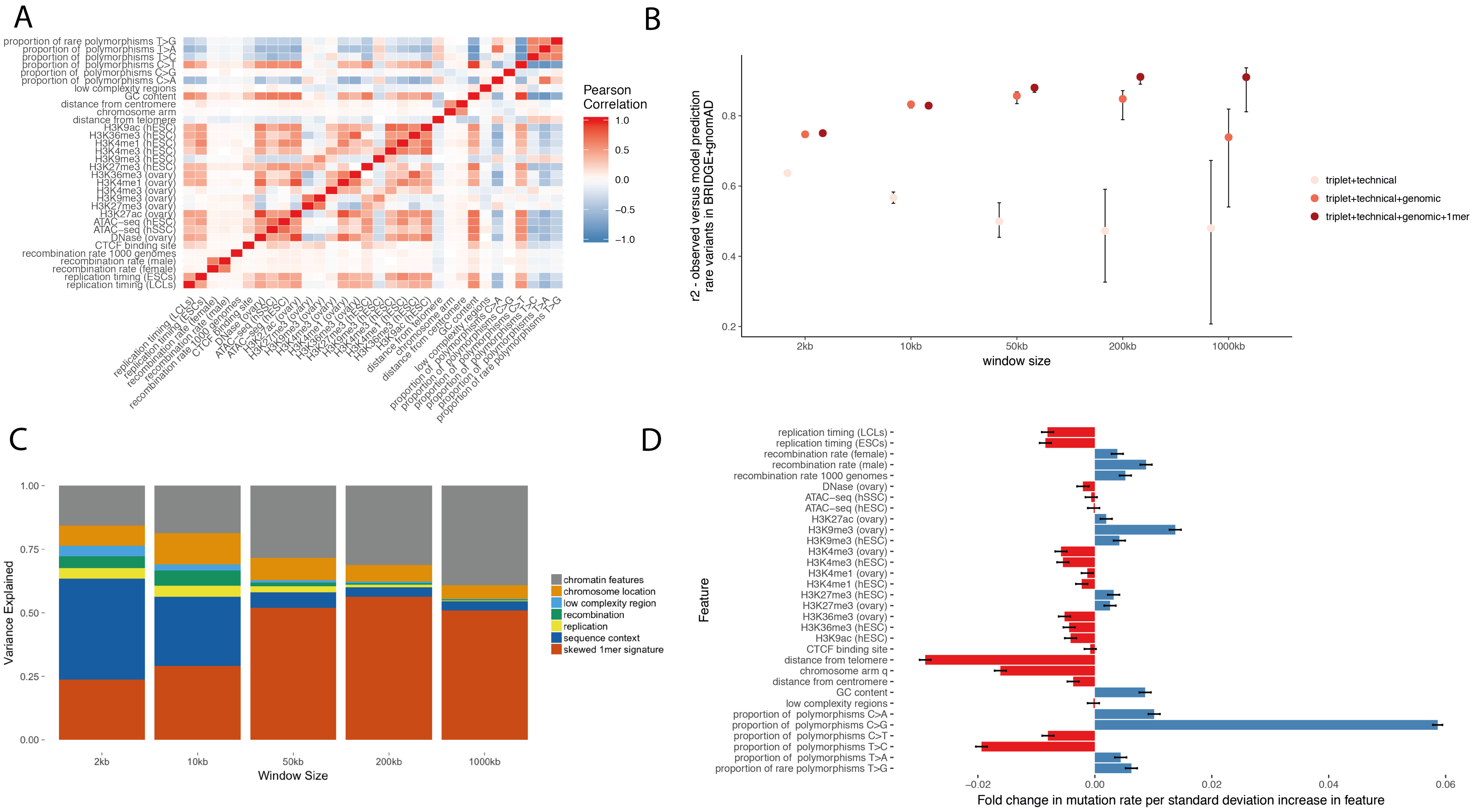


Figure 1 Modelling heterogeneity in the germline mutation rate. (a) Correlation between different model features. (b) adding genomic features and polymorphism nucleotide signatures (1mer) improves prediction of level of rare variation in different sized genomic bins. (c) Random forest regression feature importance shows decreasing importance of sequence context at larger length scales, and increased importance of chromatin features. (d) adding genomic features to mutation rate model improves prediction of *de novo* mutation rates.

Figure 2A shows the observed/expected values in 2kb bins across chromosome 8 for the model incorporating sequence context and technical covariates compared to the model incorporating sequence context, technical covariates, and genomic features. As the model was trained using rare variant data, which may not accurate reflect the underlying mutation rate, I sought to validate this model using 100,714 *de novo* SNVs from 1,548 healthy trios[6]. *De novo* SNVs will more closely reflect the mutation rate independent of selection, but are fewer in number than rare variants, motivating the choice to train using rare variant data, and validate using the DNMs. To validate the model performance, I ordered all of the 2kb bins genome-wide based on their predicted deviation from the mutation rate based on sequence context alone. Splitting the ordered bins into deciles, the expected number of *de novo* mutations was estimated for each bin using the sequence context-based model as well as the model including genomic features. The model including genomic features was a better predictor of the DNM rate than the model based on sequence context alone (Figure 2B).

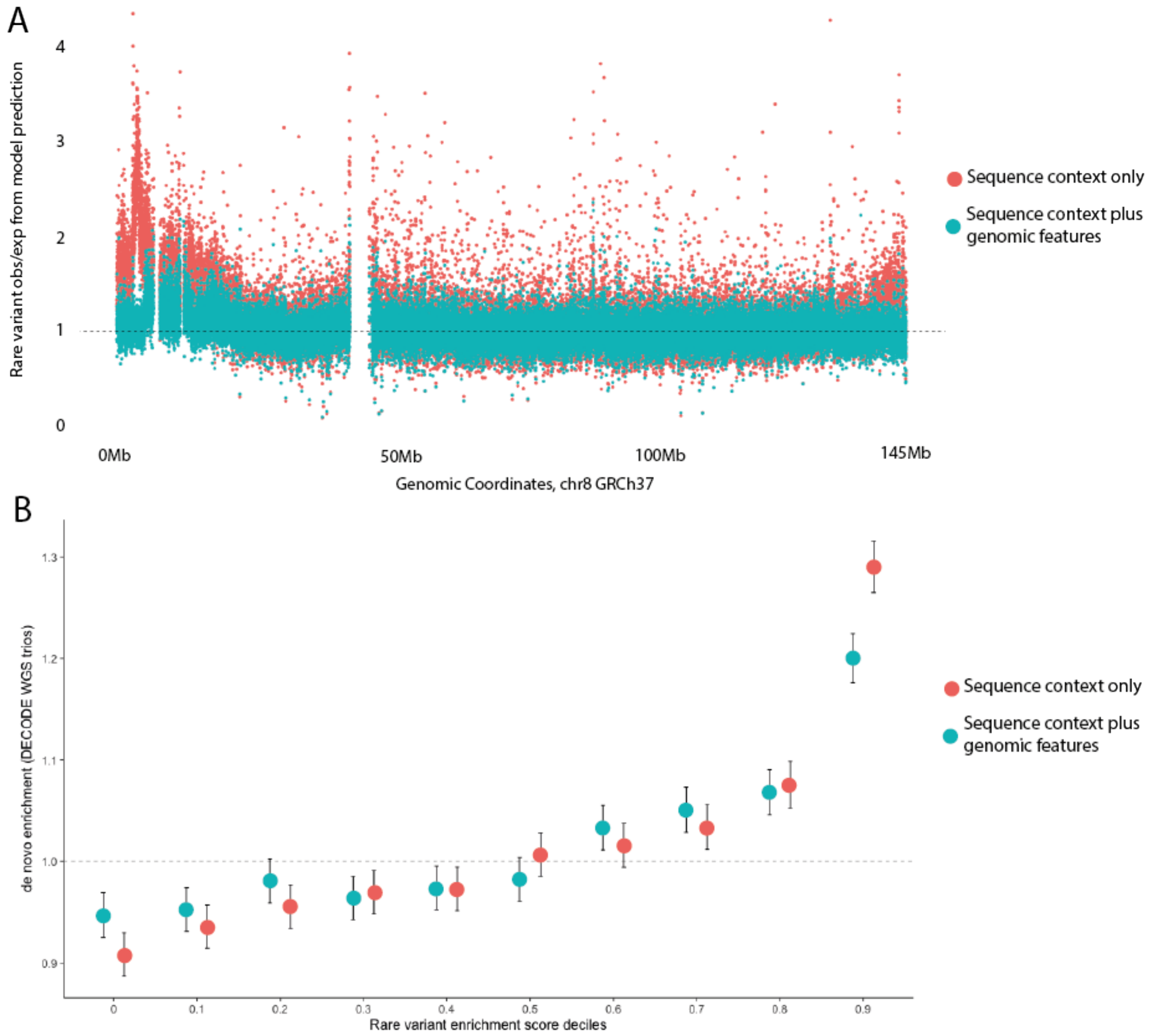


Figure 2 Visualising and validating the improved germline mutation rate model. (a) Observed over expected ratios across chromosome 2. Calculating the expectation from the sequence context only (red dots) and using the genomic features and sequence context (blue dots). (b) All 2kb elements genome-wide ordered by their observed enrichment of rare variation predicted rare variant enrichment. The enrichment of *de novo* mutations in each bin was calculated using the model based on sequence context only (red dots) and the model using genomic features and sequence context (blue dots). The mutation rate model incorporating genomic features was a closer fit to the observed number of *de novo* mutations.

Results Section 3.1.2. Genomic features associated with hypermutability

This random forest regression approach detailed in Section 3.1.1. identified significant associations between the rate of rare likely neutral variation and a number of different genomic features (Figure 1C). This analysis recapitulated several known mutagenic features. Increase in recombination rate by one standard deviation was associated with an increase in mutation rate of approximately 0.5% - 1%. Replication timing had a modest but significant effect (increase in mutation rate of 0.2% per standard deviation increase in replication timing). In somatic tissues, replication timing has been observed as one of the primary determinants of mutation rate, whereas these results suggest that replication timing has a significant, but overall modest effect on variation in mutation rate in the germline.

I observed a strong positive correlation between the proportion of C>G polymorphisms and the mutation rate (mutation rate increase of 4.6% per standard deviation increase in C>G proportion), consistent with striking sub-chromosomal mutation hot-spots identified previously on chr2, chr7, chr8, chr9, and chr16[8,9]. Re-analysing the DNMs from Jonsson et. al shows that regions with strong enrichment for maternal biased C>G also exhibit a maternal bias for other mutation signatures, and a paternal bias for C>T mutations. The DNM counts in two megabase bins on chromosome 8, which has a mutational hotspot on the first 40 megabases of the chromosome, are shown in Figure 3A, B (enrichment for rare variation is also shown in Figure 2A).

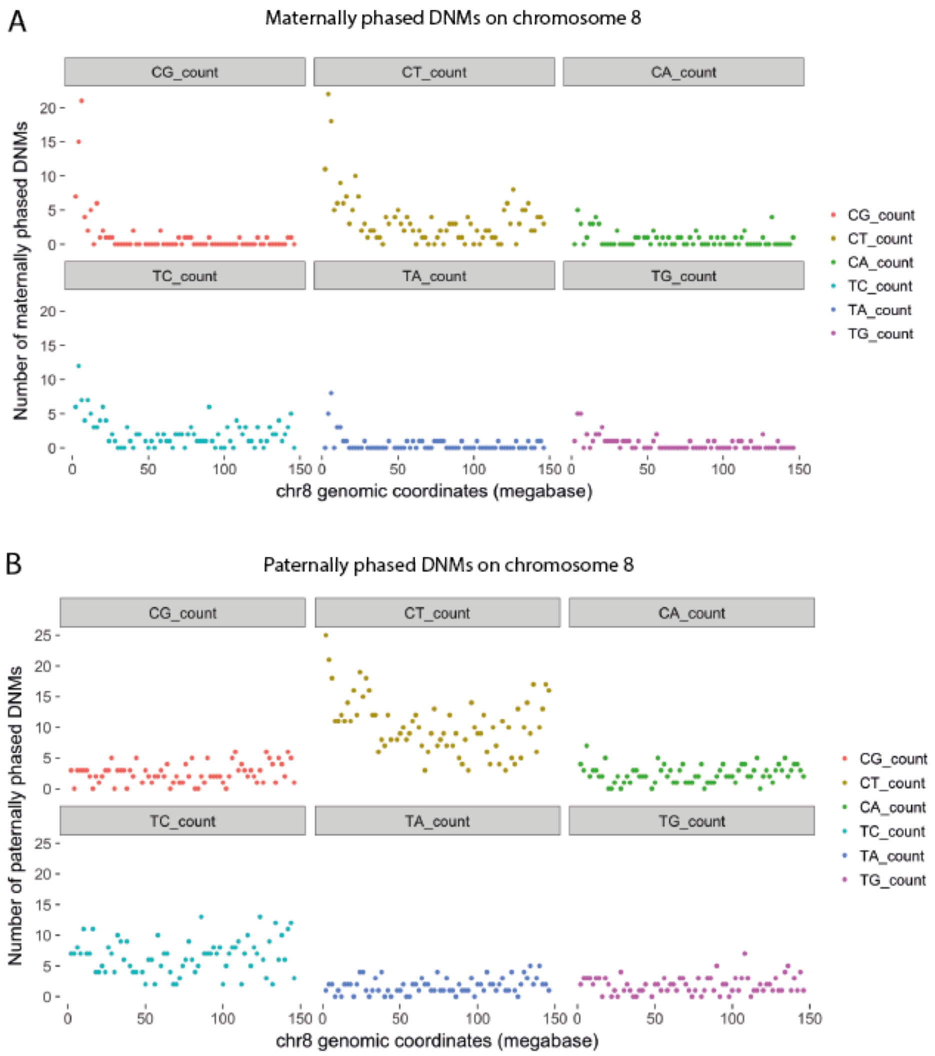


Figure 3 Hypermutable elements with C>G polymorphism signature. (a) *De novo* single nucleotide variants of maternal origin and (b) paternal origin in 2Mb bins across chromosome 8 stratified by the reference and alternate base to illustrate the difference in mutational signature between maternal and paternal derived mutations.

Even after removing chromosomes 2,7,8,9, and 16 which show large regions of extreme enrichment for C>G polymorphisms, we detect an association between this feature and increased mutation rate on other chromosomes, indicating that the phenomenon underlying these extreme events may be pervasive across the genome. Previous work describing this phenomenon has posited a role for recombination in generating these hotspots. I found a significant enrichment for maternal as well as paternal recombination hotspots in hypermutable elements with C>G polymorphism rates in the top decile, supporting previous work suggesting a role for recombination in these mutational hotspots.

I next sought to take a more unbiased approach to identify patterns underlying these hypermutable elements. I supplied maternal and paternal phased mutations to Raheleh Rahbari, a collaborator in the Voet group at the Wellcome Trust Sanger Institute, who performed non-negative matrix factorisation to detect mutation signatures separately in the maternally and paternally phased DNMs. DNMs of maternal and paternal origin both showed a strong contribution from Signature 1 (spontaneous deamination of 5-methylcytosine) and Signature 5 (unknown etiology) which were shown to predominate in the Germline by *Rahbari et. al, 2016*. The maternally phased DNMs also showed a strong contribution from Signature 3 (which has been linked with double strand break repair), supporting the hypothesis that double strand breaks in the maternal germline, perhaps due to recombination, are contributing to the enrichment for DNMs, particular C>G changes, in these regions. In contrast, there was no evidence for Signature 3 in the paternally phased DNMs or in DNMs in regions of the genome not enriched for C>G polymorphisms.

An inverse relationship between RNA expression level and mutation rate has been reported in somatic tissues based on analysis of whole-genome sequenced tumours and matched normal tissues by The Cancer Genome Atlas (TCGA)[35,36]. However, analyses in the germline based on evolutionary divergence and more recently, analysis of whole exome sequenced trios has suggested the opposite effect, whereby higher expression is associated with a higher mutation rate[37]. I tested the relationship between density of rare variation and RNA expression levels in the testis based on the GTEx data set. Consistent with previous reports in the germline, I detected a strong positive relationship between increased expression and density of rare variation as well as *de novo* mutations. Notably, this relationship was only evident in coding sequence, not in the non-coding intronic sequence. Protein-coding exons on the most highly expressed decile of transcripts had approximately 25% more DNMs than the lowest expressed decile of protein-coding exons (Figure 4).

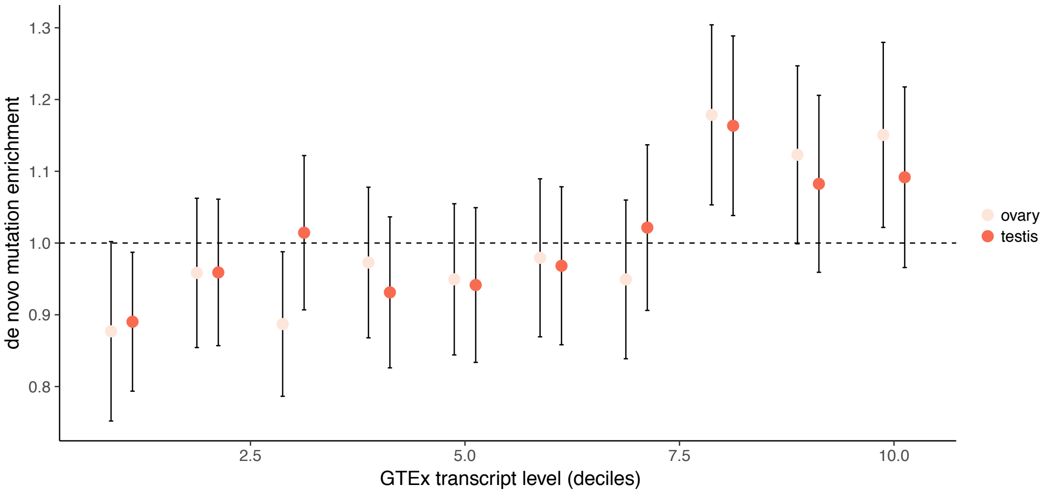


Figure 4 Association between mutation rate and expression in the ovaries and testis. Protein-coding exons were split into deciles based on transcript levels in ovary and testis from the GTeX project. Observed *de novo* mutations were compared to the expectation, generated using the random forest regression mutation rate model and show a positive correlation between transcript level and observed mutation rate that is not captured in the existing model.

It is possible that this apparent hypermutability in coding sequence could actually be explained by a fraction of mutations in highly expressed genes in the testis conferring a selective advantage by increasing spermatogonial stem cell proliferation or sperm motility. I reasoned that if this was the case, there would be an enrichment for non-synonymous changes relative to synonymous changes in the highly expressed genes compared to the lowly expressed genes. I compared the fraction of DNMs predicted to be protein-truncating, missense, and synonymous in each transcription decile and did not find any evidence for higher-than-expected rates of missense or protein-truncating mutations compared to synonymous changes that might suggest this observation was driven by a selective advantage in sperm or sperm progenitors.

After incorporating known and suspected mutation-associated genomic features into the mutation rate model, there were still genomic regions with a significant excess of rare variation. These genomic regions were also enriched for *de novo* mutations, indicating that the enrichment for rare variation is likely due to an elevated mutation rate, not selection or sources of technical variation. Thus, while the mutation rate model developed here is an improvement on the existing model based on sequence context alone, there is still opportunity for improvement, particularly in modelling finer scale variation in mutation rate. CTCF binding has been shown to be associated with increased mutation rate[13] and other transcription factors may have also increase the mutation rate when bound. Better understanding of mutation rate heterogeneity in the germline will be possible as tens of thousands of whole genome trio sequencing data sets become available – these next steps are discussed in greater detail in the Discussion.

Results Section 3.2: Measuring selective constraint in regulatory elements genome-wide

Results Section 3.2.1. Patterns of purifying selection in non-coding elements genome-wide

The degree of selective constraint on a non-coding sequence can be expressed as a ratio of observed variation divided by expected, with elements under stronger constraint having lower observed/expected ratios. The germline mutation rate model discussed in section 3.1.1 was used to calculate the mutation rate for approximately 46,000 Human/Chimp Ancestral Repeats. A linear regression was fit to predict the observed number of rare single nucleotide variants in likely neutrally evolving bases given the mutation rate in these neutral bases. This approach is analogous to the approach taken in *Lek et. al, 2016* in the coding regions using synonymous variation, which is expected to be under little purifying selection, to calibrate the model.

This model was applied to 1.7 million DNase I hypersensitive sites, 22,000 3’ UTRs, 28,000 5’ UTRS, 15,000 promoters, 88,000 exons from long non-coding RNAs (lncRNAs) and 182,000 exons from protein-coding genes. I first assessed evidence of selective constraint for rare variants in these elements regardless of any annotation of evolutionary constraint or variant effect on the nucleotide level (Figure 5A). Protein-coding exons showed the greatest degree of selective constraint, with a median depletion of 28% (observed/expected ratio = 0.72). The DNase hypersensitive sites showed evidence for a modest, but statistically significant depletion of approximately 5.5% (observed/expected ratio = 0.945). The 5’ UTRs, 3’ UTRs, and promoter sequence were not significantly different from the ancestral repeats. This was an unexpected result, but may be due to a small fraction of the nucleotides within these elements being selectively constrained, or potentially an underestimate of the mutation rate leading to a false negative call. Of note, a previous publication from *Di Iulio et. al* reported a striking level of selective constraint in promoters: 23-fold enrichment in the top 1% of elements genome-wide compared to the genome-wide average, whereas protein-coding exons were only 12-fold enriched. As mentioned previously, *Di Iulio et. al* do not model variation in the germline mutation rate[21]. This analysis does account for CpG methylation status, amongst other genomic features – as promoter elements have hypomethylated CpGs, failure to account for this feature in a sequence-context based model would drastically overestimate the mutation rate and lead to incorrect predictions of extreme selective constraint.

Long non-coding RNAs (lncRNAs) are typically identified in RNA sequencing data as transcripts with little protein-coding potential, and are often transcribed at much lower levels than protein-coding genes. There are relatively few lncRNAs with well-understood functions. Some examples include XIST, which is involved in the X-inactivation process[38,39], and MALAT1, which has been implicated as a lung cancer driver[40]. Assessing selective constraint in 15,904 lncRNA genes defined by the GENCODE consortium (v24), I find evidence for a significant depletion of rare variation in lncRNAs (Figure 5A).

Small insertions and deletions (indels), while rarer than SNVs, may have a greater functional impact than SNVs and therefore shed greater light on functional and non-functional regions of the genome[16]. However, while considerable progress has been made in modelling the single nucleotide mutation rate in the germline, modelling the rate of indels has posed a greater in challenge. This challenge is in part due to the greater complexity of these mutations, which vary may vary in size and, in the case of insertions, the sequence inserted. Furthermore, the mutation rate for indels is also ~10-fold lower than single-nucleotide variation, resulting in a smaller number of *de novo* mutations to train and validate new models. Thus, while there are several models of the indel mutation rate in development, some of which make use of convolutional neural networks using DNA sequence as an input, no well-validated models exist to date. While the lack of a validated indel mutation rate model makes predicting indel constraint for any individual element challenging, estimating constraint in the aggregate for different element classes (for example, protein-coding exons, non-coding RNA exons, and DNase hypersensitive sites) can still be achieved by using *de novo* indels from the 1,548 DECODE WGS trios to account for any underlying heterogeneity in the indel mutation rate (see Methods).

The protein-coding exons showed a very strong depletion for indels compared to the other element classes analysed, consistent with strong purifying selection against frameshift variation reported previously. Of the non-coding elements assessed, the lncRNAs, CNEs, 3' UTRs, and 5' UTRs showed the greatest degree of indel constraint, followed by promoters and ENCODE open chromatin peaks (Figure 5A).

Evolutionary studies have suggested that insertions and deletions may be more deleterious than single nucleotide variants in regulatory elements[16]. All of the element sets with sufficient *de novo* indels to accurately assess the indel mutation rate had a substantially greater degree of indel constraint than SNV constraint. This result has important implications for how constraint is assessed, as constraint methods based on indels may have greater power to detect constrained non-coding elements, despite being fewer in number. Furthermore, this result suggests that while *de novo* indels may be rarer than *de novo* SNVs, they may be more pathogenic and therefore have a non-trivial contribution to severe Mendelian disorders akin to frameshift mutations in coding regions that, despite their low mutation rate, contribute to an outsized fraction of diagnoses in severe developmental disorders.

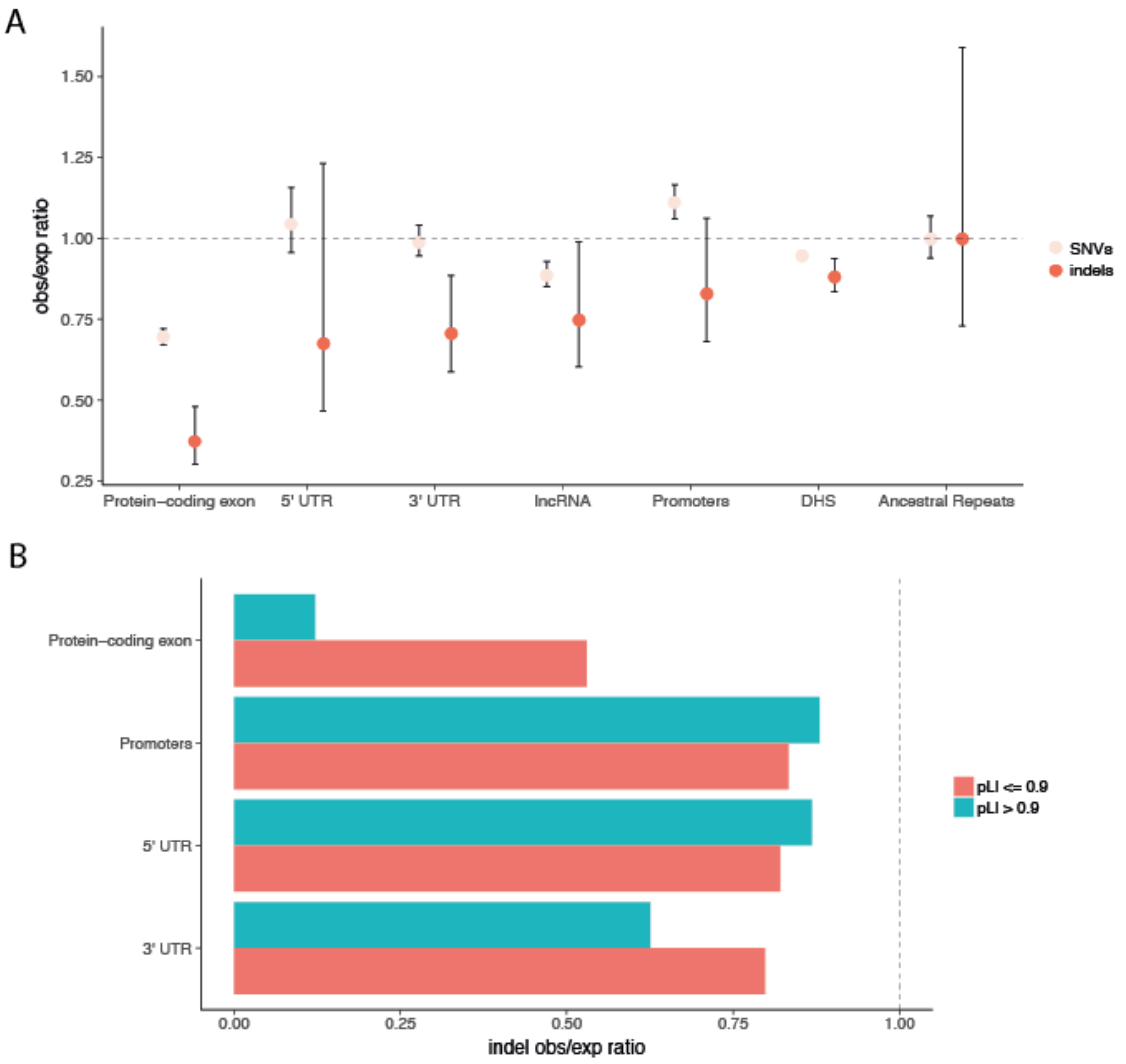


Figure 5 SNV and indel constraint in non-coding and coding elements. (a) observed over expected ratio for SNVs and indels in non-coding element sets and protein-coding exons. While the number of indel observations limits precise estimates, point estimates of indel constraint suggests stronger constraint against indels than SNVs. (b) Exons and 3' UTRs of genes with pLI > 0.9 (likely intolerant of heterozygous loss of function) show a greater degree of indel constraint than genes with pLI <= 0.9.

The exome aggregation consortium (ExAC) profiled loss of function mutations in more than 60,000 individuals and identified a depletion for loss of function mutations, including protein-truncating variants and splice variants. In this study, a subset of genes with a high probability of loss of function intolerance (pLI > 0.9) were identified. These genes have a high overlap with known haploinsufficient genes (in which loss of a single copy of the gene causes a severe phenotype). It has been proposed that high pLI genes require RNA expression above a particular threshold in order to maintain a critical function. Thus, genetic variation outside of the coding sequence causing reduction in expression may also be under selective constraint. To this end, I hypothesized that in addition to constraint on changes to the protein-coding sequence, there would be greater constraint on proximal regulatory elements such as promoters, 3' UTRs, and 5' UTRs in high PLI genes compared to other genes with less evidence for dosage sensitivity. I found a significant depletion in variation in protein-coding exons of high pLI genes compared to exons of genes with pLI < 0.9, consistent with previous reports[6]. I found evidence for a moderate depletion of rare SNVs in 3' UTRs of high pLI genes compared to genes with pLI < 0.9, but did not find any evidence for increased selective constraint on promoters or 5' UTRs of high pLI genes.

This result contradicts a relationship between selective constraint in non-coding elements and pLI score of the nearest protein-coding genes reported by *Di Iulio et. al.* However, as discussed previously, *Di Iulio et. al,* fail to account for variation in the germline mutation rate which may have biased some of their results. In particular, failure to account for variation in CpG methylation status led to a high rate of false positive predictions of selective constraint in promoter elements. Promoters of haploinsufficient genes, which have a high degree of overlap with high pLI genes, are enriched for CpG islands and have been shown to be less mutable[41], which could explain the strong enrichment for constrained promoters near high pLI genes.

I also tested the relationship between constraint and gene dosage sensitivity using indel constraint. The protein-coding exons of high pLI genes showed a very strong depletion for indels compared to exons of genes with pLI < 0.9. There is a strong correlation between 3' UTR length and gene dosage sensitivity[41] hypothesised to be due in part to a greater number of microRNA binding sites required to exert tight transcriptional control in these genes. Further work to understand the precise patterns of purifying selection, particularly for indels, is warranted in these elements. As with the SNVs, I did not observe any significant difference in indel constraint in the 5' UTRs or promoters of genes with pLI > 0.9 compared to genes with pLI < 0.9 (Figure 5B).

Beyond testing known categories of non-coding element for selective constraint, there is great interest in using selective constraint as a method to identify regulatory elements that are evolutionarily novel or have adopted a novel function in humans, such as human gained enhancers[42] (HGEs). Putative HGEs were identified by *Reilly et. al* by comparing H3K27ac and H3K4me2 levels in human to mouse and macaque. A total of 912 putative HGEs active in the fetal brain were identified. I observed moderate selective constraint on these putative HGEs, but this constraint was not significantly different from open chromatin peaks in general. Thus, while these elements do show human specific activity, there is little evidence for strong selective constraint that would imply a critical novel function. *Reilly et. al* report that the HGEs identified in their study do not show exceptional human specific changes, lending further support that the majority of HGEs may be the result of gradual changes in regulatory element function over evolutionary time, and have not undergone strong positive selection for a novel and critical function in brain development.

Human accelerated regions (HARS) are another class of regulatory element with putatively novel function in humans. HARs are non-coding elements that are highly evolutionarily conserved, but have accelerated divergence on the human lineage[43,44]. It is hypothesized that HARs have undergone recent positive selection for new and potentially human-specific function. Thus, these elements are of great interest in human evolution, developmental biology, and disease studies. HARs were originally described contemporaneously by *Prabahakar et. al* and *Pollard et. al* in 2006. The two groups relied on slightly different methodology, but both sought to identify genomic elements with extremely high sequence conservation across different vertebrate species, but a larger than expected number of point mutations that have reached fixation on the human lineage. HARs, like conserved non-coding elements in general, are enriched near genes involved in development[45]. A small number of HARs have been studied in detail, including HACNS1, an enhancer that has acquired 16 different human specific mutations[46]. HACNS1 is a weak limb enhancer in chimpanzee and macaque, but is strong enhancer in humans and is an important component of the development of the thumb. *Doan et. al* also report an enrichment for biallelic variants in HARs in Autism cases in consanguineous families[47].

I assessed the selective constraint on HARs using a set of 2,649 putative HARs compiled by *Capra et. al* in a meta-analysis of four different studies[45]. The selective constraint on HARs is greater than that of HGEs and not significantly different from other conserved non-coding elements, which are amongst the most highly constrained non-coding elements. These results indicate that the majority of HARs are likely to be functional and under selective constraint in humans. However, one particular HAR, 2x.HAR.238, has a striking three-fold enrichment for rare variation (Figure 6A,B). 2x.HAR.238 has been shown to act an enhancer for GLI2, an important gene in brain development. This element lies very close to a recombination hotspot and shows evidence for localised mutational clustering (Figure 6C), which I hypothesize may be the source of the hypermutability in this element. Thus, the enrichment for variation in 2x.HAR.238 in humans may be due to hypermutability rather than positive selection on variation enabling a novel function.

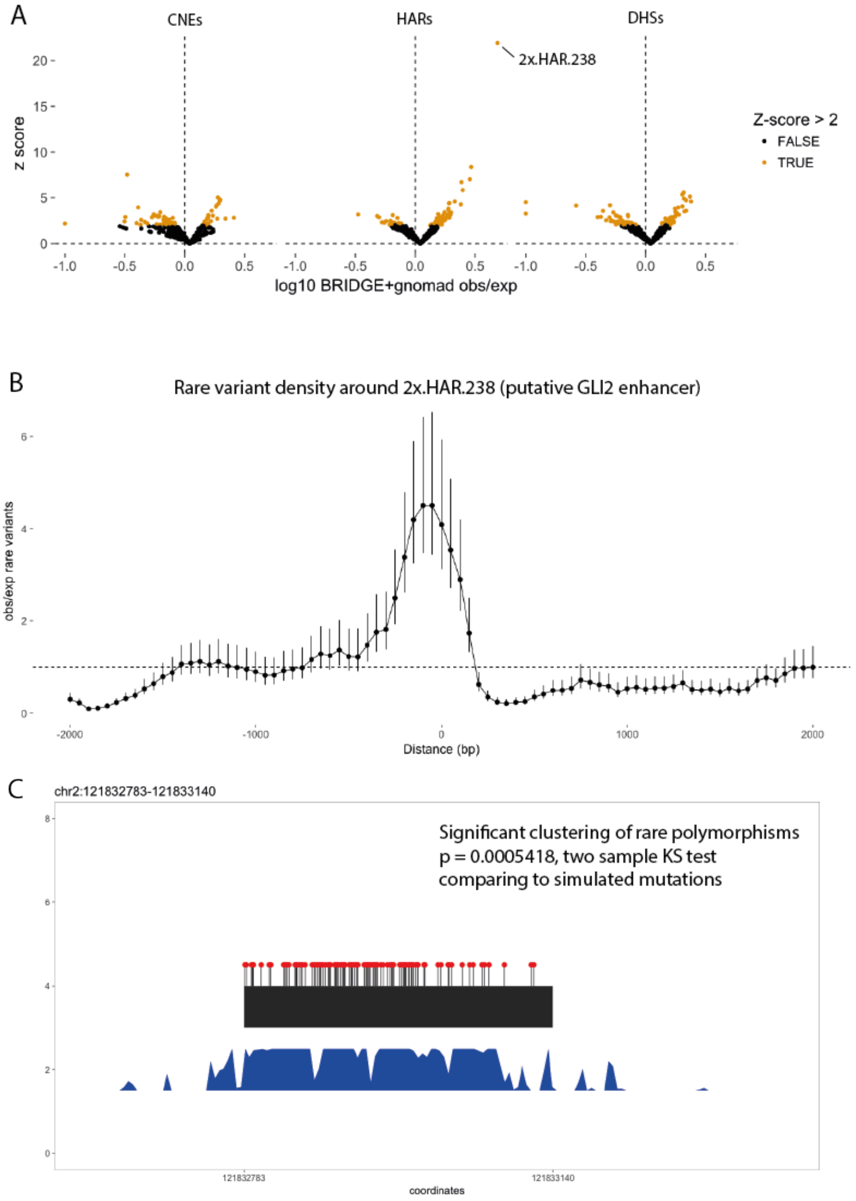


Figure 6 Density of rare variation in CNEs, HARS, and DHSs. (a) One human accelerated region (HAR), 2x.HAR.238, has a very strong enrichment for rare variation. No conserved non-coding elements (CNEs) or DNase I Hypersensitive Sites (DHSs) was observed with as extreme of an enrichment for variation. (b) Examination of the rare variation in the 2kb sequence upstream and downstream of 2x.HAR.238 reveals localised hypermutability in a window of approximately 750bp. (c) Rare variation in 2x.HAR.238 exhibits a high degree of clustering.

There are many more regulatory elements than protein-coding genes in the human genome[29]. Mammalian genes have a median of twelve associated regulatory elements, and the number of regulatory elements is positively correlated with higher expression, and greater stability of gene expression[48]. Likewise, large arrays of ultra-conserved regulatory elements have been shown to cluster near genes involved in early development, and are hypothesized to exert tight control over timing and expression levels[49]. However, the activity of more individual regulatory elements, is often restricted to a subset of tissues and developmental time points. It is not clear whether regulatory elements that drive expression in a large number of tissues, or those that operate within a relatively narrow, but potentially critical, functional window are under stronger selective constraint at the sequence level.

To test this, I used data from the Roadmap Epigenome Project to annotate activity for non-coding elements across 10 different tissue groups (see Methods) and found a significant correlation between the number of tissues in which an element is predicted to be active and the level of selective constraint (Figure 7). In contrast, we did not find strong evidence of a relationship between activity in a particular tissue or organ system (e.g. the fetal brain) and selective constraint that would imply strong constraint on enhancers active in a narrow but critical window of organismal development.

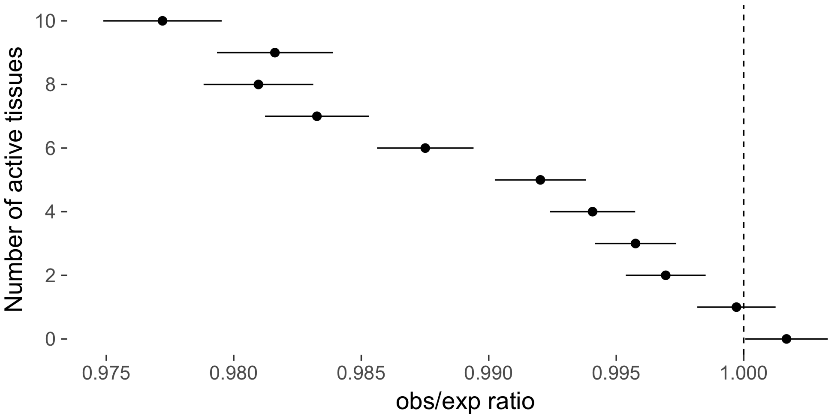


Figure 7 Activity in a greater number of tissues is correlated with increased selective constraint. Annotating DNase hypersensitive sites with predicted enhancer activity across ten different tissue groups shows a significant correlation between selective constraint and the number of tissues in which an enhancer is active

Results Section 3.2.2. Nucleotide-level conservation scores are more informative of selective constraint than locus-level scores

A large fraction of the human genome—by some estimates more than 80 percent—is biochemically active in at least one tissue. However, biochemical activity does not necessarily imply function or selective constraint. A number of different studies based on multispecies alignments suggest that just 3-15%of the genome is subject to purifying selection[16,50,51]. One potential explanation for these potentially contradictory observations is that a small fraction of biochemically active segments of the genome are functional and under selection at the sequence level, while the majority of biochemically active segments are explained by non-specific protein-binding that has no functional consequence. A second potential explanation is that a large fraction of biochemically active peaks are indeed functional, but this function is driven by a small fraction of nucleotides within these regions which are under selective constraint.

I sought to test these hypotheses on 1.7 million open chromatin regions covering nearly 400Mb of sequence outside of protein-coding exons. These regions were defined by the ENCODE consortium using DNase I hypersensitivity (DHS) assays in more than 200 different cell types and primary tissues. In combination with histone modifications and transcription factor binding sites, these open chromatin regions can be used to identify enhancers, promoters, and other functional non-coding elements.

I observed a strong correlation between evolutionary conservation and selective constraint in the ENCODE open chromatin peaks using rare variant data from the BRIDGE and gnomAD whole genome sequences (Figure 8A), which has been described previously[17,21,23]. However, even within poorly conserved peaks, there are individual nucleotides or sets of nucleotides that appear to be evolutionarily conserved within primates or vertebrates. To this end, I hypothesized that while these poorly evolutionarily conserved peaks do not appear to be under selective constraint in the aggregate, they may contain individual nucleotides that are selectively constrained.

To test this hypothesis, I selected the subset of nucleotides within the ENCODE open chromatin peaks with a PhyloP score greater than one, indicative of evolutionary conservation across vertebrates. The evolutionarily conserved nucleotides showed a depletion for rare variation regardless of the degree of conservation of the surrounding element (Figure 8A). Repeating this analysis using PhyloP scores based on primate multi-species alignments returned a similar result. Thus, while the open chromatin peaks cover nearly 400Mb of sequence, selective constraint is concentrated within a subset of approximately 120 Mb of sequence defined by nucleotide-level conservation. This sequence is spread throughout the genome, indicating that the elements underlying these peaks may be functional, albeit with a small fraction of nucleotides under selective constraint.

To further explore the relationship between element level constraint (measured by PhastCons) and nucleotide level constraint (measured by PhyloP), I annotated every nucleotide with the PhastCons100 score of the open chromatin peak in which the nucleotide is positioned and the PhyloP score of the nucleotide itself. Analysing the depletion of rare variation in a grid comprising deciles of PhyloP100 and PhastCons100 scores, it is clear that evolutionarily conserved nucleotides, even within poorly conserved peaks, are under selective constraint. Furthermore, as PhyloP score increases the degree of selective constraint also increases. Taken together, these results indicate that PhyloP score is a reasonable proxy for selective constraint in the non-coding genome, even when the sequence surrounding the conserved base is not conserved.

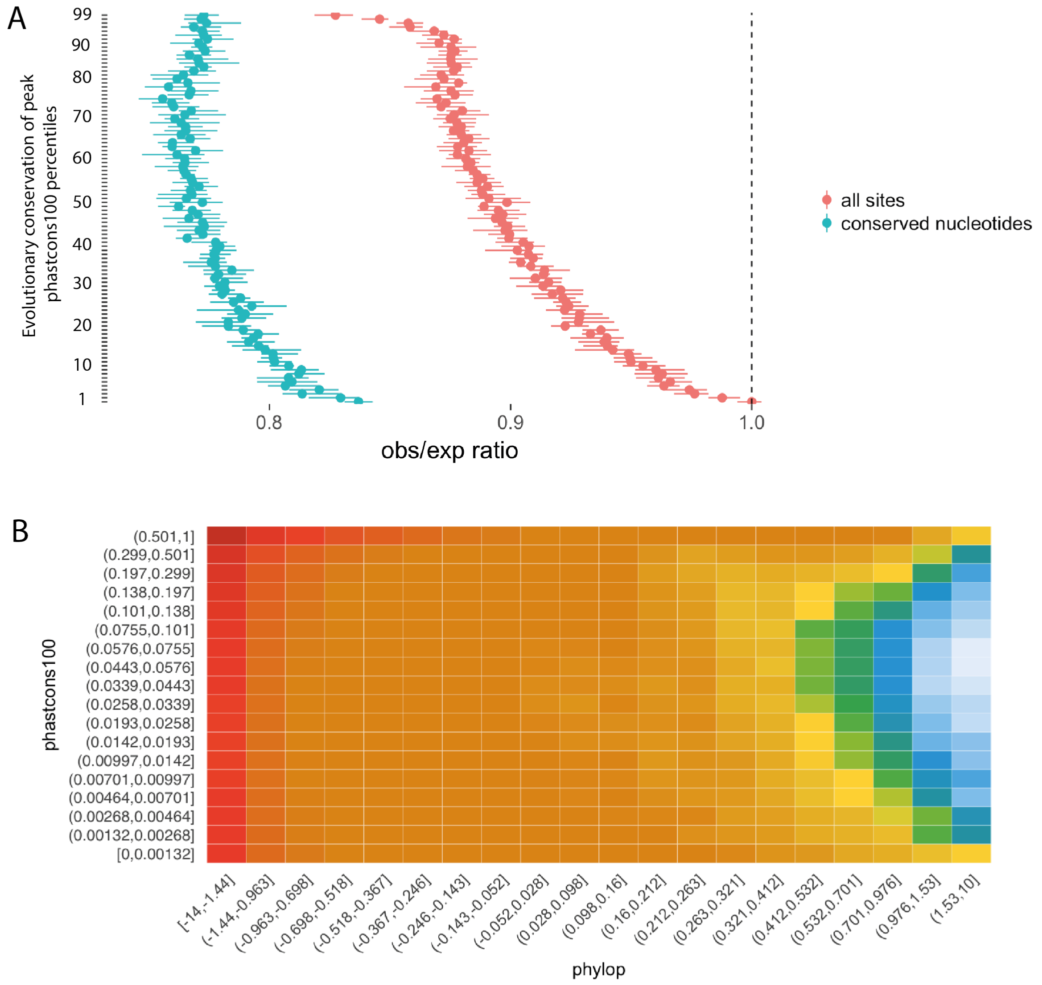


Figure 8 Nucleotide level conservation is more predictive of selective constraint than element-level conservation. (a) There is strong correlation between constraint, measured by the observed over expected ratio, and the element level correlation of DNase I Hypersensitive sites, measured by PhastCons100. Focusing on the bases with evidence for nucleotide level conservation (PhyloP > 1, denoted 'conserved nucleotides'), it is evident that even in elements with poor element-level conservation, there is constraint on conserved nucleotides. (b) All nucleotides in the 1.7 million DHSs genome-wide were annotated with nucleotide-level conservation (PhyloP) and element-level conservation (PhastCons100). Nucleotides were split into ten equal sized bins for each measure and plotted as a grid, showing a clear relationship between nucleotide-level conservation and constraint, regardless of the element-level conservation of the surrounding sequence.

I hypothesized that constraint on a small fraction of nucleotides within a DNase peak that is poorly evolutionarily conserved could be due to selective constraint on functional transcription factor binding sites within those peaks. To test this hypothesis directly, I calculated the selective constraint on a set of over two million TFBSs conserved between human, rat, and mouse (based on computational prediction of TFBS) and overlapping an open chromatin peak in at least one tissue. These TFBS are between 6bp and 30bp with a median size of 14bp. As with individual nucleotides identified by PhyloP score, the conserved TFBSs are constrained, even when they lie within poorly conserved open chromatin peaks (Figure 9A).

Testing each conserved transcription factor binding site independently, there is clear variation in the patterns of rare variation. A small fraction of the TFBS had greater levels of variation than expected. Testing these TFBS independently using *de novo* mutations from 1,548 healthy trios, I observed a 1.42-fold enrichment for mutations compared to the expectation under a null model (Figure 9B), indicating that this enrichment for rare variation is likely driven by hypermutability not captured by the updated model, perhaps due to TF binding in germline tissues as has been suggested previously for CTCF[13]. A more comprehensive analysis with greater numbers of whole genome sequenced trios will likely reveal a greater number of associations between mutation rate and transcription factor binding. On the other end of selective constraint spectrum, a number of transcription factor binding sites appear to be under strong selective constraint. The TFBS under the greatest degree of constraint is the TATA-box binding protein (TBP) motif, which is ubiquitous in mammalian promoters (Figure 9B). As this motif is likely active in a wide range of tissues, due to its general role in transcriptional activation, I hypothesize that the difference in selective constraint between TFs may be driven in part by their tissue specificity/ubiquity, in line with the observations on DHS described earlier (Figure 7), but further work is required to refine this hypothesis.

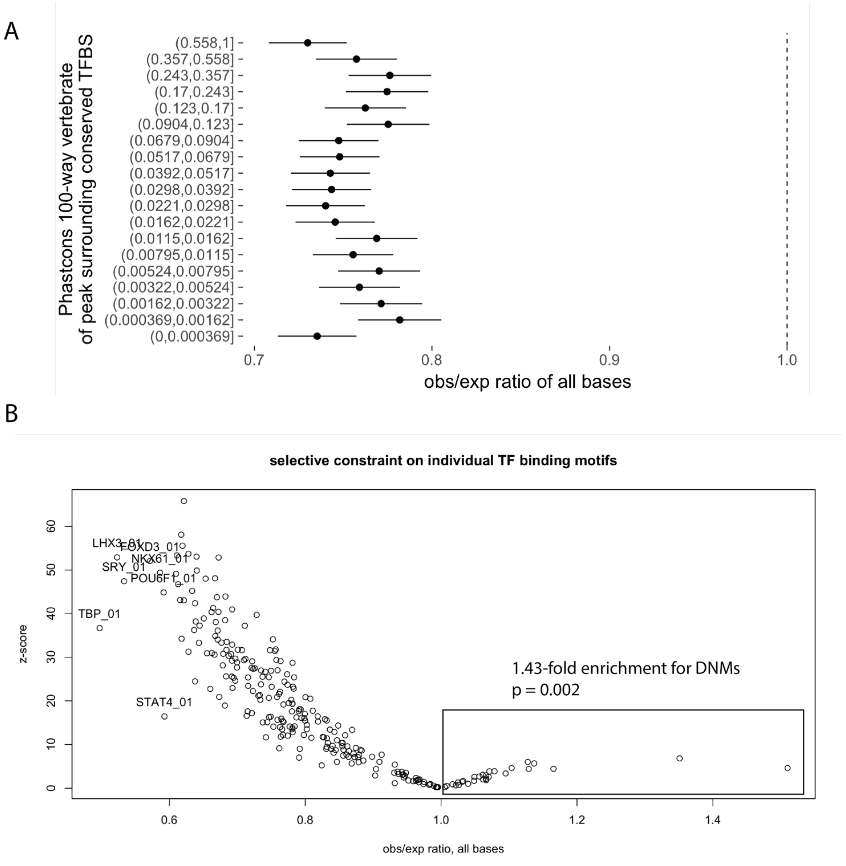


Figure 9 Constraint in conserved transcription factor binding sites. (a) Observed over expected ratio in conserved transcription factor binding sites (TFBS), stratified by the PhastCons100 score (element-level conservation) of the surrounding DHS peak. Conserved TFBS were selectively constrained to a similar degree regardless of the conservation of the peak in which they lie. (b) Different conserved TFBS showed different levels of selective constraint. TFBS with an enrichment for rare variation also showed an enrichment for *de novo* mutations, indicating that there is still variability in mutation rate that may be due to mutagenicity of TF binding not captured by the mutation rate model developed here.

Regulatory elements are more numerous and show greater evolutionary turnover compared to protein-coding genes[52,53]. Furthermore, the nucleotide level patterns denoting regulatory sequence are not yet as well understood as protein-coding genes. It has been hypothesized that with a sufficient number of deep whole genome sequences, functional regulatory elements could be detected by examining patterns of purifying selection and overlapping with biochemical signals associated with enhancers, promoters, or other regulatory sequence. However, if a relatively small fraction of nucleotides within an active enhancer or promoter are under selective constraint, the power to detect these elements will be significantly lower.

To address this question directly, I calculated the power to detect constrained non-coding sequence in this manner using 28,000 whole genomes (see Methods). At this sample size, there is sufficient power to detect large tracts of constrained non-coding sequence (e.g. 200bp or larger, Figure 10). For example, ultra-conserved non-coding elements or human accelerated regions where several hundred base pairs are under selective constraint could be detected at this sample size. However, as I have shown here, the vast majority of regulatory sequence is poorly conserved on an element-level, but likely harbours individual nucleotides or sets of nucleotides, perhaps constituting TFBSs, that are under selective constraint. Power to detect constrained sequence the size of a TFBS (e.g. 10-20bp) is extremely limited; power calculations indicate that upwards of 1 million deep WGS will be needed for this approach to succeed (Figure 10). However, the use of computational tools and biochemical assays to refine to identify likely TFBS *a priori* (analogous to using the protein code to identify variants likely to cause protein truncation) could be used to improve power.

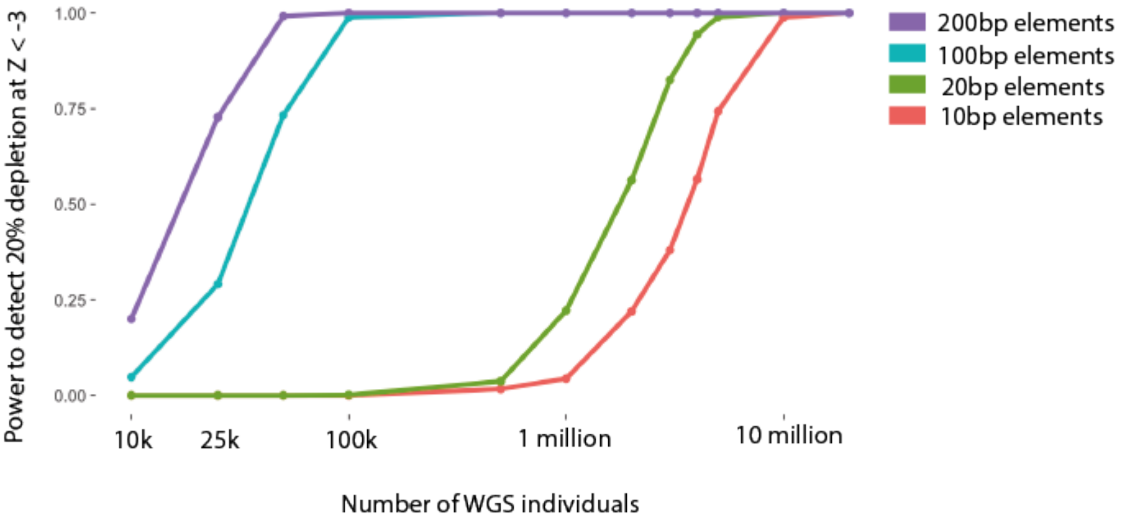


Figure 10 Power calculations for detecting constrained regulatory sequence Rare variation from 25,000 to 100,000 whole genome sequenced individuals provides substantial power to detect large tracts (e.g. 100bp – 200bp) of selectively constrained sequence. Power calculations suggest that greater than 1 million WGS

will be required to detect constraint on the level of a single TFBS (10bp-20bp).

Results Section 3.2.3. Dominance and selection in the non-coding genome

Conservation is the result of selection against fixation of non-ancestral alleles over a period of evolutionary time. Selection against fixation can manifest in a number of ways, including strong selection against heterozygotes (e.g. in the case of severe dominant disorders), weak selection against heterozygotes where the selection coefficient is greater than $1/N_e$ (the effective population size), or via selection against homozygotes, compound heterozygotes, or more complex multi-locus models.

I sought to test how selective constraint in the non-coding genome differs from constraint in the protein-coding genome, where patterns of selection and disease mechanisms are more well-understood. I annotated every nucleotide in the protein-coding exons and ENCODE open chromatin peaks with their PhyloP scores. Comparing the observed/expected ratios of protein-coding nucleotides to putative regulatory non-coding nucleotides within the same bins of evolutionary conservation reveals a pattern of selective constraint in non-coding elements that is more similar to recessive disease genes than likely dosage sensitive genes (pLI > 0.9) and known dominant disease genes (Figure 11). This result indicates that at the same degree of evolutionary conservation, regulatory elements have weaker selection on heterozygotes than protein-coding genes. This result is consistent with the limited role for *de novo* mutations in regulatory elements discussed in Chapter 2, and the predominant role for regulatory variation in common/complex disease[54,55]. However, as the ratio of observed to expected variation is most sensitive to detect strong selection against heterozygotes, these data alone do not clearly delineate between evolutionary conservation in regulatory elements being maintained by weak selection on heterozygotes, selection on homozygotes/compound heterozygotes, or oligogenic selection. Recent work in population genetics suggests that distinguishing these two modes of selection may be difficult using existing data and methodology[56].

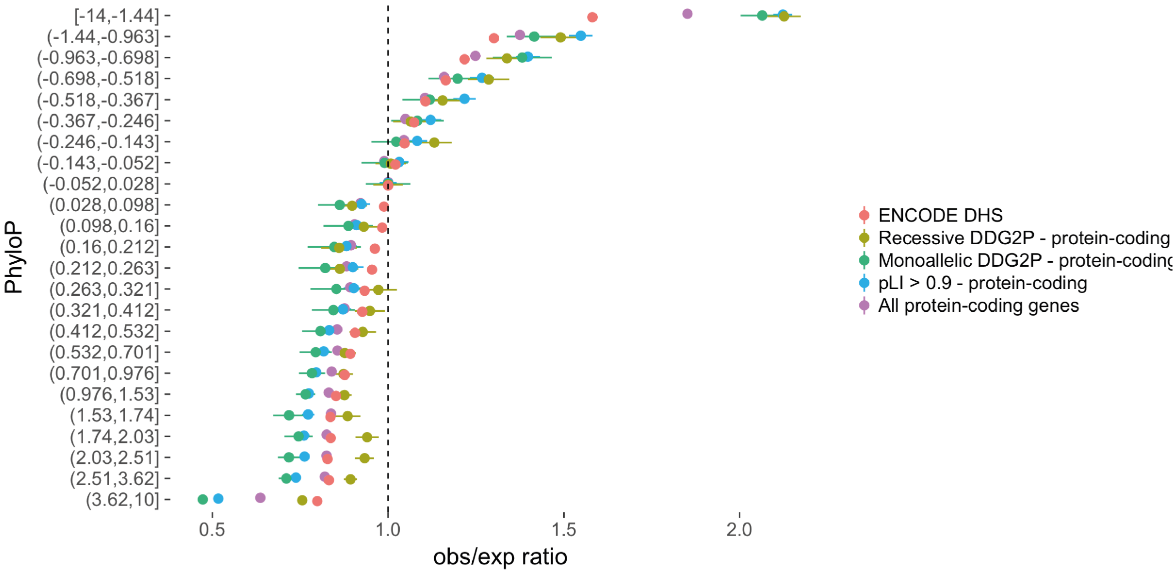


Figure 11 Comparing selective constraint in coding and non-coding elements Observed over expected ratio for various protein-coding gene sets with different hypothesized strength of selection on heterozygotes alongside the observed over expected ratio for DNase Hypersensitive Sites (DHSs). The DHSs show a pattern of selective constraint that is most similar to known recessive disease genes, whereas genes with known or suspected strong selective pressure against heterozygotes (Monoallelic DDG2P genes, genes with pLI > 0.9) show a much greater decrease in rare variation at equivalent levels of evolutionary conservation.

To further explore the strength and dominance of selection in the non-coding genome, I computed the mutability adjusted proportion of singletons (MAPS) scores for the same element sets, again stratified by PhyloP scores. I also annotated the protein-coding variants with a predicted consequence using the variant effect predictor and extracted bases predicted to result in synonymous changes, missense changes, and protein truncation. The MAPS score of non-coding variants in the top decile of nucleotide conservation was similar to missense changes in protein-coding genes genome-wide (excluding known developmental disorder genes and genes with pLI > 0.9) and consistently lower than loss of function variants (Figure 12). The results from Figures 11 and 12 together imply that there is pervasive weak selection on a small subset of evolutionarily conserved non-coding sites, and patterns of selection on these sites may be similar to those of missense changes in protein-coding exons, the majority of which have been previously reported to be weakly deleterious[57]. Further study is warranted to determine whether disease-causing alleles in regulatory elements are contributing small, additive effects, or are contributing primarily through recessive or oligogenic mechanisms.

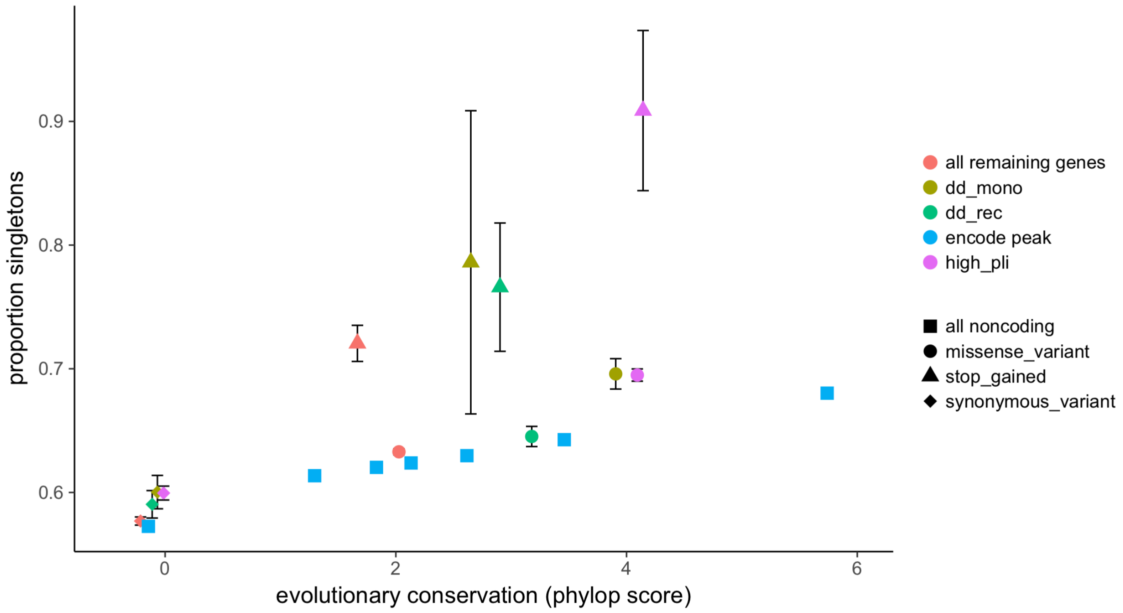


Figure 12 Comparing non-coding variation to protein-coding variation. Strength of selection measured by proportion singletons compared to nucleotide-level evolutionary conservation for protein-coding gene sets and DNase I Hypersensitive sites. Protein-coding variation was annotated with predicted effect (stop gained, missense, and synonymous) and plotted with the median PhyloP score for that category. As no such prediction is available in the non-coding genome, nucleotides in DHSs were split based on PhyloP score for comparison. Strength of selection on evolutionarily conserved nucleotides in the DHSs was similar to missense variation and substantially lower than stop-gained variants, regardless of gene set.

Discussion

A large fraction of the functional DNA in the human genome has been predicted to be non-coding. These predictions have relied primarily on evolutionary comparisons which may not be able to reveal relatively recent loss or gain of selective constraint. Furthermore, it can be challenging to deconvolute changes in mutational processes, generation times, and selective pressure across evolutionary time[2].

As large numbers of deep whole genome sequences become available, there has been great interest in applying population genetics tools to detect regions of the non-coding genome under selective constraint. However, the germline mutation rate is very heterogeneous and failure to appropriately model this heterogeneity can lead to false positives, for example in promoter regions, as well as false negatives in hypermutable regions. To this end, I have developed a new germline mutation rate model using thirty-two different features including sequence context, histone modifications in germline tissues, replication timing, and recombination rate. This model greatly improves prediction of the rate of *de novo* mutations and rare variants over the existing model which relies on sequence context alone. This mutation rate model also corroborates previously published mutation rate associated features including a recent observation of hypermutable regions with a strong bias toward maternal inheritance[8,9]. Another limitation to this mutation rate model was the lack of available chromatin data in oogonial stem cells. Chromatin marks and RNA sequencing data from ovary were used, but a tissue that is closer to the germline will likely improve the characterisation of factors influencing the maternal germline. The relationship between transcription factor binding and mutation rate should also be explored in greater detail. Analysis in this chapter on conserved transcription factor binding sites revealed an increased mutation rate in a subset of binding sites, in line with previous reports of hypermutability in CTCF binding sites in cancer[13]. Full characterisation the transcription factor binding profiles in germline tissues will likely lead to the discovery of more mutation-rate associated TF binding events and further improve understanding of mutation rate heterogeneity in the germline.

Using this improved germline mutation rate model, I modelled selective constraint in the non-coding genome using whole genome sequence data from 28,000 individuals, nearly 3-fold greater than previous non-coding constraint metrics. Furthermore, I showed that constraint on indels is much greater than SNVs across a number of coding and non-coding elements, consistent with previous results based on evolutionary divergence. I found a strong relationship between evolutionary conservation and selective constraint, but showed that this relationship was driven primarily by conservation on individual nucleotides and suggest that a substantial fraction of these sites may lie in conserved transcription factor binding sites.

While similar methodologies for detective selective constraint methods have been applied successfully to find selectively constrained protein-coding genes, the sparsity of functional nucleotides in regulatory elements presents a challenge. Power calculations suggest that 25,000 individuals provides sufficient power to detect long contiguous tracts of constrained sequence, for example ultraconserved non-coding elements, but affords little power to detect constrained TFBS within an otherwise poorly constrained element. Improvements in variant effect prediction in the non-coding genome such as identification of functional transcription factor binding sites and deleterious variation within them will improve power to detect constrained non-coding sequence substantially.

In the previous chapter, I showed that *de novo* mutations in highly evolutionarily conserved non-coding elements contribute to severe developmental disorders. Based on the analyses presented here, I hypothesize that *de novo* mutations in highly conserved bases within poorly conserved, but nonetheless active, regulatory elements may also contribute to these disorders. Identifying these functional non-coding bases is a considerable challenge and will be critical to improve power to discover pathogenic DNMs and rare variation in non-coding elements. Improvements in computational prediction of non-coding variant effects could allow for study designs that assign weights *a priori* based on a predicted functional effect (e.g. by PhyloP or a variant deleteriousness metric such as CADD), improving power akin to the implicit weighting scheme already used in protein-coding genes whereby enrichment analyses focus on missense and protein-truncating variation[58] or explicit weighting schemes such as independent hypothesis weighting[59].

Furthermore, these analyses suggest that constraint against heterozygosity in regulatory elements is in general not as strong as in coding regions. This suggests that the effect size or dominance of mutations in regulatory elements will be smaller, and there may be a greater role for recessive or oligogenic models. Results from large-scale exome-sequencing studies indicate that a substantial fraction of individuals do not carry protein-coding variant that is pathogenic with high penetrance[60,61]. Thus, it is likely that many unsolved disorders may be the result of multiple variants in the coding and non-coding regions with moderate to modest effect sizes. Approaches to integrate coding and non-coding variation, for example by analysing matched RNA-sequencing and whole genome-sequencing data, may be another strategy to interpret non-coding variant effects through their impact on transcript levels. This strategy will require sampling of the relevant tissue or cell types, as is already clinical practice in many types of cancer, or in cases where accessing the primary tissue is challenging (e.g. the developing brain), the development of cellular models or organoids to recapitulate the tissue of interest.

As whole genome sequencing is completed in tens of thousands of trios in with Autism spectrum disorder, developmental disorders, and other Mendelian disorders, there will be greater opportunities to explore the role of *de novo* mutations as well as recessive and oligogenic disease mechanisms across the whole genome in an unbiased manner.

1 Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349**, 1478-1483, doi:10.1126/science.aaa9119 (2015).

2 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**, 824-824, doi:10.1038/nrg3353 (2012).

3 Palamara, P. F. *et al.* Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *Am J Hum Genet* **97**, 775-789, doi:10.1016/j.ajhg.2015.10.006 (2015).

4 Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of Mutation Rate Variation in the Human Germline. *Annual Review of Genomics and Human Genetics* **15**, 47-70, doi:10.1146/annurev-genom-031714-125740 (2014).

5 Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *bioRxiv*, doi:10.1101/108290 (2017).

6 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

7 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**, 944-950, doi:10.1038/ng.3050 (2014).

8 Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522, doi:10.1038/nature24018 (2017).

9 Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat Genet* **50**, 487-492, doi:10.1038/s41588-018-0071-6 (2018).

10 Stamatoyannopoulos, J. a. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genetics* **41**, 393-395, doi:10.1038/ng.363 (2009).

11 Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822-826, doi:10.1038/ng.3292 (2015).

12 Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**, 337-340 (2002).

13 Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**, e1006207, doi:10.1371/journal.pgen.1006207 (2016).

14 Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10**, e1004525, doi:10.1371/journal.pgen.1004525 (2014).

15 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).

16 Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).

17 Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLOS Genetics* **11**, e1005492, doi:10.1371/journal.pgen.1005492 (2015).

18 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genetics* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013).

19 Kosmicki, J. A., Samocha, K. E., Howrigan, D. P. & Sanders, S. J. Refining the role of de novo protein truncating variants in neurodevelopmental disorders using population reference samples. *Nature Genetics*, 1-18 (2016).

20 Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, doi:10.1101/148353 (2017).

21 di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet* **50**, 333-337, doi:10.1038/s41588-018-0062-7 (2018).

22 Rhind, N. & Gilbert, D. M. DNA Replication Timing. *Csh Perspect Biol* **5**, doi:10.1101/010132 (2013).

23 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

24 Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103, doi:10.1038/nature09525 (2010).

25 Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**, 1033-1040, doi:10.1016/j.ajhg.2012.10.018 (2012).

26 Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015-1026, doi:10.1016/j.cell.2014.10.025 (2014).

27 Guo, F. *et al.* The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161**, 1437-1452, doi:10.1016/j.cell.2015.05.015 (2015).

28 Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

29 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

30 Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

31 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).

32 Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *Elife* **6**, doi:10.7554/eLife.24284 (2017).

33 Narasimhan, V. M., et. al. *Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes*. (2017).

34 Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843-2851, doi:10.1093/bioinformatics/btu356 (2014).

35 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

36 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196, doi:10.1038/nature08658 (2010).

37 Chen, C., Qi, H., Shen, Y., Pickrell, J. & Przeworski, M. Contrasting Determinants of Mutation Rates in Germline and Soma. *Genetics* **207**, 255-267, doi:10.1534/genetics.117.1114 (2017).

38 Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38, doi:10.1038/349038a0 (1991).

39 Brockdorff, N. *et al.* Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329, doi:10.1038/351329a0 (1991).

40 Gutschner, T. *et al.* The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. *Cancer Research* **73**, 1180-1189, doi:10.1158/0008-5472.Can-12-2850 (2013).

41 Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics* **6**, e1001154, doi:10.1371/journal.pgen.1001154 (2010).

42 Reilly, S. K. *et al.* Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).

43 Prabhakar, S., Noonan, J. P., Paabo, S. & Rubin, E. M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786, doi:10.1126/science.1130738 (2006).

44 Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**, 1599-1611, doi:10.1371/journal.pgen.0020168 (2006).

45 Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci* **368**, 20130025, doi:10.1098/rstb.2013.0025 (2013).

46 Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346-1350, doi:10.1126/science.1159974 (2008).

47 Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior Article Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341-354.e312, doi:10.1016/j.cell.2016.08.071 (2016).

48 Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152-163, doi:10.1038/s41559-017-0377-2 (2018).

49 Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99, doi:10.1186/1471-2164-5-99 (2004).

50 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

51 Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769-1776, doi:10.1101/gr.116814.110 (2011).

52 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).

53 Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**, 1335-1343, doi:10.1101/gr.108795.110 (2010).

54 Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

55 Hindorff, L. a. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

56 Fuller, Z., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *bioRxiv*, doi:10.1101/382481 (2018).

57 Kryukov, G. V., Pennacchio, L. a. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics* **80**, 727-739, doi:10.1086/513473 (2007).

58 Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, doi:10.1038/nature21062 (2017).

59 Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome- scale multiple testing. **13**, doi:10.1038/nmeth.3885 (2016).

60 Mcrae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438, doi:10.1038/nature21062 (2017).

61 Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. *bioRxiv*, doi:10.1101/201533 (2017).

Chapter 4: Functional characterisation of mutations in ultra-conserved elements associated with severe developmental disorders

Introduction

Methods for assessing enhancer activity using reporter constructs

While there are robust computational tools to identify putative protein-coding genes based on sequence alone, detecting transcriptional enhancers has proved more challenging. Large scale collaborations such as the ENCODE project and the Roadmap Epigenome Project have experimentally profiled hundreds of different cell types and tissues to identify putative enhancers, promoters, and other genomic elements[1,2] (these projects are discussed in greater detail in Chapter 1). While these projects have identified millions of putative regulatory elements across hundreds of different tissues and time points, these data are observational, presenting a number of drawbacks. First, regulatory elements active at a tissue or time point that is difficult to sample, for example very early embryonic development, may not be identified. Second, these efforts have been focused around assaying a small number of individuals per tissue/time point, limiting the insight into the impact of variation on enhancer function. Thus, while resources such as ENCODE or the Roadmap Epigenome Project provide a rich resource for identifying active regulatory elements, these resources are unable to predict the impact on regulatory element function by the introduction of a genetic variant.

Reporter assays are one of the most widely used methods for assessing promoter and enhancer activity. Early reporter assays made use of a modified version of the firefly luciferase gene in a mammalian expression vector[3]. Putative promoter or enhancer sequence could be inserted upstream of the luciferase gene and luciferase-induced fluorescence could be quantified as a proxy for activity. These assays have improved and diversified in many ways since their first use. To assess enhancer activity rather than promoter activity, the putative enhancer can be inserted into a vector with a minimal promoter upstream of the luciferase gene. Shortly after the introduction of the luciferase reporter gene, many other read-outs of enhancer and promoter activity were developed, including the use Green Fluorescent Protein[4] (GFP) and lacZ staining[5,6].

Mouse transgenesis assays using lacZ staining have been used to assess the impact of regulatory mutations identified in patients[7], study the role of ultraconserved regulatory elements in brain development[8], and characterise ultraconserved regulatory elements that have adopted new functions in humans[9]. Over the past decade, expression patterns in the developing mouse embryo from nearly 3,000 putative enhancers have been collected in the VISTA database[10]. While the reporter assays described above are amenable to testing dozens of putative enhancers or promoters in a single experiment and provide a powerful view into the complexity of enhancer function with respect to tissue and developmental time, they are impractical to scale to hundreds or thousands of tests. This limitation has led to a number of different high-throughput reporter assay techniques known collectively as massively parallel reporter assays (MPRAs).

Massively parallel reporter assays methods and applications

MPRAs make use of oligo synthesis to generate libraries of tens of thousands of putative enhancers or promoter elements to test in a single experiment. Each enhancer is assigned to a unique 15-20bp DNA barcode either during the oligo synthesis step[11] or added after oligo synthesis using PCR[12]. Paired-end sequencing is then used to identify the enhancer-barcode pairs. Next, a reporter gene, typically GFP is inserted between the enhancer and the barcode. As a result, the barcode is situated on the 3' end of the reporter gene and included in the gene transcript. This library of barcoded enhancers is then transfected into the cell type of interest, the RNA is harvested and reverse transcribed into cDNA, and the short barcode fragments are sequenced to quantify the transcriptional output of the upstream enhancer. Comparing the transcriptional output to the amount of DNA input, the enhancer potential of a putative regulatory sequence can be assessed. This approach, which has been applied to detect eQTLs, is described in detail in *Tewhey et. al*. An alternative approach, STARR-seq, does not make use of DNA barcodes. Instead, the enhancer is inserted downstream of the reporter gene in the viral vector[13]. As the viral vector is circular, the enhancer can drive expression of the reporter gene, transcribed along with the gene, and used as a readout for its own activity.

The first application of the MPRA was a proof of concept saturation mutagenesis experiment in a synthetic enhancer[11]. Since this proof of concept, the approach has been used in a number of different applications including construction of synthetic enhancers and dissection of enhancers with links to disease or interesting evolutionary properties. *Tewhey et. al* used an allele specific expression strategy to compare enhancer activity of the SNPs in linkage disequilibrium with the most significantly associated SNP in gene expression quantitative trait loci[12] (eQTLs). *Smith et. al,* generated approximately 5,000 synthetic enhancers containing twelve different transcription factor binding sites (TFBSs) important for function in the liver, arranged with varying degrees of complexity, from testing more than 500 sequences with a single TF binding site and more than 2,500 putative enhancers with three or more transcription factor binding sites. This work revealed general principles of enhancer function consistent with the 'billboard model' including increased expression correlating with the number and heterogeneity of TFBSs. The authors also demonstrated synergistic and antagonistic interactions between different transcription factors (TFs). *Ryu et. al,* recently used a series of MPRA to test 714 putative human accelerated regions, conserved non-coding elements with a greater number of human-specific variation than expected under a neutral model, for enhancer activity in induced pluripotent stem cell derived neuronal progenitors. Different MPRA variations and applications are reviewed in *Inoue and Ahituv* [14].

After generating MPRA data, statistical methods are needed to quantify the effect of individual variation on reporter gene activity relative to a reference sequence. Many of the first studies in this field drew on previous work in allele-specific expression (ASE) in RNA-sequencing to analyse MPRA data[11,12,15]. However, there are major differences between these two types of analysis, most notably that the input DNA sequence in a MPRA experiment may not be the same between the reference and alternate alleles, whereas in ASE studies, the ratio of DNA from the two alleles is equal. Recently, the QuASAR-MPRA approach was developed which incorporates plasmid proportions and uses a beta-binomial distribution model variance[16]. Applying this method to previously published data[12,16] indicates that this approach reduces false positives while improving power to detect ASE in MPRAs. Other statistical methods have been developed for different types of analysis. For example, Sharpr-MPRA was designed to analyse data from oligonucleotides densely tiled across a larger regulatory element to make inferences about which nucleotides are functional within the regulatory sequence[17].

While mouse transgenesis assays and MPRAs have proved to be powerful tools, they come with important caveats to consider. In the case of MPRAs, the activity of a putative enhancer or promoter will depend on the cell type or tissue in which the assay is conducted, not least due to the expression of different transcription factors in different cell types [15]. Likewise, gathering sufficient RNA to maintain library complexity requires high transfection efficiency or a large amount of cellular material to balance out low transfection efficiency. These technical limitations may limit the choice of cell types considerably and potentially bias studies toward cell lines that are experimentally tractable but less biologically relevant. Moreover, many experimental approaches involve the reporter gene being expressed in an episomal context (i.e. not integrated into a chromosome), and it has been suggested that integration into the genome, using a lentiviral vector, provides a more biologically realistic context[18]. It is currently challenging to synthesise elements longer than approximately 150bp. In the case where the true functional sequence is longer than 150bp, a MPRA may not faithfully represent *in vivo* function. For these reasons, no individual functional assay should be considered conclusive, either of enhancer function or the impact of genetic variation on enhancer function. Integrating data from evolutionary genetics, medical genetics, and multiple experimental approaches is likely to provide greatest insight into gene mis-regulation as a mechanism for disease.

This MPRA data discussed in this Chapter are derived from a pilot project designed as a follow-up to the results presented in Chapter 2. Specifically, I sought to test the evidence for enhancer activity of the CNEs sequenced in the DDD project, assess the impact of SNVs and indels in these CNEs, and to compare the impact of patient mutations to common and rare variation observed in unaffected individuals. Sebastian Gerety and Matt Hurles contributed a substantial amount of work and oversight in the study design and Sebastian Gerety and Holly Ironfield performed all of the wet-lab experiments generating the underlying data.

Methods

MPRA oligo design

Reference sequence was included in the oligo synthesis for the following element sets:

1. Conserved non-coding elements and enhancers with at least one DNM (SNV or indel) in 7,930 probands (n = 767 unique sites) from *Short et. al, 2018*.

2. Control elements assayed in U87 cells and neural progenitor cells
   - 90 elements with varying evidence for biochemical activity in Neural Progenitor Cells (30 low, 30 medium, 30 high from data kindly provided by Barak Cohen and Brett Maricque, described in Maricque et. al 2016).

3. Positive control regulatory elements
   - Three high-confidence neural *cis*-regulatory elements, derived from data kindly provided by Barak Cohen and Brett Maricque, from a LV-MPRA assay in neural progenitor cells described in Maricque et. al, 2016, chr2:72898217-72898346, chr17:44916284-44916413, and chr3:71241866-71241995 in GRCh37 coordinates.

The following subset of elements were selected for saturation (all possible SNV changes)

1. Recurrently mutated fetal brain active elements from *Short et. al, 2018*
   - 64 unique sites across 31 elements

2. Positive controls
   - Three high-confidence neural *cis*-regulatory elements, derived from data kindly provided by Barak Cohen and Brett Maricque, from a LV-MPRA assay in neural progenitor cells described in Maricque et. al, 2016, chr2:72898217-72898346, chr17:44916284-44916413, and chr3:71241866-71241995 in GRCh37 coordinates.

The following elements were selected and variants were synthesized with non-overlapping 5bp deletions across the entire element:

1. Conserved non-coding elements and enhancers with at least one DNM (SNV or indel) in 7,930 probands (n = 767 unique sites).

2. Positive controls
   - Three high-confidence neural *cis*-regulatory elements, derived from data kindly provided by Barak Cohen and Brett Maricque, from a LV-MPRA assay in neural progenitor cells described in Maricque et. al, 2016, chr2:72898217-72898346, chr17:44916284-44916413, and chr3:71241866-71241995 in GRCh37 coordinates.

All of the oligonucleotides for this experiment were synthesized by Agilent Technologies using a 244K array. Oligos are 180bp long, within 15bp of adapter sequence on the 5' and 3' end, and 150bp of genomic context in between.

Four different adapters were designed in order to allow for sub-pooling of the library to improve complexity throughout the experiment.

POOL_A_F ACTGGCCGCTTGACG
POOL_A_R CACTGCGGCTCCTGC

POOL_B_F CTGCGCCTGATGCAG
POOL_B_R GGTGCTCGCTATCGC

POOL_C_F TACGCTAGCCCGTGG
POOL_C_R TGCGTTTGGCAGGAC

POOL_D_F AGTCAGGACCGACGC
POOL_D_R AGCGCTTTCGCCCAC

The POOL_A adapter sequences were the same as the adapter used in Tewhey et. al, 2016, which did not employ any sub-pooling strategy.

The four separate adapter sequences were used to define four different pools:
Pool 1 - first 22 recurrently mutated elements (lexicographical ordering) in exome-negative probands (ref, saturation, and tiling indels)
Pool 2 - second 21 recurrently mutated elements (lexicographical ordering) in exome-negative probands (ref, saturation, and tiling indels)
Pool 3- remaining 21 recurrently mutated elements (lexicographical ordering) in exome-negative probands (ref, saturation, and tiling indels)
Pool 4 – remaining 703 elements with single DNM (ref, tiling indels from exome-positive and exome-negative)

One of each of the three positive controls from *Maricque et. al, 2016* were included in each pool. The 90 common controls were included in all three pools to allow for normalization and comparison between pools.

Reference Testing (4 oligos)

- Wildtype sequence in the forward direction with the DNM centered at +25bp, +75bp, and +125bp.
- Wildtype sequence in the reverse direction with the DNM centered at +75bp.

Saturation mutagenesis (453 oligos)

- 150bp of reference genomic sequence, with DNM at position +75, every position in sequence changed to three possible alt SNVs. (450 oligos)
- 150bp of reference genomic sequence, with DNM at position +25, only site where DNM is located is changed to alt. (1 oligo)
- 150bp of reference genomic sequence, with DNM at position +125, only site where DNM is located is changed to alt. (1 oligo)
- 150bp of reversed reference genomic sequence, with DNM at position +75, only site where DNM is located is changed to alt. (1 oligo)

- If there is a variant present in addition to the DNM within the 150bp window, one extra oligo is generated including both the variant and DNM to analyse the full haplotype.

Tiling indels (30 oligos)

- 5bp deletions beginning at position +0, +5, +10, …, +145, with respect to the DNM at position +75bp.

All oligos were generated using the oligo_design.R script in the MPRA_ddd project: https://github.com/pjshort/MPRA_ddd/

In total, 56,688 oligos were generated:

- 3,068 oligos were generated with four different ref sequences (767 unique DNMs, see above for different ref sequences).
- 363 sequences were generated with ref in the forward direction (90 common controls with each of four adapters, plus 3 positive controls with one adapter each).
- 30,150 oligos were generated with a single nucleotide of wild-type sequence changed (64 unique DNMs in recurrent elements + 3 positive controls, 450 oligos each).
- 7 oligos where child has other common or rare variants within 150bp oligo in addition to the observed DNM.
- 23,100 oligos were generated with a 5bp deletion of wild-type sequence (767 unique DNMs, 3 positive controls, 30 oligos each.

Adding barcodes and pool adapters to oligos

In order to easily identify individual oligo plasmids, 20bp barcodes were added by PCR to the oligo pools. Primers were designed to specifically amplify each subpool, adding a 3' 20-base random barcode to each fragment, and the necessary overlaps with the target vector to perform Gibson cloning. For pools A-D, these were forward primers #430-433, and reverse primers #434-437. For each pool (A-D), 300 uL PCR reactions (spread across a 96 well plate as 6 X 50ul reactions) were run using 6ng of oligo library template. Q5 NEB polymerase was used, and all reactions were done at 15 cycles: this gave us sufficient material for downstream cloning while avoiding over-amplification and potential bias in the oligo pool. The resulting PCR products were treated with ExonucleaseI (NEB) to remove unincorporated barcode primers, and then purified using standard SPRI bead methods (Ampure, Agilent).

Creating the MPRA vector library

The plasmid backbone used in this analysis (pGL4:23:ΔxbaΔluc) was provided by Ryan Tewhey and is the same backbone used in *Tewhey et. al, 2016.* This vector was prepared by PCR, incorporating sequence overlapping the library oligos (#525 and 526), thus enabling the use of Gibson assembly. 1 microgram(ug) of vector was combined with 1 ug of purified oligo in a 40ul Gibson assembly reaction, using standard conditions. The gibson reactions were purified using standard SPRI bead isolation, eluting in 20ul of elution buffer (EB, Qiagen). 10ul of this eluate was electroporated into 100 ul of high efficiency electrocompentent bacterial cells (C3020K, NEB) using recommended protocols and parameters: 0.1mm cuvettes (Biorad), with settings of 2KV,200 ohm,25uF. After recovery, cells were plated on large 22.5cm X 22.5cm agar/ampicillin plates at around 2 million CFU per plate. Plating density was confirmed by serial dilution, plating, and counting. After overnight at 37 degrees, the cells from the plates were harvested in LB broth, and plasmid DNA was prepared using two Qiagen PLUS Midiprep columns (Qiagen) per subpool. This yielded around 250ug per subpool.

These oligo-barcode libraries were sequenced in order to determine the oligo-barcode pairs for downstream analyses. This was done as described in *Tewhey et. al*, with modifications, using paired end 150bp Illumina chemistry.

To generate the final MPRA libraries, containing an open reading frame downstream of the oligo/elements, we cut 15ug of each pool DNA with Sgf1 followed by SPRI bead purification. We generated PCR amplicons containing the GFP ORF (primers #426-450), and ligated 2ug of this to 2ug of the cut vector pool using Gibson assembly. These reactions were purified using SPRI beads, re-cut with AsiSI, and purified using SPRI beads. The reactions were then electroporated into high efficiency electrocompetent bacterial cells as described above.

After recovery, these transformed cells were grown in LB+Carbenicillin for 9 hours, harvested, and purified using Qiagen PLUS maxiprep columns (4 columns per pool). The initial electroporation was also serially diluted and plated on LB/agar+AMP plates to estimate yield. Each electroporated pools each gave around 8 X 10^8 CFUs.

Transfection into HeLA and Neuroblastoma cell lines

The Neon Transfection System (ThermoFisher Technologies) was used to transfect HeLA cells and SHSy5y human neuroblastoma cells (https://www.lgcstandards-atcc.org/Products/All/CRL-2266.aspx) in triplicate. We electroporated 5 million cells with 30 ug of plasmid pool DNA per replicate, using 100uL NEON tips. For HELA, we used 1005 Volts, 35ms pulse, with 2 pulses. For SHSy5y cells, we used 1200 volds, 20ms pulse, for 3 pulses. Cells were recovered into standard growth media, and allowed to grow for two days. The presence of GFP expression was confirmed by epifluorescence microscopy. Cells were then trypsinized, spun down and snap frozen for RNA extraction.

RNA taqSeq library preparation

RNA extraction from cell pellets was done using RNeasy columns (Qiagen). All RNA samples were DNAse treated (TURBO DNA-*free* Kit, Thermofisher) to remove any residual plasmid DNA. cDNA was synthesised using 2.5 ug of total RNA, and a primer specific to the plasmid-derived 3' UTR (primer #543), thus ensuring enrichment for plasmid-derived transcripts (SuperScript IV First-Strand Synthesis System, Thermofisher). To introduce 15bp unique molecular identifiers (UMIs), we performed a second strand synthesis (primer extension reaction) using a primer that annealed 125 bases upstream of the barcode, and included 15 random bases in addition to Illumina partial adaptor sequences for library construction (primer #539). After SPRI bead purification, these uniquely labeled cDNA molecules were then amplified in two rounds of 15 cycles of PCR, to progressively add the necessary Illumina adaptors (primers #544,535), then index barcodes (Illumina 11bp index set, PE 1.0). Purified library DNA was then subjected to 25bp paired-end sequencing on an Illumina HiSeq4000 to identify barcode expression (3' end) and UMI identity (5' end). Each replicate gave around 165 million read-pairs.

The plasmid library DNA was sequenced in an identical manner, starting with primer extension step using plasmid DNA and UMI containing primer (primer #539). These plasmid counts provide the DNA input amounts with which we normalise element expression.

Linking elements to barcodes with 150bp paired-end sequencing data

First, I fused the 150bp paired-end reads to a single read using the FLASH[19] with the following flags:

```
flash -r 150 -f 220 -s 10 PoolA_R1.fastq.gz PoolA_R2.fastq.gz -o PoolA.Mar2018.Lane1
```

This will output the extended fragments as PoolA.Mar2018.Lane1.extendedFrags.fastq.

Next, the fused reads were aligned to the reference oligos using BWA mem[20] version 0.7.13:

```
bwa mem -v 0 ddd_noncoding_MPRA.refs_and_controls.dups_removed.fasta
PoolA.Mar2018.extendedFrags.fastq > PoolA.Mar2018.sam
```

A custom python script was written using pysam to reconstruct the element sequence from the alignment and populate a table of element-barcodes pairs.

Counting UMI-labelled barcodes from 25bp paired-end sequencing data

A custom python script was written using pysam to extract the barcode and UMI from the 25bp paired-end sequences and populate a table of unique barcode-UMI pairs with the total number of reads observed for each pair.

In order to correct for sequencing errors in the barcodes, any barcodes not matching a previous element-barcode, the edit distance was calculated for all known element-barcode pairs. If a barcode was within two edits of a known barcode from the previous step, it was corrected to this barcode. If there were multiple matching barcodes, the count was excluded.

Calculating normalised expression values in HeLa and Neuroblastoma and testing correlation between replicates

Elements were split into four different sub-pools (as described above) and each sub-pool was tested in HeLa and Neuroblastoma in three independent biological replicates. For each sub-pool and replicate, normalised expression was calculated as the ratio between RNA and DNA, normalised by the total number of RNA and DNA UMIs sequenced:

$$expr_{element} = \frac{\mathrm{RNA}_{\mathrm{element}} / \mathrm{RNA}_{\mathrm{total}}}{\mathrm{DNA}_{\mathrm{element}} / \mathrm{DNA}_{\mathrm{total}}}$$

Spearman rank correlation was used to assess the correlation of normalised expression values across different experimental replicates.

Results Section 4.1: Assessing disease-associated enhancer activity using massively parallel reporter assays

Results Section 4.1.1. Design of a MPRA experiment to assess enhancer activity of elements harbouring *de novo* mutations in patients

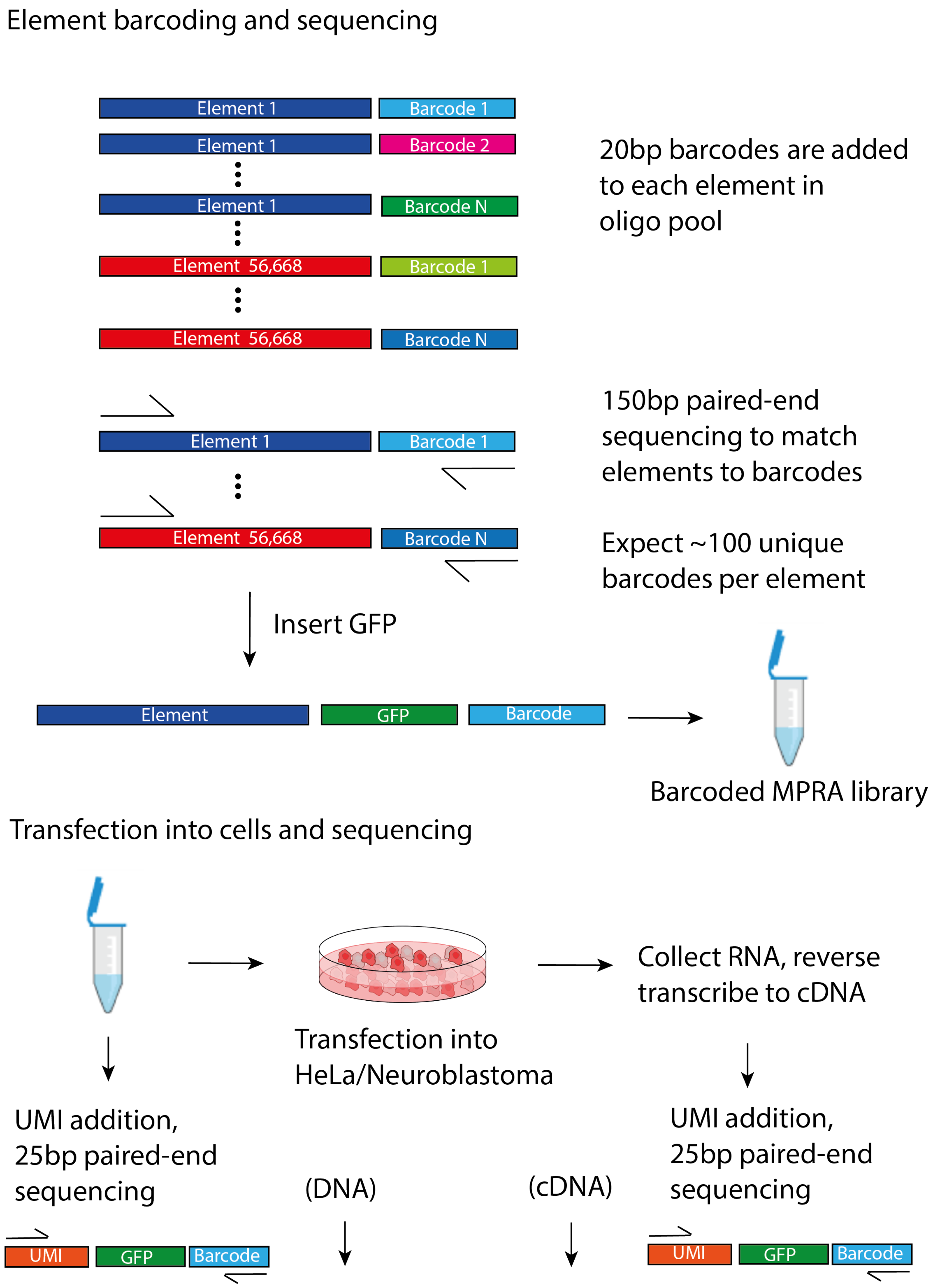
In Chapter 2, I demonstrated an enrichment for damaging *de novo* mutations (DNMs) in evolutionarily conserved non-coding elements and I showed that a substantial fraction of these elements are likely acting as enhancers. However, determining the precise effect of these regulatory DNMs remains a substantial challenge. This hampers the discovery of novel genetic associations and in resolving variants of unknown significance in patient genomes. Putative pathogenic variants in regulatory elements have been assayed previously using zebrafish models, mouse transgenesis assays, and mouse knock-ins. However, these experimental assays can only test tens of variants due to prohibitive cost or experimental complexity. Massively parallel reporter assay (MPRAs) allow traditional reporter assays to be scaled to test tens of thousands of variants in a single experiment. Thus, patient DNMs within the elements of interest can be tested alongside population variation as well as variants that have not yet been observed, providing insight not only into the impact of specific mutations discovered in patients, but also, potentially, a systematic overview of nucleotide-level functional importance within an element.

The primary aim of this experiment was to serve as a pilot experiment to shape future studies using MPRAs as a tool to evaluate putative pathogenic regulatory variation. I chose to focus on the elements that were most likely to be harbouring pathogenic regulatory variants based on the analysis in Chapter 2. While conserved non-coding elements are depleted for genetic variation, they still harbour rare and common variation that is found in healthy individuals, and one of the goals of this pilot was to compare putative pathogenic variation found in patients to variants observed in healthy individuals. I also sought to compare the impact of SNVs and deletions in these elements, with the hypothesis that deletions would be more disruptive to element function. Published work has shown a higher impact from deletions than SNVs, albeit in the context of a synthetic enhancer. If deletions do cause a more pronounced effect, they may also be a more efficient way of determining which regions in an enhancer are critical for function.

As part of the pilot project, I also ran the experiment in HeLa cells as well as neuroblastoma cells. While HeLa cells are experimentally tractable and have been used in previous MPRA work, they may not contain the relevant TFs for expression of elements likely to be active in neural tissues. In contrast, Neuroblastoma expresses many of the essential neuronal markers[21], but has lower transfection efficiency than HeLa cells. Thus, an essential part of the pilot project was determining whether we can observe cell type-specific enhancer activity, and whether choice of cell type dramatically impacts any conclusions.

To this end, I designed a series of MPRA experiments testing 56,688 different enhancer sequences based on results from Chapter 2. These 56,688 sequences included thirty-one genomic elements with DNMs observed in multiple families. A total of 64 independent DNMs were identified in these recurrently mutated elements. These 64 DNMs were included in the analysis with 150bp of genomic sequence (74bp upstream and 75bp downstream). These 64 elements were also synthesized with every possible SNV change from the reference sequence throughout the 150bp element. In the case that two DNMs fall within 150bp, a saturated element was programmed for each DNM and thus the DNMs would be evaluated twice in different contexts—once in the center of a saturated element, and a second time in the oligo centred on the other DNM. To mitigate the risk of false negatives due to the 150bp element centred on the DNM would be non-functional, I also included the wild-type sequence of the element with the DNM at the 25th base, at the 125th base, and at the 75th base with the element in reverse orientation.

A total of 767 elements, including the 64 selected for saturation SNV editing, with a DNM observed in one patient were synthesized with the reference sequence as well as five base pair deletions tiled across the element (see Methods). Control elements assayed in a previously published MPRA conducted in U87 glioblastoma cell line and neural progenitor cell line were also included[11,22] (see Methods). The experimental workflow for this analysis was based on *Tewhey et. al,* but included adaptations to the protocol most notably the addition of unique molecular identifiers (UMIs) to ensure all sequencing steps are reflective of the original input material and not subject to overcounting or PCR biases. This work was led by Sebastian Gerety in the laboratory and I worked on the experimental design, bioinformatics, and statistical analysis. An overview of the experimental design and laboratory protocol is provided in Figure 1, and described in the Methods section.


Element barcoding and sequencing
Element 1
Barcode 1
Element 1
Barcode 2
Element 1
Barcode N
Element 56,668
Barcode 1
Element 56,668
Barcode N
20bp barcodes are added to each element in oligo pool
Element 1
Barcode 1
Element 56,668
Barcode N
150bp paired-end sequencing to match elements to barcodes
Expect ~100 unique barcodes per element
Insert GFP
Element
GFP
Barcode
Barcoded MPRA library
Transfection into cells and sequencing
Collect RNA, reverse transcribe to cDNA
Transfection into HeLa/Neuroblastoma
UMI addition, 25bp paired-end sequencing
(DNA)
(cDNA)
UMI addition, 25bp paired-end sequencing
UMI
GFP
Barcode
UMI
GFP
Barcode

Barcode	DNA count	cDNA count rep 1	cDNA count rep 2	cDNA count rep 3
ATTA...TCAGT	10,322	11,354	9,932	12,004
GGGC...AATCT	1,492	12	0	85
⋮	⋮	⋮	⋮	⋮
CTTG...GTATG	4,776	5,004	5,517	5,300

Figure 1 Overview of the MPRA method adapted to include UMIs.

The oligo synthesis design included 28,800 SNVs and 23,010-thousand five base pair deletions in non-coding elements with mutations observed in patients with severe developmental disorders (see Methods). However, oligo synthesis is error-prone and as a result a far greater diversity of variation was observed than what was originally designed. In particular, 22% of oligonucleotides had a one base pair deletions and 7% had large deletions (>20bp). By comparing the sequencing reads spanning the same element-barcode pairs, I determined that these deletions were likely present in the oligo-synthesis step, not as a sequencing error. Figure 2A shows the distribution of actual observed deletions sizes relative to the reference compared to what was expected based on the array design. However, due to the quasi-random nature of the deletions, the vast majority of 'non-programmed' variants were observed at lower frequency in the barcoded library (Figure 2B).

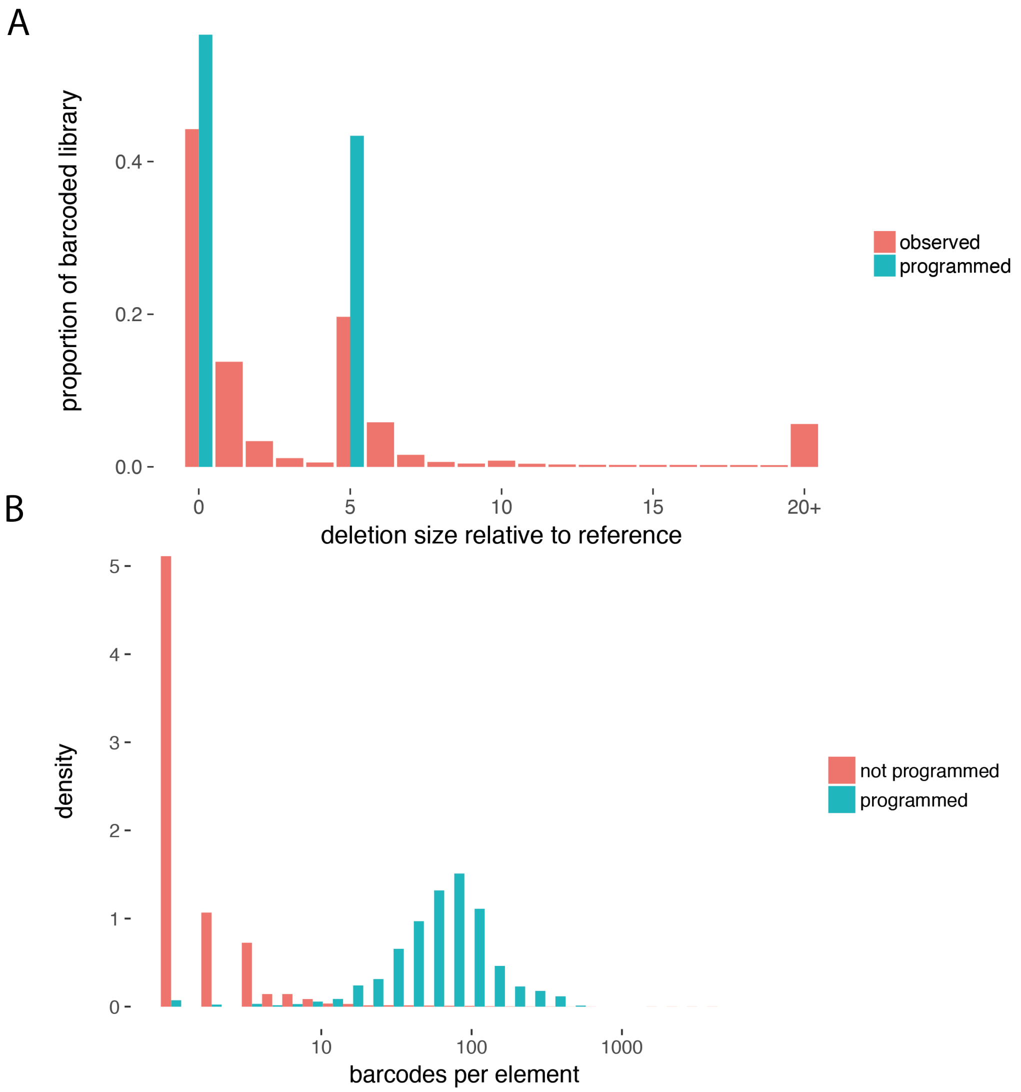


Figure 2 Linking elements to barcodes. (a) Oligo-synthesis is error-prone and introduces non-programmed variation. One base pair deletions are the most common error mode, but large deletions (>20bp) are also common. (b) Most programmed oligonucleotides are present with tens to hundreds of different barcodes in the input DNA, while non-programmed variation is at a much lower abundance in the library.

Previous MPRA work has not, to my knowledge, addressed the issue of errors in oligo-synthesis directly. Instead, they have used approximate string matching to match non-programmed variation to expected variation within a stated error tolerance (typically two edits)[12]. In cases where barcodes are synthesized on the oligo rather than ligated in a separate step, synthesis errors cannot be easily separated from sequencing errors[11,15]. In either case, pooling data from oligonucleotides sharing the same programmed edit but with different non-programmed edits may introduce noise to the experiment. For this reason, I chose to begin by strictly filtering to ensure that any inferences made about variant effect were due to the precise programmed variation and not due to error-prone oligo synthesis. I successfully recovered barcodes for 27,146/28,800 (94.2%) of SNV-containing sequences and 21,238/23,010 (92.3%) of 5bp deletion-containing sequences. Of the 64 elements selected for saturation, 61/64 elements had near complete saturation (median 447/450 SNVs recovered), while 3/64 elements were very poorly represented (Figure 3A). These three elements had long runs of repeated sequence. These elements had a high error rate during oligo synthesis and were also observed at lower abundance, indicative of low PCR efficiency. Due to these technical limitations, these three elements were excluded from further analyses. The programmed deletions were also well-represented in the library. Of the 767 elements selected for tiling 5bp deletions, >80% of elements had at least 27/30 deletions observed (Figure 3B).

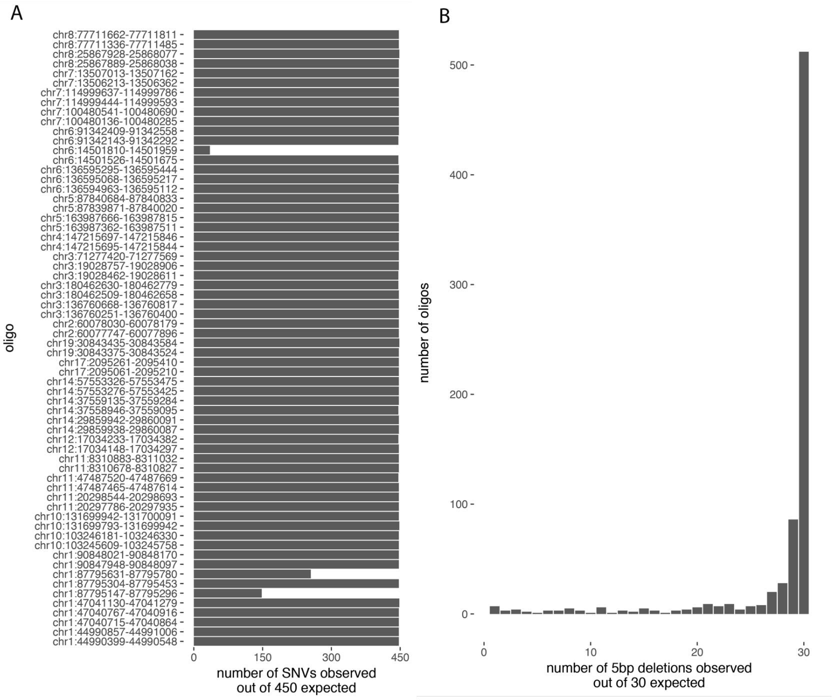


Figure 3 Proportion of programmed SNVs and deletions. (a) 61/64 elements selected for saturation had nearly all of the 450 expected SNVs observed after the barcoding step. (b) of the 767 elements

selected for tiling 5bp deletions, recovery of observed deletions was also very high.

The expression of a putative regulatory element in a massively parallel reporter assay is calculated by dividing the observed RNA levels after transfection by the input DNA levels. Previously published MPRA studies have calculated normalised expression based on dividing the proportion of sequenced reads matching element-associated barcodes in the cDNA by the proportion of those barcodes in the original library DNA library used in transfections[12]. One concern with this approach is that performing PCR before sequencing may result in PCR biases confounding expression estimates as well as redundant sequencing of the same original DNA molecule in cases where the amount of starting material is low. This has been observed in bulk RNA-sequencing[23] and single cell RNA sequencing[24], and these biases have been addressed using unique molecular identifiers (UMIs), but this approach has not yet been applied to MPRA to my knowledge. In order to test and potentially correct for overcounting the same DNA molecule, the protocol described in *Tewhey et. al* was adapted by Sebastian Gerety in the Hurles Lab to include incorporation of UMIs to each DNA molecule prior in the plasmid library prior to PCR amplification and sequencing as well as to each cDNA molecule harvested from HeLa and neuroblastoma cell lines prior to sequencing (Figure 1).

I found that in the cDNA pools, each individual molecule was sequenced, on average, just over three times, implying that the complexity of input material used in sequencing was too low. Without the use of UMIs, it would be difficult to know whether the amount of input material was sufficient. In the case of low input material, counting sequencing reads rather than UMIs would result in an artificially high level of certainty around estimates of variant effect due to overestimate of the denominator (input DNA) and numerator (output RNA) in estimates of allele specific expression (ASE). In the case of abundant input material, I expect that the UMIs will reduce the variance between independent controls by reducing the noise due to sequencing the same original molecule twice, due to sampling with replacement. For these reasons, I used the UMI counts, rather than the total read counts, in all analyses going forward.

The choice of cellular system used in MPRA has shown to impact the regulatory activity of putative enhancer sequence[22]. For this reason, I chose to test the enhancer elements in both HeLa cells and neuroblastoma cells. HeLa cells are more easily transfected and have been used extensively in MPRAs, but are lacking in many of the tissue-specific transcription factors important for activity of developmental enhancers. I sought to test the robustness of the assay in HeLa and neuroblastoma by testing the expression of putative enhancer sequences in multiple independent experiments and comparing results. I calculated the ratio of RNA output to DNA input within a given pool, normalised by the total number of unique UMIs observed (referred to as 'normalised expression' throughout) in three independent biological replicates in both HeLa cells and Neuroblastoma cells. To reduce the likelihood of PCR biases causing particular elements to be at much higher proportion in the MPRA library, all experiments were performed in four different sub-pools using specific primers included at the oligo synthesis step, resulting in a total 24 different experiments (see Methods).

I first calculated the normalised expression of each wild-type reference sequence and compared the estimated expression in independent biological replicates. The correlation was unexpectedly low compared to that reported in *Tewhey et. al* (Figure 4A, $r^2 = 0.06$, $p = 0.0002093$) and there were a large number of elements that showed evidence for expression in one replicate and no expression in another. I reasoned that this may be the result of a bottleneck causing elements at low abundance in the pool to have high variance between pools. Ranking the elements by their abundance in the plasmid pool and restricting to the top 10%, I find a much higher correlation between independent replicates (Figure 4B, $r^2 = 0.58$).

This phenomenon was not reported in *Tewhey et. al*, so I downloaded the publicly available data to see if I could reproduce the same phenomenon. I saw no evidence for drop-out or markedly high variance at low abundance plasmids in the *Tewhey et. al* data, suggesting that differences in the experimental protocol, potentially due to the lower number of cells used in our experiment (5 million cells versus 100 million) or the transfection method used (electroporation versus lipofectamine), may have resulted in a bottleneck in our data that is not present in the data from *Tewhey et. al*. As of this writing, the source of this loss of complexity in our experimental workflow has not been resolved, so all analyses going forward have been restricted to include elements above an abundance threshold of 0.0014 (see Figure 4C). As a result, approximately 40% of programmed variants were excluded from the analysis. Even at this strict abundance cutoff, the correlation between replicates in the publicly available data from *Tewhey et. al* is higher (r^2 = 0.58 versus r^2 = 0.64). Thus resolving the potential bottleneck in our experimental pipeline is critical to improve correlation between experimental replicates.

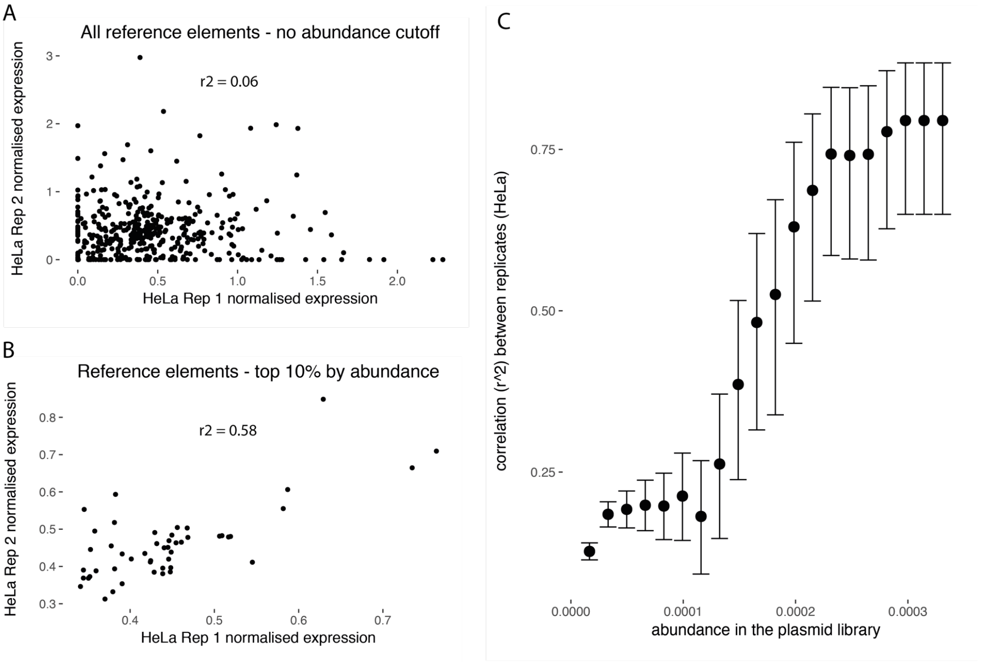


Figure 4 Correlation between biological replicates. (a) the number of unique plasmid molecules in the starting library correlates strongly with correlation between independent biological replicates. Increasing the starting material to tens of thousands of molecules per DNA sequence tested will improve reproducibility. (b) Correlation between normalised expression in HeLa cell line replicate 1 and replicate 2 (r^2 = 0.76).

After establishing criteria for reliable correlation between replicates, I next sought to test the enhancer activity for the wild-type sequence of all 767 conserved non-coding elements (CNEs) and enhancer sequences included in the MPRA design. The expression of the reference sequences assayed showed bimodal distribution in both HeLa and Neuroblastoma cell lines (Figure 5A,B), indicating that a subset of elements are likely driving robust expression, while others are not acting as enhancers in that particular cellular context.

I fit a two-component gaussian mixture model to the expression data to classify elements as likely active or inactive. In HeLa cells, 64% of elements showed strong evidence of enhancer activity and in neuroblastoma cells, 72% showed strong evidence of activity. Across the two different cellular systems, 85% of all elements were active in at least cellular context and 50% were active in both. Elements overlapping a DNase I hypersensitive site (DHS) in HeLa cells showed 1.4-fold greater expression (p = 0.00003043, Wilcoxon rank sum test) and elements overlapping a DHS in neuroblastoma showed a 1.1-fold greater expression which was not significant (p = 0.24, Wilcoxon rank sum test) (Figure 5C,D). Only one source of DHS data was available for the neuroblastoma cell line used, and these DHS data were generated as part of an early ENCODE study, thus the sensitivity may not be equivalent to the data used in HeLa cells. Elements that are active in the MPRA experiment in both neuroblastoma and HeLa cells are also 1.4-fold more likely to be overlapping open chromatin peaks in both HeLa and Neuroblastoma (p = 0.0235, chi-square test). These results support the finding from Chapter 2 that a large fraction (50-70%) of the non-coding elements sequenced in the DDD project were likely acting as enhancers.

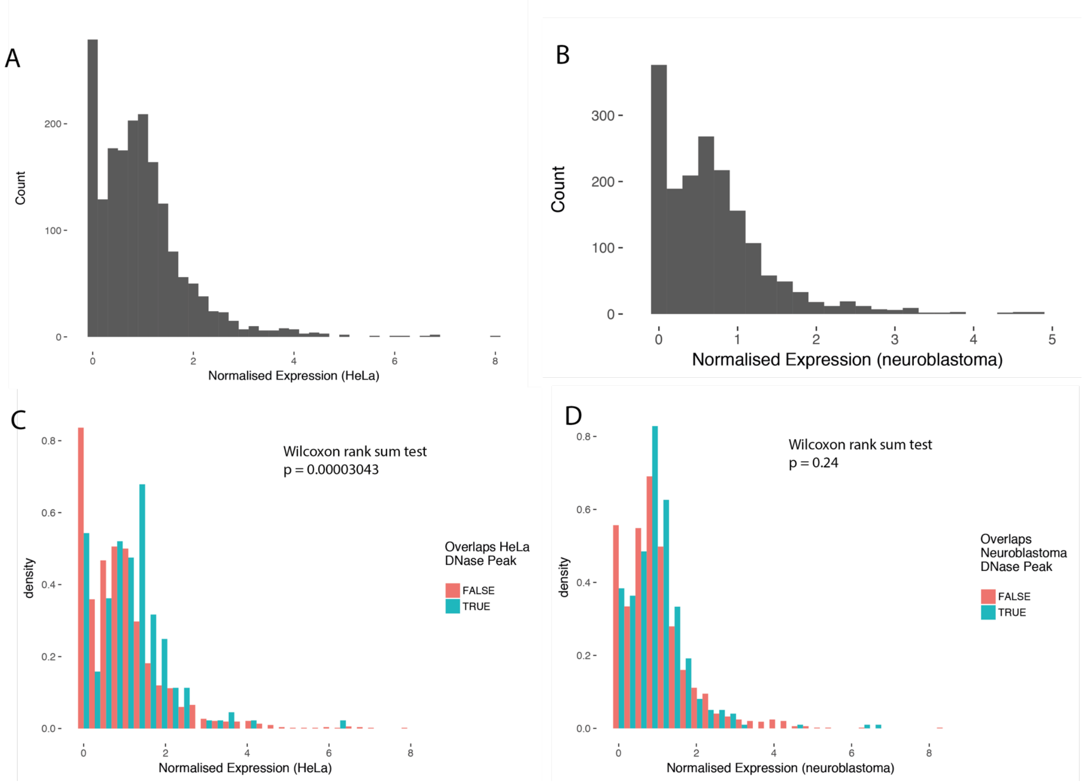


Figure 5 Activity of reference sequences in different cellular contexts. Ratio of RNA to DNA in the wildtype sequence of 767 conserved non-coding elements and enhancers in (a) HeLa cell line and (b) neuroblastoma cell line shows bimodality. (c) Elements overlapping an open chromatin peak in HeLa cells show higher expression in the MPRA experiment. (d) Elements overlapping an open chromatin

peak in Neuroblastoma cells do not show any statistically significant expression differences in the MPRA assay.

Results Section 4.1.2. Impact of SNVs and indels on ultra-conserved element function

Across all of the genetic variants tested, the majority of single nucleotide variants and indels resulted in decreased expression relative to the reference sequence (Figure 6), suggesting that the evolutionary conservation of these elements may be maintained in large part by selection against reduction in gene expression on the genes they regulate. Population genetic analyses in Chapter 3 and previous reports based on patterns of evolutionary conservation in the non-coding genome across species suggest that selective constraint on indels may be stronger than selective constraint on SNVs.

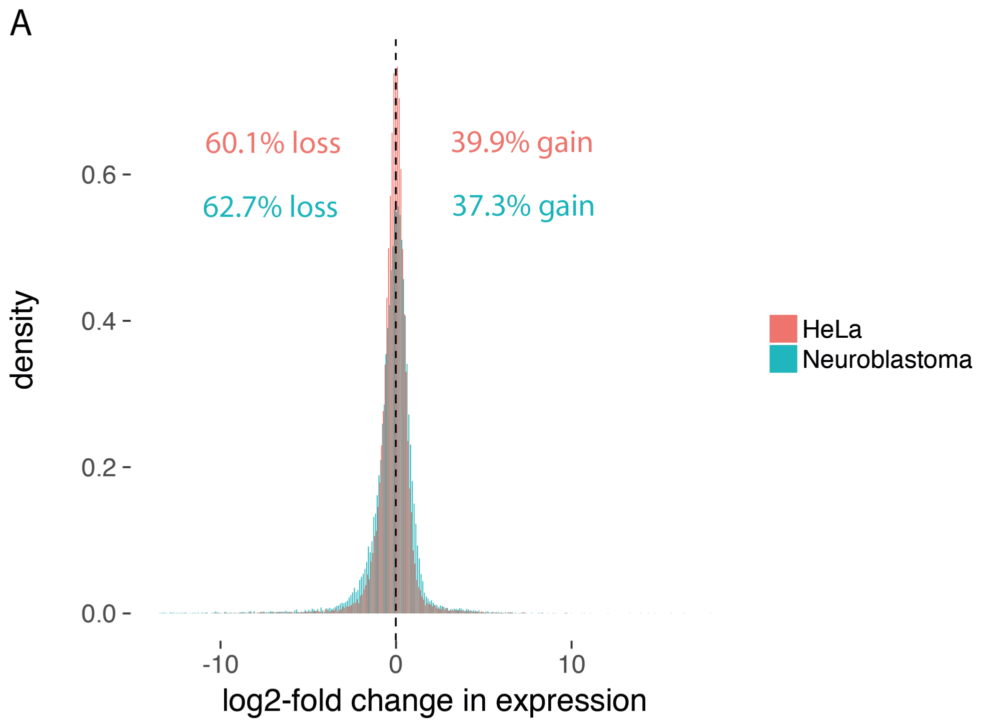


Figure 6 Impact of all SNVs on reporter gene expression. In both HeLa and neuroblastoma, approximately 60% of SNVs result in reduced expression compared to the reference sequence.

In both HeLa and neuroblastoma cells, the 5bp deletions showed substantially greater impact on expression of the wildtype sequence compared to SNVs (median 1.39-fold change from 5bp deletions compared to median 1.22-fold change for SNVs in HeLa cells p < 2.2e-16, Figure 7A, and median 1.42-fold change from 5bp deletion compared to median 1.34-fold change for SNVs in neuroblastoma cells p < 2.2e-16, Figure 7B).

While the majority of elements tested in this assay are highly evolutionarily conserved, not all nucleotides within these elements are conserved across species. I sought to test the relationship between nucleotide level evolutionary conservation across 100 vertebrate species (measured by PhyloP) and changes in reporter gene expression. Variation in sites with the highest decile of evolutionary conservation reduced expression by ~7% more than variation in sites with the lowest decile of evolutionary conservation (median 1.32-fold change versus 1.39-fold change, p = 0.000032, Figure 7C) in neuroblastoma, but the difference was not significant in HeLa. I did not find any significant difference in reporter gene expression and CADD score[25], one of the most commonly used metrics for assessing variant deleteriousness in the coding and non-coding genome (Figure 7D). The data in neuroblastoma are suggestive of a relationship between evolutionary conservation and magnitude of MPRA expression changes, but given the lack of concordance between different cell types, further study is warranted before drawing any broad conclusions.

Forces of selection acting to reduce the frequency of deleterious alleles in the population imply that rare variation is more likely to be deleterious than common variation. For example, singleton variants (observed only once in the population being studied) have been shown to be enriched for damaging variation compared to more common variation. To this end, I hypothesized that the allele frequency of variants may correlate with effect size in the reporter assay. I used the genome aggregation database (gnomAD) and the DDD unaffected parents to identify variants observed as singletons and compared their effect size to sites with two or more alleles observed. I did not find any significant difference between effect size the in the reporter assay for singleton compared to non-singleton variation in either HeLa or neuroblastoma. However, as only approximately 3% of the variants assessed in the MPRA assay have been observed in these populations, power to detect any difference is far more limited than with PhyloP, or CADD which can be applied to every tested variant.

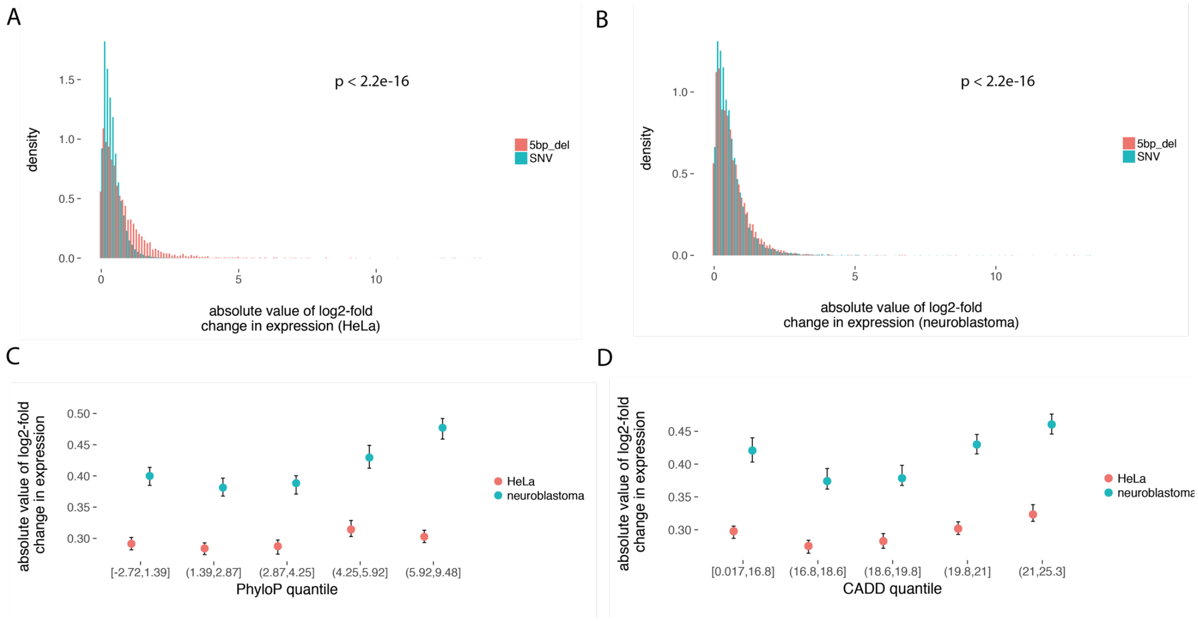


Figure 7 Impact of SNVs and deletions on reporter gene expression. Deletions result in significantly greater changes than SNVs in (a) HeLa cell line and (b) neuroblastoma cell line. (c) Increase in PhyloP score is associated with a greater impact on reporter gene expression. (d) CADD score of the variant did not show a clear relationship with impact on the reporter gene.

Results Section 4.1.3. Assessing the impact of patient mutations using MPRAs

A lack of understanding of the nucleotide-level 'grammar' of enhancers, promoters, and other regulatory sequence presents a challenge to variant interpretation in the non-coding genome. MPRAs present an opportunity to test the impact of thousands of potentially pathogenic non-coding variants in a single experiment. If these experimental assays prove robust and informative with respect to pathogenicity and faithfully recapitulate *in vivo* activity, they could serve as a 'look-up' table for variant pathogenicity, as well as a source of data to improve machine learning models for variant pathogenicity prediction.

In Chapter 2, I identified a set of 31 non-coding elements that were recurrently mutated in developmental disorder cases without a pathogenic variant in the coding regions. These 31 recurrently mutated elements harboured 64 distinct mutations which were selected for saturation mutagenesis in these MPRA experiment (see Methods). I sought to compare the impact of these patient-mutations to variation observed in the apparently healthy population. I reasoned that variation in these elements that is observed in a healthy population is likely benign or under weak selective constraint, whereas the impact of variation that had not been observed in patients, or in healthy individuals is unknown. For this analysis, I use any variation in the gnomAD database (15,796 individuals) as well as unaffected parents from the DDD project (13,192 individuals).

While median change in expression was higher for patient mutations than for sites with variation observed in unaffected individuals in neuroblastoma (1.39-fold versus 1.33-fold in neuroblastoma, Figure 8B), the difference was not statistically significant. Due to the high sample to sample variability described in Section 4.1.1, approximately half of the patient mutations and polymorphic variants had to be excluded from this analysis, and as a result this analysis is likely underpowered.

I also used QuASAR-MPRA, a statistical package for detecting allele specific expression (ASE) in MPRA data to test the patient mutations and polymorphic variation for evidence of significant ASE. Approximately 10% of the variants tested were nominally significant, but none of the variants survived multiple hypothesis test correction. QuASAR-MPRA uses a beta-binomial distribution and fits an overdispersion parameter to the data. QuASAR-MPRA has been applied successfully to MPRA data from eQTL fine-mapping studies where a large fraction of tests were expected to be negative[16]. However, in the case of the highly conserved non-coding elements tested here, where a large fraction of sites might be expected to alter gene expression, fitting the overdispersion parameter to the data may be overly conservative. Thus, modelling approaches that fit overdispersion on likely benign variation, potentially using nucleotide level conservation as a proxy as shown in Chapter 3, or modelling approaches that account for prior information about the fraction of sites expected to result in ASE may be more appropriate. This set of MPRA experiments has not shown any compelling evidence for dramatic expression changes resulting from patient mutations, but as improvements in the experimental workflow are made to ensure robust expression across experimental replicates, and larger numbers of robustly disease-associated mutations are discovered, these analyses should be re-visited.

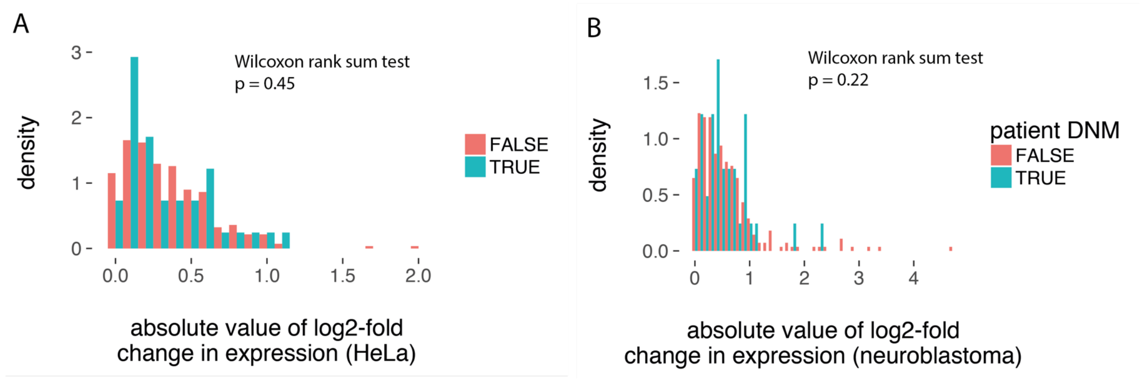


Figure 8 Assessing the impact of patient mutations in the MPRA assay. No significant difference in impact on expression of the reference sequence was observed between patient mutations and variation observed in gnomAD or the DDD unaffected parents in (a) HeLa cells and (b) neuroblastoma cells.

Results Section 4.2: Assessing disease-associated enhancer activity using mouse transgenesis assays

Results Section 4.2.1. Mouse transgenesis assays to identify putative developmental enhancers

While MPRAs have the capacity to test tens of thousands of variants in a given experiment, they are feasible only in cellular models with high transfection efficiency, and therefore may provide information only within a potentially limited cellular context. In contrast, mouse transgenesis assays are lower throughput, but provide an opportunity to test the function of an enhancer element across many tissues of a developing mouse embryo, albeit still upstream of a reporter gene.

In Chapter 1, I analysed sequence data from 7,390 trios where the child was affected with a severe developmental disorder and identified an enrichment of *de novo* mutations (DNMs) in highly evolutionary conserved non-coding elements (CNEs) with evidence for activity in the fetal brain. Thirty-one of the CNEs were recurrently mutated (DNM observed in two or more independent families). Under the null mutation rate model, we expected to see approximately fifteen recurrently mutated elements by chance. Thus, these thirty-one elements represent a source of candidate disease-associated elements with a false discovery rate of ~50%. I sought to test a subset of these candidate disease-associated elements in a mouse transgenesis assay in collaboration with Evgeny Kvon, Diane Dickel, and Len Pennachio at Lawrence Berkeley Labs.

To prioritise elements for testing, all thirty-one recurrently mutated elements identified in Chapter 1 were annotated with histone modifications and measures of open chromatin in the fetal brain and the developing mouse brain by Evgeny Kvon. In line with the analysis in Chapter 1 suggesting that the majority of the conserved non-coding elements we tested were acting as enhancers, 26/31 recurrently mutated elements overlapped an H3K27ac peak, a mark associated with enhancer activity, in either human fetal brain or mouse brain.

Based on evidence of activity in human/mouse brain and any testing in a mouse transgenesis assays, eleven of the thirty-one recurrently mutated elements were selected for testing. The wild-type sequence showed brain-specific expression at mouse developmental stage E11.5 in eight out of the eleven elements. As all eleven estimates had very strong evidence for H3K27ac and open chromatin in multiple developmental timepoint in mouse, there are a few potential explanations for the three negative elements. First, these elements may be functioning as enhancers *in vivo,* but may not be sufficient to drive expression on their own. Second, the mouse transgenesis assay was only completed at a single timepoint (E11.5) and these elements may be expressed at a different developmental timepoint. For example, one of the elements testing negative (chr10:131699490-131700091) has H3K27ac marks in the midbrain and hindbrain in E12.5, E13.5, and E14.5, but not E11.5. Ten of the eleven elements tested in the mouse transgenesis assay were tested in the MPRA and drove expression in either HeLa or Neuroblastoma (see Figure 9), including the three elements testing negative, supporting the hypothesis that these elements may be acting as enhancers in a different context other than that surveyed in the mouse transgenesis assay. The only element that tested negative was chr6:14501358-14501959, which had low representation in the pool likely due to its repetitive sequence (see Figure 3) and was excluded from further analyses.

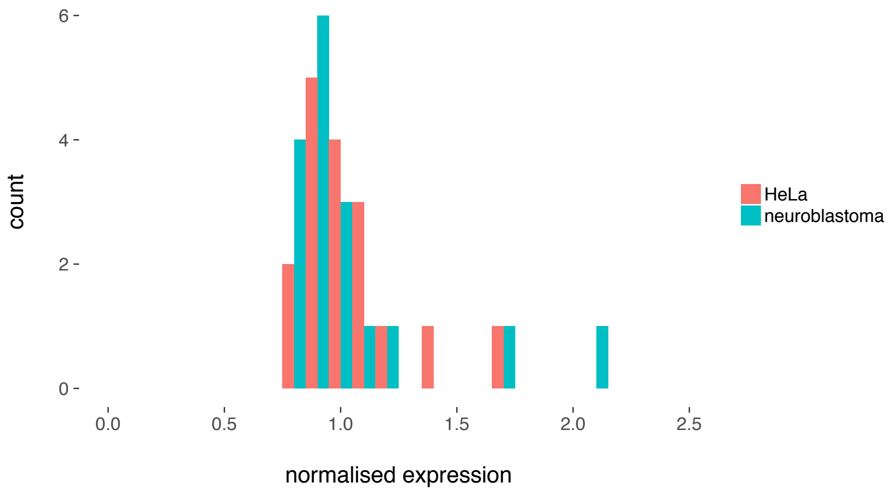


Figure 9 MPRA expression estimates for the elements tested in the mouse transgenesis assay. All of the elements that were tested in the mouse transgenesis assay showed evidence for enhancer activity in the MPRA assay in both HeLa and neuroblastoma.

The eight elements with reproducible enhancer activity in the mouse transgenesis assay were modified to include patient mutations. Three of the eight mutants substantially altered or completely ablated expression of the reporter gene. The wild-type sequence of chr14: 57553276-57553757 drove reporter gene expression in the hindbrain, but hindbrain expression was completely absent in the element containing the two patient mutations (see Table 1). This regulatory element is in a gene desert near *OTX2*, an important gene for brain and eye development expressed throughout the brain in early development. Multiple lines of evidence support a role for the non-coding element at chr14:57553276-57553757 acting as an enhancer to *OTX2*. Both of the mutations present in patients in chr14:57553276-57553757 resulted in loss of expression in the MPRA in both HeLa and neuroblastoma cells, although this difference was not significant after correcting for multiple tests. Taken together these results point toward loss of expression, perhaps exclusively in the hindbrain, of the dosage sensitive gene *OTX2* as a potential regulatory cause of severe developmental delay. Additional functional validation such as a mouse knock-in of the patient mutation is warranted to establish whether the regulatory variants observed here are indeed sufficient to cause a severe patient phenotype.

The wild-type sequence of chr6:14501358-14501959 drove expression in the midbrain and forebrain of the developing mouse embryo. After introducing the patient mutations, expression in both the forebrain and midbrain were greatly reduced. The wild-type sequence of chr7:13506147-13507336 drove expression in the forebrain, branchial arch, and trigeminal. The element containing patient mutations was active in the forebrain and branchial arch, but inactive in trigeminal. I was unable to assess concordance in the MPRA as the first element (chr6:14501358-14501959) had very low representation in the original pool due to its highly repetitive sequence (see Figure 3) and the second element (chr7:13506147-13507336) did not have sufficient observations to pass the threshold for correlation between biological replicates set in Section 4.1. As the MPRA experimental workflow is refined and more data is generated, these comparisons should be revisited.

Discussion

In this chapter, I presented preliminary data from two complementary experimental approaches to test the impact of non-coding variation on reporter gene expression. MPRAs have the potential to test tens of thousands of non-coding variants in a single experiment, but as evidenced from the data presented here, there are still experimental challenges to overcome to improve the utility of these assays. Using a strictly filtered set of elements, I find evidence for enhancer activity in a large fraction (>75%) of the developmental disorder-associated non-coding elements tested. I also find compelling evidence for 5bp deletions causing a greater change in reporter gene activity compared to SNVs. This is concordant with evidence from evolutionary studies[26], and from selective constraint in humans presented in Chapter 3, that indels are under stronger selective constraint in the non-coding genome. While MPRA assays represent a promising experimental technique for variant effect prediction in the non-coding genome, I was unable to draw any firm conclusions about the impact of patient mutations tested in this assay in large part due to low reproducibility between experimental replicates and a relatively small number of observations. As the correlation between independent replicates in published data is much higher, this is likely the result of experimental differences, possibly related to the the number of cells or the transfection method used.

Going forward, there are a number of important steps to optimise and adapt the MPRA pilot experiment conducted here. The number of cells used in this pilot experiment per replicate was much lower than that used in *Tewhey et. al*, and this likely resulted in a bottleneck due to the number of cells or the transfection method used. Adapting the method to add UMIs before each sequencing step also allowed us to determine that the amount of RNA sequenced was too low, as each input molecule was sequenced an average of three times. Increasing the number of cells used and the amount of genomic material sequenced will likely substantially improve the correlation between independent replicates and enable a larger fraction of the programmed variation to be reliably assessed.

While we have tested two different cellular models, HeLa and neuroblastoma, we have not tested whether genome integration methods provide more reliable estimates of gene expression than the episomal method employed here, as has been suggested previously[18]. Furthermore, we tested 150bp sequences created using oligo synthesis technology, but other approaches exist that allow for larger sequences to be tested including using PCR on patient samples, and the construction of larger elements out of synthesized oligos, albeit at lower throughput than the method employed here.

The current informatics pipeline does not incorporate any analysis of TFBS disruption. As published MPRAs have identified relationships between reporter assay expression and motif perturbation[11], this is an important next step toward understanding the mechanism underlying gene mis-regulation. As more developmental disorder associated elements are discovered, extending saturation mutagenesis to cover a greater number of elements beyond the 64 tested here will provide more power to discover nucleotide-level patterns associated with pathogenic mis-regulation.

In contrast to MPRAs, mouse transgenesis assays have been repeatedly validated as predictive of enhancer function, applied to thousands of different putative regulatory elements, and provide an organismal-level view of regulatory element function. However, these assays are much lower throughput in the number of putative regulatory elements they can assess and can only be cost-effectively applied to a small number of elements. For this reason, we selected eleven elements with strong prior evidence of enhancer activity in mouse and human brain, and with a high prior of association to severe DD. The majority (8/11) of the developmental disorder associated non-coding elements tested in the mouse transgenesis assay drove robust expression in at least one tissue and introducing patient mutations disrupted reporter gene expression in three out of eight of the enhancers.

One drawback of the two methods used here is that both MPRAs and mouse transgenesis assays rely on testing putative regulatory elements in a reporter construct, rather than in its endogenous locus. Ablation of expression in a reporter construct does not necessarily imply that expression will be ablated *in vivo*, as there may be compensatory effects from nearby elements[8,27]. Furthermore, mouse transgenesis assays do not provide a link from gene mis-regulation to the patient's phenotype. Thus, results from the mouse transgenesis assay should be corroborated by CRISPR knock-ins of patient mutations into the endogenous locus to establish the link between genotype and phenotype.

An additional challenge of assaying putative pathogenic variants in regulatory elements in MPRAs is that it is difficult to know based on sequence alone how a genetic variant will impact gene regulation. While variation resulting in loss-of-expression may be readily detectable in MPRAs, it is not clear how variation resulting in ectopic expression *in vivo* (e.g. in a different tissue/cell type) will manifest in a MPRA. Combining whole genome sequencing with RNA-sequencing in the relevant tissues for the patient's disorder may provide an opportunity to identify patients with abnormal expression levels attributable to a regulatory variant that can be tested in these systems. The advent of large whole genome-sequencing projects will likely uncover increasing numbers of disease-associated regulatory variants. For highly penetrant regulatory variants in particular, MPRAs, mouse transgenesis assays, and other functional assays will be critical to better understand the link between sequence and phenotypic effect.

1 Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

2 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

3 Dewet, J. R., Wood, K. V., Deluca, M., Helinski, D. R. & Subramani, S. Firefly Luciferase Gene - Structure and Expression in Mammalian-Cells. *Mol Cell Biol* **7**, 725-737, doi:Doi 10.1128/Mcb.7.2.725 (1987).

4 Chiocchetti, A., Tolosano, E., Hirsch, E., Silengo, L. & Altruda, F. Green fluorescent protein as a reporter of gene expression in transgenic mice. *Biochim Biophys Acta* **1352**, 193-202, doi:Doi 10.1016/S0167-4781(97)00010-9 (1997).

5 Kothary, R. *et al.* Inducible Expression of an Hsp68-Lacz Hybrid Gene in Transgenic Mice. *Development* **105**, 707-& (1989).

6 Kothary, R. *et al.* A Transgene Containing Lacz Inserted into the Dystonia Locus Is Expressed in Neural-Tube. *Nature* **335**, 435-437, doi:DOI 10.1038/335435a0 (1988).

7 Hill, R. E., Lettice, L. A. & Hill, R. E. Alterations to the remote control of Shh gene expression cause congenital abnormalities. *Philosophical transactions of the Royal Society of London*, doi:10.1098/rstb.2012.0357 (2013).

8 Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491-499, doi:10.1016/j.cell.2017.12.017 (2018).

9 Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346-1350, doi:10.1126/science.1159974 (2008).

10 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, 88-92, doi:10.1093/nar/gkl822 (2007).

11 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

12 Tewhey, R. *et al.* Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Resource Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529, doi:10.1016/j.cell.2016.04.027 (2016).

13 Arnold, C. D. Genome-wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **20**, R754-763, doi:10.1016/j.cub.2010.06.070 (2011).

14 Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159-164, doi:10.1016/j.ygeno.2015.06.005 (2015).

15 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 800-811, doi:10.1101/gr.144899.112 (2013).

16 Kalita, C. A. *et al.* QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* **34**, 787-794, doi:10.1093/bioinformatics/btx598 (2018).

17 Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology* **34**, 1180-1190, doi:10.1038/nbt.3678 (2016).

18 Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27**, 38-52, doi:10.1101/gr.212092.116 (2017).

19 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

20 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

21 Kovalevich, J. & Langford, D. Considerations for the use of SH-SY5Y neuroblastoma cells in neurobiology. *Methods Mol Biol* **1078**, 9-21, doi:10.1007/978-1-62703-640-5_2 (2013).

22 Maricque, B. B., Dougherty, J. D. & Cohen, B. A. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Research* 1-11, doi:10.1093/nar/gkw942 (2016).

23 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 72-74, doi:10.1038/nmeth.1778 (2011).

24 Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**, 163-166, doi:10.1038/nmeth.2772 (2014).

25 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

26 Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).

27 Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243, doi:10.1038/nature25461 (2018).

Chapter 5: Discussion

Gene regulation plays a central role in evolution, organismal development, and disease. Despite the critical importance of gene regulation throughout development, there have been few genetic variants in regulatory elements with large effects that have been robustly associated to disease (several notable examples are reviewed in Chapter 1). Even in common and complex disease where regulatory variation has estimated to account for more than 80% of disease risk attributable to genetics[1], there is still little mechanistic understanding of how this risk manifests. In this work, my overarching aim was to gain a better understanding of the contribution of genetic variation in regulatory elements to Mendelian disorders and attempted to approach this problem from three different perspectives. I first sought to assess the contribution of regulatory variation to severe developmental disorders using sequence data from 8,000 affected individuals and their parents and to identify individual elements with a high probability of harbouring pathogenic regulatory elements (Chapter 2). Next, I used data from more than 28,000 whole genome sequenced individuals to examine the forces of selection operating on non-coding elements more generally (Chapter 3). Finally, I conducted a pilot experiment to assay >50,000 different non-coding variants across more than 700 different non-coding elements, including variants observed in patients with developmental disorders in a massively parallel reporter assay (MPRA) and collaborated on an assessment of the impact of patient mutations in eleven different enhancers using mouse transgenesis assays (Chapter 4).

Analysing mutations in more than 8,000 trios with severe developmental disorders from the Deciphering Developmental Disorders (DDD) project[2,3], I showed that *de novo* single nucleotide variants (SNVs) in highly conserved non-coding elements contribute to these disorders. The DDD project sequenced approximately 5Mb of non-coding sequence in all 8,000 trios, but this sequence was biased to contain primarily highly evolutionarily conserved non-coding elements. As a result, the observations and estimates here are based primarily on highly conserved non-coding sequence, although this analysis did include more than 500 experimentally validated enhancers with varying levels of evolutionary conservation.

I estimated that 1-3% of patients in this cohort carry a pathogenic *de novo* SNV in a regulatory element active in the fetal brain. Approximately ~20-50% of these pathogenic mutations were sequenced directly in the capture, and serve as a conservative lower bound for estimates of burden in the non-coding genome. While regulatory variants have been previously associated with Mendelian disorders (reviewed in the Introduction), this is the first time, to my knowledge, that a role for damaging regulatory variation has been reported in such a large and phenotypically heterogeneous set of disorders.

It is important to note that this estimate does not include the contribution of small insertions and deletions (indels), which has not yet been assessed in the DDD cohort. The lack of robust models of the indel mutation rate presents a challenge for analyses of mutational burden. These variants play a large role in disrupting protein-coding sequence due primarily to frameshift mutations, and are also likely to be more damaging than SNVs in non-coding regions. Thus, there is a strong motivation for improved modelling of the germline mutation rate and variant calling pipelines to accurately detect this class of variation. The contribution of larger copy number variants (CNVs) in non-coding elements has also not yet been assessed on a large scale. As a number of Mendelian disorders are caused by enhancer duplications, deletions, and disrupting TAD boundaries,[4] this class of variation should be prioritised for future work within the DDD cohort and in large whole genome sequencing studies which have greater sensitivity for calling CNVs and identifying breakpoints[5].

Using data from large-scale annotations of function in the non-coding genome including the Roadmap Epigenome Project[6], the ENCODE project[7], the FANTOM5 consortium[8], and the VISTA enhancer database[9], I showed that the majority of conserved non-coding elements sequenced in the DDD patients are likely functioning as enhancers. A subset of conserved non-coding elements show little evidence of enhancer activity across a wide range of tissues, but are instead predicted to be involved in post-transcriptional regulation. Conserved non-coding elements (CNEs) have long captured the interest of the scientific community due to their near perfect sequence identity across hundreds of millions of years of evolution[10]. These elements have been shown to be under ongoing purifying selection and unlikely to be mutational coldspots[11]. I showed that conserved non-coding elements sequenced in the DDD project were under selective constraint, albeit to a lesser degree of known developmental disorder genes, consistent with these previous reports[11,12].

The distribution of *de novo* mutations (DNMs) *in* CNEs in Chapter 2 also suggests that within a disease-associated CNE, a small fraction of sites (maximum likelihood estimate of 3%) are pathogenic with a dominant mechanism for DD. In contrast, in DD-associated haploinsufficient disease genes, 8-10% of mutations result in protein truncation and many missense variants may also result in loss of function via other mechanisms[13]. Thus, even for the most highly evolutionarily conserved regulatory elements in the genome, the role these mutations play in disease and therefore the forces of selection acting upon them, differs substantially from protein-coding genes. The pattern of evolutionary conservation in these elements could be the result of pleiotropy, where a single ultraconserved regulatory element is responsible for multiple different important regulatory functions. While this model could explain the relatively low contribution to severe developmental disorders from mutations in these elements, this model would likely result in levels of selection on par or greater than protein-coding genes, which I did not observe. Furthermore, non-coding elements are typically active in fewer tissues than coding genes, making the pleiotropy argument less likely[14]. Alternatively, evolutionary conservation in these elements may be maintained by weak selection against heterozygosity, or potentially strong selection against homozygosity that prevents deleterious variation from becoming fixed in the population. I found suggestive evidence in Chapter 2 that patient mutations more frequently resulted in gain of transcription factor binding than expected under a null mutational model. Thus, I hypothesise that a large fraction of the deleterious non-coding variants in CNEs may result in gain of function, as has been observed previously[15,16]. The answer to these questions has important implications for the role these elements play in disease, the study designs required to detect them, and potential therapeutic strategies. Furthermore, these questions apply not only to CNEs, but to non-coding elements genome-wide, motivating a more comprehensive study of selective constraint in the non-coding genome.

In Chapter 3, I analysed data from more than 28,000 whole genome sequenced individuals from the gnomAD and BRIDGE projects to examine patterns of selective constraint in the non-coding genome in greater detail. I detected evidence for pervasive constraint against SNVs in the non-coding genome, most notably in DNase hypersensitive sites, long non-coding RNAs, and 3' UTRs. In each of these element sets, there was a strong correlation between evolutionary conservation of the elements and observed selective constraint. However, I found that even elements that are poorly evolutionarily conserved contain a subset of nucleotides that are evolutionary conserved which are under selective constraint. This result has important implications for the role of regulatory variation in disease. Namely, it suggests that a large fraction of active regulatory elements genome-wide may harbour variation conferring disease risk, albeit in a small fraction of the total sequence underlying the open chromatin peak or activity-associated histone mark.

If this model holds true, then it implies that efforts to understand the functional bases within a putative enhancer sequence are likely to greatly improve power to detect pathogenic variation. This also suggests that whole genome-sequencing, rather than targeted sequencing of a subset of selected regulatory elements, may be necessary to detect disease-associated variants in research or clinical applications where identifying pathogenic non-coding variation is a primary aim. Furthermore, it suggests that the conventional model of regulatory 'elements' that are hundreds or thousands of bases in length may actually be more accurately modelled as a loose collection of smaller functional sequences. As the *de novo* enrichment analyses presented in Chapter 2 focus on highly evolutionarily conserved elements, increasing genome coverage to include a greater number of less well conserved regulatory elements may provide further opportunity to answer this question. Large-scale whole genome sequencing efforts such as the Genomics England 100,000 Genomes project have recruited thousands of affected patients and their parents and these data will provide a great opportunity to assess this question and others.

Evolutionary estimates suggest that a small fraction of the genome, between 3% and 15%, is under selective constraint[17,18], while large-scale annotations of regulatory function such as the ENCODE project have suggested as much as 80% of the genome is biochemically active in at least one tissue[7]. The results presented in Chapter 3 are consistent with a model in which a large fraction of the genome may be biochemically active, but only a small fraction of functional nucleotides give rise to this activity. I have also shown that conserved transcription factor binding sites, regardless of the evolutionary conservation of the surrounding sequence, are under purifying selection. Taken together, these results lend support to the 'billboard model' of regulatory element function whereby transcription factor binding sites or sets of sites in close spatial proximity confer regulatory potential to previously inert sequences[19]. However, this analysis did not consider the potential impact of the variant (e.g. loss, gain, or neutral effect on binding), nor the impact of spacing between transcription factor binding sites. As this analysis incorporated data from approximately 28,000 healthy individuals, there is significant scope to increase this analysis using data from the UK Biobank and the 100,000 Genomes Project that is soon to be made publicly available. However, I estimated that even hundreds of thousands of deep whole genomes would not provide the resolution necessary to detect selection at 10-20bp resolution (e.g. the size of a typical transcription factor binding sites). Thus, understanding and predicting the precise patterns underlying nucleotide-level constraint in the non-coding genome represents an important area for future work that may not be resolved by population genetic approaches alone.

Evolutionary studies have suggested that regulatory elements are more intolerant of insertions and deletions (indels) than SNVs[18]. Modelling selective constraint against indels presents several challenges, most notably the lack of well-calibrated null models for the indel mutation rate. To overcome this challenge, I used *de novo* indels to directly account for variation in the underlying indel mutation rate. Results from this analysis suggest that indel constraint is substantially greater than SNV constraint in non-coding elements. However, the confidence intervals of these estimates are still large, primarily due to the relatively small number of *de novo* indels from healthy individuals used to calibrate the mutation rate model. In addition to simply collecting larger numbers of whole genome sequenced trios to more accurately quantify the indel mutation rate, a number of models are in development to predict the indel mutation rate from sequence context. Improvement of these models is critical to enable indel constraint to be quantified for individual elements, rather than aggregated groups of elements as was performed in Chapter 3.

As suggested by the analysis of selective constraint on the CNEs sequenced in the DDD project, regulatory elements show systematically lower selection on heterozygotes than protein-coding genes even at the same level of evolutionary conservation. This result, alongside the estimate that a low proportion of bases in regulatory elements are predicted to contribute to severe DD with a dominant mechanism (maximum likelihood estimate of 3%), suggests that the effect sizes of genetic variants in regulatory elements may, on average, be far smaller than protein-coding genes. Apparently weaker selection on heterozygotes in regulatory elements could also point to a greater role for selection on homozygotes maintaining evolutionary conservation. While there have been attempts to model selection on heterozygotes with an assumed dominance coefficient of one[20], there are, to my knowledge, no established methods for jointly modelling strength of selection and the dominance coefficient using human whole genome sequencing data. Recent population genetics work has suggested that it may be fundamentally difficult to separate weak selection against heterozygotes from selection against homozygotes using human polymorphism data[21]. More development in this area will improve modelling the genetic architecture of disease not only in the non-coding genome, but in protein-coding genes where incomplete penetrance or oligogenic models may play an important role[22].

Chapters 2 and 3 together provide insight into the patterns of selection shaping the non-coding genome and their role in disease. However, more in-depth functional assays incorporating mutations linked to severe disorders may provide a complementary view into the nature of gene mis-regulation and the ‘enhancer code’. Furthermore, these assays present an opportunity to test the impact of patient mutations on regulatory element function, and to validate observations from the population genetic analyses presented in Chapter 3.

For this MPRA pilot experiment, I designed a set of experiments testing more than 50,000 different putative enhancer sequences including all possible SNV changes in 64 elements, and 5bp deletions tiled across more than 700 elements. The 64 elements selected for saturation were identified as developmental disorder associated in Chapter 2. More than 75% of wild-type sequence tested showed evidence for enhancer activity in either HeLa or Neuroblastoma cells. I also found evidence for a relationship between nucleotide level conservation and magnitude of change in expression in the MPRA in neuroblastoma, but did not find any effect in HeLa cells. Deletions resulted in a larger effect in the MPRA in both HeLa and Neuroblastoma compared to SNVs, supporting the prior evolutionary work[18], and population genetic studies in Chapter 3 suggesting that indels may be more disruptive of regulatory element function, and therefore under stronger selective constraint.

While population genetic studies rely on data from only a small fraction of nucleotides where variation is observed, MPRAs provide the opportunity to test every nucleotide in a given regulatory sequence of interest. Thus, these tools may provide deeper insight into the nucleotide-level patterns responsible for regulatory element function. However, these experimental assays are still limited to cell types amenable to transfection in the laboratory and in the data from our pilot experiment, show relatively low correlation across biological replicates. As published data shows higher correlation across experimental replicates than we observed here[23], additional optimisation of laboratory protocols should yield higher quality data and allow us to make reliable inference on the impact of patient mutations. These experiments serve as a valuable pilot experiment to understand the power as well as limitations of this new experimental technique and provide a starting point for further experiments.

In addition to testing patient mutations in a series of MPRA experiments, I worked with collaborators at Lawrence Berkeley National Laboratory to test a subset of disease-associated elements in a mouse transgenesis assay. We selected eleven elements with two or more DNMs in patients for testing based on strong evidence of enhancer activity in mouse and human brain, and observed robust expression at specific brain regions at mouse developmental stage E11.5 for the wildtype sequence in 8/11 elements in the mouse transgenesis assay. Notably, I was able to test the wildtype sequence for 10/11 elements in the MPRA and found evidence for regulatory activity for all 10 in both HeLa and neuroblastoma, suggesting that the elements failing to drive expression in the mouse transgenesis assay are likely active enhancers in a different context than that tested in the mouse transgenesis assay. Three of the eight enhancers testing positive in the mouse transgenesis assay showed marked reductions in activity after introduction of patient mutations. While these assays present compelling evidence for reduction or ablation in expression in the context of a reporter gene, further work to knock-in the patient mutations to their endogenous context are needed to determine whether these gene expression changes are sufficient to cause a severe disorder. Work is now under-way in the Hurles lab to further characterise the phenotype of patients carrying these regulatory mutations and prioritise elements for mouse knock-in modelling based on the mouse transgenesis results.

Beyond evaluating the impact of patient mutations, high throughput functional assays have the potential to contribute to other longstanding problems in gene regulation. For example, predicting the target gene of a putative enhancer remains a substantial challenge that was evident in my analysis of *de novo* mutations in non-coding elements in severe developmental disorders in Chapter 2. I assessed the overlap in gene target predictions using four different approaches (Hi-C, DNase/RNA-seq correlation, evolutionary synteny, nearest gene) and found a low degree of overlap between the predictions from different methods. One challenge of using both Hi-C and predicting targets based on correlation between DNase I hypersensitivity and transcriptional output is the necessity to survey the precise tissue and timepoint of interaction. The sheer quantity of different cell types and developmental timepoints poses a fundamental challenge to observational approaches to gene target prediction. One possible solution is to use machine learning to predict gene targets in unseen tissues or timepoints by learning patterns from available data. For example, TargetFinder integrates 15 chromatin features across multiple cell types to predict gene targets[24]. For single-cell sequencing data, Cicero integrates sparse chromatin features within a single cell type to predict changes in gene expression[25]. Going forward, tools such as Hi-C or Capture-C may also be applied to directly to patient derived cell lines or tissue samples, allowing gene mis-regulation to be observed directly, rather than inferred based on gene target predictions produced from wild-type sequence.

Another approach to improve gene target prediction is to use CRISPR activation (CRISPRa) to force enhancer activity in a cell line representative of the tissue of interest and measure the impact on nearby genes[26]. In principle, this method can be used to direct the factors required for enhancer activation to the site of interest, and nearby genes can be profiled with qPCR or RNA-sequencing to test for changes in transcriptional output. However, it is still unclear how well CRISPRa recapitulates enhancer-gene interactions *in vivo*, with the exception of a small number of tested cell types[26,27]. CRISPR inactivation (CRISPRi) has also been used to selectively silence putative regulatory elements and measure the impact on the expression of nearby genes. This approach has been used to silence elements in specific loci[28] as well as genome-wide as a novel method for eQTL discovery[29].

Improvements in gene target prediction will improve power to detect disease-associated regulatory variation in several ways. Linking enhancers to the relevant gene will allow for more reliable assessment of the impact of regulatory mutations acting as a ‘second hit’ to a damaging protein-coding mutation. For patients with very distinctive disorders associated with well-characterised disease genes, but no pathogenic variant in the exome, regulatory variation is often cited as a potential source of missed diagnoses[22,30,31]. There have been a few examples of this phenomenon reported, but the prevalence cannot be reliably assessed without robust gene target maps in the relevant cell type or tissue. Analysis of transcriptomic data from patients with autism spectrum disorder and patients with cancer has suggested that cis-regulatory variation can modify risk of a coding variant by changing expression of the ‘risk haplotype’[22]. This analysis was performed using eQTLs to link regulatory variant to target genes, but in cases where gene expression is tightly controlled and depleted for eQTLs (e.g. near dosage sensitive genes[32]), establishing gene target prediction via alternative routes such as Capture-C, CRISPRa, or CRISPRi may improve power.

CNEs and enhancers are known to cluster in the genome, and these clusters often, but not always, regulate the same gene or genes[33]. Reliable gene target prediction will allow enhancers active in the same tissues and time points to be jointly analysed, improving power to discover disease associations. Genome editing in mice has already revealed examples of compensatory interactions between ultraconserved enhancers driving expression in the same brain region of a critical developmental gene[16,34]. As studies focused on individuals with Mendelian disorders begin to incorporate RNA sequencing, gene target predictions will be essential to link variation in enhancers to changes in transcriptional level. Regulatory variation that phenocopies a loss of function mutation in a gene, or decreases expression enough to cross a critical threshold, will be challenging to identify without reliable gene target predictions.

A sizeable fraction (>8%) of the possible mutations in protein-coding genes results in loss of function[13]. These variants can be aggregated together in gene burden tests, increasing power and helping to establishing clear hypotheses for disease mechanisms (e.g. haploinsufficiency leading to destabilisation of a protein-complex). Variation in the non-coding genome may prove to be more similar to missense variation in that many changes are neutral or weakly deleterious, and a small fraction or highly deleterious either acting as loss-of-functions or gain of function[13]. The MPRA experiment conducted in Chapter 4 supports this hypothesis—~62% of SNVs caused decrease in reporter gene expression in both HeLa and Neuroblastoma. Thus, assessing the impact of regulatory variation, particularly in a clinical context, where effect sizes are likely to be small and heterogeneous will require a greater understanding of the underlying 'enhancer code' or high-throughput assays to assess variant pathogenicity at scale[35]. Furthermore, there may be such a variety of risk variants or haplotypes across the population that observing any one variant or haplotype multiple times is rare. This has been the case for missense mutations in the *BRCA1* and *BRCA2* genes, for which a large proportion (>70%) of missense mutations are variants of unknown significance (VUS)[35,36]. Substantial progress has been made in separating benign from damaging missense variation using multiplexed assays of variant effect[35,36] (MAVEs). It has been suggested that these experiments could produce 'lookup tables' for clinical use, and the data can also be used to train machine learning models to better discriminate pathogenic and benign variants computationally[36]. In the non-coding genome, machine learning models have been trained using data from synthetic enhancers to predict enhancer function, but these models have not yet been applied to clinically ascertained non-coding variants[37]. MAVEs have been shown to outperform computational predictors of missense deleteriousness metrics, which are based primarily on evolutionary and physicochemical data[38-40], in some contexts[36]. Thus, it stands to reason that non-coding variant deleteriousness metrics may be improved by incorporating information from functional assays[35].

Extending the experimental strategies already successfully employed to study missense mutations in *BRCA1*[36] to the non-coding genome comes with several challenges. Assays of missense effect have been performed in haploid cell lines for which the gene of interest has been determined to be essential for survival. It may be more challenging to identify cell lines in which individual enhancers are essential for survival due to the high degree of enhancer redundancy. Even in the case that cell lines can be identified that are intolerant of deletion of an enhancer, individual genetic variants may result in change of gene expression, but not complete loss of function. Simultaneous advances in large scale genome sequencing projects in rare and common disease and high throughput experimental techniques will create tremendous opportunities for better understanding of the non-coding genome and the role of non-coding variation in disease. Going forward, approaches integrating regulatory mutations from patients, which have a higher prior probability of large effect, with high-throughput functional assays will improve our understanding of the principles of gene regulation and the mechanisms by which gene mis-regulation contributes to disease.

1 Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

2 Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet* **385**, 1305-1314, doi:10.1016/S0140-6736(14)61705-0 (2015).

3 Study, D. D. D. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228, doi:10.1038/nature14135 (2015).

4 Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics* **25**, 157-165, doi:10.1093/hmg/ddw205 (2016).

5 Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344, doi:10.1038/nature13394 (2014).

6 Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

7 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

8 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, doi:10.1038/nature12787 (2014).

9 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, 88-92, doi:10.1093/nar/gkl822 (2007).

10 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).

11 Drake, J. a. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics* **38**, 223-227, doi:10.1038/ng1710 (2006).

12 Derti, A., Roth, F. P., Church, G. M. & Wu, C.-t. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nature Genetics* **38**, 1216-1220, doi:10.1038/ng1888 (2006).

13 Kryukov, G. V., Pennacchio, L. a. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics* **80**, 727-739, doi:10.1086/513473 (2007).

14 Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152-163, doi:10.1038/s41559-017-0377-2 (2018).

15 Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**, 1725-1735, doi:10.1093/hmg/ddg180 (2003).

16 Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243, doi:10.1038/nature25461 (2018).

17 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

18 Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).

19 Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* **45**, 1021-1028, doi:10.1038/ng.2713 (2013).

20 Cassa, C. A. *et al.* Estimating the Selective Effect of Heterozygous Protein Truncating Variants from Human Exome Data. *Nature Genetics* **49**, 806-810, doi:10.1038/ng.3831 (2016).

21 Fuller, Z., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M., Measuring intolerance to mutation in human genettics. *biorXiv*, doi:10.1101/382481 (2018).

22 Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, doi:10.1038/s41588-018-0192-y (2018).

23 Tewhey, R. *et al.* Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Resource Direct Identification of

Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529, doi:10.1016/j.cell.2016.04.027 (2016).

24 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48**, 488-496, doi:10.1038/ng.3539 (2016).

25 Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*, doi:10.1016/j.molcel.2018.06.044 (2018).

26 Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology* **33**, doi:10.1038/nbt.3199 (2015).

27 Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111-115, doi:10.1038/nature23875 (2017).

28 Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773, doi:10.1126/science.aag2445 (2016).

29 Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet* **101**, 192-205, doi:10.1016/j.ajhg.2017.06.010 (2017).

30 Graf, S. *et al.* Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nature Communications* **9**, doi:10.1038/s41467-018-03672-4 (2018).

31 Alberobello, A. T. *et al.* An intronic SNP in the thyroid hormone receptor β gene is associated with pituitary cell-specific over-expression of a mutant thyroid hormone receptor β2 (R338W) in the index case of pituitary-selective resistance to thyroid hormone. *Journal of Translational Medicine* **9**, 144, doi:10.1186/1479-5876-9-144 (2011).

32 Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K., Measurement of selective constraint on human gene expression. *biorXiv* doi:10.1101/345801 (2018).

33 Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99, doi:10.1186/1471-2164-5-99 (2004).

34 Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172**, 491, doi:10.1016/j.cell.2017.12.017 (2018).

35 Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet* **101**, 315-325, doi:10.1016/j.ajhg.2017.07.014 (2017).

36 Findlay, G. M. *et al.* Accurate functional classification of thousands of BRCA1 variants with saturation genome editing. *bioRxiv*, doi:10.1101/294520 (2018).

37 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 800-811, doi:10.1101/gr.144899.112 (2013).

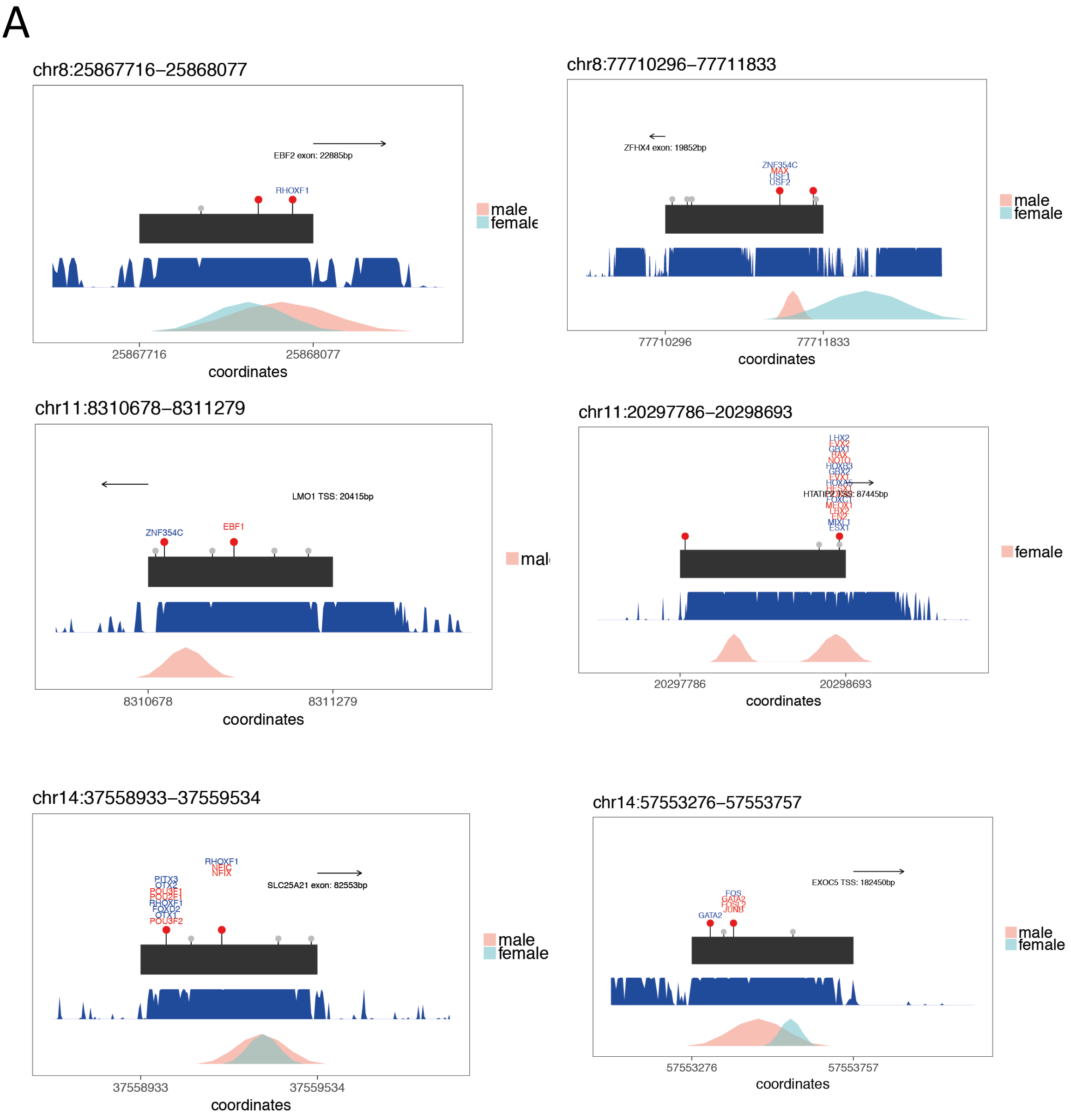
38 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

39 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nature Protocols* **11**, 1, doi:10.1038/nprot.2015.123 (2015).

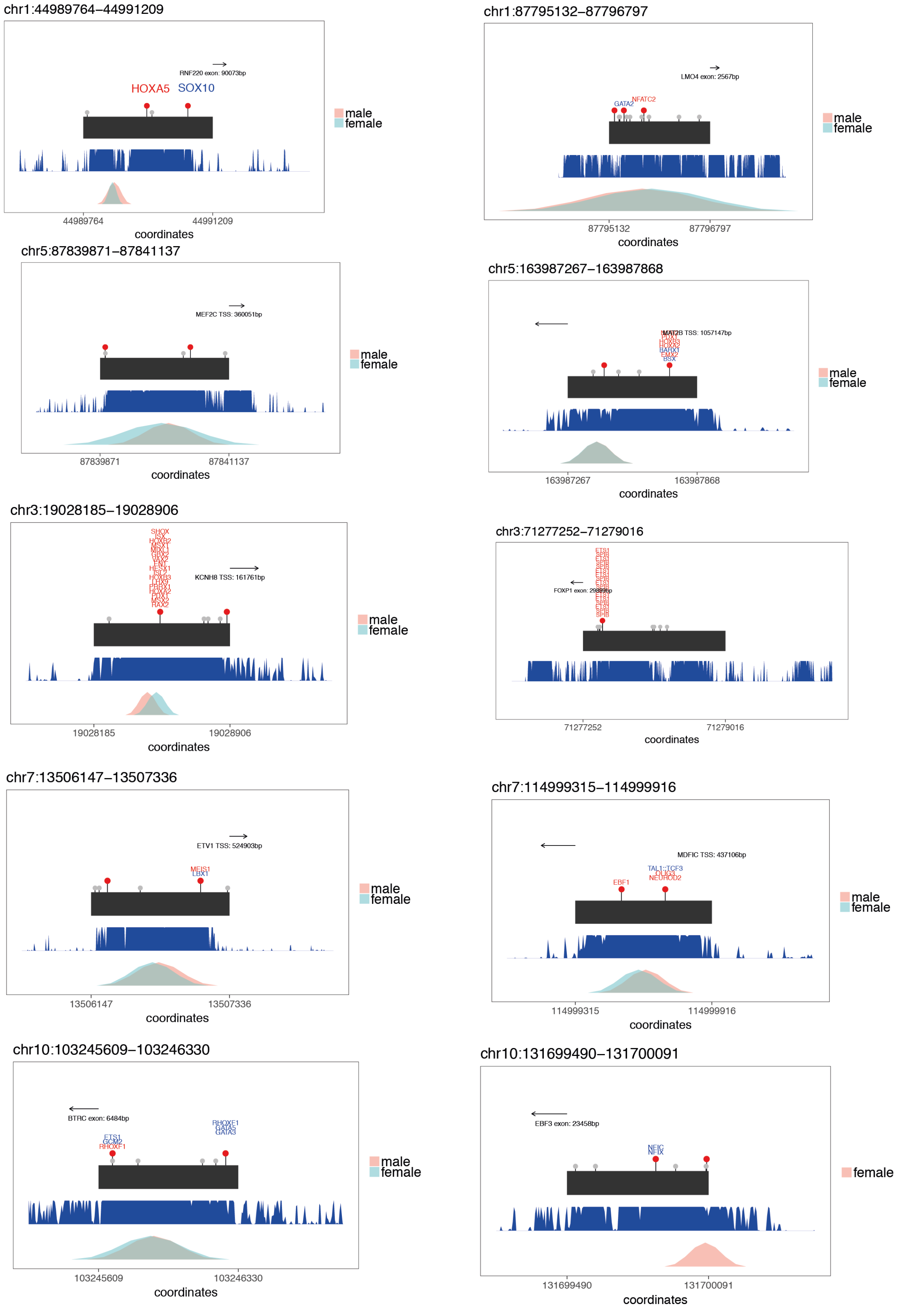
40 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

Appendix 1 – supplementary tables and figures for:

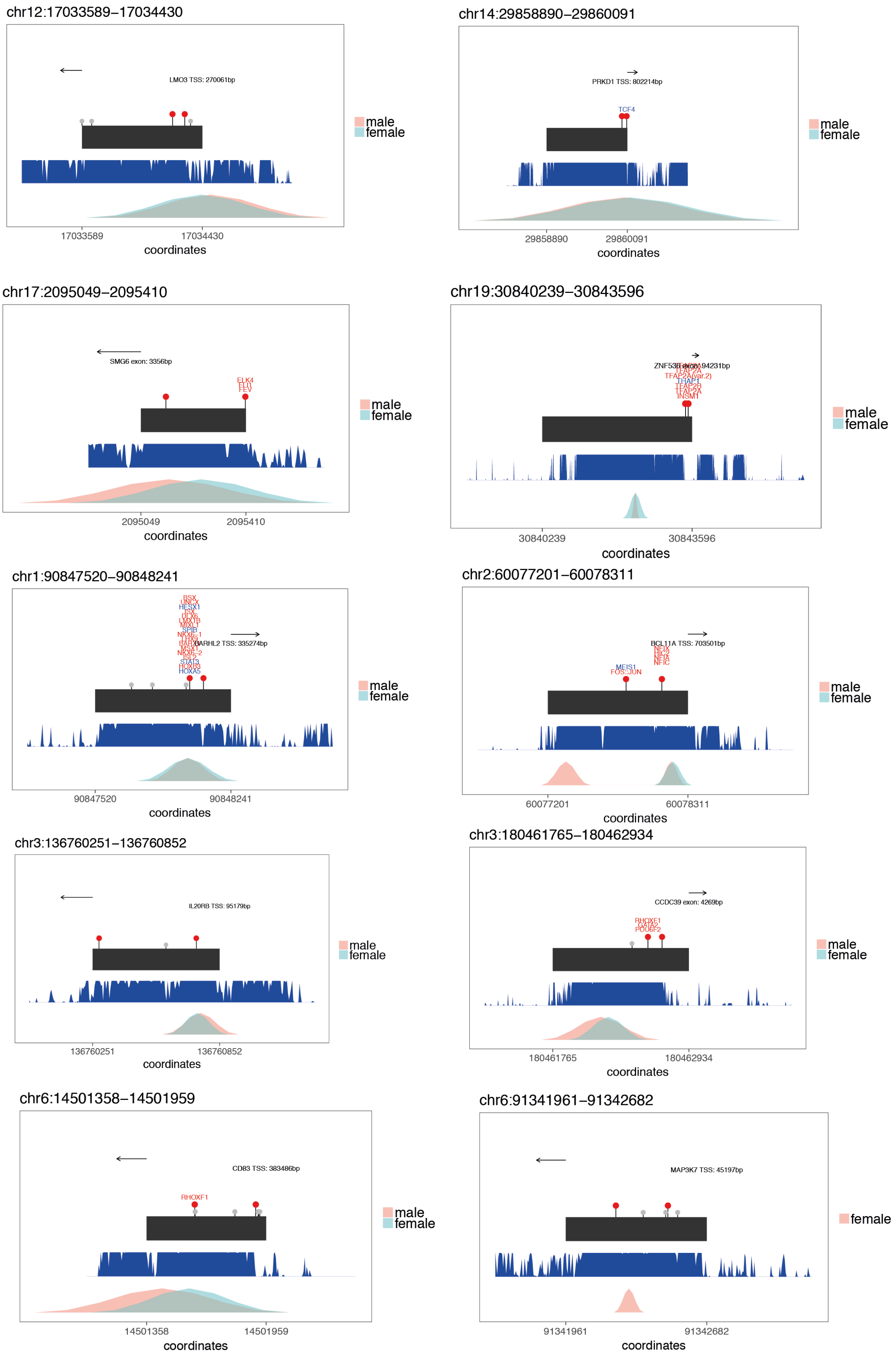
Chapter 2: *De novo* mutations in regulatory elements contribute to severe neurodevelopmental disorders



A (continued)


chr1:44989764–44991209
RNF220 exon: 90073bp
HOXA5
SOX10
44989764
44991209
coordinates
male
female
chr1:87795132–87796797
LMO4 exon: 2567bp
NFATC2
GATA2
87795132
87796797
coordinates
male
female
chr5:87839871–87841137
MEF2C TSS: 360051bp
87839871
87841137
coordinates
male
female
chr5:163987267–163987868
MAT2B TSS: 1057147bp
NKX2
PDX1
HOXB3
HOXA2
BARX1
EMX2
BSX
163987267
163987868
coordinates
male
female
chr3:19028185–19028906
SHOX
ISX
HOXB2
MSX1
MIXL1
GBX2
VAX2
ENT
HESX1
ISL2
HOXB3
LHX9
PRRX1
HOXA2
PDX1
MSX2
RAX2
KCNH8 TSS: 161761bp
19028185
19028906
coordinates
male
female
chr3:71277252–71279016
ETS1
SPIB
ETS1
SPIB
SPIB
ETS1
ETS1
SPIB
ETS1
SPIB
ETS1
ETS1
SPIB
ETS1
SPIB
SPIB
FOXP1 exon: 29899bp
71277252
71279016
coordinates
chr7:13506147–13507336
ETV1 TSS: 524903bp
MEIS1
LBX1
13506147
13507336
coordinates
male
female
chr7:114999315–114999916
MDFIC TSS: 437106bp
TAL1::TCF3
OLIG3
NEUROD2
EBF1
114999315
114999916
coordinates
male
female
chr10:103245609–103246330
BTRC exon: 6484bp
RHOXF1
GATA5
GATA3
ETS1
GCM2
RHOXF1
103245609
103246330
coordinates
male
female
chr10:131699490–131700091
EBF3 exon: 23458bp
NFIC
NFIX
131699490
131700091
coordinates
female

A (continued)


chr12:17033589–17034430
LMO3 TSS: 270061bp
male
female
17033589
17034430
coordinates
chr14:29858890–29860091
PRKD1 TSS: 802214bp
TCF4
male
female
29858890
29860091
coordinates
chr17:2095049–2095410
SMG6 exon: 3356bp
ELK4
ELF1
FEV
male
female
2095049
2095410
coordinates
chr19:30840239–30843596
ZNF536 exon: 94231bp
TFAP2A
TFAP2A
TFAP2A(var.2)
THAP1
TFAP2B
TFAP2A
INSM1
male
female
30840239
30843596
coordinates
chr1:90847520–90848241
BSX
UNCX
HESX1
ISX
DLX6
LMX1B
MIXL1
SPIB
NKX6–1
LHX9
BARX1
BARHL2 TSS: 335274bp
MSX1
NKX6–2
ISL2
STAT3
HOXB3
HOXA5
male
female
90847520
90848241
coordinates
chr2:60077201–60078311
BCL11A TSS: 703501bp
NFIX
HIC2
NFIA
NFIC
MEIS1
FOS::JUN
male
female
60077201
60078311
coordinates
chr3:136760251–136760852
IL20RB TSS: 95179bp
male
female
136760251
136760852
coordinates
chr3:180461765–180462934
CCDC39 exon: 4269bp
RHOXF1
GATA2
POU6F2
male
female
180461765
180462934
coordinates
chr6:14501358–14501959
CD83 TSS: 383486bp
RHOXF1
male
female
14501358
14501959
coordinates
chr6:91341961–91342682
MAP3K7 TSS: 45197bp
female
91341961
91342682
coordinates

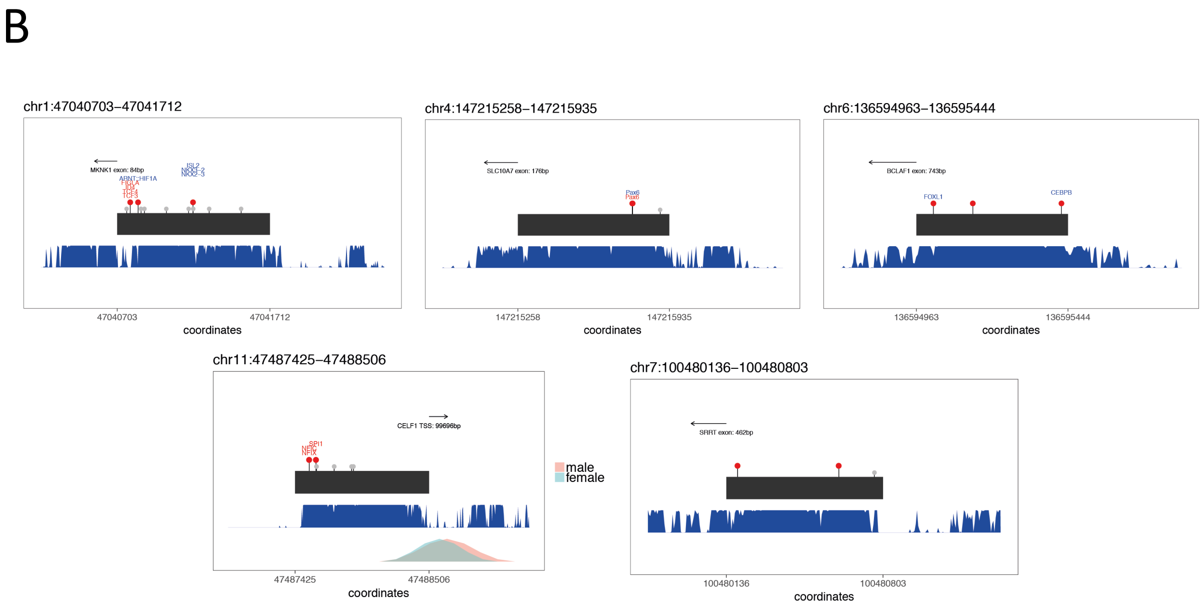


Figure S11 A) Recurrently mutated elements likely to be an enhancer. B) Recurrently mutated elements likely to be transcribed or involved in alternative splicing. The element is in black, red markers denote observed DNMs, grey markers denote observed variation at MAF > 0.1% in 7,080 unaffected parents, phastcons100 conservation score is shown in blue, and DNase hypersensitivity sites in fetal brain from the Roadmap Epigenome project are shown in blue (female) and pink (male) in the bottom track.

Region	Class	DNMs	p-value	Nearest Gene	Hi-C Gene Interactions
chr6:136594963-136595444	Conserved	3	5.30E-05	*BCLAF1*	*AHI1,MAP7*
chr1:47040703-47041712	Conserved	3	0.000640696	*MKNK1*	*NONE*
chr8:25867716-25868077	Conserved	2	0.001859799	*EBF2*	*EBF2,DOCK5*
chr17:2095049-2095410	Conserved	2	0.002057076	*SMG6*	*NONE*
chr14:57553276-57553757	Conserved	2	0.002134137	*EXOC5*	*NONE*
chr3:71277252-71279016	Enhancer	3	0.002356077	*FOXP1*	*NONE*
chr1:87795132-87796797	Enhancer	3	0.002754705	*LMO4*	*NONE*
chr7:114999315-114999916	Conserved	2	0.002818309	*MDFIC*	*FOXP2*
chr14:37558933-37559534	Conserved	2	0.0032701	*SLC25A21*	*TTC6*
chr3:136760251-136760852	Conserved	2	0.003280877	*IL20RB*	*MSL2,PPP2R3A,STAG1,CLDN18*
chr5:163987267-163987868	Conserved	2	0.003430287	*MAT2B*	*NONE*
chr4:147215258-147215935	Conserved	2	0.00352168	*SLC10A7*	*NONE*
chr6:14501358-14501959	Conserved	2	0.003823222	*CD83*	*NONE*
chr10:131699490-131700091	Conserved	2	0.004207684	*EBF3*	*NONE*
chr6:91341961-91342682	Conserved	2	0.004761199	*MAP3K7*	*NONE*
chr3:19028185-19028906	Conserved	2	0.00484094	*KCNH8*	*NONE*
chr11:8310678-8311279	Conserved	2	0.004900527	*LMO1*	*NONE*
chr1:90847520-90848241	Conserved	2	0.006338024	*BARHL2*	*NONE*
chr12:17033589-17034430	Conserved	2	0.007524004	*LMO3*	*NONE*
chr10:103245609-103246330	Conserved	2	0.008489455	*BTRC*	*LBX1*
chr7:100480136-100480803	Conserved	2	0.009844978	*SRRT*	*MUC17*
chr11:20297786-20298693	Conserved	2	0.009961525	*HTATIP2*	*NONE*
chr11:47487425-47488506	Conserved	2	0.010602029	*CELF1*	*MTCH2*
chr2:60077201-60078311	Conserved	2	0.012703131	*BCL11A*	*NONE*
chr7:13506147-13507336	Enhancer	2	0.013290101	*ETV1*	*NONE*
chr3:180461765-180462934	Enhancer	2	0.014462824	*CCDC39*	*NONE*
chr5:87839871-87841137	Conserved	2	0.014749778	*MEF2C*	*NONE*
chr14:29858890-29860091	Conserved	2	0.016188696	*PRKD1*	*FOXG1,C14orf23*
chr8:77710296-77711833	Conserved	2	0.020547115	*ZFHX4*	*NONE*
chr1:44989764-44991209	Enhancer	2	0.026516706	*RNF220*	*NONE*
chr19:30840239-30843596	Enhancer	2	0.090055315	*ZNF536*	*ZNF536*

Table 1

Genomic coordinates of the 31 recurrently mutated fetal brain active CNEs and conserved enhancers. Annotated with the nearest gene and any gene interaction reported by Hi-C data in the fetal brain. P-value is the probability of observing at least as many as the reported number of DNMs in the 6,239 exome-negative probands under the null mutation model.