Exploring non-coding variation in human diseases and disorders through targeted sequencing and functional prediction.

Lilian Elizabeth Hunt

University College London

and

The Francis Crick Institute

PhD Supervisor: Dr Greg Elgar

A thesis submitted for the degree of

Doctor of Philosophy

University College London

October 2017

Declaration

I, Lilian Hunt, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The identification of non-coding single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels) that are causative or contributory to human diseases and disorders is limited by the functional knowledge of the non-coding genome. This work demonstrates multiple approaches to elucidate functional variation in the non-coding genome by using homogenous populations or pedigrees of individuals with shared diseases and disorders, including Obesity, Schizophrenia, Anosmia and Mitochondrial Depletion Syndrome. A vast bank of non-coding variation has been created and can be utilised for population analysis. Using supporting evidence of developmental contributions to the disorders studied and genome interaction data, high coverage sequencing of targeted regions and subsequent bioinformatics analysis suggests multiple new disease-associated non-coding variants. Combining available variant function predictor tools and publicly available functional data, a selection of variants are prioritised as potentially causative or contributory and their affect on the region's function in development as an enhancer is assessed in Zebrafish. In addition, deep-sequencing and bioinformatics analysis in mouse models of MPV17 deletion contributes to the understanding of mitochondrial depletion syndrome.

Impact Statement

This work provides a significant contribution to the field of human genomics, specifically non-exonic sequencing and the variation that lies therein. The importance of exploring and understanding variation in non-exonic regions is paramount to understanding human diseases and disorders, specifically those resulting from embryonic development. Utilising genomic sequencing, as presented here, not only increases our understanding of general population variation, but also allows us to look for pathogenic mutations. The methodology of targeted sequencing and the variant calling and analysis pipelines could be applied in future in other instances of human developmental disease with a hypothesised non-coding variant cause. This top-down approach could also provide further insight to decoding the non-coding genome by continuing the work presented here, collating other functional non-coding mutations, and exploring the immediate sequence surrounding them. By using conserved non-coding elements as a base for much of this work, the functional prediction of the regions sequenced is already biased in a positive way. With further understanding of other regulatory element markers, other predicted enhancer regions could be sequenced in a similar manner.

All chapters in this work contribute to the understanding of a variety of human diseases: Isolated Congenital Anosmia, Schizophrenia, Obesity and Mitochondrial Depletion Syndrome. Understanding the underlying genetic contributions to human diseases and disorders can spark lines of investigation into treatments (such as pharmacological interventions) and preventions (including genetic counselling of prospective parents). With such a vast understanding of coding mutations and their contribution to diseases, this work looks to shift the future focus onto the other 98% of the human genome, a vast expanse of information that is yet to be fully decoded.

Acknowledgement

I'd like to acknowledge my supervisor Greg Elgar, for both supporting me and giving me the freedom to carry out this work. I've had a whale of a time. I'd also like to acknowledge Jim Smith, not only as a member of my thesis committee but as a mentor and ally. I'd like to thank Michael Simpson and Pete Scambler for their invaluable contributions as part of my thesis committee. I'd also like to acknowledge and thank Boris Noyvert and his irreplaceable bioinformatics help, as well as all the members of the Elgar lab, past and present, who I have worked with during this PhD: Joe Grice, Johanna Fischer, Stefaan Pauls, Laura Doglio, Htoo Wai, Maria Greco, Bathilde Ambroise and Chloe Moss. I'd also like to thank all of the Gilchrist lab, with whom we have shared space, time, drinks and many excellent ideas. All of this work and my survival as a student in London would not have been possible without the funding of the Medical Research Council and the support of The Francis Crick Institute and former National Institute for Medical Research.

I could not have done any of this, especially finish this thesis, without the incredible support of my fiancée Holly. Thank you.

On a personal note, there are many others who I owe this work and my PhD experience to: Rachel, Simon, Naomi, Jed, Derek, Rita, Theresa, Pete, Sissy, Sam, Alex, Rick and Morty.

I'd like to thank everyone else who has, in any way, helped me down this rabbit hole.

Table of Contents

Table of figures

List of tables

Abbreviations

1KG	1,000 Genomes Project
3C	Chromosome Conformation Capture
3C	Chromosome Conformation Capture
4C	Circularised Chromosome Conformation Capture
4C-seq	Circularised Chromosome Conformation Capture-Sequencing
A, C, G, T	Adenine, Cytosine, Guanine, Thymine
AF	Allele Frequency
AMP	Ampicillin
ANOS	Anosmia cohort
ANOVA	Analysis of Variance
BEB	Bengali from Bangladesh
BLAST	Basic Local Alignment Search Tool
BLAT	Basic Local Alignment Tool
BMI	Body Mass Index
BWA	Burrows-Wheeler Aligner software
CADD	Combined Annotation Dependent Depletion
Cas9	CRISPR associated protein 9
CASAVA	De-multiplexing software made available by Illumina
CEBPA	CCAAT/Enhancer Binding Protein Alpha
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
ChIA-PET	Chromatin Interaction Analysis with Paired-End Tag
ChIP-Seq	Chromatin ImmunoPrecipitation-Sequencing
CL±P	Cleft lip with or without cleft palate
CLP	Cleft Lip and Palate cohort
CNE	Conserved Non-coding Element
CNG	Centre National de Genotypage, Evry, France
CNV	Copy Number Variation
CONDOR	Conserved Non-coDing Orthologous Regions database
CTCF	CCCTC-binding factor
d.f	degrees of freedom
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate

dGTP	Deoxyguanosine triphosphate
DHS	Dnase 1 Hypersensitive Site
DNA	Deoxyribonucleic acid
DNase	DeoxyriboNuclease
dNTP	Deoxyribonucleoside triphosphate
dTTP	Deoxythymidine triphosphate
E.coli	*Escherichia coli*
EAF	European Allele Frequency
ENA	European Nucleotide Archive
ENCODE	ENCyclopaedia Of DNA Elements
eQTL	Expression quantitative trait loci
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FASTQ	Text-based file format comprised of FASTA sequence and quality score encoded with an ASCII character
FGF	Fibroblast Growth Factor
FGFR	Fibroblast Growth Factor Receptor
FIN	Finnish in Finland
FOXD1	Forkhead Box D1 gene
FOXO4	Forkhead Box O4
FTO	Fat mass and obesity-associated protein also known as alpha-ketoglutarate-dependent dioxygenase FTO
FWER	Family-Wise Error Rate
GAF	Global Allele Frequency
GATK	Genome Analysis ToolKit
GBR	British in England and Scotland
GERP	Genomic Evolutionary Rate Profiling
GFP	Green Fluorescent Protein
GIANT	The Genetic Investigation of ANthropometric Traits consortium
GIH	Guajarati Indian from Texas subpopulation of 1000 Genomes
GIS	Genome Institute of Singapore
GO	Gene Ontology
GOYA	Genome-Wide Population-Based Association Study of Extremely Overweight Young Adults

GWAS	Genome Wide Association Studies
H1-ESC	H1 human embryonic stem cell
HapMap	Haplotype Map
hESC	Human Embryonic Stem Cells
Hi-C	Hi-throughput chromosome confirmation capture sequencing
HMR	Human-Mouse-Rat alignment
HOXD9	Homeobox D9
hpf	Hours post fertilisation
HWE	Hardy-Weinberg Equilibrium
ICA	Isolated Congenital Anosmia
ICH	Isolated Congenital Hyposmia
IDE	Intellectual Disability and Epilepsy cohort
InDel	Short Insertion or Deletion variant
IRX1	Iroquois Homeobox 1
IRX2	Iroquois Homeobox 2
IRX3	Iroquois Homeobox 3
IRX4	Iroquois Homeobox 4
IRX5	Iroquois Homeobox 5
IRX6	Iroquois Homeobox 6
IRXA	Iroquois A gene cluster
IRXB	Iroquois B gene cluster
ITU	Indian Telugu from the UK
kb	Kilobase
KO	Knockout
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MAQ	Mapping and Assembly with Quality software
MDS	Mitochondrial Depletion Syndrome
ML	Mutation Load
MPV17	Mitochondrial Inner Membrane Protein Gene MPV17
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
MZ	MonoZygotic

NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
NOG	Noggin
NR2F1	Nuclear Receptor Subfamily 2 Group F Member 1
OMIM	Online Mendelian Inheritance in Man
PAKI	Pakistani Populations
PANSS	Positive And Negative Syndrome Scale
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PHRED	PHRED-scaled scores
PJL	Punjabi from Lahore, Pakistan
POU3F2	POU Class 3 Homeobox 2
PTU	1-phenyl 2-thiourea
RMAP	Read Mapping software
RNA	Ribonucleic acid
RNAPII	RNA Polymerase II
rNMP	Ribonucleoside Monophosphates
RPGRIP1L	RPGR-Interacting Protein 1-Like Protein
SAS	South Asian
SCHZ	Schizophrenia cohort
SCN5A	Sodium voltage-gated ChaNnel type 5 Alpha subunit
SEM	Standard Error from the Mean
SIFT	tool to Sort Intolerant From Tolerant amino acid substitutions
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SPEC	Spectinomycin
STU	Sri Lankan Tamil from the UK
T2D	Type 2 Diabetes
TAD	Topologically Associating Domain
Taq Pol	*Taq* polymerase
TF	Transcription Factor
TFAP2A	Transcription Factor AP-2 Alpha

TFBS	Transcription factor binding site
Tol2	Tol2 transposon element
TRANSFAC	TRANScription FACtor database
TRAP	TRanscription factor Affinity Prediction
UCSC	University of California Santa Cruz
UK10K	United Kingdom 10,000 Genomes Project
UNCX	UNC Homeobox
VAST	VAST BioImager system
VCF	Variant Call Format file
VEP	Variant Effect Predictor
VISTA	Vista Enhancer Browser
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
wig	Wiggle
WT	Wild-type
ZPA	Zone of Polarising Activity
ZRS	Zone of polarising activity Regulatory Sequence

Chapter 1. Introduction

1.1 Human genome sequencing

Since the Human Genome Project was completed in 2004 (International Human Genome Sequencing Consortium, 2004), remarkable progress has been made increasing the speed and efficiency, and decreasing the cost, of whole genome sequencing. We are now in a position where the field of single-cell genomics is expanding allowing a new perspective in our understanding of genetics to the cellular level (Gawad et al., 2016). Genetic sequencing has potential to add understanding to a wide variety of fields including evolutionary studies (Jones et al., 2012) and clinical diagnostics (Kingsmore and Saunders, 2011).

1.1.1 Sequencing technology developments

As the first human genome was sequenced using traditional automated Sanger sequencing techniques (Sanger et al., 1977), the National Human Genome Research Institute (NHGRI) set a goal of reducing the cost of human genome sequencing to $1000 within 10 years, funding the programme to do so itself (Collins et al., 2003). This led to the development of Next-Generation Sequencing (NGS) from multiple companies (Figure 1). The current volume of output data allows for good whole genome coverage, with 30-40x achievable for $1000. For the first human genome sequenced using Illumina short-read technology, a threshold of 15x average depth was able to detect homozygous single nucleotide variation, however 33x average depth was required to detect the same heterozygous variants (Bentley et al., 2008). Therefore a standard of 30x coverage for whole genome sequencing was quickly assumed (Ahn et al., 2009, Wang et al., 2008). This was increased to 50x in 2011 (Ajay et al., 2011) before improvements in sequencing technology decreased GC bias, delivering a more even coverage of the genome and suggesting a 35x threshold (Kozarewa et al., 2009). Uniformity of coverage, as well as depth, was shown to be essential for whole genome sequencing to identify population and individual specific variants (Sims et al., 2014). Nevertheless, all NGS sequencing platforms produce their own unique sequencing errors and biases that need identification and correction. Image analysis and cluster

amplification errors can occur at up to 1% frequency (Fox et al., 2014). Further downstream mapping errors can also occur from high frequency InDel polymorphisms, homopolymeric regions, GC- or AT-rich regions, replication bias and substitution errors (Bragg et al., 2013, Gilles et al., 2011, Huse et al., 2007).

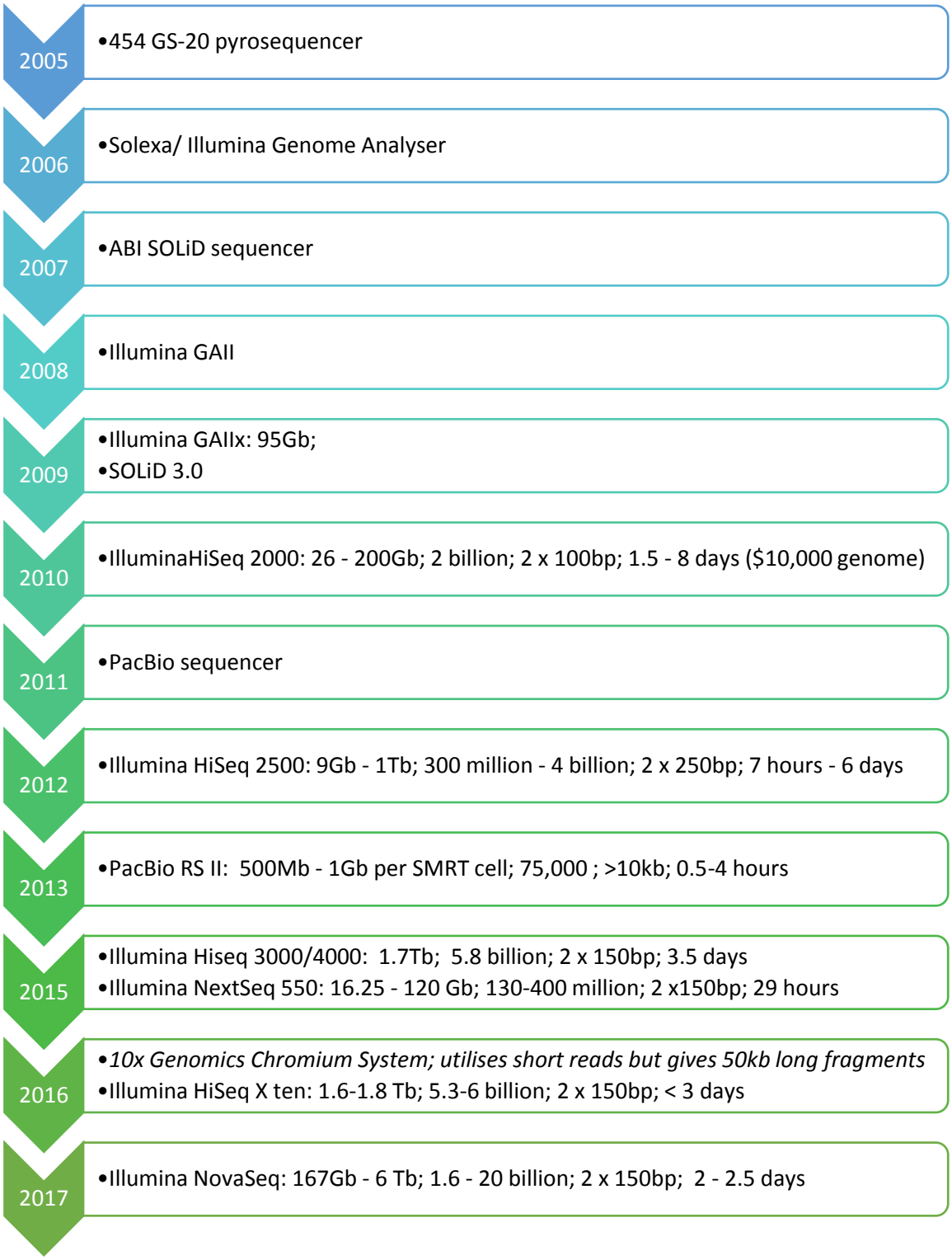


Figure 1. Timeline of key advances in sequencing platform technology. Information for each platform refers to Output range; Reads per run; Max read length; Run Time. The Illumina Hiseq X Ten system can now give 30x coverage for a whole genome for less than $1000.

In addition to the progress in whole genome sequencing, whole exome sequencing was a key target of development as reducing the size of the genome to be sequenced (~3 billion bases to 50Mb) reduces cost and memory whilst strongly enriching coverage of the exonic regions. In addition, with disease causing variants understood better in the coding regions of the genome, whole exome sequencing can be efficiently utilised in patients to drive diagnosis and medical interventions. Exons however, are GC rich (Amit et al., 2012) and the methods utilised for whole exome sequencing are less likely than whole genome sequencing to provide complete coverage of the entire coding region of the genome (Meienberg et al., 2015). With current PCR-free whole genome sequencing developments and drastic reduction in costs, whole genome sequencing is able to give better complete coverage of the coding region of the genome (Meienberg et al., 2016), and therefore clinical whole genome sequencing with downstream analysis focussing on the exome is becoming the norm (Berg et al., 2011). The underlying question of the importance of variation in the non-coding genome is, to a degree, ignored in the clinical setting. Vast amounts of sequencing data are available, yet cast aside by this process of whole genome sequencing and then exome diagnostics.

All NGS workflows utilise similar library preparation principles: DNA is fragmented, either by an enzymayic or shearing process, and these fragments are fused with platform-specific indexed adaptors. This allows multiple samples to be sequenced in one solution thanks to ‘barcode’ tag sequences on the adaptor that can be read by the sequencing platform and separated out in later computational steps. Size selection of the DNA fragments is crucial, and often PCR amplification is also utilised as a way of keeping fragments with adaptors successfully hybridised at both ends. For targeted sequencing e.g. exome, probes matching the sequence of the fragments being retained are used to pull down these fragments (often utilising biotin/streptavidin chemistry and magnetic beads).

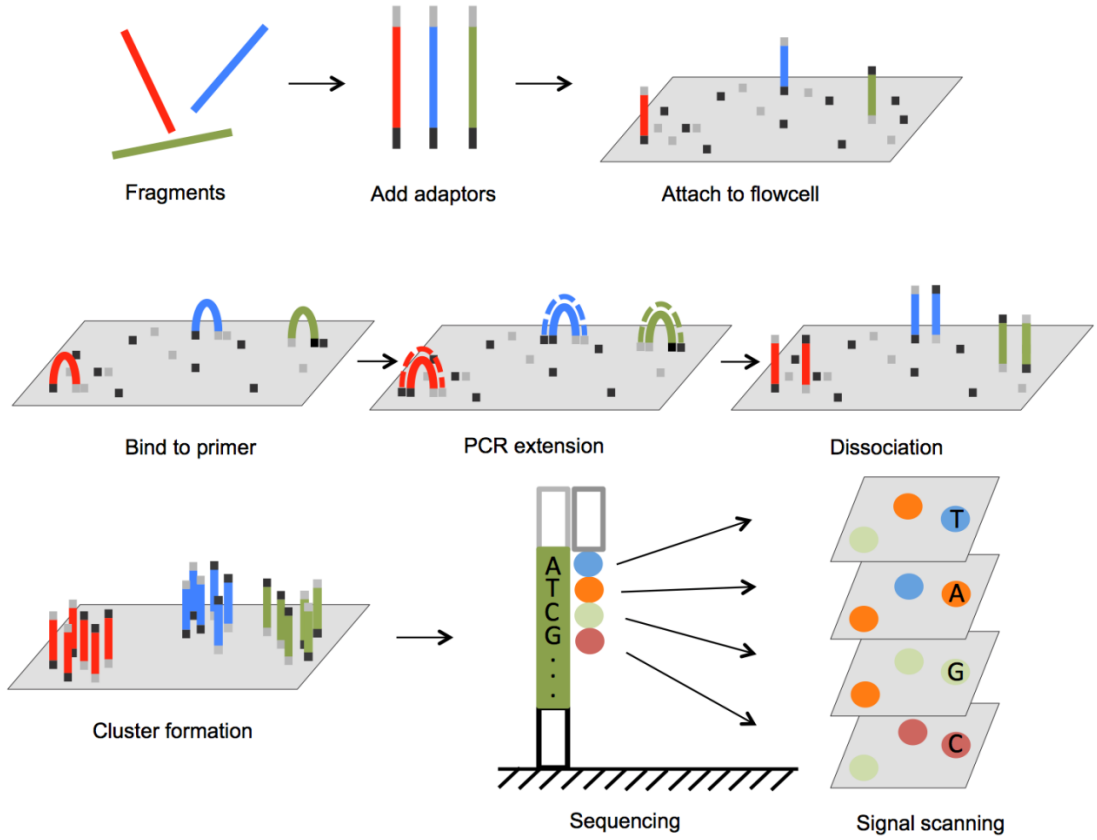


Figure 2. NGS sequencing as performed on Illumina platforms.
Figure adapted from Lu *et al*. 2016 (Lu et al., 2016). Adaptors are ligated to the ends of fragments as part of the library preparations steps. These will bind to the primer-loaded flow cell and bridge PCR then amplifies each fragment into a cluster of fragments with fluorophore attached nucleotides. Using a laser to excite the flourophores and an optic scanner to collect the signals, multiple fragments are sequenced simultaneously.

1.1.2 Bioinformatics developments

The development and improvements in NGS chemistry and sequencing platforms has only been made possible by equal advances in data storage, handling and analysis. The vast amounts of sequencing data have made demand for bioinformatics tools that can keep up with the accelerated rate of whole genome sequencing pivotal to the genomic revolution. The short reads resulting from NGS technology have resulted in new algorithms being needed for the mapping of these reads and the construction of individual whole genome sequence data (Hatem et al., 2013), as well as algorithms to overcome misread bases and correctly call variants.

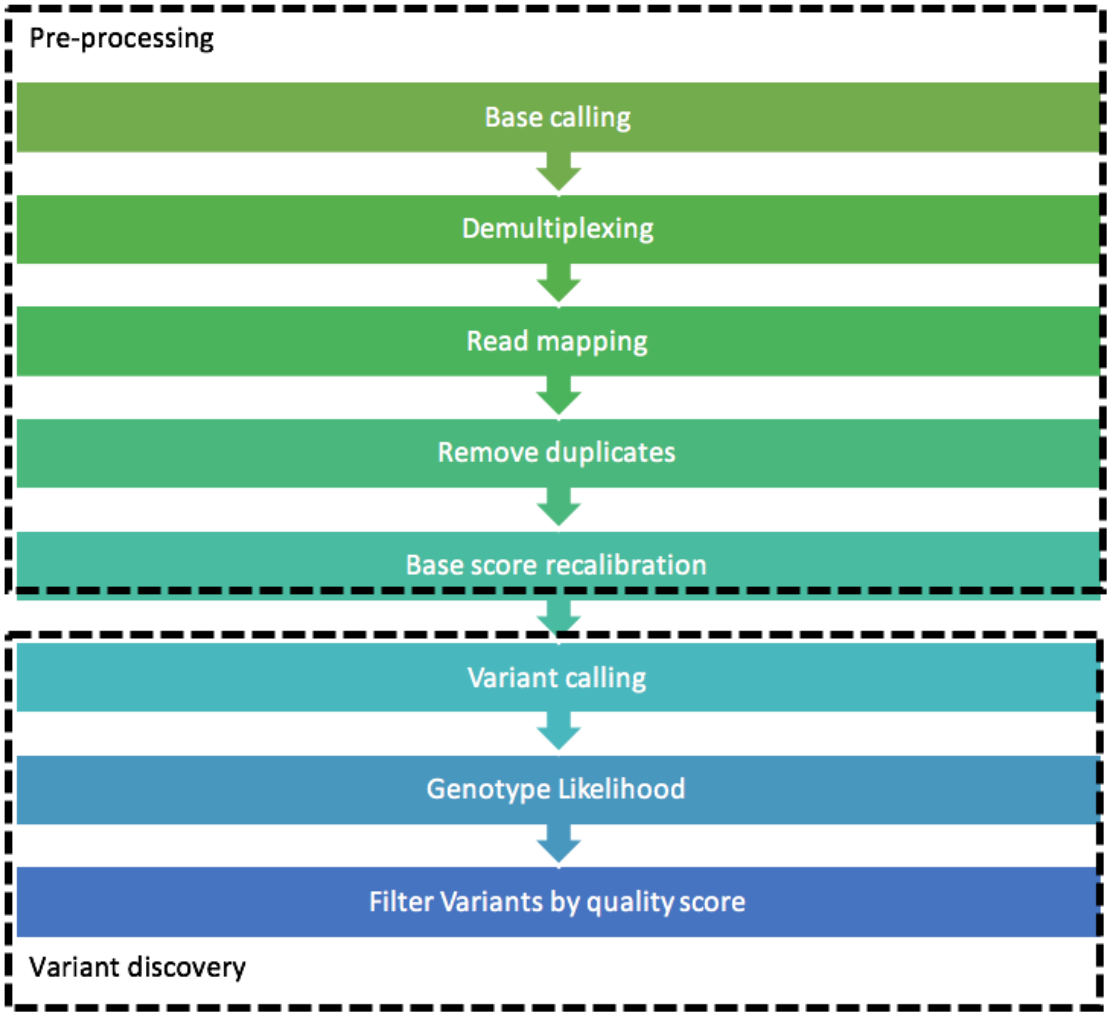


Figure 3. NGS data processing pipeline from sequence data to SNP and short InDel calling (Van der Auwera et al., 2013).

Mapping tools that can align the millions of short sequences produced form a single run vary in accuracy and speed (Hatem et al., 2013) and utilise different styles; MAQ (Li et al., 2008) and RMAP (Smith et al., 2008) build hash tables for reads whereas Bowtie2 (Langmead and Salzberg, 2012) and BWA (Li and Durbin, 2009) index the reference genome. All tools have different key performance indicators that they do well in, therefore selection of a mapping tool depends on the work being performed. Pre-processing of sequence data (de-multiplexing, removing index adaptors, mapping and marking duplicates) for whole genome sequencing currently has published best practices (Van der Auwera et al., 2013) utilising the Genome Analysis ToolKit

(McKenna et al., 2010) and mapping with BWA-MEM (Li, 2013) is a well-used process for current short-read (< 250bp) NGS data.

Once sequence data is mapped, for WGS and WES, variant identification is often the sought after concluding step. Variants can be single nucleotide polymorphisms (SNPs), short insertions or deletions (InDels), copy number variants (CNVs), large insertions or deletions, or large structural variants, including inversions and translocations approximately > 1mb in size (Sachidanandam et al., 2001, Mills et al., 2006, Freeman et al., 2006). Some early genome sequencing studies used uniformity of depth coverage and high base quality to effectively focus on small InDels and SNPs. As no mapping algorithm is perfect and there are a multitude of variants possible in each human genome, false negative and false positive variant calls are both possible and a problem for downstream analysis. This is exacerbated by low coverage such as that used in the 1000 Genomes Project Pilot Phase (Genomes Project Consortium, 2010) which relied on whole genome 3X sequencing. Low coverage sequencing can be useful as a cost-effective method for identifying variants in association studies, provided a large number of individuals are sequenced (Kim et al., 2010). Nonetheless identification of rare variants, such as individual SNPs in a rare mendelian disorder within a single family, requires a much higher depth such as >20X for WGS and >40X for WES (Meynert et al., 2014). For early SNV calling algorithms, 20X coverage worked well for calling variants at high quality bases, with the number of occurrences of each allele counted and fixed cut-offs of 20-80% alternate alleles for a heterozygous call used (Harismendy et al., 2009, Wang et al., 2008). Where the coverage is generally lower, this filtering and fixed cut-off method can lead to the under-calling of heterozygous genotypes: false negatives.

Algorithms that call SNPs and genotypes can use a probabilistic framework (Li et al., 2008, Li et al., 2009b, Li et al., 2009c), incorporating ‘genotype likelihoods’ with other prior information such as linkage disequilibrium and allele frequencies (Nielsen et al., 2011). These result in a SNV location, a genotype call, and a quality score indicating the strength of certainty of the call. This quality score can provide a statistical measure of uncertainty leading to a higher accuracy of genotype calling. By combining quality

scores, a genotype likelihood score can be calculated. The implicit assumption of independence among reads may be false due to the presence of PCR artefacts or even alignment errors. However, a weighting scheme that takes correlated errors into account can be used (Li et al., 2008). Error rates could also be estimated from each site in the read data independently, rather than using quality scores (Martin et al., 2010). Therefore, the genotype and SNP calling isn't reliant on the quality scores being accurately calculated, although the information regarding errors in the alignment process is lost this way.

In addition to variant calling directly from sequence data, SNP imputation is also utilised to fill missing data in variant data sets (Dai et al., 2006, Marchini and Howie, 2010). This takes prior information on the pattern of linkage disequilibrium surrounding sites and utilising known haplotypes. This is heavily used in the 1000 Genomes project (Genomes Project Consortium, 2012) and haplotype callers have been developed including the GATK HaplotypeCaller (Van der Auwera et al., 2013) and analysis using Haploview (Barrett et al., 2005). GATK variant discovery is particularly good and utilised by many researchers, however as a probabilistic method it can be outperformed in some areas by deterministic methods such as the string based clustering algorithm utilised by TidyVar (Noyvert, 2015). Progress is continually being made in the accuracy and speed of variant callers but false positive calls can still occur (Ribeiro et al., 2015), conflating results where they are not identified in downstream analysis. Therefore, parameters for mapping, genotyping and variant calling must be continuously re-evaluated and adapted for the specific project, such as weighing up conservative high quality variant calling compared to an increase in sensitivity (Warden et al., 2014).

Whole genome sequencing, or whole exome sequencing, and the subsequent downstream mapping and variant calling provides researchers with an exhaustive list of human variation to interpret. In the exome, this process is becoming increasingly computerised and automated through various tools such as the Ensembl Variant Effect Predictor (McLaren et al., 2016) and ANNOVAR (Wang et al., 2010). This is possible thanks to the understanding of the genetic code in proteins - the triplets of bases that directly code of amino acids (Crick et al., 1961). Utilising this knowledge, computation

of deleterious affects of amino acid changes through algorithms such as SIFT (Ng and Henikoff, 2003) (Kumar et al., 2009) and PolyPhen2 (Adzhubei et al., 2010) are possible.

1.2 Non-coding genome

Since genome sequencing has become both affordable and efficient, multiple collaborative efforts have emerged in the hope of assessing all human population variation. These include, but are not limited to, the 1000 Genomes project (The Genomes Project, 2015), the UK10K project (The, 2015), the ExAC project (Lek et al., 2016) and the current 100,000 genomes project (Mark et al., 2017). This has led us to huge amounts of genetic data becoming publicly available, allowing us to understand more about the frequency of population variation and its effect on both individuals and human evolution.

However, despite these advances in technology we are still unable to interpret the function, if any of most of the genome. Originally noted to be "junk DNA" (Ohno, 1972) and essentially useless, we now know the non-coding region of the human genome makes up close to 98% of our DNA (Venter et al., 2001) and no longer dismiss it as junk. A vast amount of work now goes towards elucidating all of its functions (Alexander et al., 2010), especially in light of evidence to suggest variation within it could be a cause for genetic disease (Alexander et al., 2010, Barr and Misener, 2016). Currently, a multitude of biochemical methods are used to define function within non-coding regions based on their interaction with DNA transcription proteins and their chromatin availability. The ENCODE project (Encode project consortium, 2007) and the Epigenome Roadmap (Kundaje *et al,*. 2015) utilise ChIP-Seq, ATAC-seq DNaseI hypersensitivity and FAIRE to determine potential regulatory elements.

1.2.1 Gene regulation

Gene expression can be measured in a variety of ways: from protein product (Burnette, 1981), RNA quantity (Alwine et al., 1977), RNA transcript quantification (Mortazavi et al., 2008) and reporter proteins (Chalfie et al., 1994). We know that different cell types have different RNA and protein profiles, therefore there must be differential gene

regulation. This is especially relevant when looking at the development of an embryo from a single fertilised egg. The regulation of the same ~20,000 genes in a human fertilised egg and the multi-cell developing embryo is the key to understanding how the process of development works and how it can go wrong.

We know that some of the instruction for gene regulation is in the vast expanse of non-coding genome (Pennacchio et al., 2006). The development of multicellular organisms depends on the precise, specific and accurate expression of genes both spatially and temporally. Genes encoding transcription factors play a critical role, as ultimately it is these proteins that are involved in the transcription of the genes (Hogan, 1996). Transcription factors help initiate and regulate the transcription of genes, including the recruitment of other transcriptional factors and opening the accessibility of the chromatin (Zaret and Carroll, 2011). Therefore, DNA binding events with transcription factors are crucial to the correct regulation of gene expression. In addition, regulation of the transcription of these factors themselves could also cause a cascade of mis-regulation of downstream genes in that transcription factor's network (Srivastava et al., 1997, Villavicencio et al., 2000).

A working model is that transcription factors bind to specific DNA motifs, but still these appear to have some wobble, allowing for some flexibility in binding (Herr and Cleary, 1995). These transcription factors recruit a co-activator complex to the DNA binding site and stabilise the transcription initiation complex at the promoter. Where this site is near the promoter of a gene, the transcription factors and mediator proteins are able to recruit RNA polymerase to initiate transcription. Sets of transcription factors can bind in co-localised regions known as cis-regulatory modules. These combinations can allow specific regulatory instructions for the nearby genes, influencing spatial-temporal transcription throughout development (Maeda and Karch, 2011) (Figure 4).

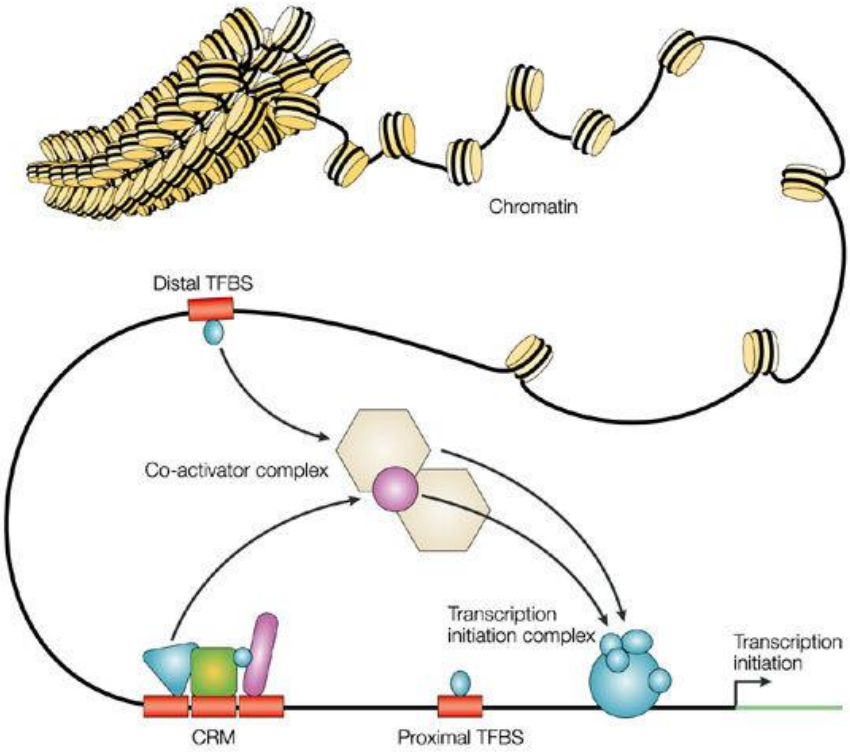


Figure 4. Regulation of gene expression by transcription factors
Adapted by permission from Macmillan Publishers Ltd: [Nature Reviews. Genetics] (Wasserman and Sandelin, 2004), copyright (2004)

Distal elements much further away from the transcription start site can act as gene regulatory elements: enhancers, insulators and silencers (Riethoven, 2010). This genomic regulation by distal transcription binding factors utilises the looping of the DNA and its 3D structure to bring distant acting enhancers close to the transcription start site in order to influence gene expression (Visel et al., 2009b). These are referred to as cis-regulatory elements or modules as they act on genes on their own chromosome. Understanding when these elements are active, what genes they act upon and even what sequence specific language or grammar they use to do so is the next vast frontier of genomics.

1.2.1.1 Conservation

One method of identifying developmental cis-regulatory elements is through comparative genomics. The hypothesis is that all vertebrate organisms share a similar

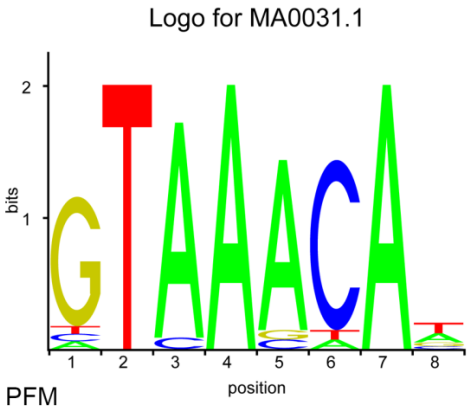
phylotypic stage in development, and the genes involved are essentially the same (Duboule, 1994, Raff, 2012). Therefore vertebrate-specific development is likely to utilise the same gene regulatory elements to control this process, with the sequence highly conserved amongst vertebrate species (Woolfe et al., 2005). Sequence conservation is most easily detected using BLAST (Altschul et al., 1990) or BLAT (Kent, 2002) software, however both of these algorithms require high identity thresholds when searching whole genomes for short sequences to obtain significant alignments. There are variations that have been developed for searching whole genomes against each other, such as MegaBLAST (Zhang et al., 2000), that makes this comparative genomics feasible and identification of large sets of non-coding sequence conservation possible (Lee et al., 2010) (Doglio et al., 2013).

Comparisons of vertebrate genomes, and specifically Fugu-human alignments (Aparicio et al., 1995) show ancient vertebrate conservation of stretches of non-coding DNA with the Fugu genome used due to its highly compact size (Brenner et al., 1993). These highly conserved non-coding elements cluster around vertebrate specific developmentally important genes (Woolfe et al., 2005). The concept that these stretches of non-coding DNA would stay near-identical throughout such a large evolutionary period suggests that variation and mutation in them would be detrimental to the organism, in a similar way that the coding genome is resistant to random mutation events due to the chance of them being disadvantageous (Drake et al., 2006) (Katzman et al., 2007). Therefore sequence comparisons are able to be utilised to identify human cis-regulatory elements (Prabhakar et al., 2006). These cis-regulatory elements identified through sequence comparisons are several hundred bases in length, therefore identification of the functional motifs within them is the next step to annotating the non-coding genome. Work has already begun to understand this cis-regulatory logic (Li et al., 2010) and more functional assays of non-coding regulatory elements are necessary to feedback into these computational predictors.

1.2.1.2 Transcription Factor Binding Sites

Sequence-specific DNA binding proteins (transcription factors) are the essential proteins utilised by cis-regulatory elements to regulate gene expression (Latchman,

1997, Chen and Rajewsky, 2007). Therefore, predicting and identifying the specific DNA motifs, or transcription factor binding sites (TFBS) could help identify cis-regulatory elements and even tissue-specific function. Computational methods have been implemented to do this (Ellrott et al., 2002) but often the number of identified sites is much greater than the number able to be functionally validated. This is made even more complex by the ability for these transcription factors to identify motifs with some 'wobble' – the exact motifs may vary by a handful of bases in some instances, the TFBS can be seen as a 'preference' rather than an exact sequence (Badis et al., 2009). For example, the JASPAR database (Mathelier et al., 2016) collates models of transcription factor binding sites in various species based on position frequency matrices (Figure 5).



PFM

A	1	0	19	20	18	1	20	7
C	1	0	1	0	1	18	0	2
G	17	0	0	0	1	0	0	3
T	1	20	0	0	0	1	0	8

Name	**FOXD1**
Class	Fork head / winged helix factors
Family	Forkhead box (FOX) factors
Species	Homo sapiens
Taxon	vertebrates
Uniprot ID	Q16676

Figure 5. An example of JASPAR database information for FOXD1 TFBS (Mathelier et al., 2016).

A key method in identifying binding sites is through capturing transcription factor binding events *in vitro*. Thanks to advances in sequencing, this combined with chromatin immunoprecipitation has allowed genome-wide ChIP-Seq to identify protein-DNA interactions (Valouev et al., 2008). Using transcription factors as 'bait' in these assays, binding events are captured and the sequences involved can be analysed for common motifs. This can be used to predict enhancers themselves (Schmidt et al., 2010, Visel et al., 2009a) as well as provide further TFBS information. The method relies on

crosslinking of transcription factors and DNA followed by using factor-specific antibodies to pull down the regions bound and detect individual binding events. Genome-wide, this one versus all method can describe a footprint of the transcription factor across the genome, and utilising a cell- or tissue-specific approach can help identify cell- or tissue-specific enhancers (Blow et al., 2010). Despite these benefits, this approach does struggle to distinguish between genuine functional binding events and those that do not instigate downstream events. Therefore, there is a high false-positive rate (Nix et al., 2008, Pickrell et al., 2011) and often the number of peaks is far too high to then functionally assay. Some progress has been made to reduce this, including comparisons of multiple TFBS peaks for common regions that could contribute to cis-regulatory modules (Zinzen et al., 2009). Crucially, understanding the specificity of transcription factor binding sites will be essential in elucidating the role a single nucleotide polymorphism could play within an evolutionary identified enhancer.

1.2.1.3 DNA-DNA interactions

Our understanding of the coding regions of the genome comes from the linear sequence of nucleotides, however the three-dimensional organisation of the chromatin has a part to play in gene regulation, bringing regions of the genome millions of bases away to close proximity in the nucleus. We are now able to capture these long-range interactions thanks to advances in a technique called chromosome conformation capture (Cope and Fraser, 2009). Crosslinking of DNA-DNA interactions and subsequent sequencing of the fragments involved allows the identification of loops and structural folding (Figure 6).

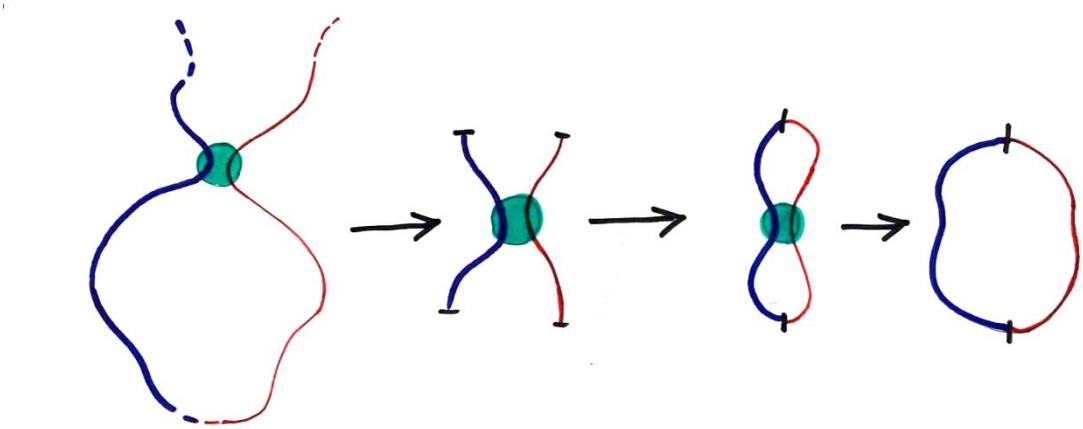


Figure 6. Chromosome conformation capture (3C) technique overview.
The interacting loci are cross-linked, capturing long range interaction. The DNA is restriction-enzyme digested and the fragments are then ligated. This occurs in dilute conditions to allow for preferential ligation to close-proximity strands. The DNA is then purified and various sequencing techniques are used to identify the interacting regions.

The chromosome conformation capture technique has evolved rapidly over the past decade, initially utilising PCR sequencing for a 'one vs one' viewpoint of interactions (Dekker, 2006). This is useful for confirming interactions between two regions. Alternatively the 4C approach of 'one vs all' (Simonis et al., 2006) allows a promoter or enhancer region to be used as a viewpoint and captures all potential interactions. This is particularly useful for identifying cis-regulatory elements for a specific gene, or the genes a cis-regulatory element might act upon. A recent iteration, Hi-C (Lieberman-Aiden et al., 2009) is an 'all vs all' high throughput sequencing approach that has since been extensively used across the whole human genome and led to the definition of topologically associating domains. Mapping these long-range interactions reveals how the human genome folds on itself and creating functionally important loops (Lieberman-Aiden et al., 2009).

Recently, evidence has shown that the underlying organisation of genomic regions can be mapped to approximately 100kb to 1mb of locally interacting DNA known as topologically associating domains (TADs) (Dixon et al., 2012). Utilising TAD information may help understand how non-coding regions affect gene regulation, as

enhancers within the same TAD as a specific gene are much more likely to act on that gene than one potentially closer but in a different TAD.

Topologically Associating Domains have been shown to have limited change when compared across species and cell types (Dixon et al., 2012, Nora et al., 2012) with key changes in chromatin architecture reorganisation in differentiating stem cells (Dixon et al., 2015). Some evidence has shown their involvement in facilitating transcriptional regulation through the integration of regulatory activities within their boundaries (Nora et al., 2013). TADs bring together the genes and cis-regulatory elements to allow cell-specific gene expression patterns characteristic of phenotypes observed (Sanyal et al., 2012). As the majority of GWAS variants are found in the non-coding genome (Manolio et al., 2009), utilising the information from TAD boundaries may help assign the correct gene these variants are acting on if found at TAD boundaries. In addition, disruption of TAD boundaries has been shown to form *de novo* enhancer-promoter interactions, gene mis-expression and resulting abnormal phenotypes (Lupiáñez et al., 2015).

A key limitation to these techniques is the resolution of interactions they obtain, mostly limited by the enzyme digestion and the sequencing depth. With deep sequencing, restriction fragment length resolution is possible with Hi-C (Jin et al., 2013, Rao et al., 2014), allowing mapping of interactions to ~1kb in cell lines. However, enhancers can often be much smaller than that and binding sites within them even smaller so. Therefore 3C, 4C and Hi-C methods all have their benefits but do not reveal the single nucleotide level of sequence contribution to cis-regulatory elements. Many different models are now developing to resolve how the TADs shape the 3D architecture of the genome and drive the networks of cis-regulatory elements that determine gene expression during development (Remeseiro et al., 2016).

1.2.2 Human non-coding variation

Within the non-coding genome, natural human variation can occur and this is markedly more often than in the exome due to the functional constraint on the sequences that transcribe proteins (Genomes Project Consortium, 2010). Since the initial 1000

Genomes data release, the third release (The Genomes Project, 2015) has presented with more opportunity to interrogate the noncoding genome for rare variants thanks to an improvement in coverage. Any individual is thought to carry 3.5-4.5 million SNPs across their whole genome, with individuals of African ancestry having the most (The Genomes Project, 2015). Therefore, any attempt to find function in the variation of the non-coding genome is met with a vast number of variants with limited information. Reducing the region of the non-coding genome to be interrogated through functional annotation of regulatory elements allows more reasonable analysis. Previous comparisons of percentage of sites containing SNPs across the non-coding genome compared to the coding sequence consistently show higher frequencies of variants in the non-coding. When comparing variation in CNEs, this shows these sequences are constrained to a similar level to non-synonymous coding variants (De Silva et al., 2014). Evidence shows that these CNEs are selectively constrained and not just mutational cold spots that have been retained over evolution (Drake et al., 2006). Outside of non-coding regulatory elements there is a large potential for genetic variation with no noticeable phenotypic effect. However there is also the potential for accumulation of mutations in the non-coding genome to influence continuous traits such as height (Lango Allen et al., 2010). The extent of non-coding regulatory variation and its contribution to evolution and natural selection is unclear (Lappalainen and Dermitzakis, 2010). However if gene regulation can be affected incrementally through regulatory variation, it gives the potential for smaller phenotypic variations to accumulate and, if beneficial, be positively selected for evolutionary. The converse could also be true if small genetic variation in regulatory elements is able to have just as severe an impact on phenotype as a nonsynonymous mutation.

1.2.3 Contribution to human disease and disorders

Most disease-associating variants that have been found in GWAS studies are located in the non-coding region of the genome (Maurano et al., 2012). Often these variants correlate to relatively small increments in risk in complex human diseases and traits (Manolio et al., 2009). There have been some examples of non-coding variation being the fundamental cause of a developmentally based disorder. A key example that also demonstrates the vast distances that gene enhancers can act across is a set of mutations

in the ZRS regulatory element that acts on the *Shh* gene (Lettice et al., 2003). Four unrelated families with congenital polydactyly were found to have point mutations within the ZRS, affecting the Shh expression in the ZPA. Since this evidence of such a strong effect of a SNP at such a large distance (~1Mb), further efforts to provide evidence for non-coding SNPs to be causative of human disease have gained traction. The use of 4C-seq analysis have shown that a common functional variant in a cardiac enhancer modulates cardiac SCN5A expression, predisposing patients to arrhythmia (van den Boogaard et al., 2014). Another approach has shown that gestational hyperglycaemia associating haplotypes disrupt regulatory element activity of the nearby HKDC1 gene, altering glucose homeostasis (Guo et al., 2015). The Deciphering Developmental Disorders project estimates that 42% of their cohort may have pathogenic mutations in the coding regions of the genome (Deciphering Developmental Disorders, 2017) and although the other half of the cohort may not be diagnosed through non-coding mutations, it is fair to predict that some of these cases may be explained this way. It is difficult to tell if multiple variants, both in the coding and noncoding regions may contribute to developmental disorders and much of the functional analysis of non-coding variants is yet to be done. In many instances, associating variants to disease phenotypes is possible, as is associating them to familial inherited disorders, however proving their effect on gene expression can be far more complex and time consuming.

1.2.4 Noncoding variation functional prediction

Noncoding variants can be pathogenic and they are also found at a higher rate than coding variants. Therefore, a great deal of resource has been put into computational methods to predict non-coding variant function and pathogenicity. This is made particularly difficult as many regulatory elements are predicted themselves based on sequence conservation, transcription factor binding sites, and chromosome conformation capture methods (see above: 1.2.1.1, 1.2.1.2, 1.2.1.3). Regulatory SNVs can affect histone modification, DNA methylation, and TF binding and all to various extents. The effect of these SNVs is very much unknown and we must rely on prediction models to sift through the vast number of personal and associating SNVs to find pathogenic variants.

Computational methods of scoring pathogenicity variants rely on available functional data sets to complement these predictors. As mentioned previously, comparing PWMs between normal and variant versions of a locus may help judge the change in putative binding affinity that a variant imposes. Utilising public databases such as TRANSFAC (Matys et al., 2003), JASPAR (Mathelier et al., 2016) and UniPROBE (Hume et al., 2014) can give these scores and their changes (Bailey et al., 2009) and predict pathogenicity based on changes in transcription factor binding. Methods such as the transcription factor affinity prediction (TRAP) utilises ChIP-seq peaks to determine the highest binding affinity transcription factors and then gives mutated sequence p-values for changes in binding sites (Thomas-Chollier et al., 2011). Much of the prediction of effect is reliant on a collection of know transcription factors and their binding sites, and is therefore limited by functional assays to feed in more information.

Further improvements to the functional prediction of SNVs can be made by integrating more publicly available experimental data, such as that of ENCODE (ENCODE Project Consortium, 2007). This uses DNaseI-hypersensitive sites (DHS) and histone modifications to predict regulatory elements. As short TFBS sequences (6-20bp generally) can be found in a large proportion of the genome, finding variants that are in active regulatory elements are more likely to be pathogenic than those in inactive regions. A database that utilises chromatin states alongside transcription factor binding (both motifs and experimentally validated data) is RegulomeDB (Boyle et al., 2012). This combines information on histone modifications, DHS, TF binding, TF motifs and conservation to score variants in categories related to predicted functional consequences (Table 1). This integration of data and emphasis on expression quantitative trait loci (eQTLs) helps to identify active regulatory elements. The eQTLs are regions of the genome shown to have an influence on gene expression level however these experiments are costly both financially and in time and therefore unable to be used for all possible variation found. In addition, RegulomeDB is somewhat limited by the known human genetic variation, currently using dbSNP build 141 (Sherry et al., 1999, Sherry et al., 2001). Its output is easily interpreted and compared across cohort groups, or case-control studies.

Table 1. RegulomeDB scoring categories for SNPs

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

An additional method to evaluate the effect of a single base change is to look at the precise genomic location conservation over evolution. This is particularly useful when looking at variants associating with developmental disorders as regulation of vertebrate-specific development appears to be conserved between species (Piasecka et al., 2013) and by conserved regulatory elements (Woolfe et al., 2005). However, within these conserved noncoding elements there is some variation in individual bases despite the high overall consensus sequence. Individual nucleotides can be in non-variable or restricted variable regions (NVRs or RVRs) (De Silva et al., 2014). The base-by-base conservation, and the probability of a variant to be pathogenic as a result, can be scored using a database like GERP++ (Davydov et al., 2010). GERP++ uses rejected substitutions and neutral rate of mutation over evolution to give base-wise scores of variants based on alignments and a model of neutral evolution. The limitations of GERP++ and other similar approaches arises from the neutral rate of mutation

estimations. They are often uncertain and can vary dependent on the alignment quality, methodology used to estimate them and the genomic region. In addition, these methods make the assumption that all sites in the sequence region are independent.

Further computational methods attempt to integrate more data for large-scale annotations and scoring. One of that is widely used is the Combined Annotation Dependent Deletion software (CADD) (Kircher et al., 2014). This software integrates multiple annotations and contrasts variants with simulated mutations and known common human variation. Machine learning software like CADD allows easy scoring of variants and ranking within a cohort set. It utilises 88 annotations from genomic and epigenomic data sets covering conservation, transcription factor binding, cell expression levels, chromatin states and histone modifications. It utilises the variant effect predictor (VEP) from ensembl (McLaren et al., 2016) for annotation as well as ENCODE data. CADD has been shown to be a valuable tool for noncoding annotation (Richardson et al., 2016) but some questions over its clinical validity remain (Mather et al., 2016). Some of the difficulty in scoring noncoding variants using CADD may result from its machine learning approach and training data that also contains coding variants. Some unsupervised approaches of integrating the same amount of annotations and scoring have also been developed (Ionita-Laza et al., 2016) with the latter more preferable in the absence of a large, representative and correctly labelled training set. CADD is able to readily integrate new information, and its upkeep in light of continued ENCODE annotation releases is a crucial benefit. The key addition to noncoding variant annotation will be tissue specific eQTLs and expression analysis, especially in light of human disease and phenotypes. However, these methods can only prioritise variants based on predicted functionality. This is a necessary step to reduce the number of variants to be functionally validated and both time and cost prevent all from being investigated. Once functional validation has been carried out, it is imperative that this information is fedback into these predictive models to continually improve their accuracy. This limiting step of validation is one of the key roadblocks in determining noncoding variation function.

1.2.5 Noncoding variation functional validation

Functionally validating noncoding variants is dependent on prior knowledge or hypothesis of how these variants function. Functional validation could be defined as showing that a genetic variant has a cause and effect relationship, affecting gene function, expression or developmental processes. This is developed through the annotation tools mentioned above (1.2.4). There are various methods that can be used to assess a variant's effect on gene expression and potentially phenotype including luciferase assays (Ozaki et al., 2002), allele-specific FAIRE assays (Smith et al., 2012), transient enhancer assays (Bessa et al., 2009), dual-reporter transgenesis (Bhatia et al., 2015), and CRISPR-Cas9 mutagenesis (Canver et al., 2015). Many of these methods are low-throughput, or the ones that are high-throughput reply on relevant cell lines and are not necessarily able to translate to the whole organism. This is particularly true for developmental enhancers, such as those predicted by conservation, and *in vivo* methods are more likely to give better evidence to the effects of a regulatory SNP on the developing embryo than *in vitro* methods. Conversely, *in vitro* methods may give a better understanding of the effect of a variant at the molecular level, such as transcription factor binding.

In vitro methods to quantify the effect of a SNV on DNA-protein interactions have been used previously to confirm non-coding variant effects in GWAS identified variants (Oldoni et al., 2016). Allele-specific formaldehyde-assisted isolation of regulatory elements (FAIRE) can show binding differences between wild-type and variant regulatory elements but gives limited information in regards to the transcription factor binding that changes and is reliant on cell-type specific nuclear extract. Luciferase reporter assays allow quantification of the changes in enhancer-driven gene expression between variants, but are again limited by appropriate cell lines and the lack of whole-organism information. These assays are fundamental as proof of concept and have been developed to be high throughput and quantitative (Smith et al., 2012, Melnikov et al., 2012).

An additional method of interrogating the relationship between a putative enhancer and gene regulation is the visualisation of a reporter gene under the control of the element *in*

vivo. The core principal that an enhancer can drive a minimal promoter underpins much of the *in vivo* enhancer assay used. This includes lacZ assays such as those in the extensive VISTA enhancer catalogue in mice (Visel et al., 2006) and the Tol2:GFP transposon mediated approach in Zebrafish (Kawakami, 2007). In addition, stable zebrafish transgenic lines using allele-specific enhancer-reporter constructs for the regulatory region of interest can show differences in *in vivo* function, although these can be hard to detect and are very low-throughput (Liu et al., 2017). The benefits of transient enhancer assays in Zebrafish come from the medium-throughput approach (multiple constructs can be analysed in a week) as long as the disease the variant associates with is developmentally relevant. Nevertheless, the mosaicism that can occur from random integration of the expression construct can make it difficult to find tangible evidence of a variant's effect between microinjections. Therefore, it is paramount to do multiple repeats, and some work has been performed to utilise dual-colour assays to allow an all-in-one approach, removing variation between injections (Bhatia et al., 2015). Therefore, there is a fine balance to be met between high-throughput and less reproducible analysis and low-throughput but highly accurate experiments.

One method of reliably assessing the function of a regulatory element variant in relation to a disease phenotype would be to create a comparable mutation in an animal model. Thanks to the advancement of CRISPR-Cas9 technology (Ran et al., 2013), the ability to mutate allele-specific sites anywhere in the genome is possible. For regulatory elements that have been discovered through comparative genomics, such as evolutionary conserved developmental enhancers, this method can be suitable as the DNA surrounding a regulatory variant is likely to be identical in human and in the vertebrate model being used. This genome-editing is costly and time consuming and therefore strong prior evidence of the variant's function must be observed, however resulting phenotypes can be verified, as well as changes in gene expression (Han et al., 2015). This single-base interrogation of regulatory elements will not only give insights into the downstream effect of a variant but also feed back into the field's computational predictions of SNV effects and help understand the language and grammar of the noncoding genome. Conversely, this approach does not take into consideration the

environment-genome interactions and the part they play in complex hereditary diseases. As it is expected that multiple noncoding variations and their accumulation are more likely to affect gene regulation due to combined effects (Cannavò et al., 2016) combinations of mutations may need to be implemented to fully understand the threshold for a disease phenotype from noncoding variation. This information will ultimately shape the way we search for noncoding variants and predict and validate their impact, in an attempt to consolidate the vast amount of sequence information we are currently producing.

Chapter 2. Materials & Methods

2.1 Methods for Chapter 3: Obesity project

2.1.1 Sequencing

Samples were curated and individuals were assessed as described previously (Klöting et al., 2008). Libraries were prepared for sequencing using Illumina Nextera Rapid Capture Custom Enrichment Kit (Cat ID FC-140-1009). The custom kit included 8,701 probes across the 2 Mb region for 288 samples (Project ID 44309). All samples were run on an Illumina HiSeq 2500 at 100 cycle pair end reads. Ninety-six multiplexed samples were run per flow cell with each multiplex being run twice on Rapid Run mode. Samples were de-multiplexed and converted to FASTQ files using Illumina software CASAVA.

2.1.2 Ethical statement

The study was approved by the regional scientific ethics committee and by the Danish Data Protection Board and fulfilled the Helsinki Declaration.

2.1.3 Availability of supporting data

Sequence data (reads) are be available through ENA at http://www.ebi.ac.uk/ena. Accession number PRJEB11794. All other data are contained within the paper or supplementary information files. All other data is fully available on request, without restriction.

2.1.4 Mapping and variant calling

Sequencing data (FASTQ) files was mapped to the hg19 assembly of the human genome, the version in human_g1k_v37.fasta file available from the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/ reference/). BWA (Burrows-Wheeler Aligner) software was used to map the reads (Li, 2013), version bwa-0.7.8, bwa-mem algorithm with default parameters. The mapped read (sam) files were then converted to bam format using samtools version 0.1.19 (Li et al., 2009a). The

reads in each bam file were then sorted by chromosome and coordinate and indexed using samtools.

Duplicate reads were marked by Picard (http://broadinstitute.github.io/picard), version 1.91, MarkDuplicates tool. Then the two bam files from different sequencing runs for each individual were merged using Picard tool MergeSamFiles. The individual bam files per sample were then processed by our in-house tool 'TidyVar' (B. Noyvert and G. Elgar, manuscript in preparation, https://github.com/boris-noyvert/TidyVar.m), which is an implementation of a novel variant calling algorithm. The algorithm uses a string matching approach to detect SNPs and short insertions and deletions, the individual genotypes are assigned using pattern recognition. A single vcf file listing all the variants found in all the individuals was produced.

2.1.5 Haplotype analysis

Haplotyping was performed with Haploview (Barrett et al., 2005) using the methods described previously for defining linkage disequilibrium blocks (Gabriel et al., 2002). For this programme, only biallelic SNPs were used across the region chr16:53,500,000-55,500,000. Comparisons over each variant over 500Kb were performed and settings altered from default to ignore Hardy-Weinberg P values, and to include only individuals with a minimum of 75% of all SNPs successfully called. Associations of individual variants and haplotypes were produced through Haploview using the case-control allelic chi-squared test with one degree of freedom for the 2 × 2 contingency table of allele counts for reference and non-reference alleles and for case and control separately (Clarke et al., 2011). The output P-values of this were used throughout this study.

2.1.6 Interaction data liftOver

The UCSC genome browser utility liftOver (http://geno me.ucsc.edu/cgi-bin/hgLiftOver) was used as the Batch Coordinate Conversion method to transfer SNP hg19 coordinates to mouse genome build mm9 coordinates using default settings. Conversion of 5,842 SNPs was successful with 5,988 SNP locations failing.

2.1.7 Genotyping and imputation of replication cohorts

This method (2.1.7) was not performed by the author but is essential to the subsequent work presented.
Genome-wide genotyping on the Illumina 610 k quad chip was carried out at the Centre National de Genotypage (CNG), Evry, France. SNPs with minor allele frequency <1 %, >5 % missing genotypes or which failed an exact test of Hardy-Weinberg equilibrium (HWE) in the controls (P <10−7) were excluded. Any individual who did not cluster with the CEU individuals (Utah residents with ancestry from northern and western Europe) in a multidimensional scaling analysis seeded with individuals from the International HapMap release 22; who had >5 % missing data; outlying heterozygosity of >35 % or <30.2 %; genetic duplicates; one of each pair of genetically related individuals; individuals with sex discrepancies and one individual whose genotyping was discordant with a previous project were excluded. Imputation to HapMap release 22 (CEU individuals) was carried out using Mach 1.0, Markov Chain Haplotyping. This method was used for both the Male and Female GOYA cohorts (Nohr et al., 2009, Paternoster et al., 2011).

Imputed genotypes for the sequenced 284 men (where available) were compared to the sequenced genotypes called by TidyVar and found to be correct 100 % for rs9939609:T>A and 98.3 % correct for the SNPs rs7186407:A>T, rs12598453:C>G and rs12596270:A>G.

2.1.8 Topological association domain comparison

The –log10(P-value) for association of each SNP with the case or control cohort was used in preparation of a variable step .wig file with a scale of 0 to 6 and each line to span 1 base. The coordinates for each SNP were converted using UCSC LiftOver from hg19 to hg18 to fit with the original scaffold used for the Hi-C data. The data used for the Hi-C tracks are limited to human embryonic stem cells (hESC). Default max [50] and min [10] values were used for the heat map visualisation ((Dixon et al., 2012) and http://yuelab.org/hi-c/).

2.1.9 Data sources

CNE locations were taken from CONDOR (Woolfe et al., 2007). Exon coordinates were taken from Ensembl Biomart release version 75 (Flicek et al., 2013). All sequence coordinates in this study are from GRCh37/hg19. 1KG (1000 Genomes) variant data are taken from the publicly available 'ALL.chr16.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz' VCF file found at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/. All dbSNP variants are taken from the NCBI publicly available 'human_9606_b142_GRCh37p13/VCF/All.vcf.gz' VCF file found at ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/. Epigenomic data were sourced and visualised through the WashU Epigenome Browser (http://epigenomegate way.wustl.edu/) using ENCODE GIS Chia-Pet publicly available data (Li et al., 2012).

2.2 Methods for Chapter 4: CNE sequencing of four cohorts

2.2.1 Selection of CNEs and probe design

CNEs were selected as described previously (Woolfe *et al*., 2007). Illumina Truseq custom enrichment probes were designed to target these regions using Illumina's DesignStudio tool (Project ID 3897).

2.2.2 Library Preparation

Libraries were prepared for sequencing using Illumina Truseq DNA Sample Preparation v2 kit (Cat ID FC-121-2003) (CLP, SCHZ, IDE) or Illumina Nano DNA Sample Prep kit (Cat ID FC-121-4001/ FC-121-4003) (ANOS, SCHZ) and Capture Custom Enrichment Kit (Cat ID FC-123-1096).

2.2.3 Custom enrichment

CNEs were enriched for from the DNA samples using Illumina Truseq Custom Enrichment Kit. The custom kit included 3542 probes across 916kb of the human genome targeting 3006 regions (coverage of ~5000 CNEs – all ≥80bp CNEs and those within 500bp of an 80bp+ CNE) (Project ID 3897). This method uses biotinylated

probes that bind to streptavidin beads in order to magnetically pull-down regions of DNA of interest.

2.2.4 Sequencing

Samples were run on an Illumina HiSeq 2500 at 100 cycle pair end reads (SCHZ, IDE, ANOS). 192 samples for were run on an Illumina GA II at 100 cycle pair end reads (CLP). Samples were de-multiplexed and converted to FASTQ files using Illumina software CASAVA.

2.2.5 Mapping and variant calling

I mapped sequencing data (FASTQ) files to the hg19 assembly of the human genome, the version in human_g1k_v37.fasta file available from the 1000 Genomes project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/). I used BWA (Burrows-Wheeler Aligner) software to map the reads (Li, 2013), version bwa-0.7.8, bwa-mem algorithm with default parameters. The mapped read (sam) files were then converted to bam format using samtools version 0.1.19 (Li et al., 2009a). The reads in each bam file were then sorted by chromosome and coordinate and indexed using samtools. Next duplicate reads were marked by Picard (http://broadinstitute.github.io/picard), version 1.91, MarkDuplicates tool. These bam files were then processed by the Elgar lab's in-house tool 'TidyVar' (B. Noyvert and G. Elgar, manuscript in preparation, https://github.com/boris-noyvert/TidyVar.m), which is an implementation of a novel variant calling algorithm. The algorithm uses string matching approach to detect SNPs and short insertions and deletions, the individual genotypes are assigned using pattern recognition. A single vcf file listing all the variants found in all the individuals was produced.

2.2.6 Prioritisation of candidate SNPs

Standard QC and QA methods were used including quality score of 100+ for TidyVarv4 coverage depth of 20. R programming software was utilised in the RStudio environment alongside Microsoft Excel in order to annotate and prioritise variants as outlined further in Figure 27 (page 96). The output vcf file from 2.2.5 was used as an input to Ensembl's Variant Effect Predictor (McLaren et al., 2016) to annotate ethnic sub-group allele

frequencies, SNP rsIDs and previously identified pathogenic variants. These annotations were used in the pipeline visualised in Figure 27, utilising the online tools and their downloaded data sets as mentioned.

2.2.7 Cloning of candidate CNEs

Table 2. PCR Primers used

Location (hg19)	Size of PCR product	Forward Primer	Reverse Primer
chr2:104496627-104496946	320bp	TGATACCTCAGCTTTCTTGGACT	GCAGCAGCGAACCATATTATCA

Primers were designed using Primer3 software (Rozen and Skaletsky, 1999). Polymerase chain reactions were set up using CNE-specific primer pairs and standard taq polymerase according to manufacturer's guidelines (Table 2). PCR products were visualised by standard agarose gel electrophoresis to confirm size and purity. PCR products were then purified using Qiagen QIAquick PCR purification kit. Purified products were cloned into the pCR8/GW/TOPO vector (Invitrogen) as per manufacturer's guidelines and transformed into Oneshot TOP10 chemically competent *E. coli* cells (Invitrogen) according to manufacturer guidelines. Outgrown cultures were then spread on agar plates containing the antibiotic spectinomycin and grown overnight at 37°C. Colonies were then picked the next morning and inoculated in 3mls of spectinomycin-containing lysogeny broth. These were incubated overnight at 37°C with agitation. The following morning, 2mls of culture was prepared using the QIAprep Spin Miniprep kit (qiagen) according to manufacturer's guidelines to obtain 50µl plasmid.

100ng of entry clone then underwent Gateway LR recombination (Invitrogen) with the pGW_tol2:cfos:egfp vector according to manufacturer's guidelines and transformed into Oneshot TOP10 chemically competent E. coli cells (Invitrogen) according to manufacturer guidelines. Outgrown cultures were then spread on agar plates containing the antibiotic ampicillin and grown overnight at 37°C. Colonies were then picked the next morning and inoculated in 3mls of ampicillin-containing lysogeny broth. These

were incubated overnight at 37°C with agitation. The following morning, 2mls of culture was prepared using the QIAprep Spin Miniprep kit (qiagen) according to manufacturer's guidelines to obtain 50µl of microinjection-ready plasmid. The correct insertion of the CNE sequence was confirmed by sanger sequencing (source bioscience).

2.2.8 Enhancer assay in Zebrafish embryos

Enhancer assays of CNEs in Zebrafish embryos is adapted and described previously (Fisher et al., 2006; Kawakami, 2007). The vector is described as Tg(cne-cfos:egfp) after cloning and further referenced as the 'expression vector'. *Tol2* transposase mRNA was transcribed *in vitro* from a linearised pCS-Tp vector containing the *Tol2* transposase ORF using the mMESSAGE m MACHINE SP6 kit (Invitrogen) according to manufacturer's guidelines. The microinjection mix totalling 5µl was prepared as follows:

- 1µl expression vector DNA (150ng/µl)
- 0.5µl transposase mRNA (300ng/µl)
- 0.5µl 0.1% Phenol Red
- 3µl ddH_2O

This mix was prepared on ice and injected into wild-type Zebrafish at the 1-cell stage. Embryos were stored at 28°C in Zebrafish embryo medium, with the addition of PTU after 24 hours, and screened for GFP expression patterns at 24hpf, 48hpf and 72hpf using fluorescence microscopy.

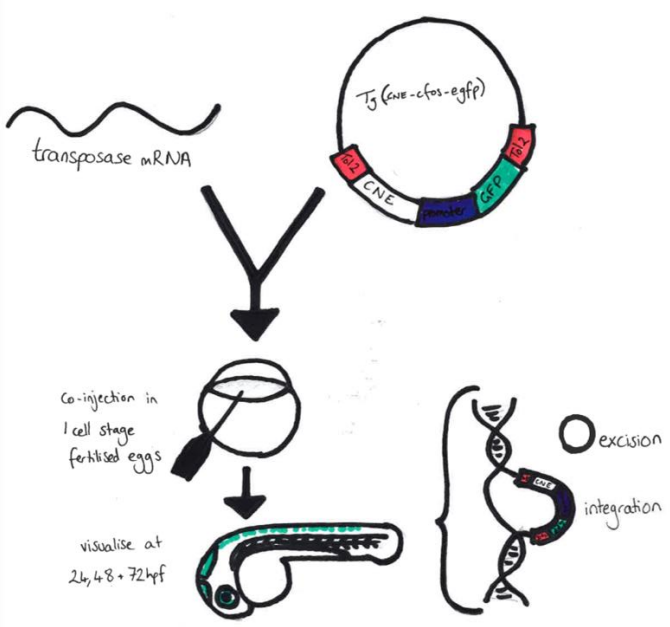


Figure 7. Transient GFP enhancer assay in Zebrafish

2.3 Methods for 0: Mitochondrial DNA sequencing

2.3.1 Mitochondrial isolation

This method 2.3.1 was not performed by the author but is essential to the subsequent work presented.

Mitochondria were isolated for mouse tissues (liver, brain) by differential centrifugation as previously described (Spinazzola et al., 2006, Gonzalez-Vioque et al., 2011).

2.3.2 Sequencing library

Part of this method (2.3.2) in italics was not performed by the author but is essential to the subsequent work presented.

Mouse mtDNA was purified from sucrose-gradient isolated mitochondria. Purified mtDNA was fragmented prior to library preparation using a Covaris S220 and the Sonolite software with settings of duty cycle 10%, intensity 5, 200 cycles for 3 minutes at 4℃. 200bp paired-end DNA libraries were prepared using the Illumina Truseq LT kit and run on the Miseq.

Sequencing data (FASTQ files) were mapped to the mm9 assembly of the mouse mitochondrial genome. Reads were mapped using BWA software (version bwa-0.7.8) (Li, 2013) using the bwa-mem algorithm with default parameters. The mapped read sam files were converted to bam format using samtools version 0.1.19 (Li et al., 2009a), and the reads sorted and indexed using samtools. Then the two bam files from different sequencing runs for each sample were merged using Picard tool MergeSamFiles.

2.3.3 Calculating coverage depths and mutation loads

The number of single nucleotide substitutions at each individual base and the overall coverage at each base position was calculated using samtools (mpileup). Dividing these numbers but the total read coverage yielded the SNP frequencies for each of the 3 possible non-reference alleles, the sum of which gave the total mutation load. 'Mutation load' is likely to be an overestimate due to false-positives. These can arise from the sequencing technology used and from the complexity of the sequence itself at that locus. Still, by only comparing samples from within a single sequencing run, the false positive error rate can remain consistent between samples.

2.3.4 Statistical analysis

Data are expressed as the mean ± the standard error of the mean (SEM). Group means were compared using parametric t-test or non-parametric Mann-Whitney test. One-way ANOVA was used to compare more than two independent groups. A P-value of <0.05 was considered to be statistically significant.

Chapter 3. Complete re-sequencing of a 2Mb topological domain encompassing the FTO/IRXB genes identifies a novel obesity-associated region upstream of IRX5

This work is published in Genome Medicine 7, no. 1 (2015): 126

"Complete re-sequencing of a 2Mb topological domain encompassing the FTO/IRXB genes identifies a novel obesity-associated region upstream of IRX5" by Hunt et al. (2015).

3.1 Background

Previous genome-wide association studies (GWAS) have consistently identified single nucleotide polymorphisms (SNPs) associated with obesity located within the first intron of the FTO gene on human chromosome 16q12.2 (Frayling et al., 2007, Hinney et al., 2007, Scuteri et al., 2007). Findings from these studies have been confirmed in meta-analyses wherein the associated SNPs are in strong linkage disequilibrium (LD) with one another. These SNPs lie in a noncoding region of the genome, resulting in some contention over their functional impact on neighbouring genes. The strongest association is found for SNP rs1121980:C>T with an odds ratio of 1.66 among 929 Caucasians (Hinney et al., 2007). This variant is in LD with a number of other SNPs ($r2 \geq 0.88$ for all), including rs9939609:T>A, which has been the most extensively genotyped. The rs9939609 risk allele (A) has an odds ratio itself of 1.34 for heterozygotes and 1.55 for homozygotes (Wellcome Trust Case Control, 2007). This association has also been identified for type 2 diabetes (T2D); however, when adjusting for body mass index (BMI), the T2D association is lost suggesting that this association is a secondary effect of BMI (Frayling et al., 2007).

The association of obesity with rs9939609: T > A has been replicated in many independent study groups across a range of different ethnicities (Dina et al., 2007, Fang et al., 2010, Hakanen et al., 2009, Hennig et al., 2009, Hotta et al., 2008, Villalobos-Comparan et al., 2008). Nevertheless, the degree of linkage disequilibrium across the entire intron 1 of FTO has prevented a single potentially functional SNP from being identified, although trans-ethnic comparison has permitted a degree of fine mapping of the region (Akiyama et al., 2014). The LD region identified in the HapMap Phase II data spans about 50 kb, covering part of the first intron of FTO, the second exon and a small portion of the second intron (International HapMap Consortium, 2007). Despite this, coding SNPs in the second exon of FTO have not been found to follow the same association patterns.

As a result of the persistent association with obesity in this region, the function of the surrounding gene, FTO, has been under close scrutiny. FTO is a ubiquitously expressed N6-methyladenosine demethylase (Jia et al., 2011) , yet there are conflicting data and models of how changes in FTO expression might affect function and phenotype. Mouse models have been informative; knockdown of FTO in mice results in reduced fat mass, suggesting that the susceptibility to obesity could be through overexpression of FTO (Church et al., 2009). A further mouse FTO knockout has been described generated through replacement of exons 2 and 3 with a neomycin STOP cassette (Fischer et al., 2009). This mouse exhibits growth retardation from postnatal day 2 onwards although it also shows a broader range of phenotypes including higher postnatal death. It supports the hypothesis that FTO is involved in energy metabolism and body weight regulation as the knockout mice show a reduction in adipose tissue and increased energy expenditure. Conversely, eQTL analyses examining the links between the associated SNPs and the expression levels of FTO have not to date identified a clear and direct correlation (Grunnet et al., 2009, Klöting et al., 2008, Wåhlén et al., 2008).

A few hundred bases upstream of FTO, and transcribed in the opposite direction, is the RPGRIP1L gene. As a result of its proximity to the LD region, the function of this gene has also been closely examined on the premise that non-coding SNPs might affect the

regulatory landscape acting in cis on this nearby gene. Some evidence to this effect has been reported (Stratigopoulos et al., 2011) and Rpgrip1l+/− mouse models gain weight more rapidly than their wild-type litter mates, as well as exhibiting increased energy intake and increased adiposity (Stratigopoulos et al., 2014).

More recently, chromosome conformation capture (3C) approaches have demonstrated that longer-range interactions occur across this region acting at both the FTO and IRX3 gene promoters (Smemo et al., 2014) although the concept of long-range regulation in this region has been speculated upon previously (Ragvin et al., 2010). These studies point to IRX3 as a further potential candidate gene that might interact with the associated SNPs in the first intron of FTO. In the paralogous IRXA cluster (encompassing IRX1, IRX2 and IRX4 at a separate genomic location on chromosome 5), there has already been some enhancer analysis that suggests co-regulation of all three genes (Tena et al., 2011). Therefore, it might be that a similar pattern of distal cis-regulation operates at this obesity-associated locus.

Further evidence to support this comes from the analysis of topologically associated domain (TAD) structure in mammalian genomes. Data from embryonic stem cells identify a TAD of approximately 2 Mb that neatly encompasses the IRXB cluster, FTO and RPGRIP1L genes (chr16:53,562,500-55,442,500) (Dixon et al., 2012). Hence perturbation of the transcriptional architecture within this region during development could potentially impact upon any or all of these genes, and lead to an altered BMI phenotype. Finally, it is of note that this region contains hundreds of deeply conserved non-coding elements (CNEs), sequences implicated in the long-range cis-regulation of genes during development, including the IRX genes. Variants in such sequences might result in altered gene expression profiles across the region. Interestingly, the locations of CNEs at the IRXB cluster span from 53.56 to 55.48 Mb, in remarkably close agreement to the boundaries of the TAD (Dixon et al., 2012, Woolfe et al., 2007).

Here, using custom enrichment, the complete sequence of 284 Danish males homozygous at rs9939609 across the 2 Mb TAD region is generated and analysed. The resulting deep and comprehensive coverage allows identification of over 14,000 SNPs

and short indels permitting the precise and complete construction of haplotypes without the need for imputation. The use of homozygotes for the FTO LD region facilitates the downstream analysis of haplotypes. A novel association that implicates the IRX5 gene region in obesity is identified, and results are compared with previously derived interaction data for the region. These findings are replicated in an expanded male cohort and in a separate female study group using accurate imputation calls, identifying an age dependent association, consistent with previous studies (Graff et al., 2013, Hardy et al., 2010). This provides a high quality, single base resolution resource for further study into the complex genetics of obesity across human chromosome 16q12.2, and a general methodology for targeted sequencing and analysis of variation across large genomic regions in general.

3.2 Results

3.2.1 Strategy and study group

I employed a custom in-solution hybridisation approach to capture and completely sequence a 2 Mb region of chromosome 16 encompassing the RPGRIP1L, FTO and IRX3, 5 and 6 genes from 288 Danish men, previously genotyped as homozygous at rs9939609 (A/A or T/T) (Jess et al., 2008). The region (53.5 to 55.5 Mb) was specifically selected to encompass a TAD defined in embryonic stem cells (53.56–55.44 Mb) (Dixon et al., 2012). The study group comprises 126 cases with a BMI of ≥31.0 kg/m2 and 162 control samples (Appendix Table 1). They originate from two larger series of men selected from the study population of Danish men (n = 362,200) examined at mandatory draft board assessment during the years 1943 through 1977(Jess et al., 2008). The case set represents all men with a BMI ≥31.0 kg/m2 at initial assessment, corresponding to those above the approximately 99.5 percentile, whereas the control group consists of a randomly selected 1% of all men in the original study population and is thus representative of the underlying population's distribution of BMI values. The case group and half of the control group have been used in several follow-up studies including one in 1998–2000 where additional blood sampling allowed extraction of high quality DNA (Berentzen et al., 2008, Jess et al., 2008, Kring et al.,

2008, Paternoster et al., 2011, Zimmermann et al., 2009, Zimmermann et al., 2011). As a result of this sampling design, this study group has a bimodal distribution of BMI values and enrichment for homozygosity across the LD region encompassing the obesity-associating SNPs. The average BMI for the controls is 21.5 compared to 33.2 for the cases (Table 1).

Table 3. Study group details

	rs9939609 T/T (%)	**A/A (%)**	**Total (%)**	**Average BMI**	**Variance (95% CI)**	**SEM**
Controls. BMI<31 kg/m^2	106 (37.3)	55 (19.4)	161 (56.7)	21.5	±0.4	4.3
Cases. BMI≥31 kg/m^2	59 (20.8)	64 (22.5)	123 (43.3)	33.2	±0.5	5.6
Total	165 (58.1)	119 (41.9)	284 (100)	26.5	±0.7	7.1

BMI values are calculated from the original draft board assessment. The rs9939609:T > A (risk) allele was present in the study group at 41.9 %. In 1000 Genomes Project (1KG) data, both the Finnish (FIN) and British (GBR) allele frequency (AF) of the minor allele is 39.3 % (Genomes Project Consortium, 2012). Therefore, despite enrichment for homozygosity, there is a similar representation of the risk allele compared to the general population. The study group also maintains the relative proportions of T/T to A/A individuals (1.9:1 in controls and 0.9:1 in cases) found in the larger case and control group from which these individuals are derived (Jess et al., 2008).

3.2.2 Sequencing and variant calling

I used 96-plex indexing to construct custom libraries for 288 samples. This generated 1.66 billion paired end reads from these libraries for a total of 166Gb of sequence. Approximately 75% of reads map back uniquely to the 2 Mb region of interest

(chr16:53,500,000-55,500,000) giving an average of 4 million reads per sample (200-fold coverage). The sequencing identified one genotyping error (a genotyped A/A individual that was actually A/T), one sample failed to run, and two samples (1 case T/T, 1 control T/T) were of low coverage and had missing genotypes for more than 50% of variants. These were removed from subsequent analyses resulting in a final set of 284 samples (161 controls and 123 cases). As expected, coverage varied extensively both between samples and across the region. Nevertheless, 277 samples have greater than 10-fold coverage across at least 90% of the region allowing comprehensive, single base resolution analysis and unequivocal variant calling (Figure 8).

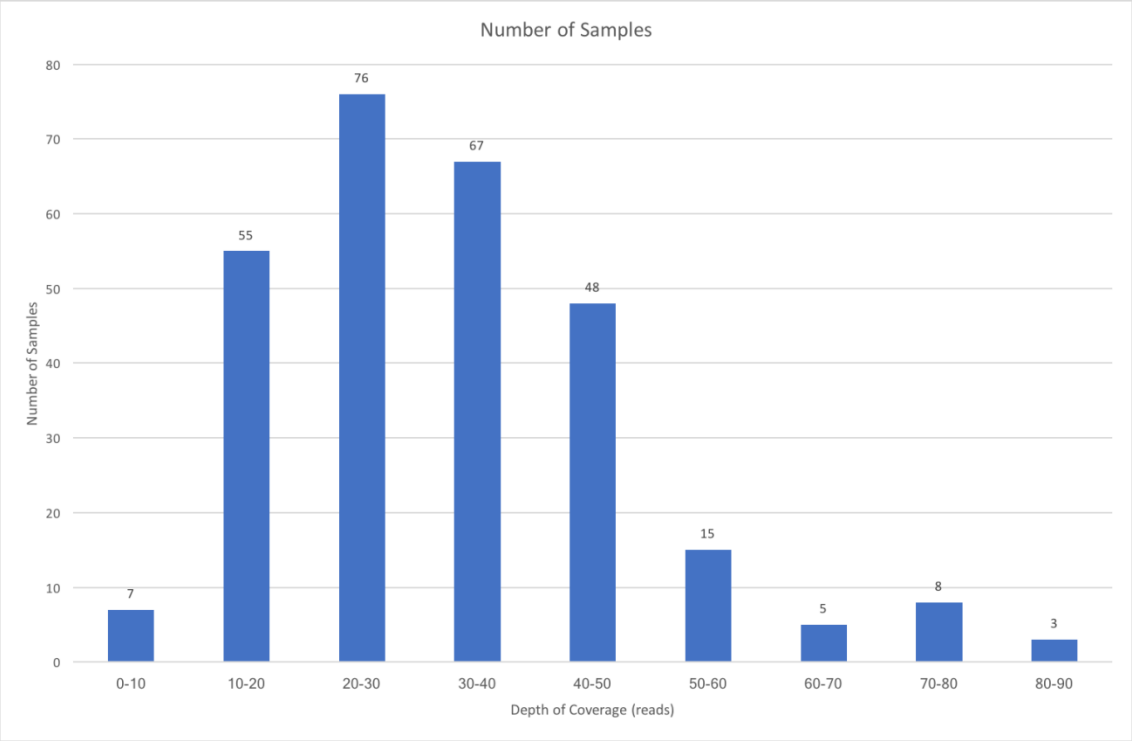


Figure 8. The number of samples where 90% of bases have the coverage of each bin value.

The following analysis utilised an in-house variant calling algorithm 'TidyVar' (methods – B. Noyvert and G. Elgar, manuscript in preparation). The algorithm is fundamentally different from that of commonly used variant calling software GATK (McKenna et al., 2010). TidyVar can be accurately deployed across any region of DNA of any size and from any species. Across the two-megabase interval, 14,101 variants passed quality control, of which 13,373 are simple (bi-allelic) and 728 are 'complex', in that they have more than one non-reference allele. Of the 13,373 simple variants, 12,392 are SNPs and 981 are indels. Fifty-nine percent of these variants are identically

catalogued in the phase 1 release of 1KG project data (Genomes Project Consortium, 2012) and 74% are identically catalogued in dbSNP build 142 (Sherry et al., 2001). On average, each individual has 2,869 variants across the region (ranging from 2,178 to 3,377).

I compared minor allele frequency (MAF) for those bi-allelic SNPs present in both the whole study group and the 1KG project (Figure 9). Reassuringly, the two datasets correlate very closely, demonstrating that despite selecting only homozygotes, the fact that I frequency matched rs9939609:T>A with the general European population results in a broadly representative set of variant frequencies. It is essential to use the comparative population from 1KG data as global allele frequencies are much more varied (Figure 10).

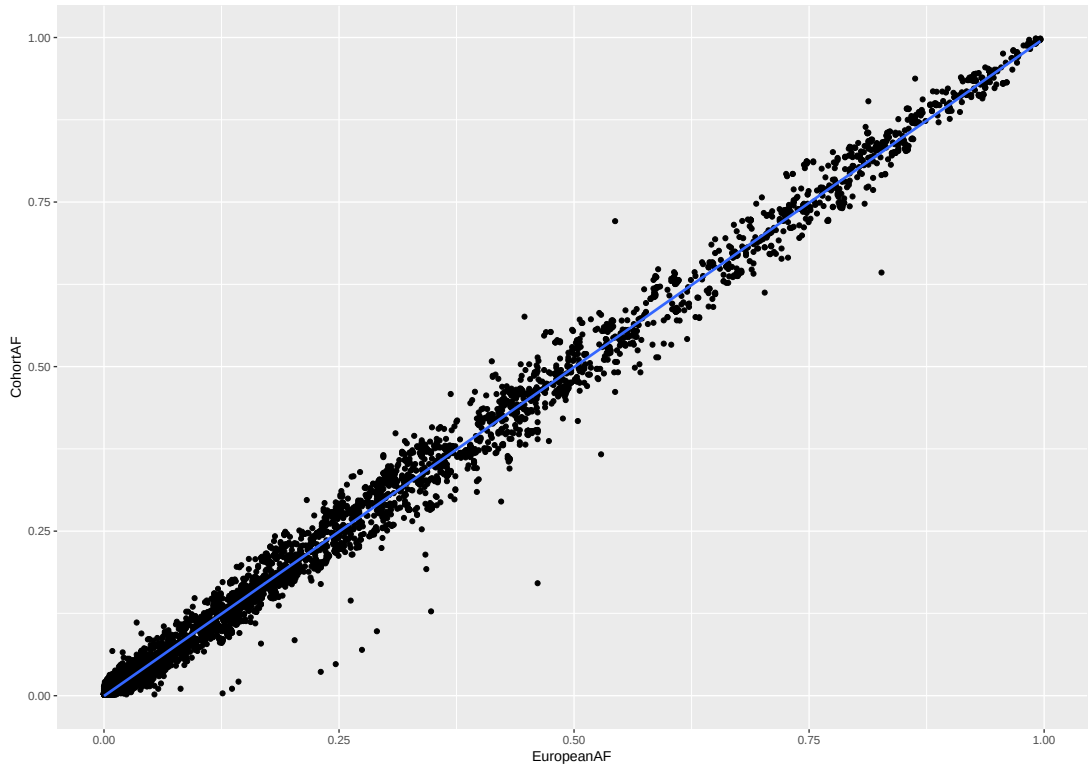


Figure 9. Variant frequencies across the 2 Mb interval.
The allele frequency of each variant in this study group is plotted against its frequency in European populations from the 1,000 Genomes Project. Only variants identified in both sets of data in the same format are directly compared (n = 9041).

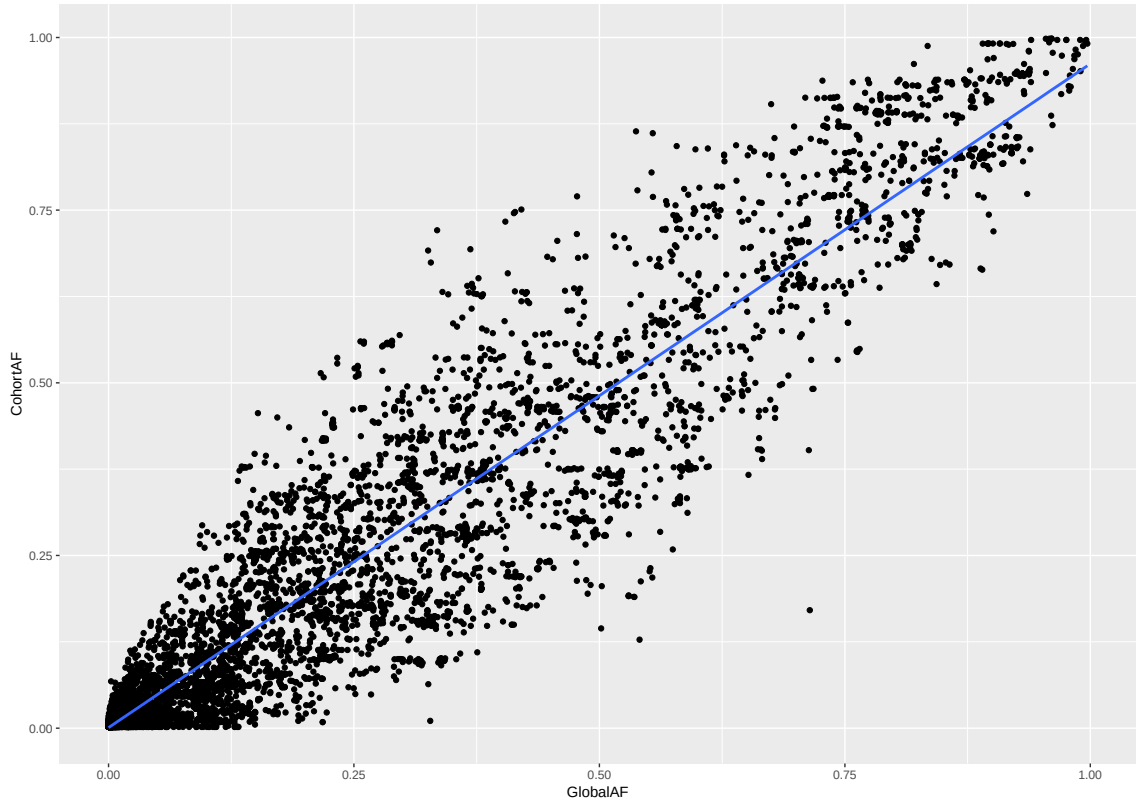


Figure 10. Global variant frequencies compared to cohort variant frequencies. Only variants identified in both sets of data in the same format are directly compared (n=9056),

3.2.3 Distribution of variants across constrained sequences

Within the 2 Mb interval sequenced in this study, 225 conserved CNEs are highly conserved between mammals and fish (CONDOR(Woolfe et al., 2007)) covering a total of more than 25kb. In addition, there is 17.2 kb of coding sequence across the region. I examined the number and distribution of SNPs in these different classes of constrained DNA (Table 4). As expected, there is a lower density of SNPs in coding sequences and to a lesser extent in CNEs, than in the remainder of the non-coding DNA across the region. SNPs in coding sequences and CNEs also have lower mean MAFs than general non-coding DNA, reflecting an excess of rare variants (Figure 11). These data reflect differing levels of functional constraint at these sites. The number of variants per individual does not differ significantly between cases and controls in any class of sequence.

Table 4. Variant summary data for chr16q12.2 classified by functional region and BMI status

Region	Size of region (kb)	Number of variant locations	Variant locations (per Kb)	Mean MAF	Av. number variants per individual	Av. number non-ref alleles per individual
CNEs	25.1	Cases: 117	4.66	0.118	20.03	27.52
		Controls: 109	4.34	0.127	19.94	27.60
		Total: 141	***5.61***	*0.098*	*19.98*	*27.57*
Coding	17.2	Cases: 38	2.21	0.075	4.61	5.37
		Controls: 56	3.26	0.051	4.75	5.40
		Total: 70	***4.07***	*0.041*	*4.69*	*5.39*
Non-Coding	1,957.7	Cases: 11014	5.51	0.181	2853	3916
		Controls: 11826	5.91	0.167	2836	3892
		Total: 13980	***6.94***	*0.142*	*2843*	*3902*

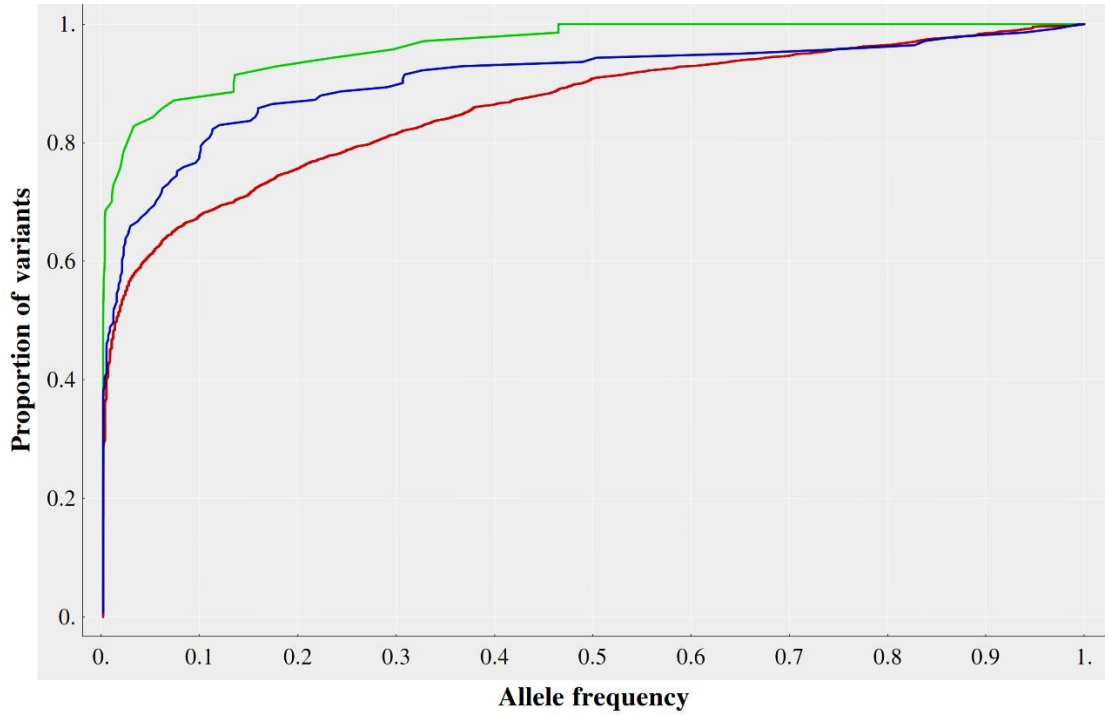


Figure 11 Cumulative frequency distribution of variants.

Blue: CNEs, Green: Coding regions, Red: All variants across the region combined.

3.2.4 Haplotype analysis

Haplotype analysis of the entire region (using pairwise comparison of SNPs up to 500kb apart) permits the identification of blocks with high LD (Figure 12), the most notable of which is the previously identified 44 kb region (A: chr16:53,799,296-53,843,533) in the first intron of the FTO gene containing rs9939609.

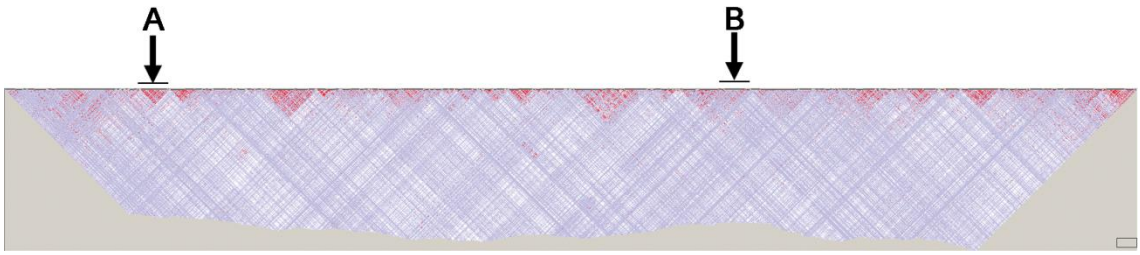


Figure 12. Full LD mountain plot of chr16:53500000-555500000 sequenced and exported from Haploview.
A: region 53.8Mb-53.85Mb. B: region 54.81Mb-54.87Mb. These regions also correspond to Figure 14 and Figure 15 respectively.

Three distinct haplotypes persist across this interval and comprise 63.5% of all haplotypes across the region. The first two (29.3% and 12.5%) differ by just one SNP (rs113191842:A>G) and account for all the rs9939609 A/A individuals (known henceforth as haplotype AH44). While the more common of these two haplotypes strongly associates with the obesity case group (Figure 13) as expected ($P = 1 \times 10^{-4}$), the second does not (P = 0.383) although this might simply reflect a lack of statistical power due to its low frequency in the study group overall. The third common haplotype (21.7%) is found only in T/T individuals, but does not show a significant association with either case or control (P = 0.086) group. Appendix Table 2 describes all the other haplotype blocks with associations to the case or control group with a frequency of >0.05. Due to the number of individuals sequenced in this study, I have focused on the two regions showing the clearest and most strongly associating variants. There are several other LD blocks containing haplotypes that also associate with either case or control outside of these regions. For this, further sequencing would be needed to establish any association of these blocks in the general population.

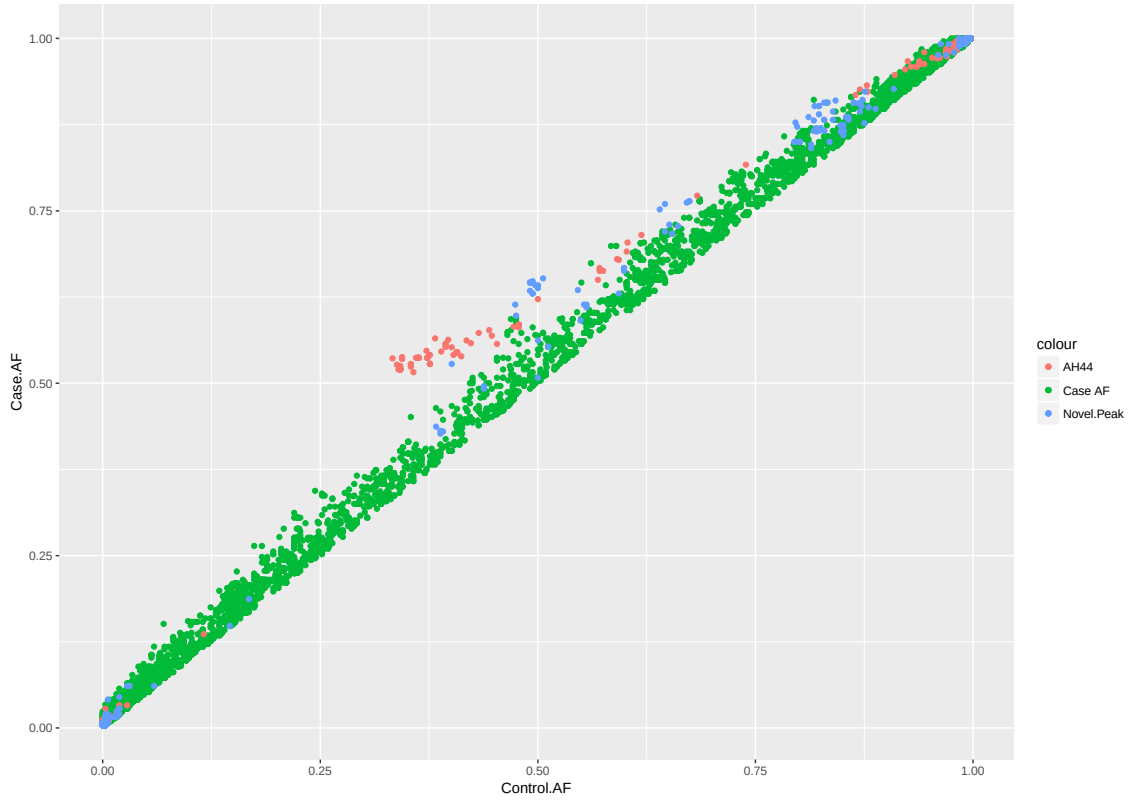


Figure 13. Minor allele frequencies for each variant across the 2 Mb interval compared between controls and cases.

Variants within the AH44 LD block are in red and variants in the second association region upstream of IRX5 are in blue. This graph is laid in perspective of the obese cohort so variant positions with alleles associating with obesity can be seen clearly. In this instance, the further above the x=y line the greater the frequency of the allele in the case cohort.

3.2.5 The AH44 haplotype

The obesity-associated haplotype, along with its almost identical sub-haplotype (referred to collectively henceforth as AH44) has a clear and distinct pattern of variation across its length when compared to other haplotypes for the same region. This haplotype encompasses many of the obesity-associated SNPs that have been identified by various GWAS studies (Berndt et al., 2013). From this in-depth analysis, 114/122 highly polymorphic SNPs (MAF >0.35) spanning 53,798,523 to 53,848,561 (50.038 kb) are in complete LD with rs9939609 in all A/A individuals in this study group. Of the remaining 8/122 SNPs, seven are uniquely heterozygous in the same individual and the final SNP also occurs just once. While these common SNPs are essentially in complete

LD across the 44 kb AH44 haplotype, there are a number of rarer variants across the region that are not in LD, indicating that while the common variants are retained, there are in fact numerous sub-haplotypes that contribute to AH44. The most frequent of these rarer alleles that separates the two AH44 haplotypes, rs113191842:A>G at 53,817,318 (just over 3 kb from rs9939609:T>A), is only present in AH44 but occurs at a frequency of 0.28 within this population. A further 26 non-unique and 21 unique variants are present within the AH44 haplotype individuals, while just one of these (rs16952522:C>G) is shared with the non-AH44 haplotype in T/T individuals.

3.2.6 Identification of a novel region associated with BMI in this study group

I used Haploview (Barrett et al., 2005) to compare the frequency of every SNP across the 2Mb interval between cases and controls and to calculate the case-control allelic association P values (Fig. 3 and methods). The known LD region in the first intron of FTO (chr16:53,797,908-53,846,168) is clearly defined (Figure 14).

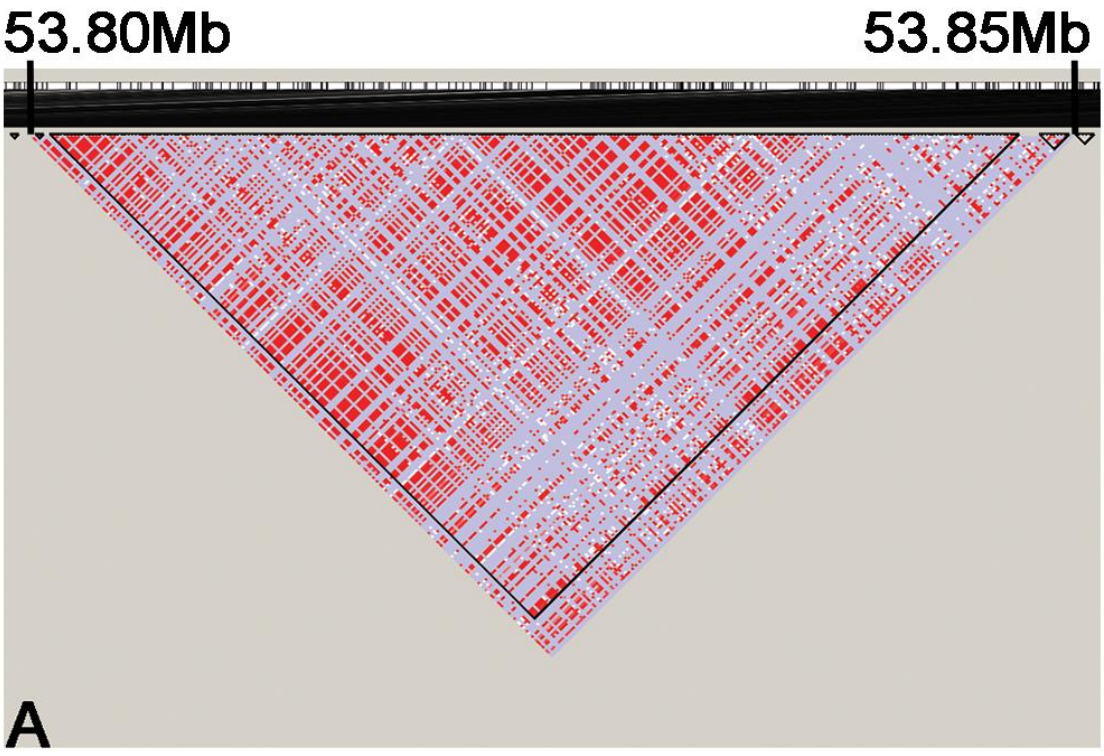


Figure 14. AH44 LD block region exported from Haploview.

In addition, there is a second peak of association approximately 1Mb away that consists of a cluster of SNPs upstream of the IRX5 gene (16:54,820,000-54,860,000). Critically, this case specific association is independent of risk allele rs9939609 and random shuffling of the cases and controls results in loss of any comparable signal across the region. The non-coding region encompasses four linkage disequilibrium blocks, the largest of which is 38 kb in size (Figure 15). In addition, PLINK conditional regression testing using GATK shows independence of association between the two haplotypes when tested against each other.

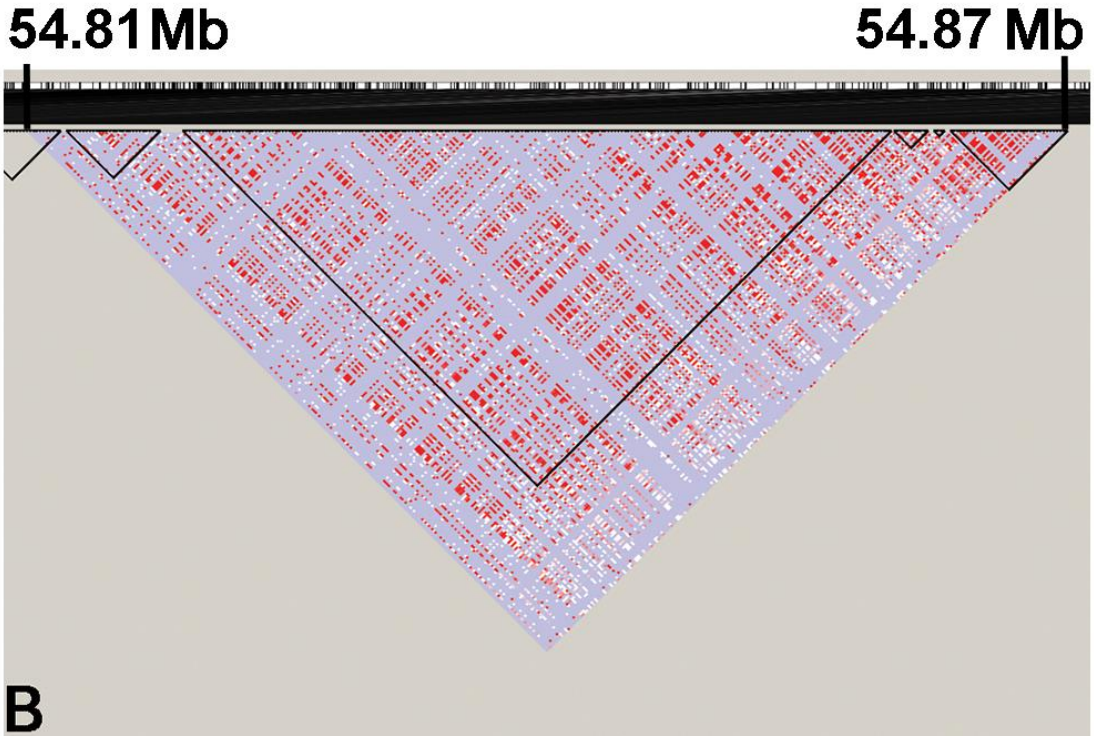


Figure 15. Novel association peak region exported from haploview

Within this LD block, there are several haplotypes identified by Haploview. The strongest obesity associating haplotype (P value =0.002) occurs at a frequency of 0.49 in the case group and 0.36 in controls. No other haplotype in this LD block has a total frequency above 0.13 suggesting that the associating haplotype is more robust in its entirety than the multiple non-associating haplotypes. The 38kb associating region encompasses 213 SNPs that have been identified in the sequencing, 78 of which are tagged by Haploview for use in haplotyping. Thirty-five SNPs within this LD block have an association P value <0.05. The lowest 10% of P values for SNPs in this region

(n=21) range from 0.0002 to 0.0073. The three highest associating SNPs in this second peak (P=0.0002) are in complete LD within 7 kb of each other in all but two individuals. These SNPs (rs7186407:A>T, rs12598453:C>G and rs12596270:A>G, hg19 coordinates chr16:54837068, 54843731, 54843981, respectively) are present at a frequency of 0.491 in controls and 0.646 in cases (0.56 in whole study group), whereas they have a wide range of derived allele frequencies in different populations in the 1KG Project, with values as low as 0.0225 to 0.036 in Japanese and Chinese populations, to greater than 0.5 in all European populations. Individuals in the study group who have neither risk region allele rs9939609:T>A (from known region) nor rs12598453:C > G from the novel association region have a mean BMI of 23.86, whereas individuals homozygous for either risk region have significantly higher mean BMIs of 27.96 (Mann-Whitney P value = 0.0062) and 27.60 (P value = 0.0088), respectively (28.90 if homozygous for both (P value = 0.00067)). Thus, both regions have a similar association with BMI.

3.2.7 Multiple testing correction

The P values presented in the previous section are not corrected for multiple testing. A naïve Bonferroni correction for 14000 variants would give a P value threshold significance of 3.5×10^{-6} (=0.05/14000) when controlling the family wise error rate (FWER) at the 5% level. Although since the variants are not independent the above correction is overly conservative. Indeed, variants belonging to the same linkage disequilibrium blocks have a strong positive correlation (consider that I identify multiple associating variants across both the known, and this novel, regions). It is therefore more appropriate to use the number of LD blocks (n = 226) identified by Haploview to estimate the corrected P value threshold. This then becomes 0.05/226 = 2.2×10^{-4}. Only the known LD region in the first intron of FTO and the second peak of association identified above pass this threshold (Figure 16), guiding the focus on these two regions only. However it is fair to consider these significant p-values as suggestive of association as the initial threshold is not met.

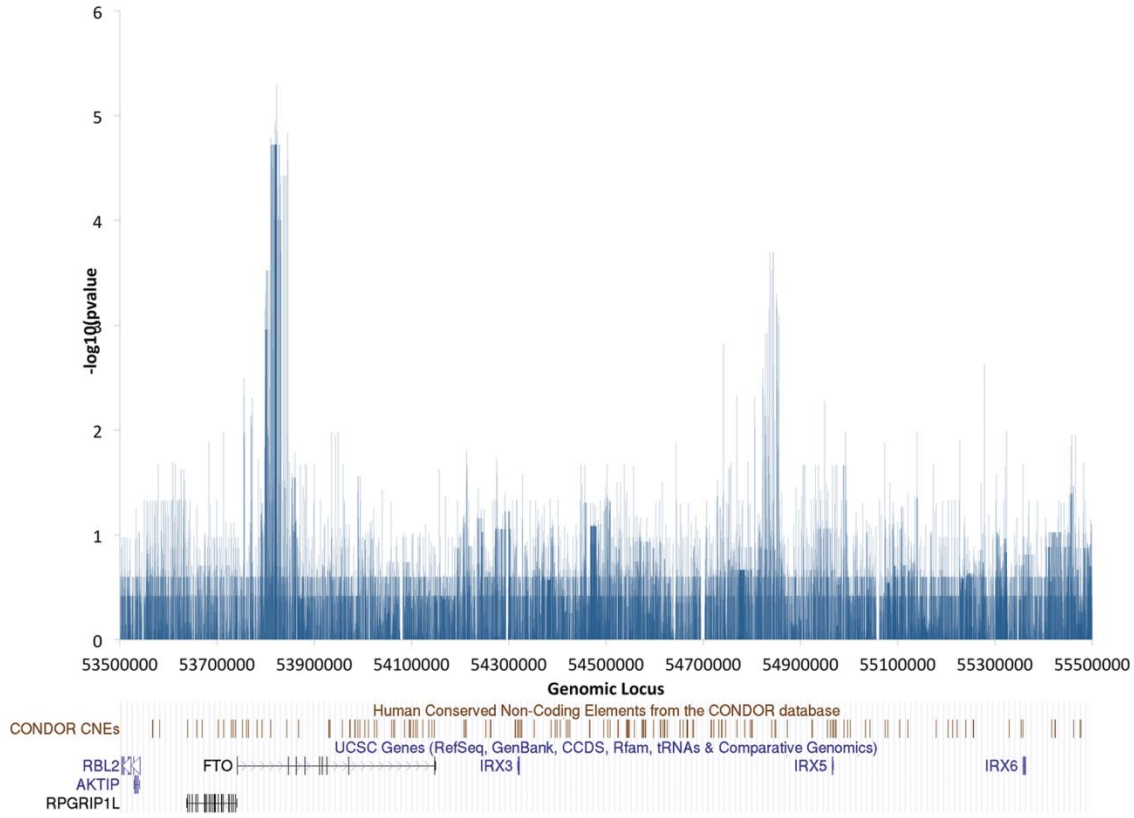


Figure 16. Association of individual SNPs to cases v controls.
Minus log10(P-value) of Case/Control association for each SNP across the 2Mb interval generated from Haploview (Barrett et al., 2005) and represented by vertical blue lines. The first peak (at 53.82 Mb) in the intron of FTO shows the known association at rs9939609:T > A and reflects the strong linkage disequilibrium across that region. The second peak (54.84 Mb) indicates a novel associated region upstream of IRX5

Since there is no exact definition of an LD block the above multiple testing correction by the number of LD blocks may be underestimated. This is why I decided to control for FWER by permuting the set of obese and control labels. This was achieved by a 100,000-permutation test in Haploview for the full set of sequenced variants across the 2Mb in this cohort of 284 men. The individual SNPs in the second peak of association have corrected P values >0.05 and therefore do not pass multiple testing correction. This is a reflection of the limited sample size and paradoxically the vast number of variants I identified through complete sequencing of the region. Therefore, it was essential that I replicate these findings in other cohorts.

3.2.8 Replications

In order to validate these findings, I replicated the case-control association tests in two larger cohorts (Table 5). The first (Male GOYA) comprises 1,450 men from the expanded cohort that this sequenced study group was initially selected from (Paternoster et al., 2011). The expanded group has imputed SNP data for the three highest associating SNPs (rs7186407:A>T, rs12598453:C>G and rs12596270:A>G) as well as for rs9939609:T>A. The three highest associating SNPs, which are in near perfect LD, were chosen to be representative of the second novel peak of association. I found that in this larger group of young men, all three representative SNPs also associate with the case group, with a P value of 0.0054 (Table 5).

In addition, I replicated the association analysis of the three representative SNPs in a large female Danish cohort (Female GOYA (Paternoster et al., 2011, Nohr et al., 2009)). I initially looked at the entire cohort of 3,908 women (1,960 extremely overweight and 1,948 control women, total average age of 29.5). In this group, using imputed data for the three representative SNPs, I cannot confidently replicate the second association peak (P values >0.05, Table 5). However, in light of previous studies that suggest genetic association to obesity at the FTO locus may be age-dependent (Graff et al., 2013, Hardy et al., 2010, Jess et al., 2008) and because of the lower, narrower age range in the male cohort (mean = 19.9), I examined the role of age in this age-diverse female cohort (range from 16 to 45). I found that the allele frequency of the three SNPs is consistently higher in cases compared with controls only for women aged under 25 years (Figure 17).

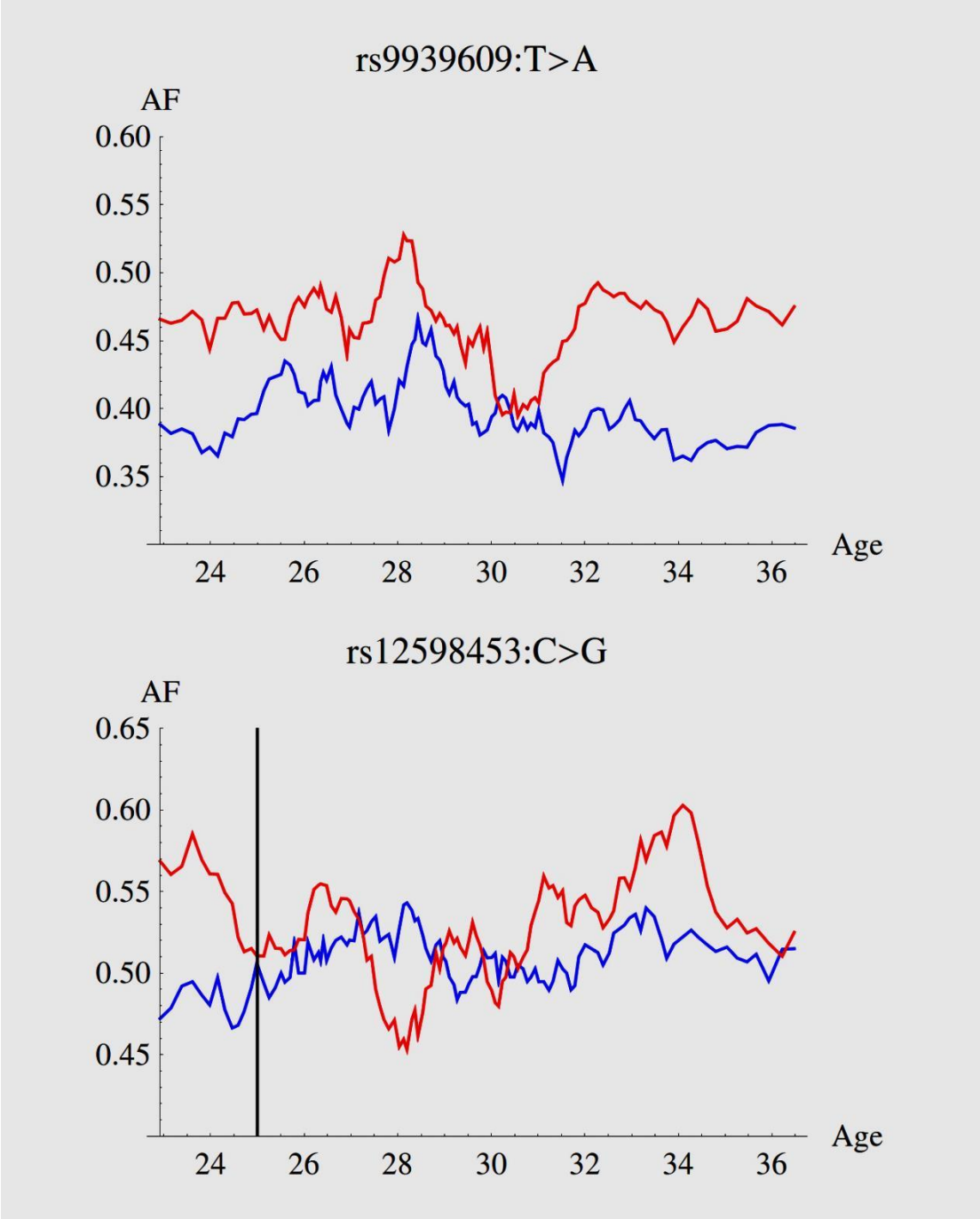


Figure 17. Allele frequencies by age in Female GOYA cohort.

The case allele frequency (AF) is shown in red, the control AF is shown in blue. The allele frequencies are calculated in groups of 400 individuals of consecutive age. Whilst the case AF is consistently larger than control AF for rs9939609:T>A across essentially the entire age range, the consistent AF difference for rs12598453:C>G is only observed in younger (up to approximately 25 years old) females. (Figure prepared by B.Noyvert)

In this smaller group of 562 individuals the three SNPs show suggestive association with obesity (P value = 0.0014, Table 5). This is consistent both with previous studies at the FTO locus (Graff et al., 2013) and with the value found in the larger Male GOYA cohort. If I consider all individuals aged less than 25 years in all cohort groups then the P value for the association of the novel peak I found is 1×10^{-5} (Figure 18) confirming that the second novel peak of association can be replicated independently in larger cohorts of the same ethnic background and similar age, regardless of gender.

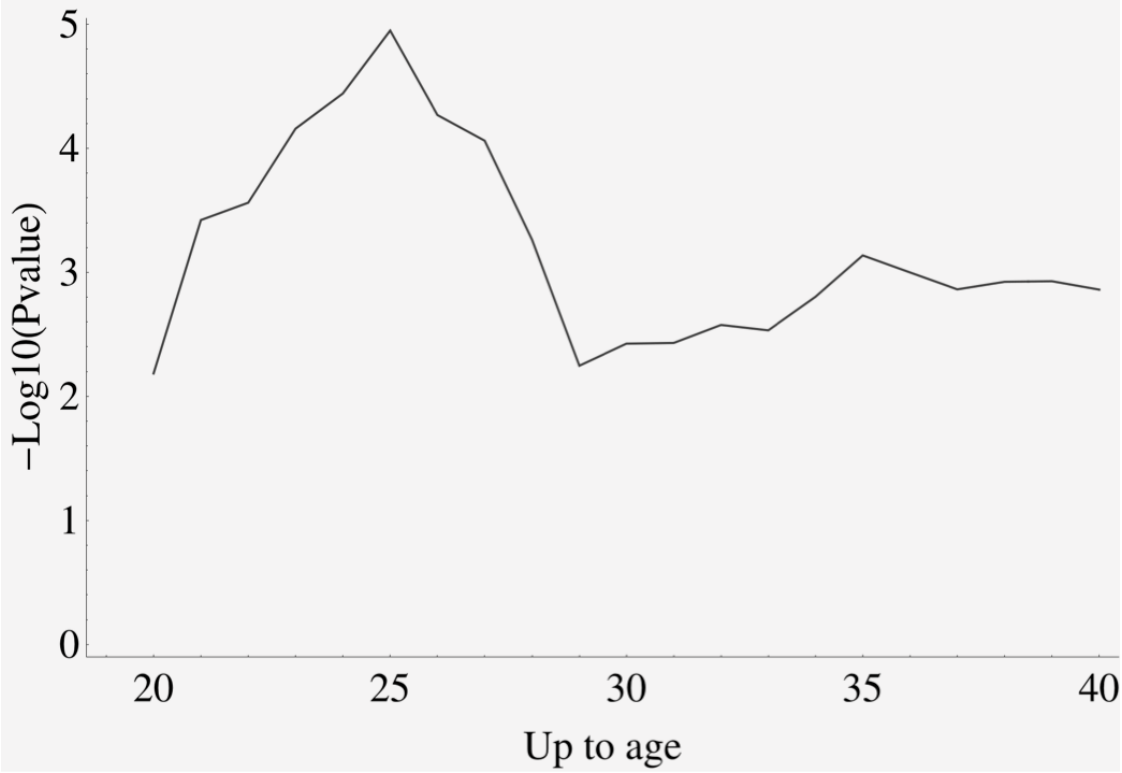


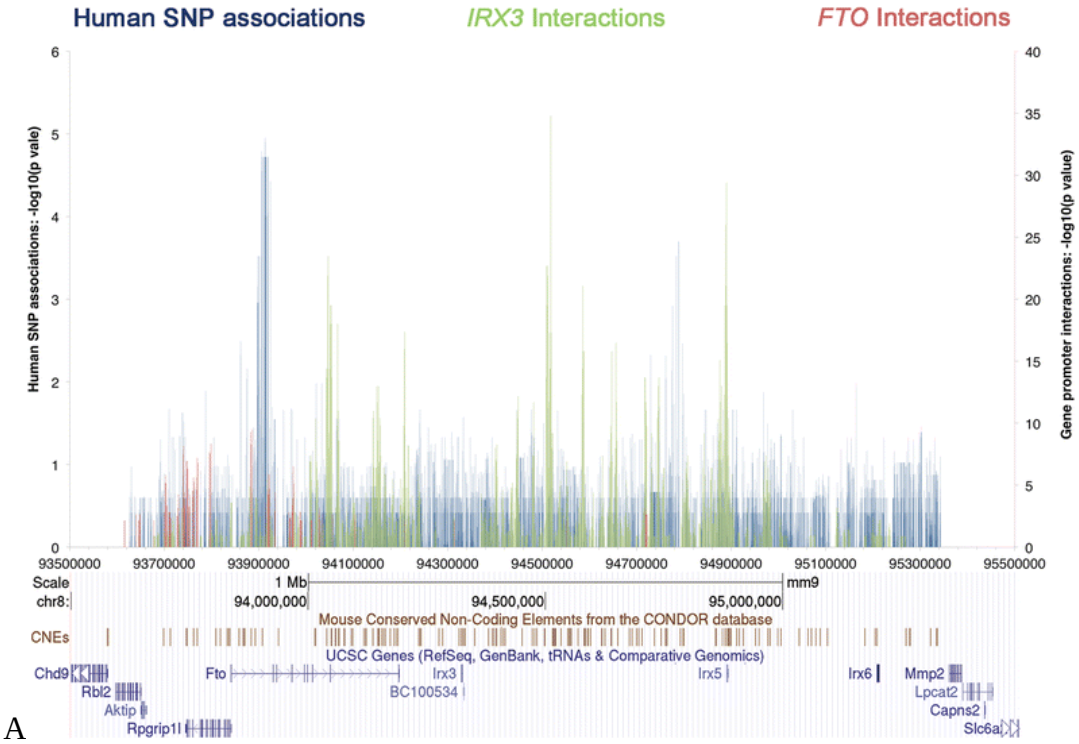
Figure 18 Age dependence of SNP rs12598453:C>G association to obesity.
Each point on the plot represents the association p-value (on y-axes) for a subgroup of combined GOYA male and female cohort younger than a certain age (on x-axes). (Figure prepared by B.Noyvert)

Table 5. Replication data using SNP rs12598453:C>G as a representative of the three SNPs referred to in the text

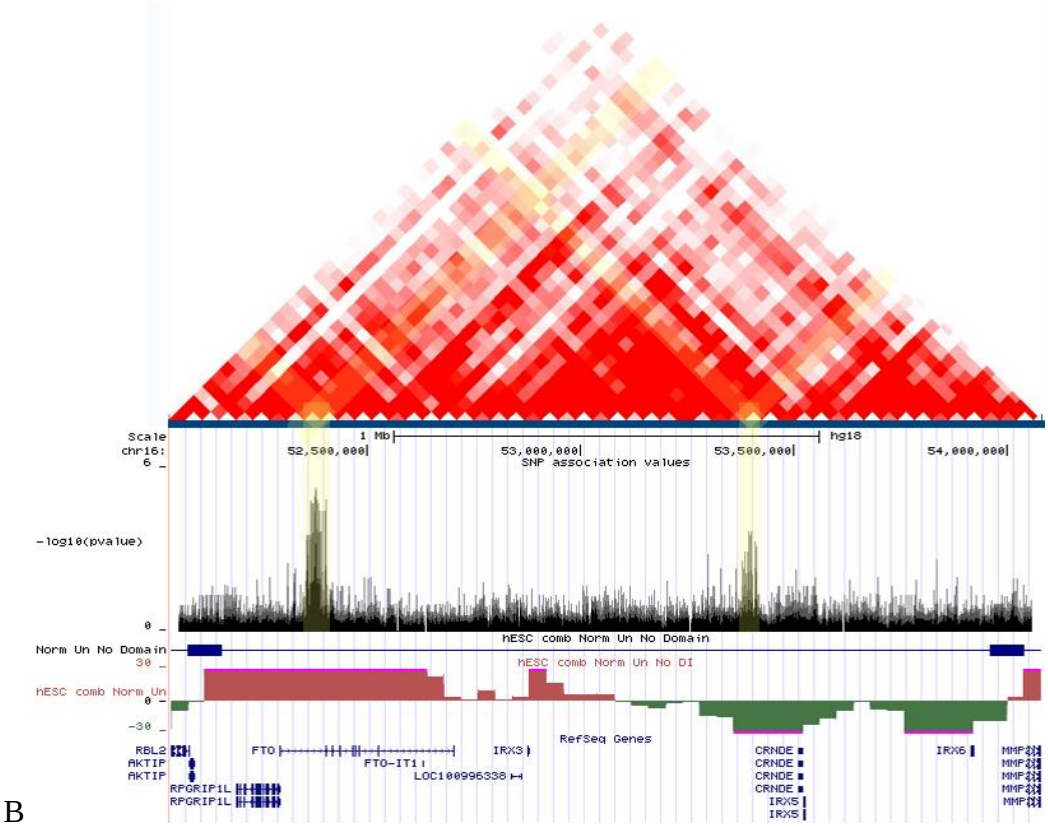
Cohort	N	Number of controls, cases	G AF in controls, cases	P-value, case-control allelic chi-squared test	BMI averages by genotype: CC CG GG
Sequenced males	284	161, 123	0.491, 0.646	0.00021	25.2 26.03 28.13
Male GOYA	1450	785, 665	0.496, 0.547	0.0054	26.51 26.89 27.39
GOYA males, younger than 25	1381	749, 632	0.493, 0.551	0.0027	26.45 26.85 27.47
Female GOYA	3908	1948, 1960	0.507, 0.529	0.056	30.00 30.21 30.46
GOYA females, younger than 25	562	255, 307	0.465, 0.560	0.0014	29.79 30.67 32.5
All combined	5401	2762, 2639	0.503, 0.534	0.0012	29.04 29.24 29.63
All combined, younger than 25	1984	1032, 952	0.486, 0.556	0.000011	27.35 27.85 28.89

3.2.9 IRX3 interactions extend beyond both BMI associated regions

Recently, long-range interactions have been experimentally defined across most of the 16q12.2 region for the FTO and IRX3 genes using chromatin conformation analysis (Smemo et al., 2014). Comparing the locations of these interactions with those of BMI-associated SNPs might help determine both a mechanism and a role for the SNP regions in the cis-regulation of the FTO or IRX3 genes. Figure 19A shows that while neither the FTO nor the IRX3 promoter-based 4Cseq data correlate strongly with the associated regions, both associating regions are within the long-range interaction architecture of IRX3, with particularly strong interactions (both with FTO and IRX3) flanking the associated region upstream of IRX5. Hi-C data from human embryonic stem cells also provides strong evidence that the novel association region upstream of IRX5 plays a role in many interactions across the TAD (Figure 19B), including with the IRX3 and FTO gene regions ((Dixon et al., 2012) and http://yuelab.org/hi-c/). Further 4C-seq analyses of non-coding association regions will contribute to understanding which genes or other non-coding regions of DNA these SNPs might be interacting with, and whether the presence of this variation changes these interaction profiles.



A



B

Figure 19. Comparison of the SNP association data with previously published 4C-seq (Smemo et al., 2016) and Hi-C data (Dixon et al., 2012). The two significant SNP association peaks lie within interacting domains within the previously defined TAD but don't appear to strongly correlate with FTO or IRX3 promoter interactions.

3.2.10 Functional predictions for the novel BMI associated region

Using publicly available data, I compared the novel BMI associated region upstream of IRX5 with gene regulatory markers and functional annotations. This includes (but is not limited to) the presence of CNEs, epigenetic marks and interaction data (HiC). Selecting a relevant cell line is a caveat of this approach, as the exact contribution of this genomic region to BMI is not fully understood. A recent study suggests the contribution of variation at the FTO locus affects adipocyte lipid accumulation through increased IRX3 and IRX5 expression (Claussnitzer et al., 2015). Within the novel region there are three CNEs (Figure 20). These highly constrained regions are strong indicators of regulatory function. One of these CNEs contains, and is surrounded by, a cluster of conserved transcription factor binding sites (HMR Conserved Transcription Factor Binding Sites).

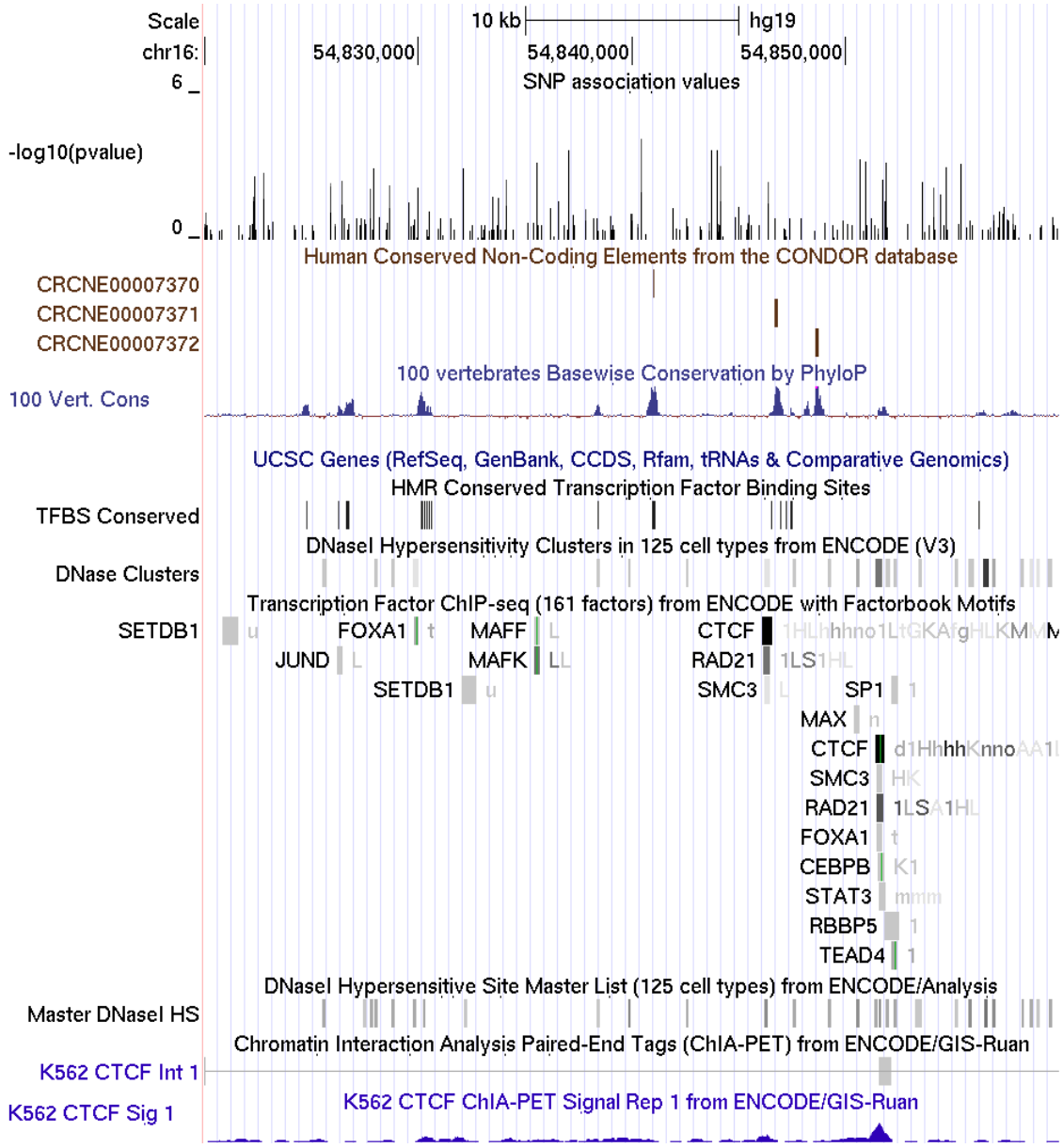


Figure 20. UCSC browser figure of the second association peak region (54820000-54860000)

In addition, ENCODE Genome Institute of Singapore ChiA-PET data show interactions in the second association peak highlighted region for RNAPII and CTCF long-range binding in two different cell lines (K562 myelogenous leukaemia cells and MCF-7 breast cancer cells). CTCF is thought to be a transcriptional regulator (Ong and Corces, 2014) and therefore the presence of long-range CTCF-mediated binding in this region suggests a potential role in either repression or activation through DNA looping. The presence of RNAPII mediated looping can also be indicative of enhancer activity in

the region. These long-range interactions across the TAD are supported by previous Hi-C data across the whole 2Mb and suggest that additional regulatory regions might contribute to the gene expression of IRX3 and IRX5 (Figure 21). Interestingly, I was unable to find any positive interaction data between the novel BMI association region and the FTO gene (or any other genes within the TAD) in the current literature.

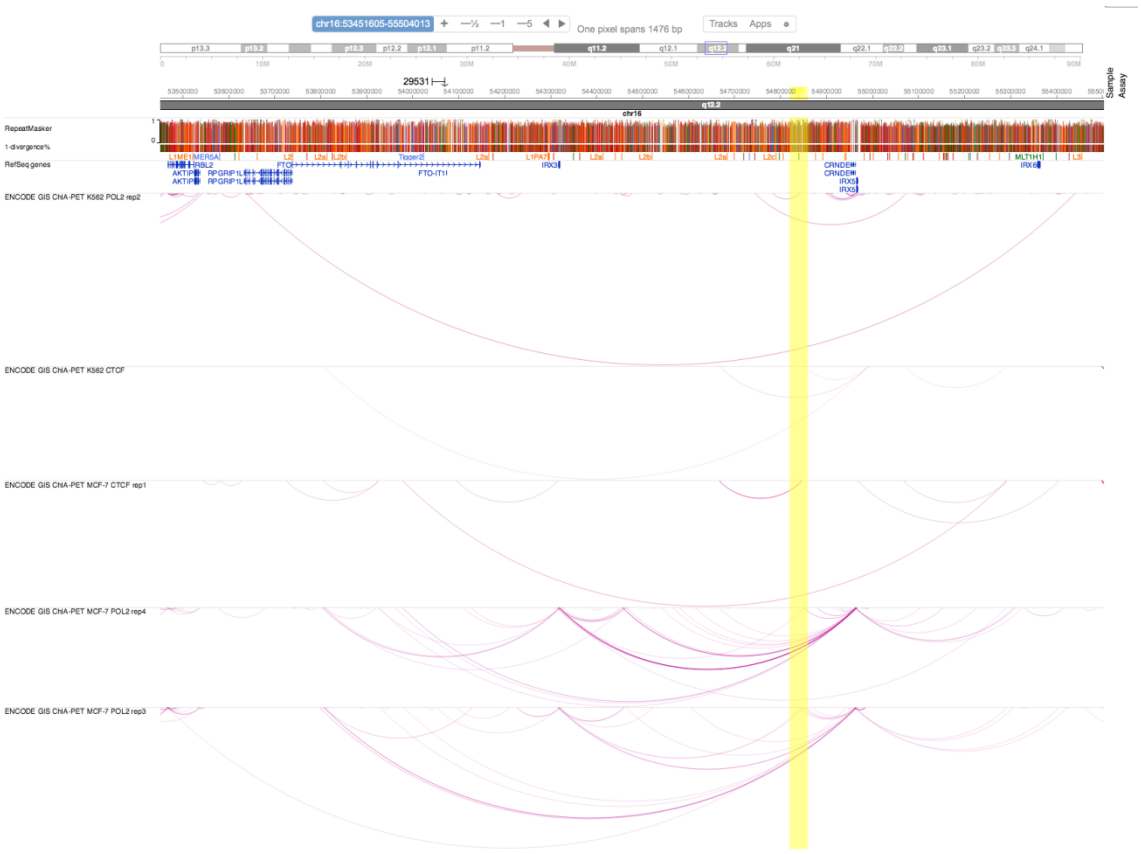


Figure 21 WashU epigenome browser figure.
The entire region sequenced is shown. Highlighted in yellow is the second novel peak of association identified.

3.3 Discussion

Recently, the selective sequencing of regions of the human genome has been achieved using hybridisation capture approaches. This has largely been exploited to sequence the coding, or exome, portions of the genome. However, the same capture approach can also be adapted to select any regions from the genome (Tewhey et al., 2009). Here,

unusually, I have employed it to capture a contiguous megabase scale region of the human genome. The 2Mb interval was selected using 8,701 probes at intervals of approximately 200bp. Since this work, this ethnically homogenous population and targeted sequencing approach has also been used to identify obesity associating variants in a Polish population (Sobalska-Kwapis et al., 2017). Seventy-two gaps in the sequence, largely repetitive and covering a total of 30.7 kb (1.5% of total), were not traversable. Between 70% and 80% of all reads map back to the region of interest in these 284 study group samples providing good coverage across 98.5% of the interval, making the capture approach considerably more efficient than whole genome sequencing. Furthermore, downstream analysis is considerably simpler and less time-consuming. As a result, I was able to generate very high coverage, (average 200-fold) which in practice means that almost every individual variant can be called with very high confidence.

At the outset only those individuals that are homozygous at rs9939609 were sequenced. This allowed very high-resolution mapping of haplotypes, particularly across the 44 kb LD region associated with this SNP. One of the aims was to determine whether, within the Danish study group I sequenced, there were any low frequency variants that contributed a significant effect within this region and would therefore allow further dissection of the association with BMI. There are multiple independent obesity association signals across this region so determining the co-occurrence or co-dependence of these would help define sub-haplotypes. It was also easier to determine whether the association of other variants across the 2 Mb interval with BMI was linked to, or independent of, the 44kb LD region. Consistent with other studies (Gabriel et al., 2002), particularly in European populations, the 44kb region is in almost complete LD. Incredibly, of 282 SNPs mapped across the 44kb, only one (rs16952522:C > G at 53,807,498) is found in common between the rs9939609 ‘A’ and ‘T’ alleles (MAF 0.045 in cases, 0.037 in controls). This implies that at least in the Danish population, recombination events in this region are historically exceptionally rare. The small size of the study group means that there is not the statistical power to evaluate whether any of the rs9939609 ‘A’ risk allele sub-haplotypes or rare variants are more associated with obesity than others.

Analysis of the constrained sequences within the region confirms that there are no coding variants nor any frequently occurring variants in highly conserved noncoding elements (CNEs) that are associated with elevated BMI. Functional constraint does have an effect on both the frequency of variant locations (4.1, 5.6 and 7.0 per kb, respectively, for coding, CNE and non-coding sequence) and the minor allele frequency of variants (85% of coding, 80% of CNE and 70% of non-coding variants have MAF<0.1).

Within the Danish male study group, I clearly identify a second, novel region associated with BMI in noncoding sequence upstream of the IRX5 gene. Individuals in this study group who are homozygous for this second region have a mean BMI elevated to a similar extent to the effect of the known FTO intron region variants. This association is independent of the FTO LD region as it is not present if A/A vs. T/T individuals are compared. This analysis was performed using data obtained at Danish draft board assessment which results in a very homogeneous study group, not only in terms of gender and ethnicity but also because all participants were of similar age (average age 19.9 years) when their BMI was measured. Interestingly, this association is strongest in younger sub-groups of the replication cohorts as well, suggesting an age-dependence aspect.

To address this idea, I first utilised imputed values for the three most highly associating variants upstream of IRX5 in the expanded male cohort, comprising 1,450 individuals, to confirm the association. Next, I used imputed values for the same three SNPs in a completely independent female Danish study group, comprising nearly 4,000 individuals. When the entire cohort is used with a higher average age, the association is not clear but in women aged under 26 years the association can be replicated. Thus, if the male and female cohorts are matched by age (as far as possible), there remains a significant association between BMI and the region upstream of IRX5.

I then searched for the 22 highest associating SNPs across the second peak of association in the GIANT consortium BMI based anthropometric data for European

populations (Locke et al., 2015). Of these 22 SNPs, 13 are included in the GIANT dataset, with over 200,000 individuals having data for these variants, yet none of these SNPs are found to have a significant association with obesity. It is impossible to discern the reason for the lack of an association in the GIANT consortium data without secondary analyses. It is unlikely that a population specific variation and association with this common variant in an outbred population from North West Europe would be missed by meta-analyses that are also biased to North West European populations. This may suggest that other factors are confounding this association in meta-analyses, potentially age or gender in this instance. Nevertheless, in these meta-analyses the association is lost as age increases, perhaps because environmental factors such as diet and levels of physical activity are likely to have an increasing impact on BMI with age, confounding the detection of some genetic associations. Conversely, if the genetic consequences of this association are established early in life, such as during development, then it is likely that a stronger association will be seen at a younger age. Given that this locus is intimately associated with complex developmental transcription factors, this would seem highly likely and reflect the life course data at the neighbouring FTO gene (Graff et al., 2013).

The IRX genes, including IRX3 and IRX5, play complex and overlapping developmental roles in multiple tissues and organs (Gaborit et al., 2012, Houweling et al., 2001). There is also evidence that both IRX clusters form complex interactions that define specific three-dimensional structures that regulate gene expression at different loci (Peters et al., 2000, Tena et al., 2011). In particular, it has been shown that the IRX3 promoter region interacts with a number of distal sites across the 2Mb region (Smemo et al., 2014) sequenced here and defined by the embryonic stem cell line TAD described previously (Dixon et al., 2012).

In order to examine this in detail, I first looked at the overlap between the interaction data for FTO and IRX3 genes and the association data across the region. As the interaction data are from mouse, I lifted the human data for the region over to the syntenic region on mouse chromosome 8. There is no strong correlation between IRX3 (or FTO) interactions and either of the BMI-associated regions although there is some

IRX3 and FTO signal across the 44kb region. Nevertheless, the fact that long-range IRX3 interactions occur up to and beyond both associated regions suggests architecture is an important aspect of gene regulation across the whole region. This is supported by Hi-C interaction data across the TAD from human embryonic stem cells (Dixon et al., 2012), which clearly show strong interactions between the FTO and IRX5 gene regions. It will therefore be important to establish the specific interaction domains of IRX5 and IRX6 in order to get a fuller picture of the complex structure of this region and to be able to place the associated regions into a fuller context.

Despite many GWAS studies (Berndt et al., 2013) and now the full sequencing of this region from a well-defined study group, there remains considerable difficulty in predicting, describing or functionally assaying the impact of non-coding variants on disease or phenotype. As a result, a number of the genes across this region have been implicated in obesity yet without any clear mechanism of regulatory control (Tung et al., 2014, Yeo, 2014). This region is particularly complex because of the presence of the IRXB cluster, a set of homologous genes that regulate many aspects of early development and are thus under tight regulatory control themselves. This control is likely to be mediated via cis-regulatory sequences that in some cases may be hundreds of kb away and even within the non-coding regions of other genes, as has been demonstrated for other genes such as Shh (Goode et al., 2005, Lettice et al., 2003).

The implication of more than one gene in the aetiology of obesity at this locus may therefore not be so surprising, neither is the identification of a second cluster of associating SNPs. The structural architecture(s) of this particular topologically associated domain (TAD) may have profound effects on the regulation of all the genes in the region at some stage, but at this juncture not enough is known about how sequence variation may alter chromatin architecture nor what the consequences might be in terms of gene expression. Nevertheless, as more insight is gained into the structure and function of the non-coding DNA in this TAD, the complete sequence of the 2Mb interval from this study group will provide a valuable resource. Furthermore, targeted region sequencing may be of great utility in examining other such complex regions in fine detail in the future.

Chapter 4. Conserved non-coding element sequencing elucidates novel mutations in regulatory regions with predicted functional consequences

4.1 Background

Conserved non-coding elements (CNEs) are small stretches of sequence that do not code for proteins and yet are highly conserved in many vertebrate organisms as defined by fish-mammal alignments (Elgar and Vavouri, 2008). Specifically, vertebrates have been shown to have many CNEs through sequence alignment against the pufferfish, Fugu (Aparicio et al., 1995, Aparicio et al., 2002), a teleost with a compact genome devoid of much of the highly repetitive sequences found between and within genes in higher organisms (methods of comparisons are reviewed in (Boffelli et al., 2004)). It is through comparing alignments of these CNEs closely that we can also define the bases within them as non-variable or restricted variable regions (De Silva et al., 2014). Completely evolutionary conserved sites are non-variable regions and those sites with at least one nucleotide substitution across any of six divergent vertebrate species (macaque, mouse, chicken, frog, zebrafish and fugu) are restricted variable regions. Due to the extreme conservation of these regions over time and evolutionary distance it is likely that they correspond to some function as otherwise we would expect them to have diverged between species through random mutation.

On further inspection of these CNEs it has been shown that at least some are enriched with specific predicted transcription factor binding sites (Parker et al., 2011) (although the increased frequency of some is matched with a decreased frequency of others). Therefore, it is suggested and widely believed that these CNEs act as regulatory elements, activating or repressing the expression of nearby genes (Featherstone, 2003, Howard and Davidson, 2004, Strahle and Rastegar, 2008). Furthermore, CNEs are seen to cluster around developmental genes, specifically those involved with head and neural development in vertebrates (Woolfe et al., 2005), suggesting they play a key role in these vertebrate morphogenetic characteristics.

CNEs by their definition have a fewer fixed mutation over evolutionary time than other areas of the genome. However, within CNEs there can be single bases that are non-variable or have restricted variance. It could be suggested that mutations in CNEs may have similar effects to those in coding sequence. Many CNEs combine to form 165 clusters and presumably control only a fraction of vertebrate specific genes (Woolfe et al., 2005), therefore there may also be some redundancy between CNEs and an accumulation of mutations may be necessary to produce a similar effect to a single coding mutation. This could occur through a low penetrance model of each CNE mutation and the resulting accumulation creating differences in gene expression. As we are yet to understand the exact grammar of these non-coding elements it is difficult to infer what a single base change at a single position may result in, however a few studies have associated phenotypes with CNE variation and mutation (Antonellis et al., 2006, Loots et al., 2005, Lettice et al., 2003, Attanasio et al., 2013). Approximately 90% of GWAS markers are found in the non-coding regions of the genome (Maurano et al., 2012). Much of the subsequent focus of these studies has looked at the effect of mutations on nearby genes, making the assumption that these GWAS variants affect gene function. This bypasses an important feature of gene regulation - how SNPs directly affect gene expression.

Many of the published examples of non-coding variants affecting long-range gene regulation have been reviewed previously (Bhatia and Kleinjan, 2014) however current methods to assess the function of a SNP or short InDel in non-coding DNA are either high-throughput and not developmentally applicable or vice versa. Therefore, the identification of non-coding causative variants is many years behind the successes of identifying coding variants and their roles in Mendelian disorders. The approach used here combines association studies and family pedigrees with functional prediction of non-coding variants in an attempt to prioritise those implicated in developmental diseases and disorders. Here I present targeted sequencing of CNEs in a number of differing cohorts as a way of elucidating functional SNPs and InDels amongst the vast non-coding genome. I go on to predict some level of functionality both through computational tools and assaying these regions for enhancer activity in the developing Zebrafish embryo using known methods (Fisher et al., 2006, Kwan et al., 2007).

Utilising the same probe set for all cohorts, CNE targeted sequencing creates a large database of non-coding variation at a read depth that matches or exceeds many exome projects allowing accurate SNPsand InDel calling in these potential regulatory regions. This will allow assessment of the extent of human variation in these regions, including meta-analysis of non-variable and restricted-variable bases within these highly conserved regions (De Silva et al., 2014), as well as create candidate lists of variants that could play a role in these developmental disorders.

4.1.1 Cohort descriptions

Table 6. Cohorts used in this chapter for CNE sequencing

Disease	Total Samples	Patients	Controls	Ethnicity	Notes
Intellectual Disability & Epilepsy	96	32	64	Danish	Used primarily to develop prioritisation pipeline
Cleft Lip and Palate	192	64	128	Dutch	
Isolated Congenital Anosmia	20	14	7	Faroese & European	Focus on pedigree tracking
Schizophrenia	265	265	0	Pakistani	Followed with Zebrafish assay

4.1.1.1 Intellectual disability and epilepsy comorbidity

Epilepsy and intellectual disability comorbidities are a common occurrence with prevalence of epilepsy ranging from 20-30% in individuals with intellectual disability (Bowley and Kerr, 2000) compared to approximately 1% of the European population having epilepsy (Forsgren et al., 2005). Epilepsy is a neurological condition and 16% of individuals with the condition also have been reported to have some level of intellectual disability (Morgan et al., 2003), a proportion that is much higher than the overall

population prevalence of <1% (Westerinen et al., 2014). The two disorders may also co-occur with many other characteristics and phenotypes, including being symptoms of known genetic disorders such as microdeletions and rearrangements (Stevenson et al., 2012) and can have a strong hereditary component (Myers and Mefford, 2015, Flint, 2001). Both are known to sometimes occur as the result of complications during pregnancy, including resulting from Fetal Alcohol Spectrum Disorder (Sumner et al., 2013). Together, this suggests a key role in developmental genetics and its misregulation in neuronal embryogenesis as a potential cause of epilepsy and intellectual disability co-occurrences.

As both disorders can have a range of severity and co-occurrences with other disorders, studying their genetic causes can be difficult. Here, 96 individuals from various familial backgrounds (7.5) with at least one child with intellectual disability and epilepsy undergo targeted conserved noncoding elements sequencing in an attempt to find vertebrate regulatory region variants that could be contributing to the clinical phenotype. The parents of 32 affected children are mostly unaffected but 17 have some diagnosed or subclinical phenotype noted by the medical examiner.

4.1.1.2 Cleft lip and cleft palate (CL±P) comorbidity

Orofacial clefts represent a heterogeneous group of defects with varying ranges of severity. Cleft lip can occur with or without cleft palate but the two are more commonly diagnosed together and incidences of cleft lip and palate are around 6.64 per 10,000 births worldwide (Group, 2011). Incidences of mortality in developed countries are relatively low (Kang et al., 2012) (but still elevated compared to those born without) the incidence of infant mortality in developing countries is significantly elevated (Bickler and Rode, 2002) and all affected children impose a substantial financial burden on healthcare (Wehby and Cassell, 2010). Twin studies have provided compelling evidence for a genetic component to CL±P (Fraser, 1970, Grosen et al., 2011) although there is also a large amount of evidence for environmental risk factors (Mossey et al., 2009) which leads to possibilities of medical interventions. These would be particularly useful if genetic risk factors could first be identified.

The occurrence of cleft lip and/or palate in humans is a result of the failure of the palate to fuse. Normal development (Figure 22) begins by the fourth week of human embryonic development around the oral cavity (a) with the nasal pits forming by the fifth (b). This leads to formation of the paired medial and lateral nasal processes (c). These in turn (c), by the sixth week, form the nasal alae. In addition, the medial nasal processes merge with the maxillary processes to form the upper lip and primary palate. The secondary palate then develops as bilateral outgrowths from the maxillary processes, which grow vertically down the side of the tongue (d). Afterwards, the palatal shelves elevate to a horizontal position above the tongue, contact one another and commence fusion (e). Fusion of the palatal shelves ultimately divides the oronasal space into separate oral and nasal cavities (f) and the failure of this as results in cleft lip and/or palate.

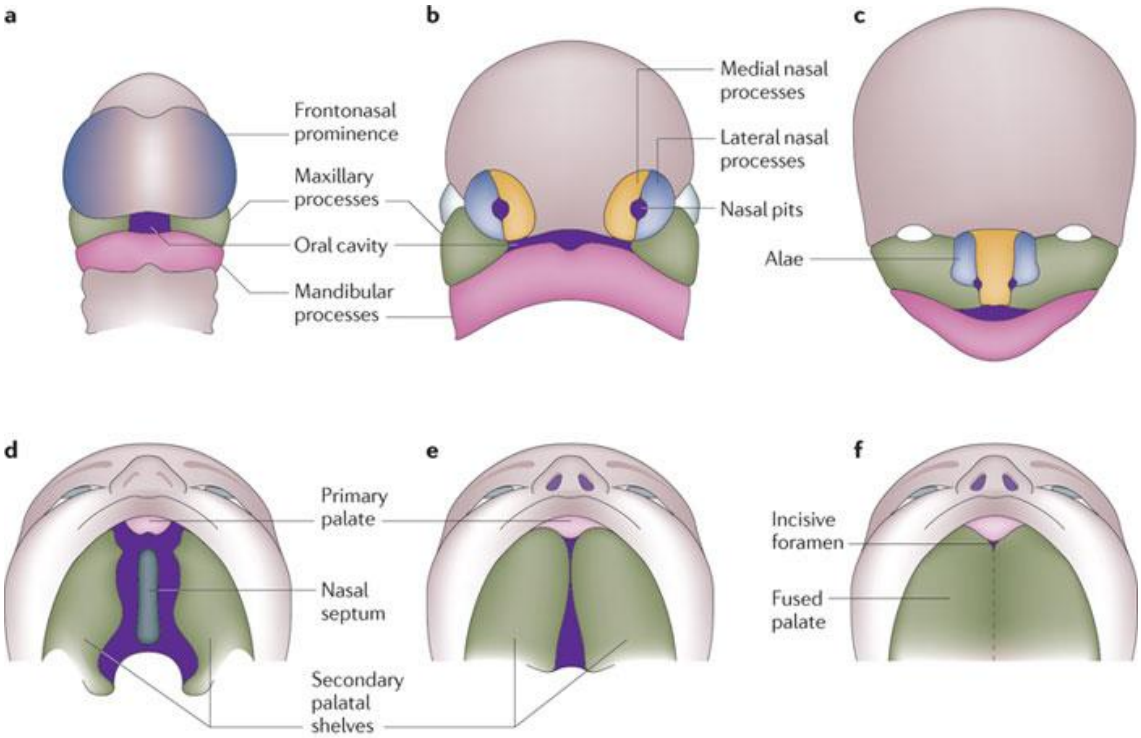


Figure 22. Development of the lip and palate in humans
Adapted by permission from Macmillan Publishers Ltd: [Nature reviews. Genetics] (Dixon et al., 2011), copyright (2011)

Environmental risk factors for CL±P stem largely from maternal exposure to a variety of elements. Strong evidence supports maternal smoking during pregnancy with a

consistent increased risk of both cleft lip with or without cleft palate and isolated cleft palate (Lammer et al., 2004, Little et al., 2004, Shi, 2007, Shi et al., 2008, van Rooij, 2001). Meta-analysis has also shown that maternal use of multivitamin supplements in early pregnancy was associated with a 25% reduction in birth prevalence of orofacial clefts (Johnson and Little, 2008). Supplements of folic acid are used to prevent the failure of the neural tube to close in development, yet despite interest in the use of folic acid to also prevent CL±P it was shown that multivitamins may give some protection, but not folic acid alone (Johnson and Little, 2008). Folic acid deficiency does cause cleft palate in rats (Asling et al., 1960) and antagonists of folic acid confers higher risk of orofacial clefts in people. This suggests a role for folic acid, but a multi-faceted approach to orofacial cleft susceptibility. What is clear is that the early embryonic development and its environment is essential to the correct formation of the palate and lip.

As CL±P is aetiologically heterogeneous, understanding the genetic influences will allow further understanding how environmental risks interact and hope to give rise to interventions. Non-syndromic CL±P is a particularly complex disorder, and the genetic variation is likely to occur in regulatory elements. It has been shown previously that long-range regulatory elements add control to craniofacial development (Attanasio et al., 2013) and although multiple mutated genes have been associated to various clefting syndromes (Ardinger, 1989, Beaty, 2010, Birnbaum, 2009, Grant, 2009, Mangold, 2010, Marazita, 2009, Pauws, 2009, Suzuki, 2009, Wehby and Cassell, 2010, Zucchero, 2004) the complexity of the traits leaves much to still be explored. One key finding with nonsyndromic cleft lip and palate is the discovery already of an accumulation of conserved noncoding variants in and around the FGF and FGFR genes that may contribute to clefting (Riley and Murray, 2007). This suggests that exploring these conserved noncoding regulatory elements for variation could contribute to the genetic understanding of CLP. In addition, functional noncoding variation that affects enhancer activity near the gene *NOG* associates with CL±P (Leslie et al., 2015) further suggesting a targeted sequencing approach of predicted developmental enhancers could discover novel contributing regulatory variants.

The cohort used in this work are 63 trios of Dutch origin where the child has been born with cleft lip and palate and neither parents are affected. These samples are from the EuroCran project (eurocran.org) and full phenotype data is attached (Appendix Table 3).

4.1.1.3 Anosmia

Isolated congenital anosmia (OMIM#107200) is the inability to smell from birth. Estimates are that 5% of the population are anosmic (Brämerson et al., 2004, Landis et al., 2004) with many of these cases a result of age, poor diet and other poor quality of life factors (Nordin and Brämerson, 2008). Congenital anosmia is also a presenting symptom of a variety of sexual and developmental abnormalities (Vowles et al., 1997) including Kallmann syndrome (Lieblich et al., 1982). Interestingly there has been evidence of family members of patients with Kallmann syndrome having isolated congenital anosmia (Pitteloud et al., 2006). This suggests a complex genetic component to the disorder, and cases of isolated congenital anosmia (without the presence of trauma or other environmental factors) is much rarer, comprising ~1% of the anosmic population (Pitteloud et al., 2006, Ciofalo et al., 2006). Although the lack sense of smell may not present as many societal and psychological difficulties for patients as blindness or deafness, its origins in neural development and genetic component could help shed light on the complex regulatory processes surrounding the formation of the vertebrate head.

Familial presentation of ICA has driven previous investigations into the underlying genetic components of the disorder. Through pedigree tracing, ICA has been observed and followed in a handful of families (Lygonis, 1969, Singh et al., 1970). This has provided some evidence towards a dominant inheritance. However, one large family in the Faroe Islands (with 28 patients having ICA across 4 generations) also provided some evidence to an alternative mode of inheritance. The isolated nature of this population has meant that otherwise recessive rare mutations may be found in much higher frequencies. The varying nature of family history makes ICA hard to study, both in the incomplete family records and the non-uniform presentation of inheritance. Generally, autosomal dominant inheritance patterns are seen but reduced penetrance is

sometimes observed with some family members presenting isolated congenital hyposmia instead (Mainland, 1945, Ghadami et al., 2004a, Singh et al., 1970, Feldmesser et al., 2006, Leopold et al., 1992).

Despite the strong evidence suggesting a genetic inheritance to ICA, few efforts to identify the causative mutations have been published. One study mapped ICA to locus 18p11.23-q12.2 using genome-wide linkage analysis of two unrelated Iranian families (Ghadami et al., 2004b). This was a small powered study with 7 patients presenting ICA as autosomal dominant with incomplete penetrance. The study then sequenced the exons and exon-intron boundaries of 8 candidate genes in this vast 30Mb region but found no mutations. A separate study looked for ICH susceptibility loci genome wide in North Americans and mapped to a 45cM region on chromosome 4 (Pinto et al., 2008).

It has been suggested that non-coding variation plays a role in reduced penetrance disorders (Ward and Kellis, 2012). Studies have used eQTL data to show a relationship between deleterious coding variants and regulatory variants that help adjust their penetrance (Lappalainen et al., 2011). Widespread potential for interactions between coding variants and regulatory variants (Montgomery et al., 2011), especially those within cis-regulatory modules, could help explain some of the reduced penetrance phenotypes seen in isolated congenital anosmia. This work looks at multiple families with some history of ICA (and 2 individuals with ICH) whose exomes have not revealed any causal variants for their phenotypes (7.5, Figure 44, Figure 45, Table 24). Therefore, without attempting to sequence and investigate the entire genome, sequencing of the CNEs reveals variants that may impact developmental gene regulation and play a part in the anosmia phenotype.

4.1.1.4 Schizophrenia

Schizophrenia (OMIM#181500) affects approximately 24 million people worldwide and has a prevalence of around 7 per 1000 adults (Saraceno and Bertolote, 2013). It is a brain disease that presents as a variety of symptoms with differing severity between individuals. It is widely diagnosed using the PANSS (positive and negative syndrome

scale) (Kay et al., 1987, Leucht et al., 2005) alongside identification of psychotic episodes, and can be treated to an extent with medication. Schizophrenia is a worldwide healthcare problem, with high morbidity, mortality (Saha et al., 2007) and societal costs (Mathers, 2008) (Knapp et al., 2004). Both diagnosis and treatment of Schizophrenia could relieve some of these societal costs and improve patient outcomes by preventing relapses (Almond et al., 2004, Hong et al., 2009). The development of Schizophrenia in patients can vary widely both in age of onset and severity, with some differences between gender (Sham et al., 1994). Despite Schizophrenia presenting in a variety of ways, all available antipsychotic drugs used as treatment exert their main therapeutic effects through blocking the type 2 dopamine receptor (Lahti et al., 2003) (Carlsson and Carlsson, 2006). Pharmacological treatments for Schizophrenia are typically low in efficacy for many patients (Leucht et al., 2013). Therefore, identifying the causes of Schizophrenia is essential for the development of new treatments and the lessening of this disease's burden on the patients and global healthcare systems.

It is believed that Schizophrenia has a significant genetic component with MZ twin concordance data showing around 48% heredity (Onstad et al., 1991). Although many studies including GWAS point to susceptibility loci, no specific gene has been seen to cause the varying symptoms of the disorder (Bray et al., 2005, Chowdari et al., 2002, Emamian et al., 2004, Jolly et al., 2013, O'Donovan et al., 2008, Riley et al., 2010, Stefansson et al., 2009, Williams et al., 2011). Recently a large GWAS study of over 36,000 cases and 113,000 controls has pointed to 108 conservatively defined loci of significant association (Schizophrenia working group of the Psychiatric Genomics Consortium, 2014). This, in addition to other research suggests that Schizophrenia is not only a complex phenotype but also polygenic in nature (Internatioal Schizophrenia Consortium, 2009). This can make pinpointing the hereditary predisposition to a particular loci, gene or variant particularly confounding. One potential solution offered here to sift through the noise is to utilise ethnically homogenous populations in association studies.

Many genome-wide association studies have found single nucleotide polymorphisms that associate with Schizophrenia in the non-coding regions of the genome

(Schizophrenia working group of the Psychiatric Genomics Consortium, 2014). The notion of polygenic contributions and large environmental influence causing the first symptoms of Schizophrenia to occur fits a gene mis-regulation model. This has been shown to happen with structural variants (Walsh et al., 2008) and as a result epigenetic changes from the early developmental environment (Dolinoy et al., 2007). Therefore, some genetic association and the underlying predisposition could come from non-coding variation and its effects on gene regulation. Data has already shown that there is an enrichment in Schizophrenia eQTL associated variants in enhancer and promoter regions (Roussos et al., 2014, Schizophrenia working group of the Psychiatric Genomics Consortium, 2014). Identification of contributing SNPs in respect to regions of high LD will only be possible with further functional analysis including enhancer assays, Hi-C information and potentially CRISPR-Cas9 genetic modification. In addition, with Schizophrenia being such a complex disease, *in vitro* methods are reliant on finding a suitable cell line which is unlikely to be feasible given the nature of neural development, networks and heterogeneity of the brain. New risk loci are being identified continually however work needs to be done to correlate these genetic identifiers with gene regulatory elements, the genes they're affecting and the role of these genes in neuronal development.

Schizophrenia is known to be a disorder resulting from altered neurodevelopmental processes (and environmental influence), the predisposition to which is set during brain development in the womb (Haijma et al., 2012, Nenadic et al., 2012, Honea et al., 2005, Collin et al., 2013, Gogtay et al., 2011).The development of a complex brain is vertebrate specific, and regulation of genes in this development must be both precise and exact. There is also the variation of gene expression and regulation across the multiple developing brain regions (Buonocore et al., 2010). Therefore, a large cohort of Schizophrenia patients from Pakistan, a disorder thought to be related to brain development (Rapoport et al., 2005, Pantelis et al., 2005) presents an opportunity to discover regulatory variation that could contribute to this disease. Using the targeted sequencing approach, vertebrate evolutionary constrained sequences are utilised as a method of identifying enhancers involved in neural development (McEwen et al., 2009).

Here, I sequence the conserved non-coding elements of 265 Schizophrenia patients of Pakistani origin and analyse disease-associating variants for potential function.

4.2 Results

4.2.1 In-solution capture probe hybridisation of CNEs produces high coverage and quality non-coding sequencing data suitable for rare variant analysis

The CNE probe regions span 769,894 bases with the distribution of bases within these regions being representative of the whole genome. Within the 1000 Genomes publicly available data across all ethnicities, there are 20,521 variant locations called within these targeted CNEs; 2.7% of bases within CNEs have annotated human variation in current publicly available datasets. The vast majority of these variants are rare (Figure 23) due to the evolutionary constraints on these sequences used in their selection.

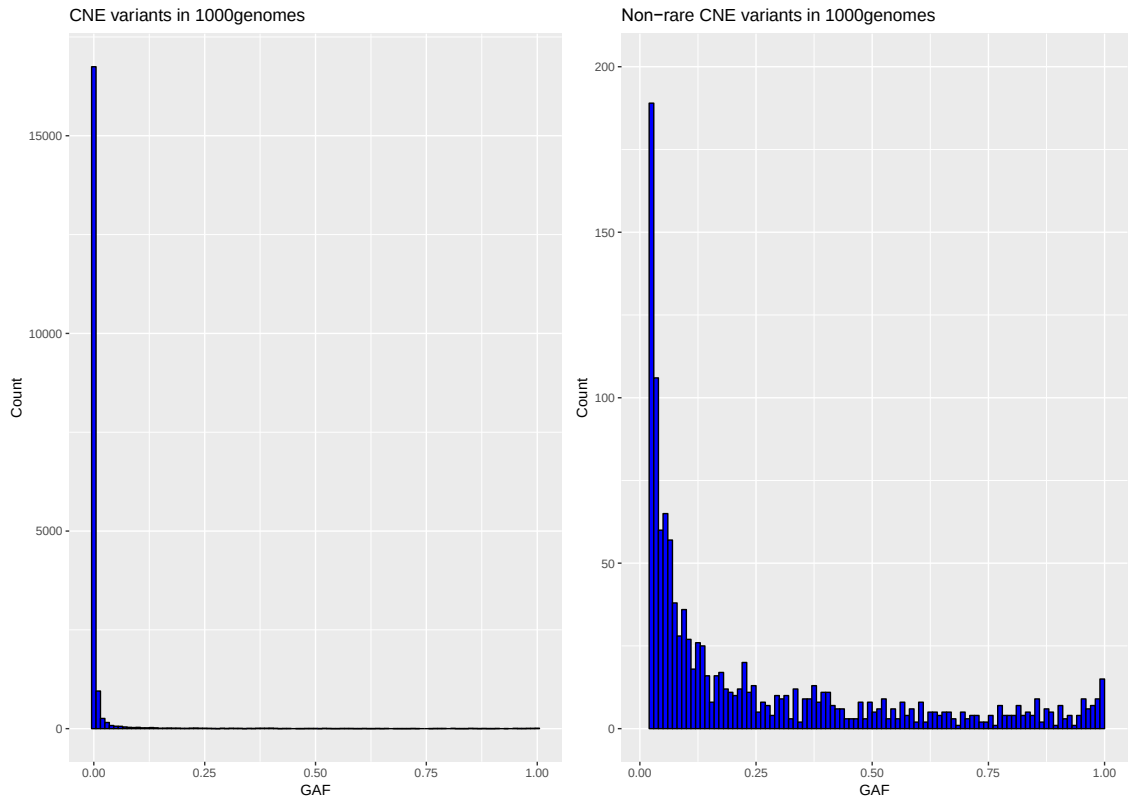


Figure 23. CNE variant frequencies currently listed in 1000 Genomes database (phase 3). All variants listed in the CNE regions targeted by the probe set used (methods 2.2.3) from 1000 Genomes phase 3 release (The Genomes Project, 2015) were extracted. Global allele

frequencies for all variants were retained and histograms plotted of their spread. GAF: Global Allele Frequency. The right-hand panel shows a detailed view of frequencies, removing the rare variants for ease of viewing. Rare variants have GAF <0.01.

The CNE targeted sequencing used produced high coverage data for 83% of all individuals sequenced in the four cohorts (Table 7, Figure 24) suitable for rare variant calling, demonstrating the success of this method (including using different methods for targeted capture and enrichment described in methods 2.2.1 and 2.2.3).

Table 7. CNE targeted sequencing coverage of all cohorts

Cohort	Samples sequenced: passed QC	Average coverage per sample
ID/E	96: 96	520
CLP	192: 159	208
ANOS	20: 19	163
SCHZ	265: 199	200

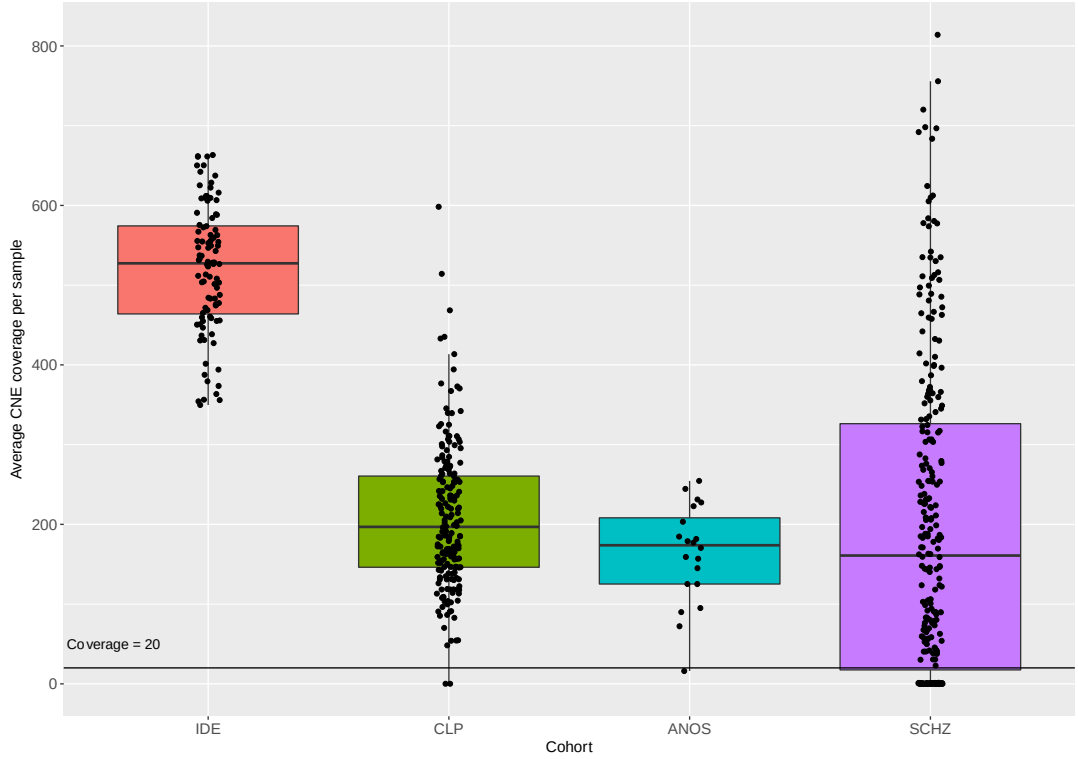


Figure 24. Spread of average coverage of samples in each sequenced cohort. Minimum average coverage cut off line = 20.

4.2.1.1 IDE

1750 SNVs were called in CNEs across 96 individuals passing all standard quality control measures. The familial inheritance patterns for these samples were complex, with multiple traits and for some, a variety of clinical notations from the medical practitioners working with the patients. Therefore, this patient group was largely used as proof of principal for targeted CNE sequencing and in the development of the variant prioritisation pipeline outlined below (Figure 27). The rich depth of clinical data could have been beneficial if a higher level of clinical input was available, including further understanding of the implications of familial phenotypes. This clinical and genetic data could be well used in future with a larger resource to investigate fully the individual cases, such as through the DDD project (Deciphering Developmental Disorders, 2017).

4.2.1.2 CLP

The cohort of trios used for the CLP study comprised of 31% female and 69% male affected children and their parents (Appendix Table 3). Once sequenced, reads mapped and variants called, a check of the patient samples for shared genetic information was performed. A simple assessment of the % of shared variation between individual samples, plotted as nodes (Figure 25) showed clear evidence of sample cross-contamination. This reduced the number of 'trusted' trios to 53 (total n= 159).

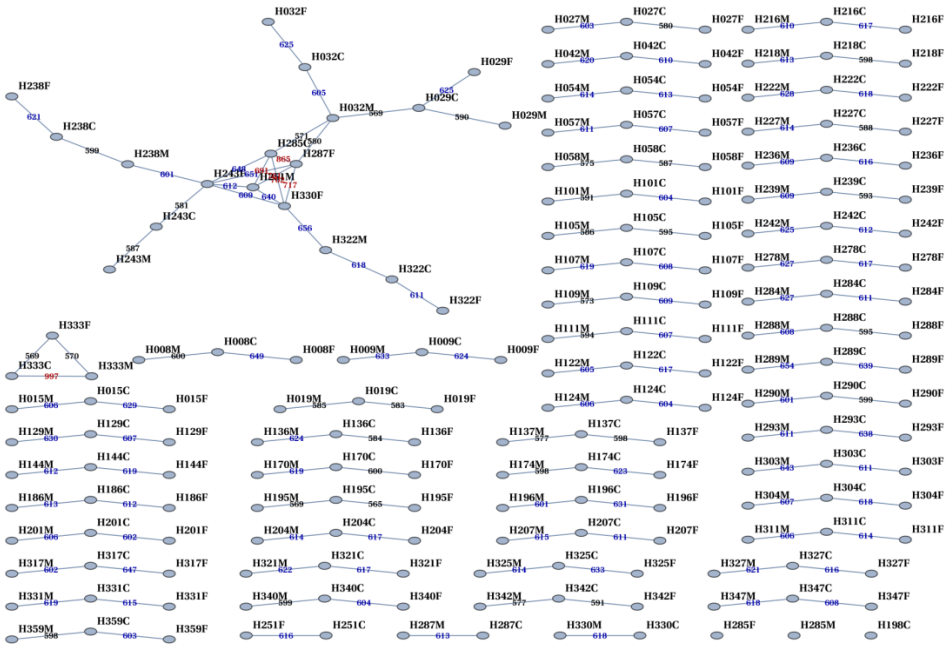


Figure 25. Common variation between CLP samples shows sample mislabelling and contamination. The kinship network diagrams were created through R Bioconductor package using shared variant genotypes amongst samples for mapping.

With such a large degree of sample mix up, any implicated families were removed from downstream analysis and samples were used to develop and perform trio-based analysis of variants through the variant prioritisation pipeline. A total of 2266 variants were identified that passed quality controls.

4.2.1.3 ANOS

The Anosmia based cohort consisted of 20 samples from various ethnicities and a variety of familial relations (see 7.6). To ascertain if there was similar sample contamination or mislabelling as per the CLP cohort noted above, variant nodal analysis was performed in the same way and samples were found to be correctly labelled and with no detectible contamination (Figure 26). A total of 4461 variants locations were found within 500bp up- or down-stream of CNE regions. One sample (AN004) had low coverage compared to the rest of the samples. A total of 1087 variants were located within the CNE regions and extended regions were used for downstream analysis

(where coverage was good enough) in an attempt to understand extended haplotypes from the CNE borders better.

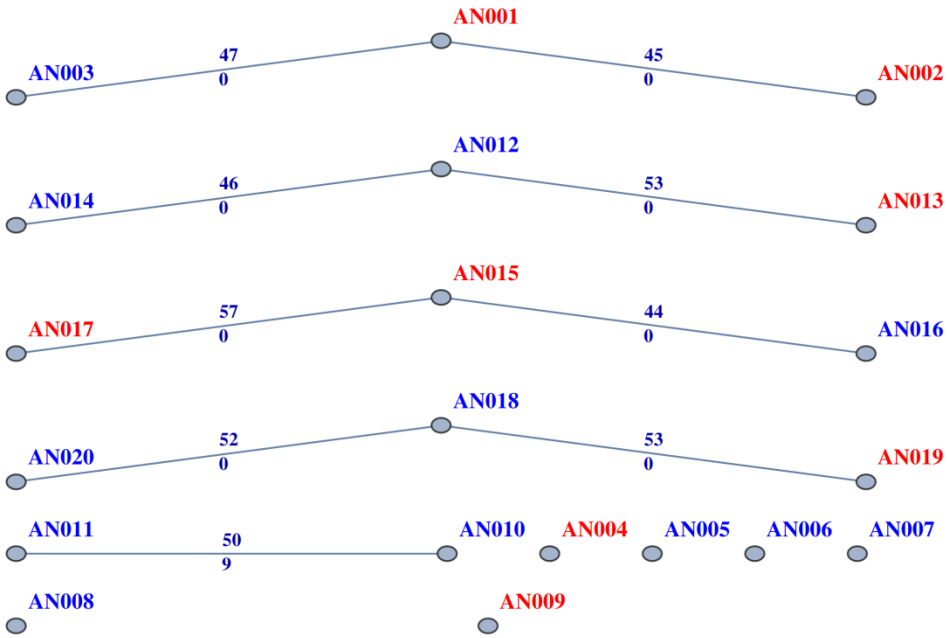


Figure 26. Identification of trios and confirmation of genders of samples using only variants called.
Numbers are % of shared variation. Red indicates females, blue indicates males. Network analysis was performed through R Bioconductor package.

4.2.1.4 SCHZ

A total 3652 variant locations were found in the 199 samples that were sequenced successfully. The Illumina Nano library preparation method may have not worked as well with the Truseq Custom Enrichment probes utilised from Illumina compared to the Truseq-Truseq compatible kit method used for the IDE and CLP cohorts, explaining the 8% failed sample rate. Samples were unrelated and similar nodal analysis of variants as performed previously confirmed this. This figure is not included as it is simply a display of individual samples with no network.

4.2.2 Development of a CNE targeted sequencing and variant prioritisation pipeline using sequence data from IDE and CLP cohorts.

Utilising cohorts of different sizes and with varying levels of inter-familial relationships and from different ethnic backgrounds, I have developed a standardised pipeline in order to sequence conserved non-coding regions and prioritise variants found within these putative developmental enhancer regions for further functional investigation (Figure 27). This method of sequential analysis of variants was developed using the IDE cohort and the CLP cohort. The CLP sequence data arose from trios where the child was affected with the most severe form of cleft lip and palate combine, yet both parents were unaffected. This allowed the variant prioritisation to focus on *de novo* heterozygous mutations amongst the affected children, or new homozygous occurrences in just affected children of rare (AF <1%) variants. The IDE cohort has complex familial information including many sub-clinical phenotypes in parents. Therefore, grouping of 'affected' and 'unaffected' individuals based on clinician recommendations (presence of absence of any neurological disorder) had to be implemented. Rare variants found only in affected individuals or homozygous in affected individuals were reported.

These two cohorts demonstrated this method's ability to reduce potential functional variants from thousands to a number small enough for human interrogation. These variants are then easily assessed against publicly available functional data for further insight into their practical validity, annotated, and where appropriate can be taken through to functional analysis.

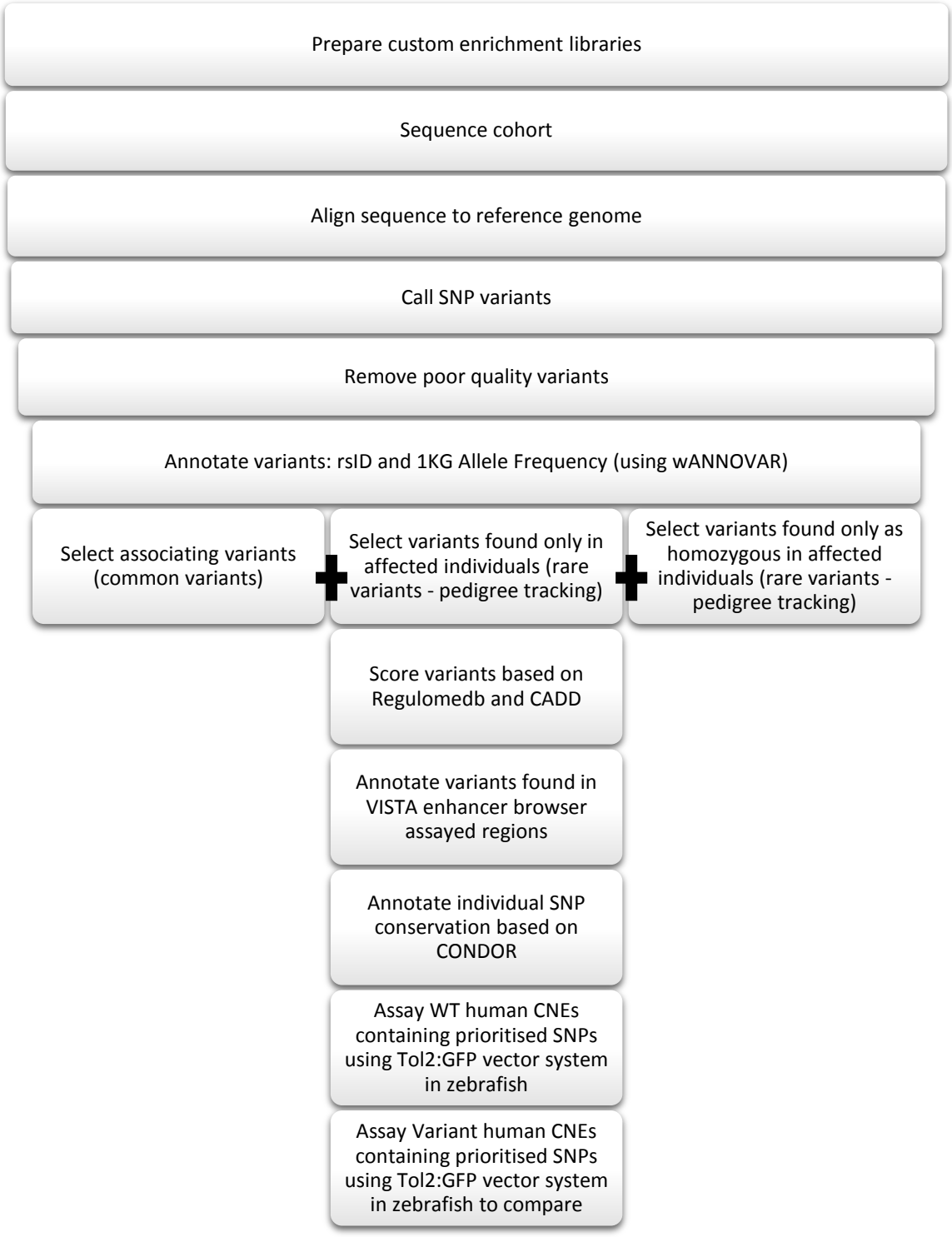


Figure 27. Visualisation of targeted sequencing and variant prioritisation pipeline. Various publicly available datasets and tools were used as follows: wANNOVAR(Wang et al., 2010, Chang and Wang, 2012) Variant Effect Predictor (VEP) (McLaren et al., 2016), Regulomedb (Boyle et al., 2012), Combined Annotation Dependent Deletion (CADD)

(Kircher et al., 2014), VISTA Enhancer Browser database (Visel et al., 2006), and COnserved Non-coDing Orthologous Regions database (CONDOR) (Woolfe et al., 2007).

After using the literature to review regulatory scoring algorithms publicly available (1.2.4), CADD and RegulomeDB were chosen to score variants on their pathogenicity. The algorithms have slightly different approaches and including both a categorical and a continuous scoring programme allows for full integration of known data and its relevance. In addition, a comparison of scores for all known CNE SNPs in dbSNP release 141 that are directly comparable in CADD and RegulomeDB (n=8273) shows multiple differences in the rank of variants based on the two approaches (Figure 28 and Figure 29). CADD scores above 20 place the variant in the top 1% of pathogenic variants, while a RegulomeDB score of 3 and above demonstrates at least two matching indicators of regulatory function at that nucleotide location.

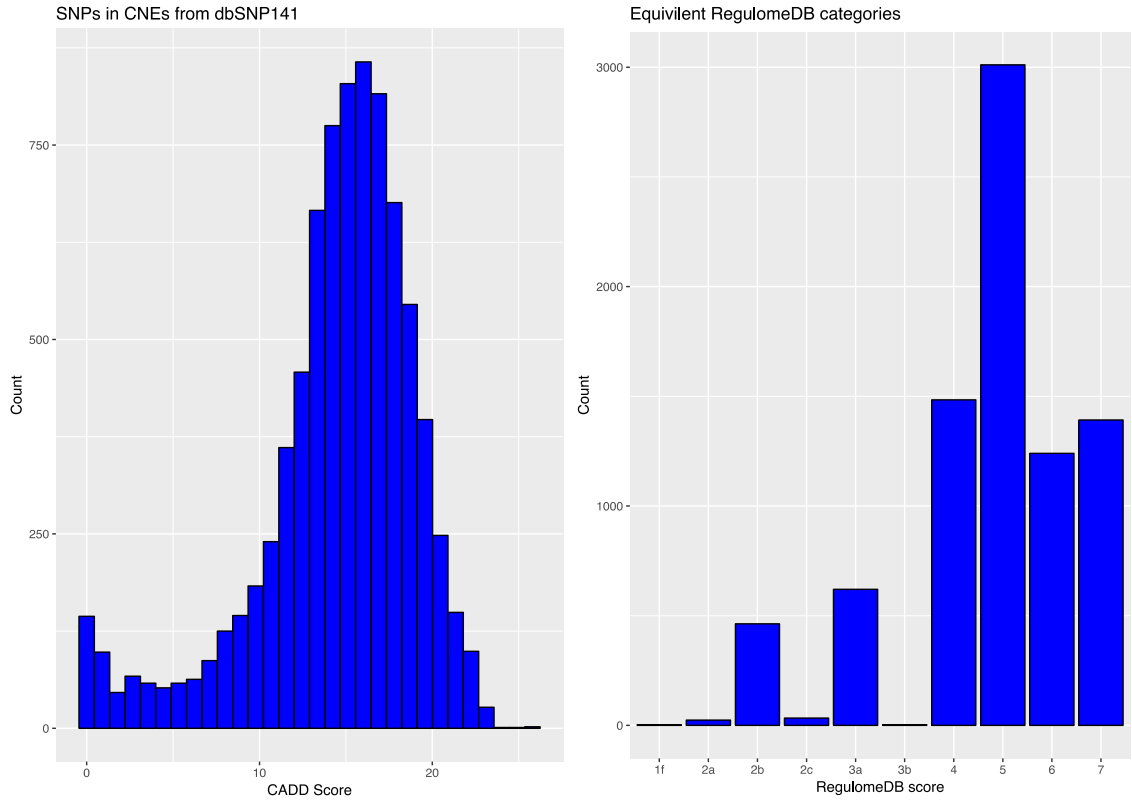


Figure 28. Comparison of spread of scores for CNE variants in dbSNP141 by CADD and RegulomeDB.

CADD scores generally place CNE variants (and non-coding variants) outside of the top 1% pathogenic and previous studies have used much lower cut off scores than 20 for non-coding variants of interest due to the way CADD is trained on all possible variants in the genome (Mather et al., 2016). Therefore its use to rank variants in order may be more informative than the overall score for CNE variants. This can be seen again when looking at the spread of CADD scores when compared to RegulomeDB scores for all CNE variants in dbSNPv141 (Figure 29) with no significant trend being observed that would suggest agreement between the two programmes.

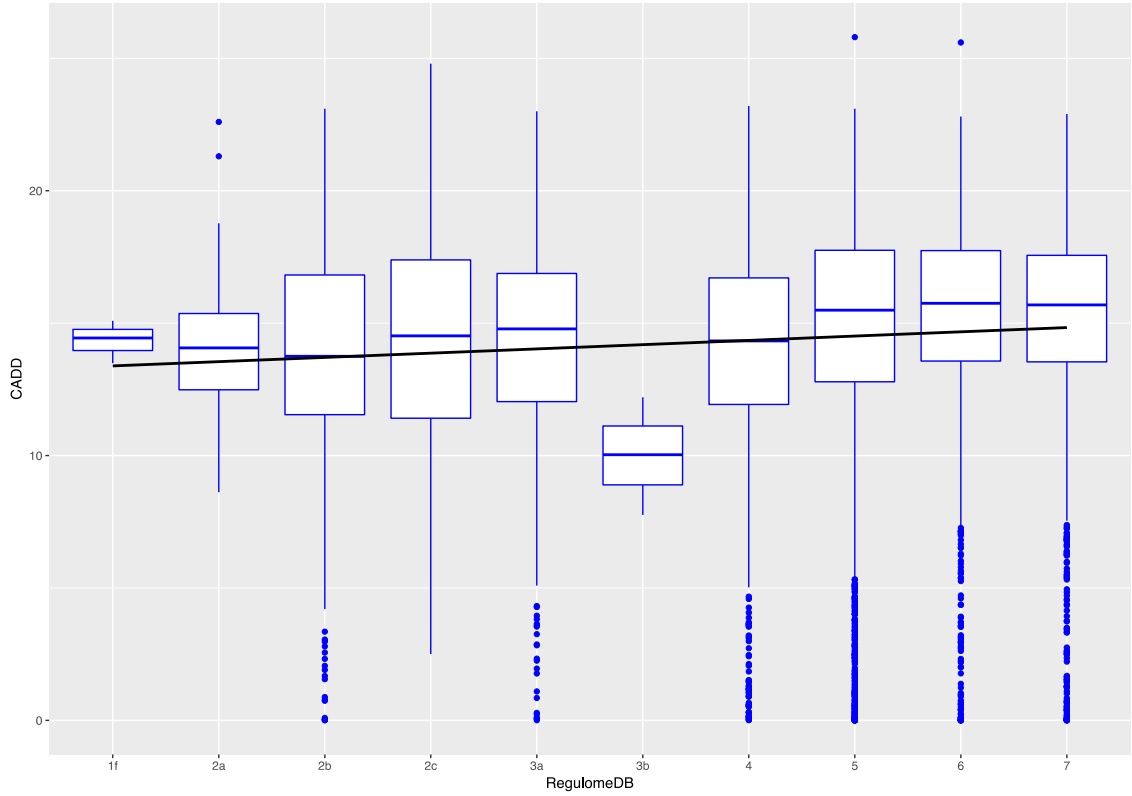


Figure 29. Comparison of the spread of CADD scores within RegulomeDB scoring categories shows no significant trend in agreement between the two sets of data

4.2.2.1 IDE variants of interest

Utilising familial data, there were 32 instances of SNPs that were only found in the variant homozygous form in affected children, and a single *de novo* heterozygous mutation in one affected child. None of these significantly associate with the phenotype

however there are 4 rare homozygous SNPs and one rare heterozygous SNP (EAF ≤ 0.01).

4.2.2.2 CLP variants of interest

Utilising trio data, there were 63 SNPs that were found homozygously only in children with CLP and one SNP that was heterozygous in an affected child only (*de nov*o het). When comparing the allele frequencies to those in 1000 Genomes phase 3 release for European cohorts, a list of 5 SNPs was curated (Table 8).

Table 8. SNPs in CNEs exclusive to CLP children, rare in 1000 Genomes European cohort. SNP 5:91019059 is a *de novo* heterozygous variant. All others are only found as homozygous in affected children.

Region	Chr	Pos	Ref	Alt	CNE	1KG EUR AF
NR2F1 REGION	*5*	*91019059*	*T*	*C*	*CRCNE00008112*	*0*
PROXIMAL TO HOXD9 AND LUNAPARK	2	176428892	C	A	CRCNE00010496	0.01
PROXIMAL TO IRX2 AND IRX1	5	2547178	C	G	CRCNE00006703	0.01
PROXIPMAL TO TFAP2A	6	10150462	C	G	CRCNE00007024	0
PROXIMAL TO POU3F2	6	97949475	G	A	CRCNE00009797	0.01

With these relatively small cohort sizes (IDE and CLP) and the small regions of conserved non-coding sequence that is studied (0.7Mb) there is a low expectation of *de novo* mutations. Methods for predicting these rates as described previously (Samocha et al., 2014) focus on exome analysis, incorporating per base function and mutation effect (synonymous or nonsynonymous) making them limited in application to this method. New per-base constraint predictions for conserved non coding regions will need to be calculated, taking into account NVR and RVR sites (De Silva et al,. 2014) alongside a deeper understanding of the non-linear code and function.

4.2.2.3 CNE sequencing discovers novel non-coding variants

Utilising such a high-coverage and targeted approach to non-coding sequencing, new variants that cannot be found in previous literature were found for each cohort. Comparisons of cohort allele frequencies against those found in 1000 Genomes publicly available data point to multiple novel variants and some disparity between more common variation as well (Figure 30 and Figure 31).

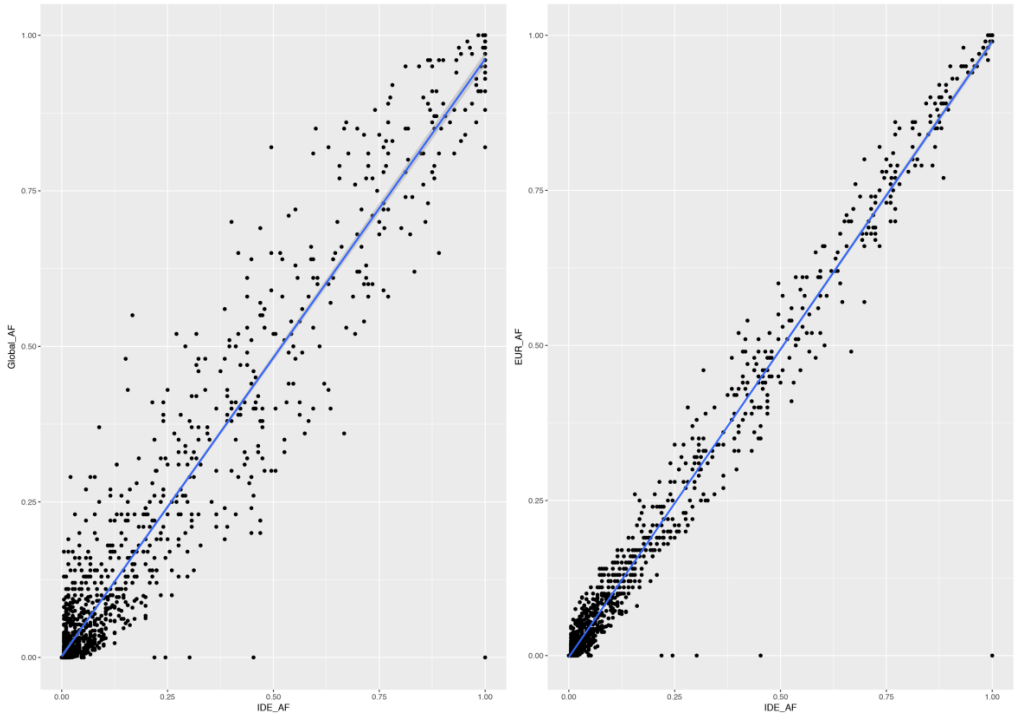


Figure 30. Comparison of IDE cohort allele frequencies and 1000 Genomes global allele frequencies identifies multiple novel variants.

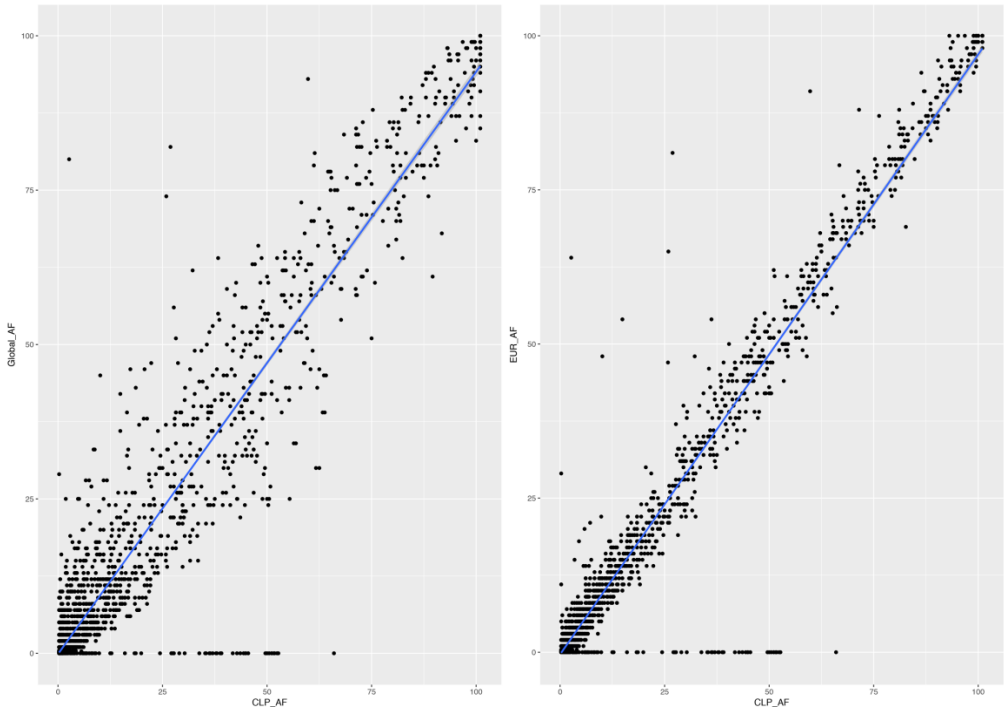


Figure 31. Comparison of CLP cohort allele frequencies and 1000 Genomes global allele frequencies identifies multiple novel variants.

Using all 4 cohorts, total number of variants annotated and those that can be deemed to be novel were listed (using wANNOVAR to annotate (Wang et al., 2010) and novel being defined as not being seen in dbSNP, gnomAD or 1000 Genomes). Targeted non-coding sequencing provides a great deal of previously unknown information on human noncoding variation (
Table 9). This is especially true in cohorts with ethnicities not used in GWAS studies as often; the Pakistani SCHZ cohort and the Faorese component to the ANOS cohort may result in higher novel mutation discovery as these ethnicities are underrepresented in current GWAS studies (Popejoy and Fullerton, 2016).

Table 9. Novel variants discovered in cohorts undergoing targeted CNE sequencing. A full list of novel variants discovered here will be made available alongside the ENA submission of sequences (vcf file).

Cohort	Total no. of variants	Novel variants	% Novel Variants	Samples Sequenced	Per sample average novel variants
IDE	1750	109	6.2%	96	1.14
CLP	2266	203	9.0%	159	1.27
ANOS	1087	48	4.4%	19	2.5

SCHZ	3652	1016	27.8%	199	5.1

4.2.3 Targeted CNE sequencing of a small Anosmia cohort identifies familial disease-associating variants with predicted functional consequences

Utilising the variant prioritisation framework developed previously (Figure 27) each family was scrutinised for variants that fit the predicted inheritance patterns they displayed. For families B1, B2, B3 and B6 (Figure 45, Table 24) these were small pedigrees with limited availability of sequence data. It was possible to predict the pattern of inheritance of contributing SNPs based off of the family pedigree information given. This plus comparing the presence/absence of variants in other family's ICA and control samples allowed the significant reduction of 'interesting' variants (
Table 10).

Table 10. Family-specific pedigree tracing, using other affected and unaffected individuals as control populations

Family	Patients sequenced	Family controls	Predicted pattern of inheritance	Number of variants that fit pedigree
B1	ICH Father ICA Daughter	Mother	Autosomal Dominant (reduced penetrance)	39
B2	ICA Brothers (2)	none	Autosomal Recessive	38
B3	ICA Father ICA Daughter	Mother	Autosomal Dominant	37
B6	ICA Daughter	Father Mother	Autosomal Dominant **or** Autosomal Recessive	7

There were no variants that were duplicated in these short lists between families. These variants were then annotated using CADD, RegulomeDB, VISTA enhancer browser and CONDOR database.

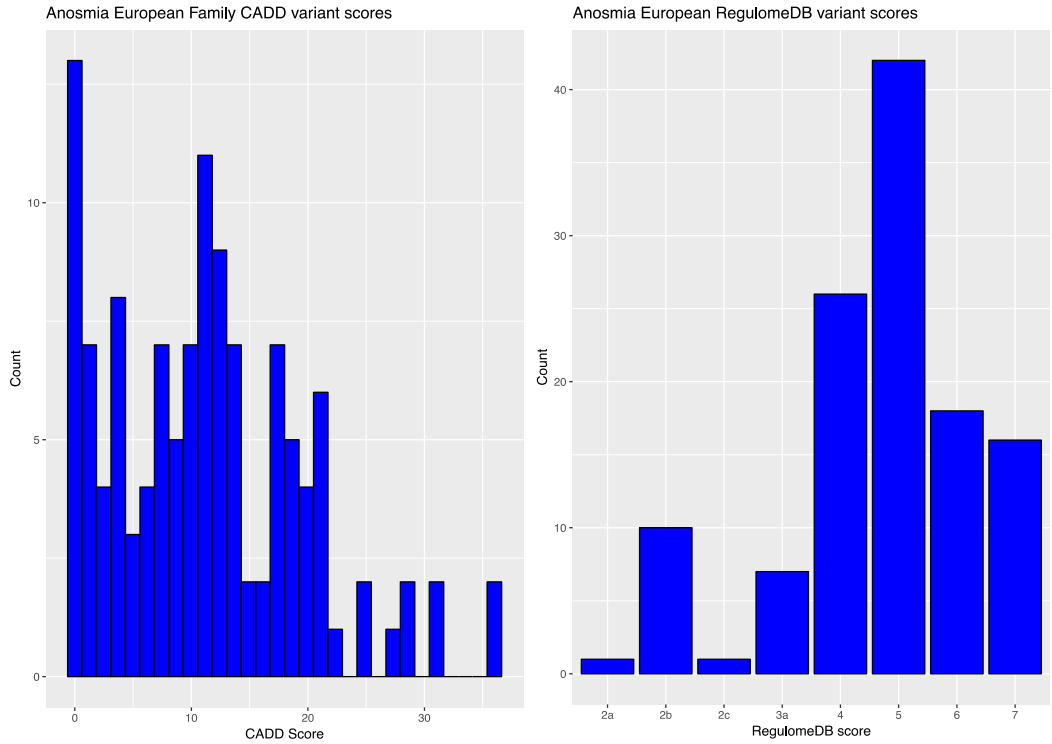


Figure 32. CADD and RegulomeDB scores for variants following inheritance patterns of Anosmia in four European families.

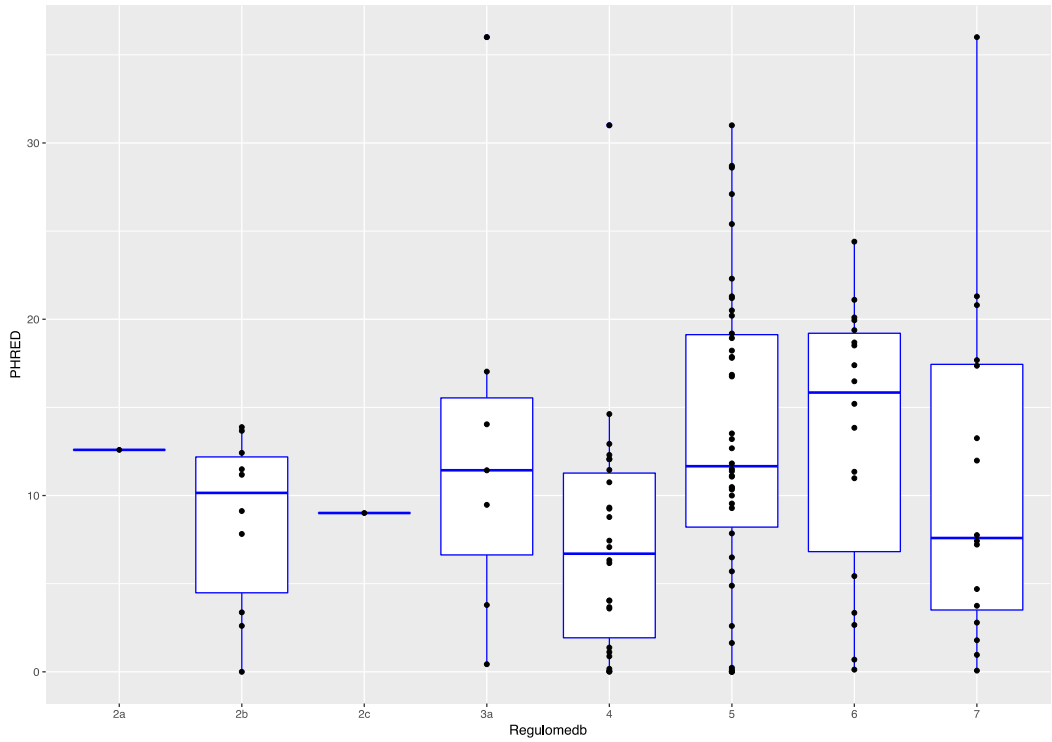


Figure 33. CADD normalised scores plotted against the RegulomeDB scores for the same variant in European Anosmia affected families

The variants with the highest pathogenic potential were compiled ready for transient Zebrafish enhancer assays going forward (Table 11).

Table 11. Prioritised familial variants in European families presenting Anosmia to be taken forward to functional studies

Family	Chrom	Pos	Ref	Alt	Vista	Regulome DB	CADD
B1	5	157959431	T	C	-	6	20.1
	3	147219331	T	C	-	5	20.2
	15	61318347	T	C	-	5	20.5
	10	78192968	C	G	-	5	22.3
	16	54524133	T	C	-	5	31
	15	96427654	G	A	-	2a	12.59
	19	31830269	A	C	-	2b	0
	7	155264279	C	T	hs1418: hindbrain, midbrain	2b	7.82
	4	147289519	T	G	-	2b	11.49
	4	147393687	C	T	-	2b	13.88
	10	78313940	C	A	-	3a	14.04
	11	16443426	C	T	-	3a	17.03
B2	10	102466902	C	T	-	7	20.8
	9	128655115	T	G	-	7	21.3
	9	128141809	G	A	-	5	28.6
	10	102469402	C	A	-	2b	2.605
	10	102475964	C	G	-	2b	9.119
	10	102475948	G	A	-	2b	11.18
	10	102475954	C	G	-	2c	9.007
	7	1308934	G	GT	-	3a	0.429
	1	3209923	A	G	-	3a	3.789
	15	98166554	A	G	-	3a	9.471
	9	128225561	T	C	-	3a	11.43

B3	2	164844248	A	T	hs421: neural tube, dorsal root ganglion	6	21.1
	1	10976340	C	T	-	5	21.2
	1	216691632	C	T	-	5	21.3
	3	147792863	G	C	-	5	25.4
	5	3326276	T	C	-	5	27.1
	16	78698869	G	T	-	5	28.7
	2	177235473	C	T	-	4	31
	10	102472468	A	T	-	3a	36
	10	102469399	C	T	-	2b	3.371
	7	156407509	A	G	-	2b	12.42
	1	63590702	GT	G	-	2b	13.67
	10	102472468	A	T	-	3a	36
B6	2	104736646	T	C	hs401: hindbrain	7	0.961

Chr2:172956687 C>T is a SNP found in two brothers (AN010 and AN011, Family B2) presenting with ICA not present in GnomAD. Although both parents are unaffected, suggesting recessive inheritance may be the cause, previous literature suggests an autosomal dominant inheritance pattern with reduced penetrance. This could be through multiple contribution variants or other environmental factors. Neither parents are CNE sequenced here, and the likelihood of both brothers randomly gaining the same SNP is rare, therefore if this variant has any contribution to the phenotype it is likely to be in addition to other variants or inherited haplotypes from the other parent. This rare variant lies within the hs422 element (chr2:172,955,879-172,957,052) tested in the vista enhancer browser (Visel et al., 2006). In this vertebrate-conserved region, the element acted as a developmental enhancer in the forebrain, midbrain and nose at mouse embryonic day 11.5 (Pennacchio et al., 2006).

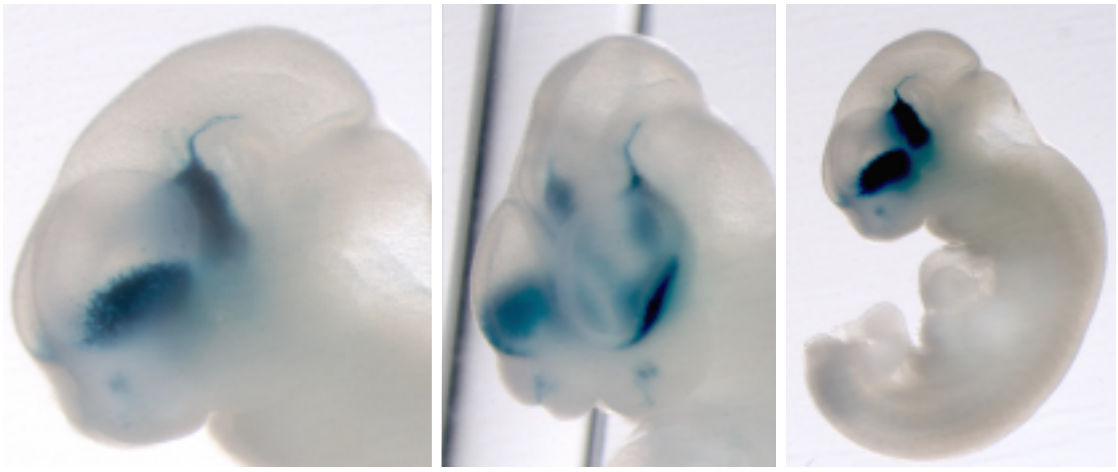


Figure 34. hs422 VISTA Enhancer element expression in e11.5 mouse embryo
Adapted by permission from Macmillan Publishers Ltd: [Nature] (Pennacchio et al., 2006), copyright (2006)

The variant has a CADD score of 22.2 suggesting it is in the top 5% of deleterious variants, although the RegulomeDB score is 5, demonstrating minimal binding evidence. The CNE region is an active enhancer in the nose but the contribution of this SNP to altered enhancer function and gene expression is unclear.

In addition to the multiple small European pedigrees, a more extensive family history was available for a Faroese family. This family had already undergone extensive genetic analysis including Karyotyping, SNP 6.0 genotyping and exome sequencing. These results were inconclusive (N. Tommerup, personal communication, 2013). Using previous evidence for dominant inheritance with reduced penetrance in a previously studied family of the same limited ethnic background (Lygonis, 1969), variants were traced that were present as heterozygous variants in affected individuals and not present in unaffected individuals. Two SNPs were found to be present in all 5 affected individuals and not seen in any unaffected individuals. These variants are in close proximity to each other: chr7:1461971 G>A and chr7:1461984 T>C. These SNPs are located 50bp downstream of CRCNE00009845, with the reduction in conservation from the CNE boundary to the variants caused by the region being missing in some vertebrate species. Both variants have regulomeDB scores of 5 and low scores from CADD (raw scores = 0.28 and 0.08 respectively) not indicating any discernible pathogenicity. Nonetheless, the presence of these two SNPs in such close proximity within this family

suggests there could be an associating haplotype in this region extending away from the CNE probe, making the consequences of these specific SNPs difficult to predict.

The cluster of CNEs in this region associate with the gene UNC Homeobox (UNCX). The UNCX Homeobox Protein is a Transcription factor involved in neurogenesis and somitogenesis. Amongst other functions, it plays a role in controlling the development of connections of hypothalamic neurons to pituitary elements. In addition, it is GO annotated to be involved in olfactory bulb interneuron differentiation (Ashburner et al., 2000, Gene Ontology Consortium, 2015). Previously, ICA patients have presented lack of olfactory cells or few olfactory sensory neurons in olfactory epithelium biopsies (Assouline et al., 1998). Therefore, changes in UNCX expression could plausibly affect the normal development of the olfactory system.

4.2.4 Targeted CNE sequencing of a Schizophrenic cohort identifies disease-associating variants with predicted functional effects

4.2.4.1 Comparative population genetics is restricted by publicly available data, including poor coverage of the non-coding genome and few ethnically comparable samples

Targeted CNE sequencing identified 3652 variant locations within the Pakistani Schizophrenic cohort group. Some of these variants were multiallelic, making the total number of variant alleles identified 3826. Using wAnnovar, these were annotated with RsIDs for dbSNP analysis and 1KG allele frequencies. A total 992 variant alleles had no RsID and no allele frequency information available (26%). Of these variant alleles, 890 (90%) are rare in the cohort (MAF<0.01) and can be classed as personal variants or a false positive variant call. This additional information on the extent of rare variation in conserved noncoding regions extensively adds to the variants identified in these regions, adding almost 1000 novel variants to the ~20,000 already called by 1KG within CNEs. An additional 102 common allele (MAF>0.01) were also identified (Appendix Table 3). Comparable ethnic cohorts are present in public data but not large enough to find variant associations.

The cohort studied is from Pakistan, therefore any comparison of allele frequencies should be between a similar ethnic background cohort where possible. Although there is a subset of the 1000 Genomes sample groups from Pakistan (n=157), utilising a larger number of individuals would gain more accurate allele frequencies for noncoding regions (where coverage is much lower). To see if the full SAS (South Asian) subgroup could be used, a PCA comparing the Pakistani cohort and the different sub-groups of 1000 Genomes was performed to identify degree of concordance between the two populations (Figure 35). This showed that the Gujarati Indian from Texas (GIH) sub-population was the most estranged, retaining some overlap. There were also outliers within the Pakistani patient cohort which could be attributed to low sequence coverage. In addition, the PJL subgroup has 2,658 variants listed across the CNE probe regions – just under 13% of the total number of variants called globally for these regions (Appendix Script 1).

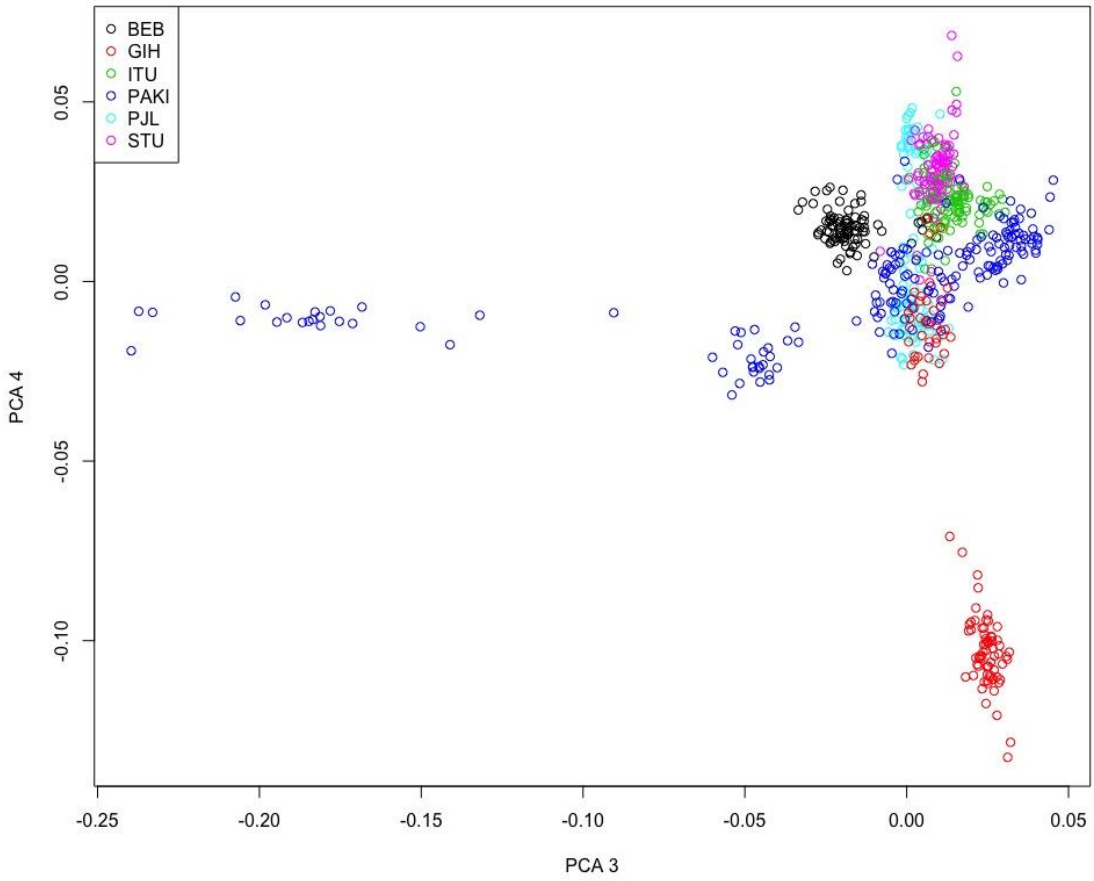


Figure 35. PCA showing overlap of Pakistani Schizophrenic cohort with SAS 1KG population. The Pakistani cohort sequenced here are PAKI in navy blue, the other

subgroups are 1KG South Asian individuals. Further analysis of the smeared samples from approximately -0.10 on PCA3 axis shows these individuals to have multiple low coverage regions. PCA performed using minor allele frequencies as described previously (Lu and Xu, 2013). The PJL subgroup are the closest geographical subgroup to the new Pakistani samples we have sequenced.

Using a 2x2 contingency table, case-'control' allelic association chi squared and p values (d.f.=1) were conducted, primarily with 1KG SAS population frequencies but also with 1000 Genomes global allele frequencies and gnomAD frequencies where values were missing. Case individuals were from our Schizophrenia cohort as described above, 'control' individuals were simply the publicly available population genetic data available for this region for 1KG SAS individuals. No variants were found to have significant association to the Schizophrenic cohort when comparing to 1KG SAS and Global Allele frequencies ($pvalue<5x10^{-8}$). This threshold was used to prioritise variants despite not necessarily being a true representation of actual statistical significance.

The gnomAD project pulls together 15,496 whole genomes from various sources including patient data, however there is no representation of South Asian origin individuals in the current data release. In addition, there is no phenotype data that can be matched with variant information. This being said, the gnomAD data provided the most comprehensive coverage of all the variants found in the case cohort, providing information on an additional 243 variants when compared to the SAS population, and 24 variants compared to the 1KG global allele frequencies. An additional 182 variants were assigned RsIDs but with no allele frequency information available. Therefore, gnomAD allele frequencies were used to perform further case-'control' association tests. The use of gnomAD as a 'control' data set is limited largely by the quality of variants available and the comparative methods used. In addition, as mentioned previously, the limit of number of individuals available for comparative ethnicities is difficult to overcome until more sequencing studies are done. GnomAD also pulls genetic information in from other disease-association studies, potentially skewing some of the data based on other disorders being studied, however the hypothesis that the large

numbers of individuals available will overcome this noise. Therefore despite being referred to as a 'control' set, it should be noted that this is just in principle for statistical analysis and variant prioritisation and the genetic data is not truly reflective of a control group.

4.2.4.2 Use of the variant-prioritisation pipeline (developed in 4.2.2) identifies Schizophrenia associating variants with predicted functional consequences

Case-control allelic association testing of cohort allele frequencies compared to gnomAD reported frequencies identified 12 variant alleles associating with the Schizophrenia cohort (p-value<$5x10^{-8}$).

Table 12. Variant alleles associating with Schizophrenia

Chrom: Chromosome; Ref: Reference Allele; C.AF: Cohort minor allele frequency; G.AF: gnomAD allele frequency

Chrom	Start	End	Ref	Alt	C.AF	G.AF	P-value
10	131556191	131556191	A	C	0.2297	9.71E-05	4.94E-238
10	77469716	77469716	A	G	0.2294	9.70E-05	1.04E-237
3	157882877	157882877	T	G	0.4615	0.0004	3.51E-233
3	71629340	71629340	T	G	0.2627	0.0002	7.14E-152
5	2113231	2113231	A	C	0.2106	0.0005	2.72E-40
2	172580364	172580364	T	G	0.2319	0.0007	4.40E-35
2	104496685	104496685	T	G	0.2308	0.0007	9.122E-35
19	30941020	30941020	C	T	0.1833	0.001	3.55E-16
10	114885262	114885262	-	A	0.029	3.27E-05	7.53E-13
13	101015915	101015915	C	G	0.0258	3.23E-05	1.43E-10
16	80143645	80143645	A	C	0.2162	0.0031	6.20E-08
16	51501890	51501890	T	A	0.0217	3.23E-05	6.94E-08

Of these variants, only chr10:114885262, chr13:101015915 and chr16:51501890 have recorded rsIDs (rs527416250, rs376094199, and rs540088918 respectively) with none

appearing in recent Schizophrenia meta-anlayses (Ripke et al., 2014). These p-values are indicators only and further sanger sequencing would be needed to confirm the variant locations and frequencies with more time and resource.

4.2.4.3 Functional predictions of disease associating SNPs prioritising variants for functional analysis

Utilising publicly available software and tools allows the novel associating variants to be prioritised for functional assays in zebrafish. Functional assays are medium-throughput with each enhancer assay taking 2-4weeks of hands-on time. The variants were initially scored using the CADD software (Kircher et al., 2014). Results were reported both as raw scores and as scaled scores (
Table 13) and variants were sorted from most likely to be pathogenic to least likely. All scaled (PHRED) scores were below 20, the suggested cut off for pathogenicity (the 1% most deleterious variants), although CADD consistently scores noncoding variants lower generally than coding variants suggesting a lower threshold level could be used.

Table 13. CADD scores for Schizophrenia associating variants

#Chrom	Pos	Ref	Alt	RawScore	PHRED
2	104496685	T	G	2.43598	19.05
16	51501890	T	A	2.427926	19
10	131556191	A	C	2.150394	17.18
16	80143645	A	C	2.005432	16.25
10	77469716	A	G	1.917527	15.7
2	172580364	T	G	1.794542	14.95
3	157882877	T	G	1.538025	13.52
13	101015915	C	G	1.34814	12.52
19	30941020	C	T	1.23543	11.93
10	114885262	-	A	1.005911	10.7
3	71629340	T	G	0.929835	10.25
5	2113231	A	C	-0.163046	1.306

Additionally, the 12 associating variants were scored using RegulomeDB (Boyle et al., 2012). RegulomeDB is specifically built for non-coding regulatory variants, scoring and annotating the SNPs into categories based on known and predicted regulatory elements. Scores are possible between 1a-6 (14 levels) with the 12 Schizophrenia variants scoring from 2b-7 (
Table 14).

Table 14. RegulomeDB scores of Scizophrenia associating variants.

Co-ordinate (0-based)	Regulomedb Score
chr3:71629341	2b
chr5:2113232	2b
chr13:101015916	2b
chr10:131556192	5
chr10:77469717	5
chr2:172580365	5
chr2:104496686	5
chr19:30941021	5
chr10:114885263	5
chr3:157882878	7
chr16:80143646	7
chr16:51501891	7

4.2.4.4 Enhancer assay in zebrafish confirms neural developmental activity of CNE surrounding an associating variant upstream of POU3F3

The highest scoring associating variant as measured by CADD is chr2:104496685 T>G (RegulomeDB score 5). This SNP lies in CRCNE00007883 which is located approximately 1Mb proximal to POU3F3. The SNP within a hESC TAD (Schmitt et al., 2016) containing two non-protein coding RNAs (LINC01796 and LINC01935), yet this TAD and the adjacent TAD downstream do share many interactions, suggesting together they form a much larger TAD (Figure 36).

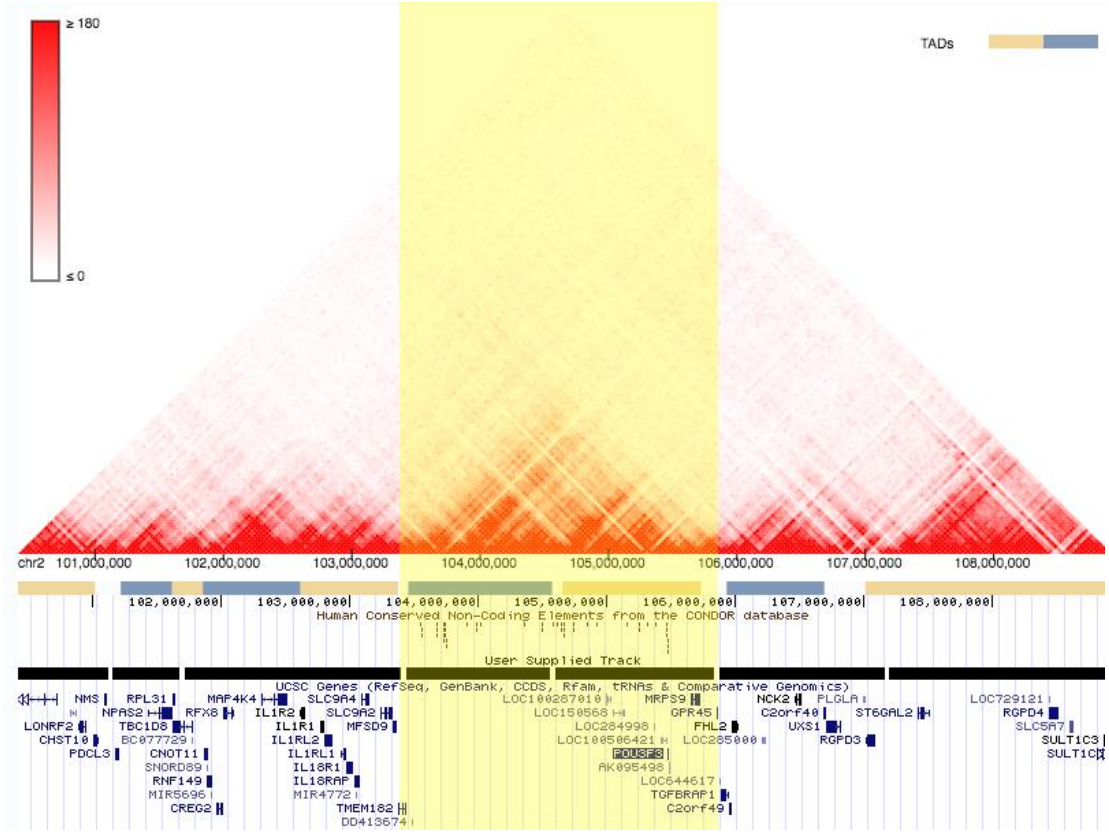


Figure 36. HiC interactions in H1-ESC cells

Blue and Yellow bars represent TADs as defined previously (Dixon et al., 2015) with hg18 to hg19 LiftOver implemented. User supplied track uses complimentary TAD boundary definitions (Schmitt et al., 2016). CNEs lie across two TAD blocks however HiC data visualised here shows strong interactions across and between both regions (highlighted). POU3F3 is highlighted as the closest neuro-developmentally active gene.

CRCNE00007883 in its WT human form was cloned upstream of a minimal cfos promoter and GFP in the Tol2 vector (methods 2.2.7). Microinjection of this construct alongside transposase mRNA into 1 cell stage fertilised zebrafish embryos allowed visualisation of enhancer function of this CNE in development (methods 2.2.8). CRCNE00007883 showed specific enhancer function in the nervous system in 29% of injected embryos, peaking at 48 hpf (Figure 37) and remaining through 72 hpf (Figure 38).

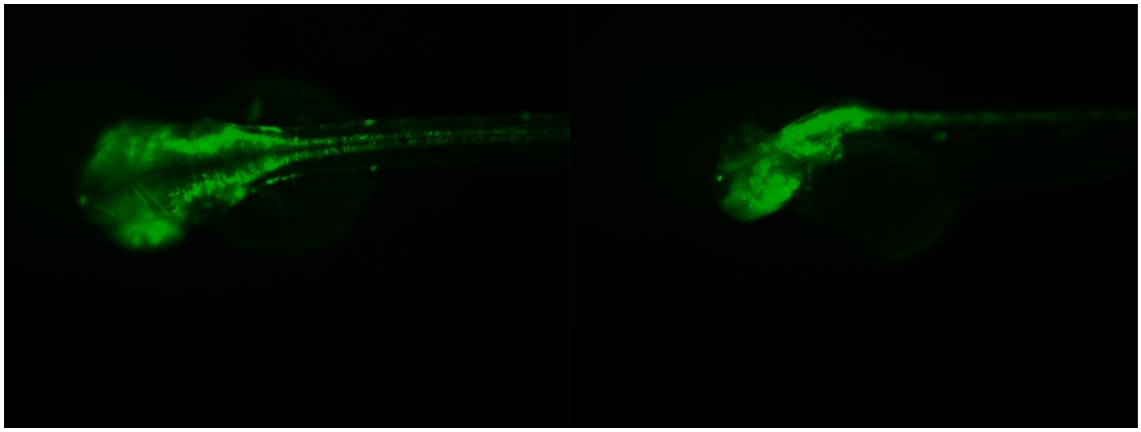


Figure 37. 48hpf zebrafish embryo with showing neuronal GFP expression after 1-cell stage microinjection of B-Tol2:GFP.

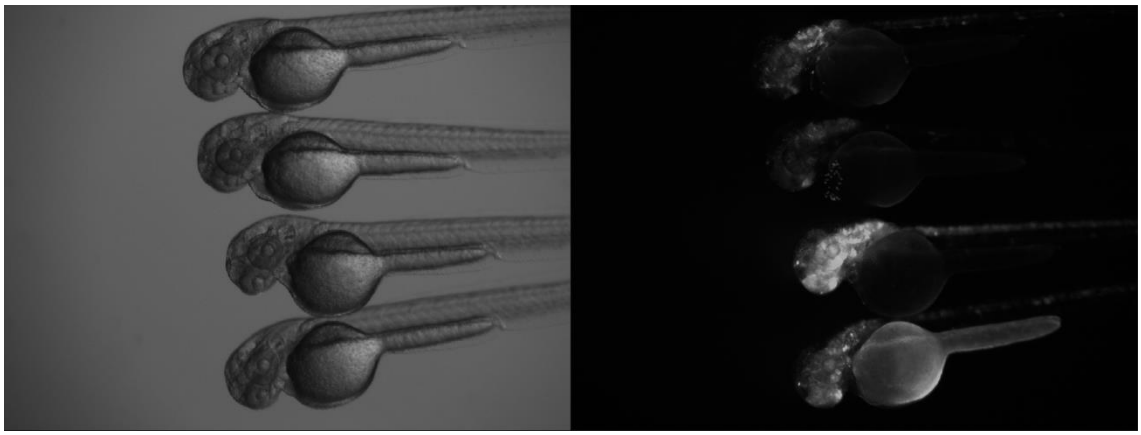


Figure 38. Enhancer signal variations between zebrafish injected with the same construct (B).

The variant version of CRCNE00007883 showed similar neuronal GFP expression patterns with no significant difference in neuronal specific GFP positive embryos when compared to the wild type construct (27%). GFP expression varied between embryos in intensity, possibly as a product of the integration of the vector. Creation of a stable transgenic line could further improve this method, although it would be at a much lower throughput.

4.2.4.5 Predicted consequences of variant allele chr2:104496685 T>G

The SNP chr2:104496685 T>G is located in a highly conserved region of the genome with the reference thymine conserved throughout all vertebrate organisms (Figure 39).

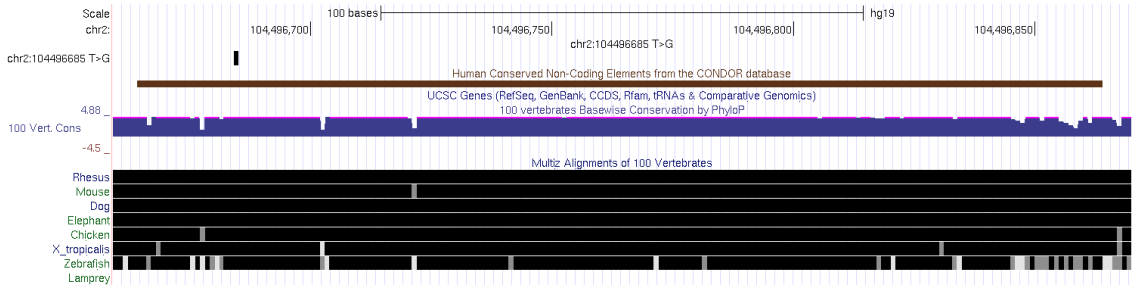


Figure 39. CRCNE00007883 sequence level conservation shows chr2:104496685 is a non-variable base amongst vertebrate organisms.

Utilising the JASPAR online tool (Mathelier et al., 2016) and the nucleotide sequence 6 bases up- and down-stream of this SNP, 8 putative transcription factor binding sites were found (Table 15). The introduction of the variant SNP altered the putative sites, introducing 4 new models above the threshold score, and removing 3 (Table 16). TRASFAC curated transcription factors that bind to the promoter of POU3F3 in functional studies from these lists are CEBPA and FOXO4 (Rouillard et al., 2016, Matys et al., 2006). The FOXO4 site is introduced as a result of the chr2:104496685 T>G base change. These predictions suggest that the variant lies within a transcription factor binding site that has the potential to regulate POU3F3, which acts as transcription factor the development of the nervous system. However the link between POU3F3 and misregulation of brain development that could contribute to schizophrenia is tenuous at best and further studies would need to be conducted to prove 1) that the associating variant found has an impact on gene regulation, 2) that this misregulation of POU3F3 has an impact on brain development, and the 3) this impact could in any way contribute to schizophrenia. This is a good example however, of a way to distil large numbers of variants into a manageable number for future lines of inquiry, in a similar way to large scale forward genetic screens as a starting point.

Table 15. JASPAR output for WT sequence chr2:104496679-104496691
Input: GGCTGGTTCAAT; relative profile score threshold: 80%. Difference to variant version are highlighted.

Model ID	Model name	Score	Relative score	Start	End	Strand	predicted site sequence
MA0745.1	SNAI2	5.08	0.8494	1	9	1	GGCTGGTTC
MA0102.3	CEBPA	3.61	0.8412	2	12	-1	ATTGAACCAGC
MA0466.2	CEBPB	2.71	0.8220	3	12	1	CTGGTTCAAT
MA0837.1	CEBPE	4.13	0.8181	3	12	1	CTGGTTCAAT
MA0099.2	FOS::JUN	5.72	0.8198	4	10	-1	TGAACCA
MA0719.1	RHOXF1	1.84	0.8410	4	11	-1	TTGAACCA
MA0847.1	FOXD2	2.56	0.8084	7	13	1	TTCAATA
MA0033.2	FOXL1	3.46	0.8390	7	13	1	TTCAATA

Table 16. JASPAR output for variant sequence chr2:104496679-104496691
Input: GGCTGGGTCAAT; relative profile score threshold: 80%. Differences compared to WT are highlighted.

Model ID	Model name	Score	Relative score	Start	End	Strand	predicted site sequence
MA0102.3	CEBPA	0.41	0.8029	2	12	-1	ATTGACCC AGC
MA0477.1	FOSL1	3.16	0.8126	2	12	-1	ATTGACCC AGC
MA0099.2	FOS::JUN	6.83	0.8602	4	10	-1	TGACCCA
MA0719.1	RHOXF1	1.84	0.8410	4	11	-1	TTGACCCA
MA0847.1	FOXD2	7.50	0.9132	7	13	1	GTCAATA
MA0042.2	FOXI1	7.33	0.8879	7	13	1	GTCAATA
MA0033.2	FOXL1	7.56	0.9140	7	13	1	GTCAATA
MA0848.1	FOXO4	8.12	0.9101	7	13	1	GTCAATA
MA0849.1	FOXO6	6.52	0.8851	7	13	1	GTCAATA

This data suggests that this cohort's Schizophrenia associating SNP chr2:104496685 T>G lies within an active enhancer in the development of the nervous system. The SNP itself has the potential to change the transcription factor binding affinity of multiple transcriptions factors. The CNE is associated to the developmental transcription factor POU3F3, with chromosome conformation capture data supporting the potential for active interaction and looping of the CNE to the gene region across this vast region. The variant allele introduces a strong binding affinity for the transcription factor FOXO4, which has already been shown previously to bind to the promoter of POU3F3 (Matys et al., 2006). POU3F3 has previously been implicated as a gene involved in the schizophrenia phenotype (Potkin et al., 2008).

POU3F3 regulates upper-layer neuronal migration and identity during development of the neocortical layers (McEvilly et al., 2002, Sugitani et al., 2002). Neuronal migration

defects have been associated with schizophrenia (Wang et al., 2011, Eastwood and Harrison, 2005), and although the evidence is not conclusive, misregulation of a neuronal development gene fits the current hypothesese that non-traumatic schizophrenia is a consequence of abnormal foetal brain development (Suddath et al., 1990).

4.3 Discussion and Conclusions

Selective sequencing of the human genome in search of mendelian inherited disease causing mutations has previously focused on the exome (Bamshad et al., 2011), however here I have demonstrated that the same hybridisation capture approach can be utilised in the noncoding genome. Specifically, using evolutionary conservation to predict vertebrate developmental enhancers, this method has taken a subsection of the noncoding genome that is likely to form an important part of the regulatory landscape. This targeted sequencing has provided extensive additional human variation information in an underrepresented region of the genome in previous sequencing efforts (The Genomes Project, 2015). This method resulted in very high coverage of the targeted regions (average = 272), allowing for confident calling of rare variants. In addition, the use of the TidyVar algorithm (Noyvert, 2015) with good quality and depth sequence data resulted in fast and accurate variant calling across the multiple cohorts.

Each clinical cohort was selected based on three key criteria – homogenous symptoms, ethnic background and evidence for a developmentally based disorder. The intellectual disability and epilepsy comorbidity study examined multiple families with children presenting with both of these disorders. Both of these disorders (when non-trauma induced) are understood to often be a result of abnormal foetal brain development (Guerrini and Dobyns, 2014). The clinical notes for the patient samples and families showed many instances of multiple other symptoms and disorders. Therefore, the cohort was mostly used to develop the variant annotation pipeline, including how it responds to various inheritance patterns. As a result, the pipeline relied on a low-information input in regards to patient status: 1 (affected) or 0 (unaffected). This reduction in information input simplified the process removing phenotypic ‘noise’, but it also removed a lot of

detail that could be important in further downstream analysis. Due to the complex presentation of symptoms, disorder-associating variants were not expected to be found, rather familial specific rare mutations. The variant prioritisation pipeline was able to reduce the 1750 CNE variants to 5 mutations to focus on in subsequent analyses.

The cleft lip and palate cohort not only had a homogenous ethnicity as a cohort, but the symptoms were also clinically the same (presentation of cleft lip and palate, the most severe of this disorder). In addition, the use of unaffected parents allowed for simplicity of variant analysis under the mendelian inheritance assumption. The sample mislabelling caused downstream problems with the reliability of data and although some variants of interest were found, functional analysis was abandoned as the results were at risk of being compromised. Sample mislabelling is a problem within the wider scientific community and therefore nodal analysis of shared variants such as that presented here should be a standard quality control measure to identify human error, especially in international and collaborative projects. In total, 64 variants of varying degrees of interest were identified, with a focus on 5 specific SNPs that both followed inheritance patterns and were very rare in publicly available data sets. No variants were found to associate significantly with the disease phenotype.

The Anosmia clinical cohort could be seen as two sub cohorts: the large familial case of Anosmia in the Faroese family, and the multiple European families. The extensive familial data allowed the variant prioritisation pipeline to accurately trace potential variants. It was able to identify two SNPs in a regulatory region near the gene UNCX. These variants were downstream of a CNE, however their close proximity and presence in all affected Faroese individuals suggests a larger haplotype could be being inherited. This result points to haplotype variation just outside of the CNEs, a region that could be important to the chromosomal architecture of the topological associating domain, bringing the CNEs in proximity with each other or the promoter region. Therefore, more research should look into the regions extending from the CNEs, supporting including 1kb either side of their boundaries in future targeted sequencing efforts. For this additional sequencing to be cost-effective, some analysis of these regions for repeat sequences or structural components should be completed first.

Although the evidence for Schizophrenia as a fetal brain development disorder is not completely conclusive (Haijma et al., 2012, Nenadic et al., 2012, Honea et al., 2005, Collin et al., 2013, Gogtay et al., 2011), the general consensus is that genetic predisposition to the disease is a factor. The large clinical cohort used here were all of Pakistani origin, an under-represented population in current human genome variation databases. Therefore, population statistics on some of the key variants found may not be entirely reliable as comparative data sets are not readily available. The scientific community should look to remedy this going forward as 81% of current GWAS studies are of European origin (Popejoy and Fullerton, 2016). Therefore genomics-driven healthcare initiatives may isolate large populations of the world, specifically those in developing countries, if primary research remains as ethnically exclusive as it currently is. Nevertheless, using global variant data available in combination with the large cohort size, high level disease-associating variants were identified. A similar number of non-syndromic individuals would need to be sequenced in such high depth in order to clarify if these variants are truly disease-associating or if they are specific to the ethnic group of the cohort (Pakistani).

In addition to under-represented populations being sequenced here, the CNE regions targeted were also underrepresented regions when comparing coverage of whole genome sequencing efforts, such as that of 1000 Genomes (The Genomes Project, 2015), where high quality exome sequence data is still prioritised. The high-quality coverage and string-based variant calling algorithm used here allows rare variants in these non-coding regions to be accurately and confidently called. This is particularly useful due to the vast number of rare variants found in the CNEs (Figure 23). The development of a streamlined variant processing pipeline greatly reduced the amount of time needed to investigate the variants called in the CNEs. By integrating inheritance patterns, known enhancer data (VISTA (Visel et al., 2006)), a category scoring method (RegulomeDB (Boyle et al., 2012)) and a quantitative scoring method (CADD (Kircher et al., 2014)), only variants showing the most promise for pathogenicity were interrogated further. Once developed using the CLP and IDE cohorts, this method was able to be easily applied to the Anosmia and Schizophrenia cohorts and could be utilised for other similar CNE sequencing cohort studies. It uses the assumption that

conserved non-coding elements all act as regulatory elements, however downstream assays used focus on validating enhancers. Therefore, not all CNEs will give positive results as not all regulatory elements are enhancers, and possibly not all CNEs are active regulatory elements. The exact number of CNEs that are not acting in this way is hard to predict. Nevertheless, most CNEs that have been assayed for enhancer function have given positive results (Grice et al., 2015, Doglio et al., 2013, Parker et al., 2011, Woolfe et al., 2005).

For all 4 cohorts sequenced, multiple novel variants were discovered in the CNEs. Novel variants were defined as those not able to be annotated using wANNOVAR (Chang and Wang, 2012), covering 1000 Genomes (The Genomes Project, 2015), hapmap (International HapMap Consortium, 2007), dbSNP (Sherry et al., 2001) and other publicly available data sets that contribute to gnomAD (Lek et al., 2016). The greatest number of novel variants discovered per individual sequenced was ~5, found in the Schizophrenia cohort. As mentioned previously, this shows the value in deep sequencing both underrepresented cohorts and underrepresented regions of the regulome. Over 1000 new variants were discovered in the CNEs as a result of this approach. This increase in understanding of the variation in these highly conserved regions will help elucidate the grammar and rules surrounding their function. Much larger efforts should be made to cover these regions in phenotypically normal individuals to help compare how much of this variation can be non-syndromic and how much could be pathogenic. It will be exciting to mine the whole genome data being generated from the 100,000 genomes project.

So far, the in-solution hybridisation CNE sequencing approach and downstream computational analysis described has been high throughput. From DNA sample preparation to a list of prioritised variants could take as little as 6 weeks depending on the resources, sequencing platform and computational power available. However, the reason behind attempting to reduce the number of called variants to a prioritised list is due to the lack of a high throughput *in vivo* method of functional validation. Some progress has been made for some tissue specific enhancers (Patwardhan et al., 2012), although comparable methods are, as of yet, unavailable for developmental enhancers.

The transient enhancer assay in zebrafish used here (Fisher et al., 2006) is potentially the most high-throughput method available for *in vivo* functional analysis of enhancer elements. The transient approach allows for enhancer function to be visualised over 24-72hpf. The nature of this method makes comparisons between constructs with subtle differences in presented GFP expression patterns difficult. In addition, the visualisation and microscopy is labour intensive. Advances in automated imaging systems such as the VAST bioimager (Pulak, 2016) may help this process moving forward, but are heavily reliant on good quality zebrafish larvae microscopes if the images are to be used for comparative analysis. The development of dual colour enhancer assays (Bhatia et al., 2015) within a single embryo will also be particularly useful for assaying both putative enhancer element function, and the effect of SNPs found within. These combined with automated imaging could allow much larger numbers of enhancer elements and their mutations to be assayed, with microinjection and the actual zebrafish development being the most time-consuming components.

The transient enhancer assay was utilised to demonstrate a schizophrenia associating variant with a high CADD score and some evidence for transcription factor binding from RegulomeDB was within an active neuronal developmental enhancer. The CNE assayed showed a high degree of specificity, however no difference between the wild-type and variant version of the CNE was found. The CNE had a low positive number of GFP expressing embryos (27%) suggesting the vector integration could be low. Computationally, transcription factor binding site predictions show that the variant chr2:104496685 T>G may have an effect on transcription factor binding. In addition, the specific base is highly conserved amongst all vertebrate species and sits within a topologically associating domain, demonstrating DNA-DNA interactions in that region. Although there is not enough evidence to say that this variant is a causative factor in schizophrenia, it is a viable variant for further analysis including assessing its impact on nearby gene expression in the developing embryo. Conserved noncoding variants may only contribute to some of the phenotype observed and may act as a part of a mutational load on a gene expression pathway. Finding the impact of each variant and attempting to predict the threshold for pathogenicity is particularly difficult, however this method may contribute to the understanding needed to do so. If further investigations were to

suggest this variant (and others identified by this method) severely affects neuronal gene expression during development, a case could be made for utilising the CRISPR-Cas9 genome editing technique to recapitulate the mutation in a model organism and assess its impact and phenotypic presentation.

Chapter 5. Targeted sequencing of mitochondrial DNA in MPV17$^{-/-}$ mice discovers no effect of dNTP insufficiency on mutational load and mtDNA replication fidelity.

Some of this work is published in PLoS Genetics 12, no. 1 (2016): e1005779

"MPV17 loss causes deoxynucleotide insufficiency and slow DNA replication in mitochondria." by Dalla Rosa, Ilaria, et al. (2016).

The authors retain copyright and all co-authors have granted permission for this work to be presented as part of this thesis. Work not performed by Lilian E Hunt has been omitted or clearly identified.

5.1 Background

As shown in the previous chapters, the use of targeted genomic sequencing can be applied in multiple ways to further our understanding of DNA variation in human diseases. Separate to our genomic DNA, mitochondrial DNA is also present in all of our cells and susceptible to variation that can cause human disease. Mitochondrial DNA is different in many ways to genomic DNA, in particular that it is found in many copies within the cell and therefore each of these copies is susceptible to variation from poor replication fidelity. Normally, misincorporation of bases can be up to two orders of magnitude higher than that for the nuclear genome (Marcelino and Thilly, 1999). It has been shown that after a misinsertion has occurred it can then be excised (Johnson and Johnson, 2001). Equimolar dNTP concentrations facilitate correct base pairing and increase replication fidelity. However, an increase in any concentration of dNTPs away from the norm will flood the replisome with a different ratio to the base composition and this can push extension and misinsertion (Song et al., 2005). Therefore, mutations that have an effect on dNTP concentration in the mitochondira can have a profound effect on mtDNA replication fidelity. There are also some genomic mutations that may also affect the mitochondrial replication programme (Alberio et al., 2007) including

those affecting mtDNA polymerase γ which has intrinsic exonuclease activity to excise these misinsertions during replication.

Mitochondrial DNA depletion syndrome (MDS) is a genetically heterogenous condition. It is characterised by decreased activities of respiratory chain enzymes and lower mtDNA copy number. There are several genes that can contain mutations that contribute to MDS (Poulton et al., 2009). These genes generally encode the proteins directly involved in mtDNA replication (Van Goethem et al., 2001, Longley et al., 2006), or factors regulating the homeostasis of the mitochondrial deoxynucleotide pool (Saada et al., 2001). The deoxynucleoside triphosphate (dNTP) pools that are used for mitochondrial DNA replications are found within the mitochondria themselves, separated from the rest of the cell. It is possible to measure mitochondrial dNTP pools (Marti et al., 2012), and in normal mitochondria environments they are found in asymmetrical concentrations, differing slightly between tissues (Mathews, 2006). Maintaining this balance and overall availability of mitochondrial dNTPs is essential for both the rate and fidelity of mtDNA replication (Mathews, 2006). Increases in the asymmetry of the mitochondrial dNTP pools can result in an increased rate of mutation in the mitochondria (Mathews and Song, 2007, Song et al., 2005).

MPV17 is a mitochondrial inner membrane protein whose loss of function phenotype causes mtDNA abnormalities, characterised in human (Uusimaa et al., 2014), mouse (Viscomi et al., 2009) and yeast (Dallabona et al., 2009). The mechanism by which MPV17 loss affects mtDNA is still unclear. Its loss results in low copy numbers of mtDNA, primarily in the liver (Viscomi et al., 2009). Previously, quantification of random mutations in the mitochondrial genome has been performed using restriction digests (Vermulstet al., 2008), however NGS can now be used to characterise mitochondrial genomic DNA heteroplasmy (Huang, 2015). Here we use this method as described previously and improve upon it by utilising a clean mitochondrial prep. Using liver tissue as a model, this work investigates the hypothesis that MPV17 deficiency alters the mitochondrial dNTP pools, causing an increase in mtDNA mutations that leads to low copy numbers. This work contributes to elucidating the true function by which MPV17 loss causes MDS.

5.2 Results

5.2.1 dNTP insufficiency does not alter the mutational load in Mpv17-/- liver mtDNA

Loss of MPV17 causes tissue-specific mtDNA depletion, with lower copy numbers in the liver and deficiencies in respiratory chain, and ATP synthase complexes (Dalla Rosa et al., 2016). In addition, $Mpv17^{-/-}$ mice have both dGTP and dTTP shortages in liver mitochondria (Dalla Rosa et al., 2016), slowing mtDNA replication. The phenotype in mice can be recapitulated in quiescent MPV17 deficient human fibroblasts (Spinazzola et al., 2006). MPV17 deficiency in human and in mice is associated with two tissue-specific mtDNA phenotypes: mtDNA copy number depletion and multiple deletions (Blakely et al., 2012). Therefore, reduced dNTP pools, as seen in MPV17 KO mice liver tissue (Dalla Rosa et al., 2016), could also affect the fidelity of mtDNA replication, causing further problems in the mtDNA RNA or protein products.

To determine the effect of the reduced dNTP pools on mtDNA fidelity, deep sequencing of purified mtDNA from the livers of two pairs of WT and Mpv17-/- mice was performed. The sequencing coverage was comprehensive for all the samples, with a small trough in the vicinity of the large non-coding region (Figure 40). With this method, the detection of 5% heteroplasmy is possible with a coverage of 1000-fold (Huang, 2015) with increasing sensitivity expected at higher levels of coverage. One sample (KO4, Table 17) did not surpass this average coverage requirement however all other samples did by substantial amounts.

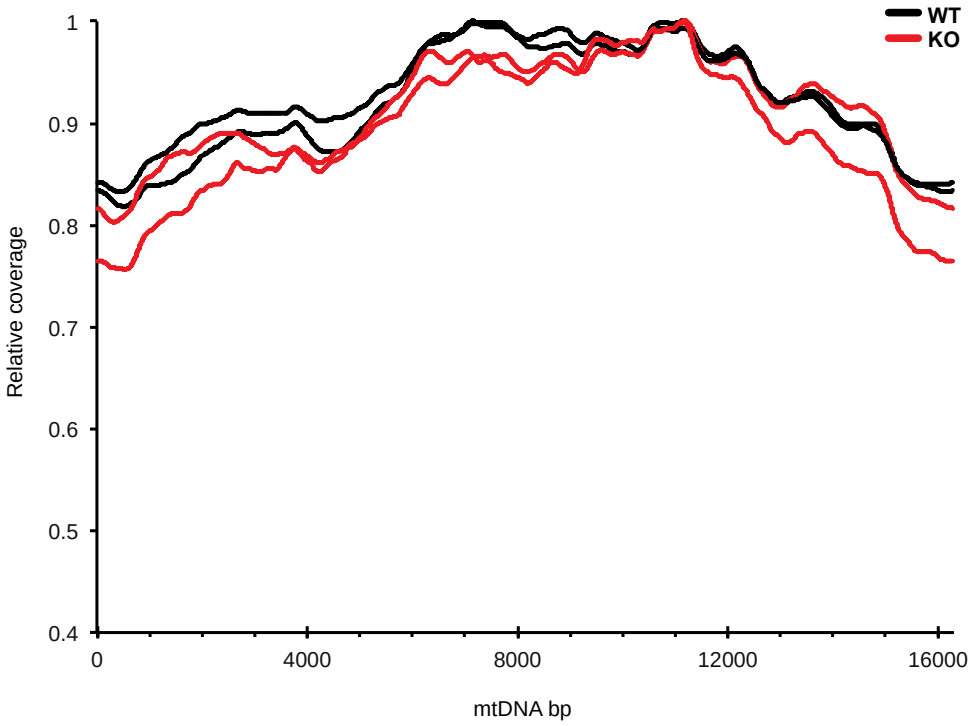


Figure 40. Mouse mtDNA samples sequence coverage.
The mitochondrial genome position (x- axis) versus sequence coverage divided by maximum coverage for each sample (relative coverage, y-axis). The coverage was calculated using a 2kb sliding window average. MtDNA of the WT and KO samples are indicated in red and black.

The error rates for the wild-type and knockout mice were similar; for one pair, the knockout mouse had a slightly lower error rate than the wild-type littermate (0.033% v 0.043%), and in the other pair a 1.7 fold higher error rate was observed in the knockout mouse (Table 17, run 1). The read depth was lowest in the second knockout animal; a replica experiment produced greater depth and confirmed the error rate as higher than the paired control (Table 17, run 2). The error rates for the four individual bases differed to similar extents in all four mtDNA samples ($P > 0.05$ using one-way ANOVA), with dGTP consistently the lowest and dATP the highest (Figure 41). Therefore, the dNTP insufficiency in the Mpv17-/- mouse appears to have little or no effect on the fidelity of mitochondrial DNA replication.

Table 17. Mutational load in purified liver mitochondrial DNA of Mpv17$^{-/-}$ mice and controls.

Misincorporation of bases is inconsistent between sequencing runs, therefore comparisons can only be made within run 1 and run 2 separately. ML- Mutation Load per site frequency. KO – Mpv17-/-, WT – wild-type littermates of KO mice. Individual bases are shown as the number of the misincorporated allele divided by total bases.

Run	Sample	Total bases	ML	A	C	G	T
1	WT1	8.76E+07	4.30E-04	1.39E-04	1.07E-04	8.20E-05	9.90E-05
1	KO2	2.50E+07	3.30E-04	1.00E-04	8.60E-05	6.70E-05	7.70E-05
1	WT3	1.13E+08	3.40E-04	1.10E-04	7.80E-05	6.20E-05	8.60E-05
1	KO4	5.97E+06	5.80E-04	2.22E-04	1.19E-04	8.20E-05	1.53E-04
2	WT3	1.82E+08	9.90E-04	3.19E-04	2.61E-04	1.58E-04	2.36E-04
2	KO4	1.16E+07	2.21E-03	7.73E-04	5.07E-04	3.40E-04	5.66E-04

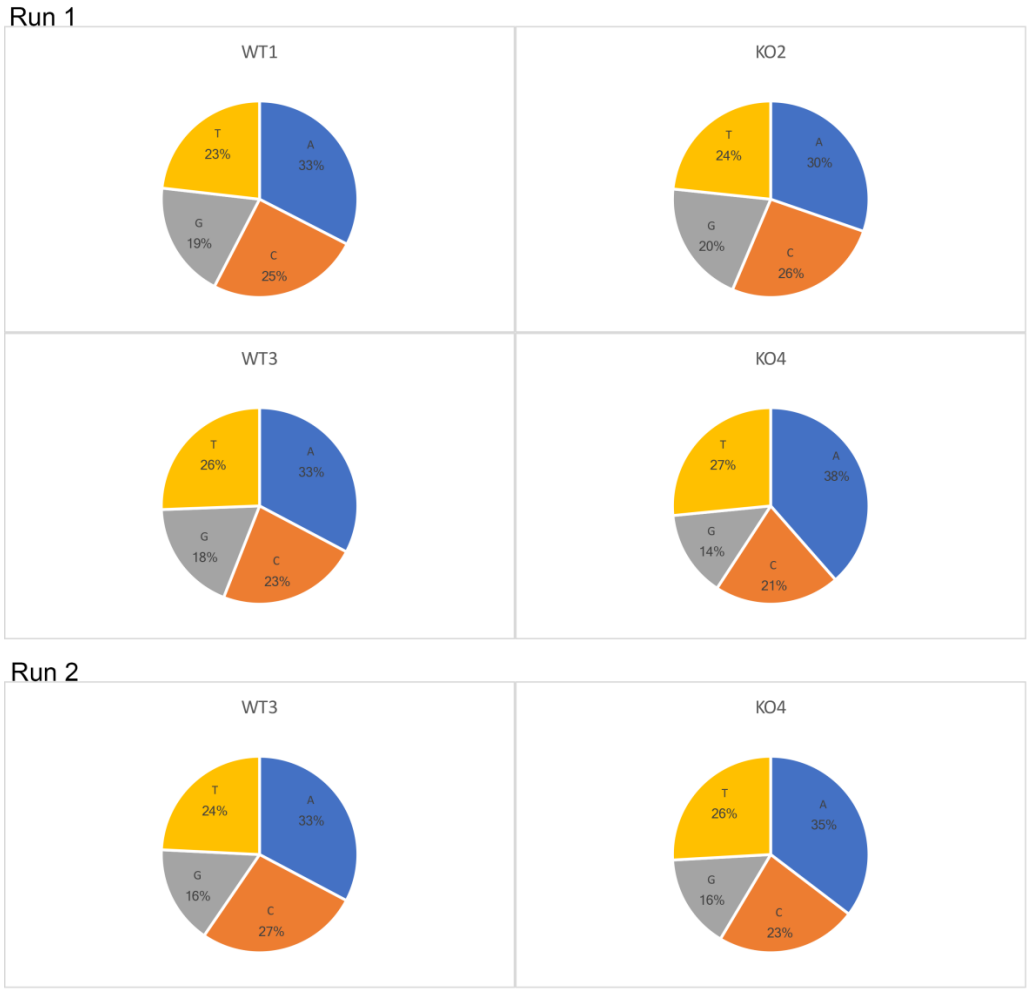


Figure 41. Proportion of misincorporated bases shown as proportions per base. No significant difference can be found between WT and KO samples.

5.2.2 MPV17 deficiency does not alter the mutant load of brain mtDNA.

MPV17 deficiency is known to cause mitochondrial copy number depletion in the liver (Spinazzola et al., 2008), however a decrease in copy number is not seen in the brain. Despite this, MDS does display a neurological phenotype (Spinazzola et al., 2008), therefore similar analysis was performed on mouse MPV17$^{-/-}$ and WT brain tissue. Conventional next generation sequencing (that does not detect rNMPs) was applied to libraries prepared from three controls (CM1-3, MPV17$^{+/+}$) and three MPV17$^{-/-}$ (CM4-6) mice of 3 months of age. There was no significant difference in the misincorporation of bases (those differing from the reference sequence) between the two groups of mice (Figure 42, Table 18).

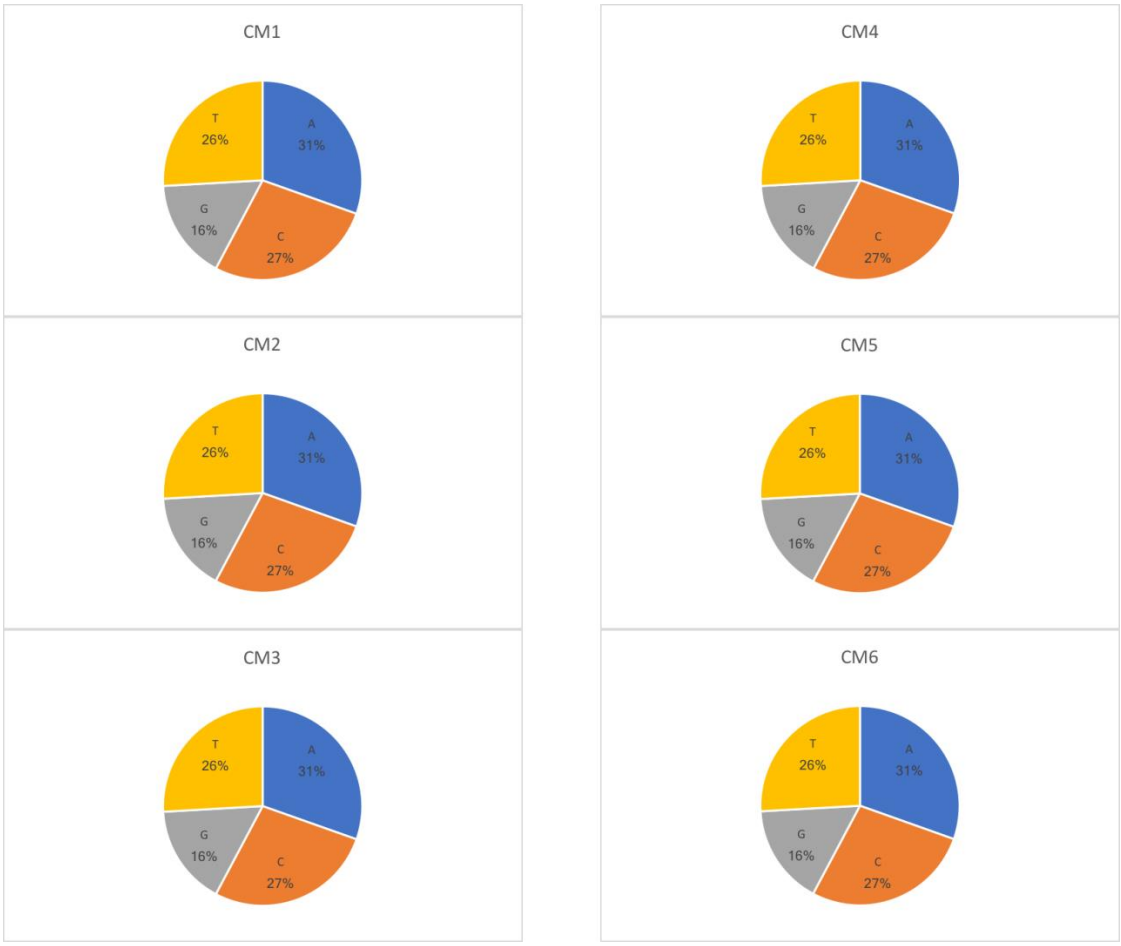


Figure 42. Proportion of misincorporated bases broken down by base.
CM1 – WT B (2 m/o); CM2 – WT B (3.5 m/o); CM3 – WT B (3.5 m/o); CM4 – MPV17 KO B (2 m/o); CM5 – MPV17 KO B (3.5 m/o); CM6 – MPV17 KO B (3.5 m/o)

Table 18. Mutational load in purified brain mitochondrial DNA of Mpv17$^{-/-}$ mice and controls.

Sample	Total bases	ML (all bases)	A	C	G	T
CM1	6.61E+08	1.26E-03	3.82E-04	3.43E-04	2.05E-04	3.25E-04
CM2	1.51E+08	5.51E-03	1.68E-03	1.51E-03	8.98E-04	1.43E-03
CM3	3.19E+08	2.60E-03	7.91E-04	7.12E-04	4.24E-04	6.74E-04
CM4	3.76E+08	2.21E-03	6.71E-04	6.04E-04	3.60E-04	5.72E-04
CM5	1.02E+08	8.16E-03	2.48E-03	2.23E-03	1.33E-03	2.11E-03
CM6	2.73E+08	3.04E-03	9.26E-04	8.33E-04	4.96E-04	7.88E-04

In addition, the spread of mutational load across the mtDNA was checked for differences between WT and KO mouse models. A comparison of misincorporated bases at each position as a factor of all reads at each position was measured for each mouse sample (Figure 43). This varied between mice and showed no significant difference between the two genotypes.

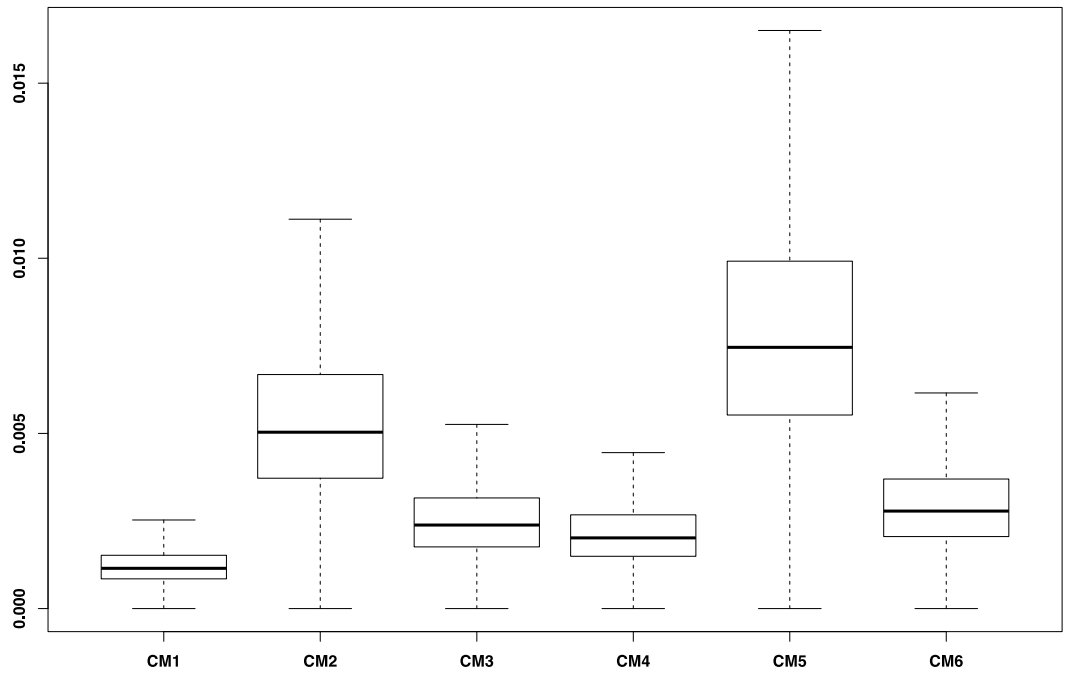


Figure 43. The rate of misincorporation of bases across each position of the mitochondrial genome in mouse brain samples.
Number of reads with mutation/total number of reads at each base as a box plot for each sample. Outliers have been removed for ease of visualisation. No significant difference is seen between the WT (CM1-3) and the KO (CM4-6).

Further to this result, unpublished data (Moss *et al.*) shows that the dNTP pools in MPV17$^{-/-}$ brain tissue are asymmetric and the same as that of the wild-type mice. This sequence data when presented alongside the dNTP pool information (work performed by C. Moss & I. Dalla Rosa) provides strong evidence that the decrease of dGTP and dTTP in MPV17 deficient mice liver does not increase mtDNA replication infidelity.

5.3 Discussion and conclusions

Targeted sequencing of regions of the genome may elucidate disease-causing variants. Where coding variants are already known, other methods of targeted sequencing can still provide crucial information to revealing the disease mechanism. This is especially true when the mutation leads to problems with the internal replication of the cell and its organelles. Here, isolation of mitochondrial DNA and its sequencing has significantly contributed to the understanding to the MPV17 phenotype, mitochondrial depletion syndrome. Alongside other work establishing dNTP insufficiency in the mitochondria as the cause of mitochondrial depletions in MPV17 deficiency, I show through sequence analysis that the mutational load in mitochondria is unaffected by the MPV17 gene mutation.

Previously studies have suggested that the mitochondrial genome instability caused by loss of MPV17 could be a result of changes in the dNTP pool available, which would fit with its function as an inner membrane protein (Spinazzola et al., 2006). This work has contributed to further understanding of MPV17 loss (Dalla Rosa et al., 2016) and utilised mouse models to reflect a highly detrimental human disorder. However, despite previous work showing that MPV17 loss causes dNTP insufficiency (Dalla Rosa et al., 2016), here it is shown that this does not affect the mutational load in liver or brain mtDNA. Imbalances of mitochondrial dNTP pools affect replication fidelity (Nishigaki et al., 2003) and have been shown to lead to a higher rate of mutation by mitochondrial DNA (Mathews, 2006, Song et al., 2005). As a result of the work presented here, MPV17 KO mice dNTP pools were measured and shown to be close to equimolar (Dalla Rosa et al., 2016). Therefore, the reduction in availability of some dNTPs has not increased the mutation rate, but rather slowed the whole process of replication. This suggests that the mechanism for MDS from MPV17 loss is from a slower rate of mtDNA replication, as suggested in a previous *in organello* model (Gonzalez-Vioque et al., 2011) rather than an increase in mtDNA mutations.

Additional validation of this increase in mutational load seen here could be performed by including a positive control – a mouse model with a known increased mutational

load in the mitochondrial DNA. The difficulty with this is that a comparative model may be hard to find as the rate of base misincorporation will vary depending on the mechanisms and severity of the mutation.

Chapter 6. Conclusions

The work presented here demonstrates three very distinct and non-conventional approaches to genome sequencing. All approaches share the common theme of searching for non-coding variation that could explain various diseases and disorders.

In the first instance (Chapter 3), targeted deep sequencing of a topologically associated domain containing known non-coding GWAS variants associated with obesity was used. This gave more detail into the human genetic variation in this 2Mb region which has been shown previously to self-interact as a regulatory block (Dixon et al., 2012). Using a very distinct ethnic cohort, novel associations were found between variants in the regulatory regions, extending further than previous interaction data (Smemo et al., 2014). This novel association peak appears to be age-dependent and previously unseen. This demonstration of extending variant associations across such a large distance demonstrates the need for a 3D perspective of the genome, especially in relation to gene regulation. In addition, this method of deep sequencing of topologically associating domains could be utilised elsewhere as a method for determining the region of variation associating with a disease phenotype and the distance at which these variants could interact. A similar technique has been implemented since the publication of this work (Sobalska-Kwapis et al., 2017) demonstrating different association signals in different ethnic populations.

In the second instance (Chapter 4), conserved non-coding elements (CNEs) previously identified using pufferfish-human genomic alignments were sequenced in four cohorts presenting different developmentally based disorders: cleft lip and palate, intellectual disability and epilepsy, anosmia, and schizophrenia. The CNEs were predicted enhancer elements conserved in the vertebrate genome that cluster around developmentally active genes (Woolfe et al., 2005), suggesting a crucial role in regulating vertebrate development. This targeted approach was used to attempt to elucidate non-coding genetic variants that either associated with the phenotype or followed a familial inheritance of the phenotype where exome sequencing had not discovered any conclusive pathogenic variants. This method identified many previously

unseen human SNVs within the noncoding genome, adding to the global map of known human genetic variation. In addition, variants were scored and prioritised by their potential for pathogenicity in an easily to follow pipeline that could be replicate between differing cohorts.

Finally, the third use of targeted deep sequencing (0) used very high coverage mitochondrial DNA targeted sequencing in mouse models of a human mitochondrial depletion syndrome: MPV17 loss. Using such high depth of coverage, mutation loads in these mice were able to be calculated and used to show how MPV17 depletion does not result in an increase of misincorporated bases in the mitochondrial DNA (Dalla Rosa et al., 2016). This proof of hypothesis allowed the true function of MPV17 loss to be determined and show that the decrease in dGTP and dTTP seen in $Mpv17^{-/-}$ mice does not affect the mtDNA replication fidelity. This was shown in both liver and brain tissue despite a reduction in mtDNA copy number in both these tissues in the mouse model. Previously, the contribution of MPV17 loss to MDS was unknown however it has now been shown that it causes deoxynucleotide insufficiency which in turn slows the rate of DNA replication in mitochondria. This method could also be applied to other mitochondrial depletion syndromes where an accumulation of mitochondrial mutations is hypothesised to be a driving factor to MDS.

All three non-conventional sequencing methods demonstrated their merit and the depth of information that can still be found in the non-exonic portions of the genome. All three methods also contributed to the further understanding of human diseases and disorders. Previously, much of the focus on coding mutations and their consequences have led to many genetic disorder diagnoses. Projects such as this also demonstrate the value in trying to understanding the non-coding portions of the genome and the regulome. The variant information included in this work will be made publicly available through the European Nucleotide Archive. Future work could further validate and consolidate these variants as well as incorporate them into other databases such as GenomAD.

The work in Chapter 3 and Chapter 4 is driven by the concept of long-range genomic regulation. Advances in understanding the 3D architecture of the genome, its spatial organisation and its compartmentalisation drive the research behind non-coding regulation. Previously distances between interacting elements of DNA were thought to be limited, however it is now known that regulatory elements can act on promoters from vast distances (Antonellis et al., 2006, Lettice et al., 2003, Lieberman-Aiden et al., 2009, Loots et al., 2005, Ragvin et al., 2010, Smemo et al., 2014). This work focuses on cis-regulatory elements, and even the clustering of CNEs around developmental genes is within a relatively small size window. There is potential for regulatory elements at much greater distances to be found, and even elements acting in trans from another chromosome. How the non-coding genome folds in on itself, its stability (or lack thereof) and how this dictates key processes such as development is still largely unknown.

Much work could be put into the creation of vertebrate models for different predicted pathogenic variants found, however this is costly and time ineffective. Dissecting the exact architecture of known regulatory elements could help determine their how they function. Though if transcription factor binding is thought to be the critical influence then every regulatory element may not follow the same set of rules. By understanding SNV influences within these regulatory elements and their link to diseases and disorders, it is hoped that the critical bases in gene regulation can be identified. Through this methodology, a large database of functionally relevant noncoding SNVs could be curated that can help sift through the non-coding genome. Once this information database is big enough, we can hope to understand the patterns around these individual bases and the 3D architecture of the genome that can interact. Inevitably, whole genome data will increase to hundreds of thousands of individuals in the near future, perhaps even millions. Mining these genomes using some of the approaches developed here will lead to a single base understanding of regulatory regions, and perhaps ultimately a much finer resolution mapping of TFBS. With this information, some dent could be made in the deciphering of the regulatory code of the human genome.

Chapter 7. Appendix

7.1 Appendix Table 1

Table 19. Obesity cohort information

IID	Case/Control (1/2)	Age	BMI
3909	1	19	20.9
4337	1	22	25.5
4347	2	19	34.0
4348	2	26	31.6
4353	2	19	32.4
4358	2	19	33.6
4362	1	19	22.3
4363	1	24	21.2
4375	1	24	23.9
4378	1	21	17.1
4380	1	19	20.6
4393	2	19	31.7
4395	2	19	31.9
4396	2	19	31.9
4397	1	18	21.1
4412	1	19	20.7
4415	1	23	21.1
4416	2	19	35.7
4462	1	19	17.8
4607	2	19	34.5
4649	2	20	36.7
4652	2	19	33.9
4653	1	19	19.8
4658	1	20	21.4
4662	1	20	26.2
4670	2	20	31.1

4672	2	19	34.2
4673	2	18	31.6
4675	2	19	32.7
4683	1	19	21.3
4704	2	22	32.8
4709	1	19	24.0
4714	2	20	31.2
4716	1	20	21.3
4719	2	24	35.4
4723	2	19	32.8
4727	2	19	31.6
4749	1	18	20.7
4751	1	19	21.4
4755	1	19	21.5
4772	1	18	19.9
4776	1	19	20.2
4779	1	19	18.9
4780	2	20	31.6
4781	1	19	17.8
4782	1	19	24.5
4792	1	25	19.6
4798	1	20	21.1
4826	2	21	32.5
4846	2	19	31.6
4860	1	19	21.7
4990	2	19	32.0
5301	1	20	20.3
5305	1	18	21.8
5308	2	19	32.7
5370	1	23	22.0
5373	1	19	21.4

5374	1	19	21.3
5397	2	18	33.0
5399	2	20	33.8
5407	2	19	33.5
5444	1	19	22.3
5469	1	20	20.7
5477	1	19	21.6
5479	2	19	32.1
5481	1	20	20.7
5482	2	20	32.1
5508	1	20	21.1
5525	2	20	37.4
5533	1	21	18.3
5550	1	20	20.7
5553	1	18	29.7
5583	1	26	20.9
5623	1	20	22.4
5695	2	20	32.0
5706	2	19	32.0
5901	1	18	23.6
6001	2	19	34.6
6884	2	20	32.8
6898	1	18	24.3
6903	2	19	33.1
6906	1	19	25.9
6908	1	19	23.3
6910	1	18	23.8
6923	2	21	31.8
6924	2	19	35.3
6932	1	20	21.1
6934	1	19	18.6

6964	2	18	36.3
6976	1	21	20.4
6981	1	20	21.2
7088	1	19	20.5
7215	2	19	31.5
7257	2	19	31.1
7269	1	18	26.3
7277	2	19	31.1
7294	2	19	31.2
7368	1	20	21.9
7404	1	19	19.8
7406	1	19	21.4
7408	1	20	24.8
7593	1	19	22.3
7647	1	28	25.1
7653	1	19	23.2
7928	2	19	32.8
7941	2	18	33.9
8048	1	19	20.0
8051	2	19	33.3
8072	2	19	33.6
8155	1	19	18.4
8277	2	18	31.2
8289	1	19	21.2
8501	1	20	22.9
8552	2	19	31.8
8553	1	20	22.6
8555	1	19	17.2
8560	2	19	43.3
8726	1	21	17.8
8792	2	25	33.0

8807	1	19	17.9
8880	1	21	24.2
8882	1	21	24.3
8921	1	18	20.9
8928	2	19	35.3
8934	2	19	32.5
8994	1	18	25.6
9016	1	19	18.9
9017	2	19	32.3
9075	1	19	20.0
9078	1	22	23.1
9108	2	19	51.8
9424	1	20	20.3
9472	2	19	31.5
9664	1	18	20.9
9769	2	19	31.5
9785	1	19	22.3
9789	1	20	19.3
9805	2	19	34.5
9808	1	19	24.2
9882	2	23	32.7
9913	2	19	33.3
9924	1	24	20.7
9957	1	19	18.0
10577	2	19	35.6
10600	2	20	33.8
10645	2	18	40.4
10713	1	19	21.7
10741	1	21	20.3
10769	2	20	34.3
10811	1	20	17.4

10835	1	18	20.9
10858	1	26	20.4
10875	1	18	19.7
10878	1	19	17.8
10881	2	19	31.4
10885	1	18	21.5
10886	1	18	26.2
10889	1	19	19.1
10898	2	21	31.4
10899	1	19	21.0
10905	2	19	31.4
10911	1	19	22.4
10912	1	20	22.9
10913	1	21	28.1
10915	2	19	31.7
10917	1	18	20.3
10942	1	18	22.1
10944	1	19	18.6
10948	1	19	19.4
10957	2	19	32.7
10971	2	19	35.1
10978	1	20	21.4
10982	1	19	19.4
10986	1	19	20.8
10989	2	20	33.0
11006	1	20	19.2
11012	1	20	21.8
11022	2	19	31.5
11025	1	21	25.6
11059	2	20	36.2
11065	2	19	36.6

11074	1	20	18.8
11078	1	19	18.2
11082	1	19	22.9
11089	2	19	31.3
11093	2	18	32.1
11096	1	21	21.5
11101	1	20	19.3
11103	2	19	31.5
11115	1	26	22.9
11124	1	20	19.8
11129	1	19	21.7
11135	1	20	19.6
11141	2	19	32.2
11145	1	20	24.8
11164	1	19	22.3
11167	1	20	21.1
11176	2	18	31.3
11183	1	19	19.1
11189	1	19	18.6
11205	1	19	22.1
11214	2	19	31.8
11234	2	21	31.0
11239	1	19	20.8
11240	2	22	31.1
11266	1	19	19.1
11276	1	19	15.9
11277	1	26	22.3
11286	2	20	31.4
11287	1	19	23.1
11297	2	19	34.0
11304	2	19	32.1

11309	2	19	34.8
11323	2	26	32.1
11324	2	20	35.0
11332	2	19	32.5
11334	1	20	22.4
11345	1	19	20.8
11348	1	19	21.0
11349	1	21	25.1
11350	2	27	35.2
11353	1	18	18.8
11356	1	21	26.6
11359	2	19	35.0
11373	1	19	24.4
11381	2	19	35.0
11383	1	19	24.3
11384	1	31	23.4
11399	1	18	22.9
11402	2	18	31.1
11408	1	19	22.5
11410	2	20	33.0
11416	1	19	19.6
11429	2	25	31.1
11439	2	20	36.2
11440	2	19	31.2
11442	2	19	37.9
11448	1	20	19.3
11453	2	18	31.3
11459	2	18	33.5
11478	2	19	35.4
11481	2	21	34.7
11497	2	19	32.0

11500	2	19	36.2
11507	2	19	31.6
11509	2	19	33.2
11512	2	19	32.5
11515	1	19	20.5
11519	1	21	23.8
11533	1	22	24.6
11535	2	26	31.6
11538	1	19	22.0
11555	2	19	31.5
11568	1	19	22.9
11570	2	19	31.7
11585	1	22	24.8
11586	2	19	31.6
11595	2	19	33.8
11602	1	25	19.3
11616	2	20	31.1
11623	1	20	19.8
11631	1	24	26.5
11642	2	19	31.2
11643	2	20	31.2
11655	2	21	32.1
11660	2	19	33.3
11666	1	23	25.0
11668	1	19	21.6
11672	2	20	31.3
11673	2	19	31.1
11675	2	19	32.3
11681	1	20	21.7
11688	2	19	32.7
11727	2	18	31.6

11736	2	19	32.5
11741	1	18	20.2
11753	1	23	18.7
11755	1	19	18.8
11756	1	20	21.4
11762	1	18	23.1
11763	1	20	19.8
11774	1	19	23.1
11779	1	20	19.2
11780	1	19	22.5

7.2 Appendix Table 2

Table 20. Obesity haplotypes

Start	End	Tagged SNP region	Freq.	Case, Control Frequencies	P Value
53606229	53739773	587-1267			
			0.264	0.315, 0.233	0.0327
53755146	53759123	1330-1349			
			0.516	0.577, 0.469	0.0105
			0.313	0.260, 0.354	0.0169
53767959	53771583	1379-1398			
			0.523	0.585, 0.475	0.0092
			0.299	0.252, 0.335	0.0316
53772346	53772626	1406-1407			
			0.525	0.589, 0.475	0.0069
			0.475	0.411, 0.525	0.0069
53774903	53786446	1427-1469			
			0.396	0.451, 0.354	0.019
			0.375	0.325, 0.413	0.0321
53793798	53795636	1510-1526			
			0.456	0.504, 0.419	0.0442
53798523	53798622	1542-1543			
			0.482	0.556, 0.424	0.0018

			0.477	0.419, 0.522	0.0148
53799296	53843533	1546-1827			
			0.293	0.384, 0.232	0.0001
53844579	53845487	1834-1841			
			0.439	0.362, 0.497	0.0013
			0.434	0.532, 0.360	0.000041
54010398	54019686	2802-2844			
			0.198	0.150, 0.235	0.0114
54211937	54213893	4008-4016			
			0.29	0.340, 0.251	0.0211
54214069	54214702	4018-4023			
			0.289	0.337, 0.252	0.0253
54268659	54271085	4349-4366			
			0.177	0.228, 0.138	0.0051
54272047	54306215	4371-4583			
			0.054	0.082, 0.034	0.0143
54327852	54328675	4707-4712			
			0.067	0.093, 0.047	0.0266
54532641	54537608	6137-6157			
			0.06	0.089, 0.037	0.0094
54542033	54546604	6183-6198			
			0.468	0.419, 0.506	0.0384
54736811	54741807	7291-7311			
			0.202	0.158, 0.236	0.0216
54753168	54774171	7389-7529			
			0.294	0.248, 0.330	0.0351
54777074	54807769	7565-7775			
			0.206	0.254, 0.171	0.016
54808449	54813519	7780-7809			
			0.295	0.356, 0.248	0.0055
54813801	54817371	7812-7842			
			0.528	0.598, 0.475	0.0038
54818762	54856786	7850-8069			
			0.412	0.493, 0.362	0.0021
			0.125	0.089, 0.156	0.019
54856933	54857871	8071-8080			
			0.584	0.638, 0.543	0.0232

54885318	54963258	8251-8666			
			0.132	0.167, 0.109	0.0466
			0.051	0.075, 0.035	0.0382
55136446	55139619	9771-9792			
			0.107	0.077, 0.130	0.0424
55140657	55141022	9799-9802			
			0.884	0.915, 0.860	0.0451
			0.116	0.085, 0.140	0.0451
55226745	55234048	10439-10508			
			0.085	0.118, 0.059	0.0124
55234072	55276574	10509-10821			
			0.1	0.131, 0.079	0.0479
55279211	55291144	10839-10899			
			0.13	0.167, 0.102	0.0243
55461759	55461854	11927-11930			
			0.655	0.703, 0.618	0.0342

7.3 Appendix Table 3

Table 21. Cleft lip/palate cohort sample and sex information

			cleft type	cleft type	cleft type
	Eurocran trio code	sex	lip	jaw	palate
H008	H008	M	Y	Y	Y
H009	H009	M	Y	Y	Y
H015	H015	F	Y	Y	Y
H019	H019	F	Y	Y	Y
H027	H027	M	Y	Y	Y
H029	H029	M	Y	Y	Y
H032	H032	F	Y	Y	Y
H042	H042	M	Y	N	Y
H054	H054	M	Y	Y	Y
H057	H057	F	Y	Y	Y
H058	H058	M	Y	Y	Y
H101	H101	M	Y	Y	Y

H105	H105	M	Y	Y	Y
H107	H107	M	Y	Y	Y
H109	H109	F	Y	Y	Y
H111	H111	F	Y	Y	Y
H122	H122	M	Y	N	Y
H124	H124	M	Y	Y	Y
H129	H129	M	Y	Y	Y
H136	H136	M	Y	Y	Y
H137	H137	M	Y	Y	Y
H144	H144	M	Y	Y	Y
H170	H170	M	Y	Y	Y
H174	H174	M	Y	Y	Y
H186	H186	M	Y	Y	Y
H195	H195	M	Y	Y	Y
H196	H196	F	Y	Y	Y
H198	H198	M	Y	Y	Y
H201	H201	F	Y	Y	Y
H204	H204	M	Y	Y	Y
H207	H207	F	Y	Y	Y
H216	H216	F	Y	Y	Y
H218	H218	F	Y	Y	Y
H222	H222	M	Y	Y	Y
H236	H236	M	Y	Y	Y
H238	H238	F	Y	Y	Y
H239	H239	F	Y	Y	Y
H243	H243	M	Y	Y	Y
H251	H251	M	Y	Y	Y
H278	H278	M	Y	Y	Y
H284	H284	M	Y	Y	Y
H285	H285	M	Y	Y	Y
H287	H287	F	Y	Y	Y

H288	H288	F	Y	Y	Y
H289	H289	M	Y	Y	Y
H290	H290	M	Y	Y	Y
H293	H293	F	Y	Y	Y
H303	H303	M	Y	Y	Y
H304	H304	M	Y	Y	Y
H311	H311	F	Y	Y	Y
H317	H317	M	Y	Y	Y
H321	H321	M	Y	Y	Y
H322	H322	M	Y	Y	Y
H325	H325	F	Y	Y	Y
H327	H327	F	Y	Y	Y
H330	H330	M	Y	Y	Y
H333	H333	M	Y	Y	Y
H340	H347	M	Y	Y	Y
H342	H352	M	Y	Y	Y
H347	H356	M	Y	Y	Y
H359	H359	F	Y	N	Y

7.4 Appendix Table 4

Table 22. Novel Schizophrenia cohort variants

Chr	Start	End	Ref	Alt	Cohort AF
1	90601727	90601727	T	-	0.8981
19	31770873	31770873	-	A	0.7549
8	106155991	106155990	-	A	0.7071
2	145202854	145202854	T	-	0.614
10	124828491	124828491	T	-	0.5486
19	31831244	31831244	-	A	0.5444
2	172956771	172956771	-	A	0.5357
16	79631698	79631698	-	T	0.5171
7	70257475	70257475	A	-	0.4659

15	37180386	37180386	G	-	0.4566
3	147099886	147099886	T	-	0.4153
18	73370970	73370979	GTTTTCTTTC	TTTTTTTTTT	0.3916
10	102501168	102501171	AAAA	-	0.3819
X	612330	612330	-	A	0.3518
2	59541122	59541122	A	-	0.3508
13	72333065	72333065	T	-	0.3507
2	59541121	59541122	AA	-	0.3325
15	98292748	98292755	ACACACAC	-	0.3109
5	158124282	158124282	T	-	0.2937
10	124828491	124828491	-	T	0.2919
3	181328198	181328198	A	-	0.2905
3	147099885	147099886	TT	-	0.2814
7	1308935	1308935	-	T	0.278
11	8351446	8351459	TTCCCCCCCCCCCA	ATCCCCCCCCCC	0.2464
9	16704711	16704711	T	G	0.2383
2	145189567	145189567	-	A	0.2365
10	78062966	78062966	T	C	0.2354
16	73093074	73093074	G	A	0.2325
6	10398152	10398152	T	-	0.2283
13	100547026	100547026	G	T	0.2271
3	181328197	181328198	AA	-	0.2264
8	71963499	71963499	T	G	0.223
13	72333065	72333065	-	T	0.2227
1	91300727	91300727	T	C	0.2195
17	35061677	35061677	A	G	0.2178
15	98292746	98292755	ACACACACAC	-	0.2176
17	35061685	35061685	A	T	0.2174
1	87800568	87800568	A	C	0.215
5	4628063	4628063	T	G	0.2133
15	98292742	98292755	ACACACACACACAC	-	0.2073
13	100547035	100547035	A	G	0.2063

6	10398132	10398132	G	T	0.2036
16	54323650	54323650	T	C	0.2019
3	181328195	181328198	AAAA	-	0.1926
1	10702242	10702242	A	G	0.1881
2	60685089	60685089	T	-	0.1854
20	21378185	21378185	T	G	0.1846
X	835430	835430	-	A	0.1845
13	36104481	36104481	T	C	0.1749
13	100547039	100547039	C	T	0.1723
10	102501169	102501171	AAA	-	0.1566
6	10397894	10397894	-	A	0.1485
5	158123012	158123012	-	A	0.1348
5	158123012	158123012	A	-	0.1275
1	63553982	63553982	A	C	0.1263
1	87801555	87801555	-	A	0.1232
3	181328196	181328198	AAA	-	0.1216
20	22565488	22565488	G	T	0.1178
3	17988437	17988437	-	T	0.1158
20	51804117	51804120	TCCC	CCCT	0.1043
X	612330	612330	-	AA	0.103
9	37034270	37034270	A	T	0.0936
15	98292744	98292755	ACACACACACAC	-	0.0907
2	60685410	60685410	-	T	0.0838
18	73370975	73370979	CTTTC	TTTTT	0.0837
1	18968447	18968447	T	C	0.0704
15	98292752	98292755	ACAC	-	0.0674
18	73370979	73370980	CT	-	0.064
15	98292750	98292755	ACACAC	-	0.0596
11	16121254	16121254	-	T	0.0478
16	73747539	73747539	T	-	0.0463
19	31915104	31915103	-	T	0.0417
10	124828491	124828491	-	TT	0.0405
15	96812001	96812001	A	-	0.0393

11	8351448	8351447	-	C	0.0386
13	95547310	95547310	T	-	0.0383
1	87801555	87801555	-	AAAA	0.0379
2	60685089	60685089	-	T	0.0366
2	172956771	172956771	-	AA	0.0357
13	100546987	100546993	AGGCGGG	GGGC	0.0332
X	612330	612330	-	AAA	0.0327
8	72110674	72110674	-	A	0.0291
7	70257602	70257602	-	T	0.0286
19	31915104	31915104	T	-	0.027
5	158124282	158124282	-	T	0.0267
10	131668257	131668257	-	A	0.0246
13	100546987	100546994	AGGCGGGG	GGGC	0.023
15	98292754	98292755	AC	-	0.0207
1	18968455	18968455	-	C	0.0201
X	612330	612330	-	AAAA	0.0201
11	33376243	33376246	AGGG	GGGT	0.0191
15	37183275	37183275	-	T	0.0191
15	98292740	98292755	ACACACACACACACAC	-	0.0181
13	100546987	100546995	AGGCGGGGG	GGGCGGGGGGC	0.0179
18	73370970	73370979	GTTTTCTTTC	TTTTTTTTT	0.0172
3	181328198	181328198	-	A	0.0169
19	30714131	30714131	-	A	0.0143
7	156407220	156407224	TCTGA	G	0.0123
16	51671786	51671786	A	G	0.0122
11	8351446	8351459	TTCCCCCCCCCCCA	ATCCCCCCCCCCCC	0.0121
3	71629324	71629324	-	A	0.0118
14	97204816	97204818	TAG	-	0.0114

7.5 Intellectual Disability and Epilepsy comorbidity sample and phenotype information

Table 23. IDE cohort clinical notes

7782 (unclassified progressive epilepsy – died two years old) scn8a
7787 mother
7786 father
EG0522 (epilepsy, autisme, intellectual disability, athrititis)
EG0521 mother
EG0523 father
EG1238 (Myoclonic astatic epilepsy, learning disability), COCO E1
EG1239 (Myoclonic astatic epilepsy, learning disability) (sister)
EG0603 (brother) (autisme)
EG0435 (mother) (depression)
EG0434 (father)
EG0540 (Juvenile myoclonic epilepsy), COCO E11
EG0542 (brother) (Juvenile myoclonic epilepsy)
EG0541 (mother)
EG0539 (father) (migraine, dyslexia)
EG0609 (epilepsy, schizophrenia, ADHD, tics) COCO E17
EG0680 (mother)
EG0819 (father)
EG0499 (Childhood absence epilepsy, ADHD), COCO E22
EG0871 (brother, Childhood absence epilepsy)
EG1253 (mother, Obsessive compulsive disorder, panic attacks)
EG0498 (father)
EG0662 (sister, epilepsy, autisme, aggresive behaviour, intellectual disability, immune defect, CP, 2 small duplications at 20p12.1) COCO E33
EG0716 (sister, epilepsy, autisme, aggresive behaviour, intellectual disability, immune defect, CP, 2 small duplications at 20p12.1)
EG1010 (mother, depression)

EG0663 (father, migraine, 2 small duplications at 20p12.1)
EG0343 (Childhood absence epilepsy, atypical autisme) COCO E39
291202 (brother, Infantil autist)
EG0344 (mother)
EG0330 (father)
EG0681 (Attension dificiet disorder, epilepsy) COCO E42
EG0694 (sister, anxiety, cutting, alice in wonderland syndrome)
EG0723 (sister, Obsessive compulsive disorder, anorexia, speech delay
EG0683 (mother, migraine, ADHD, depression, PNES (Psychogenic Non-Epileptic Seizures), anxiety, kidney disease)
EG0682 (father, depression, learning disability)
EG0750 (epilepsy, infantile autism, ADHD) COCO E43
EG0748 (sister, ADHD)
EG0751 (mother)
EG0772 (father)
EG0691 (epilepsy due to a recently discovered GRIN2A frameshift mutation) COCO E51
EG0690 (mother, epilepsy)
EG0692 (father)
EG0719 (epilepsy) COCO E52
EG0720 (brother, epilepsy, asperger syndrome)
EG0718 morther
EG0717 father
EG0850 (epilepsy) COCO E56
EG0852 (brother, epilepsy)
EG0851 (depression, anxiety)
EG0846 father
EG0725 (epilepsy, autism, intellectual disability due to a 15q11q14 duplication), COCO E64
EG0759 (brother, tourette syndrome, migraine, 15q11.2 duplication)
EG0726 mother

EG0724 (vocal tics obs tourette syndrome, 15q11.2 duplication)
EG1030 (epilepsy) COCO E82
EG1101 (sister, ADHD)
EG1031 mother,
EG1102 father (ADHD)
EG1036 (epilepsy, ADHD) COCO E83
EG1055 (brother, asperger syndrome)
EG1054 mother
EG1056 father
EG1059 (epilepsy, intellectual disability) COCO E85
EG1069 (sister, intellectual disability, autisme)
EG1058 mother
EG1068 father
EG1149 (epilepsy) COCO E88
EG1148 mother
EG1147 father
EG0632 (epilepsy, pulmonary stenosis, delayed bone growth, tourette syndrome, Obsessive compulsive disorder) COCO E89
EG0622 (sister, epilepsy)
EG0631 mother
EG630 father
EG1136 (epilepsy, speech delay, enamel dysplasia, sleep disorder)
EG1135 mother
EG1139 father
EG1118 (epilepsy, developmental delay)
EG1117 mother
EG1119 father
EG0909 (epilepsy, microcephaly)
EG0915 mother
EG0916 father
EG0413 (epilepsy, intellectual disability)

11000-10 mother
11002-10 father
EG1045 (epilepsy, intellectual disability) (maternal uncle with similar phenotype)
EG1044 mother
EG1043 father
EG0943 (epilepsy)
EG0944 (maternal halfsister, epilepsy)
EG0942 mother
EG0948 (Mathias father of EG0943)
EG1098 (epilepsy, speech delay, sleep apnea)
EG1110 (father)
EG1109 (mother)
EG0561 (epilepsy, intellectual disability, Lennox Gestault syndrome)
EG0560 (father)
EG0562 (mother)
EG0600 (epilepsy)
EG0611 (mother, epilepsy)
EG0601 (father, dyslexia)
EG1295 (Dravet syndrome due to a SCN1A mutation, extremely severely affected; suspicion of an additional genetic condition)
EG0235 also sent as 8945 father
EG0237 also sent as 8946 mother
EG0570, Aicardi syndrome
EG0531 sister,
EG0589 brother infantile Autism
EG0530 Lea mother
EG0588 Jesper father
EG0571 epilepsy, ADHD. COCO E8
EG1106 brother, obsessive compulsive disorder, anxiety, developmental delay
EG0572 (mother) ADHD, depression
EG0574 anxiety

8585 ADHD, inversion 12 COCO E10
8443 sister ADHD, inversion 12
8219 mother ADHD, epilepsy migraine, Obsessive compulsive disorder, PCDH19 mutation (suceptibility factor)
8586 father inversion 12

Sample Name	**Patient ID**
D003C	EG0152
D003F	7787
D003M	7786
D006C	EG0522
D006F	EG0523
D006M	EG0521
D008C	EG1238
D008F	EG0434
D008M	EG0435
D009C	EG0540
D009F	EG0539
D009M	EG0541
D010C	EG0609
D010F	EG0819
D010M	EG0680
D012C	EG0499
D012F	EG0498
D012M	EG1253
D015C	EG0662
D015F	EG0663
D015M	EG1010
D016C	EG0343
D016F	EG0330
D016M	EG0344

D018C	EG0681
D018F	EG0682
D018M	EG0683
D019C	EG0750
D019F	EG0772
D019M	EG0751
D020C	EG0691
D020F	EG0692
D020M	EG0690
D021C	EG0719
D021F	EG0717
D021M	EG0718
D022C	EG0850
D022F	EG0846
D022M	EG0851
D023C	EG0725
D023F	EG0724
D023M	EG0726
D026C	EG1030
D026F	EG1102
D026M	EG1031
D027C	EG1036
D027F	EG1056
D027M	EG1054
D029C	EG1059
D029F	EG1068
D029M	EG1058
D032C	EG1149
D032F	EG1147
D032M	EG1148
D033C	EG0632

D033F	EG0630
D033M	EG0631
D034C	EG1136
D034F	EG1139
D034M	EG1135
D035C	EG1118
D035F	EG1119
D035M	EG1117
D036C	EG0909
D036F	EG0916
D036M	EG0915
D037C	EG0413
D037F	11002
D037M	11000
D038C	EG1045
D038F	EG1043
D038M	EG1044
D039C	EG0943
D039F	EG0948
D039M	EG0942
D040C	EG1098
D040F	EG1110
D040M	EG1109
D041C	EG0561
D041F	EG0560
D041M	EG0562
D043C	EG0600
D043F	EG0601
D043M	EG0611
D044C	EG1295
D044F	EG0235

D044M	EG0237
D048C	EG0570
D048F	EG0588
D048M	EG0530
D052C	EG0571
D052F	EG0574
D052M	EG0572
D053C	8585
D053F	8586
D053M	8219

7.6 Isolated Congenital Anosmia patient phenotypes and pedigrees

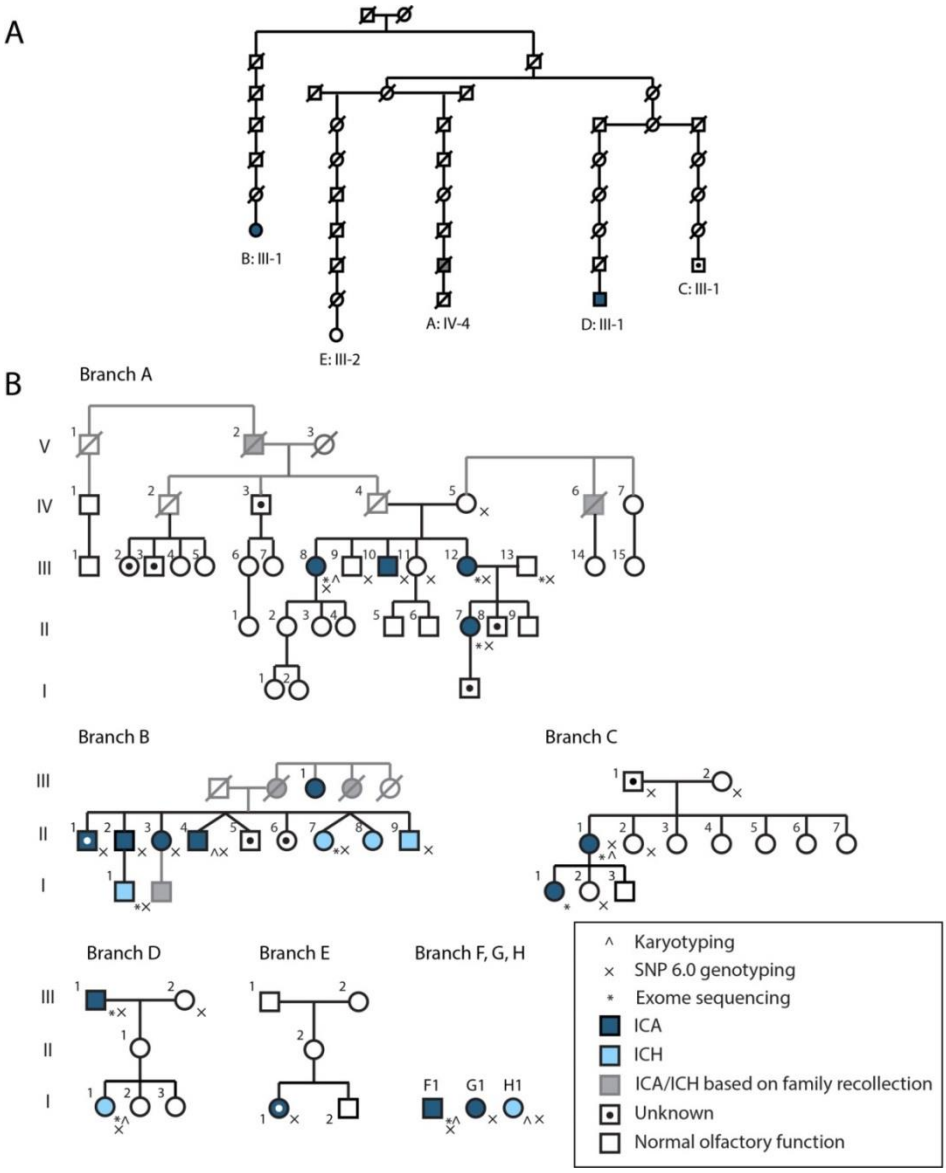


Figure 44. Faroese families

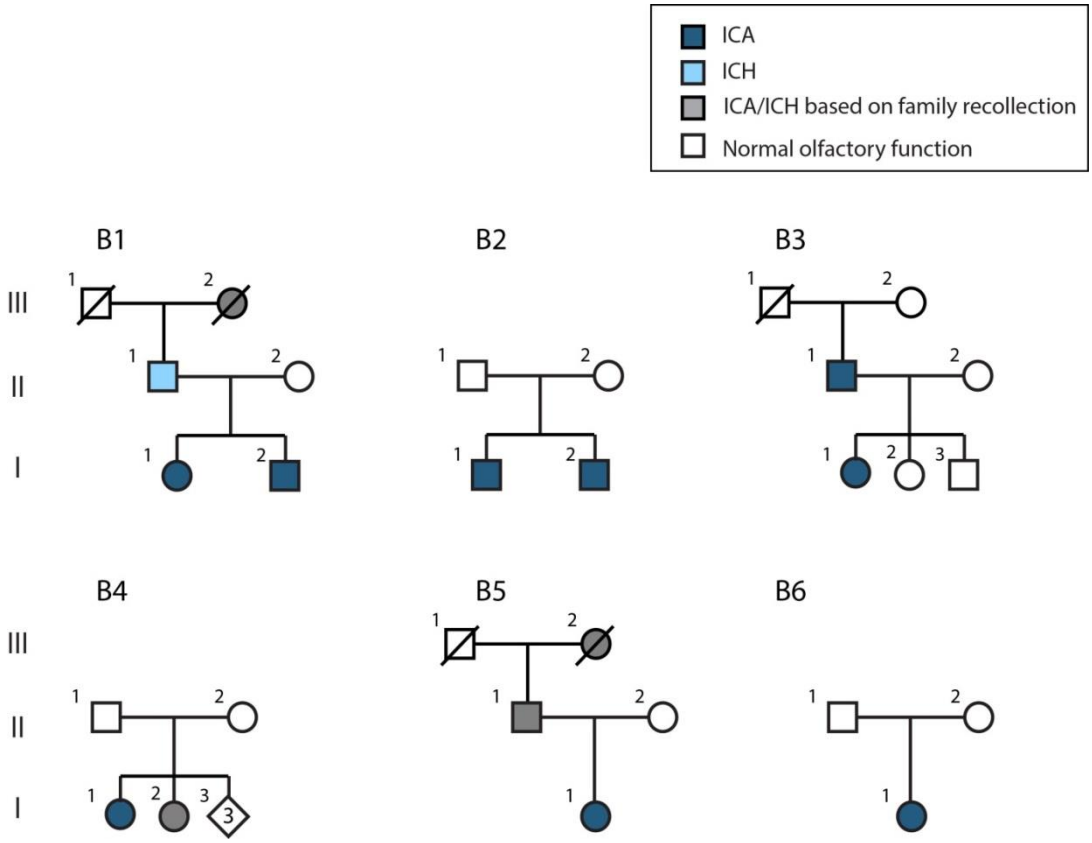


Figure 45. European families

Table 24. Anosmia sample information

ID	Affected status	Country	ID_Pedigree
AN001	ICA	Faroe Islands	BranchA_II-7
AN002	ICA	Faroe Islands	BranchA_III-12
AN003	Normal	Faroe Islands	BranchA_III-13
AN004	ICA	Faroe Islands	BranchB_III-1
AN005	ICH	Faroe Islands	BranchB_I-1
AN006	ICA	Faroe Islands	BranchD_III-1
AN007	ICA	Faroe Islands	F1
AN008	Normal	Faroe Islands	NA
AN009	Normal	Faroe Islands	NA
AN010	ICA	Germany	B2_I-1
AN011	ICA	Germany	B2_I-2
AN012	ICA	Germany	B6_I-1

AN013	Normal	Germany	B6_II-2
AN014	Normal	Germany	B6_II-1
AN015	ICA	Sweden	B3_I-1
AN016	ICA	Sweden	B3_II-1
AN017	Normal	Sweden	B3_II-2
AN018	ICA	Denmark	B1_I-2
AN019	Normal	Denmark	B1_II-2
AN020	ICH	Denmark	B1_II-1

7.7 Appendix Script 1

```
# set directory where the 1000genomes vcf files are located
dir="vcf.phase3"

# download the VCF header
~/software/htslib-1.3.2/tabix -H ${dir}/ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz  > header.txt

# Make a list of column numbers for PJL samples to remain in the final vcf.
# Assuming the order of individuals in the panel file is the same as in the vcf file.
# wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel
clmn=$(cat ${dir}/integrated_call_samples_v3.20130502.ALL.panel | awk 'BEGIN{s="1,2,3,4,5,6,7,8,9"} $2=="PJL" {s=s"," NR+8} END{print s}')
echo $clmn

# Number of samples to retain
nsamples=$(echo $clmn | awk -F"," '{print NF-9}') ; echo "Number of samples to retain: $nsamples"


# Cut "chr" from chromosome names in the bed file
cat CNE_human_goldensetfromGreg_hg19.bed | cut  --complement -c 1-3 > CNE_human_goldensetfromGreg_hg19.nochr.bed

# Write the vcf header to the file
head -n -1 header.txt > 1000g.CNE.PJL.vcf
tail -n 1 header.txt | cut -f $clmn >> 1000g.CNE.PJL.vcf

echo "Started to extract variants!"

# Extract the variants and write to the same file
cat CNE_human_goldensetfromGreg_hg19.nochr.bed |
    awk -v dr=$dir -v cl=$clmn '{cmd="~/software/htslib-1.3.2/tabix " dr "/ALL.chr"$1".phase3*vcf.gz " $1":"$2"-"$3 " | cut -f " cl " >> 1000g.CNE.PJL.vcf" ; system(cmd) }'

# Make reduced vcf, i.e. if all PJL genotypes are "0|0", remove the variant.
cat 1000g.CNE.PJL.vcf |
    awk -v s=$nsamples 'BEGIN {FS="\t"; OFS="\t"} {  for (i = 1; i <= s; ++i) { if ($(9+i)!="0|0") { print $0; break } }  }'
        > 1000g.CNE.PJL.reduced.vcf

echo "Finished!"
```

Reference List

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nature methods,* 7**,** 248-249.

AHN, S.-M., KIM, T.-H., LEE, S., KIM, D., GHANG, H., KIM, D.-S., KIM, B.-C., KIM, S.-Y., KIM, W.-Y. & KIM, C. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome research,* 19**,** 1622-1629.

AJAY, S. S., PARKER, S. C., ABAAN, H. O., FAJARDO, K. V. F. & MARGULIES, E. H. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome research,* 21**,** 1498-1505.

AKIYAMA, K., TAKEUCHI, F., ISONO, M., CHAKRAWARTHY, S., NGUYEN, Q. N., WEN, W., YAMAMOTO, K., KATSUYA, T., KASTURIRATNE, A., PHAM, S. T., ZHENG, W., MATSUSHITA, Y., KISHIMOTO, M., DO, L. D., SHU, X. O., WICKREMASINGHE, A. R., KAJIO, H. & KATO, N. 2014. Systematic fine-mapping of association with BMI and type 2 diabetes at the FTO locus by integrating results from multiple ethnic groups. *PLoS One,* 9**,** e101329.

ALBERIO, S., MINERI, R., TIRANTI, V. & ZEVIANI, M. 2007. Depletion of mtDNA: syndromes and genes. *Mitochondrion,* 7**,** 6-12.

ALEXANDER, R. P., FANG, G., ROZOWSKY, J., SNYDER, M. & GERSTEIN, M. B. 2010. Annotating non-coding regions of the genome. *Nature reviews. Genetics,* 11**,** 559.

ALMOND, S., KNAPP, M., FRANCOIS, C., TOUMI, M. & BRUGHA, T. 2004. Relapse in schizophrenia: costs, clinical outcomes and quality of life. *The British Journal of Psychiatry,* 184**,** 346-351.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

ALWINE, J. C., KEMP, D. J. & STARK, G. R. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5350-5354.

AMIT, M., DONYO, M., HOLLANDER, D., GOREN, A., KIM, E., GELFMAN, S., LEV-MAOR, G., BURSTEIN, D., SCHWARTZ, S., POSTOLSKY, B., PUPKO, T. & AST, G. 2012. Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Reports,* 1**,** 543-556.

ANTONELLIS, A., BENNETT, W. R., MENHENIOTT, T. R., PRASAD, A. B., LEE-LIN, S.-Q., GREEN, E. D., PAISLEY, D., KELSH, R. N., PAVAN, W. J. & WARD, A. 2006. Deletion of long-range sequences at Sox10 compromises developmental expression in a mouse model of Waardenburg–Shah (WS4) syndrome. *Human molecular genetics,* 15**,** 259-271.

APARICIO, S., CHAPMAN, J., STUPKA, E., PUTNAM, N., CHIA, J.-M., DEHAL, P., CHRISTOFFELS, A., RASH, S., HOON, S. & SMIT, A. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science,* 297**,** 1301-1310.

APARICIO, S., MORRISON, A., GOULD, A., GILTHORPE, J., CHAUDHURI, C., RIGBY, P., KRUMLAUF, R. & BRENNER, S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc Natl Acad Sci U S A,* 92**,** 1684-8.

ARDINGER, H. H. 1989. Association of genetic variation of the transforming growth factor-α gene with cleft lip and palate. *Am. J. Hum. Genet.,* 45**,** 348-353.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. & EPPIG, J. T. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics,* 25**,** 25.

ASLING, C., NELSON, M., DOUGHERTY, H., WEIGHT, H. & EVANS, H. 1960. The development of cleft palate resulting from maternal pteroylglutamic (folic) acid deficiency during the latter half of gestation in rats. *Surgery, Gynecology and Obstetrics,* 111**,** 19-28.

ASSOULINE, S., SHEVELL, M. I., ZATORRE, R. J., JONES-GOTMAN, M., SCHLOSS, M. D. & OUDJHANE, K. 1998. Children who can't smell the coffee: isolated congenital anosmia. *Journal of child neurology,* 13**,** 168-172.

ATTANASIO, C., NORD, A. S., ZHU, Y., BLOW, M. J., LI, Z., LIBERTON, D. K., MORRISON, H., PLAJZER-FRICK, I., HOLT, A., HOSSEINI, R., PHOUANENAVONG, S., AKIYAMA, J. A., SHOUKRY, M., AFZAL, V., RUBIN, E. M., FITZPATRICK, D. R., REN, B., HALLGRIMSSON, B., PENNACCHIO, L. A. & VISEL, A. 2013. Fine tuning of craniofacial morphology by distant-acting enhancers. *Science,* 342**,** 1241006.

BADIS, G., BERGER, M. F., PHILIPPAKIS, A. A., TALUKDER, S., GEHRKE, A. R., JAEGER, S. A., CHAN, E. T., METZLER, G., VEDENKO, A. & CHEN, X. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science,* 324**,** 1720-1723.

BAILEY, T. L., BODEN, M., BUSKE, F. A., FRITH, M., GRANT, C. E., CLEMENTI, L., REN, J., LI, W. W. & NOBLE, W. S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research,* 37**,** W202-W208.

BAMSHAD, M. J., NG, S. B., BIGHAM, A. W., TABOR, H. K., EMOND, M. J., NICKERSON, D. A. & SHENDURE, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics,* 12**,** 745.

BARR, C. & MISENER, V. 2016. Decoding the non‐coding genome: elucidating genetic risk outside the coding genome. *Genes, Brain and Behavior,* 15**,** 187-204.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.

BEATY, T. H. 2010. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genet.,* 42**,** 525-529.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L. & BIGNELL, H. R. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *nature,* 456**,** 53.

BERENTZEN, T., KRING, S. I., HOLST, C., ZIMMERMANN, E., JESS, T., HANSEN, T., PEDERSEN, O., TOUBRO, S., ASTRUP, A. & SORENSEN, T.

I. 2008. Lack of association of fatness-related FTO gene variants with energy expenditure or physical activity. *J Clin Endocrinol Metab,* 93**,** 2904-8.

BERG, J. S., KHOURY, M. J. & EVANS, J. P. 2011. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genetics in Medicine,* 13**,** 499-504.

BERNDT, S. I., GUSTAFSSON, S., MÄGI, R., GANNA, A., WHEELER, E., FEITOSA, M. F., JUSTICE, A. E., MONDA, K. L., CROTEAU-CHONKA, D. C. & DAY, F. R. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature genetics,* 45**,** 501-512.

BESSA, J., TENA, J. J., DE LA CALLE-MUSTIENES, E., FERNÁNDEZ-MIÑÁN, A., NARANJO, S., FERNÁNDEZ, A., MONTOLIU, L., AKALIN, A., LENHARD, B., CASARES, F. & GÓMEZ-SKARMETA, J. L. 2009. Zebrafish enhancer detection (ZED) vector: A new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Developmental Dynamics,* 238**,** 2409-2417.

BHATIA, S., GORDON, C. T., FOSTER, R. G., MELIN, L., ABADIE, V., BAUJAT, G., VAZQUEZ, M.-P., AMIEL, J., LYONNET, S. & VAN HEYNINGEN, V. 2015. Functional assessment of disease-associated regulatory variants in vivo using a versatile dual colour transgenesis strategy in zebrafish. *PLoS genetics,* 11**,** e1005193.

BHATIA, S. & KLEINJAN, D. A. 2014. Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Human genetics***,** 1-31.

BICKLER, S. & RODE, H. 2002. Surgical services for children in developing countries. *Bulletin of the World Health Organization,* 80**,** 829-835.

BIRNBAUM, S. 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genet.,* 41**,** 473-477.

BLAKELY, E. L., BUTTERWORTH, A., HADDEN, R. D., BODI, I., HE, L., MCFARLAND, R. & TAYLOR, R. W. 2012. MPV17 mutation causes neuropathy and leukoencephalopathy with multiple mtDNA deletions in muscle. *Neuromuscular Disorders,* 22**,** 587-591.

BLOW, M. J., MCCULLEY, D. J., LI, Z., ZHANG, T., AKIYAMA, J. A., HOLT, A., PLAJZER-FRICK, I., SHOUKRY, M., WRIGHT, C. & CHEN, F. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nature genetics,* 42**,** 806-810.

BOFFELLI, D., NOBREGA, M. A. & RUBIN, E. M. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet,* 5**,** 456-65.

BOWLEY, C. & KERR, M. 2000. Epilepsy and intellectual disability. *Journal of Intellectual Disability Research,* 44**,** 529-543.

BOYLE, A. P., HONG, E. L., HARIHARAN, M., CHENG, Y., SCHAUB, M. A., KASOWSKI, M., KARCZEWSKI, K. J., PARK, J., HITZ, B. C. & WENG, S. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research,* 22**,** 1790-1797.

BRAGG, L. M., STONE, G., BUTLER, M. K., HUGENHOLTZ, P. & TYSON, G. W. 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology,* 9**,** e1003031.

BRÄMERSON, A., JOHANSSON, L., EK, L., NORDIN, S. & BENDE, M. 2004. Prevalence of Olfactory Dysfunction: The Skövde Population‐Based Study. *The Laryngoscope,* 114**,** 733-737.

BRAY, N. J., PREECE, A., WILLIAMS, N. M., MOSKVINA, V., BUCKLAND, P. R., OWEN, M. J. & O'DONOVAN, M. C. 2005. Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. *Hum Mol Genet,* 14**,** 1947-54.

BRENNER, S., ELGAR, G., SANFORD, R., MACRAE, A., VENKATESH, B. & APARICIO, S. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature,* 366**,** 265-268.

BUONOCORE, F., HILL, M. J., CAMPBELL, C. D., OLADIMEJI, P. B., JEFFRIES, A. R., TROAKES, C., HORTOBAGYI, T., WILLIAMS, B. P., COOPER, J. D. & BRAY, N. J. 2010. Effects of cis-regulatory variation differ across regions of the adult human brain. *Hum Mol Genet,* 19**,** 4490-6.

BURNETTE, W. N. 1981. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Analytical biochemistry,* 112**,** 195-203.

CANNAVÒ, E., KHOUEIRY, P., GARFIELD, DAVID A., GEELEHER, P., ZICHNER, T., GUSTAFSON, E. H., CIGLAR, L., KORBEL, JAN O. & FURLONG, EILEEN E. M. 2016. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current Biology,* 26**,** 38-51.

CANVER, M. C., SMITH, E. C., SHER, F., PINELLO, L., SANJANA, N. E., SHALEM, O., CHEN, D. D., SCHUPP, P. G., VINJAMUR, D. S. & GARCIA, S. P. 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature,* 527**,** 192.

CARLSSON, A. & CARLSSON, M. L. 2006. A dopaminergic deficit hypothesis of schizophrenia: the path to discovery. *Dialogues in clinical neuroscience,* 8**,** 137.

CHALFIE, M., TU, Y., EUSKIRCHEN, G., WARD, W. W. & PRASHER, D. C. 1994. Green fluorescent protein as a marker for gene expression. *Science***,** 802-805.

CHANG, X. & WANG, K. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. *Journal of medical genetics***,** jmedgenet-2012-100918.

CHEN, K. & RAJEWSKY, N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature reviews. Genetics,* 8**,** 93.

CHOWDARI, K. V., MIRNICS, K., SEMWAL, P., WOOD, J., LAWRENCE, E., BHATIA, T., DESHPANDE, S. N., B, K. T., FERRELL, R. E., MIDDLETON, F. A., DEVLIN, B., LEVITT, P., LEWIS, D. A. & NIMGAONKAR, V. L. 2002. Association and linkage analyses of RGS4 polymorphisms in schizophrenia. *Hum Mol Genet,* 11**,** 1373-80.

CHURCH, C., LEE, S., BAGG, E. A., MCTAGGART, J. S., DEACON, R., GERKEN, T., LEE, A., MOIR, L., MECINOVIC, J., QUWAILID, M. M., SCHOFIELD, C. J., ASHCROFT, F. M. & COX, R. D. 2009. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet,* 5**,** e1000599.

CIOFALO, A., FILIACI, F., ROMEO, R., ZAMBETTI, G. & VESTRI, A. R. 2006. Epidemiological aspects of olfactory dysfunction. *Rhinology,* 44**,** 78-82.

CLARKE, G. M., ANDERSON, C. A., PETTERSSON, F. H., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2011. Basic statistical analysis in genetic case-control studies. *Nature protocols,* 6**,** 121-133.

CLAUSSNITZER, M., DANKEL, S. N., KIM, K.-H., QUON, G., MEULEMAN, W., HAUGEN, C., GLUNK, V., SOUSA, I. S., BEAUDRY, J. L. & PUVIINDRAN, V. 2015. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine,* 373**,** 895-907.

COLLIN, G., DE REUS, M. A., CAHN, W., POL, H. E. H., KAHN, R. S. & VAN DEN HEUVEL, M. P. 2013. Disturbed grey matter coupling in schizophrenia. *European Neuropsychopharmacology,* 23**,** 46-54.

COLLINS, F. S., GREEN, E. D., GUTTMACHER, A. E. & GUYER, M. S. 2003. A vision for the future of genomics research. *Nature,* 422**,** 835-847.

COPE, N. F. & FRASER, P. 2009. Chromosome conformation capture. *Cold Spring Harbor Protocols,* 2009**,** pdb. prot5137.

CRICK, F. H., BARNETT, L., BRENNER, S. & WATTS-TOBIN, R. J. 1961. General nature of the genetic code for proteins. *Nature,* 192**,** 1227-1232.

DAI, J. Y., RUCZINSKI, I., LEBLANC, M. & KOOPERBERG, C. 2006. Imputation methods to improve inference in SNP association studies. *Genetic epidemiology,* 30**,** 690-702.

DALLA ROSA, I., CAMARA, Y., DURIGON, R., MOSS, C. F., VIDONI, S., AKMAN, G., HUNT, L., JOHNSON, M. A., GROCOTT, S., WANG, L., THORBURN, D. R., HIRANO, M., POULTON, J., TAYLOR, R. W., ELGAR, G., MARTI, R., VOSHOL, P., HOLT, I. J. & SPINAZZOLA, A. 2016. MPV17 Loss Causes Deoxynucleotide Insufficiency and Slow DNA Replication in Mitochondria. *PLoS Genet,* 12**,** e1005779.

DALLABONA, C., MARSANO, R. M., ARZUFFI, P., GHEZZI, D., MANCINI, P., ZEVIANI, M., FERRERO, I. & DONNINI, C. 2009. Sym1, the yeast ortholog of the MPV17 human disease protein, is a stress-induced bioenergetic and morphogenetic mitochondrial modulator. *Human molecular genetics,* 19**,** 1098-1107.

DAVYDOV, E. V., GOODE, D. L., SIROTA, M., COOPER, G. M., SIDOW, A. & BATZOGLOU, S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology,* 6**,** e1001025.

DE SILVA, D. R., NICHOLS, R. & ELGAR, G. 2014. Purifying selection in deeply conserved human enhancers is more consistent than in coding sequences. *PloS one,* 9**,** e103357.

DECIPHERING DEVELOPMENTAL DISORDERS, S. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature,* 542**,** 433-438.

DEKKER, J. 2006. The three'C's of chromosome conformation capture: controls, controls, controls. *Nature methods,* 3**,** 17.

DINA, C., MEYRE, D., GALLINA, S., DURAND, E., KORNER, A., JACOBSON, P., CARLSSON, L. M., KIESS, W., VATIN, V., LECOEUR, C., DELPLANQUE, J., VAILLANT, E., PATTOU, F., RUIZ, J., WEILL, J., LEVY-MARCHAL, C., HORBER, F., POTOCZNA, N., HERCBERG, S., LE STUNFF, C., BOUGNERES, P., KOVACS, P., MARRE, M., BALKAU, B., CAUCHI, S.,

CHEVRE, J. C. & FROGUEL, P. 2007. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet,* 39**,** 724-6.

DIXON, J., SELVARAJ, S., YUE, F., KIM, A., LI, Y. & SHEN, Y. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature,* 485**,** 376 - 80.

DIXON, J. R., JUNG, I., SELVARAJ, S., SHEN, Y., ANTOSIEWICZ-BOURGET, J. E., LEE, A. Y., YE, Z., KIM, A., RAJAGOPAL, N. & XIE, W. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature,* 518**,** 331.

DIXON, M. J., MARAZITA, M. L., BEATY, T. H. & MURRAY, J. C. 2011. Cleft lip and palate: synthesizing genetic and environmental influences. *Nature reviews. Genetics,* 12**,** 167-178.

DOGLIO, L., GOODE, D. K., PELLERI, M. C., PAULS, S., FRABETTI, F., SHIMELD, S. M., VAVOURI, T. & ELGAR, G. 2013. Parallel evolution of chordate cis-regulatory code for development. *PLoS genetics,* 9**,** e1003904.

DOLINOY, D. C., WEIDMAN, J. R. & JIRTLE, R. L. 2007. Epigenetic gene regulation: Linking early developmental environment to adult disease. *Reproductive Toxicology,* 23**,** 297-307.

DRAKE, J. A., BIRD, C., NEMESH, J., THOMAS, D. J., NEWTON-CHEH, C., REYMOND, A., EXCOFFIER, L., ATTAR, H., ANTONARAKIS, S. E. & DERMITZAKIS, E. T. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature genetics,* 38**,** 223.

DUBOULE, D. 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development,* 1994**,** 135-142.

EASTWOOD, S. & HARRISON, P. 2005. Interstitial white matter neuron density in the dorsolateral prefrontal cortex and parahippocampal gyrus in schizophrenia. *Schizophrenia research,* 79**,** 181-188.

ELGAR, G. & VAVOURI, T. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet,* 24**,** 344-52.

ELLROTT, K., YANG, C., SLADEK, F. M. & JIANG, T. 2002. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics,* 18**,** S100-S109.

EMAMIAN, E. S., HALL, D., BIRNBAUM, M. J., KARAYIORGOU, M. & GOGOS, J. A. 2004. Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia. *Nat Genet,* 36**,** 131-7.

ENCODE PROJECT CONSORTIUM. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature,* 447**,** 799.

FANG, H., LI, Y., DU, S., HU, X., ZHANG, Q., LIU, A. & MA, G. 2010. Variant rs9939609 in the FTO gene is associated with body mass index among Chinese children. *BMC Med Genet,* 11**,** 136.

FEATHERSTONE, M. 2003. HOX proteins and their co-factors in transcriptional regulation. *Advances in Developmental Biology and Biochemistry,* 13**,** 1-42.

FELDMESSER, E., BERCOVICH, D., AVIDAN, N., HALBERTAL, S., HAIM, L., GROSS-ISSEROFF, R., GOSHEN, S. & LANCET, D. 2006. Mutations in olfactory signal transduction genes are not a major cause of human congenital general anosmia. *Chemical senses,* 32**,** 21-30.

FISCHER, J., KOCH, L., EMMERLING, C., VIERKOTTEN, J., PETERS, T., BRUNING, J. C. & RUTHER, U. 2009. Inactivation of the Fto gene protects from obesity. *Nature,* 458**,** 894-8.

FISHER, S., GRICE, E. A., VINTON, R. M., BESSLING, S. L., URASAKI, A., KAWAKAMI, K. & MCCALLION, A. S. 2006. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nature protocols,* 1**,** 1297.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BILLIS, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G. & FITZGERALD, S. 2013. Ensembl 2014. *Nucleic acids research***,** gkt1196.

FLINT, J. 2001. Genetic basis of cognitive disability. *Dialogues in Clinical Neuroscience,* 3**,** 37-46.

FORSGREN, L., BEGHI, E., OUN, A. & SILLANPAA, M. 2005. The epidemiology of epilepsy in Europe - a systematic review. *Eur J Neurol,* 12**,** 245-53.

FOX, E. J., REID-BAYLISS, K. S., EMOND, M. J. & LOEB, L. A. 2014. Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications,* 1**,** 1000106.

FRASER, F. 1970. The genetics of cleft lip and cleft palate. *American journal of human genetics,* 22**,** 336.

FRAYLING, T. M., TIMPSON, N. J., WEEDON, M. N., ZEGGINI, E., FREATHY, R. M., LINDGREN, C. M., PERRY, J. R., ELLIOTT, K. S., LANGO, H., RAYNER, N. W., SHIELDS, B., HARRIES, L. W., BARRETT, J. C., ELLARD, S., GROVES, C. J., KNIGHT, B., PATCH, A. M., NESS, A. R., EBRAHIM, S., LAWLOR, D. A., RING, S. M., BEN-SHLOMO, Y., JARVELIN, M. R., SOVIO, U., BENNETT, A. J., MELZER, D., FERRUCCI, L., LOOS, R. J., BARROSO, I., WAREHAM, N. J., KARPE, F., OWEN, K. R., CARDON, L. R., WALKER, M., HITMAN, G. A., PALMER, C. N., DONEY, A. S., MORRIS, A. D., SMITH, G. D., HATTERSLEY, A. T. & MCCARTHY, M. I. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* 316**,** 889-94.

FREEMAN, J. L., PERRY, G. H., FEUK, L., REDON, R., MCCARROLL, S. A., ALTSHULER, D. M., ABURATANI, H., JONES, K. W., TYLER-SMITH, C. & HURLES, M. E. 2006. Copy number variation: new insights in genome diversity. *Genome research,* 16**,** 949-961.

GABORIT, N., SAKUMA, R., WYLIE, J. N., KIM, K. H., ZHANG, S. S., HUI, C. C. & BRUNEAU, B. G. 2012. Cooperative and antagonistic roles for Irx3 and Irx5 in cardiac morphogenesis and postnatal physiology. *Development,* 139**,** 4007-19.

GABRIEL, S. B., SCHAFFNER, S. F., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGINS, J., DEFELICE, M., LOCHNER, A., FAGGART, M., LIU-CORDERO, S. N., ROTIMI, C., ADEYEMO, A., COOPER, R., WARD, R., LANDER, E. S., DALY, M. J. & ALTSHULER, D. 2002. The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-9.

GAWAD, C., KOH, W. & QUAKE, S. R. 2016. Single-cell genome sequencing: current state of the science. *Nature reviews. Genetics,* 17**,** 175.

GENE ONTOLOGY CONSORTIUM. 2015. Gene ontology consortium: going forward. Nucleic acids research, 43, D1049-D1056.

GENOMES PROJECT CONSORTIUM. 2010. A map of human genome variation from population scale sequencing. Nature, 467, 1061.

GENOMES PROJECT CONSORTIUM. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature, 491, 56 - 65.

GHADAMI, M., MAJIDZADEH ‐ A, K., MOROVVATI, S., DAMAVANDI, E., NISHIMURA, G., KOMATSU, K., KINOSHITA, A., NAJAFI, M. T., NIIKAWA, N. & YOSHIURA, K. I. 2004a. Isolated congenital anosmia with morphologically normal olfactory bulb in two Iranian families: a new clinical entity? *American Journal of Medical Genetics Part A,* 127**,** 307-309.

GHADAMI, M., MOROVVATI, S., MAJIDZADEH-A, K., DAMAVANDI, E., NISHIMURA, G., KINOSHITA, A., PASALAR, P., KOMATSU, K., NAJAFI, M. & NIIKAWA, N. 2004b. Isolated congenital anosmia locus maps to 18p11. 23-q12. 2. *Journal of medical genetics,* 41**,** 299-303.

GILLES, A., MEGLÉCZ, E., PECH, N., FERREIRA, S., MALAUSA, T. & MARTIN, J.-F. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics,* 12**,** 245.

GOGTAY, N., VYAS, N. S., TESTA, R., WOOD, S. J. & PANTELIS, C. 2011. Age of onset of schizophrenia: perspectives from structural neuroimaging studies. *Schizophrenia Bulletin,* 37**,** 504-513.

GONZALEZ-VIOQUE, E., TORRES-TORRONTERAS, J., ANDREU, A. L. & MARTÍ, R. 2011. Limited dCTP availability accounts for mitochondrial DNA depletion in mitochondrial neurogastrointestinal encephalomyopathy (MNGIE). *PLoS genetics,* 7**,** e1002035.

GOODE, D. K., SNELL, P., SMITH, S. F., COOKE, J. E. & ELGAR, G. 2005. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics,* 86**,** 172-81.

GRAFF, M., GORDON-LARSEN, P., LIM, U., FOWKE, J. H., LOVE, S.-A., FESINMEYER, M., WILKENS, L. R., VERTILUS, S., RITCHIE, M. D. & PRENTICE, R. L. 2013. The Influence of Obesity-Related Single Nucleotide Polymorphisms on BMI Across the Life Course The PAGE Study. *Diabetes,* 62**,** 1763-1767.

GRANT, S. F. 2009. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J. Pediatr.,* 155**,** 909-913.

GRICE, J., NOYVERT, B., DOGLIO, L. & ELGAR, G. 2015. A simple predictive enhancer syntax for hindbrain patterning is conserved in vertebrate genomes. *PloS one,* 10**,** e0130413.

GROSEN, D., BILLE, C., PETERSEN, I., SKYTTHE, A., VON BORNEMANN HJELMBORG, J., PEDERSEN, J. K., MURRAY, J. C. & CHRISTENSEN, K. 2011. Risk of Oral Clefts in twins. *Epidemiology (Cambridge, Mass.),* 22**,** 313-319.

GROUP, I. W. 2011. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC).

GRUNNET, L. G., NILSSON, E., LING, C., HANSEN, T., PEDERSEN, O., GROOP, L., VAAG, A. & POULSEN, P. 2009. Regulation and function of FTO mRNA expression in human skeletal muscle and subcutaneous adipose tissue. *Diabetes,* 58**,** 2402-8.

GUERRINI, R. & DOBYNS, W. B. 2014. Malformations of cortical development: clinical features and genetic causes. *The Lancet Neurology,* 13**,** 710-726.

GUO, C., LUDVIK, A. E., ARLOTTO, M. E., HAYES, M. G., ARMSTRONG, L. L., SCHOLTENS, D. M., BROWN, C. D., NEWGARD, C. B., BECKER, T. C. & LAYDEN, B. T. 2015. Coordinated Regulatory Variation Associated with Gestational Hyperglycemia Regulates Expression of the Novel Hexokinase HKDC1. *Nature communications,* 6**,** 6069.

HAIJMA, S. V., VAN HAREN, N., CAHN, W., KOOLSCHIJN, P. C. M., HULSHOFF POL, H. E. & KAHN, R. S. 2012. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia bulletin,* 39**,** 1129-1138.

HAKANEN, M., RAITAKARI, O. T., LEHTIMAKI, T., PELTONEN, N., PAHKALA, K., SILLANMAKI, L., LAGSTROM, H., VIIKARI, J., SIMELL, O. & RONNEMAA, T. 2009. FTO genotype is associated with body mass index after the age of seven years but not with energy intake or leisure-time physical activity. *J Clin Endocrinol Metab,* 94**,** 1281-7.

HAN, Y., SLIVANO, O. J., CHRISTIE, C. K., CHENG, A. W. & MIANO, J. M. 2015. CRISPR-Cas9 Genome Editing of a Single Regulatory Element Nearly Abolishes Target Gene Expression in Mice. *Arteriosclerosis, thrombosis, and vascular biology,* 35**,** 312-315.

HARDY, R., WILLS, A. K., WONG, A., ELKS, C. E., WAREHAM, N. J., LOOS, R. J., KUH, D. & ONG, K. K. 2010. Life course variations in the associations between FTO and MC4R gene variants and body size. *Human molecular genetics,* 19**,** 545-552.

HARISMENDY, O., NG, P. C., STRAUSBERG, R. L., WANG, X., STOCKWELL, T. B., BEESON, K. Y., SCHORK, N. J., MURRAY, S. S., TOPOL, E. J. & LEVY, S. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology,* 10**,** R32.

HATEM, A., BOZDAĞ, D., TOLAND, A. E. & ÇATALYÜREK, Ü. V. 2013. Benchmarking short sequence mapping tools. *BMC bioinformatics,* 14**,** 184.

HENNIG, B. J., FULFORD, A. J., SIRUGO, G., RAYCO-SOLON, P., HATTERSLEY, A. T., FRAYLING, T. M. & PRENTICE, A. M. 2009. FTO gene variation and measures of body mass in an African population. *BMC Med Genet,* 10**,** 21.

HERR, W. & CLEARY, M. A. 1995. The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes & development,* 9**,** 1679-1693.

HINNEY, A., NGUYEN, T. T., SCHERAG, A., FRIEDEL, S., BRONNER, G., MULLER, T. D., GRALLERT, H., ILLIG, T., WICHMANN, H. E., RIEF, W., SCHAFER, H. & HEBEBRAND, J. 2007. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS One,* 2**,** e1361.

HOGAN, B. 1996. Bone morphogenetic proteins: multifunctional regulators of vertebrate development. *Genes & development,* 10**,** 1580-1594.

HONEA, R., CROW, T. J., PASSINGHAM, D. & MACKAY, C. E. 2005. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *American Journal of Psychiatry,* 162**,** 2233-2245.

HONG, J., WINDMEIJER, F., NOVICK, D., HARO, J. M. & BROWN, J. 2009. The cost of relapse in patients with schizophrenia in the European SOHO

(Schizophrenia Outpatient Health Outcomes) study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* 33**,** 835-841.

HOTTA, K., NAKATA, Y., MATSUO, T., KAMOHARA, S., KOTANI, K., KOMATSU, R., ITOH, N., MINEO, I., WADA, J., MASUZAKI, H., YONEDA, M., NAKAJIMA, A., MIYAZAKI, S., TOKUNAGA, K., KAWAMOTO, M., FUNAHASHI, T., HAMAGUCHI, K., YAMADA, K., HANAFUSA, T., OIKAWA, S., YOSHIMATSU, H., NAKAO, K., SAKATA, T., MATSUZAWA, Y., TANAKA, K., KAMATANI, N. & NAKAMURA, Y. 2008. Variations in the FTO gene are associated with severe obesity in the Japanese. *J Hum Genet,* 53**,** 546-53.

HOUWELING, A. C., DILDROP, R., PETERS, T., MUMMENHOFF, J., MOORMAN, A. F., RUTHER, U. & CHRISTOFFELS, V. M. 2001. Gene and cluster-specific expression of the Iroquois family members during mouse development. *Mech Dev,* 107**,** 169-74.

HOWARD, M. L. & DAVIDSON, E. H. 2004. < i> cis</i>-Regulatory control circuits in development. *Developmental biology,* 271**,** 109-118.

HUANG, T., 2011. Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Current protocols in human genetics, pp.19-8.

HUME, M. A., BARRERA, L. A., GISSELBRECHT, S. S. & BULYK, M. L. 2014. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic acids research,* 43**,** D117-D122.

HUSE, S. M., HUBER, J. A., MORRISON, H. G., SOGIN, M. L. & WELCH, D. M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology,* 8**,** R143.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

INTERNATIONAL HAPMAP, CONSORTIUM., FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., GIBBS, R. A., BELMONT, J. W., BOUDREAU, A., HARDENBOL, P., LEAL, S. M., PASTERNAK, S., WHEELER, D. A., WILLIS, T. D., YU, F., YANG, H., ZENG, C., GAO, Y., HU, H., HU, W., LI, C., LIN, W., LIU, S., PAN, H., TANG, X., WANG, J., WANG, W., YU, J., ZHANG, B., ZHANG, Q., ZHAO, H., ZHAO, H., ZHOU, J., GABRIEL, S. B., BARRY, R., BLUMENSTIEL, B., CAMARGO, A., DEFELICE, M., FAGGART, M., GOYETTE, M., GUPTA, S., MOORE, J., NGUYEN, H., ONOFRIO, R. C., PARKIN, M., ROY, J., STAHL, E., WINCHESTER, E., ZIAUGRA, L., ALTSHULER, D., SHEN, Y., YAO, Z., HUANG, W., CHU, X., HE, Y., JIN, L., LIU, Y., SHEN, Y., SUN, W., WANG, H., WANG, Y., WANG, Y., XIONG, X., XU, L., WAYE, M. M., TSUI, S. K., XUE, H., WONG, J. T., GALVER, L. M., FAN, J. B., GUNDERSON, K., MURRAY, S. S., OLIPHANT, A. R., CHEE, M. S., MONTPETIT, A., CHAGNON, F., FERRETTI, V., LEBOEUF, M., OLIVIER, J. F., PHILLIPS, M. S., ROUMY, S., SALLEE, C., VERNER, A., HUDSON, T. J., KWOK, P. Y., CAI, D., KOBOLDT, D. C., MILLER, R. D., PAWLIKOWSKA, L., TAILLON-MILLER, P., XIAO, M., TSUI, L. C., MAK, W., SONG, Y. Q., TAM, P. K., NAKAMURA, Y., KAWAGUCHI, T., KITAMOTO, T.,

MORIZONO, T., NAGASHIMA, A., et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-61.

INTERNATIONAL SCHIZOPHRENIA CONSORTIUM. 2009. Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. *Nature,* 460**,** 748.

IONITA-LAZA, I., MCCALLUM, K., XU, B. & BUXBAUM, J. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics,* 48**,** 214.

JESS, T., ZIMMERMANN, E., KRING, S. I., BERENTZEN, T., HOLST, C., TOUBRO, S., ASTRUP, A., HANSEN, T., PEDERSEN, O. & SORENSEN, T. I. 2008. Impact on weight dynamics and general growth of the common FTO rs9939609: a longitudinal Danish cohort study. *Int J Obes (Lond),* 32**,** 1388-94.

JIA, G. F., FU, Y., ZHAO, X., DAI, Q., ZHENG, G. Q., YANG, Y., YI, C. Q., LINDAHL, T., PAN, T., YANG, Y. G. & HE, C. 2011. N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology,* 7**,** 885-887.

JIN, F., LI, Y., DIXON, J. R., SELVARAJ, S., YE, Z., LEE, A. Y., YEN, C.-A., SCHMITT, A. D., ESPINOZA, C. & REN, B. 2013. A high-resolution map of three-dimensional chromatin interactome in human cells. *Nature,* 503**,** 290.

JOHNSON, A. A. & JOHNSON, K. A. 2001. Exonuclease proofreading by human mitochondrial DNA polymerase. *Journal of Biological Chemistry,* 276**,** 38097-38107.

JOHNSON, C. Y. & LITTLE, J. 2008. Folate intake, markers of folate status and oral clefts: is the evidence converging? *International journal of epidemiology,* 37**,** 1041-1058.

JOLLY, L. A., HOMAN, C. C., JACOB, R., BARRY, S. & GECZ, J. 2013. The UPF3B gene, implicated in intellectual disability, autism, ADHD and childhood onset schizophrenia regulates neural progenitor cell behaviour and neuronal outgrowth. *Hum Mol Genet,* 22**,** 4673-87.

JONES, F. C., GRABHERR, M. G., CHAN, Y. F., RUSSELL, P., MAUCELI, E., JOHNSON, J., SWOFFORD, R., PIRUN, M., ZODY, M. C. & WHITE, S. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature,* 484**,** 55.

KANG, S. L., NARAYANAN, C. S. & KELSALL, W. 2012. Mortality among infants born with orofacial clefts in a single cleft network. *Cleft Palate Craniofac J,* 49**,** 508-11.

KATZMAN, S., KERN, A. D., BEJERANO, G., FEWELL, G., FULTON, L., WILSON, R. K., SALAMA, S. R. & HAUSSLER, D. 2007. Human genome ultraconserved elements are ultraselected. *Science,* 317**,** 915-915.

KAWAKAMI, K. 2007. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biology,* 8**,** S7-S7.

KAY, S. R., FISZBEIN, A. & OPLER, L. A. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull,* 13**,** 261-76.

KENT, W. J. 2002. BLAT—the BLAST-like alignment tool. *Genome research,* 12**,** 656-664.

KIM, S. Y., LI, Y., GUO, Y., LI, R., HOLMKVIST, J., HANSEN, T., PEDERSEN, O., WANG, J. & NIELSEN, R. 2010. Design of association studies with pooled or

un‐pooled next‐generation sequencing data. *Genetic epidemiology,* 34**,** 479-491.

KINGSMORE, S. F. & SAUNDERS, C. J. 2011. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Science translational medicine,* 3**,** 87ps23-87ps23.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics,* 46**,** 310-315.

KLÖTING, N., SCHLEINITZ, D., RUSCHKE, K., BERNDT, J., FASSHAUER, M., TÖNJES, A., SCHÖN, M., KOVACS, P., STUMVOLL, M. & BLÜHER, M. 2008. Inverse relationship between obesity and FTO gene expression in visceral adipose tissue in humans. *Diabetologia,* 51**,** 641-647.

KNAPP, M., MANGALORE, R. & SIMON, J. 2004. The global costs of schizophrenia. *Schizophrenia bulletin,* 30**,** 279-293.

KOZAREWA, I., NING, Z., QUAIL, M. A., SANDERS, M. J., BERRIMAN, M. & TURNER, D. J. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. *Nature methods,* 6**,** 291-295.

KRING, S. I., HOLST, C., ZIMMERMANN, E., JESS, T., BERENTZEN, T., TOUBRO, S., HANSEN, T., ASTRUP, A., PEDERSEN, O. & SORENSEN, T. I. 2008. FTO gene associated fatness in relation to body fat distribution and metabolic traits throughout a broad range of fatness. *PLoS One,* 3**,** e2958.

KUMAR, P., HENIKOFF, S. & NG, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols,* 4**,** 1073-1081.

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M.J. AND AMIN, V., 2015. Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), p.317.

KWAN, K. M., FUJIMOTO, E., GRABHER, C., MANGUM, B. D., HARDY, M. E., CAMPBELL, D. S., PARANT, J. M., YOST, H. J., KANKI, J. P. & CHIEN, C. B. 2007. The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. *Dev Dyn,* 236**,** 3088-99.

LAHTI, A. C., HOLCOMB, H. H., WEILER, M. A., MEDOFF, D. R. & TAMMINGA, C. A. 2003. Functional effects of antipsychotic drugs: comparing clozapine with haloperidol. *Biological psychiatry,* 53**,** 601-608.

LAMMER, E. J., SHAW, G. M., IOVANNISCI, D. M., VAN WAES, J. & FINNELL, R. H. 2004. Maternal smoking and the risk of orofacial clefts: susceptibility with NAT1 and NAT2 polymorphisms. *Epidemiology,* 15**,** 150-156.

LANDIS, B. N., KONNERTH, C. G. & HUMMEL, T. 2004. A study on the frequency of olfactory dysfunction. *The Laryngoscope,* 114**,** 1764-1769.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods,* 9**,** 357-359.

LANGO ALLEN, H., ESTRADA, K., LETTRE, G., BERNDT, S. I., WEEDON, M. N., RIVADENEIRA, F., WILLER, C. J., JACKSON, A. U., VEDANTAM, S., RAYCHAUDHURI, S., FERREIRA, T., WOOD, A. R., WEYANT, R. J., SEGRE, A. V., SPELIOTES, E. K., WHEELER, E., SORANZO, N., PARK, J. H., YANG, J., GUDBJARTSSON, D., HEARD-COSTA, N. L., RANDALL, J.

C., QI, L., VERNON SMITH, A., MAGI, R., PASTINEN, T., LIANG, L., HEID, I. M., LUAN, J., THORLEIFSSON, G., WINKLER, T. W., GODDARD, M. E., SIN LO, K., PALMER, C., WORKALEMAHU, T., AULCHENKO, Y. S., JOHANSSON, A., ZILLIKENS, M. C., FEITOSA, M. F., ESKO, T., JOHNSON, T., KETKAR, S., KRAFT, P., MANGINO, M., PROKOPENKO, I., ABSHER, D., ALBRECHT, E., ERNST, F., GLAZER, N. L., HAYWARD, C., HOTTENGA, J. J., JACOBS, K. B., KNOWLES, J. W., KUTALIK, Z., MONDA, K. L., POLASEK, O., PREUSS, M., RAYNER, N. W., ROBERTSON, N. R., STEINTHORSDOTTIR, V., TYRER, J. P., VOIGHT, B. F., WIKLUND, F., XU, J., ZHAO, J. H., NYHOLT, D. R., PELLIKKA, N., PEROLA, M., PERRY, J. R., SURAKKA, I., TAMMESOO, M. L., ALTMAIER, E. L., AMIN, N., ASPELUND, T., BHANGALE, T., BOUCHER, G., CHASMAN, D. I., CHEN, C., COIN, L., COOPER, M. N., DIXON, A. L., GIBSON, Q., GRUNDBERG, E., HAO, K., JUHANI JUNTTILA, M., KAPLAN, L. M., KETTUNEN, J., KONIG, I. R., KWAN, T., LAWRENCE, R. W., LEVINSON, D. F., LORENTZON, M., MCKNIGHT, B., MORRIS, A. P., MULLER, M., SUH NGWA, J., PURCELL, S., RAFELT, S., SALEM, R. M., SALVI, E., et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-8.

LAPPALAINEN, T. & DERMITZAKIS, E. T. 2010. Evolutionary history of regulatory variation in human populations. *Human molecular genetics,* 19**,** R197-R203.

LAPPALAINEN, T., MONTGOMERY, S. B., NICA, A. C. & DERMITZAKIS, E. T. 2011. Epistatic selection between coding and regulatory variation in human evolution and disease. *The American Journal of Human Genetics,* 89**,** 459-463.

LATCHMAN, D. S. 1997. Transcription factors: an overview. *The international journal of biochemistry & cell biology,* 29**,** 1305-1312.

LEE, A. P., KERK, S. Y., TAN, Y. Y., BRENNER, S. & VENKATESH, B. 2010. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Molecular biology and evolution,* 28**,** 1205-1215.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H.-H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION, C. 2016.

Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-291.

LEOPOLD, D. A., HORNUNG, D. E. & SCHWOB, J. E. 1992. Congenital lack of olfactory ability. *Annals of Otology, Rhinology & Laryngology,* 101**,** 229-236.

LESLIE, E. J., TAUB, M. A., LIU, H., STEINBERG, K. M., KOBOLDT, D. C., ZHANG, Q., CARLSON, J. C., HETMANSKI, J. B., WANG, H. & LARSON, D. E. 2015. Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *The American Journal of Human Genetics,* 96**,** 397-411.

LETTICE, L. A., HEANEY, S. J., PURDIE, L. A., LI, L., DE BEER, P., OOSTRA, B. A., GOODE, D., ELGAR, G., HILL, R. E. & DE GRAAFF, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet,* 12**,** 1725-35.

LEUCHT, S., CIPRIANI, A., SPINELI, L., MAVRIDIS, D., ÖREY, D., RICHTER, F., SAMARA, M., BARBUI, C., ENGEL, R. R. & GEDDES, J. R. 2013. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *The Lancet,* 382**,** 951-962.

LEUCHT, S., KANE, J. M., KISSLING, W., HAMANN, J., ETSCHEL, E. & ENGEL, R. R. 2005. What does the PANSS mean? *Schizophr Res,* 79**,** 231-8.

LI, G., RUAN, X., AUERBACH, R. K., SANDHU, K. S., ZHENG, M., WANG, P., POH, H. M., GOH, Y., LIM, J. & ZHANG, J. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell,* 148**,** 84-98.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LI, H., RUAN, J. & DURBIN, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research,* 18**,** 1851-1858.

LI, J. B., GAO, Y., AACH, J., ZHANG, K., KRYUKOV, G. V., XIE, B., AHLFORD, A., YOON, J.-K., ROSENBAUM, A. M. & ZARANEK, A. W. 2009b. Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome research,* 19**,** 1606-1615.

LI, Q., RITTER, D., YANG, N., DONG, Z., LI, H., CHUANG, J. H. & GUO, S. 2010. A systematic approach to identify functional motifs within vertebrate developmental enhancers. *Developmental biology,* 337**,** 484-495.

LI, R., LI, Y., FANG, X., YANG, H., WANG, J., KRISTIANSEN, K. & WANG, J. 2009c. SNP detection for massively parallel whole-genome resequencing. *Genome research,* 19**,** 1124-1132.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J. & DORSCHNER, M. O. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science,* 326**,** 289-293.

LIEBLICH, J. M., ROGOL, A. D., WHITE, B. J. & ROSEN, S. W. 1982. Syndrome of anosmia with hypogonadotropic hypogonadism (Kallmann syndrome): clinical

and laboratory studies in 23 cases. *The American journal of medicine,* 73**,** 506-519.

LITTLE, J., CARDY, A. & MUNGER, R. G. 2004. Tobacco smoking and oral clefts: a meta-analysis. *Bulletin of the World Health Organization,* 82**,** 213-218.

LIU, H., LESLIE, E. J., CARLSON, J. C., BEATY, T. H., MARAZITA, M. L., LIDRAL, A. C. & CORNELL, R. A. 2017. Identification of common non-coding variants at 1p22 that are functional for non-syndromic orofacial clefting. *Nature Communications,* 8**,** ncomms14759.

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E., PERS, T. H., DAY, F. R., POWELL, C., VEDANTAM, S., BUCHKOVICH, M. L. & YANG, J. 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature,* 518**,** 197-206.

LONGLEY, M. J., CLARK, S., MAN, C. Y. W., HUDSON, G., DURHAM, S. E., TAYLOR, R. W., NIGHTINGALE, S., TURNBULL, D. M., COPELAND, W. C. & CHINNERY, P. F. 2006. Mutant POLG2 disrupts DNA polymerase γ subunits and causes progressive external ophthalmoplegia. *The American Journal of Human Genetics,* 78**,** 1026-1034.

LOOTS, G. G., KNEISSEL, M., KELLER, H., BAPTIST, M., CHANG, J., COLLETTE, N. M., OVCHARENKO, D., PLAJZER-FRICK, I. & RUBIN, E. M. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res,* 15**,** 928-35.

LU, Y., SHEN, Y., WARREN, W. & WALTER, R. 2016. Next Generation Sequencing in Aquatic Models. *Next Generation Sequencing-Advances, Applications and Challenges.* InTech.

LU, D. AND XU, S., 2013. Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia. Frontiers in genetics, 4, p.127.

LUPIÁÑEZ, D. G., KRAFT, K., HEINRICH, V., KRAWITZ, P., BRANCATI, F., KLOPOCKI, E., HORN, D., KAYSERILI, H., OPITZ, J. M. & LAXOVA, R. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell,* 161**,** 1012-1025.

LYGONIS, C. S. 1969. Familiar absence of olfaction. *Hereditas,* 61**,** 413-416.

MAEDA, R. K. & KARCH, F. 2011. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Current opinion in genetics & development,* 21**,** 187-193.

MAINLAND, R. C. 1945. Absence of olfactory sensation. *Journal of Heredity,* 36**,** 143-144.

MANGOLD, E. 2010. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genet.,* 42**,** 24-26.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R. & CHAKRAVARTI, A. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747.

MARAZITA, M. L. 2009. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype specific differences in linkage and association results. *Hum. Hered.,* 68**,** 151-170.

MARCELINO, L. A. & THILLY, W. G. 1999. Mitochondrial mutagenesis in human cells and tissues. *Mutation Research/DNA Repair,* 434**,** 177-203.

MARCHINI, J. & HOWIE, B. 2010. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics,* 11**,** 499.

MARK, C., JIM, D., MARTIN, D., LEILA, E., TOM, F., SUE, H., TIM, H., LUKE, J., NICK, M., JEANNA, M.-P., GIL, M., KATRINA, N.-R., MATTHEW, P., VIVIENNE, P., AUGUSTO, R., LAURA, R., CLARE, T. & KERRIE, W. 2017. *The 100,000 Genomes Project Protocol*.

MARTI, R., DORADO, B. & HIRANO, M. 2012. Measurement of mitochondrial dNTP pools. *Methods Mol Biol,* 837**,** 135-48.

MARTIN, E. R., KINNAMON, D., SCHMIDT, M. A., POWELL, E., ZUCHNER, S. & MORRIS, R. 2010. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics,* 26**,** 2803-2810.

MATHELIER, A., FORNES, O., ARENILLAS, D. J., CHEN, C.-Y., DENAY, G., LEE, J., SHI, W., SHYR, C., TAN, G. & WORSLEY-HUNT, R. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research,* 44**,** D110-D115.

MATHER, C. A., MOONEY, S. D., SALIPANTE, S. J., SCROGGINS, S., WU, D., PRITCHARD, C. C. & SHIRTS, B. H. 2016. CADD score has limited clinical validity for the identification of pathogenic variants in non-coding regions in a hereditary cancer panel. *Genetics in medicine: official journal of the American College of Medical Genetics*.

MATHERS, C. 2008. *The global burden of disease: 2004 update*, World Health Organization.

MATHEWS, C. K. 2006. DNA precursor metabolism and genomic stability. *The FASEB Journal,* 20**,** 1300-1314.

MATHEWS, C. K. & SONG, S. 2007. Maintaining precursor pools for mitochondrial DNA replication. *Faseb j,* 21**,** 2294-303.

MATYS, V., FRICKE, E., GEFFERS, R., GÖSSLIN, E., HAUBROCK, M., HEHL, R., HORNISCHER, K., KARAS, D., KEL, A. E. & KEL-MARGOULIS, O. V. 2003. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research,* 31**,** 374-378.

MATYS, V., KEL-MARGOULIS, O. V., FRICKE, E., LIEBICH, I., LAND, S., BARRE-DIRRIE, A., REUTER, I., CHEKMENEV, D., KRULL, M. & HORNISCHER, K. 2006. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research,* 34**,** D108-D110.

MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H. & BRODY, J. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science,* 337**,** 1190-1195.

MCEVILLY, R. J., DE DIAZ, M. O., SCHONEMANN, M. D., HOOSHMAND, F. & ROSENFELD, M. G. 2002. Transcriptional regulation of cortical neuron migration by POU domain factors. *Science,* 295**,** 1528-32.

MCEWEN, G. K., GOODE, D. K., PARKER, H. J., WOOLFE, A., CALLAWAY, H. & ELGAR, G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS genetics,* 5**,** e1000762.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S. & DALY, M. 2010. The Genome Analysis Toolkit: a MapReduce framework for

analyzing next-generation DNA sequencing data. *Genome research,* 20**,** 1297-1303.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R. S., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biology,* 17**,** 122.

MEIENBERG, J., BRUGGMANN, R., OEXLE, K. & MATYAS, G. 2016. Clinical sequencing: is WGS the better WES? *Human genetics,* 135**,** 359-362.

MEIENBERG, J., ZERJAVIC, K., KELLER, I., OKONIEWSKI, M., PATRIGNANI, A., LUDIN, K., XU, Z., STEINMANN, B., CARREL, T. & RÖTHLISBERGER, B. 2015. New insights into the performance of human whole-exome capture platforms. *Nucleic acids research,* 43**,** e76-e76.

MELNIKOV, A., MURUGAN, A., ZHANG, X., TESILEANU, T., WANG, L., ROGOV, P., FEIZI, S., GNIRKE, A., CALLAN, C. G., KINNEY, J. B., KELLIS, M., LANDER, E. S. & MIKKELSEN, T. S. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotech,* 30**,** 271-277.

MEYNERT, A. M., ANSARI, M., FITZPATRICK, D. R. & TAYLOR, M. S. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC bioinformatics,* 15**,** 247.

MILLS, R. E., LUTTIG, C. T., LARKINS, C. E., BEAUCHAMP, A., TSUI, C., PITTARD, W. S. & DEVINE, S. E. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research,* 16**,** 1182-1190.

MONTGOMERY, S. B., LAPPALAINEN, T., GUTIERREZ-ARCELUS, M. & DERMITZAKIS, E. T. 2011. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS genetics,* 7**,** e1002144.

MORGAN, C. L., BAXTER, H. & KERR, M. P. 2003. Prevalence of epilepsy and associated health service utilization and mortality among patients with intellectual disability. *Am J Ment Retard,* 108**,** 293-300.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods,* 5**,** 621-628.

MOSSEY, P. A., LITTLE, J., MUNGER, R. G., DIXON, M. J. & SHAW, W. C. 2009. Cleft lip and palate. *The Lancet,* 374**,** 1773-1785.

MYERS, C. T. & MEFFORD, H. C. 2015. Advancing epilepsy genetics in the genomic era. *Genome Medicine,* 7**,** 91.

NENADIC, I., GASER, C. & SAUER, H. 2012. Heterogeneity of brain structural variation and the structural imaging endophenotypes in schizophrenia. *Neuropsychobiology,* 66**,** 44-49.

NG, P. C. & HENIKOFF, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research,* 31**,** 3812-3814.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics,* 12**,** 443.

NISHIGAKI, Y., MARTI, R., COPELAND, W. C. & HIRANO, M. 2003. Site-specific somatic mitochondrial DNA point mutations in patients with thymidine phosphorylase deficiency. *J Clin Invest,* 111**,** 1913-21.

NIX, D. A., COURDY, S. J. & BOUCHER, K. M. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC bioinformatics,* 9**,** 523.

NOHR, E. A., TIMPSON, N. J., ANDERSEN, C. S., SMITH, G. D., OLSEN, J. & SØRENSEN, T. I. 2009. Severe obesity in young women and reproductive health: the Danish National Birth Cohort. *PLoS One,* 4**,** e8444.

NORA, E. P., DEKKER, J. & HEARD, E. 2013. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays,* 35**,** 818-828.

NORA, E. P., LAJOIE, B. R., SCHULZ, E. G., GIORGETTI, L., OKAMOTO, I., SERVANT, N., PIOLOT, T., VAN BERKUM, N. L., MEISIG, J. & SEDAT, J. 2012. Spatial partitioning of the regulatory landscape of the x-inactivation center. *Nature,* 485**,** 381.

NORDIN, S. & BRÄMERSON, A. 2008. Complaints of olfactory disorders: epidemiology, assessment and clinical implications. *Current opinion in allergy and clinical immunology,* 8**,** 10-15.

NOYVERT, B. 2015. *TidyVar* [Online]. GitHub Repository. Available: https://github.com/boris-noyvert/TidyVar.m [Accessed 2017].

O'DONOVAN, M. C., CRADDOCK, N., NORTON, N., WILLIAMS, H., PEIRCE, T., MOSKVINA, V., NIKOLOV, I., HAMSHERE, M., CARROLL, L., GEORGIEVA, L., DWYER, S., HOLMANS, P., MARCHINI, J. L., SPENCER, C. C., HOWIE, B., LEUNG, H. T., HARTMANN, A. M., MOLLER, H. J., MORRIS, D. W., SHI, Y., FENG, G., HOFFMANN, P., PROPPING, P., VASILESCU, C., MAIER, W., RIETSCHEL, M., ZAMMIT, S., SCHUMACHER, J., QUINN, E. M., SCHULZE, T. G., WILLIAMS, N. M., GIEGLING, I., IWATA, N., IKEDA, M., DARVASI, A., SHIFMAN, S., HE, L., DUAN, J., SANDERS, A. R., LEVINSON, D. F., GEJMAN, P. V., CICHON, S., NOTHEN, M. M., GILL, M., CORVIN, A., RUJESCU, D., KIROV, G., OWEN, M. J., BUCCOLA, N. G., MOWRY, B. J., FREEDMAN, R., AMIN, F., BLACK, D. W., SILVERMAN, J. M., BYERLEY, W. F., CLONINGER, C. R. & MOLECULAR GENETICS OF SCHIZOPHRENIA, C. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet,* 40**,** 1053-5.

OHNO, S. So much" junk" DNA in our genome. Brookhaven symposia in biology, 1972. 366-370.

OLDONI, F., PALMEN, J., GIAMBARTOLOMEI, C., HOWARD, P., DRENOS, F., PLAGNOL, V., HUMPHRIES, S. E., TALMUD, P. J. & SMITH, A. J. 2016. Post-GWAS methodologies for localisation of functional non-coding variants: ANGPTL3. *Atherosclerosis,* 246**,** 193-201.

ONG, C.-T. & CORCES, V. G. 2014. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics,* 15**,** 234-246.

ONSTAD, S., SKRE, I., TORGERSEN, S. & KRINGLEN, E. 1991. Twin concordance for DSM-III-R schizophrenia. *Acta Psychiatr Scand,* 83**,** 395-401.

OZAKI, K., OHNISHI, Y., IIDA, A., SEKINE, A., YAMADA, R., TSUNODA, T., SATO, H., SATO, H., HORI, M. & NAKAMURA, Y. 2002. Functional SNPs in the lymphotoxin-[alpha] gene that are associated with susceptibility to myocardial infarction. *Nature genetics,* 32**,** 650.

PANTELIS, C., YUCEL, M., WOOD, S. J., VELAKOULIS, D., SUN, D., BERGER, G., STUART, G. W., YUNG, A., PHILLIPS, L. & MCGORRY, P. D. 2005. Structural brain imaging evidence for multiple pathological processes at different stages of brain development in schizophrenia. *Schizophr Bull,* 31**,** 672-96.

PARKER, H. J., PICCINELLI, P., SAUKA-SPENGLER, T., BRONNER, M. & ELGAR, G. 2011. Ancient Pbx-Hox signatures define hundreds of vertebrate developmental enhancers. *BMC Genomics,* 12**,** 637.

PATERNOSTER, L., EVANS, D. M., NOHR, E. A., HOLST, C., GABORIEAU, V., BRENNAN, P., GJESING, A. P., GRARUP, N., WITTE, D. R., JORGENSEN, T., LINNEBERG, A., LAURITZEN, T., SANDBAEK, A., HANSEN, T., PEDERSEN, O., ELLIOTT, K. S., KEMP, J. P., ST POURCAIN, B., MCMAHON, G., ZELENIKA, D., HAGER, J., LATHROP, M., TIMPSON, N. J., SMITH, G. D. & SORENSEN, T. I. 2011. Genome-wide population-based association study of extremely overweight young adults--the GOYA study. *PLoS One,* 6**,** e24303.

PATWARDHAN, R. P., HIATT, J. B., WITTEN, D. M., KIM, M. J., SMITH, R. P., MAY, D., LEE, C., ANDRIE, J. M., LEE, S.-I. & COOPER, G. M. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology,* 30**,** 265-270.

PAUWS, E. 2009. Tbx22 null mice have a submucous cleft palate due to reduced palatal bone formation and also display ankyloglossia and choanal atresia phenotypes. *Hum. Mol. Genet.,* 18**,** 4171-4179.

PENNACCHIO, L. A., AHITUV, N., MOSES, A. M., PRABHAKAR, S., NOBREGA, M. A., SHOUKRY, M., MINOVITSKY, S., DUBCHAK, I., HOLT, A. & LEWIS, K. D. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature,* 444**,** 499.

PETERS, T., DILDROP, R., AUSMEIER, K. & RUTHER, U. 2000. Organization of mouse Iroquois homeobox genes in two clusters suggests a conserved regulation and function in vertebrate development. *Genome Res,* 10**,** 1453-62.

PIASECKA, B., LICHOCKI, P., MORETTI, S., BERGMANN, S. & ROBINSON-RECHAVI, M. 2013. The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates. *PLoS genetics,* 9**,** e1003476.

PICKRELL, J. K., GAFFNEY, D. J., GILAD, Y. & PRITCHARD, J. K. 2011. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics,* 27**,** 2144-2146.

PINTO, J. M., THANAVIRATANANICH, S., HAYES, M. G., NACLERIO, R. M. & OBER, C. 2008. A genome-wide screen for hyposmia susceptibility loci. *Chemical senses,* 33**,** 319-329.

PITTELOUD, N., ACIERNO, J. S., MEYSING, A., ELISEENKOVA, A. V., MA, J., IBRAHIMI, O. A., METZGER, D. L., HAYES, F. J., DWYER, A. A. & HUGHES, V. A. 2006. Mutations in fibroblast growth factor receptor 1 cause both Kallmann syndrome and normosmic idiopathic hypogonadotropic hypogonadism. *Proceedings of the National Academy of Sciences,* 103**,** 6281-6286.

POPEJOY, A. B. & FULLERTON, S. M. 2016. Genomics is failing on diversity. *Nature,* 538**,** 161-164.

POTKIN, S. G., TURNER, J. A., GUFFANTI, G., LAKATOS, A., FALLON, J. H., NGUYEN, D. D., MATHALON, D., FORD, J., LAURIELLO, J. & MACCIARDI, F. 2008. A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophrenia bulletin,* 35**,** 96-108.

POULTON, J., HIRANO, M., SPINAZZOLA, A., HERNANDEZ, M. A., JARDEL, C., LOMBES, A., CZERMIN, B., HORVATH, R., TAANMAN, J. & ROTIG, A. 2009. Collated mutations in mitochondrial DNA (mtDNA) depletion syndrome (excluding the mitochondrial gamma polymerase, POLG1). *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* 1792**,** 1109-1112.

PRABHAKAR, S., POULIN, F., SHOUKRY, M., AFZAL, V., RUBIN, E. M., COURONNE, O. & PENNACCHIO, L. A. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome research,* 16**,** 855-863.

PULAK, R. 2016. Tools for automating the imaging of zebrafish larvae. *Methods,* 96**,** 118-126.

RAFF, R. A. 2012. *The shape of life: genes, development, and the evolution of animal form*, University of Chicago Press.

RAGVIN, A., MORO, E., FREDMAN, D., NAVRATILOVA, P., DRIVENES, O., ENGSTROM, P. G., ALONSO, M. E., DE LA CALLE MUSTIENES, E., GOMEZ SKARMETA, J. L., TAVARES, M. J., CASARES, F., MANZANARES, M., VAN HEYNINGEN, V., MOLVEN, A., NJOLSTAD, P. R., ARGENTON, F., LENHARD, B. & BECKER, T. S. 2010. Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc Natl Acad Sci U S A,* 107**,** 775-80.

RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols,* 8**,** 2281-2308.

RAO, S. S., HUNTLEY, M. H., DURAND, N. C., STAMENOVA, E. K., BOCHKOV, I. D., ROBINSON, J. T., SANBORN, A. L., MACHOL, I., OMER, A. D. & LANDER, E. S. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell,* 159**,** 1665-1680.

RAPOPORT, J. L., ADDINGTON, A. M., FRANGOU, S. & PSYCH, M. R. 2005. The neurodevelopmental model of schizophrenia: update 2005. *Mol Psychiatry,* 10**,** 434-49.

REMESEIRO, S., HÖRNBLAD, A. & SPITZ, F. 2016. Gene regulation during development in the light of topologically associating domains. *Wiley Interdisciplinary Reviews: Developmental Biology,* 5**,** 169-185.

RIBEIRO, A., GOLICZ, A., HACKETT, C. A., MILNE, I., STEPHEN, G., MARSHALL, D., FLAVELL, A. J. & BAYER, M. 2015. An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC Bioinformatics,* 16**,** 382.

RICHARDSON, T. G., CAMPBELL, C., TIMPSON, N. J. & GAUNT, T. R. 2016. Incorporating Non-Coding Annotations into Rare Variant Analysis. *PloS one,* 11**,** e0154181.

RIETHOVEN, J.-J. M. 2010. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Computational Biology of Transcription Factor Binding***,** 33-42.

RILEY, B., THISELTON, D., MAHER, B. S., BIGDELI, T., WORMLEY, B., MCMICHAEL, G. O., FANOUS, A. H., VLADIMIROV, V., O'NEILL, F. A., WALSH, D. & KENDLER, K. S. 2010. Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. *Mol Psychiatry,* 15**,** 29-37.

RILEY, B. M. & MURRAY, J. C. 2007. Sequence Evaluation of FGF and FGFR Gene Conserved Non-Coding Elements in Non-Syndromic Cleft Lip and Palate Cases. *American journal of medical genetics. Part A,* 143A**,** 3228-3234.

ROUILLARD, A. D., GUNDERSEN, G. W., FERNANDEZ, N. F., WANG, Z., MONTEIRO, C. D., MCDERMOTT, M. G. & MA'AYAN, A. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database,* 2016.

ROUSSOS, P., MITCHELL, A. C., VOLOUDAKIS, G., FULLARD, J. F., POTHULA, V. M., TSANG, J., STAHL, E. A., GEORGAKOPOULOS, A., RUDERFER, D. M. & CHARNEY, A. 2014. A role for noncoding variation in schizophrenia. *Cell reports,* 9**,** 1417-1429.

ROZEN, S. & SKALETSKY, H. 1999. Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics methods and protocols***,** 365-386.

SAADA, A., SHAAG, A., MANDEL, H., NEVO, Y., ERIKSSON, S. & ELPELEG, O. 2001. Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. *Nature genetics,* 29**,** 342.

SACHIDANANDAM, R., WEISSMAN, D., SCHMIDT, S. C., KAKOL, J. M., STEIN, L. D., MARTH, G., SHERRY, S., MULLIKIN, J. C., MORTIMORE, B. J. & WILLEY, D. L. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409**,** 928-934.

SAHA, S., CHANT, D. & MCGRATH, J. 2007. A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time? *Archives of general psychiatry,* 64**,** 1123-1131.

SAMOCHA, K.E., ROBINSON, E.B., SANDERS, S.J., STEVENS, C., SABO, A., MCGRATH, L.M., KOSMICKI, J.A., REHNSTRÖM, K., MALLICK, S., KIRBY, A. AND WALL, D.P., 2014. A framework for the interpretation of de novo mutation in human disease. Nature genetics, 46(9), p.944.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-5467.

SANYAL, A., LAJOIE, B., JAIN, G. & DEKKER, J. 2012. The long-range interaction landscape of gene promoters. *Nature,* 489**,** 109.

SARACENO, B. & BERTOLOTE, J. M. 2013. <Schizophrenia.pdf>. *World Health Organisation.*

SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511**,** 421-427.

SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P. & MACKAY, S. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science,* 328**,** 1036-1040.

SCHMITT, A. D., HU, M., JUNG, I., XU, Z., QIU, Y., TAN, C. L., LI, Y., LIN, S., LIN, Y. & BARR, C. L. 2016. A compendium of chromatin contact maps

reveals spatially active regions in the human genome. *Cell reports,* 17**,** 2042-2059.

SCUTERI, A., SANNA, S., CHEN, W. M., UDA, M., ALBAI, G., STRAIT, J., NAJJAR, S., NAGARAJA, R., ORRU, M., USALA, G., DEI, M., LAI, S., MASCHIO, A., BUSONERO, F., MULAS, A., EHRET, G. B., FINK, A. A., WEDER, A. B., COOPER, R. S., GALAN, P., CHAKRAVARTI, A., SCHLESSINGER, D., CAO, A., LAKATTA, E. & ABECASIS, G. R. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *Plos Genetics,* 3**,** 1200-1210.

SHAM, P. C., MACLEAN, C. J. & KENDLER, K. S. 1994. A typological model of schizophrenia based on age at onset, sex and familial morbidity. *Acta Psychiatr Scand,* 89**,** 135-41.

SHERRY, S. T., WARD, M. & SIROTKIN, K. 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research,* 9**,** 677-679.

SHERRY, S. T., WARD, M.-H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research,* 29**,** 308-311.

SHI, M. 2007. Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *Am. J. Hum. Genet.,* 80**,** 76-90.

SHI, M., WEHBY, G. L. & MURRAY, J. C. 2008. Review on genetic variants and maternal smoking in the etiology of oral clefts and other birth defects. *Birth Defects Res. C Embryo Today,* 84**,** 16-29.

SIMONIS, M., KLOUS, P., SPLINTER, E., MOSHKIN, Y., WILLEMSEN, R., DE WIT, E., VAN STEENSEL, B. & DE LAAT, W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics,* 38**,** 1348.

SIMS, D., SUDBERY, I., ILOTT, N. E., HEGER, A. & PONTING, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics,* 15**,** 121-132.

SINGH, N., GREWAL, M. S. & AUSTIN, J. H. 1970. Familial anosmia. *Archives of Neurology,* 22**,** 40-44.

SMEMO, S., TENA, J. J., KIM, K. H., GAMAZON, E. R., SAKABE, N. J., GOMEZ-MARIN, C., ANEAS, I., CREDIDIO, F. L., SOBREIRA, D. R., WASSERMAN, N. F., LEE, J. H., PUVIINDRAN, V., TAM, D., SHEN, M., SON, J. E., VAKILI, N. A., SUNG, H. K., NARANJO, S., ACEMEL, R. D., MANZANARES, M., NAGY, A., COX, N. J., HUI, C. C., GOMEZ-SKARMETA, J. L. & NOBREGA, M. A. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature,* 507**,** 371-5.

SMITH, A. D., XUAN, Z. & ZHANG, M. Q. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics,* 9**,** 128.

SMITH, A. J., HOWARD, P., SHAH, S., ERIKSSON, P., STENDER, S., GIAMBARTOLOMEI, C., FOLKERSEN, L., TYBJÆRG-HANSEN, A., KUMARI, M. & PALMEN, J. 2012. Use of allele-specific FAIRE to determine functional regulatory polymorphism using large-scale genotyping arrays. *PLoS genetics,* 8**,** e1002908.

SOBALSKA-KWAPIS, M., SUCHANECKA, A., SŁOMKA, M., SIEWIERSKA-GÓRSKA, A., KĘPKA, E. & STRAPAGIEL, D. 2017. Genetic association of FTO/IRX region with obesity and overweight in the Polish population. *PloS one,* 12**,** e0180295.

SONG, S., PURSELL, Z. F., COPELAND, W. C., LONGLEY, M. J., KUNKEL, T. A. & MATHEWS, C. K. 2005. DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 4990-4995.

SPINAZZOLA, A., SANTER, R., AKMAN, O. H., TSIAKAS, K., SCHAEFER, H., DING, X., KARADIMAS, C. L., SHANSKE, S., GANESH, J. & DI MAURO, S. 2008. Hepatocerebral form of mitochondrial DNA depletion syndrome: novel MPV17 mutations. *Archives of neurology,* 65**,** 1108-1113.

SPINAZZOLA, A., VISCOMI, C., FERNANDEZ-VIZARRA, E., CARRARA, F., D'ADAMO, P., CALVO, S., MARSANO, R. M., DONNINI, C., WEIHER, H. & STRISCIUGLIO, P. 2006. MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nature genetics,* 38**,** 570.

SRIVASTAVA, D., THOMAS, T., LIN, Q., KIRBY, M. L., BROWN, D. & OLSON, E. N. 1997. Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nature genetics,* 16**,** 154-160.

STEFANSSON, H., OPHOFF, R. A., STEINBERG, S., ANDREASSEN, O. A., CICHON, S., RUJESCU, D., WERGE, T., PIETILAINEN, O. P., MORS, O., MORTENSEN, P. B., SIGURDSSON, E., GUSTAFSSON, O., NYEGAARD, M., TUULIO-HENRIKSSON, A., INGASON, A., HANSEN, T., SUVISAARI, J., LONNQVIST, J., PAUNIO, T., BORGLUM, A. D., HARTMANN, A., FINK-JENSEN, A., NORDENTOFT, M., HOUGAARD, D., NORGAARD-PEDERSEN, B., BOTTCHER, Y., OLESEN, J., BREUER, R., MOLLER, H. J., GIEGLING, I., RASMUSSEN, H. B., TIMM, S., MATTHEISEN, M., BITTER, I., RETHELYI, J. M., MAGNUSDOTTIR, B. B., SIGMUNDSSON, T., OLASON, P., MASSON, G., GULCHER, J. R., HARALDSSON, M., FOSSDAL, R., THORGEIRSSON, T. E., THORSTEINSDOTTIR, U., RUGGERI, M., TOSATO, S., FRANKE, B., STRENGMAN, E., KIEMENEY, L. A., GENETIC, R., OUTCOME IN, P., MELLE, I., DJUROVIC, S., ABRAMOVA, L., KALEDA, V., SANJUAN, J., DE FRUTOS, R., BRAMON, E., VASSOS, E., FRASER, G., ETTINGER, U., PICCHIONI, M., WALKER, N., TOULOPOULOU, T., NEED, A. C., GE, D., YOON, J. L., SHIANNA, K. V., FREIMER, N. B., CANTOR, R. M., MURRAY, R., KONG, A., GOLIMBET, V., CARRACEDO, A., ARANGO, C., COSTAS, J., JONSSON, E. G., TERENIUS, L., AGARTZ, I., PETURSSON, H., NOTHEN, M. M., RIETSCHEL, M., MATTHEWS, P. M., MUGLIA, P., PELTONEN, L., ST CLAIR, D., GOLDSTEIN, D. B., STEFANSSON, K. & COLLIER, D. A. 2009. Common variants conferring risk of schizophrenia. *Nature,* 460**,** 744-7.

STEVENSON, R. E., HOLDEN, K. R., ROGERS, R. C. & SCHWARTZ, C. E. 2012. Seizures and X-linked intellectual disability. *European journal of medical genetics,* 55**,** 307-312.

STRAHLE, U. & RASTEGAR, S. 2008. Conserved non-coding sequences and transcriptional regulation. *Brain Res Bull,* 75**,** 225-30.

STRATIGOPOULOS, G., LEDUC, C. A., CREMONA, M. L., CHUNG, W. K. & LEIBEL, R. L. 2011. Cut-like homeobox 1 (CUX1) regulates expression of the fat mass and obesity-associated and retinitis pigmentosa GTPase regulator-interacting protein-1-like (RPGRIP1L) genes and coordinates leptin receptor signaling. *J Biol Chem,* 286**,** 2155-70.

STRATIGOPOULOS, G., MARTIN CARLI, J. F., O'DAY, D. R., WANG, L., LEDUC, C. A., LANZANO, P., CHUNG, W. K., ROSENBAUM, M., EGLI, D. & DOHERTY, D. A. 2014. Hypomorphism for< i> RPGRIP1L</i>, a Ciliary Gene Vicinal to the< i> FTO</i> Locus, Causes Increased Adiposity in Mice. *Cell metabolism,* 19**,** 767-779.

SUDDATH , R. L., CHRISTISON , G. W., TORREY , E. F., CASANOVA , M. F. & WEINBERGER , D. R. 1990. Anatomical Abnormalities in the Brains of Monozygotic Twins Discordant for Schizophrenia. *New England Journal of Medicine,* 322**,** 789-794.

SUGITANI, Y., NAKAI, S., MINOWA, O., NISHI, M., JISHAGE, K., KAWANO, H., MORI, K., OGAWA, M. & NODA, T. 2002. Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. *Genes Dev,* 16**,** 1760-5.

SUMNER, M., BELL, S. & HWANG, P. A. 2013. 2. Fetal alcohol spectrum disorder: Epilepsy and neuropsychiatric disorders. *Clinical Neurophysiology,* 124**,** e5.

SUZUKI, S. 2009. Mutations in BMP4 are associated with subepithelial, microform, and overt cleft lip. *Am. J. Hum. Genet.,* 84**,** 406-411.

TENA, J. J., ALONSO, M. E., DE LA CALLE-MUSTIENES, E., SPLINTER, E., DE LAAT, W., MANZANARES, M. & GOMEZ-SKARMETA, J. L. 2011. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat Commun,* 2**,** 310.

TEWHEY, R., NAKANO, M., WANG, X., PABON-PENA, C., NOVAK, B., GIUFFRE, A., LIN, E., HAPPE, S., ROBERTS, D. N., LEPROUST, E. M., TOPOL, E. J., HARISMENDY, O. & FRAZER, K. A. 2009. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol,* 10**,** R116.

THE GENOMES PROJECT, C. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

THE, U. K. K. C. 2015. The UK10K project identifies rare variants in health and disease. *Nature,* 526**,** 82-90.

THOMAS-CHOLLIER, M., HUFTON, A., HEINIG, M., O'KEEFFE, S., EL MASRI, N., ROIDER, H. G., MANKE, T. & VINGRON, M. 2011. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature protocols,* 6**,** 1860.

TUNG, Y. L., YEO, G. S., O'RAHILLY, S. & COLL, A. P. 2014. Obesity and FTO: changing focus at a complex locus. *Cell metabolism,* 20**,** 710-718.

UUSIMAA, J., EVANS, J., SMITH, C., BUTTERWORTH, A., CRAIG, K., ASHLEY, N., LIAO, C., CARVER, J., DIOT, A. & MACLEOD, L. 2014. Clinical, biochemical, cellular and molecular characterization of mitochondrial DNA depletion syndrome due to novel mutations in the MPV17 gene. *European Journal of Human Genetics,* 22**,** 184.

VALOUEV, A., JOHNSON, D. S., SUNDQUIST, A., MEDINA, C., ANTON, E., BATZOGLOU, S., MYERS, R. M. & SIDOW, A. 2008. Genome-wide analysis

of transcription factor binding sites based on ChIP-Seq data. *Nature methods,* 5**,** 829-834.

VAN DEN BOOGAARD, M., SMEMO, S., BURNICKA-TUREK, O., ARNOLDS, D. E., VAN DE WERKEN, H. J., KLOUS, P., MCKEAN, D., MUEHLSCHLEGEL, J. D., MOOSMANN, J. & TOKA, O. 2014. A common genetic variant within SCN10A modulates cardiac SCN5A expression. *The Journal of clinical investigation,* 124**,** 1844.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY‐MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D. & THIBAULT, J. 2013. From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics***,** 11.10. 1-11.10. 33.

VAN GOETHEM, G., DERMAUT, B., LÖFGREN, A., MARTIN, J.-J. & VAN BROECKHOVEN, C. 2001. Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nature genetics,* 28**,** 211.

VAN ROOIJ, I. A. 2001. Smoking, genetic polymorphisms in biotransformation enzymes, and nonsyndromic oral clefting: a gene-environment interaction. *Epidemiology,* 12**,** 502-507.

VERMULST, M., BIELAS, J.H. AND LOEB, L.A., 2008. Quantification of random mutations in the mitochondrial genome. Methods, 46(4), pp.263-268.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A. & HOLT, R. A. 2001. The sequence of the human genome. *science,* 291**,** 1304-1351.

VILLALOBOS-COMPARAN, M., TERESA FLORES-DORANTES, M., TERESA VILLARREAL-MOLINA, M., RODRIGUEZ-CRUZ, M., GARCIA-ULLOA, A. C., ROBLES, L., HUERTAS-VAZQUEZ, A., SAUCEDO-VILLARREAL, N., LOPEZ-ALARCON, M., SANCHEZ-MUNOZ, F., DOMINGUEZ-LOPEZ, A., GUTIERREZ-AGUILAR, R., MENJIVAR, M., CORAL-VAZQUEZ, R., HERNANDEZ-STENGELE, G., VITAL-REYES, V. S., ACUNA-ALONZO, V., ROMERO-HIDALGO, S., RUIZ-GOMEZ, D. G., RIANO-BARROS, D., HERRERA, M. F., GOMEZ-PEREZ, F. J., FROGUEL, P., GARCIA-GARCIA, E., TERESA TUSIE-LUNA, M., AGUILAR-SALINAS, C. A. & CANIZALES-QUINTEROS, S. 2008. The FTO gene is associated with adulthood obesity in the Mexican population. *Obesity (Silver Spring),* 16**,** 2296-301.

VILLAVICENCIO, E. H., WALTERHOUSE, D. O. & IANNACCONE, P. M. 2000. The sonic hedgehog–patched–gli pathway in human development and disease. *The American Journal of Human Genetics,* 67**,** 1047-1054.

VISCOMI, C., SPINAZZOLA, A., MAGGIONI, M., FERNANDEZ-VIZARRA, E., MASSA, V., PAGANO, C., VETTOR, R., MORA, M. & ZEVIANI, M. 2009. Early-onset liver mtDNA depletion and late-onset proteinuric nephropathy in Mpv17 knockout mice. *Human Molecular Genetics,* 18**,** 12-26.

VISEL, A., BLOW, M. J., LI, Z., ZHANG, T., AKIYAMA, J. A., HOLT, A., PLAJZER-FRICK, I., SHOUKRY, M., WRIGHT, C., CHEN, F., AFZAL, V., REN, B., RUBIN, E. M. & PENNACCHIO, L. A. 2009a. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* 457**,** 854-858.

VISEL, A., MINOVITSKY, S., DUBCHAK, I. & PENNACCHIO, L. A. 2006. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic acids research,* 35**,** D88-D92.

VISEL, A., RUBIN, E. M. & PENNACCHIO, L. A. 2009b. Genomic views of distant-acting enhancers. *Nature,* 461**,** 199.

VOWLES, R. H., BLEACH, N. R. & ROWE-JONES, J. M. 1997. Congenital anosmia. *International Journal of Pediatric Otorhinolaryngology,* 41**,** 207-214.

WÅHLÉN, K., SJÖLIN, E. & HOFFSTEDT, J. 2008. The common rs9939609 gene variant of the fat mass-and obesity-associated gene FTO is related to fat cell lipolysis. *Journal of lipid Research,* 49**,** 607-611.

WALSH, T., MCCLELLAN, J. M., MCCARTHY, S. E., ADDINGTON, A. M., PIERCE, S. B., COOPER, G. M., NORD, A. S., KUSENDA, M., MALHOTRA, D. & BHANDARI, A. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *science,* 320**,** 539-543.

WANG, J., WANG, W., LI, R., LI, Y., TIAN, G., GOODMAN, L., FAN, W., ZHANG, J., LI, J. & ZHANG, J. 2008. The diploid genome sequence of an Asian individual. *Nature,* 456**,** 60.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research,* 38**,** e164-e164.

WANG, Y., LI, G., STANCO, A., LONG, J. E., CRAWFORD, D., POTTER, G. B., PLEASURE, S. J., BEHRENS, T. & RUBENSTEIN, J. L. 2011. CXCR4 and CXCR7 have distinct functions in regulating interneuron migration. *Neuron,* 69**,** 61-76.

WARD, L. D. & KELLIS, M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology,* 30**,** 1095-1106.

WARDEN, C. D., ADAMSON, A. W., NEUHAUSEN, S. L. & WU, X. 2014. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ,* 2**,** e600.

WASSERMAN, W. W. & SANDELIN, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics,* 5**,** 276.

WEHBY, G. & CASSELL, C. H. 2010. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral diseases,* 16**,** 3-10.

WELLCOME TRUST CASE CONTROL, C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

WESTERINEN, H., KASKI, M., VIRTA, L. J., ALMQVIST, F. & IIVANAINEN, M. 2014. Age-specific prevalence of intellectual disability in Finland at the beginning of new millennium--multiple register method. *J Intellect Disabil Res,* 58**,** 285-95.

WILLIAMS, H. J., NORTON, N., DWYER, S., MOSKVINA, V., NIKOLOV, I., CARROLL, L., GEORGIEVA, L., WILLIAMS, N. M., MORRIS, D. W., QUINN, E. M., GIEGLING, I., IKEDA, M., WOOD, J., LENCZ, T., HULTMAN, C., LICHTENSTEIN, P., THISELTON, D., MAHER, B. S., MOLECULAR GENETICS OF SCHIZOPHRENIA COLLABORATION INTERNATIONAL SCHIZOPHRENIA CONSORTIUM, S.-P. G., MALHOTRA, A. K., RILEY, B., KENDLER, K. S., GILL, M., SULLIVAN, P.,

SKLAR, P., PURCELL, S., NIMGAONKAR, V. L., KIROV, G., HOLMANS, P., CORVIN, A., RUJESCU, D., CRADDOCK, N., OWEN, M. J. & O'DONOVAN, M. C. 2011. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Mol Psychiatry,* 16**,** 429-41.

WOOLFE, A., GOODE, D. K., COOKE, J., CALLAWAY, H., SMITH, S., SNELL, P., MCEWEN, G. K. & ELGAR, G. 2007. CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC developmental biology,* 7**,** 100.

WOOLFE, A., GOODSON, M., GOODE, D. K., SNELL, P., MCEWEN, G. K., VAVOURI, T., SMITH, S. F., NORTH, P., CALLAWAY, H., KELLY, K., WALTER, K., ABNIZOVA, I., GILKS, W., EDWARDS, Y. J., COOKE, J. E. & ELGAR, G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol,* 3**,** e7.

YEO, G. S. 2014. The role of the FTO (Fat Mass and Obesity Related) locus in regulating body size and composition. *Mol Cell Endocrinol*.

ZARET, K. S. & CARROLL, J. S. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes & development,* 25**,** 2227-2241.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational biology,* 7**,** 203-214.

ZIMMERMANN, E., KRING, S. I., BERENTZEN, T. L., HOLST, C., PERS, T. H., HANSEN, T., PEDERSEN, O., SØRENSEN, T. I. & JESS, T. 2009. Fatness-associated FTO gene variant increases mortality independent of fatness–in cohorts of Danish men. *PLoS One,* 4**,** e4428.

ZIMMERMANN, E., SKOGSTRAND, K., HOUGAARD, D. M., ASTRUP, A., HANSEN, T., PEDERSEN, O., SORENSEN, T. I. & JESS, T. 2011. Influences of the common FTO rs9939609 variant on inflammatory markers throughout a broad range of body mass index. *PLoS One,* 6**,** e15958.

ZINZEN, R. P., GIRARDOT, C., GAGNEUR, J., BRAUN, M. & FURLONG, E. E. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature,* 462**,** 65.

ZUCCHERO, T. M. 2004. Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *N. Engl. J. Med.,* 351**,** 769-780.