A FUNCTIONAL CHARACTERISATION OF THE PCSK6 LOCUS ASSOCIATED WITH HANDEDNESS

Robert Shore

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2016

A functional characterisation of the PCSK6 locus associated with handedness

Robert Shore

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy

at the University of St Andrews

Date of Submission: December 1 2015

1. Candidate's declarations:

I Robert Shore hereby certify that this thesis, which is approximately 40,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in August 2012 and as a candidate for the degree of PhD in August 2012; the higher study for which this is a record was carried out in the University of St Andrews between 2012 and 2015.

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date …… signature of supervisor ………

3. Permission for publication: (*to be signed by both candidate and supervisor*)

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY

No embargo on print copy

ELECTRONIC COPY

Embargo on all or part of electronic copy for a period of 2 years (maximum five) on the following ground(s): Publication would preclude future publication

Date …… signature of candidate …… signature of supervisor ………

Please note initial embargos can be requested for a maximum of five years. An embargo on a thesis submitted to the Faculty of Science or Medicine is rarely granted for more than two years in the first instance, without good justification. The Library will not lift an embargo before confirming with the student and supervisor that they do not intend to request a continuation. In the absence of an agreed response from both student and supervisor, the Head of School will be consulted. Please note that the total period of an embargo, including any continuation, is not expected to exceed ten years. Where part of a thesis is to be embargoed, please specify the part and the reason.

Dedicated to the memory of my father, Joseph Shore (1949-2013)

Ar dheis Dé go raibh a anam

Abstract

Humans display a 90% population level bias towards right-handedness, implying the vast majority of people have a left-hemisphere dominant for motor control. Although handedness presents a weak, but very consistent heritability across the literature (estimated to be approximately 25%), to date few genetic loci associated with this complex trait have been identified and replicated in subsequent studies. One such gene which has been found to be associated with handedness and subsequently replicated is PCSK6, most recently through a quantitative GWAS ($P < 0.5*10^{-8}$, Brandler *et al.* (2013)). Interestingly, PCSK6 is known to activate Nodal, a morphogen involved in a highly conserved bilaterian pathway known to regulate left-right body axis determination.

Here I present the first molecular characterisation of a handedness-associated region by conducting a detailed functional analysis of the PCSK6 locus, combining genetic analysis, *in silico* prediction and molecular assays to investigate how common genetic variants influence handedness-related phenotypes. Specifically, I defined the associated locus to be 12.7 kb in size, spanning a predicted 1.8 kb bidirectional promoter which controls the expression of both an antisense long non-coding RNA (lncRNA), and a novel short PCSK6 isoform. A series of luciferase-expressing constructs were generated to characterise the promoter, identifying a minimal sequence capable of driving transcription in a sense strand direction. I have demonstrated experimentally that one of the top associated markers in previous GWA studies, rs11855145, directly creates/disrupts a suspected transcription factor bind site in the vicinity of this bidirectional promoter.

Further functional studies of the genetic variation within PCSK6 may help explain the molecular regulatory mechanisms affecting gene expression. This project provides a model for assays to study other GWAS-nominated candidate genes, and in particular for establishing the role of noncoding variants. The findings from this study support the role of common variants in influencing complex phenotypes, such as handedness.
Thesis word count: 40,000 words

Acknowledgements

I would like to thank my supervisor Dr Silvia Paracchini for giving me not only the opportunity to come to St Andrews and complete a PhD in a topic I found enjoyable and intriguing in equal measure, but also for the advice and support you provided and the independence you allowed when required. For that I will always be grateful.

I am indebted to my thesis committee; Professor Lindsey Kent and Professor Terry Smith who helped with their encouragement, ideas and patience; I am in no doubt how compromised my endeavours would have been without your help. A sincere appreciation also goes to Dr Catherine Botting for her mass spectrometry services, Dr Kerry Pettigrew for her toils in helping define the clinical cohorts and my internal and external examiners Dr Matt Holden and Dr Nick Bray for giving up so much of their time to improve this thesis.

Finally, I would like to thank my parents, my family and my friends Liam Scally and Aleksandra Sobota for their infuriatingly persistent refusal to ever let me give up.

Table of Contents

Abstract **iv**

Acknowledgements **v**

Table of Contents **vi**

List of Figures **xi**

List of Tables **xiv**

Abbreviations **xvi**

1 Introduction **1**

1.1 Handedness 1

1.1.1 What is handedness? 1

1.1.2 Limb preference in the Animal Kingdom 2

1.1.3 Handedness in non-human primates 5

1.1.4 Right-hand dominance in the hominin lineage 6

1.1.5 How did handedness evolve? 7

1.1.6 Handedness and Language 8

1.1.7 Why study handedness? 10

1.1.8 The development of handedness 12

1.1.9 Genetic models of handedness 13

1.1.10 Previous studies and candidate genes 14

1.2 PCSK6 22

1.2.1 The PCSK6 gene 22

1.2.2 PCSK6 role in complex traits and disease 23

1.2.3 Nodal signalling pathway and cilia 24

1.2.4 RD candidate genes and cilia 27

1.2.5 Role of regulatory elements during neuronal development 27

1.2.5.1 Non-coding RNA 28

1.2.5.2 VNTRs 29

1.2.5.3 Bidirectional promoters 31

1.3 Conclusion 32

1.3.1 Overall aim of this thesis 32

1.3.2 Overview of the study methodology 33

2 ALSPAC Cohort Association Analysis 35

2.1 Abstract 35

2.2 Introduction 36

2.3 Methods 38

2.3.1 Ascertainment criteria for subgroups 38

2.3.2 Laterality measures 41

2.3.2.1 Quantitative measures of hand performance 41

2.3.2.2 Categorical measures of handedness and laterality 43

2.4 Results 44

2.4.1 ALSPAC quantative measures 44

2.4.2 Categorical measures for handedness 50

2.4.3 Correlation between quantative measures 50

2.4.4 Ambidexterity 53

2.4.5 Hand, Foot and Eye Correlation 53

2.4.6 Association analysis of rs11855415 55

2.5 Discussion 58

3 *In silico* analysis of the PCSK6 locus 64

3.1 Abstract 64

3.2 Introduction 64

3.3 Methods 67

3.3.1 Defining the region of Linkage Disequilibrium 67

3.3.2 Interrogation of the TFBS databases 67

3.3.3 Software prediction of the secondary promoter region 68

3.3.4 Identifying evolutionarily conserved sequences 68

3.4 Results 69

3.4.1 Region of Linkage Disequilibrium 69

3.4.2 Allelic effect on transcription factor binding affinity 73

3.4.3 Predicted region of secondary promoter 77

3.4.4 Evolutionary conserved sequences 79

3.5 Discussion 81

4 Functional elements at the PCSK6 locus associated with handedness 85

4.1 Abstract 85

4.2 Introduction 86

4.3 Methods 92

4.3.1 PCSK6 isoform profiling 92

4.3.2 PCSK6-AS annotation 92

4.3.3 RNA-seq analysis 93

4.3.4 PCSK6 isoform expression in the developing and adult human brain 94

4.4 Results 95

4.4.1 PCSK6 isoform profiling 95

4.4.2 Annotating the PCSK6 SI 98

4.4.3 PCSK6 SI protein prediction 101

4.4.4 RNA-seq and microarray analysis of the PCSK6 SI 103

4.4.5 RNA-seq analysis of PCSK6-AS 114

4.4.6 Annotation of the PCSK6-AS isoforms 115

4.5 Discussion 119

5 Functional Analysis of the PCSK6 Genetic Variants 123

5.1 Abstract 123

5.2 Introduction 124

5.3 Methods 127

5.3.1 PCSK6-AS Knockdown 127

5.3.2 PCSK6-AS Overexpression 129

5.3.3 Electrophoretic Mobility Shift Assay (EMSA) 130

5.3.4 Rev-ChIP 132

5.3.5 Luciferase Reporter Assay 135

5.4 Results 137

5.4.1 Knockdown of the PCSK6-AS lncRNA 137

5.4.2 Overexpression of the PCSK6-AS lncRNA 139

5.4.3 EMSA 140

5.4.4 Rev-ChIP 141

5.4.5 Luciferase assays 147

5.5 Discussion 156

6 Concluding remarks & future perspectives 161

6.1 Summary of findings 161

6.1.1 A proposed model derived from the findings of this thesis 164

6.2 Future Perspectives 166

6.3 Concluding remarks 170

APPENDICIES 173

Appendix A Primer Catalogue 173

A.1 PCSK6 isoform profiling 173

A.2 PCR primers 174

A.3 qPCR primers 175

A.4 Overexpression & Knockdown primers 176

A.5 Luciferase assay primers 177

Appendix B Supplementary Materials and Methods 179

Appendix C Isoform Analysis 181

C.1 PCSK6-AS2/AS3 cDNA 181

C.2 PCSK6-AS1 vs PCSK6-AS2 182

C.3 PCSK6 SI cDNA.. 183

C.4 PCSK6 protein sequences.. 184

Appendix D EMSA 188

D.1 Oligonucleotide Sequences.. 188

D.2 Additional EMSA images.. 189

Appendix E Luciferase Assay 191

E.1 Lipofectamine P3000 transfection protocol.. 191

E.2 Luciferase construct table 191

E.3 Luciferase assay plasmids... 192

E.4 Luciferase assay antisense strand results.. 194

E.5 Sequencing data for mutagenesis of rs11855415 TFBS 195

Appendix F RNA-Seq 197

F.1 Source and shell script developed to download RNA-Seq files....................... 197

F.2 Script for creating BAM files 198

F.3 Script for alignment with SAMtools... 199

Appendix G Rev-ChIP Supplemental Materials and Methods 200

Appendix H Association analysis 209

Appendix I Genetic variant genotyping 212

References 214

List of Figures

Figure 1.1 Broca's area 10

Figure 1.2 Graphical summary of candidate loci for hand skill or hand preference from previous studies. 18

Figure 1.3 PCSK6 gene expression in adult human tissue. 23

Figure 1.4 The prominent role of PCSK6 in the developing embryo NODAL pathway 25

Figure 1.5 Scheme indicating experimental design flow of the research project. 34

Figure 2.1 Diagram illustrating how phenotypic subgroups were identified from the ALSPAC dataset. 40

Figure 2.2 PegQ37 laterality measure distribution according to subgroup. 45

Figure 2.3 PegQ7 laterality measure distribution according to subgroup. 46

Figure 2.4 SortQ laterality measure distribution according to subgroup. 47

Figure 2.5 MarkQ laterality measure distribution according to subgroup. 48

Figure 2.6 GripQ laterality measure distribution according to subgroup. 49

Figure 2.7 MarkQ distribution in the Unaffected subgroup. 60

Figure 3.1 LD structure of the HapMap North and Western European population (CEPH) at the PCSK6 gene. 71

Figure 3.2 PCSK6 locus associated with relative hand skills 72

Figure 3.3 UCSC genome browser tracks indicating TFBSs as assayed by ChIP-seq at the PCSK6 locus of interest 76

Figure 3.4 Visualisation of the TSSs annotated in the DBTSS database on both sense and antisense strands within PCSK6 intron 14 78

Figure 3.5 Sequence conservation across a PCSK6 region associated with handedness. 80

Figure 4.1 The human corpus callosum. 89

Figure 4.2 PCSK6 isoforms tested for expression in multiple cell lines 96

Figure 4.3 PCSK6 expression across a range of cell lines according to the Human Protein Atlas (HPA). 98

Figure 4.4 Annotating the shorter PCSK6 isoform. 100

Figure 4.5 The predicted protein domains of the PCSK6 isoforms. 101

Figure 4.6 RNA-seq probes for PCSK6 regions of interest. 104

Figure 4.7 PCSK6 gene expression in the developing human embryonic brain. 108

Figure 4.8 PCSK6 expression in the developing human brain 110

Figure 4.9 PCSK6 isoform expression in the adult human brain. 112

Figure 4.10 Alternative pre-mRNA splicing of PCSK6 in various tissues. 113

Figure 4.11 PCSK6-AS1 expression across a range of tissues. 114

Figure 4.12 PCSK6-AS1 expression across a range of cell lines. 115

Figure 4.13 UCSC Genome browser image of PCSK6-AS1 and two newly discovered isoforms 116

Figure 4.14 Expression of the PCSK6 gene **(A)** and antisense lncRNA gene PCSK6-AS1 **(B)** according to the ENCODE RNA-seq dataset 118

Figure 5.1 UCSC Genome browser view of the SINE Alu element indicates overlap with the third exon of the PCSK6-AS. 124

Figure 5.2 Reverse ChIP protocol summary. 135

Figure 5.3 Location of the PCSK6 primers used for qPCR quantification. 137

Figure 5.4 siRNA knockdown of the PCSK6-AS lncRNA. 138

Figure 5.5 Overexpression of the PCSK6-AS lncRNA. 139

Figure 5.6 Overexpression of the PCSK6-AS lncRNA. 140

Figure 5.7 EMSA gel images for protein:DNA binding at the rs11855415 SNP 141

Figure 5.8 The distribution of hNSC and 1321N1 nuclear proteins identified by mass spectrometry and arranged according to molecular function. 146

Figure 5.9 Dual luciferase assay results testing for allelic effects on promoter activity. 148

Figure 5.10 Dual luciferase assay results testing for allelic effects on promoter activity. 149

Figure 5.11 Bidirectional activity of the PCSK6 secondary promoter across multiple cell lines 150

Figure 5.12 Luciferase minimal promoter analysis. 151

Figure 5.13 Dual luciferase assay results testing for allelic effects on promoter activity. 154

Figure 5.14 Luciferase-expressing constructs testing the effect of removing the TFBS centred on rs11855415. 155

Figure 6.1 A proposed model derived from thesis findings. 164

Figure C.1 cDNA sequence of the PCSK6 SI 185

Figure D.1 EMSA for SH-SY5Y nuclear extract and the rs11855415 alleles. 189

Figure D.2 EMSA for HEK293 nuclear extract and the rs11855415 alleles. 190

Figure E.1 Plasmid map of the pCR-BluntII-TOPO entry vector 192

Figure E.2 Plasmid map of the pGL4.10 luciferase reporter vector 192

Figure E.3 Plasmid map of the pGL4.10 luciferase reporter 10A vector 193

Figure E.4 Minimal Promoter construct used for allelic variation assays. 193

Figure E.5 Antisense luciferase-expressing construct results 194

Figure E.6 Chromatogram confirming the sequence of the 10A plasmid 195

Figure E.7 Chromatogram confirming the sequence of the 10A plasmid following mutagenesis 196

Figure H.1 Plot of the normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals from the Unaffected subgroup 209

Figure H.2 Plot of normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals of the RD subgroup 210

Figure H.3 Plot of normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals of the Affected subgroup. 211

Figure I.1 Genotyping of cell lines for the VNTR identified in the secondary promoter of PCSK6. 212

List of Tables

Table 1.1 Selection of animal species showing population-level asymmetry where limb preference has been investigated 4

Table 1.2 Previous genetics studies investigating hand preference and relative hand skill and the candidate loci implicated 16

Table 1.3 Enrichment of function among human genes containing VNTRs................. 30

Table 2.1 The final subgroup sizes following filtering of individuals from the ALSPAC cohort according to disorder... 41

Table 2.2 A summary of the skewness and kurtosis of the PegQ37 scores 45

Table 2.3 A summary of the skewness and kurtosis of the PegQ7 scores 46

Table 2.4 A summary of the skewness and kurtosis of the SortQ scores...................... 47

Table 2.5 A summary of the skewness and kurtosis of the MarkQ scores.................... 48

Table 2.6 A summary of the skewness and kurtosis of the GripQ scores 49

Table 2.7 Pearson's correlation between categorical measures for self-reported handedness.. 50

Table 2.8 Pearson's correlation between quantitative measures of relative hand skill and dexterity in the ALSPAC dataset subgroups.. 52

Table 2.9 Ambidexterity recorded in the quantitative measures for hand performance.53

Table 2.10 Pearson's correlation between categorical measures for self-reported handedness, footedness and eyedness 54

Table 2.11 Association analysis results for the PCSK6 genetic variant rs11855415 and hand performance measures for relative hand skill and dexterity in multiple subgroups. 56

Table 2.12 Association analysis results for the PCSK6 genetic variant rs11855415 and the PegQ7hand performance measure for relative hand skill in the Affected subgroup. 57

Table 2.13 Association analysis results for the PCSK6 genetic variant rs11855415 and categorical measures for handedness, footedness and eyedness in multiple subgroups. 58

Table 3.1 *In silico* TFBS prediction of the PCSK6 locus associated with handedness. 74

Table 3.2 PCSK6 secondary promoter prediction.. 78

Table 4.1 PCSK6 canonical and spliced isoforms according to RefSeq...................... 90

Table 4.2 Properties of the canonical PCSK6 protein and the predicted shorter PCSK6 isoform 103

Table 4.3 RNA-seq probes for the PCSK6 regions of interest. 104

Table 4.4 RNA-seq raw reads count for multiple PCSK6 exons. 106

Table 5.1 Binding preference of proteins to both rs11855415 alleles according to Rev-CHiP mass spectrometry output. 144

Table B.1 Indicates linkage disequilibrium (LD) between the SNP rs11855415 and all available SNPs within the defined PCSK6 locus 179

Table B.2 Mammalian genome assemblies as used for multi species alignment of 14 eutherian mammalian genomes 180

Table C.1 Annotation of both PCSK6-AS isoforms according to the AnnoLnc service 182

Table E.1 Luciferase assay reporter plasmids. 191

Table G.1 Full mass spectrometry output for Rev-ChIP assay rs1185545 A vs T alleles using hNSC cell line nuclear extract..201

Table G.2 Full mass spectrometry output for Rev-ChIP assay rs1185545 A vs T alleles using 1321N1 cell line nuclear extract...204

Table I.1 Genotyping of cell lines for the rs11855415 SNP and the VNTR identified in the secondary promoter of PCSK6 213

Abbreviations

ADHD	attention-deficit hyperactivity disorder
ALSPAC	Avon longitudinal study of parents and children
BF	basal forebrain
CC	corpus callosum
CCC	children’s communication checklist
cDNA	complimentary deoxyribonucleic acid
ChIP	chromatin immunoprecipitation
CNS	central nervous system
CNS	conserved non-coding sequence
C_t	cycle at threshold
DAWBA	development and well-being assessment
DSM	diagnostic and statistical manual of mental disorders
ECACC	European collection of cell cultures
EMSA	electrophoretic mobility shift assay
eQTL	expression quantitative trait loci
ER	endoplasmic-reticulum
ESS	evolutionary stable strategy
fMRI	functional magnetic resonance imaging
GWAS	genome wide association study
HKG	housekeeping gene
IQ	intelligence quotient
KBP	kilo base pair
LC/ESI	liquid chromatography and electrospray ionisation
LD	linkage disequilibrium
lncRNA	long non-coding RNA
MAF	minor allele frequency
MALDI TOF	matrix-assisted laser desorption/ionisation time-of-flight
MCS	multiple-species conserved sequence
MY	medulla oblongata
MES	mesencephalon

MET	metencephalon
MLRA	multiple linear regression analysis
MS	mass spectrometry
NAT	natural antisense transcript
ncRNA	non-coding RNA
NEB	New England Biolabs
NMR	nuclear magnetic resonance
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PCWs	post-conception weeks
PFWs	post-fertilisation weeks
qPCR	quantitative polymerase chain reaction
qRT-PCR	quantitative reverse transcriptase polymerase chain reaction
QTL	quantitative trait loci
RD	reading disability (or dyslexia)
Rev-ChIP	reverse chromatin immunoprecipitation
RNA-seq	RNA sequencing
RT	room temperature
SD	standard deviation
SILAC	stable isotope labelling by amino acids in cell culture
siRNA	short interfering RNA
SLI	specific language impairment
SNP	single nucleotide polymorphism
TFBS	transcription factor binding site
TH	thalamus
TSS	transcription start site
UTR	untranslated region
VNTR	variable number of tandem repeats
WOLD	Wechsler objective language dimensions
WHQ	Waterloo handedness questionnaire

1 Introduction

1.1 Handedness

1.1.1 What is handedness?

Handedness is a behavioural asymmetry in which dominant use is lateralised to one hand or the other. In humans this laterality of function is displayed in approximately 90% of the population who show a right hand dominant for skilled manipulation and motor functionality (Corballis, 2003). Functional lateralisation exists in multiple paired organs across the sagittal plane of the human body; our liver and our guts coil in opposite directions, the left kidney is slightly higher and larger than the right, the pancreas and spleen are asymmetrically placed right and left of the midline respectively and even the organs themselves can be asymmetrical; our heart must be larger on one side if that side is to passage blood around the entire body. The brain is no exception, exhibiting both functional asymmetries and anatomical asymmetries, e.g. the right hemisphere is wider in the anterior region while the left hemisphere is wider in the posterior region.

Handedness is a heritable continuum ranging from extreme right to extreme left-handedness and although substantial evidence exists to support a consistent right hand bias for multiple tasks in different societies (Raymond *et al.*, 1996, Perelle and Ehrman, 1994), the bias has been shown to fluctuate across both time and space. Stock *et al.* (2013), demonstrated this change in a study in which skeletal analysis was used to report an 80% right-lateralisation bias in a British medieval population. Additionally, Faurie and Raymond (2005) investigated functional specialisation in traditional societies and concluded that the prevalence of left-handers reported fluctuates between 3 and 27%, a figure which varies in Western societies between 2 and 13%. Gender has also been shown to influence the incidence of handedness. In a meta-analysis of 144 studies (N = 1,787,629) Papadatou-Pastou *et al.* (2008) demonstrated that the sex difference is both significant and robust with males 23% more likely to be left-handed than females. In short, it remains unclear whether the preference for one hand over the other is learned and influenced by environment, experience and enforcement (Provins, 1997) or if handedness is innate and genetically controlled (Corballis, 1997), or a combination of the above. Furthermore, why a stable, albeit fluctuating, 10% of the population is non-right-handed has yet to be fully explained. Yahagi and Kasai (1999) demonstrated in a study involving visual skills not subject to social control that left-handers show less functional asymmetry than right-handers (N = 24), that is, the degree to which left handers use their preferred hand is less than that of right-handers. Therefore, differences in performance between the two hands are significantly smaller for left-handers, resulting in an advantage in the use of the non-preferred hand. Such an advantage could maintain a stable ratio of non-right-handers in a population as a result of an evolutionarily stable strategy (ESS) (Ghirlanda and Vallortigara, 2004).

1.1.2 Limb preference in the Animal Kingdom

If handedness does represent an ESS then one might expect to see its repeated evolution in multiple lineages. Until quite recently humans were considered to be the only species that exhibited such a population-level bias in limb preference direction (Hopkins *et al.*, 2011) however mounting empirical evidence from a variety of vertebrate and invertebrate species has begun to challenge this assumption (Rogers *et al.*, 2013).

Ströckens *et al.* (2013) recently reviewed the lateralised behaviours of non-human vertebrates and found 68% of the 119 species reviewed show lateral bias, 51% of which are at a population-level (Table 1.1). One of the most well-known examples of forelimb asymmetry in non-human vertebrates are parrots which show a left foot bias of almost 90% in picking up objects (Rogers, 2008). As with humans however, this distribution might be context specific; parrots, like chicks, use their feet to manipulate or scratch for food whereas pigeons do not. In a study which involved scratching off adhesive tape placed on a bill, Güntürkün (1988) found no foot preference in pigeons for removing the tape while Rogers and Workman (1993) found 82% of chicks favoured the right foot, a result supported by Andrew *et al.* (2000) who demonstrated chicks have a right bias for discriminating food from non-food in a foraging task. Together these results suggest limb preference in birds could arise in species that use their forelimbs for manipulation.

Other vertebrate species also exhibit a dominant forelimb preference, though in both cats and mice this does not seem biased to either the right or left paw at a population level (Fabre-Thorpe *et al*., 1993, Bulman-Fleming *et al*., 1997). Interestingly, different mouse strains exhibit differences in the direction and strength of paw preference, indicating that genetic background has a contributory effect on this behaviour (Waters and Denenberg, 1994).

Non-human vertebrates have also been observed to display asymmetric behaviours, overt or otherwise; for example several species of fish display population-level biases, that is, the majority of a population will consistently turn to the left or the right to avoid predation (Bisazza *et al*., 2000), an analogous trait reflected in humans where the right hemisphere is specialised for avoidance and the left tends to be specialised for approach (Davidson, 2004). A left-hemisphere dominance for vocal production appears in both canaries (Halle *et al*., 2003) and frogs (Bauer, 1993) while mice have shown a left-hemispheric advantage for the recognition of ultrasonic communication calls (Ehret, 1987). As noted by Corballis (2009), this left hemisphere control of vocalisation production and perception might be a precursor to left hemisphere dominance for language processing in humans.

Table 1.1 Selection of animal species showing population-level asymmetry where limb preference has been investigated

			Preferred side %			
species	*Measure*	*N*	*left*	*right*	*none*	*citation*
spitting spiders	limb loss	36	75	25	-	(Ades and Ramires, 2002)
Eurasian curlew	roosting on one foot	310	45	55	-	(Randler, 2007)
red kangaroo	bipedal feeding	21	86	4	10	(Giljov *et al.*, 2015)
ostrich	forward foot resting posture	65	22	68	10	(Baciadonna *et al.*, 2010)
chicks	foot preference	50	6	84	10	(Casey and Martino, 2000)
buzzard	foot to grasp prey	34	32	53	15	(Csermely, 2004)
cane toads	righting behaviour	42	10	90	-	(Robins and Rogers, 2002)
leatherback turtles	Flipperedness	361	46	54	-	(Sieg *et al.*, 2010)
tortoises	righting behaviour	34	18	53	29	(Stancher *et al.*, 2006)
wallabies	bipedal feeding	27	74	7	19	(Giljov *et al.*, 2013)
bottlenose dolphins	flipper rubbing	111	45	-	55	(Sakai *et al.*, 2006)
bats	Climbing	25	24	76	-	(Zucca *et al.*, 2010)
lions	forelimb stand preference	24	21	75	4	(Zucca *et al.*, 2011)
rats	paw preference	198	20	73	7	(Guven *et al.*, 2003)
whales	rolling feeding behaviour	11	10	90	-	(Canning *et al.*, 2011)
gorilla	hand preference (various)	76	22	54	24	(Hopkins *et al.*, 2011)
lemurs	food reaching	194	47	33	20	(Ward *et al.*, 1990)
spider monkey	food reaching	13	79	21	-	(Laska, 1996)
olive baboon	communicative gestures	60	17	58	25	(Meguerditchian and Vauclair, 2006)
chimpanzees	hand preference (various)	536	29	50	21	(Hopkins *et al.*, 2011)
Japanese macaques	hand preference (various)	394	38	30	32	(Itani *et al.*, 1963)

- dash indicates where no value was recorded or appropriate

Asymmetry at the population-level can also occur among invertebrates, supporting growing evidence that the specialisation of the brain's hemispheres was already in place prior to the arrival of the vertebrates half a billion years ago (MacNeilage *et al.*, 2009). The *Bombus terrestris* honeybee favours the right antenna in responding to learned odours (Anfora *et al.*, 2011), a similar population-level lateralisation to the recall of olfactory memory found in honeybees and several species of Australian stingless bees (Frasnelli *et al.*, 2012). The study of invertebrate asymmetry is also important if we are to understand the role of environment on how limb asymmetries develop (Palmer, 2012). Both the water bug *Belostoma flumineum* (Kight *et al.*, 2008) and *Temnothorax albipennis* ant (Hunt *et al.*, 2014) display a population-level left turn bias, a possible result of a lateralised central nervous system (CNS), while the spectacularly asymmetric forelimbs of some crustaceans such as the fiddler crab (Backwell *et al*., 2007) show how significant behaviour can be in prompting and orienting morphological, and consequently functional, asymmetries (Versace and Vallortigara, 2015).

Collectively, these studies provide important insights into the functional and anatomical asymmetries that exist across a wide range of taxa and species. However due to the phylogenetic distances involved and the considerable number of studies that report a lack of functional asymmetry at a population level (Hook, 2004), it is difficult to say which if any of these asymmetries could be precursors to handedness in humans. To do so requires a more detailed look at our closest relatives.

1.1.3 Handedness in non-human primates

If the laterality of hand function can be shown to exist in non-human primates (chimpanzees in particular) it could make them a useful model in offering insights into the emergence of human handedness. However, the expression of manual laterality in non-human primates at a population level and their potential continuities with Homo sapiens remains less than conclusive. Drawing on an extensive range of sources, Meguerditchian *et al*., (2013) reviewed the collective findings and concluded that for bimanual behaviours, that is, the engagement of two hands in a coordinated and asymmetrical manner, right hand dominance for manipulation seems to extend to all terrestrial primates, but not to arboreal species. The frequency of right-hand bias is consistently lower than that found in humans and in a meta-analysis of 1,524 wild and captive great apes Hopkins (2006) did find right-hand bias at the population level, though never exceeding 55% bias for all samples.

Though some studies report no population-level asymmetries in the nonhuman primates (McGrew and Marchant, 1997, Papademetriou *et al*., 2005, Cashmore *et al*., 2008), behavioural observations for forelimb asymmetries have been shown with a similar right-hand bias for bimanual behaviours in baboons (Vauclair *et al*., 2005), chimpanzees and gorillas (Llorente *et al*., 2011), and in adult bonobos (Chapelain *et al*., 2011). To confound matters, most data are derived from great apes kept in captivity, with only two studies investigating bimanual coordinated behaviours in wild groups (chimpanzees: Corp and Byrne (2004) and gorillas: Byrne and Byrne (1991)). This poses a problem since evidence of population-level hand bias in wild primates is rare (Parnell, 2001, Byrne and Byrne, 1991). The discrepancy in findings between captive and wild apes, has prompted some to argue that captive apes may have been influenced by human handlers in a predominantly right-handed environment (Palmer, 2002) or that the limited sample sizes of wild-population studies are denied the statistical power larger captive cohorts do not lack (Hopkins and Cantalupo, 2005).

In conclusion, where the literature shows a hand preference in great apes, it does so at a reduced ratio and not the consistent 9:1 we see at the population level in humans. The high level of right-hand dominance displayed by humans should therefore be considered a unique, derived trait (Gibson, 1993). By implication, the inconsistent population-level handedness seen in our nearest relative, the chimpanzee, suggests that the trait is derived in *Homo sapiens* and that population-level right-handedness must have emerged sometime after the divergence from our last common ancestor approximately 5.4 million years ago (Patterson *et al*., 2006). The search for the emergence of handedness must therefore focus on hominin species prior to *Homo sapiens*.

1.1.4 Right-hand dominance in the hominin lineage

Paleoanthropological research in to the handedness of ancient humans employs a range of excavated data including skeletal elements, prehistoric art, tool-manufacturing techniques and directional cut marks on faunal material for inferring hand preference (Frayer *et al*., 2012). For example, by comparing Upper Palaeolithic (ca. 50,000-10,000 years ago) hand stencils found on the walls of French caves to modern students' technique for producing similar representations, Faurie and Raymond (2004) found an identical ratio of at least 77% right-handedness in both groups. Further back in the hominin lineage, an analysis of skeletal asymmetry (e.g. rigidity of the second metacarpal) by Stock *et al*. (2013) was used to reflect habitual mechanical loading and infer hand dominance among thirteen hunter-gatherers populations as far back as the late Pleistocene age (ca. 126,000 years ago), confirming a 62.5% right bias.

Electron microscopy provides a means of detecting enamel scratches that occur on fossilised teeth as a result of tool manipulation performed at the front of the mouth (Lozano-Ruiz *et al.*, 2004). Using this technique, Frayer *et al.* (2012) were able to infer a left hand preference of 12% for a group of 17 *Homo neanderthalensis* specimens 30,000-130,000 years old. This conclusion is supported by Uomini (2011) whose comprehensive analysis of the fossil and archaeological records suggested 8–20% of *Homo neanderthalensis* to be left-handed. Ancestral to *Homo neanderthalensis*, the same study indicated that by the Middle Pleistocene age (ca. 781,000 to 126,000 years ago) there was already a strong bias towards right hand dominance. Finally, by analysing the wear and tear of cleavers and hand axes dating from roughly one million years ago, Phillipson (1997) was able to make inferences on ancient human handedness by reconstructing grip types, observing a majority of right hand users in the process. Taken together, these studies help establish that by half a million years ago significant lateralisation had occurred in the hominin lineage (McGrew and Marchant, 1997), though most behavioural scientists are in general agreement that right-handedness could have evolved even earlier, approximately 2.5 million years ago.

1.1.5 How did handedness evolve?

Several theories exist to explain how handedness may have evolved in humans since the divergence from our last common ancestor. Though now largely discounted, Corballis (1997) proposed a genetic mutation unique to the evolution of *Homo sapiens* approximately 150,000 - 200,000 years ago prompted a stable right hand bias. This theory fits the genetic data and offered a reasonable explanation for the stable 9:1 bias ratio found in humans but doesn't explain the mounting evidence for population-level asymmetries in other species. One theory posited by Rogers and Andrew (2002) does take this in to account and proposes that 500 million years ago lateralisation for perceptual and motor control of feeding emerged in the earliest vertebrates, organisms whose mouth was located on the head's left side. In support of this theory, motor control by the left hemisphere of right side organs/forelimbs has been identified in all classes of vertebrate from fish (Bisazza *et al.*, 1998) to non-human primates (Hopkins, 2006).

The task complexity hypothesis (Fagot and Vauclair, 1991, Bradshaw and Rogers, 1993) proposes that demanding bimanual behaviours enhance a stronger degree of handedness than simple unimanual tasks, as observed in Australian parrots that show a left foot bias of up to 90% (Brown and Magat, 2011) when using coordinated foot-beak actions, a bias not recorded in birds that eat seeds with beak alone. In addition, Forrester *et al.* (2013) found that chimpanzees, gorillas and children all show a stronger right bias in response to inanimate objects but not to animate objects, suggesting any right hand bias in primates is a result of a left-hemisphere that originally specialised for tool use.

Another theory that looks beyond the order of primates suggests that a species posture can influence forelimb bias. The postural origin theory (MacNeilage *et al.*, 1987) proposes bipedalism facilitates bimanual object manipulation and thereafter manual laterality, with a positive correlation between the degree of upright posture and forelimb laterality as seen in non-human primates as well as marsupials (Hopkins *et al.*, 2011); with the more quadrupedal species less likely to express manual lateralisation than their bipedal counterparts (Giljov *et al.*, 2015).

Ultimately, handedness is a behavioural laterality whose control is dominated by one hemisphere, a feature shared by another functional lateralisation, and a uniquely human one, language.

1.1.6 Handedness and Language

The majority of both left and right-handed individuals possess a left-hemisphere dominant for language comprehension and production (Rasmussen and Milner, 1977), though right-handers (88%) are more likely to demonstrate a left hemisphere dominance than left-handers (78%) (Mazoyer *et al.*, 2014). For the remainder of the population, language lateralisation is specific to the right-hemisphere, or is distributed symmetrically across both hemispheres (Knecht *et al*., 2000).

This left hemisphere dominance for both language and handedness in the majority of the population previously led some to suggest handedness and language coevolved as uniquely human traits (Corballis, 2003) however the ever-increasing evidence for population-level forelimb asymmetries in phylogenetically distant species has since led to a general dismissal of this hypothesis. Instead, a more widely held theory posits that a human left-hemisphere dominance for both language and handedness is an outcome of the lateralisation of motor functionality that arose earlier in human evolution (Meguerditchian *et al*., 2013). If a gestural language was the precursor to speech then the circuits controlling this gesturing hand may have developed over time so as to take control of the neuronal circuits involved in language (Thomas, 2006, Pollick and de Waal, 2007). Such an integration between spoken language and communicative gestures has been recorded in the activation of different brain areas (Willems *et al*., 2007, Andric *et al*., 2013) with gestures, such as gestural hand movement, found to be asymmetrical in children (Trevarthen, 1996) as well as nonhuman primates (Meunier *et al*., 2013, Hopkins *et al*., 2003). To this effect, a well-known region in the frontal lobe of the human brain associated with speech production, Broca's area (Figure 1.1), is in essence a premotor module which coordinates muscle contraction patterns that are related to other functions besides language (Rorden *et al*., 2008). Interestingly, Broca's area has a distinctive degree of anatomical asymmetry, a lateralisation also found in the analogous region of the chimpanzee brain, even though it does not yet correspond to language capability (Cantalupo and Hopkins, 2001).

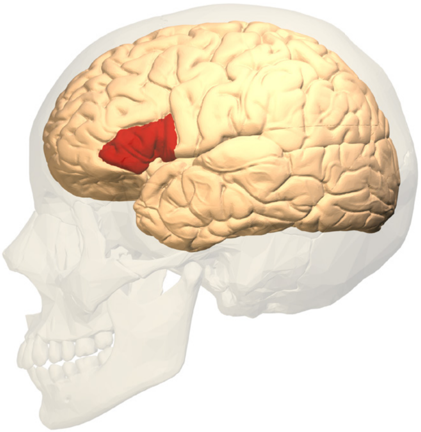


Figure 1.1 Broca's area marked in red is a region in the human frontal lobe of the dominant hemisphere with functions linked to speech production

1.1.7 Why study handedness?

Understanding how right-hand dominance evolved and is sustained at a population level is still a complex challenge, though hand bias may have offered an evolutionary advantage by allowing for non-redundant cerebral functionality and enlarged brain capacity through the increased speed of unihemispheric processing (Ocklenburg and Gunturkun, 2012).

The shared left hemisphere dominance for both motor control and language in the majority of the population has led some researchers over the years (Eglinton and Annett, 1994, Hynd *et al*., 1990, Tonnessen *et al*., 1993) to posit a possible link between hand preference and disorders that affect language development such as dyslexia, a relatively common disorder with approximately 5-10% of children affected (Pennington and Bishop, 2009). Although no consistent association has been found between hand preference and dyslexia (Francks *et al*., 2003c, Brandler *et al*., 2013) evidence exists to suggest there may be atypical cerebral asymmetry in dyslexic individuals (for a review see Richlan *et al*. (2009) and Friederici (2006)). In a series of eight consecutive post-mortem studies Galaburda *et al*. reported individuals with dyslexia displayed reduced planum temporale asymmetry (Galaburda *et al*., 1985, Galaburda, 1989). Several *in vivo* structural magnetic resonance imaging (MRI) studies initially supported this finding

(Hynd *et al.*, 1990, Duara *et al.*, 1991) although most controlled studies have not found such an association (Best and Demb, 1999, Eckert *et al.*, 2003, Leonard *et al.*, 2002). One finding which has been replicated is based on an independent post-mortem study by Witelson (1985) which indicated the corpus callosum (CC), the main fibre tract connecting the two cerebral hemispheres, to be larger in lefthanders in the general population. This suggestion that the CC plays a role in the expression of handedness is supported by genetic evidence in a study by Brandler *et al.* (2013) (see section 1.2.3). Additionally, a functional MRI meta-analysis of six studies confirmed general activity appears greater in the right hemisphere of dyslexic individuals (Maisog *et al.*, 2008) with subsequent functional transcranial Doppler ultrasound confirming this finding (Illingworth and Bishop, 2009). In short, the exact relationship between dyslexia, handedness and cerebral asymmetry remains elusive.

Several other studies have observed a link between a reduced frequency of right-handedness and neuropsychiatric disorders. In a meta-analysis of 16 studies (N = 3,175 patients with 65,284 control subjects) Dragovic and Hammond (2005) indicated that schizophrenia patients are significantly more left-handed than controls (odds ratio of 1.81, 95% confidence interval 1.6-2.1). Multiple meta-review articles incorporating dozens of studies have confirmed this consistent leftward shift (e.g. Hirnstein and Hugdahl (2014), Satz and Green (1999), Sommer *et al.* (2001)). Although null findings (Wahl, 1976, Oddy and Lobstein, 1972) and even reports of fewer non-right-handers in schizophrenia samples (Taylor *et al.*, 1980) have been found in earlier individual studies, in all, the vast weight of empirical evidence supports the establishment of a significant, reproducible link between schizophrenia and left-handedness.

Thus, studying the aetiology of handedness may provide us with important insights in to the development and trajectories of other forms of functional laterality and aspects of psychological functioning that are more difficult to study, particularly during infancy and childhood.

1.1.8 The development of handedness

Long before language function has developed, anatomical bias is observed even at the early fetal stages, with the right hand more developed than the left by 7 post fertilisation weeks (PFWs) (O'Rahilly and Muller, 2010) and the temporal gyri to be asymmetrical in fetal brains from 10–44 PFWs (Galaburda *et al.*, 1978). Additionally it has been shown that the development of the right hemisphere precedes that of the left in the brains of infants between 12 and 36 months of age, though this pattern is reversed towards the left hemisphere during language development at approximately 3 years of age (Chiron *et al.*, 1997).

In an oft-cited study Hepper *et al.* (1991) used ultrasound to observe that most foetuses at 15 PFWs prefer to suck their right thumb, following up this study 14 years later with the finding that 67% of left-handed foetuses remained left-handed while 100% of right-handed foetuses remained right-handed (Hepper *et al.*, 2005). Tan and Tan (1999) have also suggested that lateralised motor behaviour in early gestation is predictive of hand preference in adulthood, comparing neonate palmar grasp reflex strength to adult hand preference. Whether this early hand bias is controlled by spontaneous movement regulation in the spinal cord or by high-level regulation in the left hemisphere (Sun and Walsh, 2006), is still unknown however if valid, then such results demonstrate that foetuses early in the gestation period are manifesting lateralisation bias for handedness similar to those displayed later in childhood.

As a result of the infant's position *in utero*, most infants prefer to lay their heads to one side when supine for the first 12-24 weeks of life (Michel and Goodwin, 1979). Though hand preference during infancy is highly flexible (Corbetta *et al.*, 2006), it has been posited that infant postural preferences can influence the development of hand bias (Michel, 1981); recordings of hand preference for grasping objects (Michel *et al.*, 2006) and reaching (Marschik *et al.*, 2008) have been observed to mirror hand preference distributions firmly established by 7 years of age.

The studies presented thus far provide evidence that both functional and anatomical brain asymmetry precede the absorption of information from environment and cognitive development; some indeterminate intrinsic controls regulate human handedness very early in foetal development, a manifestation most likely derived from both environmental and genetic factors.

1.1.9 Genetic models of handedness

Although historically there have been non-genetic theories proposed for the determinism of hand preference (Provins, 1997, Morgan and Corballis, 1978), empirical support for the genetic contribution to handedness first came from twin studies which showed that monozygotic twins are more likely to be concordant for handedness than dizygotic twin pairs (Bryden *et al*., 1997, Sicotte *et al*., 1999). A study conducted by Medland *et al*. (2009b) of almost 25,000 twins subsequently concluded that hand preference is a weak genetic trait with a heritability of almost 25%, thereby ruling out exclusively non-genetic arguments in the process.

After several attempts to accommodate all the data (for a review see Harris (1992)) two competing but similar genetic causal models by McManus (1985a) and Annet (1998) proposed that hand preference is controlled by a single hypothesised gene with two alleles. Other simple genetic models were also proposed including the random-recessive model (Klar, 2003) and the X-linked three alleles model (McKeever, 2004), however all simple models, though they accommodate the heritability estimates (Klar, 1996), have been largely discarded. Traditional molecular approaches such as linkage analysis have failed to identify a single-gene locus that can be replicated across studies while evidence from the large cohorts of modern genome-wide association studies (GWASs) conclude in a rejection of the simple genetic models. One such GWAS meta-analysis of 10 studies (Armour *et al*., 2014) had 99% power to detect a single locus for the simple models of McManus and Annett but found no significant associations. Thus, how handedness is determined genetically is not straightforward and must involve multiple genes or other unknown factors (Mackay, 2014).

1.1.10 Previous studies and candidate genes

The handedness trait in humans appears to have been under the selective pressures of multiple factors; bipedal posture, communicative gesturing and task complexity may be manifested in a complex genetic architecture and influenced by the possibility of multiple evolutionary origins (Table 1.1). Furthermore, in a meta-analysis of GWASs McManus *et al*. (2013) concluded that handedness is not determined by a single genetic locus but instead estimate a minimum of 40 loci to be involved in determining this complex trait. Genetic linkage studies have not identified a single locus, instead proposing multiple regions connected with handedness (Figure 1.2). This may be due to differences in how handedness was measured between studies or, alternatively, provides further evidence to support the notion that handedness is a polygenic trait.

Handedness has typically been measured using a questionnaire approach (e.g. (Annett, 1970, Oldfield, 1971)) in which participants are asked to provide a simple binary 'left' or 'right' answer. Due to language bias restrictions and the subjective nature of this methodology many researchers have also employed a quantitate-centric approach, instead utilising hand performance measures to infer relative hand skill by deriving a score which indicates the difference in performance between the two hands. One such measure of hand performance is PegQ, a peg-board task which measures the time taken by individuals to move a row of pegs from one location to another with the left hand and right hand separately, after which a laterality quotient score is derived (see section 2.2 for a full discussion).

The first genome-wide linkage screen was conducted by Francks *et al*. (2002) on 195 dyslexic sibling pairs, identifying two quantitative trait loci (QTLs) for the PegQ measure on chromosome regions 17p11–q23 and 2p11.2–12. The latter QTL was subsequently confirmed in a study of left-handed brothers (Francks *et al*., 2003a) and revealed a parent-of-origin effect (Francks *et al*., 2003b). Assessed by questionnaire, Warren *et al*. (2006) identified in 584 Mexican–Americans a linkage signal for writing hand preference within chromosome region 12q21–23. Using a similar questionnaire-centred approach to derive a laterality quotient, Van Agtmael (2002) found genetic linkage for handedness on chromosome region 10q26 in a study of 25 Australian nuclear families. Finally, in a cohort of 180 pairs of left-handed brothers, Laval *et al*. (1998) proposed linkage between relative hand skill as measured by questionnaire and the Annett Peg board test (Annett, 1994) and a marker on the X chromosome (Xq21). For a summary of previous genetics studies investigating hand preference see Table 1.2

Table 1.2 Previous genetics studies investigating hand preference and relative hand skill and the candidate loci implicated

Linkage Analysis Studies				
	Sample set characteristics			
Reference	Country	Composition	Phenotype	Findings summary
Laval *et al.* (1998)	UK	N = 180 pairs of left-handed brothers	Measure of relative hand skill	A weak linkage finding for handedness close to the Xq21.3 region was observed (maximum LOD score of 2.8). No evidence for a locus linked to increased likelihood of left-handedness.
Van Agtmael *et al.* (2002)	Australia	N = 173	Hand preference	Segregation analysis in an extended pedigree identified allele sharing in the NODAL candidate region on chromosome 10 for left handers
Francks *et al.* (2002)	UK	N = 195 reading-disabled sibling pairs	Measure of relative hand skill	Locus on chromosome 2p11.2-12 yielded strong evidence for linkage to a measure of relative hand skill, PegQ (empirical P=.00007), and another locus on 17p11-q23 was also identified (empirical P=.002)
Francks *et al.* (2003a)	UK	N = 105 pairs of left handed adult brothers	Measure of relative hand skill	Further evidence for linkage of the 2p12-q11 locus to a measure of relative hand skill (P=.00035), which exceeded the critical value of P=0.01
Warren *et al.* (2006)	US	N = 584	Hand preference for writing	Using genome-wide multipoint linkage screens using 382 highly informative autosomal STR markers, suggestive linkage signals for drawing (LOD 2.10) and writing (LOD 2.00) hand preference were identified on chromosome 12q21–23

Association Studies

		Sample set characteristics			
Gene	Reference	Country	Composition	Phenotype	Findings summary
AR	Medland *et al.* (2005)	Australia/Holland	N = 783 adults	Hand preference for writing	A greater number of CAG repeats in the androgen receptor gene *AR* was correlated with a lower incidence of left-handedness
LRRTM1	Francks *et al.* (2007)	UK	N = 222 dyslexic siblings	Quantitative measure for handedness	LRRTM1 (Leucine-rich repeat transmembrane neuronal 1) is a maternally suppressed gene that is associated paternally with handedness
COMT	Savitz *et al.* (2007)	South Africa	N = 240 (55 bipolar disease patients and relatives)	Hand preference/ Relative hand skill	Significant association between the catechol-O-methyltransferase gene *COMT* Val158Met polymorphism and relative hand skill but not hand preference, with the M*et al*lele of this polymorphism being associated with greater right-handed skill
APOE	Bloss *et al.* (2010)	US	N = 147 children	Hand preference for writing	Apolipoprotein E gene *APOE* epsilon 2 allele carriers showed a significantly higher prevalence of left-handedness compared to other allele carriers
PCSK6	Scerri *et al.* (2011a)/Brandler *et al.* (2013)	UK	N = 192/728 dyslexic individuals	Quantitative measure for handedness	Proprotein convertase subtilisin/kexin type 6 PCSK6 gene is associated with handedness in individuals with dyslexia
AR	Hampson and Sankar (2012)	Canada	N = 180 adult males	Hand preference	Mixed-handers carry a significantly longer CAG repeat in the AR gene than either strong left- or strong right-handers who did not differ significantly from each other in terms of CAG repeat length
PCSK6	Arning *et al.* (2013)	Germany	N = 1113 adults	Hand preference	PCSK6 is associated with degree of handedness (consistent vs. inconsistent handedness), but not direction (left-handedness vs. right-handedness)
N/A	Medland *et al.* (2009a)	Australia/UK	N = 23,433	Hand preference for writing	No significant associations found. Promising results were found on chromosome 5 (P=2.455 x 10^{-7}) and 13 (P=6.149 x 10^{-7}), in regions that encompass SLIT3, MAB21L1 and NBEA

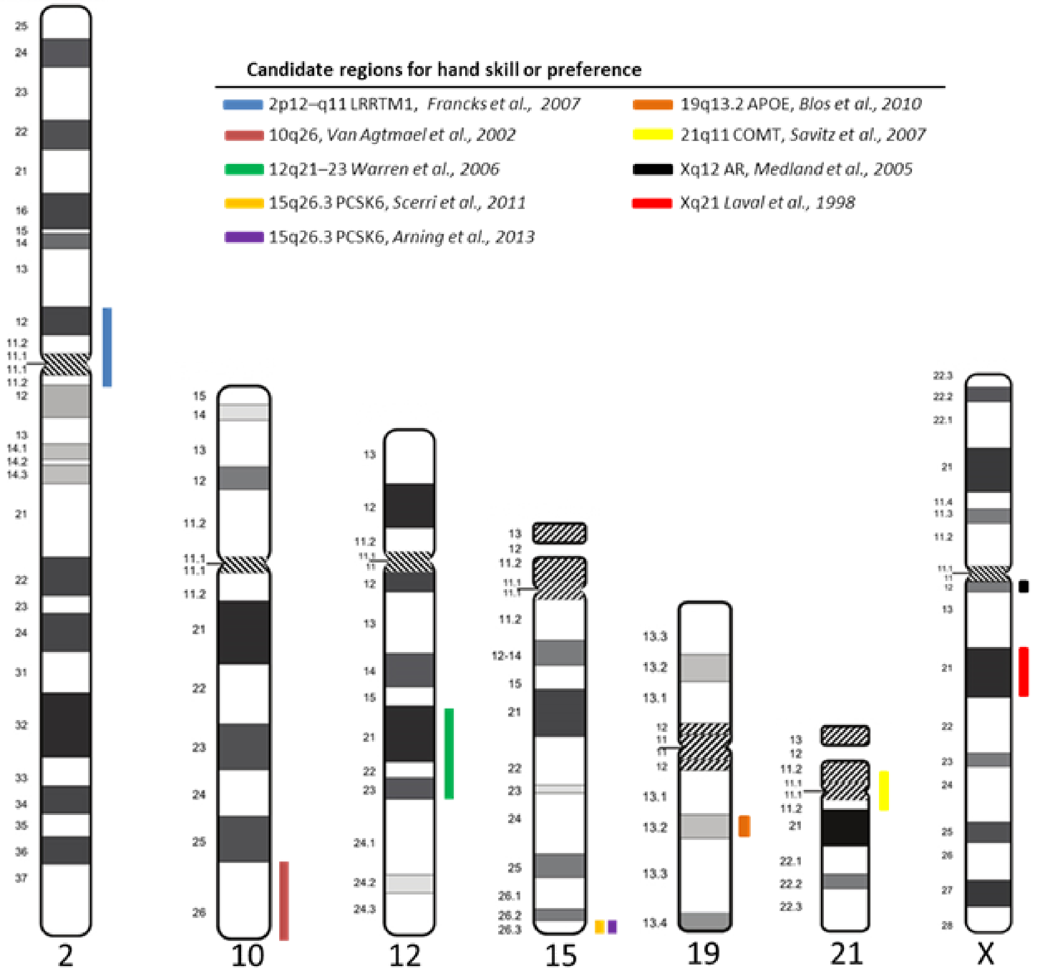

Candidate regions for hand skill or preference
2p12–q11 LRRTM1, *Francks et al., 2007*
10q26, *Van Agtmael et al., 2002*
12q21–23 *Warren et al., 2006*
15q26.3 PCSK6, *Scerri et al., 2011*
15q26.3 PCSK6, *Arning et al., 2013*
19q13.2 APOE, *Blos et al., 2010*
21q11 COMT, *Savitz et al., 2007*
Xq12 AR, *Medland et al., 2005*
Xq21 *Laval et al., 1998*
25
24
23
22
21
16
15
14
13
12
11.2
11.1
11.1
11.2
12
13
14.1
14.2
14.3
21
22
23
24
31
32
33
34
35
36
37
2
15
14
13
12
11.2
11.1
11.1
11.2
21
22
23
24
25
26
10
13
12
11.2
11.1
11
12
13
14
15
21
22
23
24.1
24.2
24.3
12
13
12
11.2
11.1
11.1
11.2
12-14
15
21
22
23
24
25
26.1
26.2
26.3
15
13.3
13.2
13.1
12
11
11
12
13.1
13.2
13.3
13.4
19
13
12
11.2
11.1
11.1
11.2
21
22.1
22.2
22.3
21
22.3
22.2
22.1
21
11.4
11.3
11.2
11.1
11
12
13
21
22
23
24
25
26
27
28
X

Figure 1.2 Graphical summary of candidate loci for hand skill or hand preference from previous studies. Each coloured line represents the approximate location on each numbered chromosome of a locus implicated either through previous linkage analysis or association studies. See Table 1.2 for further details on individual studies.

Linkage analysis, though suitable for detecting single gene defects in families with a clear cut pattern of inheritance, is impractical when analysing complex traits where it is hypothesised many loci make small contributions to the trait aetiology (Bush and Moore, 2012). Instead, GWAS is a hypothesis-free approach that typically compares the allele frequency of genetic variants across the genome to check which variants are associated with a trait or disease. For quantitative traits, correlation between alleles of each variant and the trait is measured in population or clinical-based cohorts. Genetic variants are usually referred to as common (Minor Allele Frequency or MAF > 1% in a human population) or rare (MAF < 1%) and divided broadly in to several classes: Single Nucleotide Polymorphisms (SNPs), indels and structural variants (e.g., Variable Number Tandem Repeats (VNTRs) and Copy Number Variations (CNVs) among others (Eichler *et al.*, 2007)). SNPs are the most prevalent class of genetic variation among individuals; as of September 2015 dbSNP build 144 contains 97,535,033 validated SNPs (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi). Whether common complex traits are primarily due to common variants with small effect size (Reich and Lander, 2001) or are the result of rare, high-penetrance variants (Bodmer and Bonilla, 2008) is still being debated (Gibson, 2012, Saint Pierre and Génin, 2014). In any case, the importance of common and rare variants in common complex phenotypic traits has been firmly established with each variant type contributing to various degrees depending on the heritability and epidemiology of the particular trait under consideration (Wellcome Trust Case Control, 2007, Sanna *et al.*, 2008, Visscher *et al.*, 2012, Rivas *et al.*, 2011).

An earlier study which investigated the fine mapping of linkage regions was conducted by Francks *et al.* (2007) who reported the significant association of a quantitative measure of human handedness with a haplotype upstream of the Leucine rich repeat transmembrane neuronal 1 (LRRTM1) gene in a set of 222 dyslexic siblings when the haplotype was inherited paternally (though this finding could not be replicated in the same study). In a study involving a similar quantative measure, Savitz *et al.* (2007) conducted a family-based genetic association analysis of a bipolar cohort with family members (N = 240) and found relative hand skill was significantly associated with a functional variant in the catechol-O-methyltransferase (COMT) gene.

It has been hypothesised that prenatal exposure to testosterone may affect lateralisation by influencing cell death in the foetal brain (Bauer *et al*., 2013, Haudry *et al*., 2013). Testosterone binds to the X chromosome-linked androgen receptor (AR), which contains a polymorphic CAG repeat, the length of which positively correlates with testosterone levels in males, and negatively correlates in females. In a candidate gene approach involving 783 twin pairs Medland *et al*. (2005) found that the length of the repeating CAG sequence has reverse effects on the probability of left-handedness in females and males. The same genetic marker was subsequently found to be associated with mixed-handedness in a smaller cohort of 180 adult males (Hampson and Sankar, 2012). For a visual summary of previous findings see Figure 1.2.

One of the most significant, intriguing and recurring associations discovered to date has been between various measures of handedness and a chromosome region (15q26) within the gene proprotein convertase subtilisin/kexin type 6 (PCSK6). In the first GWAS for a quantitative measure of human handedness (as assessed by the peg-board task, PegQ), (Scerri *et al.,* 2011a) found the first SNP for handedness to be identified at a genome-wide significant level ($P < 5 \text{ x } 10^{-8}$). Although no SNPs gave P-values below 5×10^{-8} in an initial RD cohort (N = 192), a meta-analysis of this and a subsequent two RD cohorts (N = 744) showed rs11855415, a SNP located within an intron of PCSK6, to be significantly associated with the PegQ measure ($P = 2 \text{ x } 10^{-8}$). The minor 'A' allele of rs11855415 was shown to confer increased relative right-hand skill in the dyslexic cohort and showed a nominally significant trend towards reduced laterality of hand skill in the general population (P=0.002, N = 2,666). The observation that the increase in relative right-hand skill associated with rs11855415 was specific to the dyslexic cohort led the authors to suggest there may be epistatic interaction between PCSK6 and dyslexia candidate genes thought to be involved in axon guidance and neuronal migration such as KIAA0319 (Dennis *et al*., 2009), DCDC2 (Schumacher *et al*., 2006) and DYX1C1 (Wang *et al*., 2006). A GWAS for the same quantitative measure of relative hand skill (PegQ) was conducted by Brandler *et al*. (2013) though this time combining the previous 3 RD cohorts of Scerri *et al*. (N = 728) and found the most strongly associated variant, rs7182874 ($P = 8.68 \times 10^{-9}$), to be located within the same PCSK6 locus*,* further supporting the previous rs11855415 association reported by

Scerri *et al.* (2011a). The rs7182874 association was specific to a dyslexic cohort (N = 728) and was not associated with handedness in a general population cohort (N = 2,666); the most highly associated SNP in the general population cohort was rs7883190 ($P = 2.08 \times 10^{-6}$) which is located 6 kb upstream of the gene GPC3, a gene that causes visceral organ asymmetry defects when disrupted in mice. In an independent study of a German general population (N = 1,113) Arning *et al.* (2013) showed a 33 bp variable-number tandem repeat (VNTR) within the same PCSK6 15q26 locus to be significantly associated ($P < 0.0025$) with the degree of handedness rather than the direction, as assessed by a hand preference questionnaire (Oldfield, 1971). The authors also investigated for the effect of rs1185415 SNP and found it failed to reach nominally significant levels of association ($P = 0.14$) though a failure to replicate can occur for numerous reasons including inadequate sample size or variability in phenotype definitions across independent samples (Greene *et al.*, 2009), as was the case here. However like Scerri *et al.* they concluded that carriers of the minor 'A' allele show reduced variability in relative hand skill in the general population, though they did not indicate how significant this association to be. Since no significant association was found to exist between the VNTR and handedness direction, it may be that handedness consistency (or strength of preference) and direction represent distinct phenotypes. Support for this is provided in multiple studies which show handedness consistency, as opposed to handedness direction, is a reliable predictor of hand performance in several cognitive domains, e.g. risk perception (Westfall *et al.*, 2012) and episodic memory retrieval (Propper *et al.*, 2005) (for review see Prichard *et al.* (2013)).

Curiously, the notion that strength and direction of preference is represented by two discrete phenotypes is also supported by studies in zebrafish. In this species (*Danio rerio)*, behavioural lateralisation is modulated by structural asymmetries in the epithalamus, a part of the dorsal forebrain, which includes the pineal gland (Bianco and Wilson, 2009). The occurrence of these anatomical asymmetries is regulated during embryonic development by several genes within the Nodal pathway, a highly conserved pathway across the *Bilateria* phyla, which determines left-right asymmetry and anteroposterior body asymmetries early in development (Levin, 2005, Grande and Patel, 2009). Disruption of this pathway does not affect the structural asymmetry itself, but leads to a randomised direction of the asymmetry and not leftward as in the wild-type

(Concha *et al.*, 2000). Therefore, it seems reasonable to suggest that the strength and direction of these hemispheric asymmetries in zebrafish might be controlled by two different genetic pathways.

Significant evidence from multiple large-scale studies suggests PCSK6 is important in the development of handedness, but its exact contribution to phenotype expression remains unknown. One possible mechanism by which PCSK6 could influence in the ontogenesis of handedness is its primary role in the LR-determining Nodal pathway (Shen, 2007). To understand this role it is necessary to take a look at this intriguing handedness candidate in greater detail.

1.2 PCSK6

1.2.1 The PCSK6 gene

Secretory proteins such as receptors, hormones, neuropeptides, adhesion molecules, growth factors and enzymes are essential for cellular function and are biologically activated through post-translational processing in the form of endoproteolytic cleavage (Artenstein and Opal, 2011). Such a mechanism has allowed evolutionarily complex species to maintain homeostasis, responding to challenges where and when required. PCSK6, previously known as PACE4, is one such cleavage enzyme and belongs to a family of furin-like proprotein convertases which proteolytically activate substrates through the recognition of specific single or paired basic amino acids found within the substrates proregion (Layne, 2013). PCSK6 is constitutively secreted from the trans-Golgi network into the extracellular matrix and primarily expressed in the liver, brain, testis and colon (Figure 1.3). There are 8 recognised spliced transcript variants (Pruitt *et al.*, 2014), some of which are assumed to be inactive in cleavage due to missing protein domains, and others which are expressed only in certain tissues (Tsuji *et al.*, 1997).

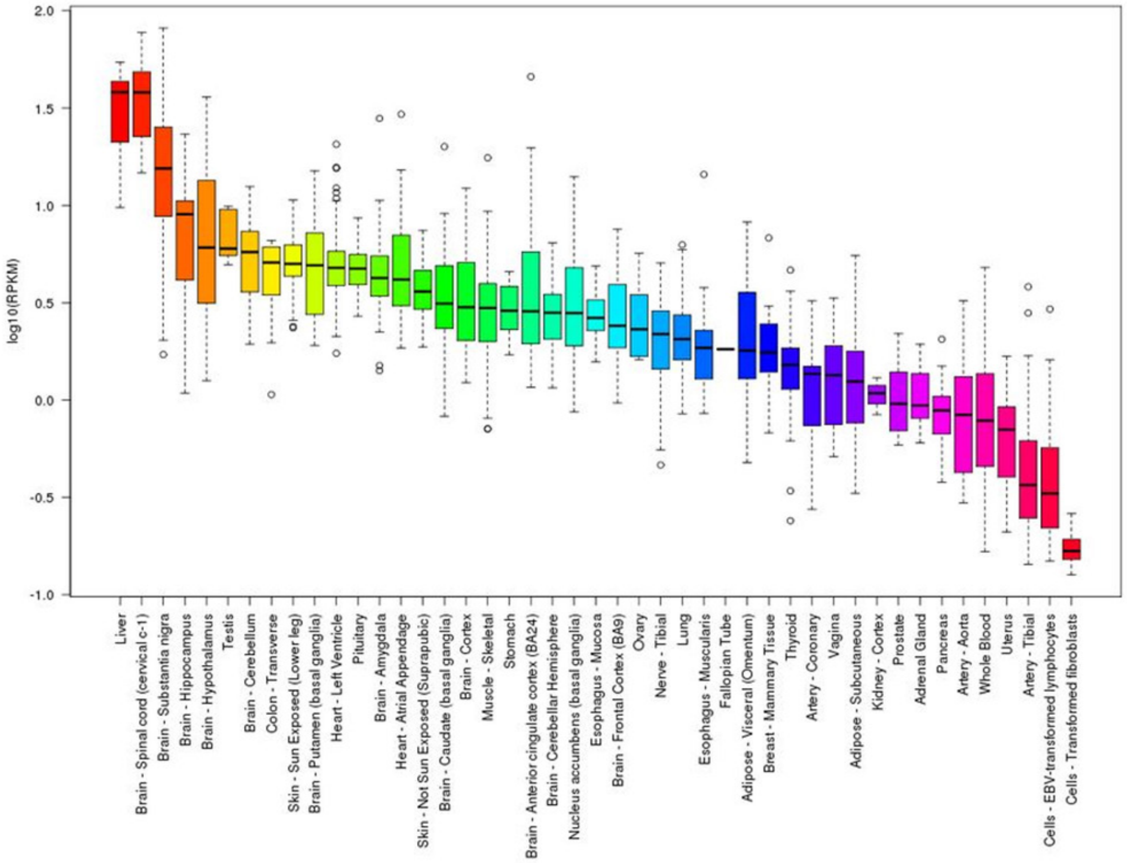


Figure 1.3 PCSK6 gene expression in adult human tissue. The plot was generated by the Gtex portal www.gtexportal.org and represents relative gene expression for PCSK6 in reads per kilobase per million (RPKM) across a range of adult tissue types. See (Lonsdale *et al.,* 2013) for details on data generation.

1.2.2 PCSK6 role in complex traits and disease

PCSK6 is thought to influence glioblastoma (GBM) tumour progression (Delic *et al*., 2012) and the development of prostate (D'Anjou *et al*., 2011), breast (Cheng *et al*., 1997) and skin (Bassi *et al*., 2010) cancers. Interestingly, the gene has also been found to be a direct target of the transcription factor forkhead box protein p2 (FOXP2). Vernes *et al*. (2007) demonstrated a reduction of FOXP2 negatively affects the development and function of a variety of neural circuits, including those involved in speech and language acquisition. This finding supported previous research by (Lai *et al*., 2001) who demonstrated disruption of the FOXP2 in a three-generation pedigree with severe speech and language disorders, proposing FOXP2 assumes a primary role in the developmental process that culminates in speech and language.

PCSK6 is known to have several downstream substrates, one of which is ADAMTS4, whose own substrate BCAN is localised to the surface of neurons in the brain, suggesting a primary role for this enzyme in the central nervous system (CNS) (Frischknecht and Seidenbecher, 2012). However, one of the more interesting findings related to PCSK6 is its role in embryogenesis and development. Knockout mice studies show the absence of PCSK6 leads to substantial anatomical defects in axis development, visceral orientation and craniofacial abnormalities (Constam and Robertson, 2000). Intriguingly, these defects in the left and right axis are preceded at embryonic day 8.5 by abnormal mRNA levels of another PCSK6 substrate; the axis determining TGF beta-like growth factor Nodal (Constam and Robertson, 2000).

1.2.3 Nodal signalling pathway and cilia

PCSK6 is known to cleave the axis-determining protein Nodal which plays a vital role in determining LR and anteroposterior body asymmetries early in development (Levin, 2005, Grande and Patel, 2009). Specifically, Nodal mRNA produces an immature protein form of Nodal that is subsequently cleaved by PCSK6 in order to generate a mature morphogen; in effect PCSK6 is regulating Nodal conversion from an inactive to an active form (Figure 1.4).

Nodal expression is largely restricted to embryonic tissues and is a critical factor in normal embryonic development for a range of species across the entire bilaterian phyla (Schier and Shen, 2000). For each of the vertebrate species in which Nodal signalling has been studied, Nodal must be activated on the left side to specify the left–right axis of the developing body plan. In the absence of Nodal signalling, visceral organs are distributed in a random (*heterotaxia*) or reversed (*situs inversus*) manner and not the default distribution (*situs solitus*), demonstrating that the Nodal pathway determines the direction of asymmetries but not their early establishment (Concha *et al*., 2000). To illustrate the role of PCSK6 in the Nodal pathway Constam and Robertson (2000) were able to induce PCSK6 knockdown mouse embryos to bilaterally express the normally asymmetrically expressed Nodal morphogen.

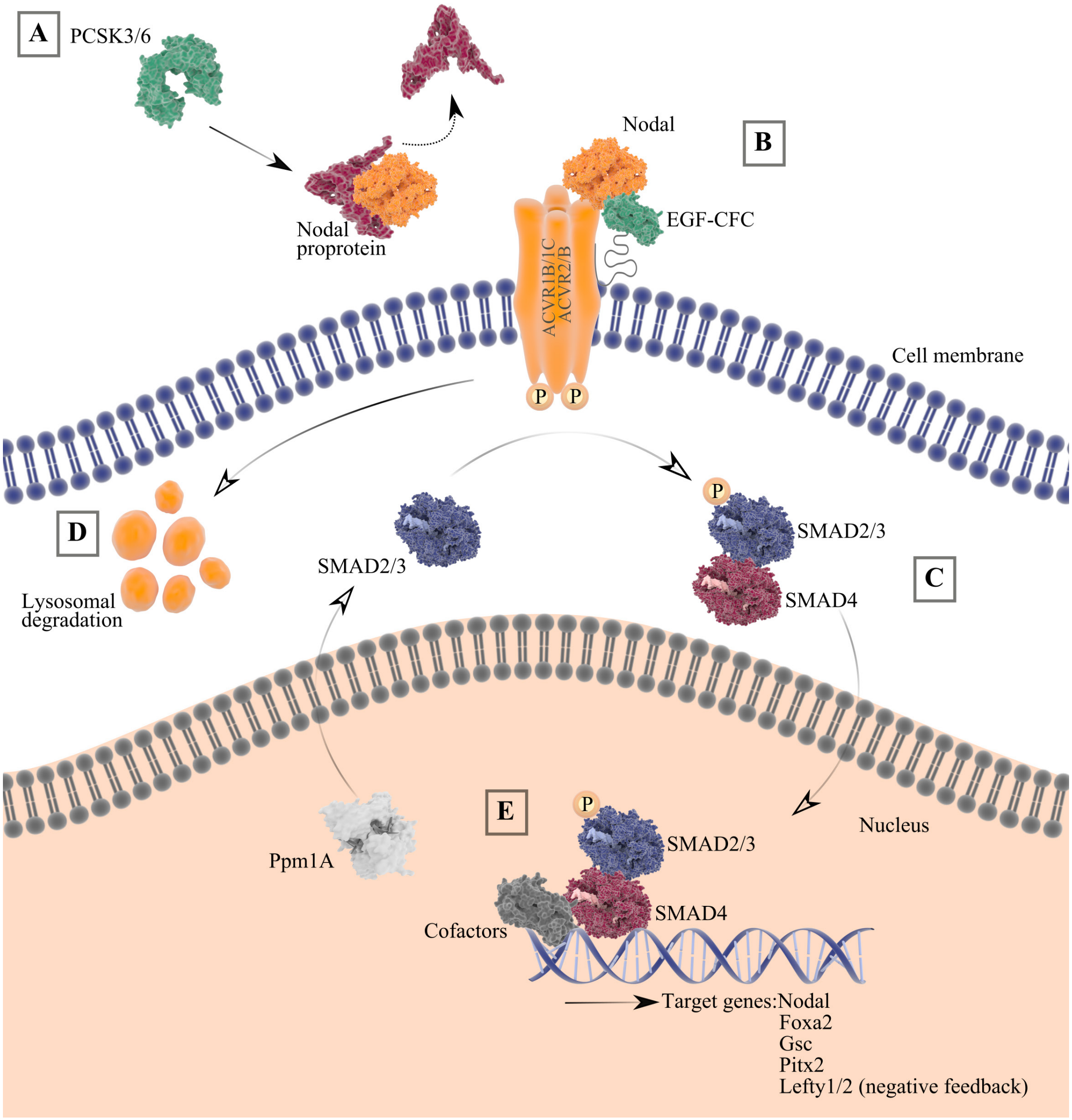


Figure 1.4 The prominent role of PCSK6 in the developing embryo NODAL pathway **(A)** The nodal precursor is expressed as a homodimeric proprotein, and is cleaved extracellularly to generate the carboxy-terminal ligand by Furin (also known as PCSK3) and PCSK6. **(B)** Mature Nodal ligands bind to an EGF-CFC co-receptor which is anchored to the membrane by a glycolipid tether and activates the type I (ACVR1B/ACVR1C) and type II receptors (ACVR2A or ACVR2B dimers), which phosphorylates the downstream effectors SMAD2/3. **(C)** Activated receptor SMADs associate with SMAD4 and translocate to the nucleus to regulate target gene expression (Arnold and Robertson, 2009). Meanwhile, the receptor complex undergoes internalization into endosomes and can be targeted for lysosomal degradation **(D)**. **(E)** Within the nucleus transcription cofactors, including the winged helix factors FOXA2 and FOXH1, function cooperatively to target promoters leading to transcription that activates expression of Nodal target genes, thereby specifying the cell is on the left side of the embryo (Wrana *et al.*, 1994). Nodal up-regulates its own expression through a SMAD–FOXH1-dependent auto-regulatory enhancer, but also triggers a negative-feedback circuit by prompting the expression of soluble antagonists left–right determination factor 1 (LEFTY1) or LEFTY2, which reduce Nodal signalling by interacting with EGF-CFC co-receptors to inhibit their function. The Smad phosphatase PPM1A, promotes efficient SMAD2 and SMAD3 turnover by promoting the nuclear export of SMAD2 and possibly targeting it for proteasomal degradation

One of the most interesting questions arising from this research was whether asymmetrical patterning of the visceral organs and the phenomenon of brain lateralisation are controlled by the same biological mechanisms or pathways. A significant analysis and discussion on the subject was presented by Hirokawa *et al.* (2006) who investigated LR asymmetry and the role played by motile cilia; clockwise rotating whip-like structures that project from the cell body and create a leftward flow at the ventral node thereby inducing asymmetrical expression of genes such as NODAL. To demonstrate if visceral and brain lateralisation were controlled by the same pathway, McManus *et al.* (2004) performed an analysis of patients with Primary Ciliary Dyskinesia (PCD), a rare genetic disorder that causes a defect in the action of the motile cilia lining the respiratory tract. Results showed 50% of people with PCD have *situs inversus*; an expected phenotype, if the motile cilia fail to asymmetrically express the Nodal morphogen. However out of 88 individuals with PCD, only 15% of 46 individuals with *situs inversus*, and 14% of 42 individuals with *situs solitus*, were left handed. If visceral and brain lateralisation were controlled by the same pathway then one would expect an equal ratio of handedness in *situs inversus* patients though clearly cerebral lateralisation is still present, and at a ratio approximate to the general population.

It would seem reasonable therefore to suggest that the regulatory mechanisms involved in establishing brain asymmetries appear to be distinct from those that establish visceral organ asymmetry (Sun and Walsh, 2006). The evidence is mixed however and some studies suggest a shared model where handedness is under the control of many variants, some of which are in genes that also contribute to the determination of body LR asymmetry. For example, using gene-set enrichment analysis (GSEA) of GWAS data Brandler *et al.* (2013) showed polymorphisms within particular LR asymmetry genes to be associated with relative hand skill (e.g. PCSK6 P = 3.9 x 10^{-8}, PKD2 P = 3.4 x 10^{-4} and MNS1 P = 8.7 x 10^{-4}). The mice homologues of these same genes, when knocked out, result in an absent corpus callosum and the LR asymmetry phenotypes *heterotaxia* and *situs inversus*.

1.2.4 RD candidate genes and cilia

Previous research has suggested a potential role for dyslexia candidate genes in the function and structure of cilia. A study conducted by Massinen *et al.* (2011) demonstrated the protein of the dyslexia candidate gene DCDC2 localises to the primary cilium in rat hippocampal neurons while an overexpression of the same gene was shown to increase *ciliary* length. Using a zebrafish model null for DYX1C1, Chandrasekar *et al.* (2013), produced pleiotropic phenotypes characteristically associated with cilia defects such as *situs inversus* and kidney cysts. These findings, suggesting a role for DYX1C1 and DCDC2 in ciliogenesis, were replicated in a large-scale meta-analysis of gene co-expression networks by Ivliev *et al.* (2012) who also proposed the dyslexia candidate gene KIAA0319 to influence the development of cilia.

Ciliopathies influence not only LR body asymmetry phenotypes (Fliegauf *et al.*, 2007), but also a broad spectrum of disorders affecting multiple organs including brain midline phenotypes such as absent cerebellar vermis and missing corpus callosum (Badano *et al.*, 2006). Furthermore, some ciliary disorders affect the development of the CNS and thus can have an effect on cognitive functions. One such disorder is Bardet-Biedl syndrome where patients are characterised by *situs inversus* and language and learning deficits, an interesting phenotype in the context of dyslexia (Allendorf and Luikart, 2007). Indeed, the specificity of the Scerri *et al.*, 2011a) and Brandler *et al.* (2013) findings; that PCSK6 is associated with a measure for relative hand skill in a dyslexic cohort, is suggestive of a possible epistatic interaction between PCSK6 and dyslexia candidate genes. In conclusion, it would seem that handedness is determined in part by the mechanisms that establish LR asymmetry early in development, such as ciliogenesis and the PCSK6-dependant Nodal signalling pathway (Brandler and Paracchini, 2014).

1.2.5 Role of regulatory elements during neuronal development

The evolution of human-specific anatomical and functional asymmetries has occurred without a significant change in the number of protein coding genes, suggesting the regulation of gene expression during the development of the CNS plays an important role (Wray, 2007). Specific to this project, the role of non-coding regulatory elements are of particular interest since neither of the previous relevant GWA handedness studies (Scerri *et al.,* 2011a, Brandler *et al.,* 2013) implicated a coding variant, but instead reported the strongest association close to a predicted intronic regulatory region of PCSK6 thought to contain non-coding RNA (1.2.5.1), Variable Number Tandem Repeats (1.2.5.2) and bidirectional promoters (1.2.5.3).

1.2.5.1 Non-coding RNA

The complexity of the human brain arises from the capacity to produce functional diversity through mechanisms such as RNA editing and alternative splicing; processes more widespread in the CNS than in any other tissue (Ramskold *et al*., 2009). The role of RNA as the relay of genetic information for protein synthesis has long been known. However multiple studies have also demonstrated the functional ability of non-coding RNAs (ncRNAs) and their substantial role in regulating gene expression (for review see Iyengar *et al*. (2014)).

ncRNA are often spliced, polyadenylated and roughly classified into two groups: long (> 400 nucleotides) and short (< 400 nucleotides) though many ncRNAs have mixed characteristics and do not clearly fall into any one category (Salta and De Strooper, 2012). Long ncRNAs (lncRNAs) play key roles in many important biological processes such as mammalian X-inactivation (Tian *et al*., 2010), cell differentiation, immunological responses and complex human disease (Ponting *et al*., 2009). One class of lncRNA that is of particular relevance to this project are the Natural Antisense Transcripts (NATs).

NATs are RNA polymerase II transcripts that originate from the antisense strand of protein-coding sense mRNAs and almost 70% of human and mouse genes have been reported to undergo antisense transcription (Werner and Sayer, 2009). Interestingly PCSK6 is known to have at least one corresponding NAT. How exactly NATs affect their sense counterpart's RNA level is complex but may occur at multiple levels to produce inhibition of the protein product (Katayama *et al*., 2005). Several studies suggest they are important in the fine-tuning of gene expression and NATs are known to play a role in neurological disorders such as Parkinson's (Morais *et al*., 2009), Huntington's (Johnson *et al*., 2010), and Alzheimer's disease (Seitz *et al*., 2005). NAT involvement in important neuronal processes such as oligodendrocyte differentiation (Pollard *et al*., 2006), cortical neuron specification and migration (Ling *et al*., 2011, Ramos *et al*., 2013), plasticity and long-term memory formation (Pruunsild *et al*., 2007) has also been shown.

Human lncRNA genes tend to be less well-conserved than protein-coding genes (Pang *et al*., 2006) and can give rise to unique transcripts not found in other species (Lipovich *et al*., 2012). This evolved specificity characteristic of lncRNA is shared by another mechanism for gene regulation; VNTRs (Variable Number of Tandem Repeats).

1.2.5.2 VNTRs

The VNTR genetic variation is of particular interest to this project. As discussed previously, Arning *et al*. (2013) demonstrated that a VNTR (rs10523972) at the same PCSK6 locus identified in the previous GWA studies (Scerri *et al.,* 2011a, Brandler *et al.,* 2013) was significantly associated with handedness consistency but not with handedness direction in a general population cohort (P < 0.0025, N = 1,113). VNTRs are repetitive copies of the same DNA sequence lying in tandem with one another and often occur within regulatory and coding regions (Li *et al*., 2002). VNTRs are not uniformly represented in all classes of genes and a study by Legendre *et al.* (2007) demonstrated that genes in particular categories of biological function are enriched for VNTRs (Table 1.3). Two ontological classes emerge from the categories presented in Table 1.3: transcriptional regulation and development.

For a long time these repeating DNA sequences were thought to have no biological function (Doolittle and Sapienza, 1980). However due to their highly unstable and dynamic nature (mutation rates up to 100,000 times higher than in other areas of the genome (Vogler *et al*., 2007)), VNTRs are increasingly seen as being linked to variation in function with particular emphasis on transcription regulation. For example, 10% -

20% of eukaryotic promoters and genes are estimated to contain a repeating DNA motif (Gemayel *et al.*, 2010). VNTRs in regulatory sequences can affect gene expression since a variation in the number of repeats can have a significant physical impact on the organisation of the DNA structure, thus introducing changes in spacing between critical promoter elements and altering the number of transcription factor binding sites (TFBSs) (Gemayel *et al.*, 2012).

Table 1.3 Enrichment of function among human genes containing VNTRs. Genes in the human genome containing VNTRs are enriched for particular functions and processes. Shown are the most enriched categories and the corresponding value of statistical significance. Adapted from Legendre *et al.* (2007). RNA Polymerase II (RNAPII).

Biological Process	*Adjusted P-value*
Regulation of transcription from RNAPII promoter	8.05×10^{-9}
Positive regulation of transcription; DNA dependent	6.09×10^{-4}
Forebrain development	4.15×10^{-3}
Negative regulation of metabolic processes	3.35×10^{-3}
Embryonic morphogenesis	7.90×10^{-3}
mRNA metabolic processes	9.28×10^{-3}
Sensory organ development	1.11×10^{-2}
Cell fate commitment	1.96×10^{-2}
Base-excision repair; DNA ligation	1.96×10^{-2}
Chromatin remodelling	2.26×10^{-2}
Organ morphogenesis	2.45×10^{-2}
Neurogenesis	2.55×10^{-2}
Anterior/posterior pattern formation	3.01×10^{-2}
Ribosome assembly	3.51×10^{-2}

An interesting perspective on the involvement of VNTRs in development is presented in the work of Allendorf *et al.* (2013), who provide evidence that VNTRs in key regulators of morphological development can drive diversity in dog breeds. Genetic variation in VNTRs often located within or next to human genes can lead to neurodegenerative diseases such as Fragile-X syndrome (Usdin, 2008) and Huntington's disease, where the presence of glutamine repeats corresponds with the gaining of a regulatory activity regulating neural adhesion in the complex mammalian CNS (Gemayel *et al.*, 2012). VNTRs have also been found to influence neurodevelopmental disorders such as Attention deficit/hyperactivity disorder (ADHD). Hedrick (2005) performed a meta-analysis of the association between the *S*LC6A3/DAT1 gene with persistent ADHD in 1,440 patients and 1,769 controls. A 9/9 genotype of the VNTR in the 3′-untranslated region (UTR) of the gene was found to be associated with persistent ADHD (OR 1.34, 95% CI 1.03–1.76, P = 0.03). However associations between the DAT1 VNTR and ADHD have been mostly inconsistent (Hedrick, 2011, Ammerman and Cavalli-Sforza, 1984). Alternatively VNTRs can also confer beneficial phenotypic variability, including plasticity in skeletal morphology (Jin *et al.*, 2009) and cell surface variability (Fidalgo *et al.*, 2006).

In short, the abundance of VNTRs in both coding and regulatory regions, across a range of organisms, suggests that VNTR genetic variation might be a common mechanism in influencing phenotypes, with a large number of studies reporting correlations between VNTRs and changes in gene expression (Fuke *et al.*, 2001, Hranilovic *et al.*, 2004, Fiskerstrand *et al.*, 1999).

1.2.5.3 Bidirectional promoters

Research suggests most promoters display residual bidirectional transcription; more than 10% of protein-coding genes in humans are transcribed from bidirectional promoters, that is, a genomic region of DNA that initiates transcription in both orientations (Piontkivska *et al.*, 2009). Such bidirectional promoters are the origin of pervasive lncRNAs and lead to the transcription of unique, lineage-specific transcripts (Piontkivska *et al.*, 2009). Furthermore, in human embryonic stem cells (ESCs), half of all expressed lncRNAs represent transcription from bidirectional promoters of known protein-coding genes (Sigova *et al.*, 2013), with a significant portion transcribed intragenically (Yang *et al.*, 2007).

Interestingly, rs11855415, the first SNP for handedness to be identified at a genome-wide significant level (Scerri *et al.* (2011a), $P < 5 \times 10^{-8}$), is located within an intron of *PCSK6* that is suspected to contain an intragenic bidirectional promoter; possibly driving the transcription of a shortened PCSK6 isoform and a lncRNA gene PCSK6-AS1. Ponjavic *et al.* (2009) found brain-expressed lncRNAs in mice originating from such promoters are significantly enriched for predicted RNA secondary structures and are more frequently conserved. Additionally, these lncRNAs are usually located adjacent to protein-coding genes that are (1) also found to be expressed in the brain and (2) contribute to nervous system development or in transcriptional regulation.

The outcome of ncRNA transcription from a bidirectional promoter depends on the subsequent transcript sequence, length and stability (Wei *et al*., 2011). ncRNAs generated from bidirectional promoters have demonstrated several functional roles including gene expression regulation acting at multiple levels; from modifying local chromatin (Hirota *et al*., 2008) to enabling regional signal spreading and more distal regulation (Xu *et al*., 2011).

To conclude, such complex regulatory architecture means that when mapping a phenotype to a locus, it is essential to consider bidirectional promoters and ncRNAs as sources of phenotypic variability (Wei *et al*., 2011). In the case of lncRNAs, Gong *et al.* (2015) recently developed a database which identified a total of 495,729 SNPs in more than 30,000 human lncRNA transcripts. Furthermore, by mapping SNPs to GWAS results from the NHGRI GWAS Catalogue (~10,000 SNPs linked to human traits/disease as of July 2015), they found 40% of the lncRNA SNPs were within ±500 kilo base pairs of 142 GWAS-identified SNPs at a genome-wide significant level.

1.3 Conclusion

1.3.1 Overall aim of this thesis

Previous findings by Scerri *et al.* (2011a) and Brandler *et al*. (2013) have identified PCSK6 as the first gene associated with a quantitative measure for relative hand skill (PegQ) at a statistically significant level ($P < 0.5 \times 10^{-8}$). The overall aim of the research presented in this thesis is to identify and characterise the causal variant(s) which may be affecting PCSK6 expression and to define what role regulatory mechanisms (e.g. lncRNA, VNTRs) might have to this effect.

1.3.2 Overview of the study methodology

The work described in the following chapters attempts to identify the mechanisms behind functional genetic change in PCSK6 in order to better understand the biology of the handedness trait, and more broadly, to further our understanding of the role of non-coding genetic variants in complex genetic diseases.

To do this, I performed association analysis for PCSK6 genetic variation with various measures for laterality and handedness (Chapter 2), employed bioinformatics analysis to refine the region of interest within PCSK6 (Chapter 3), performed sequencing and isoform profiling to characterise the refined locus of interest (Chapter 4) before considering a functional analysis on the effects of allelic variation on gene expression (Chapter 5). A flow chart of the work done in this project is summarised in Figure 1.5.

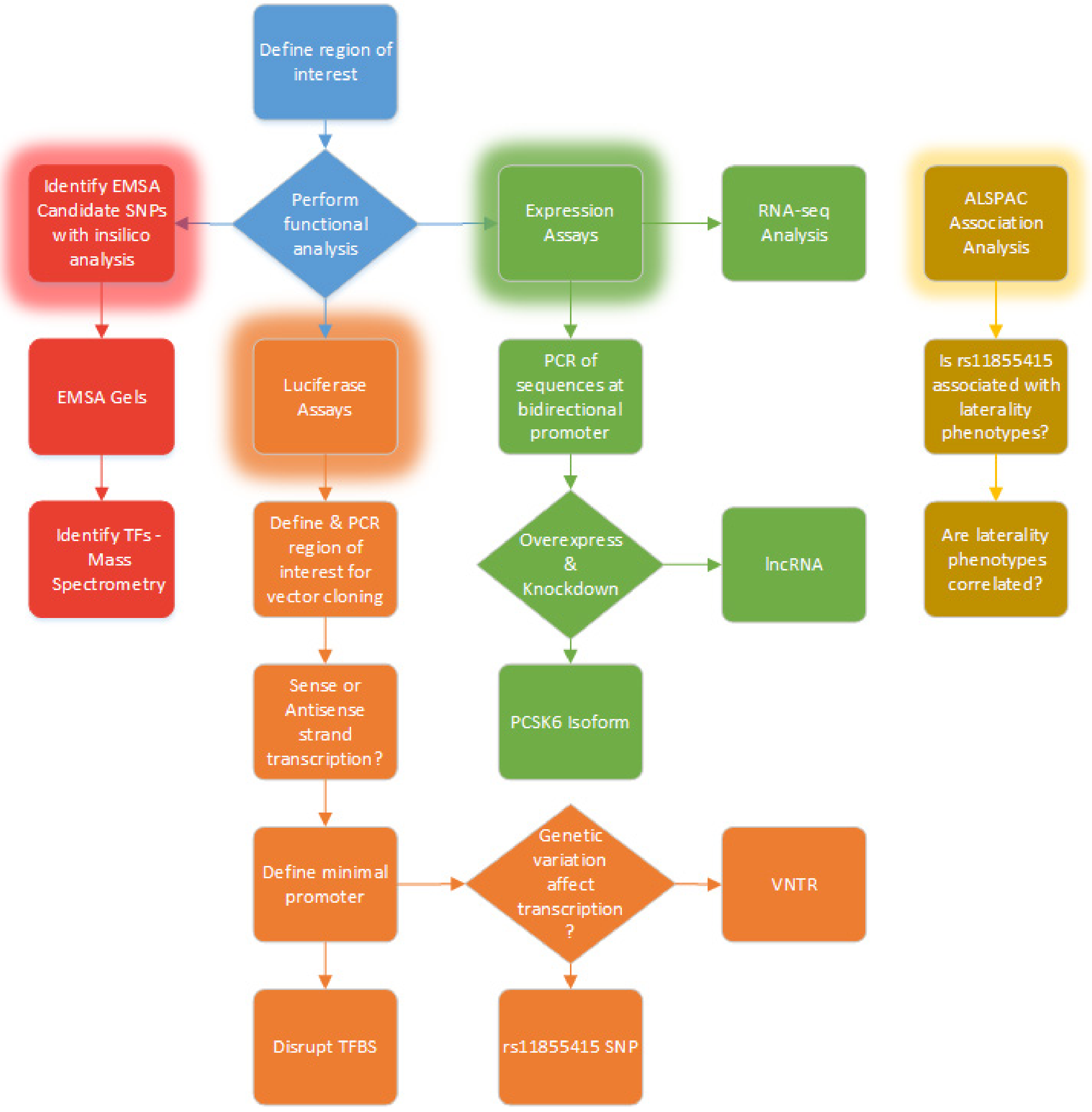

Define region of interest
Identify EMSA Candidate SNPs with insilico analysis
Perform functional analysis
Expression Assays
RNA-seq Analysis
ALSPAC Association Analysis
EMSA Gels
Luciferase Assays
PCR of sequences at bidirectional promoter
Is rs11855415 associated with laterality phenotypes?
Identify TFs - Mass Spectrometry
Define & PCR region of interest for vector cloning
Overexpress & Knockdown
lncRNA
Are laterality phenotypes correlated?
Sense or Antisense strand transcription?
PCSK6 Isoform
Define minimal promoter
Genetic variation affect transcription ?
VNTR
Disrupt TFBS
rs11855415 SNP

Figure 1.5 Scheme indicating experimental design flow of the research project. Different coloured boxes indicate the primary methodology shared by experiments in that section – red indicates gel-based electrophoresis analysis, orange includes all luciferase-based experiments, green represents expression assays which ultimately use polymerase chain reaction (PCR) and quantitative PCR (qPCR) for quantification while gold indicates software-based association analysis. Acronyms used are Electrophoresis Mobility Shift Assay (EMSA), transcription factor (TF), transcription factor bind site (TFBS), Single Nucleotide Polymorphism (SNP), long non-coding RNA (lncRNA), Variable Number Tandem Repeat (VNTR), Proprotein Convertase Serine Kinase 6 (PCSK6).

2 ALSPAC Cohort Association Analysis

2.1 Abstract

Scerri *et al.* (2011a) conducted a genome wide association study (GWAS) for a quantitative measure of relative hand skill (PegQ) in a dyslexic cohort and found the most highly associated marker, the Single Nucleotide Polymorphism (SNP) rs11855415 ($P < 0.5 \times 10^{-8}$), within a locus of PCSK6. The minor allele 'A' of rs11855415 was found to confer greater relative right-hand skill in a dyslexic cohort while in the general population (N = 2,666), the same minor allele showed a trend towards reduced laterality of hand skill (P = 0.002). This initial finding was supported by Brandler *et al.* (2013) in a subsequent GWAS who found multiple markers from the same PCSK6 locus to be associated with handedness in the same dyslexic cohort (N = 728), the most significant of which was rs7182874 ($P = 8.68 \times 10^{-9}$).

I have followed up these GWAS findings by conducting an association analysis between rs11855415 and various measures of relative hand skill and laterality in both neurodevelopmental and general population subgroups using the same cohorts as previously (Scerri *et al.,* 2011a, Brandler *et al.,* 2013). I report a nominally significant association between rs11855415 and both a quantitative measure of relative hand skill (PegQ7) and a categorical measure of hand preference (Hand11) in a dyslexic subgroup, but not in the general population cohort.

Finally, I derived a predictive model of handedness, with the hand performance measure MarkQ (r = 0.66), rather than PegQ7 (r = 0.35), contributing most to the variance of the handedness phenotype as measured by a categorical measure for hand preference at age 7 (Hand7).

2.2 Introduction

A recurrent finding from previous studies investigating handedness was the significant association with a PCSK6 locus in both the general population (Arning *et al.*, 2013) and dyslexic cohorts (Scerri *et al.,* 2011a, Brandler *et al.,* 2013). The specificity of the latter findings is interesting in the context of a meta-analysis of 25 studies by Eglinton and Annett (1994) who found a small but consistent increase in the prevalence of non-right handers among dyslexics. An updated meta-analysis of 44 studies by Koufaki (2010) supported the earlier findings of Eglinton & Annett and confirmed a statistically significant increase in non-right-handedness among dyslexics compared with non-dyslexics (N = 16,561 OR = 1.57, 95% confidence interval = 1.24 – 1.99). The reporting by both Scerri *et al.* (2011a) and Brandler *et al.* (2013) of a significant association between PegQ and PCSK6 in a dyslexic cohort (though not the general population) raises several interesting questions I aim to address in this chapter, namely:

1. Does PCSK6 associate with alternative measures which assess for relative hand skill? If not, then what are the properties (gripping, dexterity etc.) of the PegQ measure that support such a significant association?
2. Why is the reported significant association between PCSK6 and PegQ specific to a dyslexic cohort and, by extension, does such an association exist in other clinical cohorts?

To begin answering these questions it is useful to define what exactly the term dyslexia refers to and how we distinguish cohorts for future analysis. Derived from the Greek language meaning 'difficulty in reading', dyslexia (or reading disability, RD) can be defined as a specific difficulty in learning to read that cannot be attributed to age, intellectual deficiency, sensory disorders, or inadequate schooling (Ramus, 2014). RD affects approximately 6 - 10% of the population (DeFries, 1989) and although twin studies clearly show a genetic basis with a heritability of 0.53 - 0.82 (Wadsworth *et al.*, 2010, Olson *et al.*, 1989, Petrill *et al.*, 2007), the inheritance mode remains unclear. RD shows substantial overlapping predisposition with other neurodevelopmental disorders; approximately 33 - 45% of individuals with attention deficit hyperactivity disorder

(ADHD) show comorbidity for RD (Field *et al*., 2013) while it is estimated that 43% of children with Specific Language Impairment (SLI) are later diagnosed with RD (Snowling *et al*., 2000). When considering individuals for inclusion in association analysis studies it is important to recognise such comorbidities and filter where necessary. As such, I have included, with the help of my colleague Dr Kerry Pettigrew, both general population and neurodevelopmental disorder subgroups in my association analysis of the Avon Longitudinal Study of Parents and Children (ALSPAC) dataset (Figure 2.1).

The ALSPAC Children of the 90s project is a long-term birth cohort study that began in the early 1990s when it recruited more than 14,000 pregnant women who were due to give birth between April 1991 and December 1992. Further information on the ALSPAC cohort and the available data can be accessed via the study website (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/) and in the literature (Boyd *et al*., 2013, Team, 2001). Briefly, children from 4 weeks to 18 years of age are assessed across a wide range of behavioural, physical and neuropsychological phenotypes, including reading and laterality-related measures, such as handedness.

The assessment of handedness has traditionally been measured through the use of various questionnaires including the Waterloo Handedness Questionnaire (WHQ, Steenhuis *et al*. (1990)), the Edinburgh Handedness Inventory (Oldfield, 1971) and Annett's handedness questionnaire (Annett, 1970) however these suffer inherent limitations and do not recognise the multifactorial nature of handedness when they fail to assess for fine motor tasks or dexterity. To complement this approach researchers have employed performance-based measures such as grip strength, manual sorting, dot filling etc. as a cost-effective and accessible method to evaluate handedness. The advantages of using performance-based measures are their objectivity and quantitative nature, as opposed to the subjectivity and binary (or 'categorical') nature of a self-reporting handedness questionnaire. Indeed, since the strength of hand preference can vary by task (Willems *et al*., 2014), any assessment of one's handedness based on a combination of tasks is preferable to those based on a single task (e.g. asking which hand a person writes with) or to assessments that are made without reference to any specific task (e.g. such as a questionnaire's 'are you right or left handed?'). Furthermore, analysis by Corey *et al.* (2001) using unimanual motor tasks (moving pegs on a board and finger tapping) concluded that handedness is not a one-dimensional behaviour and must be defined using multiple measures that assess different aspects of hand preference and performance. As such, it is important to determine which performance measures accurately predict hand preference scores to facilitate research based on distinct handedness groups. In conclusion and in addition to the two primary objectives previously stated, this chapter also seeks to establish:

3. What is the correlation between quantitative measures of relative hand skill and if such measures can be used as predictors in assessing handedness and its stability over time?
4. Whether the SNP rs1185545, previously identified to be associated with a complex trait (relative hand skill), also contributes to other measures of laterality such as footedness or eyedness and whether these categorical measures themselves correlate with handedness?

2.3 Methods

Below is a detailed methodology which has been used to arrange individuals from the ALSPAC dataset in to RD, ADHD, SLI and general population subgroups and how these subgroups were subsequently phenotyped for multiple measures of laterality.

2.3.1 Ascertainment criteria for subgroups

From the entire ALSPAC children dataset (N = 15,443 as of July 2014) we identified a cohort using criteria defined by Scerri *et al.* (Scerri *et al.*, 2011b) that included only individuals with a near complete data set on all the measures used for sample assignment, Intelligence Quotient (IQ) and ethnicity. The same criteria were also used for the two previous GWA studies of relevance (Scerri *et al.,* 2011a, Brandler *et al.,* 2013). To avoid effects of population stratification, we excluded individuals that did not have a white European ethnicity based on a self-reported ethnicity of 'non-white'. Any individuals who had a performance IQ ≤ 85 at the age of 8.5 years or a score ≤ -3 standard deviation (SD) for a composite score of 7 measures from the Children's Communication Checklist (CCC) were removed; this second filter was to rule out individuals with autistic features at 7.5 years. These exclusion criteria removed individuals that may have performed badly on the psychometric tests for reasons other than specific reading or language impairment. Finally, we only included one child from any twin pairs for further analysis.

After filtering we were left with a sample baseline of N = 3,747 (F1, Figure 2.1) from which to ascertain subgroups. It is worth noting that this sample baseline differs slightly to Scerri *et al*. (2011a) and Brandler *et al*. (2013) (N = 3,725) since additional samples were added to the ALSPAC cohort in the most recent data release (July 2014). For subgroup filtering, individuals were assigned by an unaffected status or by neurodevelopmental disorder (RD, SLI, ADHD or any of the four comorbid combinations of these three disorders) giving eight subgroups in total (Table 2.1).

RD was identified based on an age-adjusted single-word reading score 1 SD below the mean for both a 40-word reading test at 7.5 years and a 12-word reading test at 9.5 years. Two time points were used to correct for random error on each individual measure. Individuals missing any data points were excluded. An assignment of ADHD was based on a DAWBA DSM-IV clinical diagnosis at 7.5 years of age. To record the different components of language impairment, an assignment of SLI was given if an individual scored positive for at least two of the following four criteria:

- an age adjusted non-word repetition score ≤ -1 SD at 8.5 years
- a composite score of 7 measures from the CCC pragmatic aspects of communication ≥ -3 SD and ≤ -1 SD at age 7.5 years
- an age adjusted WOLD comprehension score ≤ -1 SD given at 8.5 years
- a questionnaire given at 9.5 years asking if the individual has ever had language/speech therapy

These four criteria target different aspects of language problems and while each of them might over-identify impairment, two concomitant low scores have been shown to be a valid strategy to predict clinical diagnosis (Catts *et al*., 2005). In total, there were 442 affected individuals who met any of the assignment criteria detailed above for RD, SLI or ADHD (Affected Subgroup, Figure 2.1).

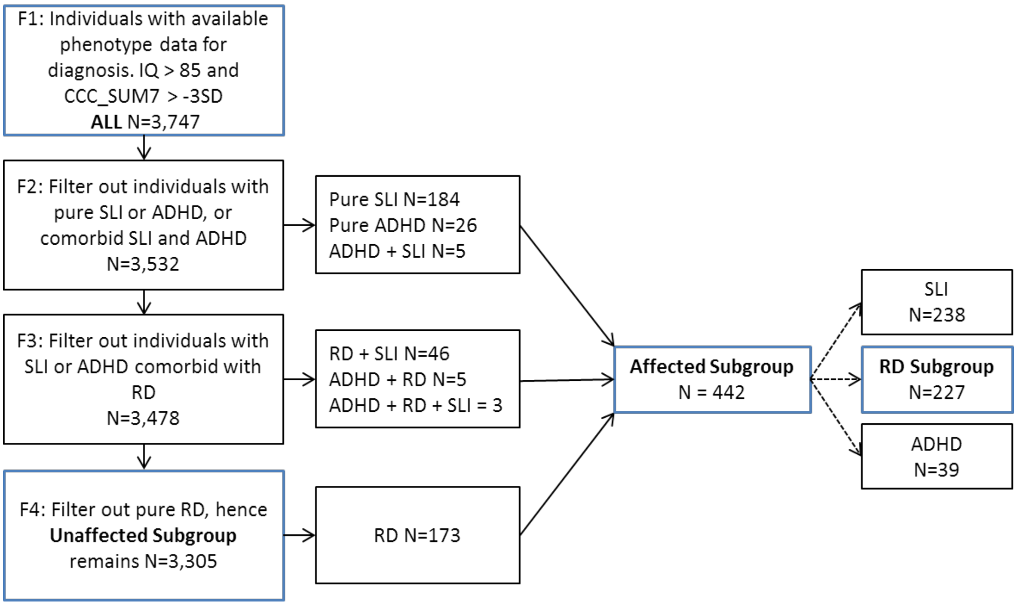

F1: Individuals with available phenotype data for diagnosis. IQ > 85 and CCC_SUM7 > -3SD
ALL N=3,747
F2: Filter out individuals with pure SLI or ADHD, or comorbid SLI and ADHD N=3,532
Pure SLI N=184
Pure ADHD N=26
ADHD + SLI N=5
F3: Filter out individuals with SLI or ADHD comorbid with RD N=3,478
RD + SLI N=46
ADHD + RD N=5
ADHD + RD + SLI = 3
F4: Filter out pure RD, hence Unaffected Subgroup remains N=3,305
RD N=173
Affected Subgroup N = 442
SLI N=238
RD Subgroup N=227
ADHD N=39

Figure 2.1 Diagram illustrating how phenotypic subgroups were identified from the ALSPAC dataset. Dashed line represents non-mutually exclusive subgrouping (comorbidity). The subgroups used throughout this chapter are highlighted in blue: All, Unaffected, RD and Affected. See Table 2.1 for full breakdown. CCC_SUM7 is the sum of first seven scales from the Children's Communication Checklist, Intelligence Quotient (IQ), Reading Disability (RD), Specific Language Impairment (SLI), and Attention-Deficit/Hyperactivity Disorder (ADHD).

The three derived subgroups relevant to this chapter are as follows (blue boxes, Figure 2.1):

(1) Unaffected Subgroup (N = 3,305): This subgroup was the remainder of the ALSPAC cohort after filtering for IQ, ethnicity, missing data and disorder

(2) RD Subgroup (N = 227): This subgroup consists of individuals with either pure RD or those with RD and a comorbidity for SLI and/or ADHD

(3) Affected Subgroup (N = 442): This subgroup contains all RD, SLI and ADHD individuals

Table 2.1 The final subgroup sizes following filtering of individuals from the ALSPAC cohort according to disorder

Subgroup	*N*
Pure SLI	184
Pure RD	173
Pure ADHD	26
RD + SLI	46
ADHD + SLI	5
ADHD + RD	5
ADHD + RD + SLI	3
Unaffected	3305
Excluded from subgroup filtering	11696

2.3.2 Laterality measures

Phenotype data for the following quantative and categorical measures for handedness were made available from the ALSPAC dataset for use in association analysis.

2.3.2.1 Quantitative measures of hand performance

PegQ

The pegboard task (Annett, 1985) is a quantitative measure for relative hand skill that involves the measurement of time taken by an individual to move a row of ten pegs from one side of a board to the other using one hand. A relative hand skill measure (PegQ) for each subject is derived by calculating the difference between the average times for the left hand (L) and the right hand (R), divided by the average time for both hands combined:

$$PegQ = \frac{\mathrm{L - R}}{(L + R)/2}$$

A pegboard task similar to the Annet pegboard was performed in the ALSPAC sample as part of a battery of manual dexterity tests known as Movement ABC (Henderson *et al*., 2007). In this case there were 12 pegs on the table that the child picked up one-at-a-time and placed in a pegboard. After a trial practice, children performed the test once with each hand. The pegboard task was performed at 37 months and at 7 years of age (only 10 pegs used in the 37 months task) with a laterality quotient score derived for each as above (PegQ37 and PegQ7 respectively). One outlier (PegQ7 = 6.99) was removed from further analyses.

Handedness is a multifactorial trait and as such, in addition to the pegboard task, the ALSPAC dataset allows the measuring of fine motor performance and manual dexterity by assessing grip strength, the sorting of matches and marking squares. For the purpose of consistency, I derived a similar laterality quotient score for each quantitative measure using the equation above.

GripQ

Grip strength was assessed at 11 years of age using the Jamar hand dynamometer, which measures isometric strength in kilograms. The child was given a practice squeeze of the dynamometer (not recorded) and then encouraged to squeeze as long and as tightly as possible; the higher the reading, the stronger the grip. The child repeated the measurement with the left hand and two further measurements were taken with each hand, alternating sides to give three readings in total for each side. An average for all 3 readings was used to derive the GripQ score. Three outliers were removed from the Unaffected subgroup (GripQ = 7.36, 7.83, 8.2).

MarkQ

At 10 years of age the child was asked to make a short dash on a piece of paper which has a grid marked on it consisting of rows of 20 squares. They were asked to start at the top left hand side of the squared paper, working across it. When the first line is completed the child should start on the left side of the next row with the objective to see how many squares can be marked in 60 seconds. When the child has completed the test with the preferred hand, they change hands and turn the paper around and repeat. Unlike the other quantative measures, MarkQ has a negative mean since the majority are expected to score higher with their right hand than their left (i.e. Mean = -0.28, SD = ±0.27, N = 7,351 for all individuals with MarkQ data available). For ease of interpretation I transformed MarkQ to a positive mean by simply applying a factor of -1 to each phenotypic score.

SortQ

Both MarkQ and SortQ tests are a repetition of those used in the National Child Development Study, 1958 Cohort (Leask and Crow, 2001). At 10 years of age the child was asked to take matches one at a time out of a full match box (on their right) and transfer them directly to an empty match box (on their left) using one hand only, starting with the preferred hand. The children were allowed one practice go (not recorded) before they started the task. The tester timed how long it took for the child to transfer all the matches from one box to the other. The test was then repeated with the non-preferred hand. One outlier was removed from further analyses (SortQ = 5.46).

For further analyses, all quantitative measures were normalised to a mean of 0 and SD of 1. Thus, a positive score indicates superior relative right-hand skill while a negative score indicates superior relative left-hand skill.

2.3.2.2 Categorical measures of handedness and laterality

There were also several categorical measures of laterality available from the ALSPAC dataset, that is, measures that only provide a binary ‘left’ or ‘right’ result:

Hand7: The child was asked at 7 years of age which hand they wrote with

Hand10: At 10 years of age the child was asked prior to a computer session which was their dominant hand

Hand11: Prior to the GripQ session at 11 years of age the child was asked to specify which their dominant hand was

FOOT: At 37 months the child was invited to kick a ball at a set of skittles, with the foot used to kick noted. The ball was placed in front of the child at a midline position between the two feet. This task was repeated three times and an average taken.

EYE: Three coloured boxes were stood on a table at the child's eye height. When the child was first brought into the room they were asked to look through a hole in each box and asked a simple question about the contents in order to maintain interest. The examiner noted which eye was used on each occasion. The sample was only considered for further analyses if the child used the same eye 3 times.

All correlative analysis was conducted using SPSS v21. Association analysis was performed for an allelic model using Plink v1.9 (Purcell *et al.*, 2007).

2.4 Results

2.4.1 ALSPAC quantative measures

The PegQ7, PegQ37, SortQ and GripQ measures show a unimodal distribution in the Unaffected subgroup and are continuous and approximately normal, thus making them suitable quantitative phenotypes (Figures 2.2 - 2.6). There is no evidence for significant skewness or kurtosis (Tables 2.2 - 2.6). The MarkQ distribution displays a slight bimodal distribution and higher skewness (0.71, see Discussion section 2.5).

PegQ37

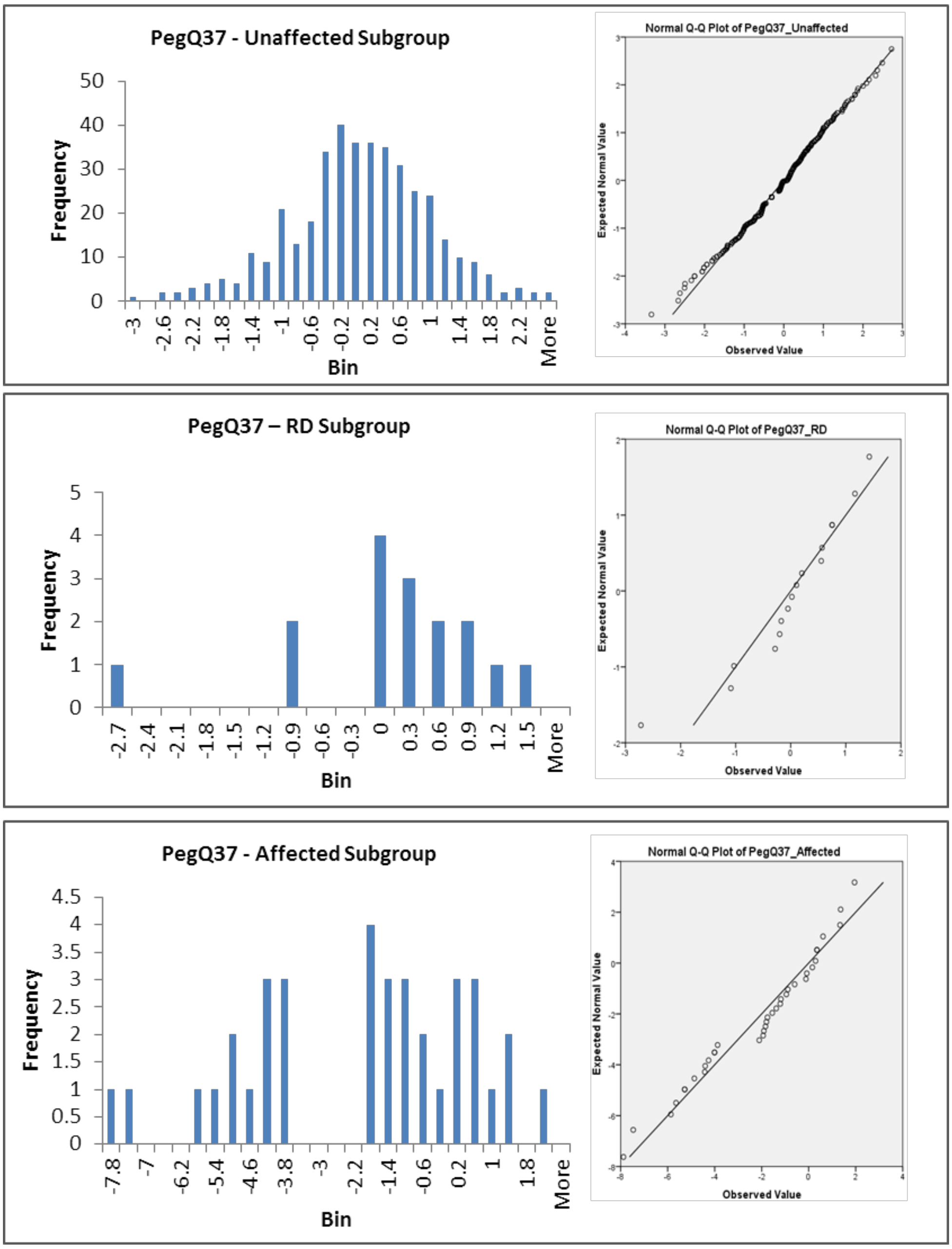


Figure 2.2 PegQ37 laterality measure distribution according to subgroup. Q-Q plots were generated with SPSS v21. RD subgroup includes those individuals showing comorbidity with ADHD and SLI. Affected subgroup includes all individuals with a neurodevelopmental disorder (ADHD, SLI or RD). Scores are normalised to a mean of 0 and a standard deviation of 1

Table 2.2 A summary of the skewness and kurtosis of the PegQ37 scores

Subgroup	*N*	*Skewness*	*Kurtosis*
Unaffected	402	-0.19	0.44
RD	16	-1.28	2.68
Affected	36	-0.46	-0.63

PegQ7

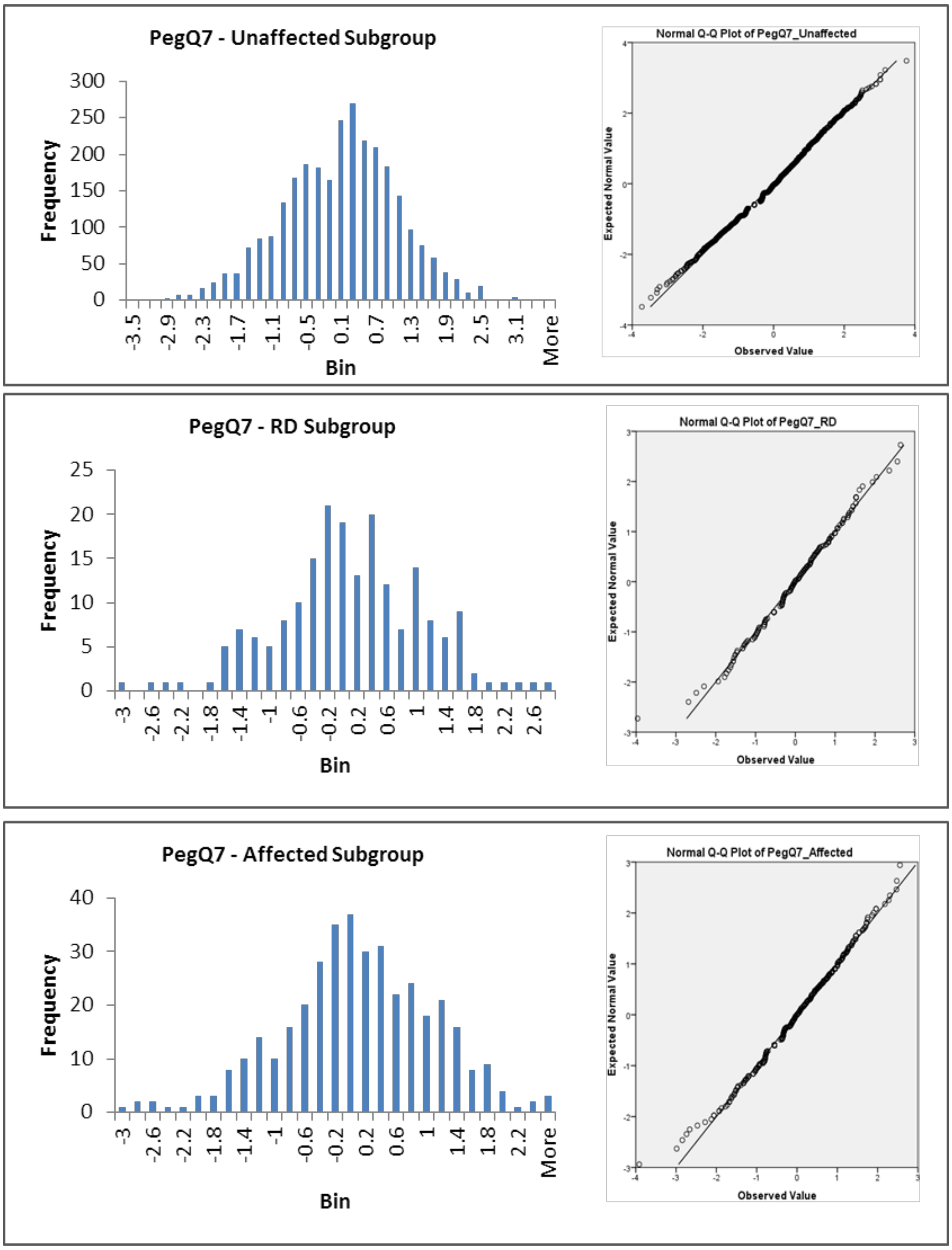


Figure 2.3 PegQ7 laterality measure distribution according to subgroup. Q-Q plots were generated with SPSS v21. RD subgroup includes those individuals showing comorbidity with ADHD and SLI. Affected subgroup includes all individuals with a neurodevelopmental disorder (ADHD, SLI or RD). Scores are normalised to a mean of 0 and a standard deviation of 1.

Table 2.3 A summary of the skewness and kurtosis of the PegQ7 scores

Subgroup	*N*	*Skewness*	*Kurtosis*
Unaffected	2825	-0.04	0.94
RD	197	-0.28	0.88
Affected	380	-0.25	0.45

SortQ

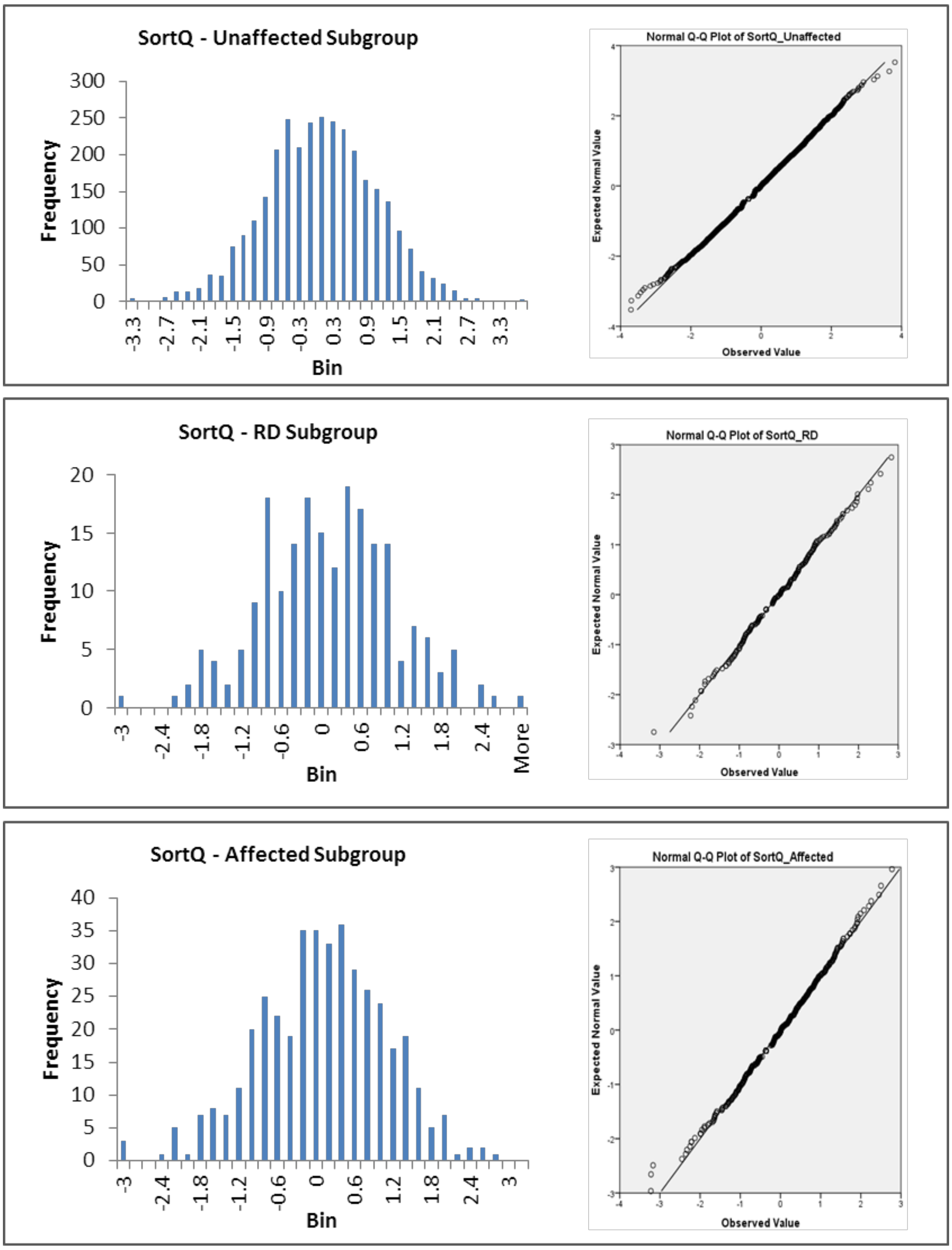


Figure 2.4 SortQ laterality measure distribution according to subgroup. Q-Q plots were generated with SPSS v21. RD subgroup includes those individuals showing comorbidity with ADHD and SLI. Affected subgroup includes all individuals with a neurodevelopmental disorder (ADHD, SLI or RD). Scores are normalised to a mean of 0 and a standard deviation of 1

Table 2.4 A summary of the skewness and kurtosis of the SortQ scores

Subgroup	*N*	*Skewness*	*Kurtosis*
Unaffected	3147	-0.01	0.42
RD	209	0.02	0.12
Affected	412	-0.2	0.17

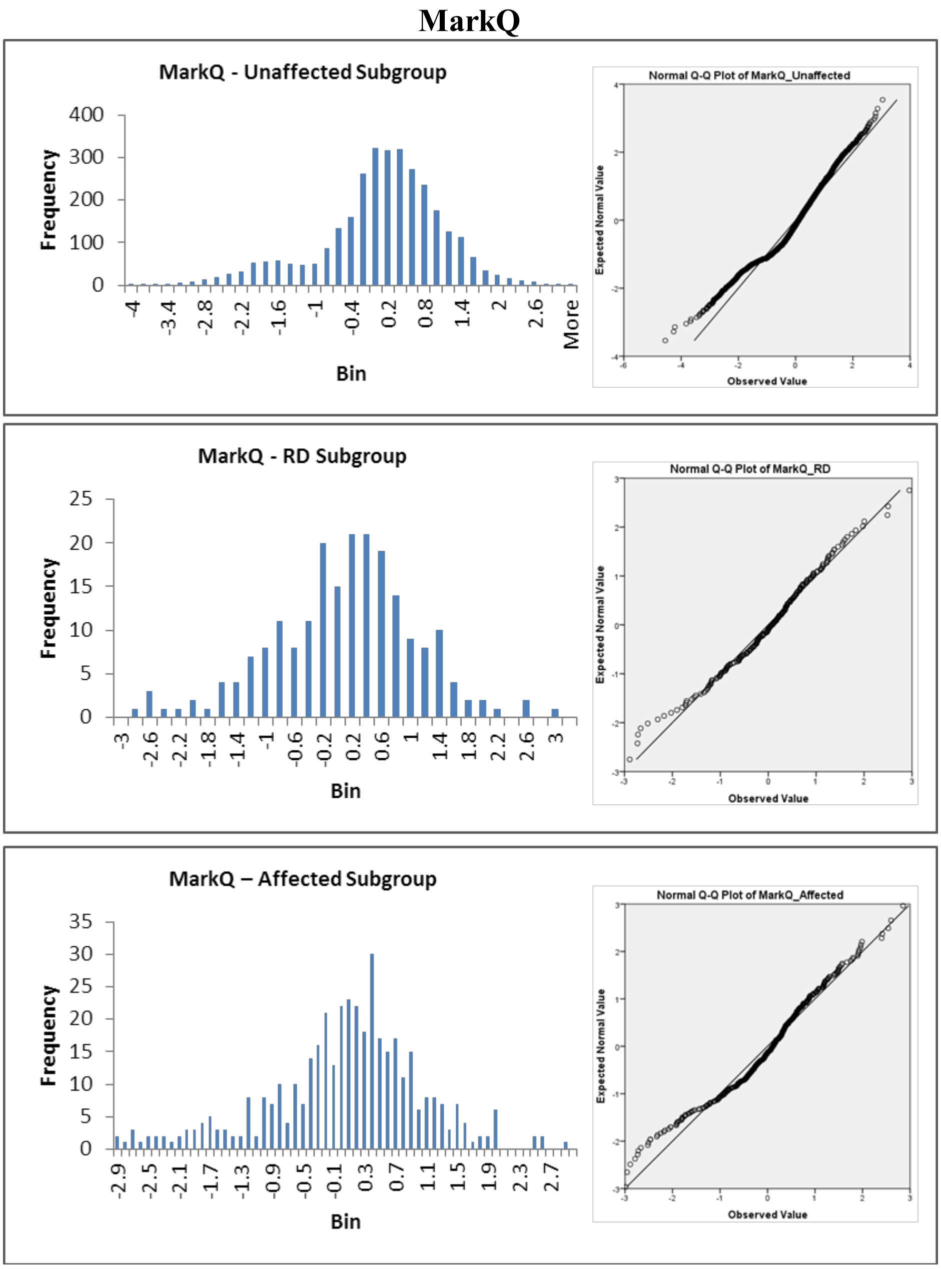


Figure 2.5 MarkQ laterality measure distribution according to subgroup. Note the secondary peak in Affected and Unaffected subgroups represents a subpopulation of left-handers. See Figure 2.7 for separated distributions. Q-Q plots were generated with SPSS v21. RD subgroup includes those individuals showing comorbidity with ADHD and SLI. Affected subgroup includes all individuals with a neurodevelopmental disorder (ADHD, SLI or RD). Scores are normalised to a mean of 0 and a standard deviation of 1.

Table 2.5 A summary of the skewness and kurtosis of the MarkQ scores

Subgroup	*N*	*Skewness*	*Kurtosis*
Unaffected	3115	0.71	1.03
RD	211	0.29	0.55
Affected	410	0.47	0.71

GripQ

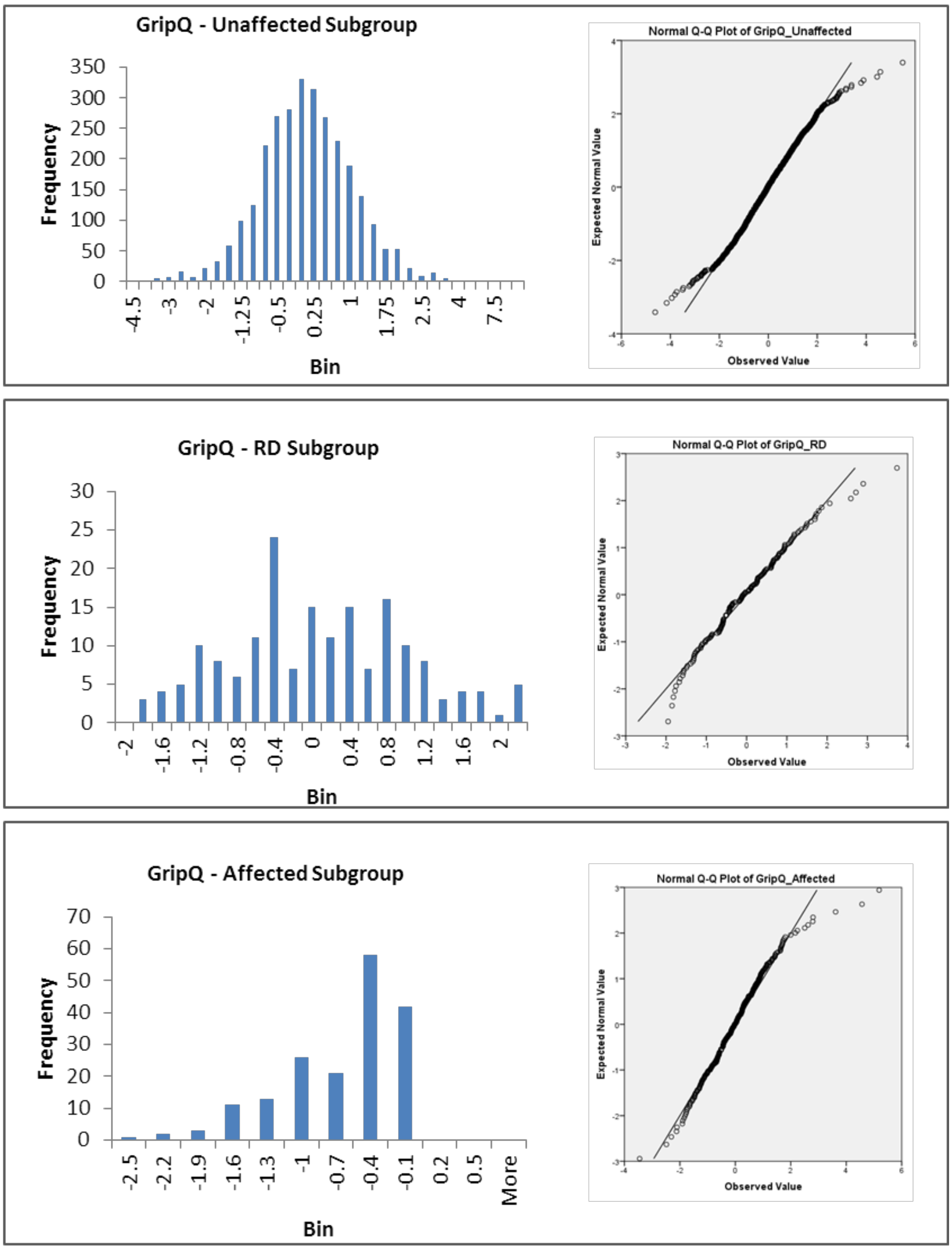


Figure 2.6 GripQ laterality measure distribution according to subgroup. Q-Q plots were generated with SPSS v21. RD subgroup includes those individuals showing comorbidity with ADHD and SLI. Affected subgroup includes all individuals with a neurodevelopmental disorder (ADHD, SLI or RD). Scores are normalised to a mean of 0 and a standard deviation of 1.

Table 2.6 A summary of the skewness and kurtosis of the GripQ scores

Subgroup	*N*	*Skewness*	*Kurtosis*
Unaffected	2859	0.54	4.82
RD	177	0.54	0.65
Affected	382	0.69	3.03

2.4.2 Categorical measures for handedness

The ALSPAC dataset provides several measures of categorical hand preference recorded over a four year period (age 7-11 years). Three categorical measures for self-reported handedness (Hand7, Hand10 and Hand11) were tested for correlation within three subgroups; Unaffected, RD and Affected. Table 2.7 indicates all three measures to be highly correlated with self-reported handedness a stable measure over the time period, irrespective of subgroup status. As the measure with the most phenotype data available, Hand7 (N = 8,084) was selected as a stable self-reported measure of hand preference for subsequent analyses. Any individuals (N = 46) that displayed variation in self-reported hand preference were removed from further analyses.

Table 2.7 Pearson's correlation between categorical measures for self-reported handedness at 7 (Hand7), 10 (Hand10) and 11 (Hand11) years of age in the Unaffected (*italics),* RD (underlined) and Affected (**bold**) subgroups. All correlations are significant at the 0.01 level (2-tailed).

	Hand10	Hand7	Hand11
Hand10	1	*.95, (N = 2807)*	*.95 (N = 2807)*
		.93 (N = 189)	.95 (N = 189)
		.91 (N = 365)	**.95 (N = 365)**
Hand7		1	*.96 (N = 2807)*
			.98 (N = 189)
			.96 (N = 365)
Hand11			1

2.4.3 Correlation between quantative measures

To reduce confounding variables such as the subjective nature inherent in self-reporting measures, researchers use performance-based measures which assess the differences between the two hands on a given task. An analysis of five such quantitative measures of relative hand skill in the ALSPAC dataset (GripQ, SortQ, MarkQ, PegQ37 and PegQ7, Table 2.8) indicates a weak correlation between each measure across all subgroups. There is a slight increase in correlation between the measures in the Affected subgroup ($r = 0.07 - 0.25$) compared to the Unaffected subgroup ($r = 0.06 - 0.2$). The highest correlated pair of measures in all 3 subgroups was between PegQ37 and PegQ7 (Unaffected, RD and Affected subgroups were $r = 0.15$, $r = 0.58$ and $r = 0.22$ respectively).

The current analysis serves to determine which of the five performance-based measures produces the strongest prediction of hand preference as measured by the self-reporting Hand7 categorical measure. To this effect, I developed a predictive model of handedness using stepwise multiple linear regression analysis (MLRA) using Hand7 as the dependent variable and four hand performance measures as independent variables. Note the PegQ37 measure was removed from the analysis due to the reduced number of data points available (N = 402) since models which impute missing data have been shown to cause subsequent relationships to be over-identified (Allison, 2002).

An examination of the subsequent correlation matrix (grey bars, Table 2.8) reveals MarkQ to have the highest correlation with the Hand7 measure for hand preference across all subgroups. The Pearson coefficient indicates a consistent inverse correlation since all performance measures have a positive mean score while the Hand7 hand preference variable is categorical (1= right-handed 2= left-handed); as the Hand7 dependent variable increases to ‘2’ (signifying left-handedness) then the relevant predictor will undergo an equivalent shift to the left on the normalised distribution curve. All performance measures, except the SortQ task in the Unaffected cohort, contributed to the prediction of hand preference. In the Unaffected cohort the MLRA model (Hand7= 1.11 - 0.19(MarkQ) -0.05(PegQ7) -0.03(GripQ)) produced an explained variance of 48% (adjusted R^2= 0.48). In the RD cohort the model (Hand7= 1.12 - 0.148(MarkQ) - 0.052(PegQ7) + 0.007(GripQ) -.022(SortQ)) produced an adjusted R^2= 0.22 while in the Affected cohort a model (Hand7= 1.13 - 0.168(MarkQ) – 0.057(PegQ7) -0.017(GripQ) – 0.019(SortQ)) produced an adjusted R^2= 0.35.

Table 2.8 Pearson's correlation between quantitative measures of relative hand skill and dexterity in the ALSPAC dataset subgroups Unaffected (A), RD (B) and Affected (C). The grey box in each table indicates the Pearson's correlation between the dependent variable Hand7 (hand preference) and multiple measures of hand performance (independent variables) when using MLRA. Note PegQ37 was not included due to missing data.

(A) Unaffected Subgroup (N = 2,353)

	GripQ	*SortQ*	*MarkQ*	*PegQ37*	*PegQ7*
GripQ	1	.06**	.2**	.043	.14**
SortQ		1	.15**	.12*	.13**
MarkQ			1	.13*	.3**
PegQ37				1	.15**
PegQ7					1
Hand7	-.22**	-.12**	-.66**	-	-.35**

(B) RD Subgroup (N = 144)

	GripQ	*SortQ*	*MarkQ*	*PegQ37*	*PegQ7*
GripQ	1	-.04	.16*	.07	.16
SortQ		1	.04	.29	.03
MarkQ			1	.09	.19**
PegQ37				1	.58*
PegQ7					1
Hand7	-.072	-.06	-.47**	-	-.264**

(C) Affected Subgroup (N = 307)

	GripQ	*SortQ*	*MarkQ*	*PegQ37*	*PegQ7*
GripQ	1	.07	.1	.25	.2**
SortQ		1	.08	.07	-.01
MarkQ			1	-.22	.2**
PegQ37				1	.22
PegQ7					1
Hand7	-.16*	-.1	-.57**	-	-.3**

Correlation is significant at the **0.01 level (2-tailed) and *0.05 level (2-tailed).

2.4.4 Ambidexterity

Deriving a laterality quotient for the five quantitative measures also enabled the recording of which individuals displayed ambidexterity, that is, the state of being equally adept in the use of both hands at a particular task. A phenotypic association between ambidexterity and neurodevelopmental disorders has been reported previously in RD (Tonnessen *et al.*, 1993), developmental disorder (Goez and Zelnik, 2008) and schizophrenia (Barrantes-Vidal *et al.*, 2013, Tran *et al.*, 2015), though several earlier studies did not observe this association (Pennington *et al.*, 1987, Gilger *et al.*, 1992). My data do not support a straightforward relationship between relative hand skill and neurodevelopmental subgroups (Table 2.9).

Table 2.9 Ambidexterity recorded in the quantitative measures for hand performance. Sample sizes (N) included in brackets. PegQ37 is not included due to low sample sizes (RD, N=16). Ambidexterity was recorded as an individual with a normalised performance score within the range ± 0.05

Measure	*All (N)*	*Unaffected (N)*	*RD (N)*	*Affected (N)*
GripQ	4.3% (288/6,686)	4.3% (123/2,862)	1.7% (3/178)	4.2% (16/383)
SortQ	4.7% (349/7,383)	5.1% (143/2,793)	2.9% (6/209)	4.4% (18/412)
MarkQ	4.7% (345/7,360)	4.8% (151/3,117)	5.2% (11/211)	3.6% (15/411)
PegQ7	3.2% (221/7,002)	3.6% (102/2,827)	3.0% (6/197)	3.4% (13/380)

2.4.5 Hand, Foot and Eye Correlation

Questionnaires such as the Waterloo Handedness Questionnaire (WHQ, Steenhuis *et al.* (1990)) and the Edinburgh Handedness Inventory (Oldfield, 1971) typically focus on hand preference and fail to explore dominance for foot or eye, even though footedness has been reported as a more robust predictor of language lateralisation than handedness. In a limited study, Elias and Bryden (1998) demonstrated footedness to be a far better predictor of language lateralisation than handedness (N = 32, P < .001). Similarly, in a cohort of 37 epileptic candidates for temporal lobe resection, Watson *et al.* (1998) found in Wada testing neither handedness alone nor the interaction of handedness and footedness contributed substantially more than footedness alone to the prediction of language laterality. Other research also suggests that non-right-footedness plays a role in schizophrenia (Tran *et al.*, 2015, Schiffman *et al.*, 2005), though these findings have been inconsistent (Nicholls *et al.*, 2005, Asai *et al.*, 2011). Analysis of the Hand7, Foot and Eye categorical measures of laterality across all three subgroups (Table 2.10) shows hand preference is only weakly correlated with foot preference (Unaffected r = 0.28) however both RD and Affected subgroups were too underpowered to reliably perform MLRA on due to low sample size (N = 18 and 33). In total, just 10.4% of the variance as measured by Hand7 can be explained when using Foot and Eye variables as predictors in the Unaffected subgroup (MLRA model Hand7 = 0.71 + 0.22(Foot) + 0.11(Eye)). My data support previous findings (Teng *et al.*, 1976, Hoosain, 1990), showing the correlation between handedness and eyedness (r = 0.22) to be weaker than the correlation between handedness and footedness (r = 0.28, Table 2.10).

Table 2.10 Pearson's correlation between categorical measures for self-reported handedness, footedness and eyedness in the Unaffected (*italics*), RD (underlined) and Affected (**bold**) subgroups. ** Correlation is significant at the 0.01 level (2-tailed).

	Hand7	Foot	Eye
Hand7	1	*.28** (N = 375)* -.09 (N = 18) **-.09 (N = 33)**	*.22** (N = 375)* -.19 (N = 18) **.21 (N = 33)**
Foot		1	*.2** (N = 375)* .081 (N = 18) **-.11 (N = 33)**
Eye			1

2.4.6 Association analysis of rs11855415

Finally, I investigated if the significant association between SNP rs11855415 and handedness recorded previously (P = 4.7×10^{-7}, Scerri *et al.* (2011a); P= 6.96×10^{-8}, Brandler *et al.* (2013)) was exclusive to a pegboard measure in an RD subgroup or whether this finding could be extended to other measures of relative hand skill across multiple subgroups. The only measure showing nominally significant association with rs11855415 was PegQ7 in the RD (P= 0.002) and Affected (P= 0.02) subgroups as reported in Table 2.11. These results act as a benchmark in confirming this analysis was performed on a very similar subset to previous GWA studies i.e. Scerri *et al.* (2011a) in their analysis of an RD cohort (N = 368, P= 0.033, β= 0.19). Furthermore, individuals with the minor (derived) 'A' allele of rs11855415 were shown to have significantly greater relative right-hand skill compared with those carrying the major 'T' (ancestral) allele. The mean effect size of each copy of the minor allele is 0.28 standard deviations (SD) to the positive (right-handed) end of the PegQ7 distribution (see 'β' in Table 2.11) in the Affected subgroup and β= 0.29 in the RD subgroup. In the Unaffected subgroup (N = 2,597), I did not detect significant association between rs11855415 and relative hand skill (P = 0.32, β = −0.034).

Table 2.11 Association analysis results for the PCSK6 genetic variant rs11855415 and hand performance measures for relative hand skill and dexterity in multiple subgroups. Each result represents a P-value and subgroup size. Further details of the nominally significant PegQ7 associations are included below the main table. β is the mean effect size of each copy of the minor allele measured in standard deviations. SE is the standard error. R^2, the regression r-squared, estimates the proportion of the phenotypic variation that is explained by the rs11855415 marker. Highlighted in bold are the nominally significant associations between rs11855415 and PegQ7 in the RD and Affected subgroups.

	Group			
Measure	All	Unaffected	RD	Affected
GripQ	0.31 (N=5,956)	0.46 (N=2,636)	0.17 (N=167)	0.11 (N=346)
PegQ37	0.33 (N=568)	0.91 (N=378)	0.88 (N=15)	0.38 (N=33)
PegQ7	0.36 (N=6,101)	0.32 (N=2,597)	**0.02 (N=183)**	**0.002 (N=341)**
MarkQ	0.87 (N=6,498)	0.69 (N=2,860)	0.15 (N=195)	0.13 (N=369)
SortQ	0.52 (N=1,955)	0.37 (N=2,890)	0.35 (N=193)	0.94 (N=370)

Measure	*Subgroup (N)*	*P-value*	*β*	*SE*	*R^2*
PegQ7	RD (183)	0.02	0.29	0.12	0.02
PegQ7	Affected (326)	0.002	0.28	0.28	0.09

As defined in the Methods section, the RD subgroup in Table 2.11consists of individuals with either pure RD or those with RD and comorbidity for SLI and/or ADHD while the Affected subgroup contains all RD, SLI and ADHD individuals (and combinations of comorbidities). The significant level of association found between PegQ7 and rs11855415 in the Affected subgroup (P = 0.002, N = 341, Table 2.11) might therefore be as a result of the RD cohort within the Affected subgroup driving this association. However, the further division of the Affected subgroup (N = 341) in to its constituent cohorts of pure SLI (N = 154), RD (N = 148) and ADHD (N = 24) resulted in no one single clinical cohort displaying significant association between rs11855415 and the PegQ7 measure (Table 2.12). When combining both RD and ADHD cohorts from the Affected subgroup there is a significant association (P = 0.02, N = 161) between rs11855415 and the PegQ7 measure (Table 2.12).

Table 2.12 Association analysis results for the PCSK6 genetic variant rs11855415 and the PegQ7hand performance measure for relative hand skill in the Affected subgroup. The Affected subgroup was divided in to its constituent cohorts according to disorder. Each result represents a P-value and subgroup size. β is the mean effect size of each copy of the minor allele measured in standard deviations. SE is the standard error. R^2, the regression r-squared, estimates the proportion of the phenotypic variation that is explained by the rs11855415 marker.

Measure	*Subgroup (N)*	*P-value*	β	*SE*	R^2
PegQ7	ADHD (20)	0.08	0.61	0.33	0.16
PegQ7	RD (141)	0.15	0.19	0.13	0.01
PegQ7	SLI (135)	0.32	0.16	0.16	0.01
PegQ7	SLI & RD (37)*	0.08	0.6	0.32	0.08
PegQ7	SLI + ADHD (155)†	0.06	0.27	0.15	0.02
PegQ7	RD + ADHD (161) †	**0.02**	0.29	0.12	0.03
PegQ7	SLI + RD (276) †	0.10	0.17	0.10	0.01

* Indicates individuals displaying comorbidity for both disorders rather than combined subgroups

† Indicates the sum of combining subgroups and not co-morbidity

In conclusion, a test for association between rs11855415 and the categorical measures for laterality was also conducted using the Hand7, Hand10, Hand11, Foot and Eye measures (Table 2.13). Although Hand7 does not display nominal significance, Hand7 and Hand11 are highly correlated in the RD and Affected subgroups (r = 0.975 and 0.964 respectively, see Table 2.7 previously) where Hand11 does show nominal significance ($P \leq 0.02$). In this basic allelic model the minor A allele of rs11855415 has a low-mid effect size in both RD and Affected subgroups (OR = 0.29 and 0.47 respectively). Hand11 carries an effect in the same direction as the previous association analysis between rs11855415 and PegQ7 ($\beta = 0.28 - 0.29$, Table 2.13). That is, a shift towards right-handedness for each additional minor A allele.

Table 2.13 Association analysis results for the PCSK6 genetic variant rs11855415 and categorical measures for handedness, footedness and eyedness in multiple subgroups. Each result represents a P-value and subgroup size for that association tested. Further details of the nominally significant Hand11 associations are included below the main table. OR is the odds ratio, Chi is the chi-square basic allelic test and F is the frequency of the minor A allele for rs11855415.

	Group			
Measure	All	Unaffected	RD	Affected
Hand7	0.17 (N=8,084)	0.27 (N=3,284)	0.08 (N=224)	0.12 (N=441)
Hand10	0.70 (N=7,498)	0.52 (N=3,177)	0.09 (N=213)	0.11 (N=418)
Hand11	0.89 (N=6,688)	0.52 (N=2,870)	**0.01 (N=201)**	**0.02 (N=382)**
Foot	0.80 (N=1,028)	0.36 (N=417)	0.7 (N=19)	0.36 (N=38)
Eye	0.72 (N=7,438)	0.07 (N=2,982)	0.2 (N=206)	0.35 (N=408)

Measure	*Subgroup (N)*	*P-value*	*Chi*	*OR*	*F*
Hand11	RD (201)	0.01	5.97	0.29	0.24
Hand11	Affected (382)	0.02	5.27	0.47	0.23

2.5 Discussion

Quantative measures offer several advantages over self-reporting hand preference questionnaires (e.g. elimination of both language reliance and the subjective nature of self-reporting), though such hand performance measures display their own inherent limitations. An analysis of five quantitative measures of relative hand skill in the ALSPAC dataset (GripQ, SortQ, MarkQ, PegQ37 and PegQ7) demonstrated poor intra-correlation (Table 2.8), suggesting each unimanual task likely records only one aspect of manual performance abilities (e.g., speed, accuracy or dexterity). As such, motor behaviour is not a unitary trait and should be defined using multiple measures of both hand performance and hand preference.

Through the use of MLRA modelling I examined which of the five hand performance measures might best predict hand preference as measured by Hand7. The use of MLRA allowed the decomposing of each predictor of hand preference such that each Pearson correlation represented the unique predictive capacity of each individual variable (Table

2.8). MarkQ is the majority contributor to the regression equation followed by PegQ7 and GripQ with an adjusted $R^2 = 0.48$, or in other words, 48% of the variability in Hand7 can be accounted for by just three of the measures of relative hand skill in the Unaffected subgroup ($P < 0.001$, N = 2,353). Removal of the GripQ independent variable only reduced the predicted variance by 1% to 47%. Such a minor contribution to the variance is to be expected since grip strength has been consistently reported to be weakly lateralised (Borod *et al.*, 1984, Steenhuis *et al.*, 1990). The SortQ measure was an excluded predictor from the model. Though the adjusted R^2 value may not appear high (0.48), since all predictors are statistically significant ($P < 0.001$), one can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.

The correlation between hand performance and hand preference (Hand7) also depends strongly on the task being measured. Correlation ranges from r = 0.12 (SortQ) to r = 0.66 (MarkQ) in the Unaffected subgroup (Table 2.8). This suggests a positive relationship between complexity of task and correlation score since both MarkQ and PegQ require skilled and learned manipulation of pen and peg respectively in comparison to the unimodal nature of the SortQ measure.

Interestingly, significantly different distributions exist for hand performance and hand preference (Nicholls *et al.*, 2010). Hand preference yields a bimodal (or J-shaped) distribution with a large number of strongly right-handed individuals, a smaller number of strongly left-handed individuals, and few individuals in between (Ocklenburg *et al.*, 2014). That MarkQ is the best predictor of hand preference in this data likely arises from the fact that the distribution of scores appears bimodal (Figure 2.5); tasks producing bimodal distributions provide clear support for distinct handedness groups (left vs right) whereas unimodal distributions suggest a continuum or a series of groups (strong , weak and mixed) might best describe handedness. A demonstration of this can be seen in Table 2.5 which indicates the distribution of MarkQ scores to be markedly skewed in the Unaffected subgroup (skewness = 0.71). When the subjects are subdivided into left- and right-handers, the skewness is significantly reduced (right hand skew= 0.1, N = 2,082; left hand skew = -0.3, N = 267). In fact, data from both right and left-handers can be adequately described by normal distributions of differing means but having the same variance (Figure 2.7: right mean = 0.24, SD = ±0.74; left mean = -1.8, SD = ±0.75). A similar trend exists when repeating this left-right division in the other subgroups. This would support the notion that left and right-handers represent two distinct subsamples in a population as proposed by McManus (2002).

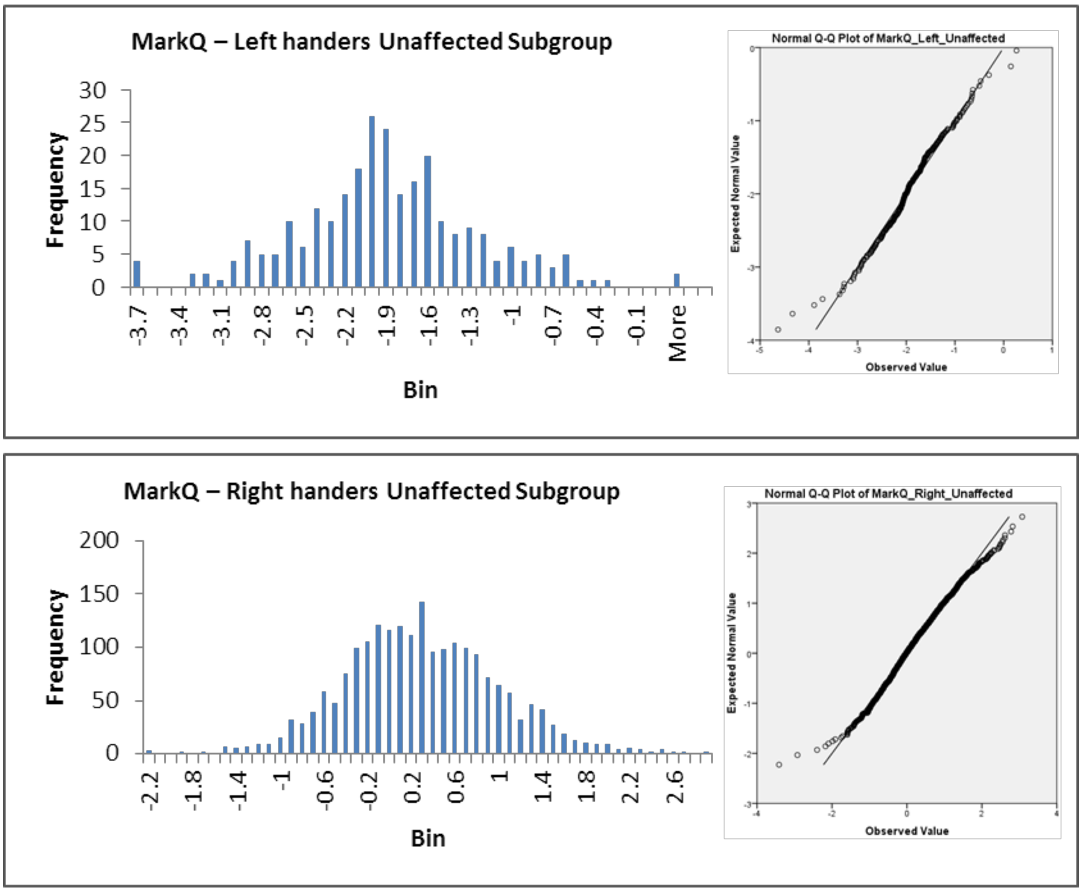


Figure 2.7 MarkQ distribution in the Unaffected subgroup. This quantative measure of relative hand skill has been subdivided in to left (top) and right (bottom) handers according to the self-reporting hand preference measure Hand7. Scores are normalised to a mean of 0 and a standard deviation of 1. Q-Q plots were generated with SPSS v21. For the original combined distribution of left and right handers in the Unaffected subgroup see Figure 2.5

The distribution of hand performance data seems to be task-dependent to a large extent, with some tasks clearly showing more bimodal distributions (e.g. handwriting (Provins *et al*., 1982); dots (Tapley and Bryden, 1985) and punching holes (Annett, 1992). In contrast, hand performance measured with the peg board task typically shows a unimodal distribution (Annett and Kilshaw, 1984), though the argument exists that the peg board data are also bimodal since a smaller distribution of left-handers might be concealed within the tail of the larger distribution of the right-handers (McManus, 1985b). Unlike the MarkQ data which appears to represent two distinct subsamples in a population, my pegboard task data for the Unaffected subgroup (N = 2,825, skewness = -0.04, kurtosis = 0.94) demonstrate an unskewed platykurtic distribution with a wider spread of values around the mean indicative of a classic unimodal distribution as a visual inspection confirms (Figure 2.3).

One limitation of my interpretation of the hand performance data in this chapter was the use of laterality quotients (Oldfield, 1971). Such laterality quotients are only sensitive to the direction of handedness whereas alternative methodologies such as a laterality score (LS) have been shown to be sensitive to both the degree and direction of handedness (p. 169, Mandal *et al.* (2000)).

As discussed in the chapter introduction, it is of interest to see whether the PCSK6 SNP rs11855415 which Brandler *et al.* (2013) and Scerri *et al.* (2011a) found to be significantly associated with PegQ also shows association with other laterality measures in both clinical and general population cohorts. When testing for association between the SNP rs11855415 and multiple quantative measures of relative hand skill (see Table 2.11), only PegQ7 showed nominal significance in the Affected (P = 0.002, N = 341) and RD subgroups (P = 0.02, N = 183). This finding results in several points worthy of discussion, namely:

(1) The specificity of this finding - that only PegQ is significantly associated with rs11855415 - may relate to the notion that handedness is a multivariable trait which can be viewed as a multiple of separate phenotypes each representing hand strength, speed, dexterity etc. Since each of the performance measures record different aspects of handedness, this association might just represent a relationship between a genetic variation and one specific phenotype of handedness. Alternatively, the peg-board task might simply provide a measure (PegQ) which represents a broader aggregate of handedness aspects (grip, manual dexterity, spatial awareness etc.) than the other hand-performance measures. For example, the highest correlation in all 3 subgroups between a PegQ measure and a non-peg measure for relative hand skill was between GripQ and

PegQ7 (RD and Affected subgroups were r = 0.16 and r = 0.2 respectively), while the equivalent values between GripQ and MarkQ show an overall weaker correlation (r = 0.16 and r = 0.1 for the RD and Affected subgroups respectively).

(2) Why a significant association between rs11855415 and PegQ was observed in only the RD and Affected subgroups and not the general population remains to be clarified. One possible reason why both Brandler *et al.* (2013) and Scerri *et al.* (2011a) report significant association specific to the RD subgroup might be that such a finding represents a false positive; an artefact as a result of the slight variation in how the peg-board task was conducted between the RD and general population cohorts (see Brandler *et al.* (2013) for peg-board task details). An analysis of Table 2.3 shows a discrepancy in skewness between the Unaffected (-0.04, N = 2825), RD (-0.28, N = 197) and Affected (-0.25, N = 380) subgroups though none record a skewness score which indicates the data is significantly skewed (≤-1 or ≥1). As such, all subgroups are considered to have a normal distribution. A visual inspection of Figure 2.3 also fails to report any substantial difference in the tails of the Unaffected, RD and Affected subgroup distributions, indicating a similar spread of extreme lateralised individuals regardless of the subgroup. In any case, the observation that the increase in relative right-hand skill associated with rs11855415 in PSCK6 is specific to the RD and Affected subgroups may represent epistatic interaction between PCSK6 and RD susceptibility genes.

(3) A further division of the Affected subgroup in to its constituent disorder cohorts of pure SLI, ADHD and RD (Table 2.12) showed no one disorder cohort to display significant association between rs11855415 and PegQ7. Therefore from these data at least, SLI individuals can be discounted from driving the signal we see for significant association between rs11855415 and PegQ7 in the Affected subgroup. Since the RD Subgroup in Table 2.11 displayed significant association between rs11855415 and PegQ7 (P=0.02), it would therefore appear the signal in the Affected subgroup is driven by the pure RD cohort with the possible addition of pure ADHD individuals and individuals co-morbid for both disorders (ADHD+RD, N = 5, Table 2.1). Although the sample sizes are small (e.g. ADHD, N = 20), the association in Table 2.12 is in the same allelic trend i.e. individuals with the minor 'A' allele of rs11855415 have significantly

greater relative right-hand skill compared with those carrying the major ‘T’ (ancestral) allele. This differs to the allelic trend of the general population which displays a trend towards reduced laterality of hand skill for the minor allele ($\beta = -0.03$, $N = 2{,}597$).

If the ADHD cohort is influencing the significant association in the Affected subgroup, as the trend in Table 2.12 suggests, then the argument exists that PegQ data should be collected in children across a range of neurodevelopmental disorders rather than just in dyslexic individuals. It would be of great interest to see what effect an increase in clinical sample numbers might have on such trends and what ramifications this has in defining the link between handedness, language and brain asymmetries.

In conclusion, this chapter supports previous GWAS findings that the SNP rs11855415 shows nominally significant association with a performance measure for relative hand skill in dyslexic (RD) and affected (RD, SLI, ADHD) subgroups. The remainder of this thesis will investigate this association by first defining the extent of the PCSK6 locus in which the SNP is found (Chapter 3 & 4) and thereafter performing a functional analysis of the genetic variants within this region of association (Chapter 5).

3 *In silico* analysis of the PCSK6 locus

3.1 Abstract

Chapter 1 introduced two genome wide association (GWA) studies which have reported significant genetic association implicating a PCSK6 locus with handedness (Brandler *et al.,* 2013, Scerri *et al.,* 2011a). Due to the nature of their design, GWA studies typically return candidate Single Nucleotide Polymorphisms (SNPs) which may or may not be the causal genetic variant driving the reported association. We are thus required to define the PCSK6 locus of interest and determine the list of genetic variants within it for further study via functional genomic analyses. These analyses are performed using a combination of linkage disequilibrium (LD) mapping, evolutionary conserved sequence identification and the prioritisation of putative functional SNPs and regulatory elements by the integration of bioinformatics methods including interrogation of the Encyclopaedia of DNA Elements (ENCODE) project datasets and promoter database prediction software. The data presented in this chapter provide evidence that the functional variant(s) contributing to previous GWAS signals at the PCSK6 locus reside within a 12.7 kb sequence which is predicted to contain a 1.8 kb secondary promoter.

3.2 Introduction

GWA studies indicate associations between specific genomic loci and normal or pathological traits via a set of genetic marker SNPs designed to tag all known common variants in the genome in a hypothesis-neutral framework (Hirschhorn and Daly, 2005). However, pinpointing the strongest candidate causal variants from GWAS-associated loci and revealing the biological relevance of these associations remains a significant challenge. So far, there has only been limited scope for functional investigation in to the many biological hypotheses turned up by GWA studies (for example studies see Vernes *et al*. (2008) and Massinen *et al*. (2009)). This is the first study to perform a functional analysis of a locus previously detected to be significantly associated with a measure of handedness.

Post-GWAS functional follow-up studies typically involve the selection of variants that are the most likely to be the causal SNP underlying GWAS results. Several approaches have evolved over time – from a simple selection of markers to follow up based on their ranked marginal association test statistics to a more elegant approximate Bayesian procedure, based on posterior probabilities that each marker is the causal marker (Thompson *et al.*, 2013). This chapter uses an approach that involves linkage disequilibrium (LD), the notion that alleles at neighbouring loci tend to be co-inherited, or in other words, a non-random association of alleles that is influenced by selection, the rate of recombination and population structure among other factors (Ardlie *et al.*, 2002). As a starting point, a complete catalogue of all variants at the associated locus that are in LD is required.

The public availability of the HapMap project data (http://hapmap.ncbi.nlm.nih.gov/) allows access to sufficiently comprehensive genotyping data from which to compile such a list of genetic variants; 108,502 quality controlled SNPs are available alone for chromosome 15 on which PCSK6 is positioned. All previous association studies relevant to this project (Arning *et al.*, 2013, Brandler *et al.,* 2013, Scerri *et al.,* 2011a) were conducted on cohorts of white European ethnicity and as such all further analysis and discussion is consigned to this population.

The challenge of identifying functional variants from GWAS-identified loci is further complicated by variant types; while coding variants are easier to detect and annotate since they have a direct effect on protein structure and subsequent function, a common feature of GWAS findings is the vast majority of reported markers lie in intergenic or intronic regions (~88%, Welter *et al.* (2014)). Noncoding variants usually reside in regions whose annotations are characteristic of regulatory sites, such as sequence conservation, transcription factor binding sites (TFBSs), epigenetic markings etc. and are thought to influence gene expression through either transcriptional, post-transcriptional or post-translational mechanisms (Ward and Kellis, 2012). The process of transcription is a complex one which begins when RNA polymerase II assembles at a gene promoter with the basal transcription machinery and begins to catalyse production of the complementary RNA (Alberts *et al.*, 2008). Polymerases are large enzymes composed of almost a dozen subunits, typically complexed with numerous transcription factors (TFs) by associating with regulatory sequences. The accessibility of such TFs is largely dependent upon chromatin structural changes which are controlled by epigenetic histone modifications, such as acetylation and methylation (Strahl and Allis, 2000).

The advent of large-scale studies designed to identify regulatory elements in humans such as the ENCODE project (Consortium *et al.*, 2007) has vastly improved our ability to annotate putative *cis*-regulatory variants, thereby facilitating the design of downstream functional analyses and testable hypotheses. For example, the ENCODE project includes chromatin immunoprecipitation with sequencing (ChIP-SEQ) data, a method used to analyse protein interactions with DNA. ENCODE tracks are useful in displaying ChIP-SEQ epigenetic data, since any non-coding variants found within regions marked with histone modifications are useful in reliably marking regulatory regions (Visel *et al.*, 2009).

If regulatory elements are shown to contain non-coding variants, then one can use comparative genomics to infer conserved noncoding sequences (CNSs) perform functions, which places varying degrees of constraint on their evolution (Jegga and Aronow, 2001). In other words, such CNSs can be informative since functional sequences are more likely to be retained through evolution, compared with non-functional sequences (Haudry *et al.*, 2013, Burgess and Freeling, 2014), as demonstrated by Thomas *et al.* (2003) who used conservation of bases across a region to successfully identify functional elements. In keeping with a definition used previously (Duret *et al.*, 1993, Loots *et al.*, 2000, Dermitzakis *et al.*, 2002), in my analysis conserved sequences are defined as having 70% identity over at least 100 bp of ungapped alignment (which is above the average rate of neutral conservation) between human and orthologous sequences from other species.

In conclusion, in this chapter I define the extent of a PCSK6 locus previously identified to be associated with handedness. Additionally I provide a catalogue of candidate variants that fall within this region and use comparative genomics and ENCODE project data to identify CNSs and regulatory elements which contain genetic variants that may ultimately affect PCSK6 expression.

3.3 Methods

3.3.1 Defining the region of Linkage Disequilibrium

Haploview v4.2 (Barrett *et al.*, 2005) was used to visually inspect the haplotype structure and frequency of the downloaded genotype data for the canonical PCSK6 gene (NM_002570.4) in the HapMap CEPH population (Rel28 PhaseII+III Aug10 NCBI36, dbSNP126). SNP genotype data for the forward (fwd) strand was used for this purpose. Note that although the more expansive 1000 Genomes Project data (Genomes Project *et al.*, 2012) was publicly available at the time of querying, the HapMap dataset was used for consistency purposes since this was the dataset originally used by Scerri *et al.* (2011a) for their imputation of SNPs. Haploview enables the analysis of LD patterning through the use of colour coding which indicates the pairwise correlation *D'* or r^2 coefficients (measures of LD) between SNPs. Pairwise comparisons of markers greater than 500 kb apart and individuals with more than 50% missing genotypes were ignored.

3.3.2 Interrogation of the TFBS databases

All 22 SNPs of the defined 12.7 kb HapMap block were queried for TFBSs in 20-base pair sequences centred on each SNP allele using the TRANSFAC v2014.4 (Matys *et al.*, 2006) and Genomatix MatInspector v8.0.6 (Cartharius *et al.*, 2005) databases. Only those SNPs that displayed significant TFBS gain/loss on allelic variation were considered for further functional analysis (Table 3.1). TRANSFAC used a minFP profile to minimise false positives reported by the interrogated vertebrate taxonomic group. 1000 Genome Pilot project data offers superior SNP resolution to the HapMap Rel28 PhaseII+III dataset and as such was used with SNAP v2.2 (Johnson *et al.*, 2008) to create a list of SNPs (see Appendix B) that were in LD (> 0.4 r^2) with rs11855415, the highest associated PCSK6 marker from a previous handedness GWA study (Scerri *et al.*, 2011a).

3.3.3 Software prediction of the secondary promoter region

Beginning with an initial spread of 1 kb either side (chr15:101873604-101875513, hg19) of the H3K27Ac acetylation markings (indicative of an active promoter) at the PCSK6 locus (track C, Figure 3.2), several software packages were used to create a consensus prediction for the extent of a suspected secondary promoter in both strand directions. Promoter2.0 (Knudsen, 1999) predicts transcription start sites (TSSs) of vertebrate RNA Polymerase II promoters in DNA sequences while Proscan v1.7 predicts promoter regions based on scoring homologies with putative eukaryotic RNA Polymerase II promoter sequences. The Mammalian Promoter Database (MPromDb) (Gupta *et al.*, 2011) and the DataBase of Transcriptional Start Sites (DBTSS, Suzuki *et al.*, 2015) are both curated databases that provide exact positions of TSSs identified from ChIP-Seq experiment and TSS-seq results respectively.

3.3.4 Identifying evolutionarily conserved sequences

To identify conserved sequences across the previously defined region of LD (chr15:101863220-101875949, hg19, see Results section 3.4.3), the comparative genomics tool of the Ensembl genome browser (http://www.ensembl.org/) was used to obtain a multiple-sequence alignment for the human reference (hg19) and 13 eutherian mammals; chimp, gorilla, orang-utan, rhesus, baboon, marmoset, mouse, rat, rabbit, cow, dog, cat and horse (genome releases listed in Appendix B). Note that I have restricted analyses to genome assemblies that have been sequenced at relatively high coverage (> 6 X) to minimise the impact of sequencing and assembly errors. Sequence conservation was analysed by pair-wise alignment using the mVISTA service (default parameters, http://genome.lbl.gov/vista/mvista/submit.shtml) and the Vertebrate Multiz Alignment & Conservation (100 Species, Element Conservation by phastCons) track from the UCSC Genome Browser (Siepel *et al.*, 2005). For the mVista service any variant that fell within the 12.7 kb associated region in an interval ≥ 70% conservation across a 100bp sliding window in any of the species was considered conserved.

3.4 Results

3.4.1 Region of Linkage Disequilibrium

As discussed previously in the introductory chapter, two GWA studies (Brandler *et al.,* 2013, Scerri *et al.,* 2011a) have implicated a locus of PCSK6 to be associated with handedness at a genome wide significant level (P-values below 5 × 10^{-8}) in a dyslexic cohort. These studies were followed up by an independent association study of a general population cohort which also reported a genetic variant at that locus, a VNTR (rs10523972), to be associated with the degree of handedness (P = 0.001, significance threshold: P < 0.0025, adjusted for multiple comparisons). Together these studies suggest a PCSK6 locus of unknown size, contains the causal variant(s) driving an association signal between PCSK6 and handedness. It is possible to define such a locus by taking note of the location of the most significantly-associated genetic variants uncovered in previous GWA studies (rs11855415 P = 2.0 × 10^{-8} and rs7182874 P = 8.68 ×10^{-9}, see track D Figure 3.2) and to annotate all SNPs in LD with these variants at that locus. Such a locus would encapsulate not just the tagged genetic variants but also the causal variant(s). For this purpose I downloaded data from the International HapMap project to map the LD among SNPs in the PCSK6 gene. There are two popular LD measures provided in the HapMap data, *D'* and r^2; higher values of each imply stronger LD among the SNPs. I used *D'* as the basis for partitioning the PCSK6 gene into regions of LD since this measure is normalised for allele frequencies, thereby making it better suited than r^2 for estimating the overall LD across pairs of multi-allelic loci (Zapata, 2000).

The blue triangle in Figure 3.1 indicates a 12.7 kb block of LD containing 22 SNPs (for a zoomed-in version see track D, Figure 3.2) including the highest associated PCSK6 markers from previous GWA studies and is defined to be in the region chr15:101863220-101875949, hg19. This region spans from rs11630012 (within intron 13 of PCSK6) to rs1871975 (intron 17), and contains intriguing epigenetic markings (track C, Figure 3.2) around the highest associated SNP rs11855415 and the VNTR. The high LD amongst the 22 SNPs (Appendix B) in the region precludes identifying the precise causal variant(s) through association analysis alone. All 22 SNPs are included on the Illumina Omni Express SNP array (as used in the latest Brandler *et al*. (2013) GWAS) and have a minor allele frequency exceeding 5%.

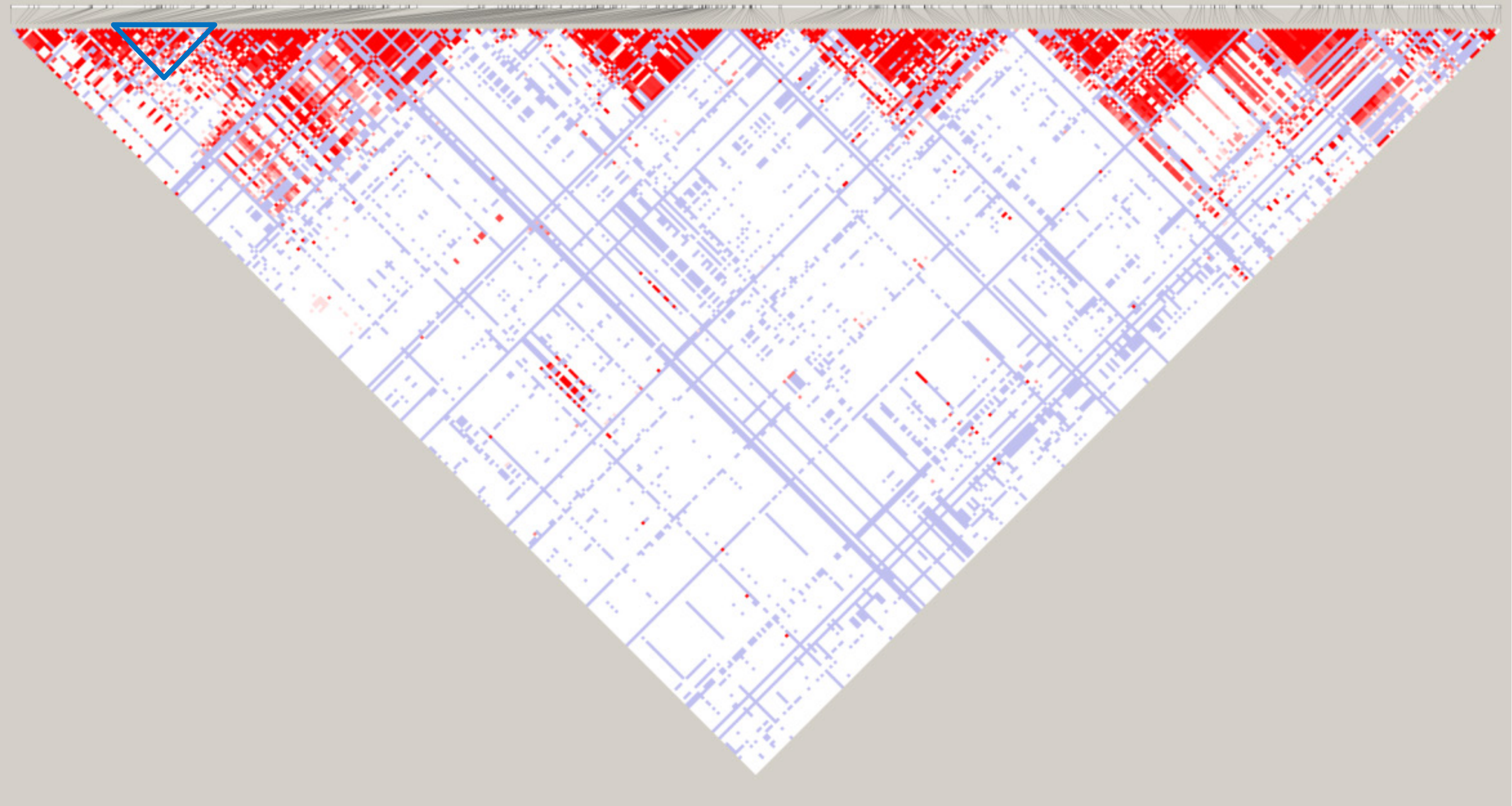


Figure 3.1 LD structure of the HapMap North and Western European population (CEPH) at the PCSK6 gene.SNP genotype data for the PCSK6 gene (accession: NM_002570) was downloaded from the HapMap data source for the CEU population (Rel28 PhaseII+III Aug10 NCBI36, dbSNP126) and evaluated with Haploview v4.2 (Barrett *et al.*, 2005). A red box indicates the absolute D prime (*D'*) between two loci while an empty blue box represents low LD. The navy triangle indicates the genomic region harbouring the genetic variants as represented in Figure 3.2.

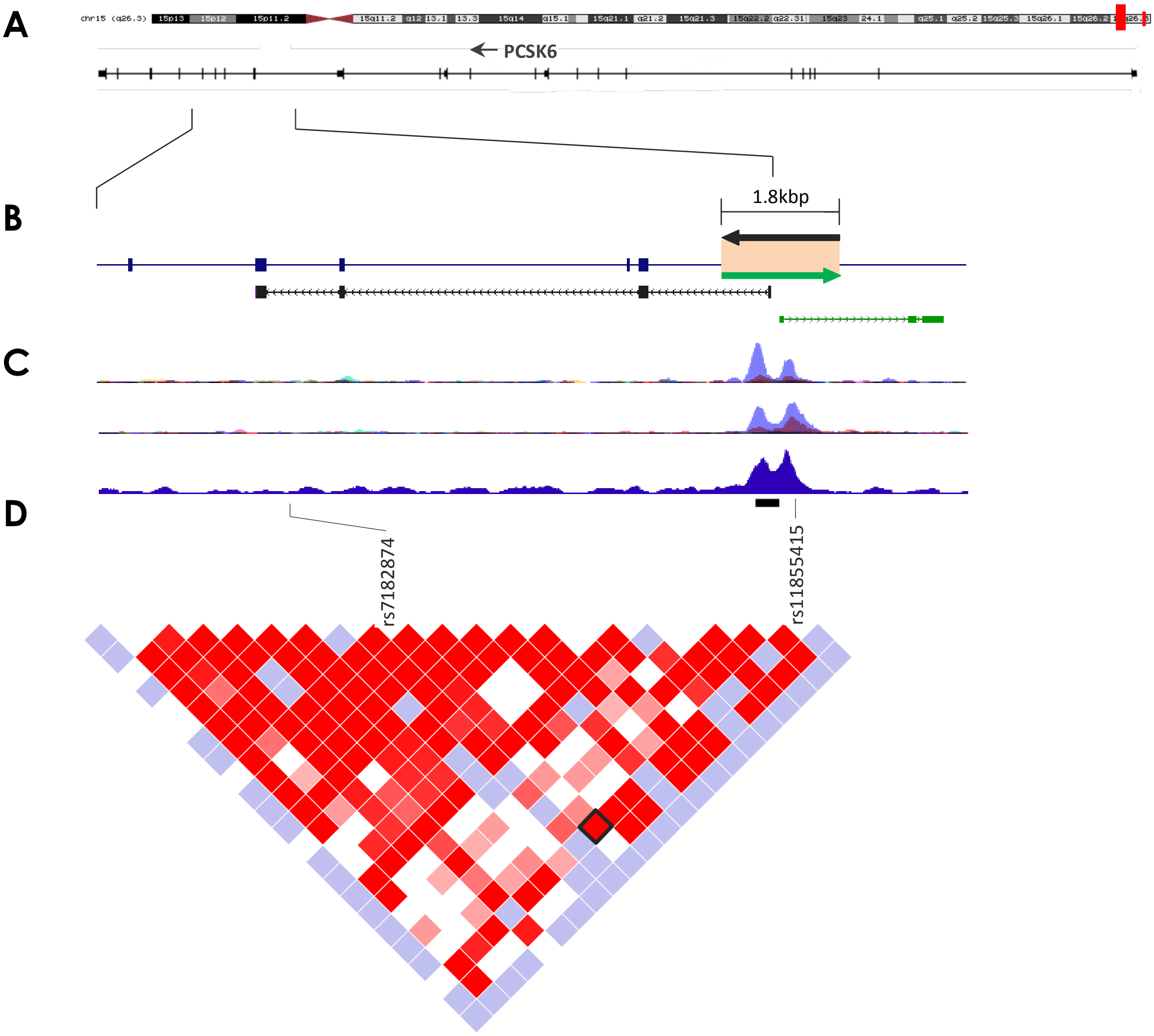


Figure 3.2 PCSK6 locus associated with relative hand skills
(A) PCSK6 is located on chromosome 15q26.3 indicated by the red box **(B)** Zoomed-in view of the PCSK6 region associated with relative hand skill (Brandler *et al.,* 2013, Scerri *et al.,* 2011a). A 1.8kbp region (beige box) is predicted to be a regulatory element involved in driving transcription in a sense (black transcript, accession DB023826) and antisense (green transcript, RP11-299G20.3) direction **(C)** Tracks from the UCSC Genome Browser (http://genome-euro.ucsc.edu) showing Chip-SEQ ENCODE data indicating putative promoter activity in (B). From top, the tracks show H3K27Ac and H3K4Me3 histone marks as determined in different cell lines. The highest peak (violet) indicates epigenetic markings in the K562 cell line. The bottom track displays RNA Polymerase II binding for the K562 cell line which shows a higher signal compared to other cell lines. **(D)** The genetic associations cluster within a 12.7 kb linkage disequilibrium (LD) block as defined by HapMap CEPH data (chr15:101863220 – 101875949). For the purpose of clarity only the highest-associated GWAS-detected markers are included here. The black bar indicates the rs10523972 VNTR position within the predicted promoter. rs7182874, the highest-associated marker in the most recent GWAS (Brandler *et al*., 2013) is in high LD (black diamond) with rs11855415 (Scerri *et al*., 2011a). A red box indicates the absolute D prime (*D’*) between two loci while an empty blue box represents low LD

3.4.2 Allelic effect on transcription factor binding affinity

All 22 SNPs from the HapMap data within the 12.7 kb LD region were tested for allelic effects on the loss/gain of TFBSs using the TRANSFAC (Matys *et al.*, 2006) and Genomatix MatInspector (Cartharius *et al.*, 2005) databases (Table 3.1). The marker rs11855415 was predicted to affect the largest number of TFBSs and is located in close proximity (< 500 bp) to a region predicted to act as a secondary bidirectional promoter (track C, Figure 3.2), the annotation of which is discussed in the following section (3.4.3). The SNP rs11855415 displays the most significant loss/gain of TFBSs on allelic variation. It is also worth noting that the VNTR rs10523972 also lies within the same 12.7 kb LD region and is predicted to have two TFBSs at the junction of each tandem repeat; c-Ets-1 is an enhancer-binding protein that activates transcription (Wasylyk *et al.*, 1990) while Tel-2 has been shown to be a transcription repressor (Gu *et al.*, 2001) however an extensive search of the literature suggests neither TF to function in multiplex, a relevant mechanism of action if each tandem repeat were acting as a bind site.

Table 3.1 *In silico* TFBS prediction of the PCSK6 locus associated with handedness. Combined TRANSFAC (v2014.4) and Matinspector (MatInspector Release 8.0.6) prediction of allelic effects on TFBSs for SNPs at the PCSK6 locus associated with the PegQ measure of handedness.

SNP	*allele*	*transcription factors*
rs3825921	A	None
rs3825921	G	None
rs1871975	C	None
rs1871975	T	None
rs1871976	A	None
rs1871976	G	None
rs1871978	C	HSF 1 NF-E2-related factor 1/Transcription Factor MafG heterodimers binding to subclass of AP1 sites
rs1871978	T	MyT1 RREB-1
rs9806218	A	None
rs9806218	G	None
rs9806256	C	Cardiotrophin-1
rs9806256	T	Nuclear transcription factor Y Binding site for a Pbx1/Meis1 heterodimer
rs4965830	A	Homeobox C10/Hox-3iota Muscle TATA box
rs4965830	T	Intestine specific homeodomain factor for Homeobox protein CDX-1 Homeobox protein Hox-B9 Homeobox protein Hox-C13 Homeobox protein Hox-D12
rs2220055	A	None
rs2220055	G	None
rs2277593	C	None
rs2277593	G	None
rs2277593	T	None
rs7182874	C	Paired box protein Pax-5 Protein BANP Krueppel-like factor 6
rs7182874	T	T-cell acute lymphocytic leukemia protein 1
rs12901236	C	Inhibitor of growth protein 4
rs12901236	T	Inhibitor of growth protein 4
rs1471656	C	None
rs1471656	T	None
rs1947942	A	Xvent-1 protein

rs1947942	G	None
rs752028	C	None
rs752028	T	Breast cancer type 1 susceptibility protein: Upstream stimulatory factor 2 complex
rs882422	A	Krueppel-like factor 6
rs882422	G	Krueppel-like factor 6 Transcription intermediary factor 1-beta BEN domain-containing protein
rs752026	A	None
rs752026	G	None
rs755867	A	None
rs755867	G	None
rs2073592	C	None
rs2073592	T	None
rs2239858	A	Killer cell lectin-like receptor subfamily G member 1
rs2239858	G	Killer cell lectin-like receptor subfamily G member 1
rs12916087	A	None
rs12916087	G	None
rs12900794	C	None
rs12900794	T	None
rs11855415	A	Sex-determining region Y protein POU domain, class 6, transcription factor 1 Homeobox protein Hox-A5 POU domain, class 4, transcription factor 3 Homeobox protein BarH-like 2 POU domain, class 2, transcription factor 1 Homeobox protein Hox-B3 LIM-homeodomain transcription factor Homeobox protein Nkx-6.3 DNA-binding protein SATB1 GS homeobox 1 Zinc finger protein 333
rs11855415	T	POU domain, class 3, transcription factor 2 Intestine specific homeodomain factor for Homeobox protein CDX-1 Homeobox B8 / Hox-2delta Spalt-like transcription factor 1 Special AT-rich sequence-binding protein 1, predominantly expressed in thymocytes, binds to matrix attachment regions (MARs) Homeobox D10

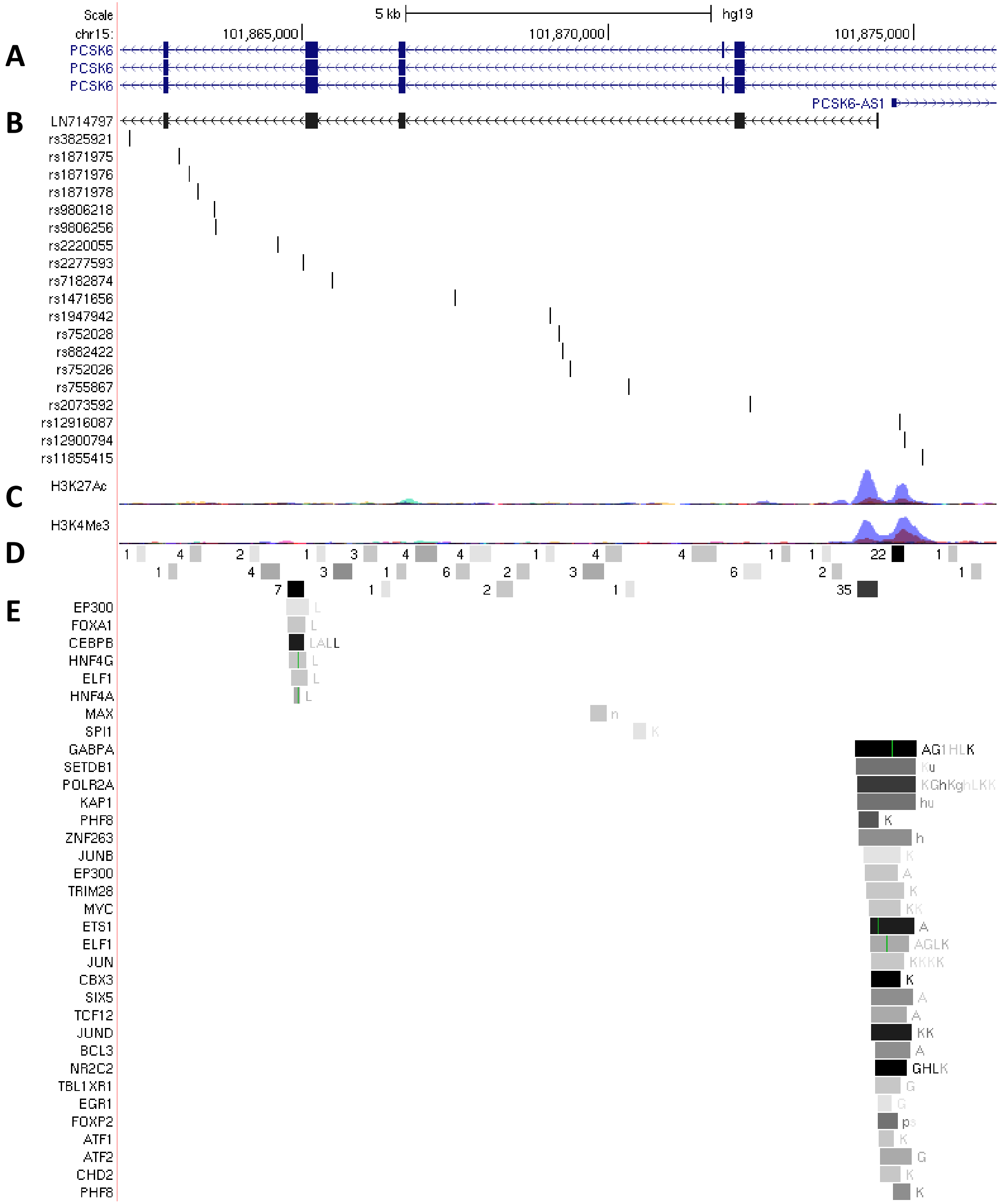


Figure 3.3 UCSC genome browser tracks indicating TFBSs as assayed by ChIP-seq at the PCSK6 locus of interest (chr15:101863220-101875949) according to the ENCODE project datasets (Gerstein *et al.*, 2012). PCSK6 canonical (accession: NM_002570) and shorter isoform (accession: LN714797) genes are displayed relative to the PCSK6-AS1 lncRNA transcript **(A)**. SNPs within the 12.7 kbp region of LD (see Figure 3.2 & Table 3.1) **(B)** lie in close proximity to epigenetic markings (H3K27Ac and H3K4Me3 histone marks, **(C))** indicative of a bidirectional promoter. Numbering adjacent to the DNAse clusters **(D)** indicate combined data from the peaks of multiple cell lines (from a total of 125 cell types assayed). Green vertical bars within the transcription factor bind sites **(E)** represent DNA binding motifs. ChIP-seq experiments were performed by the ENCODE project on 91 cell types, with transcription factors binding at the indicated locations in the following cell types: A549 (A), GM12878 (G), IMR90 (I), HeLa-S3 (H), HepG2 (L), K562 (K), U87 (u), HEK293 (h), GM12872 (g) and PANC-1 (p).

3.4.3 Predicted region of secondary promoter

The location of two previously identified genetic variants found to be associated with handedness (SNP rs11855415 and VNTR rs10523972) is within a region that displays enrichment for histone modifications (track C, Figure 3.2). Specifically, the ENCODE project tracks represent epigenetic markings including trimethylation of Lys4 of histone H3 (H3K4Me3, associated with promoters that are active or poised to be activated); acetylation of Lys27 of histone H3 (H3K27Ac, often found near active regulatory elements since its thought to enhance transcription possibly by blocking the spread of the repressive histone mark H3K27Me3) and monomethylation of Lys4 of histone H3 (H3K4Me1, associated with enhancers and with DNA regions downstream of transcription starts). As such, the obvious bimodal distribution of both H3K27Me3 and H3K27Ac markings at this location are highly suggestive for the presence of a putative bidirectional secondary promoter.

Taking an initial spread of 1 kb either side of the H3K27Ac acetylation markings at the PCSK6 locus (chr15:101873604-101875513, hg19) I used several databases and software algorithms to predict the minimum sequence required to drive transcription in both strand directions. From the analysis indicated (Table 3.2), the consensus is that the TSS for a bidirectional promoter is located between chr15:101873808-101874718 (hg19). After primer design considerations that might be required in downstream assays, the sequence of interest was extended to 1,806 bps spanning chr15:101873803-101875608 (hg19).

This putative secondary promoter (beige box, track B, Figure 3.2) is within the vicinity of two expressed sequence tags (ESTs) that are potentially driven by it: 1) The RP11-299G20.3 EST is a suspected long non-coding RNA (lncRNA) type gene, a regulatory element transcribed in an antisense direction and suspected to have an effect on gene expression on the sense strand (green transcript, track B, Figure 3.2) and 2) the DB023826 EST appears to be a 4-exon shorter isoform of the PCSK6 gene (black transcript, track B, Figure 3.2). Further characterisation of this suspected secondary promoter region is pursued in the chapters which follow.

Table 3.2 PCSK6 secondary promoter prediction. Several software packages were used to form a consensus of which sequence within an initial 2 kbp spread (chr15:101873604-101875513, hg19) would act as a promoter. Note DBTSS and Promoter 2.0 offer predictions on both strands. DBTSS provides a visualisation of the TSSs predicted in both a sense and antisense strand direction in Figure 3.4.

software	*strand*	*TSS*	*Notes*
Promoter 2.0	Sense (-)	chr15: 101874713	Highly likely prediction (1.091)
	Antisense (+)	chr15:101873983	Marginal prediction (0.57)
Proscan v1.7	Sense (-) & Antisense (+)	chr15:101873808	Promoter Score: 56.47 (Cutoff = 53)
MPromDb	Sense (-) & Antisense (+)	chr15:101874396-101874718	Annotated as a bidirectional novel promoter in lymphoblastoid CD4+T cells
DBTSS	Sense (-)	chr15:101874407	Cell/tissue: PC3, testis & HEK293
	Antisense (+)	chr15:101874697	Cell/tissue: RERF-LC-MS, testis & HEK293

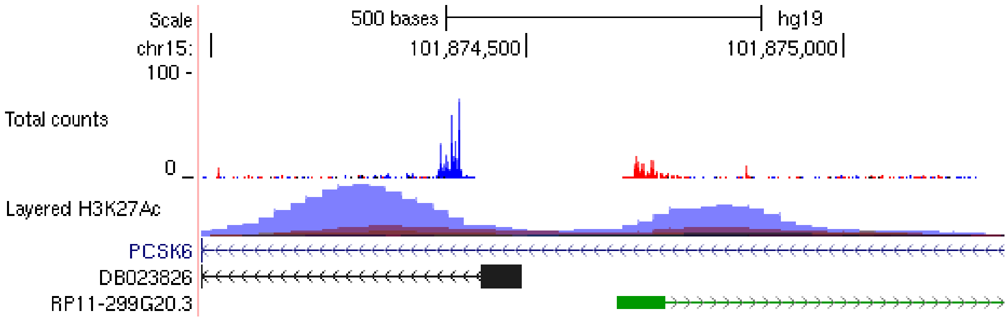


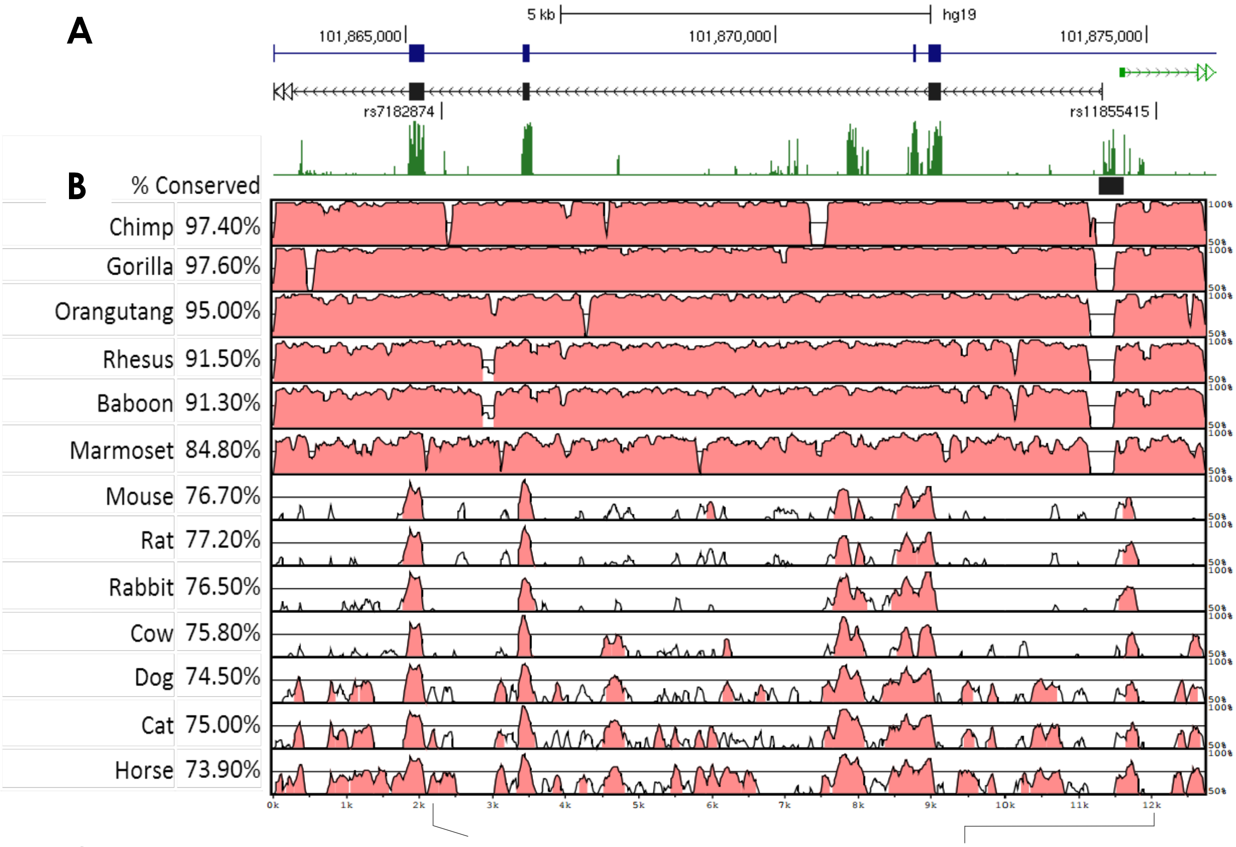
Figure 3.4 Visualisation of the TSSs annotated in the DBTSS database on both sense and antisense strands within PCSK6 intron 14. The total counts track represents TSS capped analysis of gene expression (CAGE) peaks which are identified by DPI (decomposition based peak identification) from adult testis samples. See http://fantom.gsc.riken.jp/5/ for further details and data exploration. H3K27Ac markings are indicative of a bidirectional promoter driving transcription of a short PCSK6 isoform on the sense strand (black) and a lncRNA on the antisense strand (green).

3.4.4 Evolutionary conserved sequences

The functional importance of the 22 noncoding SNPs previously identified at the 12.7 kb PCSK6 locus can be explored in terms of constrained elements. Previous analyses of conserved noncoding sequences in the human genome (Pennacchio *et al*., 2006, Bejerano *et al*., 2004) have defined ultraconserved elements among mammals (sequences at least 200bp in length that are 100% identical among human/mouse/rat) as useful indicators of sequences with an increased likelihood of demonstrating gene regulatory activity. This was demonstrated in a recent study by Forrest *et al*. (2014) who showed regulatory regions such as promoters often overlap evolutionarily conserved sites in mammals. Interestingly within the 12.7 kb LD region previously identified (track D, Figure 3.2), both the known PCSK6 exons and the region encompassing the suspected secondary promoter satisfy this definition of ultraconserved elements (track B, Figure 3.5).

Of the 22 SNPs identified, only 5 lie within VISTA pair-wise conserved sequences across all 14 mammalian genomes (see Appendix B). None of these variants are exonic though three SNPs, the GWAS-detected rs11855415, rs12900794 and rs12916087 lie within the predicted bidirectional secondary promoter. Apart from rs11855415, none of the SNPs are predicted to affect a TFBS (Table 3.1) and as such were discounted from downstream functional analysis. Analysis of the rs11855415 SNP reveals that the major T allele is conserved across all species except rabbit, while the rs7182874 SNP C allele is conserved across all species except mouse, rat and cat (track C, Figure 3.5).

Both the SNP rs11855415 and the VNTR were previously shown to lie within epigenetic markings indicative of promoter activity (track C, Figure 3.2) though the later also appears to disrupt an evolutionary conserved sequence (black bar, track A, Figure 3.5); if this is the case, rather than a sequencing artefact, it would make the VNTR an interesting candidate for functional analysis.



C

GGCCCC**C**GCTGG	Human	TCATT**T**TTATT
GGCCCC**C**GCTGG	Chimp	TCATT**T**TTATT
GGCCCC**C**GCTGG	Gorilla	TCATT**T**TTATT
GGCCCC**C**ACTGG	Orangutan	TCATT**T**TTATT
GGCCCC**C**ACTGA	Rhesus	TCATT**T**TTATT
GGCCCC**C**ACTGA	Baboon	TCATT**T**TTATT
GGCCCC**C**ATTGG	Marmoset	TCATT**T**TTATT
- - - - - - - - - - - -	Mouse	ACACT**T**TTATT
- - - - - - - - - - - -	Rat	TCAAT**T**TTAAT
GGCCCC**C**AGCGG	Rabbit	--------**-**-------C
TCCC-**C**AAGGG	Cow	TCATT**T**TTATT
TCCC-**C**AGAGG	Dog	TCATT**T**TTACT
T-----------------	Cat	TCATT**T**TTACT
GCCC-**C**AGGGG	Horse	TCATT**T**TTAAT

Figure 3.5 Sequence conservation across a PCSK6 region associated with handedness. **(A)** Blue track displays exons 14 - 17 of the PCSK6 gene (accession: NM_002570) with a shorter PCSK6 isoform (black) and antisense lncRNA PCSK6-AS1 (green) originating from a putative secondary promoter. For purposes of clarity only the two highest associated SNPs from previous GWAS, rs11855415 and rs7182874 (Scerri *et al*., 2011a, Brandler *et al*., 2013), are shown above the 100 vertebrate conservation (phastcons) track from the UCSC Genome Browser. The VNTR range is represented by a black horizontal bar. **(B)** Plot of conservation spanning between the PCSK6 LD block (chr15:101863220-101875949, hg19) and the reference sequences from 13 eutherian mammals. The mVista service was used to visualise alignment for the identification of sequence similarity. Pink coloured regions are > 70% conserved and indicate conserved non-coding sequence (CNS) between the human and query sequence using a calculation window of 100 bp. **(C)** Aligned sequence from 14 species for SNPs rs7182874 and rs11855415.

3.5 Discussion

This chapter describes an approach to define the extent of a PCSK6 locus previously reported to be associated with handedness. LD maps are a useful parameter in guiding post-GWA functional studies and have been used by researchers to untangle the evolutionary history of humans, including population growth and structure, selection, genetic drift, migration, recombination rates and mutations (Hedrick, 2011). In this chapter, I visually inspected LD patterning at a PCSK6 locus showing association with handedness to determine the extent of the association signal and the position relative to the PCSK6 gene. I defined a 12.7 kb block containing 22 SNPs in modest to high LD (*D'* = 0.55 – 1, Appendix B) spanning introns 13 – 17 (chr15:101863220 – 101875949). Patterns of LD have been previously known to be noisy where nearby pairs of sites from the same region might be in weak LD and pairs of sites that are many kbps apart might be in almost complete LD (Wall and Pritchard, 2003). Analysis of the distribution of LD across the PCSK6 gene in the CEU population (Figure 3.1) indicates, to the contrary, multiple distinct neighbouring regions of LD (for criteria see 3.3.1) separated by inferred recombination events (Gabriel *et al*., 2002).

All 22 SNPs within the 12.7 kb LD region were tested for allelic effects on the loss/gain of TFBSs with the marker rs11855415 predicted to affect the largest number of TFBSs. Interestingly some of the TFs predicted to bind at this location are known to have an effect during 'multicellular organismal development' (GO:0007275) and 'anterior/posterior pattern specification' (GO:0009952) such as HOXA5, HOXB3, HOXB8, HOXD10 and HOXB13 (UniProt, 2015). Another TF predicted to bind to the A but not the T allele of rs11855415 is a protein known to play a role in the formation of radial glia, the cells that provide a scaffold structure for neuronal migration (Kiyota *et al*., 2008). Results from the TFBS predictions suggest rs11855415 to be a primary candidate for functional analysis. Combining the identification of sites containing known binding motifs with sequences of evolutionary conservation is a powerful approach to identifying SNPs which may have a functional effect. Many studies are increasingly integrating knowledge of cross-species conserved regions in to the selection process when considering SNPs for functional studies – whether it is to identify evidence of eQTLs (Sillé *et al.*, 2012), potential splice sites (Burgess and Freeling, 2014) or insertions/deletions (indels) (Ajawatanawong *et al.*, 2012). Five of the 22 SNPs appeared within a CNS (see Appendix B), though of these only rs11855415 was predicted to contain a TFBS and as such was retained for further analysis.

The defined 12.7 kb LD region appears to contain a suspected regulatory element as indicated by sequence conservation (UCSC phastcons track, Figure 3.5). Analysis of the histone modification markings at the region (track B, Figure 3.2) provides a complementary approach; one of the parameters that controls gene transcription is the interplay of regulatory events between gene promoters and gene-distal regulatory elements called enhancers (Andersson, 2015). Enhancers are typically short (50 - 1500 bp) sequences of DNA that bind activator proteins which interact at a promoter to begin gene transcription. To aid the functional characterisation of this PCSK6 locus and the generation of hypotheses it is essential to define the regulatory element as either an enhancer or a promoter. Fortunately, ENCODE project data provides distinct chromatin signatures for these different regulatory elements; the substantial bimodal distribution for H3K27Me3 markings is highly suggestive for the presence of a secondary promoter internal to the PCSK6 gene. In contrast, enhancers typically display low H3K4Me3 markings relative to H3K4Me1 markings. Further evidence suggesting this element to be a promoter rather than a strict enhancer is the H3K27Ac and RNAPII markings (track C, Figure 3.2) which are much less pronounced at enhancers.

Inspection of the UCSC genome browser indicated there are two spliced expressed sequence tags (ESTs) with exons originating at this locus, one in a sense direction to the gene, and the other antisense (track B, Figure 3.2). The transcription of such a bidirectional gene pair, coupled with the epigenetic markings previously discussed is indicative of a bidirectional promoter, a common feature of mammalian genomes (Koyanagi *et al.*, 2005). *In silico* predictions and promoter databases provide further support for a bidirectional promoter (Table 3.2 & Figure 3.4) however *de novo* prediction of regulatory elements has its limitations, for example, not all promoters in eukaryotes have the same characteristic elements (e.g. a TATA box). I have attempted to offset this limitation with the addition of two curated database services which annotate gene promoters identified from ChIP-Seq experiment results (PromDB) and the sequencing of cDNA from humans and mice (DBTSS). It should also be noted that many tools rely on ENCODE data, which only has a limited scope of certain TFs and cell types so a high probability of false positives exists where SNPs might influence signals in irrelevant cell types. Ultimately, these *in silico* tools represent a useful starting point to guide the design of functional assays.

From the 22 SNPs, 5 were identified to lie within a CNS when comparing conserved sequences across 14 eutherian mammals (rs2073592, rs2277593, rs12916087, rs12900794 and rs11855415). Only rs11855415 indicated substantial predicted effects on allelic variation (Table 3.1) which in addition to the VNTR lies in close proximity to the suspected bidirectional promoter and as such may have a functional effect by affecting the transcriptional machinery that can bind at that location and thereby modulating transcriptional output. It is reasonable to posit that SNPs within conserved regions may be more likely to have phenotypic effects than SNPs in non-conserved sequences. In such a context, conserved sequence can be used as a putative annotation for genomic regions that may have functional importance, even when the exact nature of their function is unknown. The close proximity of rs11855415 and the VNTR to a suspected regulatory sequence in a region of strong conservation (phastCons track, Figure 3.5) across 13 orthologous species' sequences is a possible indication of an orthologous functional regulatory element that has been evolutionarily conserved. Though this regulatory element might be conserved, evidence supports an association between bidirectional promoters and lineage-specific novel transcripts in mammals (Piontkivska *et al.*, 2009). Furthermore, Gotea *et al.* (2013) argue that the lineage-specific activation of bidirectional promoters is an important mechanism for the emergence of novel transcripts which provide a molecular pool for functional diversification and adaptive change.

VNTRs display mutation rates up to 100,000 times higher than in other areas of the genome (Vogler *et al.*, 2007). The VNTR at the PCSK6 locus was shown to disrupt the sequence alignment between humans and all other aligned mammals (track B, Figure

3.5), though this finding most likely represents an artefact of genome assembly rather than a genuinely novel sequence that has evolved since the last common ancestor of humans and chimpanzees. In conclusion, SNP rs11855415 and the VNTR fall in a genomic region with relatively higher sequence conservation in 14 mammals investigated compared to neighbouring genomic areas. Of the 22 SNPs systematically analysed *in silico*, rs11855415 is predicted to affect the largest number of TFBSs on allelic variation - no other SNP shown to lie within a CNS was predicted to affect the binding of TFs. In addition, rs11855415 and the VNTR are both predicted to lie within a bidirectional promoter. The SNP rs7182874, identified to be significantly associated with handedness by Brandler *et al.* (2013), is not within a CNS or a suspected regulatory element. All things considered, both the SNP rs11855415 and the VNTR make interesting functional candidates for future analyses.

For now, building on the findings of this chapter I proceed with a functional approach to identifying the etiological variant(s) driving the association between a 12.7 kb region of LD and handedness, and the role, if any, a 1.8 kb predicted secondary promoter may have.

4 Functional elements at the PCSK6 locus associated with handedness

4.1 Abstract

Despite the abundance in availability of genome-scale data, the relationship between genome sequence and complex trait phenotype is still neither straightforward nor entirely understood. A functional analysis of the genetic variants within the 12.7 kb PCSK6 locus is preceded in this chapter by a cataloguing of the expression of all the RefSeq-recognised PCSK6 isoforms across a battery of cell lines. In Chapter 3 the K562 cell line displayed epigenetic markings indicative of an intronic bidirectional promoter which my findings here suggest drives the transcription of several novel long non-coding RNA (lncRNA) gene isoforms (PCSK6-AS1/2/3) and a previously unknown PCSK6 shorter isoform (SI) complete with signal peptide and 3'UTR (accession #LN714797). Conformational DNA sequencing has enabled prediction of the SI's protein function and though thought to be inactive and not secreted by the cell, may be exerting some phenotypic effect through its predicted growth factor-like binding and PLAC domains. Interrogation of the ENCODE project RNA sequencing (RNA-seq) datasets suggest the SI is retained in the nucleus in most cell lines, though the human embryonic stem cell line (hESC) differed in its predominant export of the SI to the cytosol for presumed further processing. Microarray datasets for both the developing and adult brains suggest a relatively high expression of the SI in the corpus callosum (CC), a region of the brain thought to exert an influence on functional laterality. The results I present in this chapter make the 12.7 kb PCSK6 locus associated with handedness more amenable to detailed functional studies which will be required to characterise the role, if any, of the PCSK6 gene's isoforms in the development of the handedness trait.

4.2 Introduction

Strategies for the prioritisation of SNPs from GWAS signals and thus the progression from GWAS-identified tagged SNP to functional SNP to mechanism typically begin with a linkage disequilibrium (LD) approach as presented in Chapter 3. While the region to be functionally annotated can be guided by LD structure, there are challenges to this approach if the strength of the correlation between the tagged SNP and the functional genetic variant is low due to noise or otherwise. To understand how genetic variants identified within the 12.7 kb LD block might affect function, I characterised part of the regulatory landscape by analysing histone modification signalling and evolutionary conserved sequences. Such an approach yielded a short DNA stretch of 1.8 kb which was putatively annotated as a bidirectional promoter predicted to drive the transcription of a 3-exon Natural Antisense Transcript (NAT) PCSK6-AS1, a lncRNA on the antisense strand, and a shorter version of the canonical PCSK6 gene (NM_002570) on the sense strand. This shortened PCSK6 isoform (DB023826.1) is thought to be four exons in length, the first of which is not included in any other PCSK6 isoform according to the NCBI Reference Sequence database, RefSeq (http://www.ncbi.nlm.nih.gov/refseq/). Little evidence exists for either lncRNA or SI transcript in the literature; Kimura *et al*. (2006) previously detected a single cDNA clone of the PCSK6 SI when investigating alternative promoters in adult testis while the existence of the PCSK6-AS1 lncRNA was first posited with the release of the GENCODE gene sets arising from the ENCODE project (Harrow *et al*., 2012).

The transcription of two adjacent genes coded on opposite strands, with their 5' ends oriented toward one another are known as bidirectional gene pairs (BGPs) (Yang *et al*., 2007). In humans, 10% of protein-coding genes and more than half of all expressed lncRNAs represent divergent transcription from such promoters (Sigova *et al*., 2013, Trinklein *et al*., 2004). BGPs are typically expressed in a highly spatial- and temporally-specific manner (Djebali *et al*., 2012) with previous studies indicating a positive correlation between the majority of identified sense/antisense gene pairs expressed, though inversely correlated pairs are also known to exist (Katayama *et al*., 2005, Morrissy *et al*., 2011, Mizuta *et al.,* 2013). It is worth noting in the context of PCSK6-

AS1 that expression of lncRNAs from bidirectional promoters are generally highly enriched in neuronal genes (Hu *et al*., 2014) and human embryonic stem cells (hESCs) where > 60% promoters might be bidirectional and associated with divergent lncRNAs (Sigova *et al*., 2013). Conversely, most human promoters are thought to bind polymerase complexes in a bidirectional manner and are therefore capable of initiating transcription in both strand directions (Wei *et al*., 2011). Thus, we cannot exclude that the presence of lncRNAs at some bidirectional promoters may represent a passive by-product of gene transcription. In short, it is still unclear how exactly NATs affect their sense counterpart's RNA expression level though it is likely that other lncRNAs will be identified at loci identified by GWASs as more RNA-seq data and methodologies for detecting rare transcripts become available (Han *et al*., 2015). A primary aim of this chapter is therefore to confirm and annotate the existence of such a BGP at the previously identified PCSK6 locus.

Analysis of expression pattern is often one of the first steps in understanding a gene's function however the novel first exon of the shorter PCSK6 isoform means a paucity of datasets from which we can query expression since no microarray data exist which have been probed at this genomic location. Fortunately the public availability of the ENCODE project's RNA-seq data allows the detection of alternative splicing and the quantification of expression levels down to the level of individual transcript isoform (Feng *et al*., 2013). In higher eukaryotes, the vast majority of protein-coding genes express multiple transcript isoforms (Nagasaki *et al*., 2005) and PCSK6 is known to have at least 8 confirmed and annotated isoforms according to RefSeq (see Table 4.1).

PCSK6 is widely expressed in humans, with particularly high expression in the liver, spinal cord (Figure 1.3) and corpus callosum (CC, the broad band of nerve fibres joining both hemispheres, Figure 4.1). Callosal involvement has long been suspected in neurodevelopmental disorders such as autism (Frazier *et al*., 2012), ADHD (Paul, 2011) and dyslexia; the CC in dyslexic individuals is of different size and shape (von Plessen *et al*., 2002, Hynd *et al*., 1995) and inter-hemispheric transfer is less efficient compared to normal subjects (Sotozaki and Parlow, 2006). The callosal transfer deficit hypothesis (Fabbro *et al*., 2002) suggests dyslexia may be accounted for by an abnormal CC and, consequently, defective transfer and direction of neural impulses in the brain which are responsible for anomalous performance e.g. in written language processing. Interestingly, the front portion of the CC has been reported to be 11% larger in left-handed and ambidextrous people than right-handed people (for a meta-analysis see Driesen and Raz (1995)). If previous research suggests both dyslexia and handedness are influenced by the size and function of the CC, then might there be a pleiotropic gene or function of the CC influencing both phenotypes simultaneously?

Although a significant proportion of transcript isoforms are most likely the result of noise in the splicing process (Chern *et al.*, 2006, Melamud and Moult, 2009), several prominent examples of isoform switching resulting in a large impact on cellular phenotypes are known to exist in the mammalian brain (da Cruz e Silva *et al.*, 1995, Cheung *et al.*, 2007). For example, Flames *et al.* (2004) demonstrated different isoforms of Neuregulin-1 (NRG1), a schizophrenia candidate gene involved in controlling neuronal migration, are expressed in the developing cortex. Intriguingly, NRG1 is involved in glycogen synthase kinase 3 (GSK3) signalling, disruption of which affects many fundamental processes of brain development (Lovestone *et al.*, 2007). Furthermore, Hur and Zhou (2010) showed that only one of the GSK3 isoforms is expressed specifically in the nervous system, with the highest levels found during development. While the functional relevance of most spliced isoforms localised to the brain or otherwise, remains unknown (Tress *et al.*, 2007), these findings demonstrate the importance of analysing altered isoform expression rather than considering just the canonical gene expression in developmental traits and diseases. As a consequence, this chapter will explore PCSK6 isoform expression in both the developing and adult human brain, with a particular emphasis on the CC.

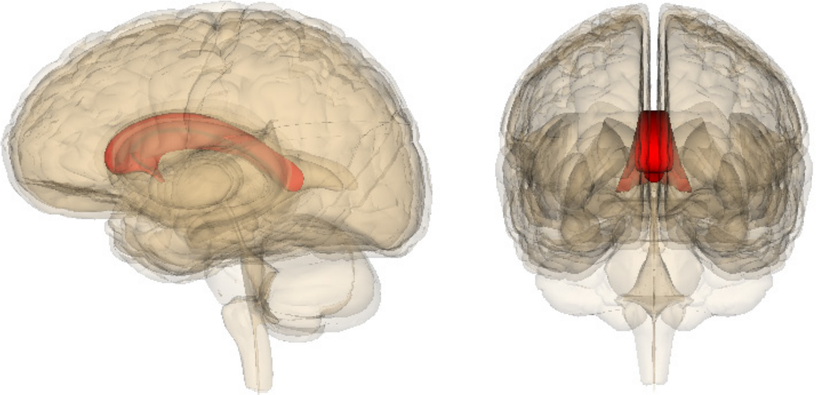


Figure 4.1 The human corpus callosum. Sagittal and anterior view of the human cerebrum within which the corpus callosum marked in red connects both hemispheres via a thick band of nerve fibres

Table 4.1 PCSK6 canonical and spliced isoforms according to RefSeq. Length of the isoform is given in base pair (bp) cDNA and translated protein amino acid (AA) length. Signal peptide prediction was provided by the SignalP 3.0 service (Bendtsen *et al.,* 2004). Comments supplemented by RefSeq (accessed August 2015). See Figure 4.2 for a visual representation of the isoforms.

Accession Number	*RefSeq Isoform*	*Length (bp,AA)*	*Enzymatically active*	*Signal Peptide*	*Comments*
NM_002570	PACE4A-I	992,969	Yes, secreted	Yes	Encoding the predominant canonical isoform, PACE4A-I precursor protein seems to exist in the reticulum endoplasmic as both a monomer and a dimer-sized complex whereas mature PACE4A-I exists only as a monomer, suggesting that propeptide cleavage affects its tertiary or quaternary structure. Isoform PACE4A-I is expressed in heart, brain, placenta, lung, skeletal muscle, kidney, pancreas, but at comparatively higher levels in the liver.
NM_138319	PACE4A-II	998,956	Yes, secreted	Yes	PACE4A-II is at least expressed in placenta.
NM_138320	PACE4E-II	927,962	Possibly, Retained intracellularly	Yes	Endomembrane system; Peripheral membrane protein. Note=Retained intracellularly probably through a hydrophobic cluster in their C-terminus. Isoform PACE4E-II is at least present in cerebellum.
NM_138321	PACE4E-I	921,975	Possibly, Retained intracellularly	Yes	Endomembrane system; Peripheral membrane protein. Note=Retained intracellularly probably through a hydrophobic cluster in their C-terminus. Isoform PACE4E-I is expressed in cerebellum, placenta and pituitary
NM_138324	PACE4C	128,652	Probably not, not secreted	Yes	Not secreted, remains probably in zymogen form in endoplasmic reticulum. Placenta.
NM_138323	PACE4CS	368,623	Probably not, not secreted	Yes	Not secreted, remains probably in zymogen form in endoplasmic reticulum
NM_138325	PACE4D	341,497	Probably not, not secreted	No	PACE4D is at least expressed in placenta.
NM_138322	PACE4B	611,487	Probably not	Yes	Predicted to be secreted

The choice of cell type is important when considering any functional analysis of candidate genetic variants since regulatory elements such as bidirectional promoters are known to be highly tissue- and cell-type specific (Kippner *et al*., 2014). For example, a recent study detailed very different activities of eleven enhancers across four mammary epithelial cell lines, emphasising the necessity of performing these assays in various cellular contexts (Rhie *et al*., 2013). An essentially healthy, non-aberrant tissue is perhaps the hardest context to replicate and maintain in a laboratory situation; many of the cell lines used throughout this study are cancerous in etiology (e.g. HeLa, SH-SY5Y), used here not by design but rather for practical purposes; the myeloid cell line K562 is a less than ideal model for neurogenesis however the cells are well-documented and easy to culture/transfect and so can be used in experiments in which the macro behaviour of the cell itself *per se* is not of interest. Several curated databases such as the Human Protein Atlas (Uhlen *et al*., 2015) do exist in which PCSK6 expression in individual cell lines is detailed (Figure 4.3) however such databases tend to aggregate RNA expression across all transcripts rather than offer specific expression profiles for individual isoforms. Therefore an empirical assessment of PCSK6 isoform expression across multiple cell lines is required if we are to understand the role, if any, of the PCSK6 SI.

In summary, by using multiple techniques this chapter will annotate in detail a locus of the PCSK6 gene associated with handedness which is predicted (see Chapter 3) to contain a bidirectional promoter driving expression of a lncRNA (PCSK6-AS1) and a shorter PCSK6 isoform (DB023826). To begin analysing the PCSK6 locus in a functional capacity requires the compilation of a catalogue of PCSK6 isoforms across various cell lines (4.4.1). Confirmation of the existence of the SI and lncRNA involved Polymerase Chain Reaction (PCR) assays and Sanger sequencing (4.4.2 and 4.4.6). Acquiring the cDNA sequence also enabled estimates of protein structure and therefore function (4.4.3), if any, in addition to a querying of RNA-seq and microarray databases for gene isoform expression profiling (4.4.4 and 4.4.5). The shorter PCSK6 isoform is predicted to express a novel exon and little proof of its existence is evident in the literature, however inference from existing gene expression data is possible (4.4.4).

4.3 Methods

4.3.1 PCSK6 isoform profiling

Analysis of PCSK6 gene isoform expression was performed by extracting and purifying total RNA (tRNA) from the range of cell lines indicated in Figure 4.2 (RNeasy Mini Kit, Qiagen) which were all cultured according to ATCC guidelines except for the H9 hESC-derived cells which were cultured according to the manufacturer’s instructions (Life Technologies). Five µg of DNase-treated RNA (Ambion DNA-free Kit, Invitrogen) was used for cDNA synthesis using random hexamer primers as part of the Superscript III assay (Invitrogen). Each PCSK6 isoform was PCR amplified as follows: MyTaq™ DNA Polymerase (Bioline) was used for PCR reactions (4 µl Buffer Master Mix, 1 µl cDNA, 400 nM of each primer, 1 Unit Taq polymerase and made up to 20 µl with H_2O) with the following conditions: 60 s at 95 °C followed by 40 cycles of 10 s at 95 °C, 20 s at 60 °C, 30 s at 72 °C before finishing with 5 min at 72 °C. 1 µl of 6 X Orange-G loading dye was added to 5 µl PCR product and run on a 0.8% TAE agarose gel. PCR product clean-up was performed (2 µl of ExoSAP-IT for every 5 µl PCR product, 37 °C 15 min then 80 °C 15 min) before confirming all sequences by Sanger sequencing (DNA Sequencing and Services, Dundee). To minimise technical variation all cell lines were profiled using the same amount of starting tRNA, an equal amount of template cDNA in PCR and run on the same thermal cycler together. All primers throughout this project were designed using Primer3 (Rozen and Skaletsky, 2000), tested for primer specificity with Primer-BLAST (Ye *et al.*, 2012) and are listed in Appendix A.

4.3.2 PCSK6-AS annotation

RNA extraction, cDNA synthesis and PCR components were all conducted as above. The primer pair 5’-GGTGCAGAAAACAAGCCTG and 5’-CTTCCCTGCTGGCGTTTTTG was originally used to detect PCSK6-AS1 (spanning intron 1 from exon 1 to exon 2 of PCSK6-AS1, see Figure 4.12). When it became apparent there was more than one PCSK6-AS isoform I designed the primer pair 5’-GGTGCAGAAAACAAGCCTG 5’-AAAGGCAGGAAAACCAAAGT and 5’-GGTGCAGAAAACAAGCCTG 5’-TGCCAAAAGAGTTATAGGTGATT to distinguish between PCSK6-AS1 and PCSK6-AS2/3 lncRNA respectively. Conditions for both PCRs were as follows: 60 s at 95 °C followed by 40 cycles of 15 s at 95 °C, 15 s at 56 °C, 20 s at 72 °C before finishing with 5 min at 72 °C. 1 µl of 6 X Orange-G loading dye was added to 5 µl PCR product and run on a 1.8% TAE agarose gel. Bands were excised using QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer’s instructions, quantified on a Nanodrop 2000 and diluted to 20 ng/µl for Sanger sequencing (DNA Sequencing and Services, Dundee).

4.3.3 RNA-seq analysis

For RNA-seq analysis of the PCSK6 SI’s novel exon RNA-seq BAM files from Cold Spring Harbor Laboratories and Caltech were downloaded from the ENCODE Consortium’s UCSC data source (http://genome.ucsc.edu/ENCODE/downloads.html). All RNA reads were more than 200 nucleotides in length and were obtained as short reads from the Illumina GAIIx platform. The average depth of sequencing was ~200 million reads (100 million paired-ends). Biological replicates across multiple cell lines were obtained representing Poly-A+ and Poly-A- RNA from whole cells and subcellular compartments where available. For more information on ENCODE library preparation and sequencing methodology refer to Parkhomchuk *et al.* (2009). BWA v0.7.12 was used for mapping raw sequences against the human genome (hg38), after which SAMtools v1.2 was used for post-processing all RNA sequence read alignments in the BAM format. SAMtools view facility enabled the alignment of the 5 PCSK6 exons listed in Table 4.3. A shell script was developed with the help of Dr Miguel Pinheiro for the quantification of RNA-seq reads (Appendix F). For the RNA-seq analysis of PCSK6-AS1 I used the Human Body Map data: RNA-seq reads for over 16 tissue types were made publicly available by the EMBL-EBI (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/) and visualised via the NONCODE database (Xie *et al.*, 2014). Briefly, RNA was prepared for sequencing as follows: 1 µg of tRNA was subjected to two rounds of oligo-dT beads binding. Purified mRNA was fragmented and random primed for cDNA synthesis. cDNA fragments were end repaired, dATP added, and ligated to the Illumina pair-end sample prep adaptor. The ligated material was amplified by PCR for 15 cycles, and used for Illumina sequencing on a HiSeq 2000. A full protocol is detailed in Derrien *et al*. (2012). Additional samples submitted by the Rinn lab (Harvard) were also represented in Figure 4.11. These RNA-seq reads were also paired end but were sequenced on an Illumina GA2 rather than a Hiseq 2000, with 50,000 reads per sample. Full details for the Rinn lab samples submitted can be found in Cabili *et al*. (2011).

4.3.4 PCSK6 isoform expression in the developing and adult human brain

PCSK6 gene expression in the developing human embryonic brain was performed by querying early-late prenatal (8-38 post-conception weeks, PCWs) developmental transcriptome data of the Brainspan project (http://www.brainspan.org/rnaseq/search/index.html). Prenatal LMD microarray data for a male donor (H376.IIIA.02) aged 15 PCWs was also interrogated http://www.brainspan.org/lcm/search/index.html. Expression analysis of the SI in the adult human brain (Donor H0351.2001: 24 year old male) was performed using the Allen Brain Atlas (http://www.brain-map.org/).Relevant PCSK6 probes from the datasets were A_23_P151907, A_23_P390006, A_24_P189997 and CUST_9258_PI416261804. Gene expression was viewed using the Brain Explorer 2 (v2.3.5). For additional corpus callosum analysis pre-mRNA splicing patterns of PCSK6 in various tissues was available from an independent microarray dataset (NCBI dataset record GDS832). Oligonucleotide probes 36 nucleotides in length, centrally positioned with respect to each exon-exon junction were designed for all human RefSeq mRNA sequences having at least one exon-exon junction. For further details on library preparation see Johnson *et al*. (2003).

4.4 Results

4.4.1 PCSK6 isoform profiling

To define which cell lines might be suitable for the functional analysis of genetic variants at the PCSK6 locus I began by cataloguing alternatively spliced PCSK6 mRNA isoforms across a broad range of cells representative of brain and non-brain tissues alike (Figure 4.2). Different protein isoforms are generated through alternative splicing of pre-mRNA and genes tend to express many isoforms simultaneously (Djebali *et al.*, 2012). By analysing the different gene isoforms observed within and across cell lines we can also discern in which cell/tissue type the novel short PCSK6 isoform may be expressed; if a cell line does not express isoforms containing exons downstream of the bidirectional promoter (i.e. lanes 2-5, all panels, Figure 4.2) then it is unlikely to express the PCSK6 SI either.

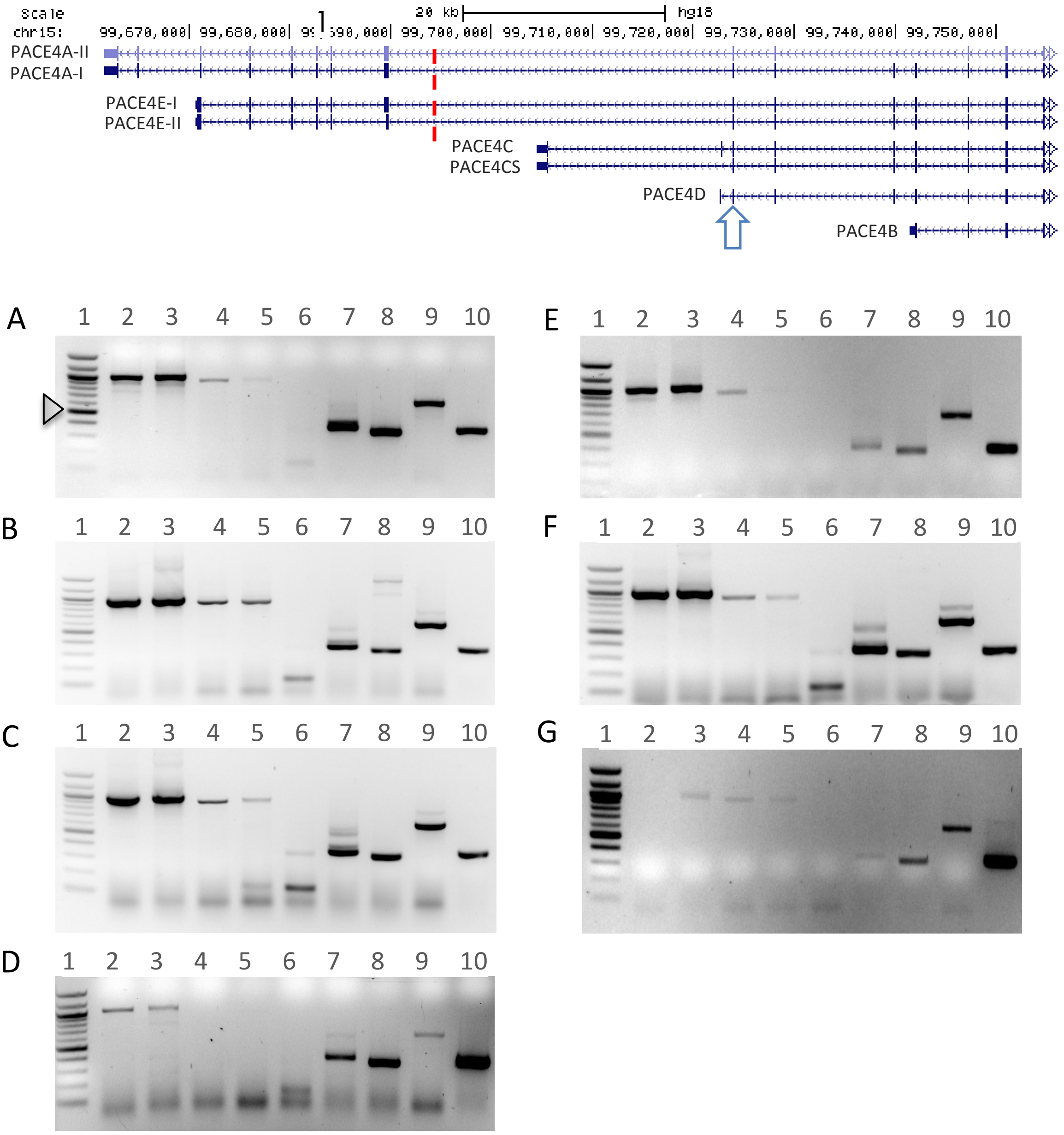
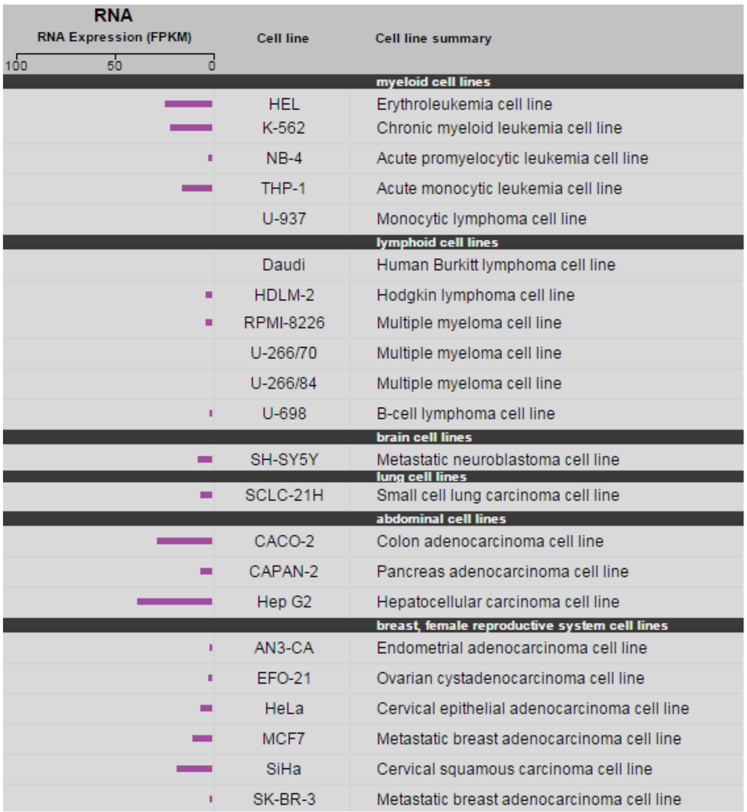


Figure 4.2 PCSK6 isoforms tested for expression in multiple cell lines. Only the 3' end of all transcripts has been included in the UCSC RefSeq tracks in blue since all recognised RefSeq isoforms share the same PCSK6 5' end. Dashed red line marks the approximate location of the bidirectional promoter and the novel exon. The canonical PCSK6 gene is labelled PACE4A-I. cDNA for the following cell lines was used for expression analysis: SH-SY5Y **(A)**, RPE-1 **(B)**, K562 **(C)**, hNSC **(D)**, HeLa **(E)**, HEK293 **(F)** and 1321N1 **(G)**. Lanes in each of the agarose gels are (Lane 2)PACE4-AI 998 bp, (3)PACE4-AII 992 bp, (4)PACE4E-II 927 bp, (5)PACE4E-I 921 bp, (6)PACE4C 128 bp, (7)PACE4CS 368 bp, (8)PACE4B 341 bp, (9)PACE4 611 bp and (10) β-Actin positive control 352 bp. The 100bp Gene O'Ruler (Lane 1) is marked by a grey triangle at 500bp in panel A. The blue arrow indicates exon 13 of the canonical PCSK6 gene.

Figure 4.2 shows the neuroblastoma SH-SY5Y (Panel A) and non-neuronal RPE-1 (B), K562 (C) and HEK293 (F) cell lines to express to varying degrees all eight PCSK6 isoforms discussed in Table 4.1. Only 3 isoforms are expressed in all cell lines; PACE4CS, PACE4B and PACE4D, and all are upstream of the predicted bidirectional promoter (red line, Figure 4.2). The astrocytoma glial 1321N1 is the only cell line not to express the canonical PCSK6 isoform at a detectable level (Lane 2, Panel G, Figure 4.2). However the primary difference between all 7 cell lines appears to be the extent of which bands in lanes 5 and 6 (isoforms PACE4E-I, and PACE4C respectively) are expressed. For example neither hNSC nor HeLa express these spliced isoforms while both K562 and HEK293 display an additional unidentified band in lane 6, possibly a hitherto unknown spliced isoform. It is worth noting lane 6 (PACE4C) is the isoform immediately upstream of the genetic variants associated with handedness (rs11855415, VNTR) and the predicted bidirectional promoter.

The broad range of isoform expression coupled with the histone modification markings discussed in Chapter 3 (Figure 3.2) and the comparatively high protein expression as indicated by the Human Protein Atlas data in Figure 4.3 makes K562 a suitable candidate cell line for the functional testing of genetic variants within PCSK6.



Cell line	Cell line summary
	myeloid cell lines
HEL	Erythroleukemia cell line
K-562	Chronic myeloid leukemia cell line
NB-4	Acute promyelocytic leukemia cell line
THP-1	Acute monocytic leukemia cell line
U-937	Monocytic lymphoma cell line
	lymphoid cell lines
Daudi	Human Burkitt lymphoma cell line
HDLM-2	Hodgkin lymphoma cell line
RPMI-8226	Multiple myeloma cell line
U-266/70	Multiple myeloma cell line
U-266/84	Multiple myeloma cell line
U-698	B-cell lymphoma cell line
	brain cell lines
SH-SY5Y	Metastatic neuroblastoma cell line
	lung cell lines
SCLC-21H	Small cell lung carcinoma cell line
	abdominal cell lines
CACO-2	Colon adenocarcinoma cell line
CAPAN-2	Pancreas adenocarcinoma cell line
Hep G2	Hepatocellular carcinoma cell line
	breast, female reproductive system cell lines
AN3-CA	Endometrial adenocarcinoma cell line
EFO-21	Ovarian cystadenocarcinoma cell line
HeLa	Cervical epithelial adenocarcinoma cell line
MCF7	Metastatic breast adenocarcinoma cell line
SiHa	Cervical squamous carcinoma cell line
SK-BR-3	Metastatic breast adenocarcinoma cell line

Figure 4.3 PCSK6 expression across a range of cell lines according to the Human Protein Atlas (HPA). The HPA (Uhlen *et al*., 2015) displays the RNA-levels in purple bars for each cell line, which represent an aggregate expression of multiple isoforms expressed in that cell line. The cell lines are ordered according to cellular origin. RNA values are represented as FPKM (Fragments Per Kilobase of transcript per Million fragments sequenced) and further described here: http://www.proteinatlas.org/about/assays+annotation#rna

4.4.2 Annotating the PCSK6 SI

Analysis of the PCSK6 isoforms also shows the canonical gene's exon 13 (blue arrow, Figure 4.2) to be present in all cell lines and therefore acts as a useful marker; the novel exon of the shorter PCSK6 isoform is predicted to be the first of that transcript and a PCR using a primer pair (green arrows & Panel C, Figure 4.4) spanning from the novel exon to exon 13 demonstrates the novel exon is not part of a longer known isoform but belongs to a transcript whose transcriptional start site (TSS) originates within the bidirectional promoter. A second primer pair spanning from the novel exon to the 3'UTR of the canonical PCSK6 gene (yellow line, Figure 4.4) further indicates the shorter PCSK6 isoform not to be four exons in length as previously thought (Kimura *et al*., 2006) but fully extends the full length of the PCSK6 gene. Sanger sequencing of the

PCR product (accession #LN714797) confirmed the PCSK6 SI isn’t a truncated product of random bidirectional promoter activity, but a full-length gene isoform complete with 3’ UTR and poly-A tail. This PCR analysis provided the first evidence that the predicted bidirectional promoter within a locus associated with handedness is driving transcription of a novel PCSK6 SI.

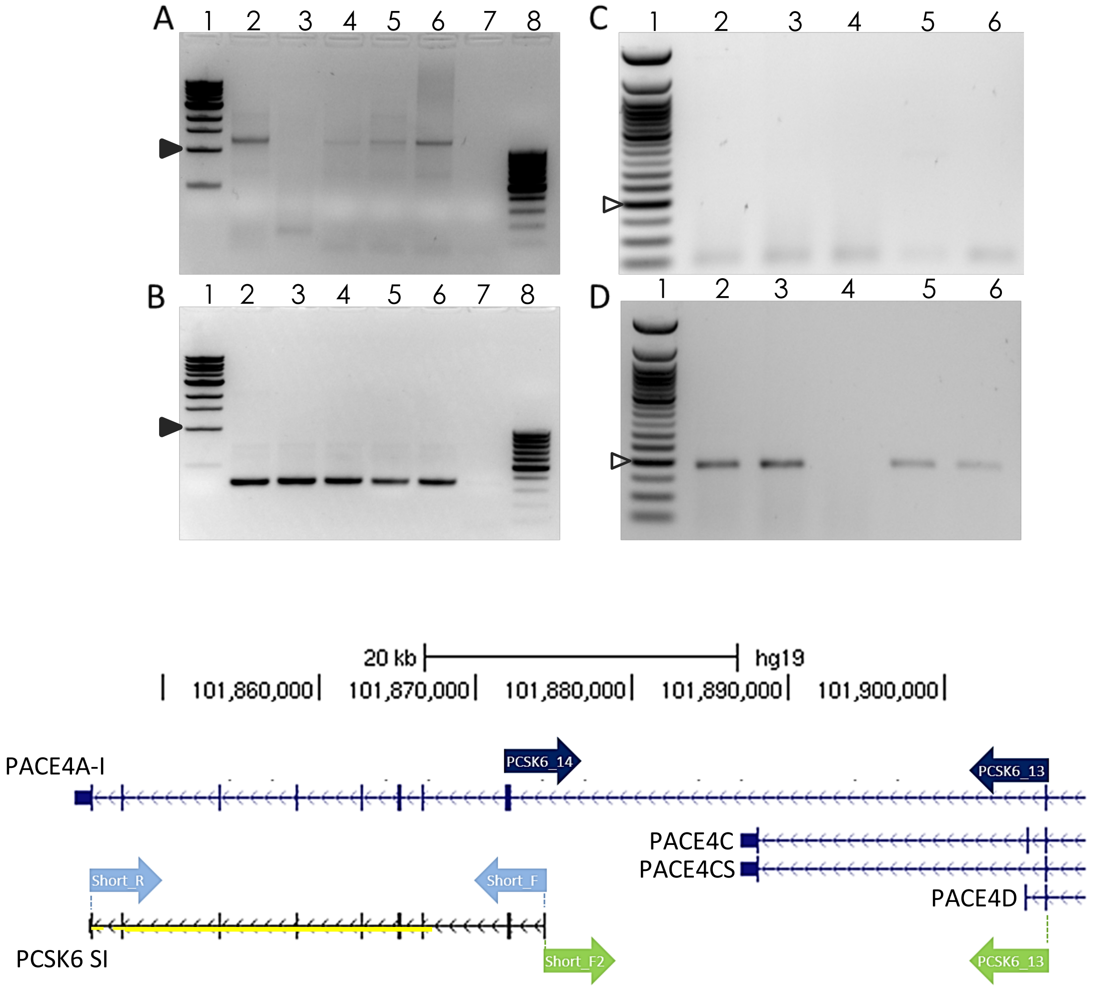


Figure 4.4 Annotating the shorter PCSK6 isoform.All PCR assays used oligo d(T) cDNA derived from multiple cell lines; K562 (Lane 2, HeLa(3), 1321N1(4), hNSC(5) and SH-SY5Y(6). In panels A&B there are two ladders; black triangle indicates 1,000 bp on a 1 kb NEB ladder in lane 1, 100 bp Gene O'Ruler ladder in lane 8. In panels C&D a white triangle indicates the 200 bp step on a 50 bp NEB ladder. **(A)** Represents the PCR product of the light blue arrows, the primer pair used for gene expression analysis of the novel PCSK6 SI (#LN714797). This product (1,247 bp) spans from the first exon of the novel SI to the 3'UTR of the canonical PCSK6 gene. The yellow line indicates the full extent of the transcript, exceeding the four exon length previously found (Kimura *et al.*, 2006) **(B)** A β-Actin primer pair (product 353 bp) was used as a positive control for cDNA synthesis **(C)** PCR product of the green arrows which represent the primer pair (product approximately 300 bp) used to indicate the shorter PCSK6 isoform does not share the upstream exon 13 belonging to the canonical PCSK6 gene **(D)** PCR product of the dark blue arrows which indicate a primer pair used as a positive control to demonstrate a product of 211 bp spanning exon 13 and 14 was present in all but one of the cell lines tested (1321N1, Lane 3). See Appendix A for all labelled primer sequences. Labelled are the PCSK6 isoforms of relevance (see Figure 4.2).

4.4.3 PCSK6 SI protein prediction

The function of the PCSK6 SI is still unknown, however sequencing of the SI (blue arrows, Figure 4.4) provides a 9-exon 1.2 kb sequence (Chromosome 15: 101305125-101334254, see Appendix C.3) which can be translated and used as the basis for function prediction according to Interproscan (Mitchell *et al*., 2015), thereby enabling the functional analysis of the SI by classifying the sequence into known families, predicting domains and cleavage sites in the process (Figure 4.5).

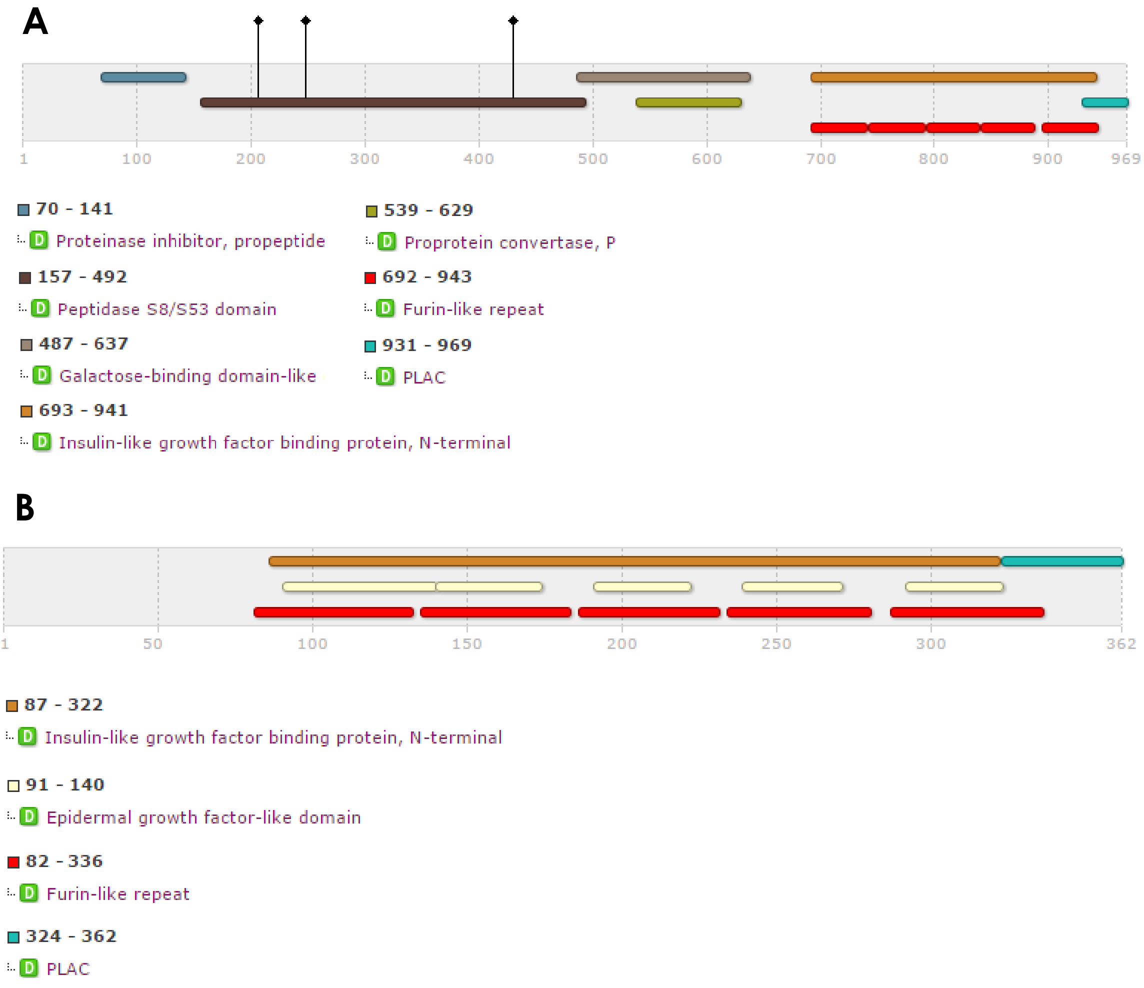

A
1
100
200
300
400
500
600
700
800
900
969
70 - 141
D Proteinase inhibitor, propeptide
157 - 492
D Peptidase S8/S53 domain
487 - 637
D Galactose-binding domain-like
693 - 941
D Insulin-like growth factor binding protein, N-terminal
539 - 629
D Proprotein convertase, P
692 - 943
D Furin-like repeat
931 - 969
D PLAC
B
1
50
100
150
200
250
300
362
87 - 322
D Insulin-like growth factor binding protein, N-terminal
91 - 140
D Epidermal growth factor-like domain
82 - 336
D Furin-like repeat
324 - 362
D PLAC

Figure 4.5 The predicted protein domains of the PCSK6 isoforms.The top panel displays the canonical isoform **(A)** and the bottom panel displays the shorter PCSK6 isoform **(B)** as identified by Interproscan. The amino acid (aa) length on the x-axis indicates the location of each predicted domain. The active sites predicted by the Scanprosite software (de Castro *et al*., 2006)) at aa 205, 246 and 420 are shown (black diamonds): these represent conserved aspartate, histidine and serine – the catalytic triad. Protein sequences for both isoforms are included in the Appendix C.4. See Table 4.2 for additional protein properties.

Several substantial differences are predicted to exist between the canonical PCSK6 protein (4,409 bp/969 aa) and the SI (1,222 bp/362 aa). As a proprotein convertase PCSK6 is synthesised as an inactive precursor (or 'zymogen') that is chaperoned through the cell by its prodomain, of which the SI lacks. Prodomains are known be involved in protein folding and activation (Thomas, 2002), the loss of which results in a lack of an autoproteolytic initial cleavage step thereby rendering the protein inactive and localised to the endoplasmic reticulum (Leduc *et al.*, 1992). For the full length canonical PCSK6 protein this is followed by a pH- and calcium-dependent cleavage event in the trans-Golgi network–endosomal compartments that completes its activation (Anderson *et al.*, 2002). The SI is also predicted to lack a catalytic domain which contains the catalytic triad necessary for cleavage. Both PCSK6 isoforms are predicted at a minimum to have a signal peptide/transmembrane domain at the N-terminus (Table 4.3) which directs translocation in to the endoplasmic-reticulum (ER), an insulin-like growth factor domain which confers protein-protein interaction properties and directs cell surface tethering and a PLAC (protease and lacunin) domain, a Cys-His-rich domain of about 40 residues at the carboxyl-terminus.

In short, analysis of the SI at a predicted protein level suggests its molecular function may involve a growth factor receptor domain IV (Figure 4.5 & Table 4.2), the fourth extracellular domain common in all receptor tyrosine protein kinases which influences the regulation of ligand binding to receptor domains (Cho and Leahy, 2002).

Table 4.2 Properties of the canonical PCSK6 protein and the predicted shorter PCSK6 isoform

Protein	*PCSK6 (NP_002561.1)*	*PCSK6 SI (#LN714797)*
Gene Isoform	PACE4A-I	To be confirmed
Length (cDNA bp/protein aa)	4409bp/969 aa	1222bp/362 aa
Biological Process	Release of mature proteins from their proproteins by cleavage at the RX(K/R)R consensus motif	None predicted
PTM (aa)	Phosphorylation: S291,Y298,S303,Y637,T755,S757, T815 Ubiquitylation: K136 Glycosilation: N259, N914, N932	Unknown
Molecular Function	Peptidase_S8 (PF00082) P_proprotein (PF01483) S8_pro-domain (PF16470) GF_recep_IV (PF14843) PLAC (PF08686)	GF_recep_IV (PF14843) PLAC (PF08686)
Signal peptide	Prediction: Yes Probability: 0.999	Prediction: Yes Probability: 0.958
Enzymatically active	Yes, secreted	Unknown

Note: ExPasy Translate (http://web.expasy.org/translate/) was used to translate DNA to the equivalent amino acid sequence. Post translational modification (PTM) identification was via the curated Phosphosite service (www.phosphosite.org, Hornbeck *et al.*, 2004) and the molecular function was predicted using Pfam (pfam.xfam.org, Finn *et al.*, 2014). Signal peptide (http://www.cbs.dtu.dk/services/SignalP-3.0/, reference here) predicted cleavage probability. The protein sequences used to predict the properties for both isoforms are included in the Appendix C.4.

4.4.4 RNA-seq and microarray analysis of the PCSK6 SI

Given the predicated catalytic inactivity of the SI at the protein level, I focused my analysis on whether the SI might have a regulatory effect at the RNA level and the cellular subcompartment in which this might occur since an interrogation of different subcellular RNA fractions can provide a snapshot of the status of an mRNA transcript along the RNA processing pathway. The large dynamic range, public availability of data and the ability to detect previously unknown transcripts means RNA-seq is the method of choice for the profiling of the novel first exon, a sequence of mRNA that differentiates the SI from all other known PCSK6 isoforms. Five sequences were designed for querying the ENCODE Consortium's RNA-seq data which enabled quantification of exons upstream and downstream of the novel exon (Table 4.3 & Figure 4.6) in separate cellular fractions (nucleus, cytosol and whole cell). The analysis of RNAs isolated from different subcellular fractions also provides data concerning compartment-specific relative steady-state abundance and the post transcriptional processing state for each of the detected transcripts. For example, the ENCODE RNA-seq data also divides quantification in to poly(A)$^+$ (enriched with mature mRNA) and poly(A)$^-$ signals (a variety of other RNA types such as pre-mRNA and ncRNA which usually lack a poly(A) tail). Attached to the 3' end of mRNAs, the poly(A) tail protects the mRNA molecule from enzymatic degradation in the cytoplasm and aids in transcription termination, nuclear export and translation (Guhaniyogi and Brewer, 2001).

Table 4.3 RNA-seq probes for the PCSK6 regions of interest. PCSK6 sequences were designated for interrogating the ENCODE RNA-seq datasets for quantification of the novel SI exon (3) and exons upstream (1,2) and downstream (4,5). Sequence probe locations were based on previous PCR primer sequences used in the 'PCSK6 isoform profiling' assay (see Figure 4.2) and as such known to be expressed.

number	*exon of canonical PCSK6*	*location (chr15,hg38)*	*span (bp)*
1	1 of 22 (5'UTR)	101489374-101489984	611
2	7 of 22	101398404-101398576	173
3	Not Applicable	101334191-101334254	63
4	17 of 22	101324850-101325046	197
5	22 of 22 (3'UTR)	101303928-101305355	1428

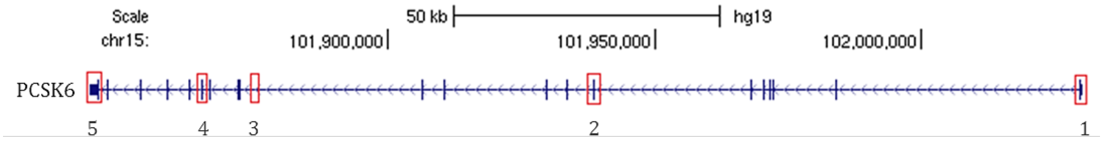


Figure 4.6 RNA-seq probes for PCSK6 regions of interest. Red boxes indicate the location of the RNA-seq probes discussed in Table 4.3

Interrogation of the RNA-seq database indicates a consistent bias in the number of reads quantified at the 3’ UTR in all cell lines (Table 4.4). This most likely relates to selection bias due to RNA fragmentation during the poly(A) mRNA selection before the random priming cDNA synthesis. This inverse relationship between RNA stability and bias towards 3' UTRs is an effect well documented in the literature (e.g. Wang *et al*. (2012)) and especially prevalent when a combined oligo-dT and random hexamer approach is used during cDNA synthesis, as is the case for the ENCODE RNA-seq datasets. A bias towards reads mapping to the 3’ UTR can also occur due to the often larger size of the 3’UTR in comparison to other exons.

Table 4.4 RNA-seq raw reads count for multiple PCSK6 exons. Each numbered entry indicates average count value of raw sequencing reads mapped to the respective exon in each cell line (see Table 4.3 & Figure 4.6 for probe locations).

		3'UTR	Exon 17	Novel	Exon 7	5'UTR
H1hESC	Cell Poly -	68.75	10	0.75	30.5	0
	Cell Poly +	166.25	20.25	0.5	18.75	6.25
	Cytosol Poly -	7.5	1	0	2.5	3
	Cytosol Poly +	178	12	0	15.5	7
	Nucleus Poly -	47	4.5	0.5	8	4.5
	Nucleus Poly +	145	21.5	2.5	20	10
K562	Cell Poly -	187.5	29	24	110.25	12.25
	Cell Poly +	1031.8	161.5	0.75	196.75	22.5
	Cytosol Poly -	22.5	6.5	0	11.5	6.75
	Cytosol Poly +	859.5	147.5	0.5	346.25	83.5
	Nucleus Poly -	126.75	31.5	11.25	94.5	43.75
	Nucleus Poly +	1789	397.75	6.25	307.5	117.25
HeLa S3	Cell Poly -	1950.3	475.25	59.25	412.25	21.5
	Cell Poly +	1187.3	142.75	0.5	88.5	16.25
	Cytosol Poly -	0.5	0	0	0	2
	Cytosol Poly +	1435	335	0	264.25	68
	Nucleus Poly -	173.25	60.75	5.75	64.75	48
	Nucleus Poly +	1111.5	295.5	1.5	155.5	51.25
Hep G2	Cell Poly -	2380.5	625.5	113.5	332.5	55.25
	Cell Poly +	9301.5	1504.3	1	936.75	216.75
	Cytosol Poly -	369.5	109.75	0.25	93.75	60.25
	Cytosol Poly +	8979.5	1798.8	0	1419.8	434
	Nucleus Poly -	500.5	132.25	19.25	168.25	79.5
	Nucleus Poly +	3652	810	18.5	633.75	190.25
SK-N-SH	Cell Poly +	76	0.25	0.5	1.5	13
	Cytosol Poly +	112.75	6.25	0	3.75	12.75
	Nucleus Poly +	552.5	0	0	0.75	8.5
A549	Cell Poly -	230.5	35	21.5	65.25	8.25
	Cell Poly +	948.5	169.5	0.5	91.25	23.5
	Cytosol Poly +	793.5	230	0	296.25	82.75
	Nucleus Poly +	722.25	174.75	0	181	36
PFSK1*	Cell Poly +	375.5	85	0	142.5	100.5
BE2C*	Cell Poly +	102.5	26.5	0	36.5	22.5

Note: Single 75 nucleotide (nt) directed reads and paired 75nt reads were combined to give an average value for mRNA with (Poly+) and without (Poly-) a Poly-A tail. Source data was provided by the ENCODE project Cold Spring Harbour lab and where marked (*), Caltech. Heat map indicates low (green) to high (red) expression levels.

Previous findings which suggest a signal peptide for K562 cell mRNAs containing the novel exon (Table 4.2) are not at odds with the results in Table 4.4; in the K562 RNA-seq data 6.25 reads quantified have a poly(A) tail (compared to 11.25 poly(A)$^-$ reads), indicating possible export to the cytoplasm.

So far a mixed picture emerges for the possible function of the SI: protein analysis suggests an inactive enzyme with a signal peptide that is probably not secreted but is predicted to possess a domain which influences the regulation of ligand binding to receptor domains. RNA-seq data allows for an export of SI mRNA to the cytoplasm in human embryonic stem cells, though in other cell lines a substantial bias exists for poly(A)$^-$ transcripts, indicating likely retention in the nucleus for, as yet, unknown reasons.

This cursory finding, that the nuclear export of the short isoform is dominant in hESCs only, is curious in the context of the known role PCSK6 plays in the left-right axis determining pathway during embryonic development. An extensive search of the scientific literature provides minimal elucidation on PCSK6 isoform expression in the developing embryo though several recent large scale efforts such as the Brainspan atlas (BrainSpan, 2011) allow a survey of gene expression in specific brain regions.

The Brainspan atlas of the developing human brain uses post-mortem human brain specimens for studying transcriptional mechanisms and gene expression involved in human brain development. Figure 4.7 displays the results following an interrogation of the Brainspan atlas RNA-seq dataset for the canonical PCSK6 gene expression across 20 donors. Results are displayed in a tempo-spatial manner according to anatomical region and developmental period (8-38 PCWs). A first inspection of the data show PCSK6 is expressed primarily in the hippocampus (HIP) up to and including 13 PCWs after which expression shifts to the medial frontal cortex (MFC) and cerebellar cortex (CBC). There is also a spike in PCSK6 expression at week 16 in the dorsolateral prefrontal cortex (DFC), primary motor cortex (M1C) and ventrolateral prefrontal cortex (VFC).

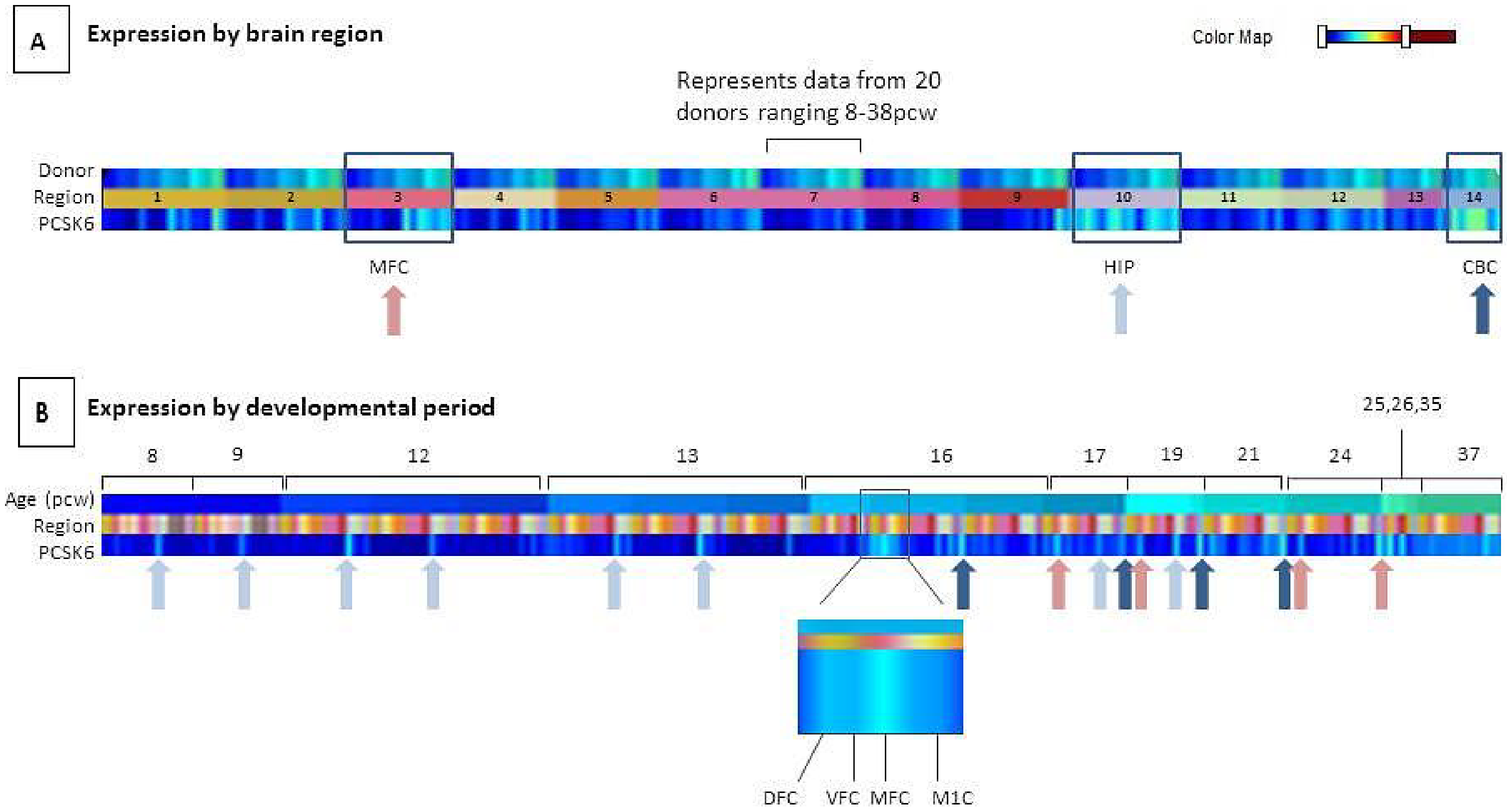
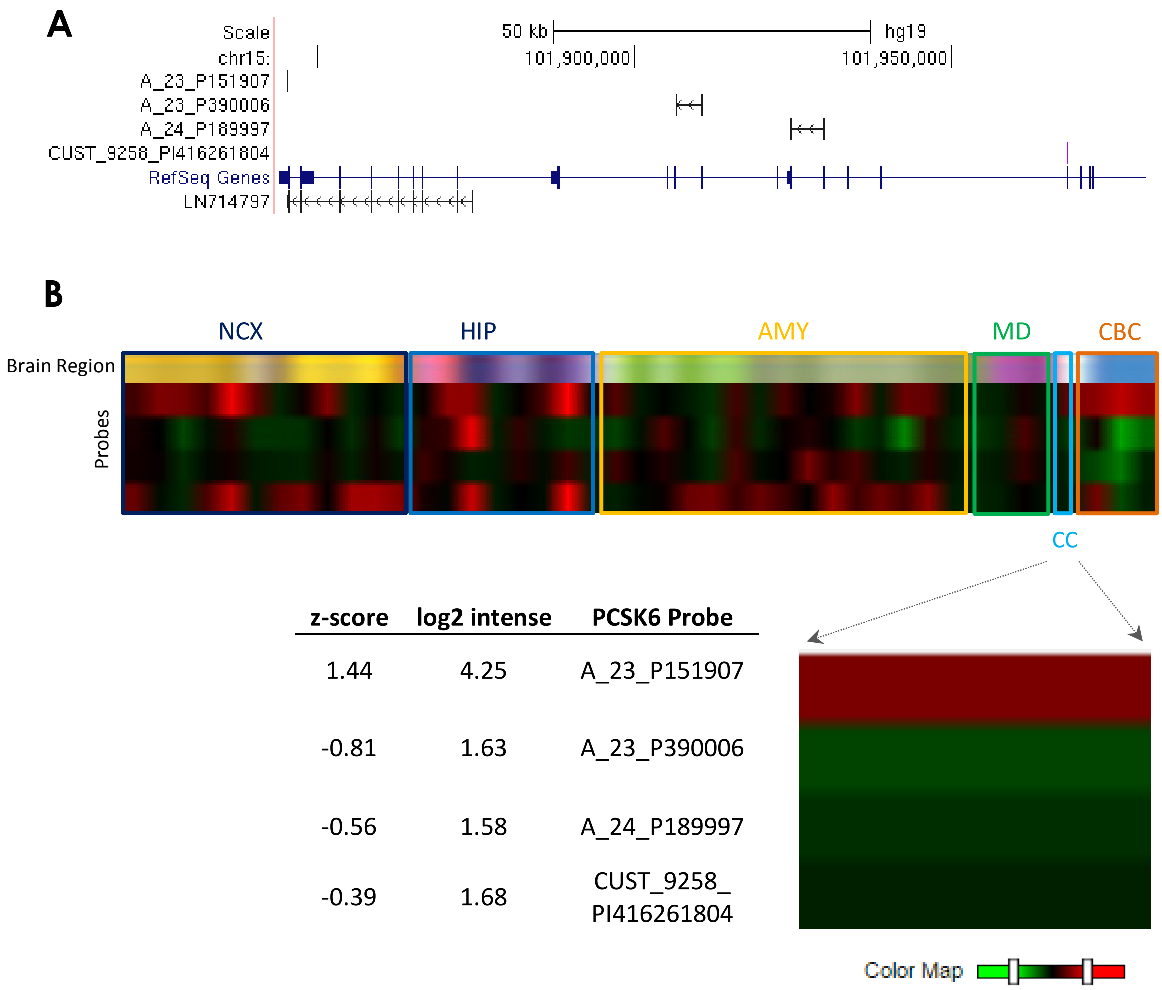


Figure 4.7 PCSK6 gene expression in the developing human embryonic brain. **(A)** PCSK6 gene expression is displayed according to developmental transcriptome data of the Brainspan project and organised by anatomical region with blue boxes indicating higher relative expression in the Medial Frontal Cortex ((3) MFC, pink arrow), Hippocampus ((10), HIP, light blue arrow) and Cerebellar Cortex ((14), CBC, dark blue arrow) throughout the developmental period 8-38 PCW. Other regions interrogated are numbered as follows: (1) Dorsolateral prefontal cortex, (2) Ventrolateral profontal cortex, (4) Orbital frontal cortex, (5) Inferior parietal cortex, (6) Primary auditory cortex, (7) Posterior superior temporal cortex, (8) Inferolateral temporal cortex, (9) Primary visual cortex, (11) Amygdaloid complex, (12) Striatum, (13) Mediodorsal nucleus of thalamus **(B)** PCSK6 gene expression is displayed by developmental period. The heat map represents the normalised gene level RNA-Seq expression data in RPKM (Reads Per Kilobase of exon model per Million mapped reads) ranging from lower expression (dark blue) towards higher expression (light blue/turquoise). Dorsolateral prefrontal cortex (DFC), primary motor cortex (M1C), ventrolateral prefrontal cortex (VFC). Colour map scale ranges from dark blue (- log2 RPKM: 0.13) to bright turquoise (- log2 RPKM: 2.7).

The Brainspan atlas also provides for the analysis of gene isoform expression in the developing embryonic brain via exon microarray datasets (Figure 4.8). Four probes were available for PCSK6 gene expression analysis; 3 upstream of the novel exon and 1 at the 3' UTR. Such an arrangement allows an approximate estimation of the distribution of the SI throughout the brain's anatomical regions since, unlike previous RNA-seq analysis, higher expression of the 3'UTR probe (A_23_P151907) isn't likely to be a bias of cDNA synthesis since the RNA was fragmented in such a way so as to provide a more even distribution of reads across the length of the RNA molecules. In addition, random hexamer rather than oligo-dT primers were used throughout first-strand cDNA synthesis.



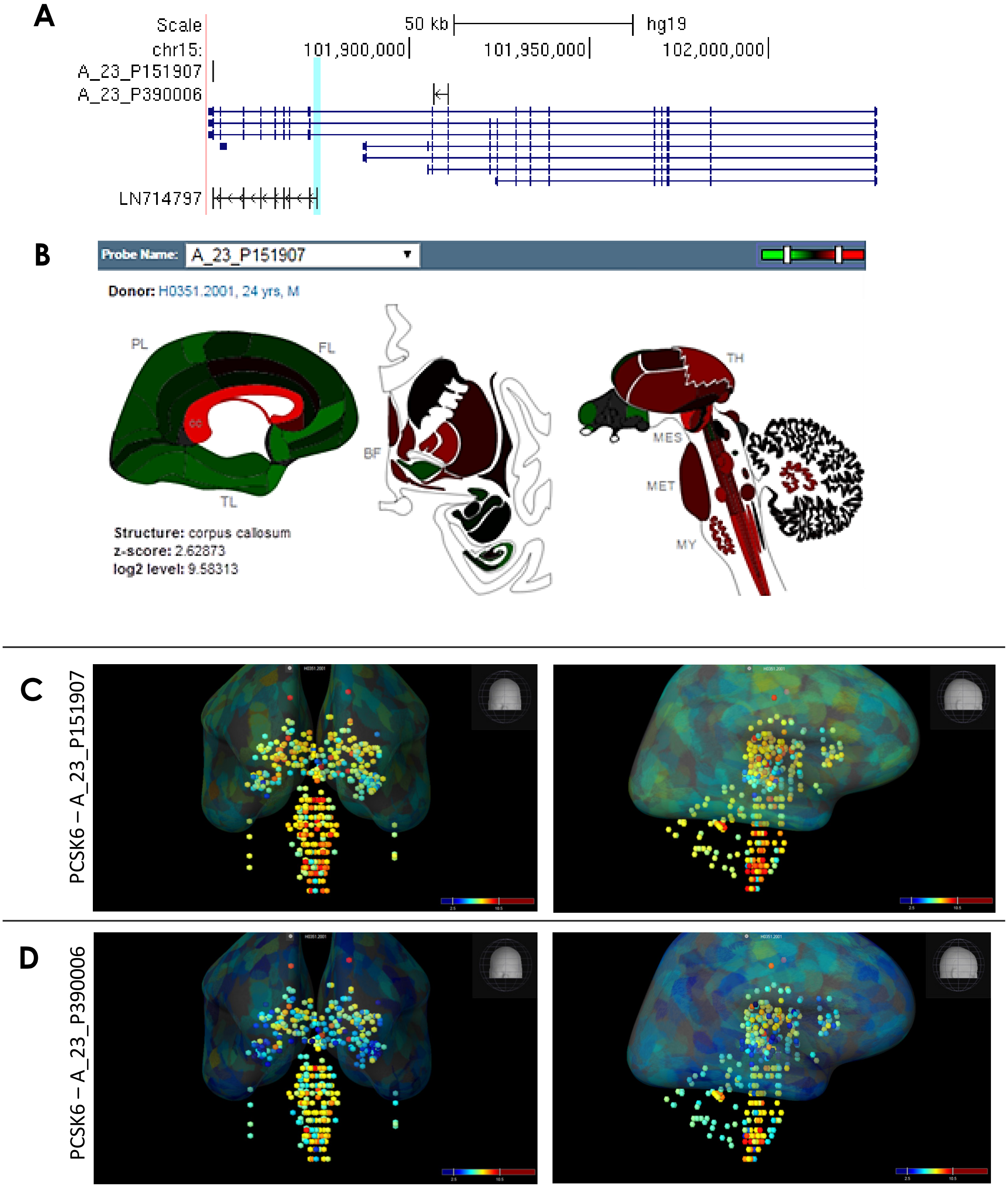
z-score	log2 intense	PCSK6 Probe
1.44	4.25	A_23_P151907
-0.81	1.63	A_23_P390006
-0.56	1.58	A_24_P189997
-0.39	1.68	CUST_9258_ PI416261804

Figure 4.8 PCSK6 expression in the developing human brain **(A)** The location of the 4 PCSK6 probes as used in the Brainspan prenatal LMD microarray are displayed in the UCSC genome browser (A_23_P151907, A_23_P390006, A_24_P189997 and CUST_9258_PI416261804). The shorter PCSK6 isoform LN714797 (black) is included relative to the 3' end of PCSK6 (blue) **(B)** Indicates PCSK6 probe expression in the corpus callosum (CC) in a male donor (H376.IIIA.02) aged 15 PCW. Microarray data is presented in a heat map format in the green - red scale; green represents low expression values while red represents high expression values. CC region displays higher expression of the 3'UTR probe (A_23_P151907) than probes upstream. Other brain regions displayed are neocortex (NCX), hippocampus (HIP), amygdala (AMY), medial dorsal nucleus (MD) and cerebellar cortex (CBC). Colour map FPKM values range from bright green (- z-score: -1.93, log2 intensity: 0.78) to bright red (- z-score: 3.05, log2 intensity: 7.78).

On the whole, the neocortex (NCX) and HIP show equal expression of probes placed at the 3'UTR and the 5' end of the gene. This differs to both the corpus callosum (CC) and CBC, which display a marked difference in expression between the higher expression of 3'UTR probe (A_23_P151907) and the relatively lower expression of all other probes upstream of the novel exon. In this context, the previous Brainspan RNA-seq signal (track B, Figure 4.7) indicating PCSK6 expression in the CBC in the late prenatal developmental period (16 - 24 PCWs), may be driven in part by the shorter PCSK6 isoform.

The Allen Brain Atlas microarray analysis of PCSK6 isoform expression in adult human brain also displays relatively high expression in the CC of a probe downstream to the novel exon (track A&B, Figure 4.9). The blue tone of Figure 4.9 track D indicates a lower expression of the full length probe relative to the shorter probe visualised in track C. This graphic interpretation of probe quantification provides further confirmation of transcripts probed at the 3' UTR to be more abundant across the midbrain, CC and spinal cord.

Finally, I accessed an independent microarray dataset (NCBI dataset record GDS832, Johnson *et al*. (2003)) to interrogate for alternative pre-mRNA splicing of PCSK6 in various tissues. Track A of Figure 4.10 shows the location of the three microarray probes of interest spanning Exons 1-2 (5'UTR), Exons 13-14 (spanning the predicted secondary promoter) and Exons 21-22 (3' UTR). Note exon 14 of the canonical PCSK6 gene is also exon 2 of the shorter PCSK6 isoform. Interestingly, and in support of previous findings, of all 51 cell and tissue types analysed, the probe spanning exons 13-14 of the canonical PCSK6 gene (and by extension the bidirectional promoter, rs11855415 SNP and VNTR) is highest in the CC (track C, Figure 4.10). The probe spanning exons 21-22 is second highest in expression terms in the CC. These results suggest exon 2 of the SI may be contributing towards the increased overall probe signal. The 5'UTR probe on the other hand reports the highest probe signal in the spinal cord, as reported previously (Figure 4.10, track B & Figure 1.3).



A_23_P151907

Donor: H0351.2001, 24 yrs, M

Structure: corpus callosum
z-score: 2.62873
log2 level: 9.58313

Figure 4.9 PCSK6 isoform expression in the adult human brain. **(A)** Microarray probes from the Allen Brain Atlas (Hawrylycz *et al*., 2012) both downstream (A_23_P151907) and upstream (A_23_P390006) of the novel exon are shown in the UCSC genome browser relative to the shorter PCSK6 isoform (black). Turquoise bar indicates predicted secondary promoter **(B)** Allen brain atlas microarray data indicates relatively high expression of the PCSK6 gene in the corpus callosum (CC), as indicated in bright red. Other areas of note for high PCSK6 expression are the basal forebrain (BF), thalamus (TH), mesencephalon (MES), metencephalon (MET) and medulla oblongata (MY). **(C)** The probe downstream of the novel exon (A_23_P151907) shows a higher expression in a 3-D representation for the sampled brain (Donor H0351.2001: 24yrs, male) relative to **(D)** a probe upstream of the novel exon (A_23_P390006). Colours indicate the level of expression for the PCSK6 probe selected with lower (blue), mid (yellow) and high (red) expression indicated across multiple regions. Anterior (left) and sagittal (right) views were generated using Brain Explorer 2 (v2.3.5).

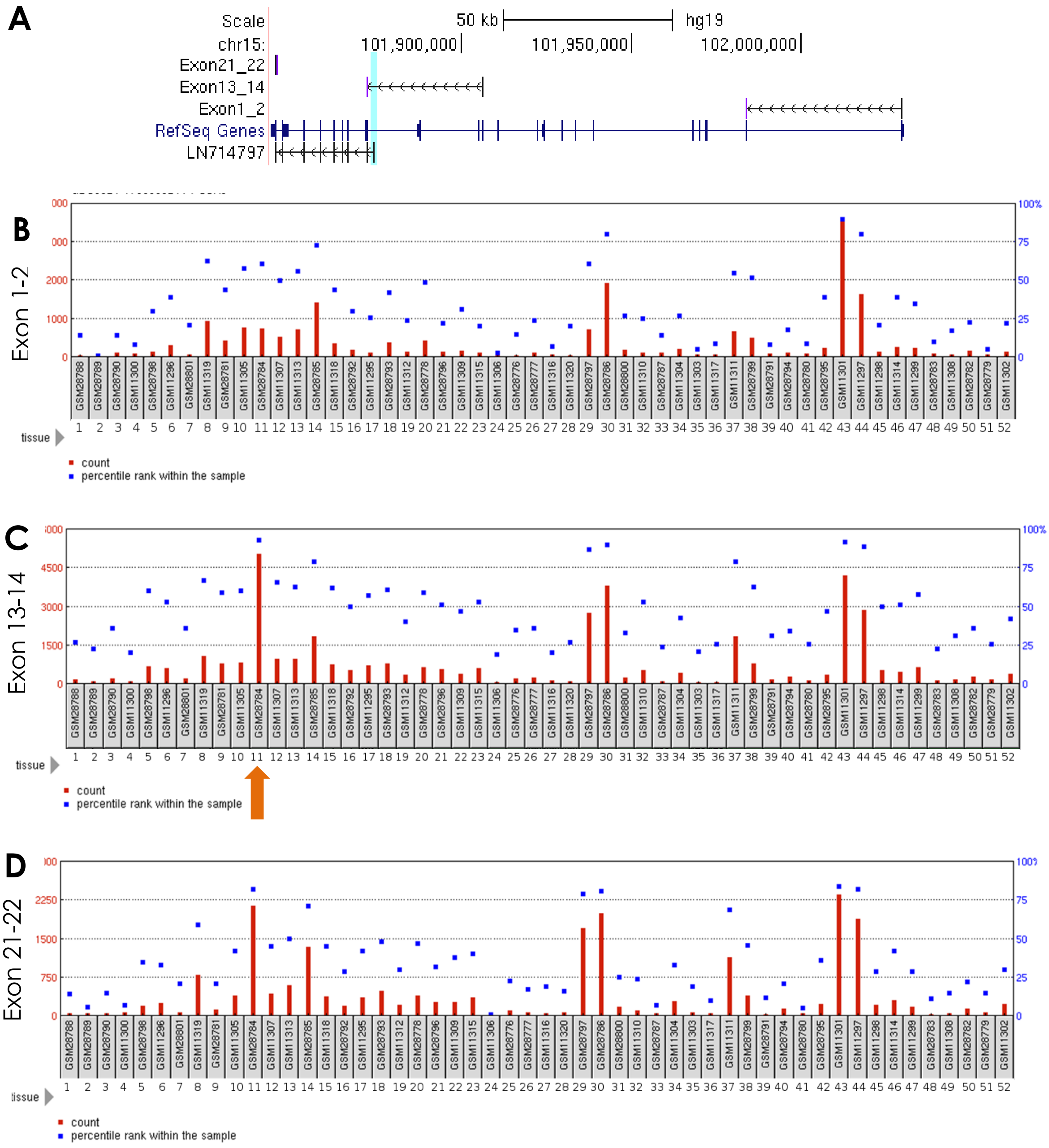


Figure 4.10 Alternative pre-mRNA splicing of PCSK6 in various tissues. **(A)** Several PCSK6 microarray probes are shown relative to the 3' end of the PCSK6 gene (blue) and shorter PCSK6 isoform (black). These tracks show the expression of the PCSK6 gene probed in the NCBI dataset record GDS832 at **(B)** Exons 1-2 (5'UTR), **(C)** Exons 13-14 (spanning the predicted secondary promoter) and **(D)** Exons 21-22 (3' UTR). The orange arrow in track C indicates PCSK6 expression in the corpus callosum (sample GSM28784, grey code). See (Johnson *et al.*, 2003) for further details on RNA preparation protocol. 1.adrenal cortex 2.adrenal medulla 3.bladder 4.bone marrow 5.brain 6.amygdala 7.brain fetal 8.caudate nucleus 9.cerebellum 10.cerebral cortex 11.corpus callosum 12.hippocampus 13.post central gyrus 14.thalamus 15.colon descending 16.colon transverse 17.colorectal adenocarcinoma 18.duodenum 19.epididymis 20.heart 21.ileum 22&23.jejunum 24.kidney 25.kidney fetal 26.leukemia chronic myeloid 27.leukemia lymphoblastic 28.leukemia promyelocytic 29.liver 30.liver fetal 31.lung 32.lung fetal 33.lung carcinoma A549 34.lymph node 35.lymphoma burkitt daudi 36.lymphoma burkitt ra 37.mislabelled 38.placenta 39.prostate 40.retina 41.salivary gland 42.skeletal muscle 43.spinal cord 44.spleen 45.stomach 46.testes 47.thymus 48.thyroid 49.tonsil 50.trachea 51.uterus 52.uterus corpus

4.4.5 RNA-seq analysis of PCSK6-AS

Much of this chapter has been concerned with expression of the novel exon and by extension the shorter PCSK6 isoform, with results so far highly suggestive of transcription being driven from a secondary bidirectional promoter. On the antisense strand, the bidirectional promoter is suspected of driving transcription of a 3-exon lncRNA PCSK6-AS1 (639bp, chr15:101874641-101877633), first reported by the ENCODE sub-project GENCODE (Harrow *et al*., 2012). Like the novel exon, no microarray data exists in the literature for this lncRNA since no probes specific to the 3 exons exist. As such we rely on existing RNA-seq data to help define the expression profile for PCSK6-AS1 (Figure 4.11).

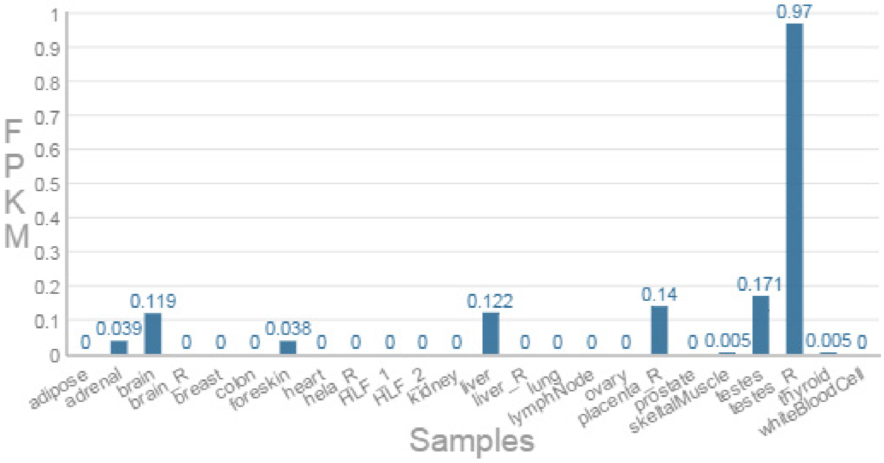


Figure 4.11 PCSK6-AS1 expression across a range of tissues. RNA-seq data expressed by Fragments Per Kilobase of transcript per Million mapped reads (FPKM) was sourced from the Human Body Map 2.0 (Flicek *et al*., 2014). HLF=human lung fibroblasts. Samples labelled with R represent additional samples as submitted by the Rinn lab (Harvard).

4.4.6 Annotation of the PCSK6-AS isoforms

Among human tissues, both protein-coding and lncRNA transcripts are reported to show the greatest heterogeneity in testes and brain (Ramskold *et al.*, 2009) with PCSK6-AS1 displaying a similar trend (Figure 4.11). Furthermore, as a class in general, lncRNAs are preferentially localised to the nucleus and chromatin of the cell (Derrien *et al.*, 2012). Although PCSK6-AS1 does appear in publicly available databases (e.g. http://www.noncode.org/), the UCSC genome browser still only classes PCSK6-AS1 as having a transcription support level of 'tsl3', meaning support for the existence of this lncRNA is from a single EST. As such, as a start point for any regulatory cis-acting role PCSK6-AS1 may have on the sense transcript, I confirmed both the existence (Figure 4.12) and sequence (see Appendix C.1) of this lncRNA. Cell lines were selected based on previous PCSK6 profiling and RNA-seq findings in an effort to maintain consistency across experiments.

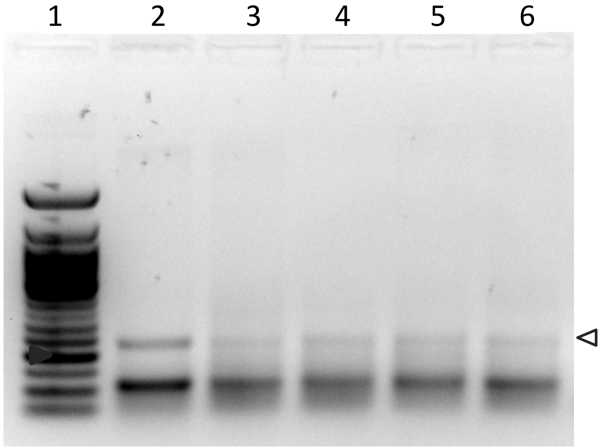


Figure 4.12 PCSK6-AS1 expression across a range of cell lines. Lane (2)K562 (3)HeLa (4)1321N1 (5)hNSC (6)SH-SY5Y. Lane 1 50 bp NEB ladder where black triangle marks 200 bp. Expected product is 112 bp. Second upper band (white triangle) sequenced and annotated as PCSK6-AS2, 232 bp. 10 µl of PCR product was ran on a 2% TAE gel.

Intriguingly, a second band was detected just below the 250 bp ladder band (Figure 4.12). This band was excised from Lane 2 (K562) and sequenced; alignment appears to represent a second PCSK6-AS isoform (232 bp, PCSK6-AS2, accession #LN713952, white arrow Figure 4.12). This second isoform is also a 3-exon lncRNA but both exon 1 and 3 are extended in length in comparison to the ‘canonical’ PCSK6-AS1. A third isoform (PCSK6-AS3, accession #LN713953) was subsequently isolated and sequenced in the same manner; this isoform is similar to PCSK6-AS2, but has a middle exon spliced out. A graphic indicating the varying exon sizes of the three PCSK6-AS isoforms are shown in Figure 4.13. The function of each lncRNA isoform remains unknown however several databases allow annotation of the transcripts using existing datasets, including AnnoLnc (http://annolnc.cbi.pku.edu.cn/) which provides some curious differences between the AS1 and AS2 isoforms (PCSK6-AS3 wasn’t annotated since the sequence of the first exon is largely unknown). Antisense transcripts are similar to other long ncRNAs in that they can contain specific domains that interact with DNA, RNA or proteins. Appendix C.2 displays the TFs that are predicted to bind at the TSS of the lncRNA. This predictive service differs to the previous TRANSFAC analysis (Table 3.1) in that it predicts TF overlap at the lncRNA TSS rather than at a sequence centred on the rs11855415. Unsurprisingly, none of the TFs from either list overlap. Substantial differences exist between the two isoforms in the miRNA bind sites predicted on each. Additionally, curated CLIP-seq datasets which are used to annotate RNA-protein interactions indicate none for PCSK6-AS1 while PCSK6-AS2 has been shown to interact with FUS and ELAVL. According to RefSeq the FUS protein has been implicated in cellular processes that include regulation of gene expression, maintenance of genomic integrity and mRNA/microRNA processing while the ELAVL protein is known known to stabilise mRNAs. Interestingly the TSSs of both lncRNA are found to overlap with the predicted bind site of the FOXP2 transcription factor in neuronal cell lines (Appendix C.2). The PCSK6 gene’s main promoter has been found to be a direct target of FOXP2, a gene in which a mutation is thought to lead to severe forms of language or speech impairment (Lai *et al.*, 2001).

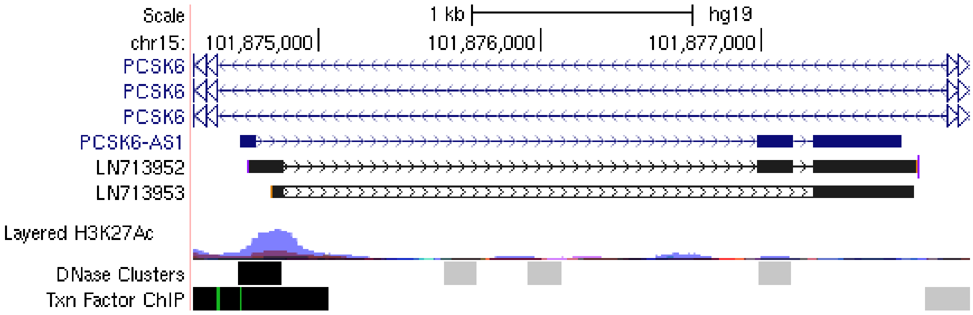


Figure 4.13 UCSC Genome browser image of PCSK6-AS1 and two newly discovered isoforms PCSK6-AS2 (accession #LN713952) and PCSK6-AS3 (accession #LN713953). Intron 13 of the canonical PCSK6 gene is shown in the top track in blue with PCSK6-AS1, PCSK6-AS2 and PCSK6-AS3 below. The ENCODE project H3K27Ac markings are indicative of promoter activity, with DNase clustering and TFBSs providing further evidence for a regulatory element at this location.

Having detected PCSK6-AS1 in a number of cell lines and confirmed its sequence, I used publicly available ENCODE project RNA-seq data to validate these results and to posit which cell line(s) could be used in future functional analysis. Analysis of Figure 4.14 suggests K562 is an obvious candidate since both PCSK6 and the lncRNA PCSK6-AS are expressed in all cellular subcompartments analysed: whole cell, nucleus and cytosol (similar to the novel exon previously analysed in RNA-seq data). Apart from in the K562 cell line, PCSK6-AS is exclusively localised to the nucleus, as might be anticipated of a lncRNA. As expected PCSK6 was most prevalent in the HepG2 cell line (Figure 4.3, Table 4.4) however, unlike K562, it is notoriously difficult to transfect (p128, Cemazar *et al*. (2010)) and does not express cytosolic PCSK6-AS. Since the function and effect of PCSK6-AS remains, as yet, unknown, preference exists towards use of K562 as a cell model in which to perform functional analysis.

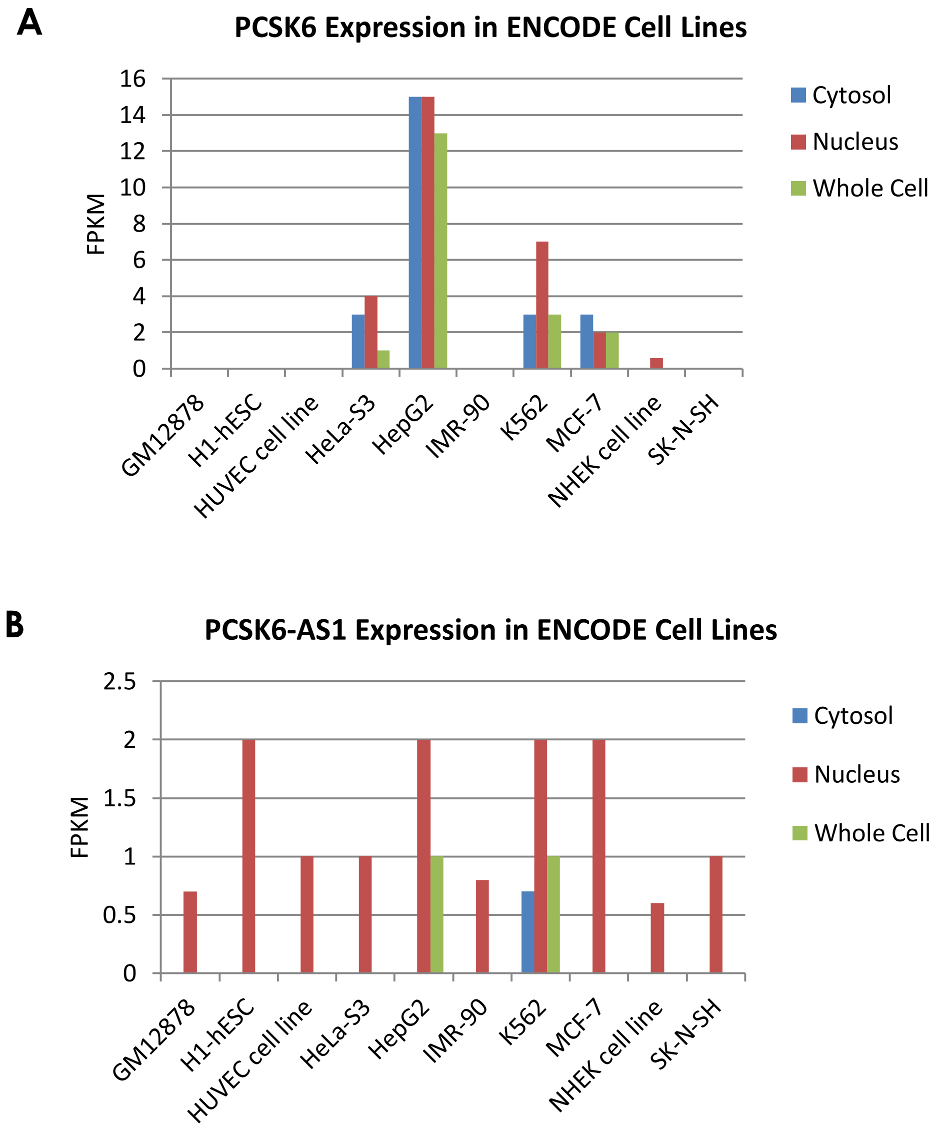


Figure 4.14 Expression of the PCSK6 gene **(A)** and antisense lncRNA gene PCSK6-AS1 **(B)** according to the ENCODE RNA-seq dataset (Djebali *et al.*, 2012) as expressed by Fragments Per Kilobase of transcript per Million mapped reads (FPKM) in cytosol (blue), nucleus (red) and whole cell (green). The minimum expression level cut-off is 0.5 FPKM. This data supports K562 as a model cell line for PCSK6-AS1 functional testing.

4.5 Discussion

Annotating all possible gene isoforms and their *in vivo* expression patterns in specific cell populations is a necessary first step in understanding the isoform-specific functions of a gene and the differentiation between trait-relevant gene isoforms and 'bystander' alternative forms of a gene. PCR analysis of the different PCSK6 gene isoforms observed within and across cell lines is an easily accessible though limited approach to cataloguing such transcription heterogeneity. Initial PCR results (Figure 4.2) would suggest PCSK6 does not display a consistent predominance of one isoform across cell lines and differences in gene expression levels were as equally varied as differences in splicing. Clearly then, PCSK6 undergoes alternative splicing and employs alternative transcription start and termination sites across multiple cell types and ontogenetic stages. Djebali *et al.* (2012) found that in general there was a plateau of 10 - 12 expressed isoforms per gene per cell line analysed. Inspection of the 8 PCSK6 isoforms currently recognised by RefSeq (Table 4.1) suggests there might still be several isoforms left to be annotated, one of which may be the PCSK6 short isoform (SI) discussed throughout this chapter. Several cell lines were shown to express all RefSeq PCSK6 isoforms (K562, HEK293 and RPE-1) and as such make suitable candidate model cell lines for future functional analysis of genetic variation within the PCSK6 locus.

Results from a study by Pal *et al.* (2011) propose gene isoforms are predominantly generated via alternative transcriptional rather than splicing mechanisms, highlighting alternative promoters as primary sources of transcriptome diversity. Furthermore the authors suggest the majority of genes associated with neurological diseases expressed multiple transcripts through alternative promoters. One such type of promoter is the bidirectional gene pair promoter (BGP) driving transcription of a short PCSK6 isoform and a lncRNA. BGPs are a common feature of the human genome, and have also been described in drosophila (Graveley *et al.*, 2011), mouse (Kanhere *et al.*, 2010) and plants (Li *et al.*, 2006) and a key point of discussion arising from this chapter centres on what the influence of these BGP divergent transcripts may have on PCSK6 gene expression and regulation.

Publicly available empirical data specific to the PCSK6-AS are quite limited, however a search of the literature can shed light on the possible function of this lncRNA. Hu *et al.* (2014) found that most of the lncRNAs expressed in the prefrontal cortex (39.8%) localize in close proximity (<4 kb) to known protein-coding genes, while lncRNAs such as PCSK6-AS which are located upstream of the protein-coding genes on the antisense strand, are particularly significant; these natural antisense transcripts (NATs) were shown to originate from a specific class of bidirectional promoters showing unique epigenetic features (see histone modifications, Figure 3.2), were highly enriched upstream of genes that are expressed in neurons and involved in neuronal functions (see PCSK6 expression in hNSC, Figure 4.2) and show a significantly positive correlation with the expression of the upstream protein-coding genes. Results presented in this chapter show the previous 4-exon transcript not to be a truncated divergent transcript, the result of stochasticity in the splicing process, but rather a fully extended, though predicted to be inactive, PCSK6 isoform complete with poly-A tail and signal peptide (Table 4.2).

A prediction of function on a protein level suggests the SI undergoes nuclear export though not secretion from the cell. The exact mechanism by which an inactive enzyme such as the PCSK6 SI might have an effect remains unknown though a large-scale analysis by Pils and Schultz (2004) into the function and evolution of inactive enzyme-homologues revealed startling insights, suggesting that inactive enzymes are conserved among metazoan species and even though they have lost their catalytic activity, they have evolved new functions; predominantly regulatory processes where the inactive enzyme-homologue regulates its active counterpart. It is worth noting that isoforms differing in their functional domains and driven by different promoters (such as the canonical PCSK6 and 'short' isoforms) are known to possess significantly different expression profiles. For example, Dijkmans *et al.* (2010) showed DCLK1, a member of the doublecortin gene family involved in neurogenesis and neuronal migration, is expressed from two distinct promoters and generates four transcript variants/protein isoforms. The first two isoforms are driven by the upstream promoter and are highly expressed in early P0–P5 stages while the downstream promoter derived isoforms 3 and 4 that are P15 and adult specific. We know from results presented previously in this chapter that PCSK6 is differentially expressed in various regions of the brain throughout development, namely hippocampus from 8-15 PCW and thereafter predominantly in the CBC and MFC up until 37 PCW. However we cannot state definitely which isoform(s) are driving these signals; microarray probes allow us to infer the isoform but there is still a shortage of publicly available data that allows one to explicitly track PCSK6 SI expression over developmental time.

One method of definitively profiling the SI expression is by interrogating RNA-seq datasets for the presence of the shorter isoform's novel exon, which previous PCR results suggest is not expressed in any other known RefSeq PCSK6 isoform. Any biological interpretation of the RNA-seq results in Table 4.4 should include the caveat that due to the initial RNA fragmentation step when generating data, longer transcripts will contribute to more fragments and are thus more likely to be sequenced (see 3'UTR, Table 4.4). These read counts would usually be normalised for transcript length and sequence depth when quantifying transcripts by employing an expression level metric such as reads per kilobase and million mappable reads (RPKM). However for the purposes of this particular study, it was detection of the novel exon itself that was of primary importance, rather than relative quantification between the exons. Indeed, Ramskold *et al*. (2009) found that it is more accurate to simply exclude 3' UTRs completely from gene models when calculating RPKM expression levels. Additionally, it is unlikely an absolute quantification of the novel exon is possible since the accuracy of transcript expression level estimates degrades progressively from high to low expressed transcripts (Kanitz *et al*., 2015). The novel exon also has extensive sequence overlap with the VNTR; reads that map to multiple genomic locations are known to present a problem when quantifying (Ramsköld *et al*., 2012). Results in Table 4.4 show quantification of the novel exon was consistently highest in the nucleus and in poly(A)$^-$ transcripts (K562, HeLa and Hep G2) though interestingly this trend is reversed for all other probes in the nucleus e.g. the nucleus poly(A)$^+$ signal without exception exceeds the nucleus poly(A)$^-$ signal for all other exon probes (1,2,4,5). Together these data suggest the function of the SI, in K562, HeLa and Hep G2 cells at least, is localised to the nucleus. In hESC cells quantification was highest in transcripts with poly(A) tails which is consistent with all other probes i.e. nucleus poly(A)$^+$ signal consistently exceeds the poly(A)$^-$ signal, supporting the notion that in hESC cells the novel exon is present in transcripts that are exported to the cytosol for further processing. Finally, RNA-seq data also showed PCSK6-AS1 to display the highest relative expression in testes and brain (Figure 4.11), a potentially interesting finding for PCSK6 in light of the results published by Hu *et al*. (2014) who found divergent transcription of lncRNAs from bidirectional promoters to be highly enriched in neuronal genes.

One of the inherent limitations of microarray data is the inability to define spliced isoforms if an exon is not probed for though the presence of an unprobed exon can still be inferred by analysis of exons upstream and downstream to the missing exon, as performed in Figure 4.8. In this microarray analysis of a 15 PCWs developing fetal brain I interrogated the Brainspan dataset for a visual quantification of probes upstream and downstream of the novel exon. There was a marked higher expression of the PCSK6 3' UTR probe relative to all other probes upstream of the novel exon in the CC and CBC. The notion that the shorter PCSK6 isoform may be contributing to the relatively higher expression of the PCSK6 signal in the CC is supported by data from the Allen Brain atlas which also showed probes upstream of the novel exon to display relatively higher expression in the adult brain (Figure 4.9). Brain-specific isoforms are known to exert substantial phenotypic effects. For example, the aberrant expression of developmentally regulated mRNA isoforms has been observed in the cerebella of patients with psychiatric disorders like schizophrenia, autism, anxiety, attention-deficit hyperactivity disorders (Ten Donkelaar and Lammens, 2009, Grimmer and Weiss, 2006). As such, given the high expression of PCSK6 within the corpus callosum, aberrant expression of PCSK6 isoforms could have an effect on CC development which in turn could influence a variety of traits and disorders including handedness.

In summary, this chapter has shown that the bidirectional promoter drives the transcription of a bidirectional gene pair: a novel, though predicted to be inactive, PCSK6 isoform in the sense strand direction and several previously unknown lncRNA antisense transcripts. RNA-seq data suggests this PCSK6 shorter isoform to show relatively high expression in both the developing and adult corpus callosum.

5 Functional Analysis of the PCSK6 Genetic Variants

5.1 Abstract

The results of several functional assays which investigate the effect of lncRNA regulation on PCSK6 isoform expression in addition to a systematic analysis of the effects of genetic variation on bidirectional promoter activity at the PCSK6 locus associated with handedness are reported and discussed in this chapter.

Electrophoretic mobility shift assay (EMSA) results provide conclusive *in vitro* evidence that the rs11855415 SNP shows significant differences in protein:DNA complex binding on allelic variation in all nuclear extracts analysed (SH-SY5Y, hNSC, hNSC-derived Neuronal and K562). A reverse-chromatin immunoprecipitation (Rev-ChIP) assay using hNSC nuclear extract posited a number of proteins to bind to the rs11855415 minor A allele but not to the major T allele, one of which was a member of the SRY (sex determining region Y)-box (SOX) TF family, as previously predicted by *in silico* analysis (Chapter 3).

A luciferase assay was used to test for bidirectional promoter activity, with results indicating a strong transcriptional bias in the sense strand direction. Luciferase results also indicate allelic variation at the rs11855415 SNP to have a weak effect on promoter activity in the HeLa and neural 1321N1 cell lines ($P < 0.02$) but not the K562 cell line. No significant difference in promoter activity was observed between the VNTR 6,9 and 10 alleles in either K562 or 1321N1 cell lines however a slight increase in luciferase was reported for the HeLa 10 allele ($P = 0.02$, ANOVA single factor).

In conclusion, the rs11855415 SNP is the most likely candidate to affect expression of the bidirectional gene pair through the creation/disruption of TFBSs. How this specifically relates to PCSK6 expression remains unknown; results from pilot PCSK6-AS knockdown and overexpression assays on an SH-SY5Y cell line were inconclusive and it remains to be clarified if and how the PCSK6-AS lncRNA regulates PCSK6 isoform expression.

5.2 Introduction

The preceding chapters provide evidence for the existence of several previously unknown PCSK6-AS lncRNA gene isoforms in addition to a novel PCSK6 shorter isoform (SI), both of which are positively correlated in expression and driven from an RNA Polymerase II-binding bidirectional promoter. Most human promoters are thought to bind polymerase complexes in a bidirectional manner and are therefore capable of initiating transcription in both directions (Core *et al.*, 2008). Thus, we cannot exclude that the expression of PCSK6-AS lncRNA may represent a passive by-product of gene transcription. As such, a series of experimental assays are required to elucidate both functionality and the effect genetic variation might have on the transcription of PCSK6-AS from the bidirectional promoter. The lack of a poly(A) tail usually suggests an unstable transcript that is retained in the nucleus, a subcellular compartment lncRNAs as a class predominantly localise to. However the PCR detection of PCSK6-AS using oligo-dT synthesised cDNA suggests PCSK6-AS1, under certain conditions at least undergoes exportation to the cytosol. To this effect, Mercer and Mattick (2013) demonstrated a substantial proportion of lncRNAs reside within, or are dynamically shuttled, to the cytoplasm where they regulate protein localisation, mRNA translation and stability. For the PCSK6 SI, RNA-seq data for the hNSC cell line (Table 4.4) and the prediction of a signal peptide (Table 4.2) support the export of this transcript to the cytosol. As such, both sense and antisense strand transcripts are in theory amenable to short interfering RNA (siRNA) knockdown assays in which cells can be transfected with a molecule targeting the transcript and the resulting change in mRNA transcript abundance measured by quantitative PCR (qPCR). The novel exon, as yet, is the only known exonic difference between the SI and the 3’ sequences of PCSK6 isoforms currently recognised by RefSeq. Such a distinction severely restricts the sequence to which a siRNA can target (<100 bp), thereby rendering any siRNA knockdown approach unfeasible. In addition, much of the third exon of the lncRNA appears to be an Alu short interspersed element (SINE) (Figure 5.1). Alu elements are ~300 bp long and are the most abundant transposable element in the human genome; about 10.7% of which consists of Alu sequences (Deininger, 2011).

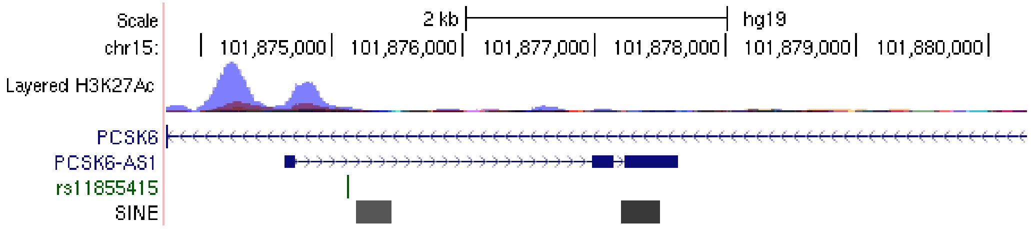


Figure 5.1 UCSC Genome browser view of the SINE Alu element indicates overlap with the third exon of the PCSK6-AS. The H3K27Ac markings indicate the position of the bidirectional promoter relative to the PCSK6-AS1 lncRNA (blue) and the rs11855415 SNP (green). Alu SINE transposable elements are marked by black boxes.

Alu elements have been shown to bind to the 3'UTR of actively transcribed target genes. For example, Batista and Chang (2013) showed that under stress conditions the lncRNA antisense to Ubiquitin Carboxyl-Terminal Esterase L1 (Uchl1) moves from the nucleus to the cytoplasm and binds to the 5’ end of the Uchl1 mRNA to promote its translation. Additionally, Alu elements in cytoplasmic lncRNA can form imperfect complementary RNA duplexes with Alu elements in the 3′ UTRs of target mRNAs (Gong and Maquat, 2011). In any case, the Alu overlap with the third exon of PCSK6-AS effectively restricts any siRNA design to the remaining ~250 bp sequence of exon 1 and 2 due to an inevitable non-specific knockdown for a probe designed against such a repetitive element.

In theory, the function of the PCSK6-AS lncRNA can also be assessed via overexpression of the lncRNA in a relevant cell line – an overexpression in the nucleus should highlight what effect the antisense transcript has on sense strand expression and whether the mechanism of function is via the act of transcribing the lncRNA transcript itself rather than, for example, chromatin remodelling (Gupta *et al.*, 2010).

Previous results have highlighted rs11855415 as a primary candidate for functional analysis for a number of reasons: (1) its consistent presence in handedness GWAS results and association with PegQ (P = 2 x 10^{-8}, Scerri *et al.* 2011a) (2) presence of the SNP within an evolutionary conserved sequence (Figure 3.5) (3) the substantial number of TFBSs predicted via *in silico* analysis to bind at that location (Table 3.1) (4) the proximity of the SNP to a bidirectional promoter and its intronic location within a lncRNA (Figure 5.1). Its core location within the bidirectional promoter and significant association with degree of handedness (P = 0.001, Arning *et al*., 2013) means the VNTR is also a candidate for functional analysis, though to a lesser extent since only two TFs (Ets-1, Tel-2) are predicted to bind at the tandem boundaries and neither in multiplex. An assessment of genetic variation and allele-specific protein binding is vital given that the majority of regulatory functions (such as transactivation and chromatin looping) are mediated through TFs and other proteins (Edwards *et al*., 2013).

A common approach to evaluating the potential of a genomic region to drive transcription involves the use of reporter assays such as luciferase products whose luminescence can be used to infer promoter activity of a cloned region. This is particularly useful when limited information regarding the regulatory potential is available, as in this case where the products of a putative promoter have limited support in the literature and whose directional bias is unknown. The findings of Chapter 3 provided a defined region of interest approximately 1.8 kb in size, a sequence short enough to clone via traditional cloning techniques. As previously discussed, the choice of cell type is also considered since *cis*-regulatory elements are highly tissue- and cell-type specific.

Several other approaches can be used to elucidate the effects of allelic variation *in vitro*. One important technique for determining protein:DNA interactions is the electrophoretic mobility shift assay (EMSA), an approach here which uses a 21 bp oligomer centred on the allele of interest that acts as a bait to which proteins will bind (Revzin, 1989). Another advantage of the EMSA is the source of the DNA-binding protein may be inexpensive crude nuclear or whole cell extract. Chromatin immunoprecipitation sequencing (ChIP-seq) technologies provide a complementary approach to predicting TF-binding sites however they are limited in that each experiment profiles just one TF. As a way around this constraint and based on an assay originally provided by Dr Ashwin Unnikrishnan (Tursky et al., 2015), I have performed a series of Rev-ChIP experiments to utilise quantitative mass spectroscopy in the screening of SNPs for differential TF binding. An advantage of this powerful technique is that multiple SNPs can be assayed simultaneously if necessary and the binding TF(s) identified in a single experiment.

To conclude, previous findings suggest the PCSK6 SI may be expressed in both the developing and adult human brain. If this is the case, the bidirectional promoter and its inherent genetic variants would be in a primary position to influence such an expression profile, possibly via epigenetic regulation since lncRNA are known to play a pivotal role in embryonic and adult neurogenesis (Yao and Jin, 2014). The purpose of this chapter therefore is not necessarily to take forward the previously identified strongest genetic variants to test their causative association with the handedness trait *per se,* but to understand their functional impact on a molecular level, thereby offering insights into the complex regulatory network between gene, isoform splicing and their regulating lncRNAs.

5.3 Methods

5.3.1 PCSK6-AS Knockdown

The 20 nmole Stealth siRNA (Invitrogen) was suspended in 1 ml RNase-free H_2O to make a 20 µM solution (20 pmol/µl). Transfection with the BLOCK-iT fluorescent oligo (Invitrogen) was performed to assess the transfection efficiency of siRNAs in SH-SY5Y cell line. SH-SY5Y cells were seeded at a $1x10^5$ density 24 hours before transfection on a 12-well plate (CLS3513, Corning). The Stealth RNAi and the BLOCK-iT fluorescent oligo were transfected into SH-SY5Y cells using RNAiMAX (Invitrogen) as follows: 60 pmol Stealth RNAi (5'-GGGUUUCAGAAUGUUUGCCAGGAUG) and 60 pmol of BLOCK-iT fluorescent oligo were each diluted in 100 µl Opti-MEM I Reduced Serum Medium (Invitrogen) and mixed gently. RNAiMAX (2 µl) was diluted in 100 µl Opti-MEM I Reduced Serum Medium and incubated for 15 minutes at room temperature (RT). After a 15 minute incubation, the diluted Stealth RNAi and BLOCK-iT fluorescent oligo were combined with the diluted RNAiMAX and incubated for 15 minutes at RT to allow the oligomer-RNAiMAX complex formation. The culture medium was changed with 1 ml fresh DMEM/F12 without antibiotics before transfection, and the oligomer-RNAiMAX complexes were added to each well and mixed gently. After transfection, the cells were incubated at 37 °C in a CO_2 incubator for 24 hours after which fluorescent uptake was observed under an immunofluorescence microscope. The cells were harvested and the RNA extracted (RNeasy Mini Kit, Qiagen). Five µg of DNase-treated RNA (Ambion DNA-free Kit, Invitrogen) was used for cDNA synthesis using oligo-dT primers as part of the Superscript III assay (Invitrogen). Quantitative real-time PCR (qPCR) was performed on the Applied Biosystems ViiA™ 7 Real-Time PCR System using SYBR Select Master Mix (Invitrogen). Results represent four independent plates (MicroAmp Fast 96-Well Reaction Plate, Invitrogen) with each sample in technical triplicate. Each 15 µl qPCR reaction consisted of SYBR Select 2 X (Invitrogen), 0.3 µM of each primer, 50 ng cDNA and up to 15 µl H_2O. Primer pair for PCSK6-AS quantification were 5’-GGTGCAGAAAACAAGCCTG and 5’- CTTCCCTGCTGGCGTTTTTG. For the PCSK6 SI (see 5.3.2) the primer pair 5’-GCAGCGGTGAGAACAACTT and 5’-CTGATGGGCACTGAAGGTGT were used. Reaction conditions were as follows: 2 min at 50 °C, 2 min at 95 °C then 40 cycles of 95 °C for 1 s and 60 °C for 30 s. The data was normalised using a housekeeping gene (HKG) primer pair for Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and DNA-directed RNA polymerases I, II, and III subunit RPABC2 (POLR2F). Primers were designed to avoid binding to SNP locations and to bridge exon-exon junctions in order to avoid amplification of human genomic DNA (gDNA). qPCR reactions were run on 1.8% agarose gels to verify the correct product size and the melt curves also checked for evidence of nonspecific amplification. All primers were designed using Primer3 (Rozen and Skaletsky, 2000), tested for primer specificity with Primer-BLAST (Ye *et al*., 2012) and are listed in Appendix A. To confirm the knockdown efficacy of the assay a known positive control siRNA for use in SH-SY5Y cells was purchased (SNCA gene, Invitrogen). To minimize non-specific amplification, primer pair concentration was optimised and efficiency validated with a standard curve for GAPDH (400 nmole forward primer, 600 nmole reverse: 102% amplification efficiency), POLR2F (400:600, 103%), SHORT (600:400, 98%), BOTH (600:600, 96%), LONG (300:600, 103%), SNCA positive control (600:200, 90%) and PCSK6-AS (300:300,100%). As with all qPCR analysis conducted, melt curve analysis ensured the qPCR signal was not driven by primer-dimer pairs.

5.3.2 PCSK6-AS Overexpression

For overexpression of PCSK6-AS both gDNA and cDNA sequences of the lncRNA were cloned in to a pcDNA3.1 vector (V790-20, Addgene Vector Database) and transfected in to a neuronal cell line (SH-SY5Y). In summary, a primer pair was designed to isolate the gDNA sequence 3,425 bp in size. The DNA had been genotyped previously for SNP rs11855415 (T/T alleles) and VNTR (6/6 alleles). The primer pair (5'-CCCGGGGATCCGGTGACAGCGACACAGGAA and 5'-CCCGGGCTCGAGAGGAAAGAGCCCAGGAGGAA) includes BamHI and XhoI restriction enzyme (RE) sites respectively for downstream cloning. PCR was performed using Phusion High-Fidelity DNA Polymerase (NEB, 7.5 µl HF buffer, 300 nmole of each primer, 1 unit of polymerase, 50 ng of gDNA and H_2O for a total volume of 15 µl) under the following thermo-cycler conditions: 30 sec at 98 °C followed by 35 cycles of 10 s at 98 °C, 20 s at 54.5 °C, and 2 min at 72 °C before finishing on 10 min at 72 °C. The PCSK6-AS1 cDNA was purchased and arrived as part of a lyophilised 2,450 bp pEX-A2 plasmid (Eurofins). 11.4 µl TE was added to bring it to a 500 ng/µl concentration and performed a double-digest on the backbone RE sites for Bam and XhoI to remove the PCSK6-AS1 cDNA (3 µl of NEB buffer 3.1, 2 µg DNA, 21 µl H_20, 1 µl of each RE, 37 °C 30 min) and thereafter performed a gel extraction for clean-up of the PCSK6-AS1 cDNA using the QIAquick Gel extraction kit according to the manufacturer's instructions. Both the PCSK6-AS1 gDNA PCR product and the elute from the gel extraction kit (PCSK6-AS1 cDNA) were separately cloned into the pcDNA3.1 vector via the BamHI and XhoI RE sites and ligated using T4 ligase according to the manufacturer's instructions (NEB). pcDNA3.1 plasmids containing the 3,425 bp PCSK6-AS1 gDNA and 655 bp cDNA were then transformed into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen) before plating on ampicillin-resistant plates and overnight incubation. Colonies were tested for PCSK6-AS1 expression using the primer pair 5'-GGTGCAGAAAACAAGCCTG and 5'-TTGGTCCCACTGCTTCTTCC with the same PCR conditions as previously. Colonies were suspended in 10 µl H_20 and then immediately transferred to 10 ml luria broth (12795-027, Invitrogen) with 0.01% Ampicillin, cultured at 200 rpm 37 °C overnight before extraction of the plasmids 12 hours later using the Qiagen Plasmid Midi extraction kit according to the manufacturer's instructions. Plasmids were quantified using a Nanodrop 2000 and brought to a concentration of 500 ng/µl in TE for storage. SH-SY5Y cells were seeded 4 x 10^5 per well on a 12-well plate (CLS3513, Corning) 24 hours before transfection. 1.5 µg of DNA was added to 2.5 µl of P3000 reagent and Opti-MEM (Invitrogen) to give a total 5 µl volume which was allowed to incubate for up to 10 minutes. To this, 0.4 µl Lipofectamine 3000 and 4.6 µl Opti-MEM were added to give a total volume of 10 µl which was then added to each well after 5 minutes incubation at RT. The cells were harvested after 24 hours and the RNA extracted (RNeasy Mini Kit, Qiagen). Five µg of DNase-treated RNA (Ambion DNA-free Kit, Invitrogen) was used for cDNA synthesis using oligo-dT primers as part of the Superscript III assay (Invitrogen). For qPCR quantification GAPDH was used as the HKG using conditions as described previously (4.3.5). All primer pair sequences have been included in Appendix A.

5.3.3 Electrophoretic Mobility Shift Assay (EMSA)

SH-SY5Y, K562 and HEK-293 cells were maintained to ECACC guidelines in DMEM/F12 supplemented with 10% FBS and 5% penicillin/streptomycin at 37 °C with 5% CO_2 in TC-treated T-75 flasks (Nunc 156499). Human Neural Stem Cells (hNSC) were H9 hESC-derived, cultured and differentiated into neurons according to manufacturer's protocol (N7800-100, Life Technologies). Isolation of nuclear protein extract was performed using the Nuclear Extraction Kit (SK-000, Signosis) according to the manufacturer's protocol. Briefly, up to 10^7 cells were washed in 1 x PBS before the addition of Buffer I (7.5 ml 1 X Buffer 1 solution, 75 µl DTT solution and 75 µl Protease inhibitor), placed on ice and rocked at 200 rpm for 10 minutes. The cells were released using a sterile scraper, transferred to a 15 ml falcon tube and centrifuged at 10 x g for 5 minutes at 4 °C. The supernatant was discarded before the addition of Buffer II (250 µl 1 X Buffer II solution, 2.5 µl DTT and 2.5 µl Protease inhibitor). The pellet was resuspended and transferred to a 1.5 ml microcentrifuge tube and placed vertically in an icebox on a shaking platform for 2 hours at 200 rpm. Finally the sample was centrifuged at 10 x g for 5 minutes at 4 °C and the nuclear extract supernatant transferred to a new tube. Protein concentrations were determined using a Qubit fluorometer and Qubit Protein Assay (Invitrogen). Single-stranded probes 21 bp in length and centred on the SNP allele of interest (for sequences see Appendix D) were synthesized with 5′-biotin labels and purified by HPLC (Eurofins MWG operon, Ebersberg, Germany). Equimolar amounts of double-stranded oligonucleotide probe stock (100 mmol) were prepared by annealing 5'-biotin labelled oligonucleotides in annealing buffer (10 mM Tris pH 7.5-8, 1 mM EDTA, 50 mM NaCl). The annealing oligos were heated at 95 °C for 2 minutes before being left to cool overnight to RT and storage at 4 °C. A 15 µl binding reaction of nuclear protein extract (7.5 µg), 0.5 µg poly d(I-C) (Sigma Aldrich, UK) and 3 µl of 5 X Binding Buffer (1 mL 1 M Hepes pH 8.0, 2.5 mL 1 M KCl, 25 µl 1 M DTT, 5 µl 0.5 M EDTA, 50 µl 1 M $MgCl_2$, 2.5 mL glycerol dH_2O to 10 mL) was incubated with labelled probe (20 fmol) alone or with 10 X/100 X unlabelled competitor or scrambled probe at RT for 20 minutes. Following incubation, 2 µl of 6 X Orange-G dye was added to each 15 µl sample before electrophoresis at 200V for 5 minutes followed by 100V for 30 minutes. The samples were electrophoresed through non-denaturing 5% polyacrylamide minigels (8x8x0.1 cm 4.46 mL ddH_20, 591 µl 5 X TBE, 895 µl 40% acrylamide/bisacrylamide (29:1), 60 µl 10% APS, 4.18 µl TEMED). Note gels had previously been left to polymerise overnight at 4 °C wrapped in 5 X TBE soaked white tissue and cellophane. The gels were then electroblotted to a Nylon B positive membrane (Thermo Scientific) for 45 minutes at 100 V in 4 °C 0.5 X TBE buffer. The protein:DNA complexes were autocross-linked to the nylon membrane with a Stratalinker 1800 UV transilluminator using 312 nm bulbs. Blocking of the membrane and subsequent chemiluminescent detection of the biotin-labelled DNA was performed using an enhanced luminol substrate for horseradish peroxidase (HRP) (Thermo Scientific) and visualised using a Fuji LAS-3000 imaging system. EMSA experiments were performed at least four times for each cell line.

5.3.4 Rev-ChIP

Cell cultures

HeLa (ATCC CCL-2) and 1321N1 (86030402 Sigma) cells were maintained to ECACC guidelines in DMEM/F12 supplemented with 10% FBS and 5% penicillin/streptomycin at 37°C with 5% CO_2 in TC-treated T-75 flasks (156499, Nunc). Human Neural Stem Cells (hNSC) were cultured according to the manufacturer's protocol (Cat N7800-100, Life Technologies). Isolation of nuclear protein extract was performed using the Signosis Nuclear Extraction Kit (SK-0001, Signosis) according to the manufacturer's protocol.

Preparation of bait oligonucleotides

Double-stranded oligonucleotide probe were prepared by annealing 700 ng of 5'-biotin labelled oligonucleotides with 800 ng of non-labelled oligonucleotide in annealing buffer (10 mM Tris pH 7.5-8, 1 mM EDTA, 50 mM NaCl). The annealing oligomers were heated at 95 °C for 2 minutes before being left to cool overnight to RT. Double-strand annealing was confirmed on a 2% TAE agarose gel before storage at -20°C. See Appendix G for bait oligonucleotide sequences.

Beads preparation

Dynabeads M-280 Streptavidin (30 µl) (11205D, Invitrogen) were concentrated on a magnetic particle concentrator and the supernatant removed before washing twice with TE + 0.01% NP-40 (120 µl/wash) and twice with Buffer DW (250 µl/wash, 20 mM Tris, pH 8.0, 2 M NaCl, 0.5 mM EDTA, 0.03% NP-40). Beads were resuspended in Buffer DW (120 µl Buffer DW for 30 µl beads), DNA added (1.5 µg for 30 µl beads) and rotated on a wheel at RT. After 3 hours, beads were washed once with 120 µl TE + 0.02% NP-40 and 3 times with Buffer DW (120 µl/wash). For blocking 200 µl of Blocking Buffer (20 mM HEPES-NaOH pH7.9, 0.05 mg/ml BSA, 0.3M KCL, 5 mg/ml Polyvinylpyrrolidone, 0.05 mg/ml Glycogen, 2.5 mM fresh DTT, fresh Protease Inhibitors (P8340, Sigma Aldrich) and fresh Phosphatase Inhibitors (P0044, Sigma Aldrich)) was added to the beads and left to incubate on a rotator for 1 hour at RT. The beads were concentrated, supernatant removed and washed in 200 µl of NEB Buffer 3 +

0.02% NP-40 followed by 2 washes (400 µl/wash) of Buffer G (20 mM Tris HCL pH7.4, 10% Glycerol, 0.1 M KCL, 0.2 mM EDTA, 10 mM Potassium Glutamate, 0.04% NP-40, 2 mM fresh DTT, 4 µl fresh Protease Inhibitors and 4 µl fresh Phosphatase Inhibitors). The beads were stored on ice while the protein nuclear extract was prepared.

Clearing and incubating nuclear extract

A graphic summarising the Rev-ChIP protocol is provided in Figure 5.2. Briefly, nuclear extract (200 µg) was thawed on ice before spinning at 15,000 g at 4 °C for 10 minutes to remove denatured insoluble material. Supernatant was removed to a fresh tube and adjusted to a final concentration of 10 mM Potassium Glutamate. The extract was diluted with one volume of Buffer G (with 0.2 mg/ml poly dA.dT (P0883, Sigma Aldrich), fresh Protease Inhibitors and fresh Phosphatase Inhibitors). The extract was again centrifuged at 15,000 g for ten minutes at 4 °C to remove the insoluble pellet and the supernatant moved to a fresh tube and placed on ice. Next a fresh 30 µl aliquot of Dynabeads® was washed with 100 µl TE containing 0.02% NP-40, concentrated and the supernatant removed before a 100 µl wash of Buffer DW and a 100 µl wash of Buffer G (add fresh Protease Inhibitor and Phosphatase Inhibitor to both buffers). The Buffer G wash was repeated before combining the beads with the previously prepared extract on a rotator at 4 °C for 1 hour. The beads were concentrated and the supernatant (cleared extract) removed to a fresh tube and incubated overnight on a rotator with the previously blocked DNA-conjugated beads at 4 °C. Next the beads were concentrated and the supernatant saved as unbound extract. The beads were washed 5 times (500 µl/wash) in Buffer GS (20 mM Tris HCl pH 7.4, 0.1 M KCl, 0.2 mM EDTA, fresh 2 mM DTT and 4.2 µl fresh Phosphatase Inhibitor).

Digestion and LC-MS mass spectrometry

Beads were briefly washed in 50 mM ammonium bicarbonate and resuspended in 20 µl ammonium bicarbonate. For reduction, 2.5 µl of 8 M Urea (final concentration 1 M) and 1 µl of 100 mM DTT (final concentration 5 mM) were added before incubation at 56 °C for 40 minutes. The sample was alkylated by adding 2.4 µl of freshly prepared iodoacetamide (final concentration 5 mM) followed by incubation for 30 minutes in the dark at RT. 1 µl of 100 mM DTT was added and left for 5 minutes, before adding 1 µl of 20 mM $CaCl_2$. Freshly prepared trypsin at 1:20-1:50 concentration to protein (i.e. if 0.1 mg/mL protein, add 1 µl etc.) was added (V5111, Promega) and incubated at 30°C overnight in a thermoshaker at 400 rpm. To halt the reaction 3.5 µl of 10% acetic acid was added and the beads removed prior to sending for mass spectrometry analysis.

nanoLC-ESI (electrospray) tandem mass spectrometry analysis on the TripleTOF 5600+

The peptides were then separated on an Acclaim PepMap 100 C18 trap and an Acclaim PepMap RSLC C18 column (ThermoFisher Scientific), using a nanoLC Ultra 2D plus loading pump and nanoLC as-2 autosampler (Eksigent). The peptides were eluted with a gradient of increasing acetonitrile, containing 0.1 % formic acid (5-50% acetonitrile in 90 minutes, 50-95% in a further 1 minute, followed by 95% acetonitrile to clean the column, before re-equilibration to 5% acetonitrile). The eluent was sprayed into a TripleTOF 5600+ electrospray tandem mass spectrometer (ABSciex, Foster City, CA) and analysed in Information Dependent Acquisition (IDA) mode, performing cycles of 250 msec of MS followed by 100 msec MSMS analyses on the 15 most intense peaks seen by MS. The MS/MS data file generated via the 'Create mgf file' script in PeakView (ABSciex) was analysed using the Mascot algorithm (Matrix Science), against the NCBInr database Apr 2015 both restricted to human and with no species restriction, trypsin as the cleavage enzyme and carbamidomethyl as a fixed modification of cysteines and methionine oxidation as a variable modification. The Mascot search results were exported a Mascot.dat files and loaded into Scaffold v4.4.5 for further interrogation. All common contaminants (keratin etc.) were retained.

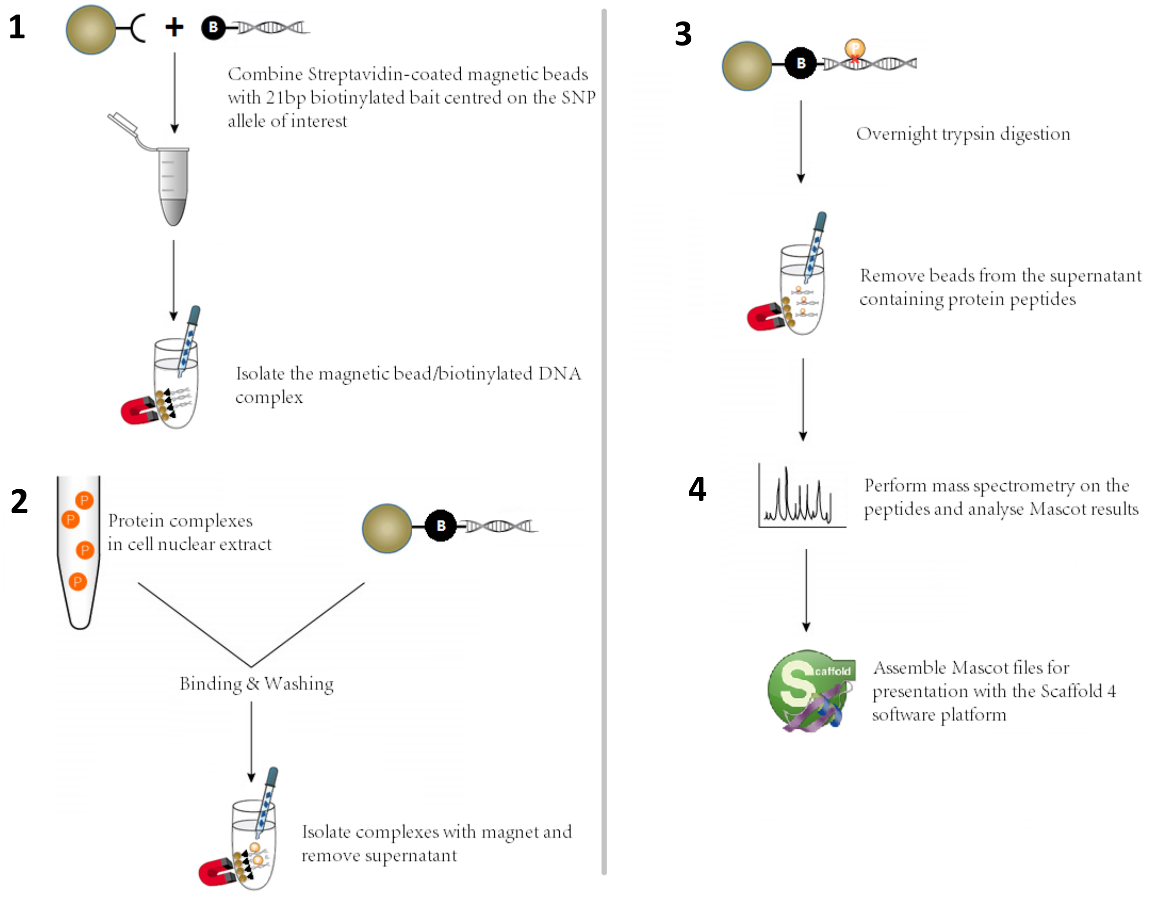

1
B
Combine Streptavidin-coated magnetic beads with 21bp biotinylated bait centred on the SNP allele of interest
Isolate the magnetic bead/biotinylated DNA complex
2
P
Protein complexes in cell nuclear extract
B
Binding & Washing
Isolate complexes with magnet and remove supernatant
3
B
P
Overnight trypsin digestion
Remove beads from the supernatant containing protein peptides
4
Perform mass spectrometry on the peptides and analyse Mascot results
Scaffold
Assemble Mascot files for presentation with the Scaffold 4 software platform

Figure 5.2 Reverse ChIP protocol summary. The protocol consists primarily of 4 major steps: (1) The annealing of biotinylated double-stranded oligonucleotide baits to streptavidin magnetic beads. Note the baits match those as used in the EMSA assay (see 5.3.3) (2) binding the nuclear extract protein from the relevant cell line with the biotinylated DNA/streptavidin complex (3) performing on-bead in-solution trypsin digest and isolation of the resulting protein peptides (4) performing mass spectrometry and the preparation of results for analysis. See Methods (5.3.4) for full protocol.

5.3.5 Luciferase Reporter Assay

Previous *in silico* analysis predicted a locus within PCSK6 to contain a secondary bidirectional promoter (chr15:101873803-101875608, hg19). A PCR using gDNA homozygous for both SNP rs11855415 and VNTR rs10523972 was performed (5’-CTGGCTCTAAATGGCAGCCT and 5’-ACCCCGAGTACTACTGCTTTT) to produce an amplicon for luciferase reporter cloning. PCR product sizes depended on the VNTR alleles: 1,707 bp for the 6 allele, 1,806 bp for the 9 allele and 1,839 for the 10 allele. PCR was performed using Phusion High-Fidelity DNA Polymerase (NEB, 7.5 µl HF buffer, 300 nmole each primer, 1 unit polymerase, 50 ng of gDNA and H_2O for a total volume of 15 µl) under the following cycle conditions: 1 min at 98 °C followed by 35 cycles of 10 s at 98 °C, 20 s at 63.2 °C, and 2 min at 72 °C before finishing on 5 min at 72 °C. Resulting blunt-end products were cloned into the pCR™-Blunt II-TOPO vector (Invitrogen) and transformed into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen) before plating on ampicillin-resistant plates and overnight incubation. Colonies were suspended in 10 µl H_20 and immediately transferred to 10 ml luria broth (12795-027, Invitrogen) with 0.01% Ampicillin, cultured at 200 rpm 37 °C overnight before extraction of the plasmids 12 hours later using the Qiagen Plasmid Maxi extraction kit according to the manufacturer's instructions.. A double digest was performed between the KpnI and XhoI restriction sites and the resulting sticky-end PCR products were then cloned in both directions (see plasmid maps Appendix E) into a pGL4.10 luciferase vector (Promega) and cells seeded in 96-well clear-bottom white plates (VWR, 734-1610) at 3×10^4 cells per well 24 hours prior to transfection. Cells were transiently co-transfected in quadruplicate in antibiotic-free medium with 80 ng pGL4.10 promoter construct and 20 ng pRL-TK renilla luciferase control plasmid using Lipofectamine 3000™ (Invitrogen) according to the P3000 protocol (Appendix E). Cells were assayed 24 hours after transfection using the Dual-Luciferase System (Promega). Relative luciferase activity (RLA) was determined using a MicroBeta2 Plate Counter (Perkin-Elmer) and normalised to the pRL-TK luciferase activity. A minimum of 3 independent transfection experiments were performed. For the assessment of a minimal promoter constructs containing accumulative deletions were designed by removing the sequence between the restriction sites for KpnI (-1,171 bp from the TSS of the novel PCSK6 isoform) and the following: Ndel (-1,067 bp), AvrII (-783 bp), SanDI (-637 bp), StuI (-507 bp), PfIMI (-378/-246 bp), AcII (-155 bp) and BsaAI (+100bp). Orientation and allele identity of recombinant clones were verified by Sanger sequencing (DNA Sequencing and Services, Dundee); mismatches relative to the reference genome were permitted if represented in dbSNP Build 143 (Smigielski *et al.*, 2000). Site-directed mutagenesis using GeneArt Site-Directed Mutagenesis System (Invitrogen) was performed to synthesise constructs not available from source gDNA. Constructs for 6/6A and 10/10T were synthesised from the previously cloned 6/6T and 10/10A plasmids respectively (see Appendix A for sequences). Mutagenesis protocol was according to manufacturer's instructions using 50 ng of source plasmid DNA. PCR was performed using Phusion High-Fidelity DNA Polymerase (NEB, 7.5 µl HF buffer, 300 nmole each primer, 1 unit polymerase, 50 ng of gDNA and H_2O for a total volume of 15 µl) under the following cycle conditions: 1 min at 98 °C followed by 35 cycles of 15 s at 98 °C, 20 s at 57 °C, and 2 min at 72 °C before finishing on 5 min at 72 °C.

5.4 Results

5.4.1 Knockdown of the PCSK6-AS lncRNA

To detect any potential regulatory effect PCSK6-AS has on PCSK6 expression I performed a siRNA-mediated knock-down of the PCSK6-AS lncRNA in a neuronal cell line (SH-SY5Y) and measured the effect via qPCR quantification (see Figure 5.3 for the location of primer pairs used). This siRNA knockdown reduced PCSK6-AS expression by approximately half (versus untransfected) and showed no discernible impact on either the 'both' or 'long' PCSK6 expression. There was an upregulation in the 'short' primer pair (Figure 5.4). A similar siRNA knockdown on the shorter isoform was not possible due to the restrictions imposed by the short length of the novel exon. siRNA target mRNA in the cytoplasm but considering (1) lncRNA as a class typically localise to the nucleus and (2) previous PCSK6-AS RNA-seq analysis showed relatively high expression in the brain (Figure 4.11), I repeated the knockdown but this time using an antisense oligonucleotide (ASO, Integrated DNA Technologies) in the same neuronal cell line (SH-SY5Y). ASOs are oligomers 15 – 25 bp in length and designed to block expression of specific targeted proteins in the nucleus by degradation of the targeted mRNA. The results were inconsistent and difficult to interpret (not shown), most likely due to the knockdown probe targeting of the third exon which mostly consists of an SINE Alu transposable element, thereby rendering the target sequence non-specific. The difficulties discussed make the siRNA approach to knocking down PCSK6-AS and PCSK6 SI untenable.

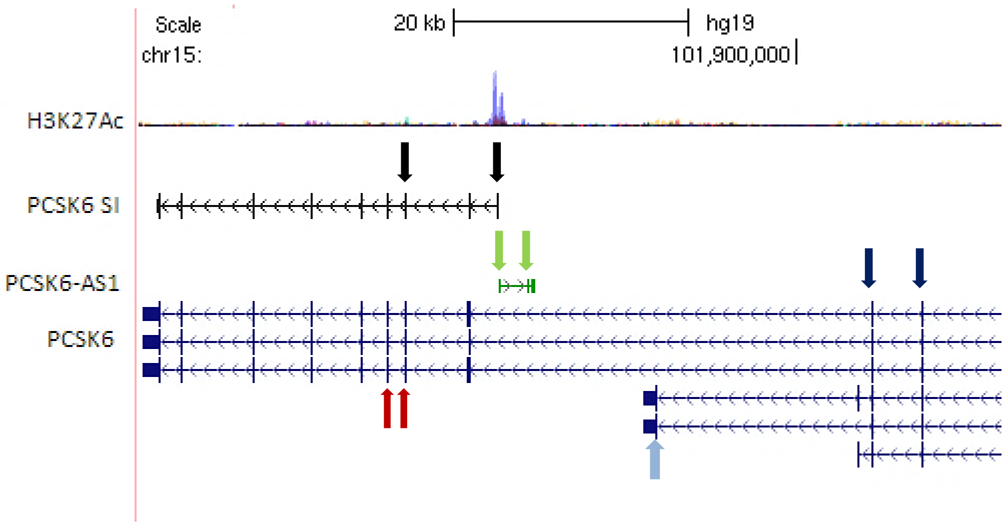


Figure 5.3 Location of the PCSK6 primers used for qPCR quantification. PCSK6 SI (black arrows), PCSK6-AS1 (green) and the isoforms already recognised by RefSeq ranked by length: 'Short' (light blue), 'Long' (red) and short and long combined ('Both', dark blue). Top track indicates acetylation marking from the UCSC genome browsers ENCODE track and indicates the approximate location of the bidirectional promoter.

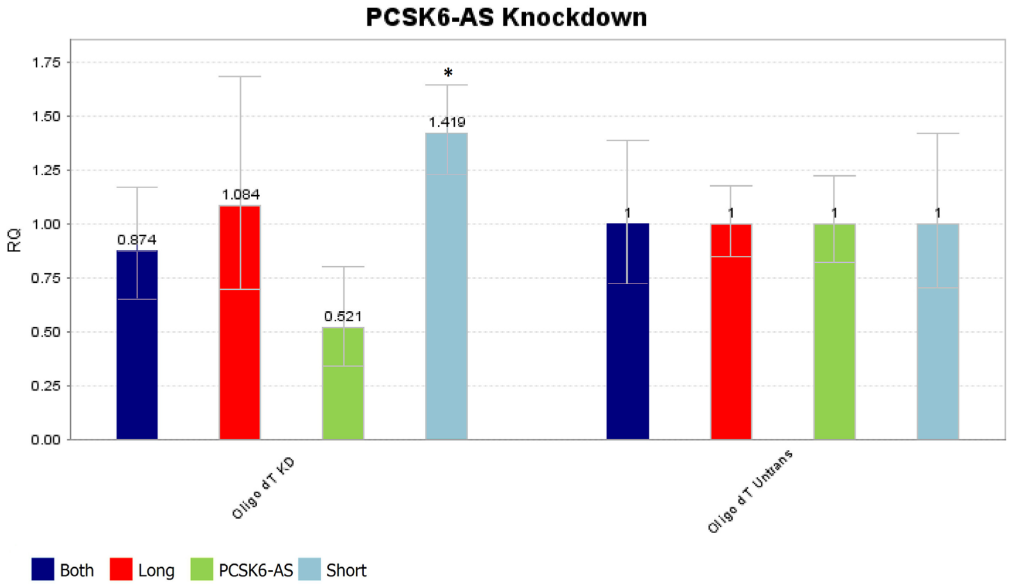


Figure 5.4 siRNA knockdown of the PCSK6-AS lncRNA. Quantification was for Short (light blue), Long (red), Both (dark blue) and PCSK6-AS (green). Bar colour reflects primer pair used to quantify expression (arrows in Figure 5.3). PCR products are from SH-SY5Y-derived cDNA synthesised using oligo-dT primers. qPCR experiments were performed in triplicate on 3 independent plates (N = 3) and quantified relative to the untransfected cells equal to '1'. Bars indicate mean (SD). * indicates significant difference to the untransfected at a value of P≤0.05 (Student's T-test, two-tailed unpaired).

5.4.2 Overexpression of the PCSK6-AS lncRNA

To investigate whether the PCSK6-AS lncRNA affected expression of the PCSK6 SI and other known PCSK6 isoforms *in vitro*, I overexpressed the lncRNA by cloning both the gDNA and cDNA sequence in to the CMV-driven pcDNA3.1 vector followed by a transient transfection of the SH-SY5Y cell line. qPCR quantification confirms that when compared with the mock and untransfected cells, an induced overexpression of the lncRNA records no significant effect in any other PCSK6 isoform's expression. Both gDNA and cDNA sequences for the PCSK6-AS1 lncRNA were cloned since the mechanism of action of the lncRNA is still unknown, though clearly the lncRNA undergoes alternative splicing (Figure 4.13). Since lncRNA stability is affected by genetic variation, all haplotype combinations of both rs11855415 and VNTR were cloned with all showing similar results. For the purpose of clarity only the gDNA plasmid containing the rs11855415 T allele and VNTR 6 allele has been included in Figure 5.5 and Figure 5.6.

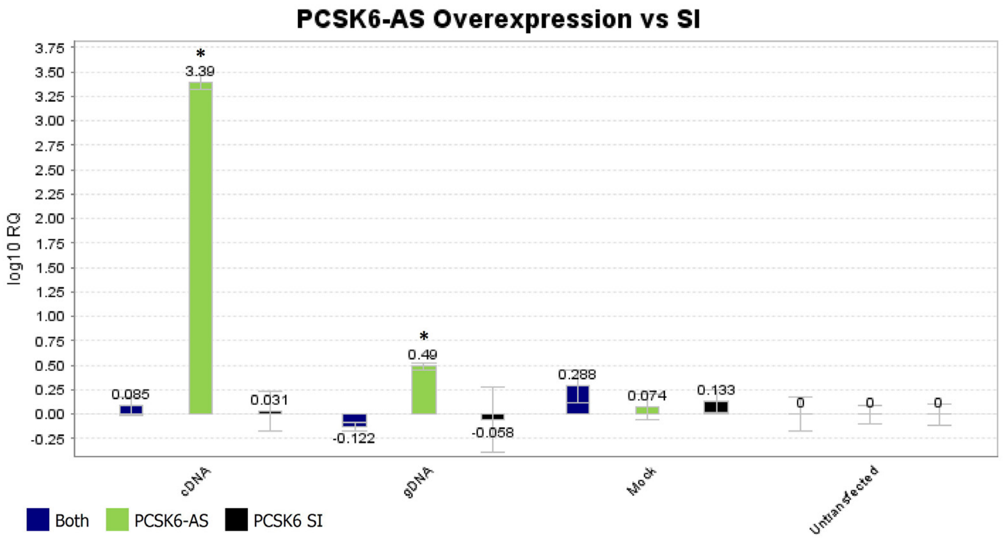


Figure 5.5 Overexpression of the PCSK6-AS lncRNA. qPCR quantification for Both (blue), PCSK6-AS (green) and PCSK6 SI (black) PCR products from SH-SY5Y-derived cDNA synthesised using oligo-dT primers. qPCR experiments were performed in triplicate on 3 independent plates (N = 3). Results were $\log_{10}$ expressed due to the vast discrepancy between PCSK6-AS and the other results and measured relative to the untransfected cells. Bars indicate mean (SD). See Figure 5.3 for primer pair locations. * indicates significant difference to the untransfected at a value of $P \leq 0.01$ (Student's T-test, two-tailed unpaired) prior to log transformation.

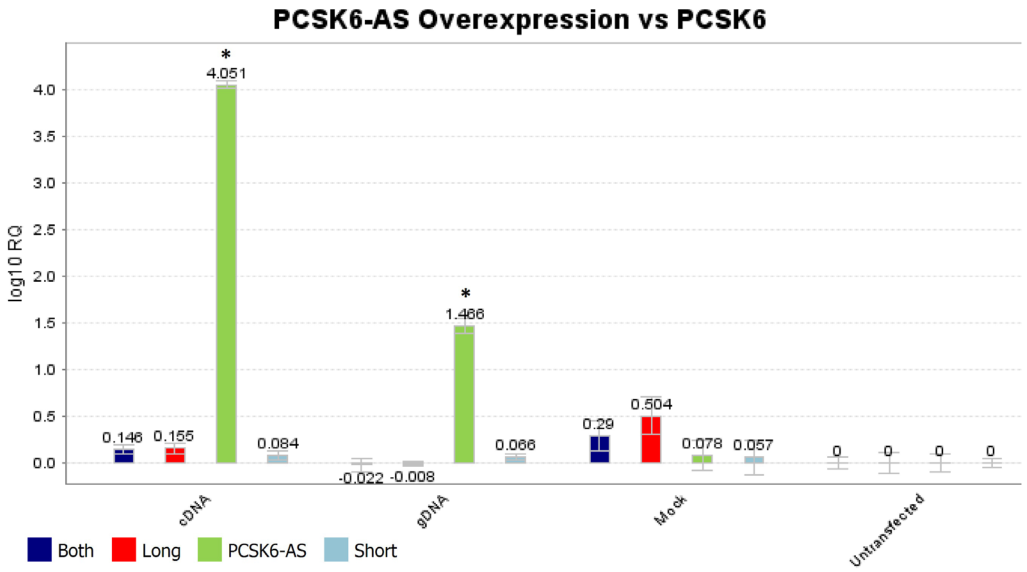


Figure 5.6 Overexpression of the PCSK6-AS lncRNA. qPCR quantification for Both (dark blue), Long (red), PCSK6-AS (green) and Short (light blue) PCR products from SH-SY5Y-derived cDNA synthesised using oligo-dT primers. qPCR experiments were performed in triplicate on 3 independent plates (N = 3). Results were $\log_{10}$ expressed due to the vast discrepancy between PCSK6-AS and the other results and measured relative to the untransfected cells. Bars indicate mean (SD). See Figure 5.3 for primer pair locations. * indicates significant difference to the untransfected at a value of P≤0.01 (Student's T-test, two-tailed unpaired) prior to log transformation.

5.4.3 EMSA

Based primarily on previous *in silico* predictions (Table 3.1) and GWAS findings (Brandler *et al*., 2013), rs11855415 and rs7182874 were selected for an analysis of each SNP's ability to modulate TF binding. EMSAs were conducted primarily with embryonic hNSC nuclear extract which provides a good model to test the effect of genes expressed in early development as in the case of PCSK6. No difference in binding affinity was observed for rs7182874 while a significant allelic difference was observed for rs11855415 in the hNSC cell line in addition to the K562 and Neuronal cell lines (Figure 5.7). Specifically, the rs11855415 minor allele A created a protein binding band which was absent for the major T allele. The specificity of this finding was confirmed by the failure of a 100-fold excess of specific-competitor probe or a 10-fold excess of scrambled probe to significantly affect the visible band (e.g. lanes 4 and 5, panel C, Figure 5.7). A similar pattern was observed for rs11855415 across a range of cell lines (see Appendix D.2 for additional cell lines analysed). These EMSA results support previous *in silico* predictions and show rs11855415 to have a substantial effect on *in vitro* protein:DNA interaction.

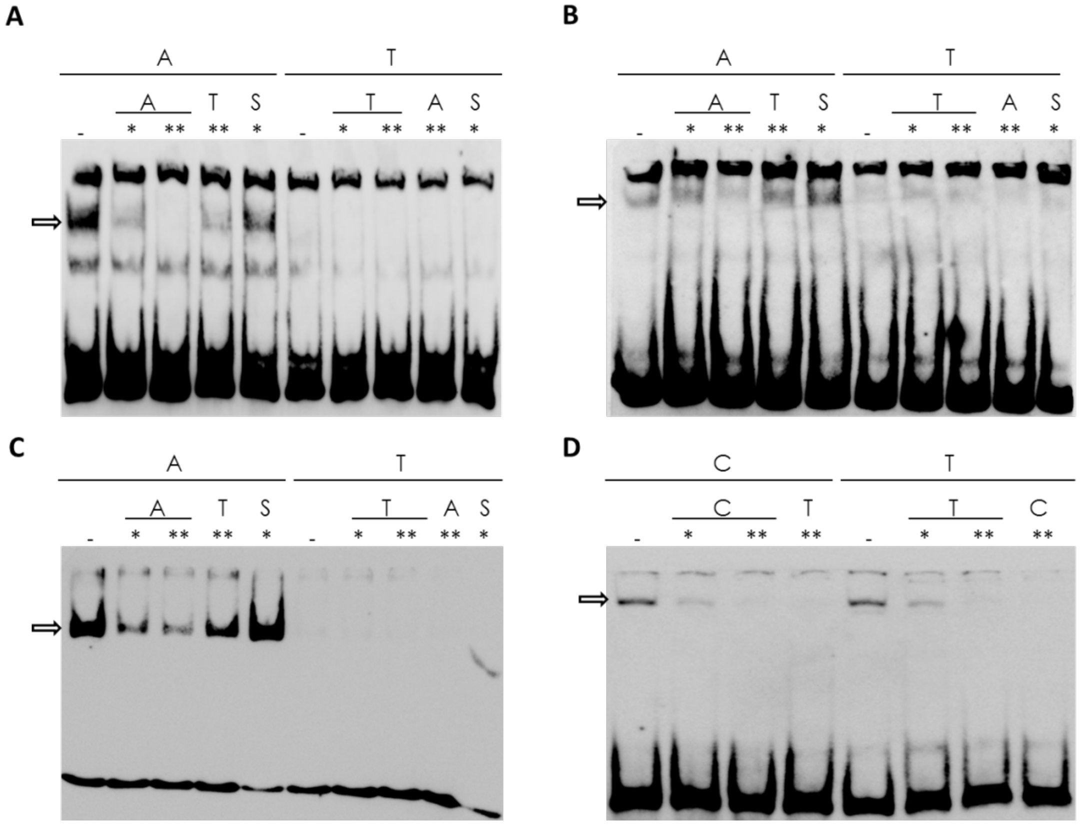


Figure 5.7 EMSA gel images for protein:DNA binding at the rs11855415 SNP. Displayed is the binding of hNSC nuclear extract to probes containing the rs11855415 SNP A versus T alleles **(A)**, the binding of Neuronal nuclear extract to probes containing the rs11855415 SNP A versus T alleles **(B)**, the binding of K562 nuclear extract to probes containing the rs11855415 SNP A versus T alleles **(C)** and the binding of SH-SY5Y nuclear extract to probes containing the rs7182874 SNP C versus T alleles **(D)**. Arrow indicates the protein:DNA complex band. The presence of a competitor is denoted above each lane: -, no competitor; S, scrambled competitor; and *, 10-fold and **, 100-fold excess of competitor respectively.

5.4.4 Rev-ChIP

Rev-ChIP is a method which allows for the detection of SNP sequences that differentially bind protein in an allele-specific manner using all the TFs of a nuclear crude extract at once. Since Rev-ChIP employs an agnostic approach in the identification of multiple proteins on a single DNA sequence using a powerful mass-spectrometry resolution, one should expect to see non-specific binding in subsequent results. To address this I implemented a stringent protocol to reduce the identification of such false positives – all proteins identified were required to have at least two peptide sequences identified by Mascot (Perkins *et al.*, 1999) and any proteins that were found to bind to the A or T allele and a non-related control oligo were removed from further consideration. Since we are primarily interested in how genetic variation affects TF binding at a bidirectional promoter I suspect to be active (though not exclusively) in fetal and adult neurogenesis, I restricted my analysis to neuronal cell line nuclear extracts (hNSC and 1321N1).

For the hNSC cell line the most intriguing result was the finding that the SRY (sex determining region Y)-box 5, a member of the SOX TF family as predicted by *in silico* analysis (Table 3.1) also appears here to bind to the minor A allele and not the T or control allele (Table 5.1). Another interesting result from the hNSC analysis and the only result found specific to the T allele was what Mascot defines as 'signal recognition particle 14 kDa (homologous Alu RNA binding protein), isoform CRA_b '. Note in Figure 5.1 the proximity of the SNP to the nearby Alu element (~60 bp). This is an interesting correlate for the possible function of the PCSK6-AS and a future line of investigation might be to perform a similar pull down assay using the whole third exon sequence of the PCSK6-AS, a region which displays substantial overlap with an Alu element. Suitably controlled, such an assay would provide a list of prime candidates which interact with the PCSK6-AS lncRNA. For the 1321N1 cell line's A allele Mascot identified 'forkhead box C2 (MFH-1, mesenchyme forkhead 1), isoform CRA_b (FOXC2)', the specific function of which has yet to be determined though it is known to be involved in the 'heart development' pathway (http://www.ncbi.nlm.nih.gov/biosystems/198802) where it interacts directly with Sonic Hedgehog (SHH), a gene instrumental in patterning the early embryo. No one single TF was observed to bind to either rs11855415 allele in both nuclear extracts indicating no overlap between the nuclear extracts (Table 5.1).

Using the Mascot results list from Table 5.1, I arranged the proteins according to molecular function using the PantherDB functional classification service (http://www.pantherdb.org/). All alleles display the binding of proteins with catalytic activity. hNSC A allele analysis also suggests the proteins binding to have enzyme regulator activity (GO:0030234) while the T allele in 1321N1 cells is thought to display receptor activity (GO:0004872), possibly indicative of the ‘Chain A, Core Of The Alu Domain Of The Mammalian Srp’ binding as indicated in Table 5.1 previously. Such findings seem contrary to the results from the EMSA assay where use of the same hNSC nuclear extract displayed little or no binding to the T allele (Lanes 6 – 10, Figure 5.7A). However this discrepancy is most likely due to the vast difference in resolution offered by both assays; the mass spectrometry-based assay’s ability to detect single protein-binding events versus the limited visual inspection of protein:DNA binding offered by EMSA. Though it is difficult to interpret these data since the identity of the proteins binding have not yet been validated, the most significant difference in terms of molecular functional classification between the A and T-binding proteins in both 1321N1 and hNSC nuclear extracts is the presence of translation/enzyme regulator activity for the A but not the T alleles (see cyan/turquoise sectors in Figure 5.8).

Table 5.1 Binding preference of proteins to both rs11855415 alleles according to Rev-CHiP mass spectrometry output. At least 2 peptides were required for each protein identified (99.5% homology with the reference sequence, 1% FDR).Mass spectrometry data were generated for amino acid (AA) identification in both hNSC and 1321N1 cell lines. – indicates no protein detected and * indicates unrelated biotinylated molecule. Dashed boxes highlight the proteins discussed in text.

		rs11855415 allele								
		A			*T*			Control*		
Cell	Protein	# unique peptides	% AAs identified	Quantitative value	# unique peptides	% AAs identified	Quantitative value	# unique peptides	% AAs identified	Quantitative value
hNSC	unnamed protein product	2	8%	5	-	-	-	0	0	0
	microtubule-associated protein 1B, isoform CRA_a	2	1%	3	-	-	-	0	0	0
	SRY (sex determining region Y)-box 5, isoform CRA_d	2	4%	3	-	-	-	0	0	0
	signal recognition particle 14kDa (homologous Alu RNA binding protein), isoform CRA_b	-	-	-	2	26%	2	0	0	0
	ribosomal protein S16, isoform CRA_b	-	-	-	2	12%	2	2	14%	1
	CUG triplet repeat, RNA binding protein 1, isoform CRA_e	-	-	-	2	7%	2	2	9%	1
	hnRNP-E1	-	-	-	4	18%	6	4	23%	4
	poly(rC)-binding protein 2 isoform b	-	-	-	7	28%	8	6	23%	4

		rs11855415 allele								
		A			*T*			*Control**		
Cell	Protein	# unique peptides	% AAs identified	Quantitative value	# unique peptides	% AAs identified	Quantitative value	# unique peptides	% AAs identified	Quantitative value
1321N1	microtubule-associated protein 4	6	7%	9	-	-	-	0	0	0
	forkhead box C2 (MFH-1, mesenchyme forkhead 1), isoform CRA_b	5	14%	7	-	-	-	0	0	0
	heat shock-related 70 kDa protein 2	2	4%	4	-	-	-	0	0	0
	general transcription factor II, i, isoform CRA_a	2	7%	3	-	-	-	1	3%	1
	actinin, alpha 4, isoform CRA_c	2	3%	3	-	-	-	0	0	0
	dermcidin preproprotein	2	20%	3	-	-	-	0	0	0
	unnamed protein product	2	3%	3	-	-	-	0	0	0
	Chain A, Core Of The Alu Domain Of The Mammalian Srp	-	-	-	2	26%	3	0	0	0
	DNA topoisomerase 1	-	-	-	2	2%	3	3	4%	2
	ribosomal protein S19, partial	-	-	-	2	12%	3	2	15%	1
	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	-	-	-	2	16%	3	0	0	0
	cold inducible RNA binding protein, isoform CRA_b	-	-	-	3	20%	4	4	14%	2
	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a-like 1	-	-	-	5	8%	7	4	5%	2
	unnamed protein product	-	-	-	8	30%	12	5	30%	3

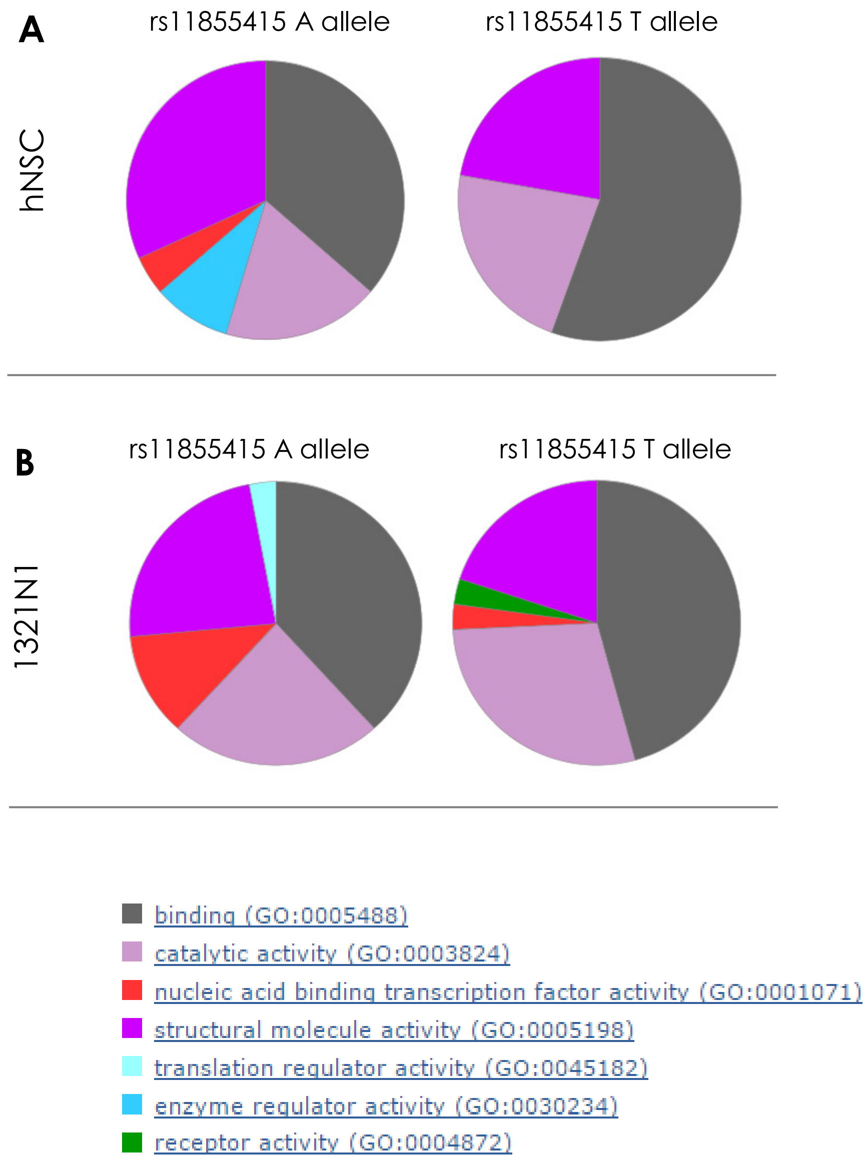


Figure 5.8 The distribution of hNSC and 1321N1 nuclear proteins identified by mass spectrometry and arranged according to molecular function. Each cell line hNSC (A) and 1321N1 (B) is subdivided in to proteins showing greater binding affinity to the rs11855415 A or T allele bait (see Table 5.1). Charts were generated by the PantherDB functional classification service (http://www.pantherdb.org/).

5.4.5 Luciferase assays

The presence of a genetic variant can perturb the function and expression of a gene product through a range of mechanisms including mRNA splicing, translation efficiency and TF binding (Pal *et al.*, 2015). However the presence of a transcription regulatory factor binding site as predicted by *in silico* analysis does not imply that binding affects transcription, and in many instances that is likely not the case (Doolittle, 2013, Graur *et al.*, 2013). As such, it is imperative to empirically test for the effects of genetic variation on gene expression using a range of functional assays, one of which is the luciferase reporter assay.

As previously discussed, the SNP rs11855415 is located in close proximity to a predicted secondary promoter within intron 13 of PCSK6, a region which appears to bidirectionally regulate the PCSK6-AS lncRNA and the PCSK6 SI. To assess whether this potential promoter is active in either a unidirectional or bidirectional manner and what effect genetic variation has on such expression I cloned the locus in to a luciferase gene reporter construct in both strand directions (beige box, track B, Figure 3.2). The cloned fragment was up to 1,839 bp long, spanning the predicted promoter region and nearby genetic markers associated with handedness, including rs11855415 and the VNTR. In total, I cloned 6 constructs in both a sense and antisense strand direction carrying all possible rs11855415/VNTR allele combinations including those not available from human gDNA (acquired via site-directed mutagenesis) (see Appendix E.2). Transfections were conducted in both neuronal (1321N1, hNSC) and non-neuronal (K562, HeLa) cell lines. Neuronal cells were chosen due to the high expression of PCSK6 in the central nervous system (Johnson *et al.*, 2003, Tsuji *et al.*, 1997), HeLa is a commonly used cellular model easily transfected while K562 showed the strongest signals for potential promoter activity in the ENCODE tracks at this locus (Bernstein *et al.*, 2012). No significant difference in luciferase expression was observed on allelic variation for either the rs11855415 SNP or the VNTR. The subsequent luciferase expression is displayed in Figures 5.9 and 5.10.

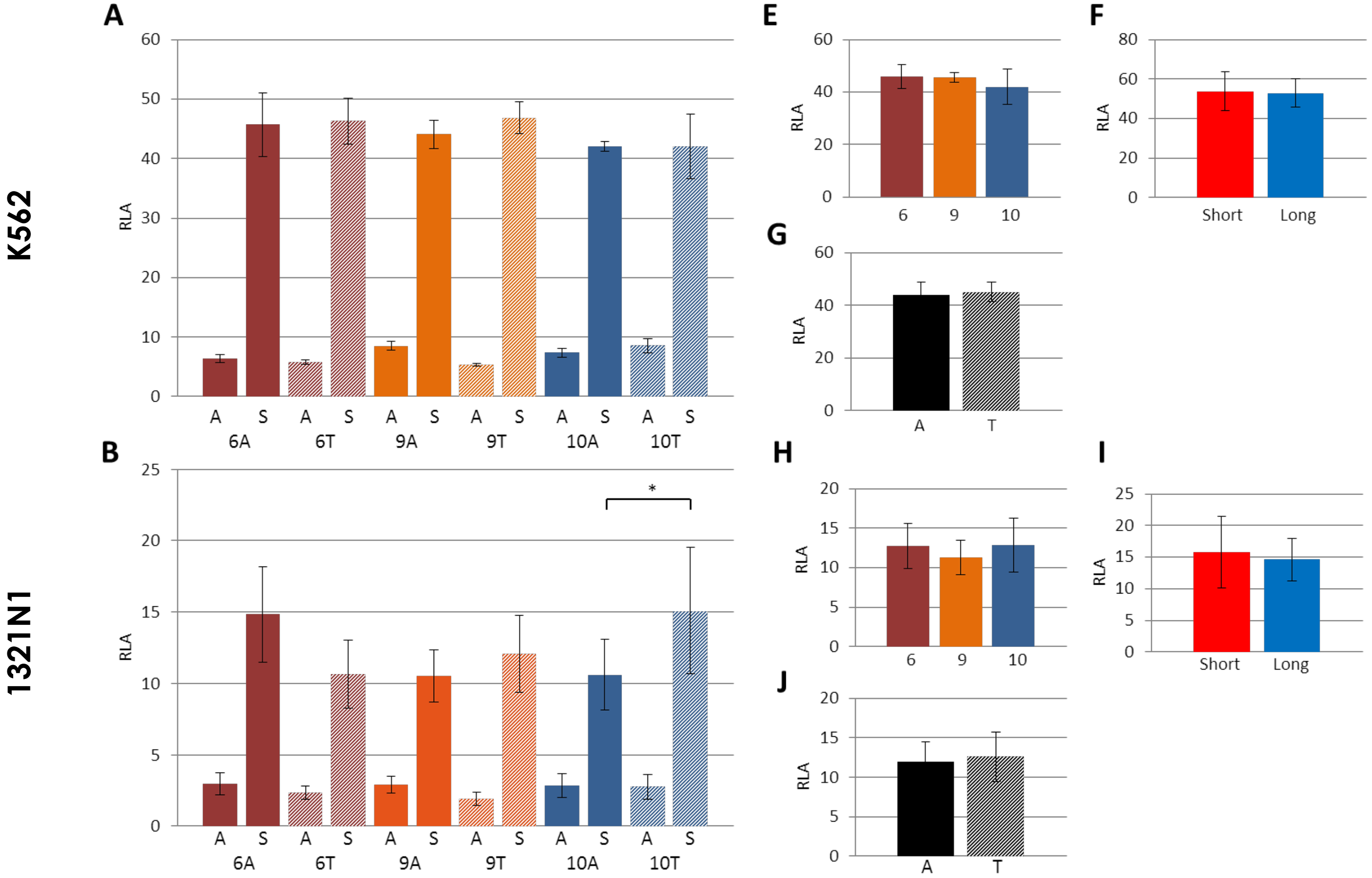


Figure 5.9 Dual luciferase assay results testing for allelic effects on promoter activity. Luciferase assay were conducted in K562 **(A, E, F** and **G)** and 1321N1 (**B, H, I** and **J**) cell lines. The alleles at rs1185415 (A as solid bars; T as stripe bars) and at the VNTR (6 is red; 9 is orange; and 10 is blue) were compared in different haplotypic combinations both in the antisense (A) and sense (S) direction (**A** and **B**). The rs11855415 alleles were compared regardless of the VNTR background (**G** and **J**). The VNTR alleles were analysed individually (**E** and **H**) or as short (6 repeats in bright red) and long (9 and 10 repeats in light blue; **F** and **I**), regardless of the SNP background. Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation. Data are representative of at least 3 independent experiments performed in triplicate and are expressed as Mean±SD of normalized luciferase activity (N = 3).*P-value of less than 0.05.

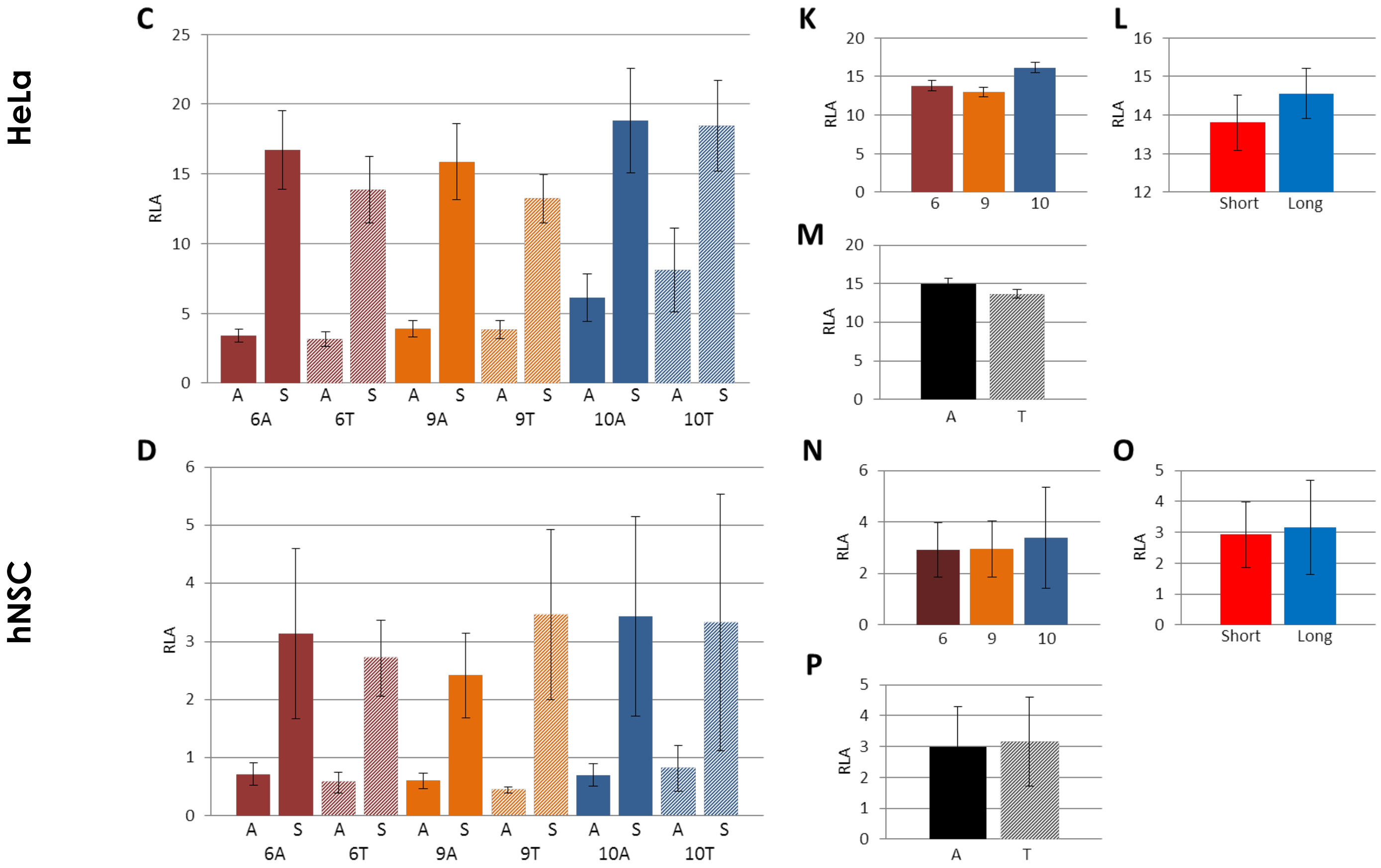


Figure 5.10 Dual luciferase assay results testing for allelic effects on promoter activity. Luciferase assay were conducted in HeLa (**C**, **K**, **L** and **M**) and hNSC (**D**, **N**, **O** and **P**) cell lines. The alleles at rs1185415 (A as solid bars; T as stripe bars) and at the VNTR (6 is red; 9 is orange; and 10 is blue) were compared in different haplotypic combinations both in the antisense (A) and sense (S) direction (**C** and **D**). The rs11855415 alleles were compared regardless of the VNTR background (**M** and **P**). The VNTR alleles were analysed individually (**K** and **N**) or as short (6 repeats in bright red) and long (9 and 10 repeats in light blue; **L** and **O**), regardless of the SNP background. Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation. Data are representative of at least 3 independent experiments performed in triplicate and are expressed as Mean±SD of normalized luciferase activity (N = 3).*P-value of less than 0.05.

Analysis of luciferase expression in Figure 5.9 (A, B) and Figure 5.10 (C, D) showed strongest promoter activity in the sense strand direction. A combination of all sense and all antisense luciferase expression in each cell line is summarised in Figure 5.11. In agreement with the ENCODE data (track C, Figure 3.2), the K562 cells were shown to have the most active promoter according to luciferase product detected.

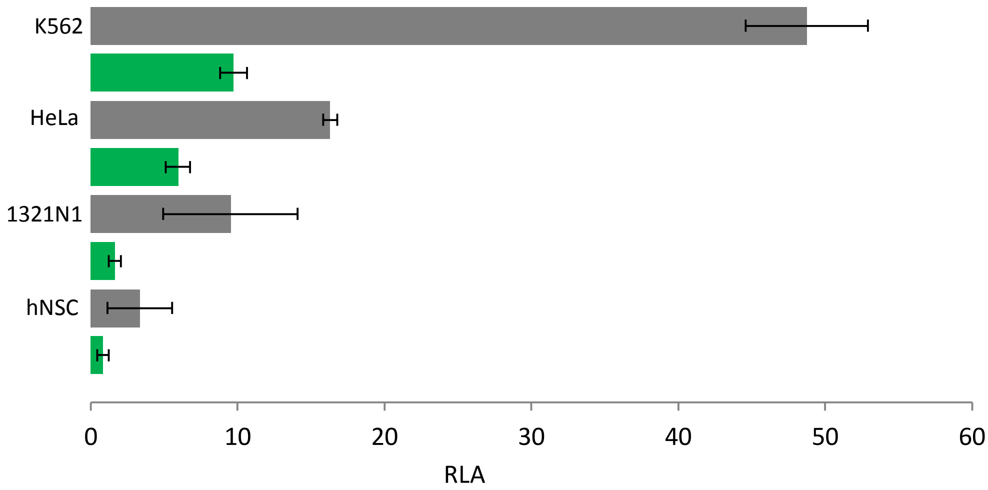


Figure 5.11 Bidirectional activity of the PCSK6 secondary promoter across multiple cell lines. The genomic region spanning the regulatory region at the PCSK6 locus was cloned into a luciferase reporter vector in both sense (grey) and antisense (green) strand directions. The bars are an average value of the sense and antisense luciferase expression as displayed in the four cell lines in Figure 5.10 (A,B) and Figure 5.11 (C,D). Bars show relative luciferase activity (RLA): mean fold change of luciferase expression following renilla normalisation and expression relative to the empty pGL4 vector. All assays were performed in triplicate and repeated at least four times. Error bars indicate Mean (SD) (N = 4)

To further characterise this promoter I generated a series of progressive deletions from the 1,839 bp-long cloned sequence using the 10T construct in the sense direction as the baseline (top bar, Figure 5.12). Analysis was conducted in the top three cell lines according to promoter activity in Figure 5.11 (K562, HeLa and 1321N1). The analysis revealed a minimal core promoter located in the region up to -155 bp relative to the TSS of the shorter PCSK6 isoform in all cell lines (Figure 5.12).

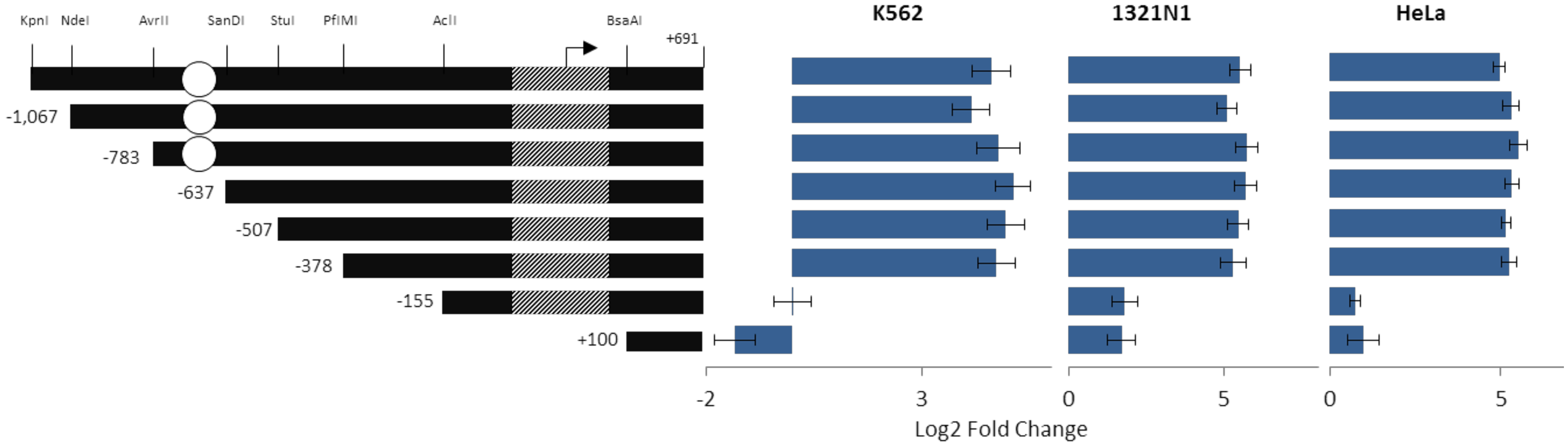


Figure 5.12 Luciferase minimal promoter analysis. Luciferase-expressing constructs containing different segments of the region upstream to the short sense PCSK6 isoform TSS were analysed for promoter activity in K562, 1321N1 cells. Restriction sites used to create the segments are shown on the top segment. The rs11855415 SNP and the VNTR location are indicated by a white circle and a stripe bar, respectively. An arrow indicates the position of the TSS of the PCSK6 SI. Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation and log transformed. All assays were performed in triplicate and repeated at least three times. Error bars represent the Mean (SD) (N = 3).

So far, luciferase assay results indicate a bidirectional promoter with a minimal presence within 155 bp of the SI TSS to display a bias in transcriptional orientation towards the sense strand direction.

Analysis of the minimal promoter results indicates the construct displaying the highest luciferase expression in both HeLa and 1321N1 cell lines is the shortest construct that contains both the rs11855415 SNP and the VNTR (3rd horizontal bar, Figure 5.12), a construct 284 bp shorter than the 1.8 kb insert of Figures 5.9 and 5.10. As such given the higher luciferase expression, I was interested to see what effect, if any, genetic variation of SNP rs11855415 and the VNTR would have on luciferase expression but this time using a shorter construct than previously. I tested for a difference in luciferase expression upon variation of the SNP rs11855415 alleles using the AvrII (-783 bp) construct from the Minimal Promoter assay (3rd horizontal bar, Figure 5.12). Results indicate the only allelic difference was observed for the rs11855415 A vs T constructs in the 1321N1 and HeLa cell lines. However this was not consistent across different tests. Overall the experiments do not demonstrate any allelic effects. Curiously, as in the longer constructs (Figures 5.9 & 5.10), HeLa displays an allelic difference in the opposite direction to all other cell lines tested (Figure 5.13, panel A). K562 shows the same direction of effect as 1321N1 though is not significant. In analysing what effect the VNTR might have on inferred transcriptional activity at the bidirectional promoter, no substantial allelic difference was observed in either K562 or 1321N1 cell lines however a slight increase in luciferase was reported for the HeLa 10 allele (P = 0.02, ANOVA single factor, CI = 0.89), though it is important to remember the PCSK6 SI was not detected in the HeLa cell line in previous analysis (lane 3, Panel A, Figure 4.4). There was no difference in the 1321N1 or K562 cell lines either when the VNTR data was defined as ‘short’ or ‘long’ according to the definition used by Arning *et al.*, 2013 which reported association between the rs11855415 SNP and degree of handedness. Note a similar analysis for all experiments has also been conducted in the antisense strand direction however, as indicated in Figure 5.11, luciferase expression was very low (RLA < 5) in all but one cell line and as such difficult to interpret (see Appendix E.4).

Finally, I performed site-directed mutagenesis on the TFBS centred at chr15:101875123 (site of rs11855415), replacing 6 bp with a scrambled sequence not predicted to contain a TFBS of any known TF according to TRANSFAC v2014.4 (see Appendix A for mutagenesis sequences, Appendix E.5 for confirmation Sanger sequencing and chromatograms of post-mutagenesis plasmids). Removal of the 6bp centred at chr15:101875123 (SNP rs11855415 location) leads to a significant increase in luciferase expression when comparing the rs11855415 A allele to the scrambled 6 bp construct in all cell lines tested (K562, HeLa, 1321N1 and hNSC). Refer to Figure 5.14 for luciferase expression results.

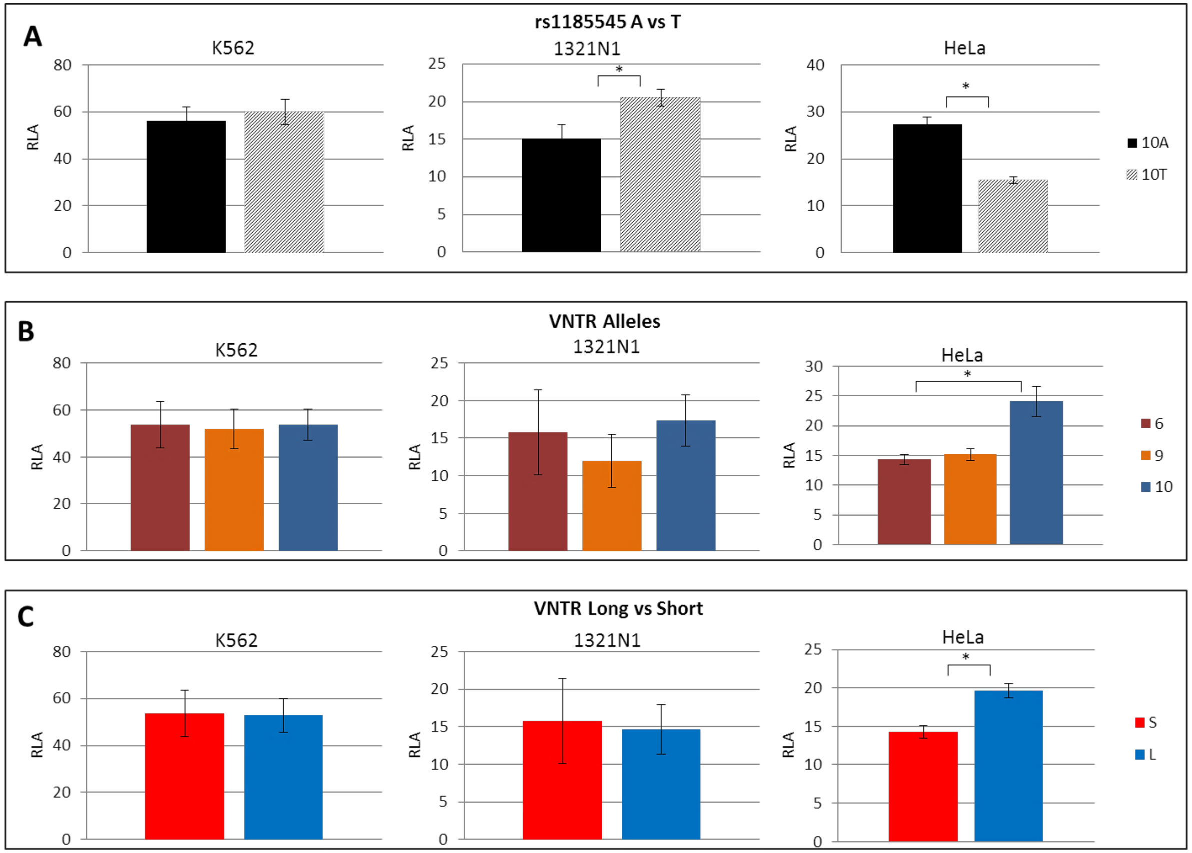


Figure 5.13 Dual luciferase assay results testing for allelic effects on promoter activity. Luciferase assay were conducted in K562, 1321N1 and HeLa cell lines using a shorter insert (3rd horizontal bar Figure 5.13) than the 1.8 kb insert of Figures 5.10 and 5.11; this was the shortest construct that would contain both rs11855415 SNP and VNTR. The alleles at rs1185415 (A as black bars; T as grey stripe bars) and at the VNTR (6 red; 9 orange; and 10 blue) were compared in the sense strand direction. The rs11855415 alleles were compared regardless of the VNTR background (A). The VNTR alleles were analysed individually (B) or as short (6 repeats in bright red) and long (9 and 10 repeats in light blue; C), regardless of the SNP background. Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation. Data are representative of at least 3 independent experiments performed in triplicate and are expressed as Mean±SD of normalized luciferase activity (N = 3).*P-value of less than 0.05.

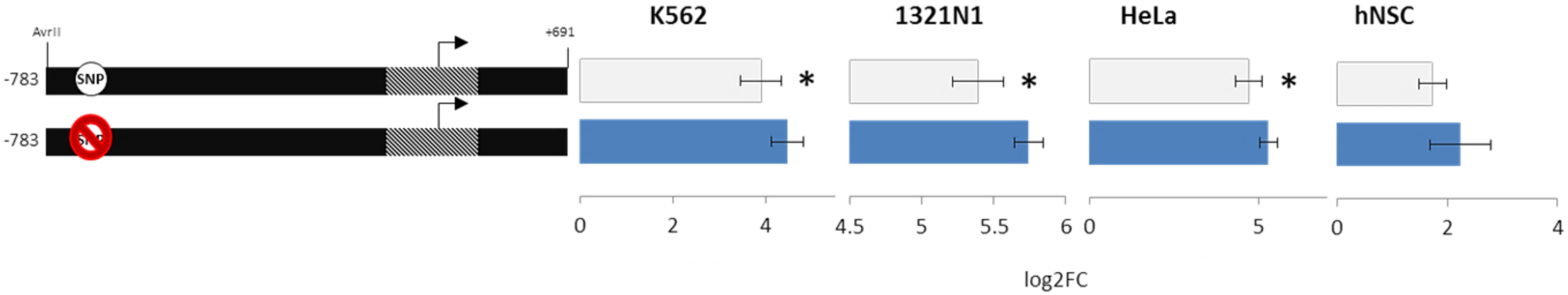


Figure 5.14 Luciferase-expressing constructs testing the effect of removing the TFBS centred on rs11855415. Both constructs were based on the 10A AvrII (-783 bp) construct from the Minimal Promoter assay with site-directed mutagenesis used to disrupt the 6 base pairs centred at rs11855415. An arrow indicates the position of the TSS of the PCSK6 SI. 10A construct is pale blue, post-mutagenesis construct in dark blue. Luciferase expression was measured relative to luciferase activity (RLA) of the empty pGL4 vector following renilla normalisation. This fold change (FC) has been log transformed for easier interpretation ($\log_2$ fold change). White circle and striped bar indicate rs11855415 SNP and VNTR respectively. All assays were performed in triplicate and repeated at least three times. Error bars represent the Mean(SD) (N = 3). * indicates significant difference at a value of P≤0.05 (Student's T-test, two-tailed unpaired).

5.5 Discussion

There exist several mechanisms by which a SNP or other genetic variant may affect the level of activity of a protein nearby in the DNA sequence. Any effect rs11855415 might have as a functional variant is most likely not on the RNA regulatory level; PCSK6-AS has previously been shown to undergo splicing in to multiple isoforms in which the intronic SNP is removed. miRNA databases (e.g. http://www.mirbase.org/) also return no predicted or curated miRNA bind sites within the lncRNA transcript. Alternatively, rs11855415 allelic variation may have a direct impact on a splice junction. Therefore any functional effect is most likely to be at the DNA level through effects on TF binding.

EMSA results provided in this chapter strongly support rs11855415, one of the top associated SNPs in previous handedness GWA studies, to have a significant direct effect on protein:DNA binding affinity (Figure 5.7). Results clearly demonstrate that the rs1185415 alleles create/disrupt the binding of transcription factor/s across a range of neuronal and non-neuronal cell lines in an *in vitro* context. Although an EMSA is useful in demonstrating the effect of allelic variation on the gain/loss of TF binding sites, it lacks the specificity to determine which TF exactly is binding to the probe and whether it is in complex or not; for example up to 12 different TFBSs are predicted for a 21 bp oligomer centred on the rs11855415 A allele (Table 3.1).

A subsequent Rev-ChIP assay using nuclear extract from the hNSC cell line identified, among others, SOX5 to bind to a probe centred on the rs11855415 A allele. A repeat of the experiment using the 1321N1 nuclear extract did not identify the SOX5 TF. A more robust approach to reducing background noise would be desirable in tailoring future Rev-ChIP experiments; for example to eliminate background and false positives typically seen in mass spectrometry results (e.g. keratin) researchers have used the stable isotope labelling by amino acids in cell culture (SILAC**)** method (see Ong and Mann (2006) for details). Nevertheless, Rev-ChIP results do support a previous *in silico* prediction (Table 3.1) in suggesting a binding of a SOX TF family member to the A allele and not the T allele at the rs11855415 site, suggesting a functional mechanism by which the SNP rs11855415 A allele might affect gene expression from the bidirectional promoter through the creation of a binding site for the SOX5 TF. Future experiments should be directed towards testing the binding of SOX5 to the rs11855415 A allele through use of a supershift EMSA assay, an assay similar to EMSA but with the addition of an antibody of choice to confirm binding specificity.

Luciferase assay results suggest that the PCSK6 locus containing a bidirectional promoter is within a region that is transcriptionally complex, with genetic variants and choice of cell line both influencing transcription from the promoter in a significant way. The term ‘bidirectional promoter’ may be considered a misnomer since nascent sense strand transcripts are at least eight times more abundant than divergent transcripts at more than half of yeast promoters (Churchman and Weissman, 2011); approximations that are in line with my own findings of sense vs antisense transcriptional activity (Figure 5.11). Luciferase results also support a rejection of the ‘steric interference’ model in which transcribing RNA PolII transcription units are thought to collide when transcribing on opposite strands. Polymerase collision is most likely when there are two strong convergent transcription units however my data shows a strong bias for transcription in the sense strand direction suggesting the collision of transcriptional complexes during simultaneous transcription to be unlikely at this PCSK6 locus. Furthermore, luciferase data supports previous EMSA findings in identifying a difference between the rs11855415 A and T alleles though why this result occurs in an opposite direction in the HeLa cell line (Figure 5.14, panel A) or whether this simply represents a false positive result remains unknown. Whether this reversal is related to the notion that HeLa expresses no isoforms at this locus (Figure 4.2, panel E) also remains to be clarified. Overall and across cell lines, for both short and long luciferase constructs there was no consistent difference observed between alleles for either the rs11855415 SNP or VNTR genetic variants. While these results did not support an allelic effect on promoter activity they do not rule it out either. Intronic promoters, lncRNAs, enhancers and other non-coding functional elements are likely to have roles that are tightly regulated in a tempo-spatial manner (Corradin and Scacheri, 2014, Rinn and Chang, 2012) and therefore putative allele-specific modulations might be detected only when all the relevant transcriptional machinery is in place. Therefore, it is possible that any genuine allele-specific effects might not be detectable through in vitro methodologies such as a luciferase assay, or might require specific cell types not used in this study.

Results displayed in Figure 5.14, in which luciferase expression from a construct with the 6 bp centred on the rs11855415 A allele is compared to a scrambled 'null' construct with the same 6 bp removed, suggest the TF(s) binding to the rs11855415 A allele *in vitro* have a prohibitive effect on transcription at the bidirectional promoter in the cell lines tested. When interpreting these results it is worth noting that the EMSA assay showed significant protein binding at the rs11855415 A allele, and not the T allele. This is a potentially interesting finding considering the previous *in silico* TFBS analysis (Table 3.1) and hNSC Rev-ChIP results (Table 5.1) which suggest the protein binding at the rs11855415 A allele may be SOX5, a TF known to act as a transcriptional silencer (Huang *et al*., 2008, Kwan *et al*., 2008), though further validation is required.

Luciferase assay results, coupled with a lack of predicted TFBSs within the tandem repeats support the VNTR to have no effect; the difference between the long and short VNTR alleles observed in the HeLa cell line (Figure 5.13, panel B&C) was not recorded in any other cell line or in the original 'longer' luciferase constructs (Figures 5.9 & 5.10) and remains ambiguous.

Finally, in this chapter I sought to identify whether the PCSK6-AS lncRNA has an effect on sense strand PCSK6 expression and if so what that mechanism of action might be. Using siRNA to affect a knockdown of PCSK6-AS did suggest a potential upregulation in transcripts upstream of the bidirectional promoter though all other isoforms remained unaffected. The difficulty in interpreting any such data however is that lncRNAs are expressed at very low levels, sometimes only at specific developmental stages and in specific tissues and, unlike most other RNA molecules, the function of a lncRNA antisense transcript can be mediated by either the transcript itself or via the act of its transcription. Antisense transcripts can remain at their location of transcription (such as through stalled polymerases, triple helices or R-loops which protect promoters from *de novo* methylation) which allows the RNA to exert its function in cis (Pelechano and Steinmetz, 2013). Therefore if PCSK6-AS1 does exert a regulatory influence through steric interference (though unlikely as discussed) then a cytosol or nucleus-directed knockdown will have little effect on the expression of sense-strand transcripts downstream of the PCSK6-AS site such as the PCSK6 long isoforms. Additionally, the limiting of the knockdown assay to one cell line, the modest transfection efficiency of a 'difficult to transfect' cell line (SH-SY5Y) and designing an effective siRNA knockdown probe in a restricted sequence which includes an Alu transposable element means the knockdown assay provides ambiguous results at best.

Results from the overexpression of the PCSK6-AS lncRNA in the SH-SY5Y cell line show no effect on expression levels of either the PCSK6 SI or any of the RefSeq-recognised PCSK6 isoforms, although future experiments should investigate the effects of overexpression across a range of both neuronal and non-neuronal cell lines. Combining previous epigenetic markings at the bidirectional secondary promoter's location (Figure 3.2, track C) with both knockdown and overexpression data, it's reasonable to suggest that the PCSK6-AS might have a regulatory effect not through a steric interference model but rather as an anchor point which recruits epigenetic remodelling proteins to allow transcription of the sense strand's PCSK6 SI to commence. Alternatively the very act of antisense transcription itself, rather than the produced transcript, has also been shown to induce chromatin modifications which are deposited during transcription and subsequently regulate the expression of the modified regions (Su *et al*., 2012). Such a model fits the overexpression assay data though whether this occurs *in vivo* and what exactly the functional effect of the PCSK6 SI isoform is remains unknown. However considering lncRNAs have been shown to regulate multiple major biological processes, including development (Ponting *et al*., 2009), differentiation (Guttman *et al*., 2011) and carcinogenesis (Gupta *et al*., 2010) such a hypothesis would appear plausible. Using constructs that harbour stretches of DNA for overexpression gene studies represents a useful though increasingly anachronistic tool in accessing the regulation of genes. An attempt to understand the role of the PCSK6 SI through overexpression was attempted using the CRISPR-Cas9 system (Ran *et al*., 2013) however owing to its novelty as an assay with unestablished design parameters, meant inconclusive and enigmatic results (data not shown).

In conclusion, the results of this chapter indicate that the SNP rs11855415 has a functional effect on sense strand expression, most likely by affecting a TFBS in the proximity of the bidirectional promoter and, by extension, the expression of the transcribed SI and lncRNA.

6 Concluding remarks & future perspectives

6.1 Summary of findings

A primary objective of this thesis was to identify and characterise the functional variant(s) contributing to the GWAS signal previously found to associate a locus of the PCSK6 gene with handedness.

The results in this project demonstrate that a common polymorphism (rs11855415) residing upstream of a secondary bidirectional promoter within intron 13 of PCSK6 has an effect on transcription factors which bind at that location. By influencing promoter activity, genetic variation could affect expression of the PCSK6 shorter isoform, a transcript which RNA-seq and microarray analysis of the developing and adult human brain suggest to be expressed in the corpus callosum, a region of the brain thought to influence the handedness phenotype.

Handedness is a complex phenotype which represents multiple characteristics including manual dexterity, visual-spatial awareness and grip. In Chapter 2 an examination of a correlation matrix between 5 such performance-based measures for relative hand skill revealed a marking task and not the peg-board task to have the highest correlation with a measure for hand preference (Hand7) across all subgroups analysed. A weak correlation between the PegQ and other measures for hand performance was observed across all subgroups (r = -0.1 – 0.3) with little variation in correlation between the subgroups. The only measure for hand performance to display significant genetic association with the SNP rs11855415 was the PegQ measure (RD & Affected subgroups, Table 2.11). A further filtering of the Affected subgroup in to its constituent disorder cohorts of pure SLI, ADHD and RD resulted in no one disorder cohort displaying significant association between rs11855415 and PegQ7. The same allelic trend was observed across these disorder cohorts i.e. individuals with the minor ‘A’ allele of rs11855415 have significantly greater relative right-hand skill compared with those carrying the major ‘T’ (ancestral) allele. This is in contrast to the general population which displays a trend towards reduced laterality of hand skill for the minor allele. Such a finding supports the notion that PegQ data should be collected in children across a range of neurodevelopmental disorders rather than just in dyslexic individuals. My data also shows hand preference to be only weakly correlated with foot preference, although the correlation between handedness and eyedness was weaker than the correlation between handedness and footedness.

The data generated in Chapter 3 defined a 12.7 kb region with the marker rs11855415 predicted to affect the largest number of TFBSs. A multi-sequence alignment also suggests the first exon of PCSK6-AS to be within a CNS, though there appears to be a lack of homology for the lncRNA on the whole; an unsurprising finding given evolutionary evidence for selected effect functionality of lncRNAs, in general, is meagre (Haerty and Ponting, 2013) and the proportion of lncRNA sequence that is under purifying selection appears to be small, approximately 5% (Ponjavic *et al.*, 2007). Chapter 4 provided confirmation for the first time of the existence of a novel, though predicted to be inactive, PCSK6 isoform whose transcription is driven by a secondary bidirectional promoter within the 12.7 kb region previously defined. This promoter was also shown to generate several novel gene isoforms of a lncRNA PCSK6-AS in the antisense strand direction. Exploratory RNA-seq data suggests PCSK6 SI expression to be relatively high in the developing and adult corpus callosum.

Finally, the EMSA protocol in Chapter 5 provided a relatively cost-effective and accessible way of validating candidate genetic variants for further study. A significant and robust allelic difference was observed for rs11855415 in all neuronal and non-neuronal cell lines analysed, results which supported previous *in silico* predictions and show rs11855415 to have a substantial effect on *in vitro* protein:DNA interaction. The Rev-ChIP assay also demonstrated the usefulness in employing an unbiased approach linking detected SNPs from GWA studies directly to a TF protein without *a priori* knowledge, though as a hypothesis-generating assay any results still require validation (via, for example, the supershift EMSA assay). This assay, using hNSC nuclear extract, posited a number of proteins to exclusively bind to the rs11855415 minor A allele but not to the major T allele; one of which was a member of the SRY (sex determining region Y)-box (SOX) TF family, as previously predicted by *in silico* analysis in Chapter 3. Reporter gene studies identified a minimal promoter capable of controlling expression of the PCSK6 SI and show that replacement of the 6 bp centred at chr15:101875123 (SNP rs11855415 location) with a scrambled ‘null’ sequence leads to a significant increase in luciferase expression when comparing the rs11855415 A allele to the scrambled 6 bp construct in all cell lines tested (K562, HeLa, 1321N1 and hNSC) suggesting the TFs which bind, *in vitro* at least, to have a prohibitive effect on luciferase expression and, by inference, transcriptional activity at the secondary promoter. The luciferase assay also supports the notion that overall across all cell lines there was no consistent difference observed between alleles for either the rs11855415 SNP or VNTR genetic variants. While these results did not support an allelic effect on promoter activity they do not rule it out either. Though inconclusive, siRNA-mediated knock-down of PCSK6-AS in the SH-SY5Y neuronal cell line also showed a specific increase in PCSK6 shorter isoforms, while overexpression of the PCSK6-AS had no regulatory effect on PCSK6 expression.

6.1.1 A proposed model derived from the findings of this thesis

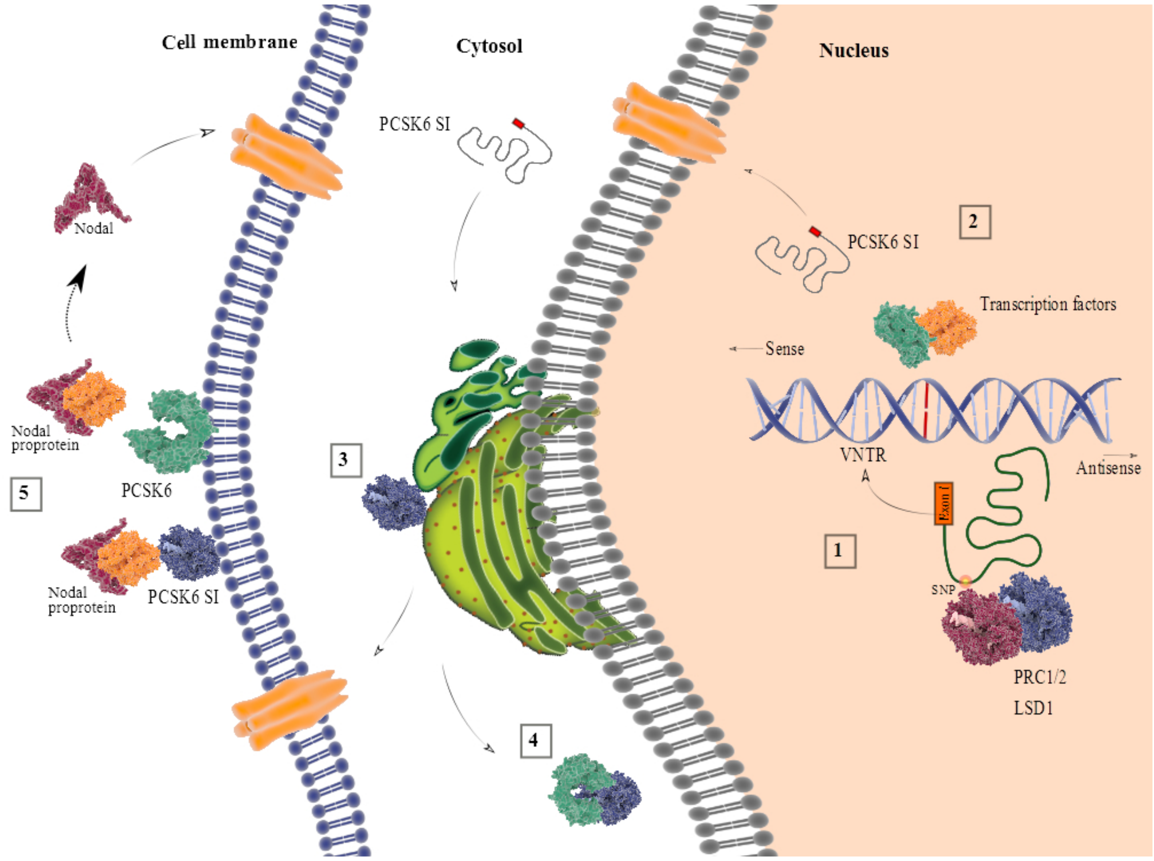


Figure 6.1 A proposed model derived from thesis findings. The PCSK6-AS lncRNA (green transcript) is transcribed from the PCSK6 secondary bidirectional promoter in an antisense strand direction. Allelic variation of the rs11855415 SNP (red nucleotides) may have an effect on chromatin-modifying enzymes such as PRC1/2 and LSD1 binding to the subsequent transcribed lncRNA **(1)** or could affect transcription factors in activating/repressing transcription in the sense strand direction **(2)**. Several scenarios might exist once the sense strand PCSK6 SI is transcribed and exported to the cytosol for further processing: remain in zymogen form in the ER **(3)**, interact with substrate proteins within the cell **(4)** or tether to the cell membrane where the canonical PCSK6 protein is known to cleave the Nodal morphogen during early embryogenesis **(5)**. An extended discussion on the figure can be found in the discussion immediately below.

EMSA results from Chapter 5 indicated a significant difference on allelic variation in the proteins that will bind to the sequence centred on SNP rs11855415. The downstream ramifications of such a finding are best approached by discussing potential effects according to the direction of transcription. In the antisense strand direction (Figure 6.1 **(1)**) a specific secondary structure could potentially permit the PCSK6-AS lncRNA (shown in green) to act as a scaffold and interact with different chromatin-modifying enzymes (e.g. PRC1/2, LSD1), thereby coordinating their action and directing specific epigenetic modifications of the nearby chromatin (altering the accessibility of the genome to RNA Polymerase II and its associated factors is arguably the most efficient means to activate or repress transcription broadly). Multiple lncRNAs have been found to mediate changes in chromatin structure (e.g. HOTAIR, Gupta *et al.* (2010)) and this is currently the fastest growing class of lncRNAs known to regulate transcription. Sequencing of PCSK6-AS confirmed Exon 1 of the lncRNA transcript to have substantial overlap with the VNTR's tandem repeat sequence, in theory permitting the Exon 1 to bind and act as an anchor for the lncRNA to interact with the chromatin-modifying enzymes. Such a model would not only explain why an overexpression/knockdown of the lncRNA had no effect on PCSK6 SI expression but also why VNTR allelic variation in the luciferase assay had no effect on expression in the majority of cell lines analysed and why our genetic association studies discounted the VNTR from having any effect. Such a model would also accommodate the positive correlation between PCSK6 SI and PCSK6-AS expression while discounting the 'steric interference' model (i.e. when RNA polymerase collide when transcribing on opposite strands) such a correlation suggests.

Alternatively in the sense strand direction (Figure 6.1 **(2)**), and if we are to consider the presence of the lncRNA as an expected by-product of endogenous bidirectional behaviour thought to exist at all promoters, then SNP rs11855415 allelic variation (nucleotides shown in red) could simply be influencing transcription factor bind sites (TFBSs), as demonstrated in the EMSA assay in Chapter 5. Disruption of this TFBS in the luciferase assay suggests the TFs to bind to the rs11855415 A allele to have a repressive effect; Rev-ChIP results propose this to be SOX5, a TF known to have a repressive effect on transcription although further validation is required. It's also worth noting that lncRNAs can themselves act in an enzymatic manner by functioning as ligands for transcription factors (e.g. MALAT1, Ma *et al.* (2015)) and though PCSK6-AS1 is not predicted to be active, the newly-discovered PCSK6-AS2 is predicted to interact directly with other proteins (see Appendix C.2).

RNA-seq data from a human neural stem cell (hNSC) line in addition to protein prediction software support PCSK6 SI to have a signal peptide/transmembrane domain at the N-terminus (Figure 6.1 **(2)**, red box), which directs translocation in to the endoplasmic-reticulum (ER) for further processing. Based on software analysis the PCSK6 SI could enact several roles. The loss of a prodomain means PCSK6 SI lacks an autoproteolytic initial cleavage step, rendering the protein inactive and localised to the ER as in Figure 6.1 **(3)**; several PCSK6 isoforms are thought to remain in zymogen form in the ER (see Table 4.1). The SI is predicted to lack a catalytic domain which contains the catalytic triad necessary for cleavage however an insulin-like growth factor receptor domain IV which confers protein-protein interaction properties (Figure 6.1 **(4)**) and directs cell surface tethering (Figure 6.1 **(5)**) is retained. This fourth extracellular domain is common to all receptor tyrosine protein kinases and regulates ligand binding to receptor domains (Cho and Leahy, 2002). As discussed in Figure 1.4 since Nodal is known to regulate its own expression via a feedback circuit (Shen, 2007) then PCSK6 SI could play a regulatory role early in embryogenesis in the Nodal pathway by binding the Nodal proprotein at the cell surface without cleaving it, thereby acting as a tempo-spatial fine-tuning mechanism for controlling available Nodal morphogens. Such a scenario clearly extends my hypothesis beyond the empirical findings of this project but such a model could accommodate my results while acknowledging how allelic variation at a single nucleotide polymorphism could induce phenotypic variation for a complex trait such as handedness.

6.2 Future Perspectives

The findings and subsequent discussion arising from Chapter 2 raise several points worthy of note here. For example, the MLRA model derived in this chapter employed a combination of performance measures to best predict hand preference and provides a complimentary approach to a questionnaire-based measure of preference such as the WHQ (Elias *et al.*, 1998) or the Edinburgh Handedness Inventory (Oldfield, 1971). Such an alternative approach may prove useful in future studies since no handedness questionnaire has been designed for explicit use with younger and special populations even though several challenges exist, for example the inherent verbal requirements and inability to assess children's familiarity of specific items and tasks are particularly problematic. Another challenge facing researchers is the modality of the hand performance distribution; because of their asymmetry and bimodality, J-shaped distributions such as hand preference cannot be easily analysed with parametric measures of central tendency such as means - the failure of the MarkQ hand performance measure to show significant genetic association with SNP rs1185545 may arise from its inherent bimodal distribution (see Discussion section 2.5). Alternatively, SNP rs11855415 might just influence a particular aspect of handedness or combination of traits (fine-motor control, spatial awareness etc.) which are more accurately recorded with the peg-board task. Either way it is clear that researchers must choose carefully which performance measures to use in analysing hand skill differences, particularly in children.

Chapter 4 displayed results for PCSK6 SI expression by interrogating the ENCODE datasets related to RNA-seq, a technology that is quickly making microarrays increasingly obsolete for gene expression analysis such that just a few million reads are needed for detecting expressed transcripts with a sensitivity below a single transcript per cell (Ramskold *et al*., 2009). Nevertheless, the functional relevance of the majority of gene isoforms, including PCSK6 SI, remains difficult to ascertain, particularly in light of the rapid change of isoform usage in evolution, indicating relatively weak selection pressure in the process (Barbosa-Morais *et al*., 2012). Future analyses therefore will primarily involve defining both the functional role and mechanism of the PCSK6 SI, particularly in the corpus callosum; a region of the brain in which it is thought to be highly expressed. In general, the increased availability of RNA-seq datasets derived from a broad range of cell and tissue types will help broaden our understanding of the transcriptome's influence on complex phenotypes such as handedness.

Like gene isoforms, our interpretation of the role of bidirectional promoters and lncRNAs in influencing complex phenotypes is continuously evolving. Several studies have shown that lncRNAs are transcribed from genomic regions associated with disease risk and complex traits. For example, Cunnington *et al*. (2010) showed modulation of

ANRIL lncRNA expression mediates susceptibility to several important human diseases including coronary disease, stroke, diabetes, melanoma, and glioma though further research is required for elucidating exactly how genetic variation affects lncRNA function. Future experiments involving genome editing tools such as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), a genome editing system which can be used for the introduction of specific variants into engineered cell lines, might be useful in removing or overexpressing PCSK6-AS and measuring subsequent gene expression. For example, an investigation in to the effects of overexpression via the pcDNA-dCas9-p300 Core CRISPR plasmid (#61357, Addgene) would be useful in measuring the effect acetylation has on the secondary promoter region in a cell line that previously displayed low PCSK6-SI expression (such as HeLa).

Currently, there is a scarcity of structural information illustrating lncRNAs bound to their protein targets. Crystal or nuclear magnetic resonance (NMR) structures of lncRNA/protein complexes, even consisting of the minimal domains that interact, will provide highly valuable pictures of the complexes, enabling experiments to test directly structure/function relationships. Less labour-intensive computational approaches are also beginning to predict how lncRNAs interact with DNA or chromatin by suggesting the involvement of lncRNA in transcriptional repression/enhancement by identifying complementary DNA sequences within lncRNA-associated regions that might indicate direct RNA–DNA–DNA triplex formation (Buske *et al.*, 2012, Vance *et al.*, 2014). An alternative method by which the PCSK6-AS may be having an effect is via the lncRNA's third exon, an Alu repeat that might be mediating intermolecular interactions between RNA molecules and leading to functional consequences; several other systems have previously provided similar examples of Alu repeats and other abundant repeated sequences in mammalian genomes to have a role in gene expression regulation (Wang *et al.*, 2013, Gong *et al.*, 2013, Holdt *et al.*, 2013).

Many post-GWAS studies have focused on *cis*-regulatory variation to explain disease associations. However it is becoming increasingly clear that genetic variants can also affect the expression or function of not just lncRNAs such as PCSK6-AS but other ncRNAs too. For example, miRNAs typically regulate gene expression through binding to 3′ UTRs of target mRNAs to direct their post-transcriptional repression and several studies (Richardson *et al.*, 2011, Gamazon *et al.*, 2012) have demonstrated genetic variants within the 3′ UTRs of susceptibility genes at miR-binding sites are associated with disease risk and should be routinely considered in post-GWAS functional studies. miRNA databases such as miRBASE (www.mirbase.org) fail to return a predicted or curated miRNA bind site within the PCSK6-AS lncRNA transcript though there does appear to be a miRNA bind site centred at rs1030, a SNP in the 3'UTR of the PACE4C and PACE4CS isoforms (Figure 4.2) and in moderate LD with rs11855415 (r2 = 0.59, D' = 0.95, CEU population HapMap 3rel2 dataset). rs1030 was imputed in the previous handedness GWAS (Brandler *et al.* 2013) though owing to its weaker association with the PegQ phenotype (P = 0.0002, MAF = 0.26) it is perhaps unlikely to represent the etiologic variant.

Allele-specific protein binding effects via the EMSA and Rev-ChIP assays were investigated in Chapter 5. However, due in part to their *in vitro* nature, such assays will have a propensity for giving false-positive results. Future experiments might involve the combination of technologies such as 5C/Hi-C, assays which assess chromosome interactions on a genome-wide scale, and CRISPR may prove highly influential in identifying and validating the gene(s) directly affected by candidate variants detected by GWAS.

Modelling the effect of validated variants via *in vitro* and *in vivo* model organism experiments will provide further avenues for studying genetic traits and disease since one of the biggest challenges in post-GWAS validation is being able to accurately evaluate the effects of SNPs and their associated genes in an intact organism. The mouse is usually the mammalian model of choice because of the multiple methodologies which allow for genetic manipulation, its genome similarity and the ability to mimic human multifactorial disease phenotypes (Cho *et al.*, 2013). Additionally, selected mice lines have shown to be a useful mammalian genetic resource for studying the neurobiology of cerebral lateralisation since multiple factors of handedness, previously identified in humans and other primates, also exist in mice (Li *et al.*, 2013). Zebrafish is another popular model organism for studying the effect of allelic variation on function and has a number of advantages for post-GWAS analysis over its rodent mammalian counterpart including ease of genetic manipulation, rapid production of large numbers of organisms of a specific genotype, and the capacity to study tissue-specific gene expression in live organisms (Edwards *et al.*, 2013). For zebrafish, regulatory elements can also be assessed by the generation of transgenic zebrafish by means of reporter constructs and the microinjection of mRNA, DNA, or morpholinos into early embryos; a strategy being pursued by our group through the work of my colleague Monika Gostic.

6.3 Concluding remarks

Functional assays are the major bottleneck in the identification and functional characterisation of causal variants responsible for the association signals detected by GWAS. While EMSA, ChIP and luciferase assays are widely applied and are the backbone to many molecular genetic analyses, the development of high-throughput screening methods that can accurately and sensitively screen functional candidate variants, possibly via the field of genome engineering, would represent a significant breakthrough.

There are several key findings from this project and points of discussion elsewhere that I have considered in deriving my concluding hypothesis (see Figure 6.1):

- PCSK6 has a known fundamental role in the body LR patterning Nodal pathway. Brandler *et al.* (2013), provide support for the notion that the mechanisms responsible for setting up LR body asymmetry might influence handedness and brain asymmetry.

- The corpus callosum (CC) is a fundamental component in the neurodevelopmental process of cerebral midline development. Previous research suggests both dyslexia and handedness are influenced by the size and function of the CC e.g. the CC in dyslexic individuals is of different size and shape (see section 4.2).

- Brain-specific isoforms are known to exert substantial phenotypic effects (e.g. the morphogenic signalling of CDC42-palm in hippocampal neurons, Wirth *et al.* (2013)). Though it is possible to detect PCSK6 SI in various cell lines, the *in vivo* expression profile of this novel gene isoform and its specificity to brain tissue is as yet unknown.

- Data from this project support a high expression of PCSK6 SI relative to other PCSK6 isoforms in the corpus callosum of the developing brain, aberrant expression of such a PCSK6 isoform could have an effect on CC development which in turn could influence a variety of traits and disorders including handedness.

As such it seems reasonable to hypothesise that allelic variation of the rs11855415 SNP affects the transcription factors binding at that location, thereby effecting promoter activity at the bidirectional promoter which influences expression levels of PCSK6 SI, a transcript thought to be highly expressed in the developing corpus callosum.

Conversely, as indicated in Figure 6.1, PCSK6 SI may enact its presence much earlier during embryogenesis through interaction with the Nodal morphogen. Ultimately, and on a broader scale, it may be that the functional change in PCSK6 subtly alters the initial left–right patterning of the early embryo, and this has a downstream effect during neuronal migration on the development of cerebral asymmetry.

In summary, this project described the first functional characterisation of a locus associated with human handedness within PCSK6, a gene controlling the establishment of structural asymmetries. Data suggest that the association between handedness and common variants is mediated by an intronic bidirectional promoter controlling both sense (PCSK6 SI) and antisense (PCSK6-AS) transcripts; observations which are in agreement with the increasing evidence that support the role of genetic variants within non-coding regions in influencing complex phenotypes (Ward and Kellis, 2012). The reduction in sequencing costs and the expansion of the publicly-available RNA-seq datasets will ultimately enable differential expression of the PCSK6 SI in both clinical and general population cohorts across a range of tissue types and thus enable a greater understanding of the complex web of association between handedness, dyslexia and cerebral/body asymmetry pathways. Future studies will aim to understand the function of the transcripts regulated by the bidirectional promoter in order to elucidate the mechanisms by which genes controlling structural laterality are also implicated in behavioural and functional asymmetries.

APPENDICIES

Appendix A Primer Catalogue

A.1 PCSK6 isoform profiling

Isoform PACE4-AI

5’- CACCCCAGGCTCTGCTAATA

5’-ATGCTGCTCCTGGGGAGATA

Isoform PACE4-AII

5’- TCCTGAAGATGAGGAAGATTACAC

5’-ATGCTGCTCCTGGGGAGATA

Isoform PACE4B

5’- CTCGGGAACCAAGTCTCAAC

5’-TTGGAGGACTCGCACTTTCT

Isoform PACE4C

5’- TCCTGTTGCAAATCAACTGACC

5’-TGGCTTTGGTCATCTGTCCC

Isoform PACE4CS

5’- GAGCATCCCCTTAGTGCAGG

5’-TGTTCAATCTGCCACCGGAA

Isoform PACE4D

5’- CCTGGGCTCCATTTTCGTCT

5’-GTGACCTGAGGGTTCTTCCG

Isoform PACE4E-I

5’- TCCTGAAGATGAGGAAGATTACAC

5’-GTCCACCAATGGGGTGTGAG

Isoform PACE4E-II

5’- CACCCCAGGCTCTGCTAATA

5’-GTCCACCAATGGGGTGTGAG

A.2 PCR primers

See Figure 4.13 for primer locations

PCSK6-AS1 (Exon 1 – Exon 2)

5’-GGTGCAGAAAACAAGCCTG

5’- CTTCCCTGCTGGCGTTTTTG

PCSK6-AS1 (Exon 1 – Exon 3)

5’-GGTGCAGAAAACAAGCCTG

5’-AAAGGCAGGAAAACCAAAGT

PCSK6-AS2/3

5’-GGTGCAGAAAACAAGCCTG

5’-TGCCAAAAGAGTTATAGGTGATT

Annotating the shorter PCSK6 isoform (Fig 4.4)

PCSK6 Exon 13 -14 (dark blue arrows)

5’- GTTGCTGGATCTTTCCAATG

5’- CTGATGGGCACTGAAGGTGT

PCSK6 Short F – Short R (light blue arrows)

5'-GAACAACTTCCTGTGTCACTGC

5'-ATGCTGCTCCTGGGGAGATA

Novel exon – Exon 13

5'-CGCTGCAGCAGTGACACAGGA

5'-ATGCTGCTCCTGGGGAGATA

Beta-Actin

5'- GCTCGTCGTCGACAACGGCTC

5'- AAACATGATCTGGGTCATCTTCTC

VNTR genotyping (Figure 4.16)

5'- ACAGGGCTCGGTTCATTAAG

5'- TCGGAATGTGGCTGTAACTG

A.3 qPCR primers

See Figure 5.3 for PCSK6 primer locations

PCSK6-AS

5'-GGTGCAGAAAACAAGCCTG

5'-TTGGTCCCACTGCTTCTTCC

PCSK6 shorter isoform (SI)

5'- GCAGCGGTGAGAACAACTT

5'- CTGATGGGCACTGAAGGTGT

FANCC Housekeeping gene

5'- AGCTGCGGTTTGCACTCA

5'- GTCCCCGAGGGATATCTTGA

GAPDH housekeeping gene

5’-TCTATAAATTGAGCCCGCAGCC

5’-GACCAAATCCGTTGACTCCG

POLR2F housekeeping gene

5’-CCCGAAAGATCCCCATCAT

5’-CACCCCCCAGTCTTCATAGC

PCSK6 ‘short’

5’-TGACGCCTTTCCCCAAAACT

5’-TTGGTTGCATTTCTCCCCGA

PCSK6 ‘both’

5’-CTGGTTTCTCCCTCGGGAAC

5’-CCTGGGATGGCAGATCTTGG

PCSK6 ‘long’

5’-GGAGTGTGGTGACAAAGGCT

5’-TGCTGTGTCCCCAAAGTAGC

TaqMan rs11855415 genotyping probe (Figure 4.15)

VIC:

ACTGGAATGGAAGAGAGACTTCATT**A**TTATTACACTCTCTGTTTGACTTTA

FAM:

ACTGGAATGGAAGAGAGACTTCATT**T**TTATTACACTCTCTGTTTGACTTTA

A.4 Overexpression & Knockdown primers

PCSK6-AS Stealth siRNA (targets Exon 2 of PCSK6-AS1)

5’-GGGUUUCAGAAUGUUUGCCAGGAUG

PCSK6-AS ASO Knockdown (targets Exon 3 of PCSK6-AS1)

5'-GCGTGCCTCCCAAAGTGCTG

For isolating PCSK6-AS gDNA for overexpression

5'-CCCGGGGGATCCGGTGACAGCGACACAGGAA

5'- CCCGGGCTCGAGAGGAAAGAGCCCAGGAGGAA

SNCA gene positive control for Stealth siRNA knockdown

5'-CAUGCUUCCAGAGAAUGCAUAUUCU

For SNCA Stealth siRNA positive control knockdown qPCR quantification

5'-TAAAACCTGCAAATTCACATCTTC

5'-AAGTAGGTAAGTAGGGCAGTGCAT

A.5 Luciferase assay primers

Produce an amplicon for cloning into the pCR™-Blunt II-TOPO vector

5'-CTGGCTCTAAATGGCAGCCT

5'-ACCCCGAGTACTACTGCTTTT

Mutagenesis from 6/6T -> 6/6A for rs11855415

5'-AGAGAGACTTCATTATTATTACACTCTCT

5'-AGAGAGTGTAATAATAATGAAGTCTCTCT

Mutagenesis from 10/10A -> 10/10T for rs11855415

5'-AGAGAGACTTCATTTTTATTACACTCTCT

5'-AGAGAGTGTAATAAAAATGAAGTCTCTCT

Mutagenesis to disrupt the TFBS centred on rs11855415

Note: The 6 base pairs in bold indicate the sequence replacing the original 6 base pairs at that location. The replacement oligomer sequence was not predicted to create a TFBS according to TRANSFAC v2014.4. See Appendix E.5 for sequencing chromatograms confirming this replacement sequence.

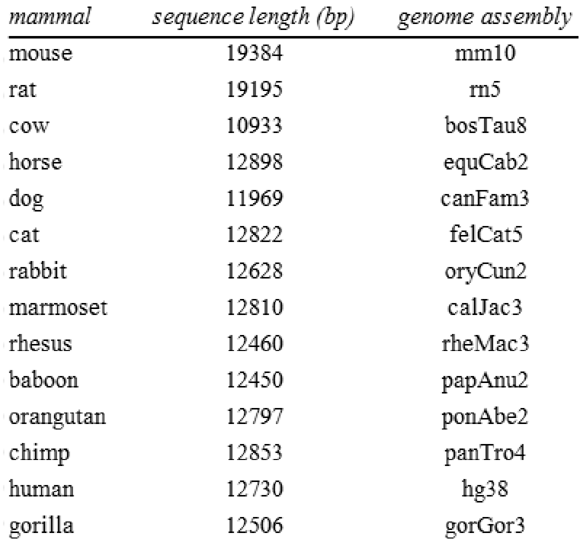
5’-TCAAACAGAGAGTGTAA**GCTAGC**TGAAGTCTCTCTTC
5’-GAAGAGAGACTTCA**GCTAGC**TTACACTCTCTGTTTGA

Appendix B Supplementary Materials and Methods

Table B.1 Indicates linkage disequilibrium (LD) between the SNP rs11855415 and all available SNPs within the defined PCSK6 locus of interest (chr15:101863220-101875949,hg19) from a CEPH population. Genotype data was downloaded from the 1000 Genomes Pilot project (Genomes Project *et al.*, 2010) and LD calculated using the Broad Institute SNAP v2.2 service (Johnson *et al.*, 2008). Bold indicates SNPs located within a non-coding conserved sequence across 14 eutherian mammals (see Methods, Chapter 3). Where available a P-value has been provided indicating that marker's association with the PegQ measure of relative hand skill (taken from a GWA meta-analysis of individuals with reading disability, Brandler *et al.* (2013))

SNP	*Distance*	*r^2*	*D'*	*Position(hg18)*	*Major*	*Minor*	*MAF*	*CNS*	*P-value*
rs9806256	11550	0.8	1	99681096	T	C	0.2	No	1.7×10^{-7}
rs1871978	11834	0.8	1	99680812	C	T	0.2	No	1.23×10^{-7}
rs7182874	9651	0.6	1	99682995	C	T	0.25	No	8.68×10^{-9}
rs1871976	11991	0.6	1	99680655	A	G	0.25	No	
rs752028	5938	0.527	1	99686708	C	T	0.275	No	5.02×10^{-8}
rs3825921	12960	0.385	1	99679686	C	T	0.342	No	4.69×10^{-6}
rs12900794	282	0.29	1	99692364	C	T	0.408	Yes	
rs882422	5868	0.2	1	99686778	G	A	0.5	No	
rs752026	5752	0.175	1	99686894	A	G	0.467	No	
rs1471656	7636	0.175	1	99685010	C	T	0.467	No	
rs2220055	10527	0.167	0.55	99682119	G	A	0.1	No	
rs12916087	360	0.093	1	99692286	A	G	0.317	Yes	
rs1871975	12153	0.089	0.80	99680493	T	C	0.408	No	
rs9806218	11569	0.079	1	99681077	G	A	0.283	No	
rs2073592	2817	0.076	1	99689829	G	A	0.275	Yes	
rs2277593	10126	0.053	1	99682520	G	A	0.208	Yes	
rs1947942	6072	0.05	1	99686574	A	G	0.2	No	
rs755867	4795	0.047	1	99687851	C	T	0.192	No	
rs11855415	0	1	1	99692646	T	A	0.21	Yes	6.96×10^{-8}

Table B.2 Mammalian genome assemblies as used for multi species alignment of 14 eutherian mammalian genomes (see section 3.3.4)



mammal	*sequence length (bp)*	*genome assembly*
mouse	19384	mm10
rat	19195	rn5
cow	10933	bosTau8
horse	12898	equCab2
dog	11969	canFam3
cat	12822	felCat5
rabbit	12628	oryCun2
marmoset	12810	calJac3
rhesus	12460	rheMac3
baboon	12450	papAnu2
orangutan	12797	ponAbe2
chimp	12853	panTro4
human	12730	hg38
gorilla	12506	gorGor3

Appendix C Isoform Analysis

C.1 PCSK6-AS2/AS3 cDNA

The following sequences were acquired through Sanger sequencing using the primers indicated. See Section 4.3.2 for further details.

Forward: 5’-GGTGCAGAAAACAAGCCTG
Reverse: 5’-TGCCAAAAGAGTTATAGGTGATT

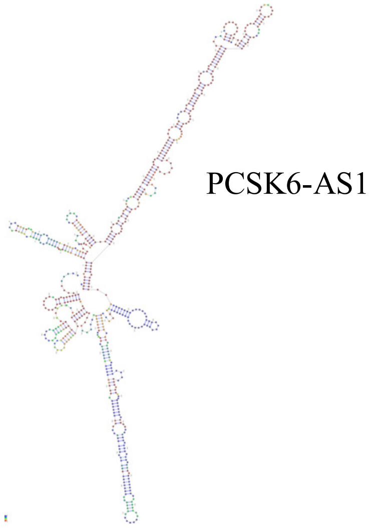
```
>PCSK6-AS2 accession LN713952.1
ccattcgaccattaagagtgtttcactccattggaagataaatggggaat
ctttacataaccgggggtacaacaagaagttgttctcacccccccggggat
cccacaggaagttgttctcaccgctgcaggtgcagaaaacaagcctggtg
aggaacctctgactctcctcagctccttagggtccagttacggccacatt
ccgaccacaaaggaatccgagcactttaaccaccaagtggtgcactgaga
ttggctggggttgtgatgacgatactcatgacagcctatgagggggccagg
cactgagctaacaacctgcggagctgagagctgggagctccaaaaacgcc
agcagggaagaagcagtgggaccaaagcaacccctccctgcatgtgcctc
caaaagagacctttccttttctaatagatggtgtctcgctctgttgcccg
gctggagtgcagtggcaccatctcagctcactgcaagctccgcctcctgg
gttcacgccattctcctgcctcagcctcctgagtagctgggactacaggc
gcccgccaccatgcccggctaattttttgtatttttagtagagatggggtt
tcagaatgtttgccaggatggtcttggtctcttgaccttgtgatccgcgt
gcctcccaaagtgctgggattacaggcatgagccactgcacctggcctat
ctcccctttctagtacttaaatgctttttttcactttctcaaccaagggag
tcactttggttttcctgcctttggaagacgtaaaaatgagaattccatac
ctatggcataaagtgtatggcataaatttgaagagtgattctttttttaaa
attacttttccctagttagaataaaaattattaaatgttgaagatttta
aagggaaa
```

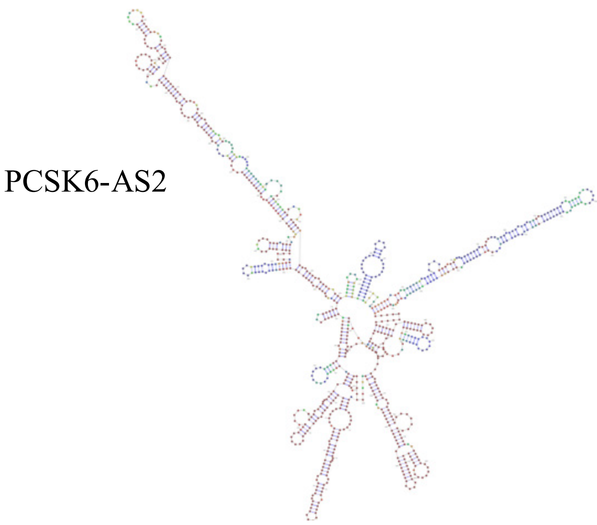
```
>PCSK6-AS3 accession LN713953.1
aaggaaacccgaacactttttaccaccaagtggtgcactgagattggctgg
ggttgtgatgacgatggtgtctcgctctgttgcccggctggagtgcagtg
gcaccatctcagctcactgcaagctccgcctcctgggttcacgccattct
cctgcctcagcctcctgagtagctgggactacaggcgcccgccaccatgc
ccggctaattttttgtatttttagtagagatggggtttcagaatgtttgcc
aggatggtcttggtctcttgaccttgtgatccgcgtgcctcccaaagtgc
tgggattacaggcatgagccactgcacctggcctatctcccctttctagt
acttaaatgctttttttcactttctcaaccaagggagtcactttggttttc
ctgcctttggaagacgtaaaaatgagaattccatacctatggcataaagt
gtatggcataaatttgaacaatgattctttttttaaaaatttttttccta
attaaaataaaaattattaaat
```

C.2 PCSK6-AS1 vs PCSK6-AS2

Table C.1 Annotation of both PCSK6-AS isoforms according to the AnnoLnc serviceTargetScan (Agarwal *et al*., 2015) used to predict miRNA-binding sites. For TF binding annotation 498 ChIP-Seq datasets covering 159 transcription factors (TFs) in 45 cell lines from the ENCODE project were used. For CLIP-Seq annotation 112 CLIP-Seq datasets covering 51 RNA binding proteins (RBPs) were collected and cross-linking sites calculated (P-values indicated). The differing secondary structure for PCSK6-AS1 and PCSK6-AS2 are indicated below. Structures were predicted using the RNAfold algorithm in the Vienna Package (Lorenz *et al*., 2011).

	PCSK6-AS1 (639bp)	PCSK6-AS2 (871bp)
location (hg19)	chr15:101874642-101877633 Exon 1:101874642-101874718 Exon 2:101876983-101877142 Exon 3:101877232-101877633	chr15:101874642-101877744 Exon 1:101874642-101874840 Exon 2:101876983-101877142 Exon 3:101877232-101877744
miRNA families	miR-148ab-3p/152, miR-146ac/146b-5p, miR-133abc, miR-146ac/146b-5p, miR-9/9ab	miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac, miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d, miR-148ab-3p/152, miR-216a, miR-146ac/146b-5p, miR-148ab-3p/152, miR-1ab/206/613, miR-133abc, miR-146ac/146b-5p, miR-9/9ab
TFs (cell type)	JunB (K562), Pol2 (HCT-116, K562) TRIM28 (K562), ATF1 (K562), ELF1 (GM12878, HepG2, K562) and FOXP2 (PFSK-1, SK-N-MC)	JunB (K562), Pol2 (HCT-116, K562) TRIM28 (K562), ATF1 (K562), ELF1 (GM12878, HepG2, K562) and FOXP2 (PFSK-1, SK-N-MC)
CLIP-Seq	None	FUS (P=1.943158e-45, HEK293), ELAVL (P=1.951124e-11, HEK293)





C.3 PCSK6 SI cDNA

Below is the cDNA sequence acquired through Sanger sequencing and submitted (accession #LN714797). Matching bases in coding regions of cDNA and genomic sequences are coloured blue and capitalized. Matching bases in UTR regions of cDNA and genomic sequences are coloured red and capitalised. Light blue (coding) or orange (UTR) bases mark the boundaries of gaps in either sequence (often splice sites).

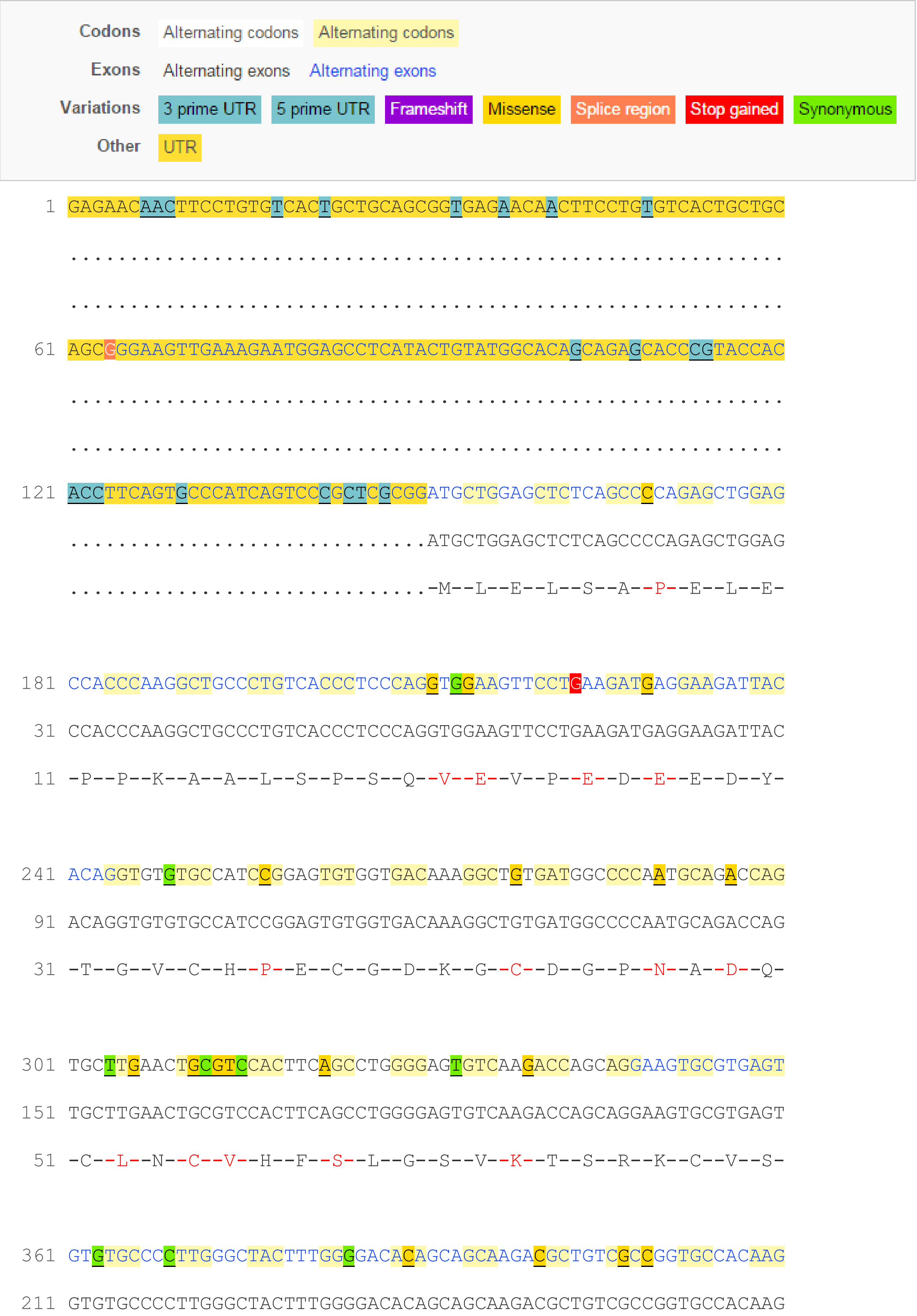
```
GAACAACTTC CTGTGTCACT GCTGCAGCGG GGAAGTTGAA AGAATGGAGC   50
CTCATACTGT ATGGCACAGC AGAGCACCCG TACCACACCT TCAGTGCCCA   100
TCAGTCCCGC TCGCGGATGC TGGAGCTCTC AGCCCCAGAG CTGGAGCCAC   150
CCAAGGCTGC CCTGTCACCC TCCCAGGTGG AAGTTCCTGA AGATGAGGAA   200
GATTACACAG GTGTGTGCCA TCCGGAGTGT GGTGACAAAG GCTGTGATGG   250
CCCCAATGCA GACCAGTGCT TGAACTGCGT CCACTTCAGC CTGGGGAGTG   300
TCAAGACCAG CAGGAAGTGC GTGAGTGTGT GCCCCTTGGG CTACTTTGGG   350
GACACAGCAG CAAGACGCTG TCGCCGGTGC CACAAGGGGT GTGAGACCTG   400
CTCCAGCAGA GCTGCGACGC AGTGCCTGTC TTGCCGCCGC GGGTTCTATC   450
ACCACCAGGA GATGAACACC TGTGTGACCC TCTGTCCTGC AGGATTTTAT   500
GCTGATGAAA GTCAGAAAAA TTGCCTTAAA TGCCACCCAA GCTGTAAAAA   550
GTGCGTGGAT GAACCTGAGA AATGTACTGT CTGTAAAGAA GGATTCAGCC   600
TTGCACGGGG CAGCTGCATT CCTGACTGTG AGCCAGGCAC CTACTTTGAC   650
TCAGAGCTGA TCAGATGTGG GGAATGCCAT CACACCTGCG GAACCTGCGT   700
GGGGCCAGGC AGAGAAGAGT GCATTCACTG TGCGAAAAAC TTCCACTTCC   750
ACGACTGGAA GTGTGTGCCA GCCTGTGGTG AGGGCTTCTA CCCAGAAGAG   800
ATGCCGGGCT TGCCCCACAA AGTGTGTCGA AGGTGTGACG AGAACTGCTT   850
GAGCTGTGCA GGCTCCAGCA GGAACTGTAG CAGGTGTAAG ACGGGCTTCA   900
CACAGCTGGG GACCTCCTGC ATCACCAACC ACACGTGCAG CAACGCTGAC   950
GAGACATTCT GCGAGATGGT GAAGTCCAAC CGGCTGTGCG AACGGAAGCT   1000
CTTCATTCAG TTCTGCTGCC GCACGTGCCT CCTGGCCGGG TAAGGGTGCC   1050
TAGCTGCCCA CAGAGGGCAG GCACTCCCAT CCATCCATCC GTCCACCTTC   1100
CTCCAGACTG TCGGCCAGAG TCTGTTTCAG GAGCGGCGCC CTGCACCTGA   1150
CAGCTTTATC TCCCCAGGAG CAGCAT
```

C.4 PCSK6 protein sequences

Protein sequences used for PCSK6 canonical Vs PCSK6 SI prediction (see 4.4.3)

```
>Canonical PCSK6 969aa Isoform PACE4A-I (identifier: P29122-1)
MPPRAPPAPGPRPPPRAAAATDTAAGAGGAGGAGGAGGPGFRPLAPRPWRWLLLLALPAA
CSAPPPRPVYTNHWAVQVLGGPAEADRVAAAHGYLNLGQIGNLEDYYHFYHSKTFKRSTL
SSRGPHTFLRMDPQVKWLQQQEVKRRVKRQVRSDPQALYFNDPIWSNMWYLHCGDKNSRC
RSEMNVQAAWKRGYTGKNVVVTILDDGIERNHPDLAPNYDSYASYDVNGNDYDPSPRYDA
SNENKHGTRCAGEVAASANNSYCIVGIAYNAKIGGIRMLDGDVTDVVEAKSLGIRPNYID
IYSASWGPDDDGKTVDGPGRLAKQAFEYGIKKGRQGLGSIFVWASGNGGREGDYCSCDGY
TNSIYTISVSSATENGYKPWYLEECASTLATTYSSGAFYERKIVTTDLRQRCTDGHTGTS
VSAPMVAGIIALALEANSQLTWRDVQHLLVKTSRPAHLKASDWKVNGAGHKVSHFYGFGL
VDAEALVVEAKKWTAVPSQHMCVAASDKRPRSIPLVQVLRTTALTSACAEHSDQRVVYLE
HVVVRTSISHPRRGDLQIYLVSPSGTKSQLLAKRLLDLSNEGFTNWEFMTVHCWGEKAEG
QWTLEIQDLPSQVRNPEKQGKLKEWSLILYGTAEHPYHTFSAHQSRSRMLELSAPELEPP
KAALSPSQVEVPEDEEDYTAQSTPGSANILQTSVCHPECGDKGCDGPNADQCLNCVHFSL
GSVKTSRKCVSVCPLGYFGDTAARRCRRCHKGCETCSSRAATQCLSCRRGFYHHQEMNTC
VTLCPAGFYADESQKNCLKCHPSCKKCVDEPEKCTVCKEGFSLARGSCIPDCEPGTYFDS
ELIRCGECHHTCGTCVGPGREECIHCAKNFHFHDWKCVPACGEGFYPEEMPGLPHKVCRR
CDENCLSCAGSSRNCSRCKTGFTQLGTSCITNHTCSNADETFCEMVKSNRLCERKLFIQF
CCRTCLLAG
```

Figure C.1 cDNA sequence of the PCSK6 SI(accession #LN714797) overlaid with codon/protein prediction according to Ensembl (Transcript: PCSK6-020, ENST00000632686). Top row represents the queried sequence, middle row reference sequence (hg19) and bottom row amino acid predicted.



```
      Codons   Alternating codons  Alternating codons
       Exons   Alternating exons  Alternating exons
  Variations   3 prime UTR  5 prime UTR  Frameshift  Missense  Splice region  Stop gained  Synonymous
       Other   UTR

  1 GAGAACAACTTCCTGTGTCACTGCTGCAGCGGTGAGAACAACTTCCTGTGTCACTGCTGC

    ............................................................

    ............................................................

 61 AGCGGGAAGTTGAAAGAATGGAGCCTCATACTGTATGGCACAGCAGAGCACCCGTACCAC

    ............................................................

    ............................................................

121 ACCTTCAGTGCCCATCAGTCCCGCTCGCGGATGCTGGAGCTCTCAGCCCCAGAGCTGGAG

    ..............................ATGCTGGAGCTCTCAGCCCCAGAGCTGGAG

    ..............................-M--L--E--L--S--A--P--E--L--E-


181 CCACCCAAGGCTGCCCTGTCACCCTCCCAGGTGGAAGTTCCTGAAGATGAGGAAGATTAC

 31 CCACCCAAGGCTGCCCTGTCACCCTCCCAGGTGGAAGTTCCTGAAGATGAGGAAGATTAC

 11 -P--P--K--A--A--L--S--P--S--Q--V--E--V--P--E--D--E--E--D--Y-


241 ACAGGTGTGTGCCATCCGGAGTGTGGTGACAAAGGCTGTGATGGCCCCAATGCAGACCAG

 91 ACAGGTGTGTGCCATCCGGAGTGTGGTGACAAAGGCTGTGATGGCCCCAATGCAGACCAG

 31 -T--G--V--C--H--P--E--C--G--D--K--G--C--D--G--P--N--A--D--Q-


301 TGCTTGAACTGCGTCCACTTCAGCCTGGGGAGTGTCAAGACCAGCAGGAAGTGCGTGAGT

151 TGCTTGAACTGCGTCCACTTCAGCCTGGGGAGTGTCAAGACCAGCAGGAAGTGCGTGAGT

 51 -C--L--N--C--V--H--F--S--L--G--S--V--K--T--S--R--K--C--V--S-


361 GTGTGCCCCTTGGGCTACTTTGGGGACACAGCAGCAAGACGCTGTCGCCGGTGCCACAAG

211 GTGTGCCCCTTGGGCTACTTTGGGGACACAGCAGCAAGACGCTGTCGCCGGTGCCACAAG
```

```
 71 -V--C--P--L--G--Y--F--G--D--T--A--A--R--R--C--R--R--C--H--K-

421 GGGTGTGAGACCTGCTCCAGCAGAGCTGCGACGCAGTGCCTGTCTTGCCGCCGCGGGTTC
271 GGGTGTGAGACCTGCTCCAGCAGAGCTGCGACGCAGTGCCTGTCTTGCCGCCGCGGGTTC
 91 -G--C--E--T--C--S--S--R--A--A--T--Q--C--L--S--C--R--R--G--F-

481 TATCACCACCAGGAGATGAACACCTGTGTGACCCTCTGTCCTGCAGGATTTTATGCTGAT
331 TATCACCACCAGGAGATGAACACCTGTGTGACCCTCTGTCCTGCAGGATTTTATGCTGAT
111 -Y--H--H--Q--E--M--N--T--C--V--T--L--C--P--A--G--F--Y--A--D-

541 GAAAGTCAGAAAAATTGCCTTAAATGCCACCCAAGCTGTAAAAAGTGCGTGGATGAACCT
391 GAAAGTCAGAAAAATTGCCTTAAATGCCACCCAAGCTGTAAAAAGTGCGTGGATGAACCT
131 -E--S--Q--K--N--C--L--K--C--H--P--S--C--K--K--C--V--D--E--P-

601 GAGAAATGTACTGTCTGTAAAGAAGGATTCAGCCTTGCACGGGGCAGCTGCATTCCTGAC
451 GAGAAATGTACTGTCTGTAAAGAAGGATTCAGCCTTGCACGGGGCAGCTGCATTCCTGAC
151 -E--K--C--T--V--C--K--E--G--F--S--L--A--R--G--S--C--I--P--D-

661 TGTGAGCCAGGCACCTACTTTGACTCAGAGCTGATCAGATGTGGGGAATGCCATCACACC
511 TGTGAGCCAGGCACCTACTTTGACTCAGAGCTGATCAGATGTGGGGAATGCCATCACACC
171 -C--E--P--G--T--Y--F--D--S--E--L--I--R--C--G--E--C--H--H--T-

721 TGCGGAACCTGCGTGGGGCCAGGCAGAGAAGAGTGCATTCACTGTGCGAAAAACTTCCAC
571 TGCGGAACCTGCGTGGGGCCAGGCAGAGAAGAGTGCATTCACTGTGCGAAAAACTTCCAC
191 -C--G--T--C--V--G--P--G--R--E--E--C--I--H--C--A--K--N--F--H-

781 TTCCACGACTGGAAGTGTGTGCCAGCCTGTGGTGAGGGCTTCTACCCAGAAGAGATGCCG
631 TTCCACGACTGGAAGTGTGTGCCAGCCTGTGGTGAGGGCTTCTACCCAGAAGAGATGCCG
211 -F--H--D--W--K--C--V--P--A--C--G--E--G--F--Y--P--E--E--M--P-
```

```
 841 GGCTTGCCCCACAAAGTGTGTCGAAGGTGTGACGAGAACTGCTTGAGCTGTGCAGGCTCC

 691 GGCTTGCCCCACAAAGTGTGTCGAAGGTGTGACGAGAACTGCTTGAGCTGTGCAGGCTCC

 231 -G--L--P--H--K--V--C--R--R--C--D--E--N--C--L--S--C--A--G--S-


901 AGCAGGAACTGTAGCAGGTGTAAGACGGGCTTCACACAGCTGGGGACCTCCTGCATCACC

 751 AGCAGGAACTGTAGCAGGTGTAAGACGGGCTTCACACAGCTGGGGACCTCCTGCATCACC

 251 -S--R--N--C--S--R--C--K--T--G--F--T--Q--L--G--T--S--C--I--T-


 961 AACCACACGTGCAGCAACGCTGACGAGACATTCTGCGAGATGGTGAAGTCCAACCGGCTG

 811 AACCACACGTGCAGCAACGCTGACGAGACATTCTGCGAGATGGTGAAGTCCAACCGGCTG

 271 -N--H--T--C--S--N--A--D--E--T--F--C--E--M--V--K--S--N--R--L-


1021 TGCGAACGGAAGCTCTTCATTCAGTTCTGCTGCCGCACGTGCCTCCTGGCCGGGTAAGGG

 871 TGCGAACGGAAGCTCTTCATTCAGTTCTGCTGCCGCACGTGCCTCCTGGCCGGGTAA...

 291 -C--E--R--K--L--F--I--Q--F--C--C--R--T--C--L--L--A--G--*-...


1081 TGCCTAGCTGCCCACAGAGGGCAGGCACTCCCATCCATCCATCCGTCCACCTTCCTCCAG

     ............................................................

     ............................................................


1141 ACTGTCGGCCAGAGTCTGTTTCAGGAGCGGCGCCCTGCACCTGACAGCTTTATCTCCCCA

     ............................................................

     ............................................................


1201 GGAGCAGCAT

     ..........

     ..........
```

Appendix D EMSA

D.1 Oligonucleotide Sequences

(1) 5’-BIO-AGACTTCATTATTATTGGACT

(2) 5’-AGTCCAATAAT**T**AATGAAGTCT

(3) 5’-BIO-AGACTTCATTTTTATTGGACT

(4) 5’-AGTCCAATAA**A**AATGAAGTCT

(5) 5’-AGACTTCATTATTATTGGACT

(6) 5’-AGACTTCATTTTTATTGGACT

(7) 5’-GCCTGTCACCCGTCATGTAT

(8) 5’-ATACATGACGGGTGACAGGC

(9) 5’-BIO-AAGCTGGCCCCGCTGGAAGG

(10) 5’-BIO-AAGCTGGCCCTGCTGGAAGG

(11) 5’-CCTTCCAGC**A**GGGCCAGCTT

(12) 5’-CCTTCCAGC**G**GGGCCAGCTT

(13) 5’-AAGCTGGCCCTGCTGGAAGG

(14) 5’-AAGCTGGCCCCGCTGGAAGG

All double-stranded EMSA probes are combinations of the above, allele values (in bold) refer to the sense strand (-) hg19.

rs11855415 biotinylated probes = T: (1)+(2), A: (3)+(4)
rs11855415 cold probes = T: (2)+(5), A: (4)+(6)
rs7182874 biotinylated probes = G: (9)+(12), A: (10)+(11)
rs7182874 cold probes = G: (12)+(14), A: (11)+(13)
Scrambled probe: (7)+(8)

D.2 Additional EMSA images

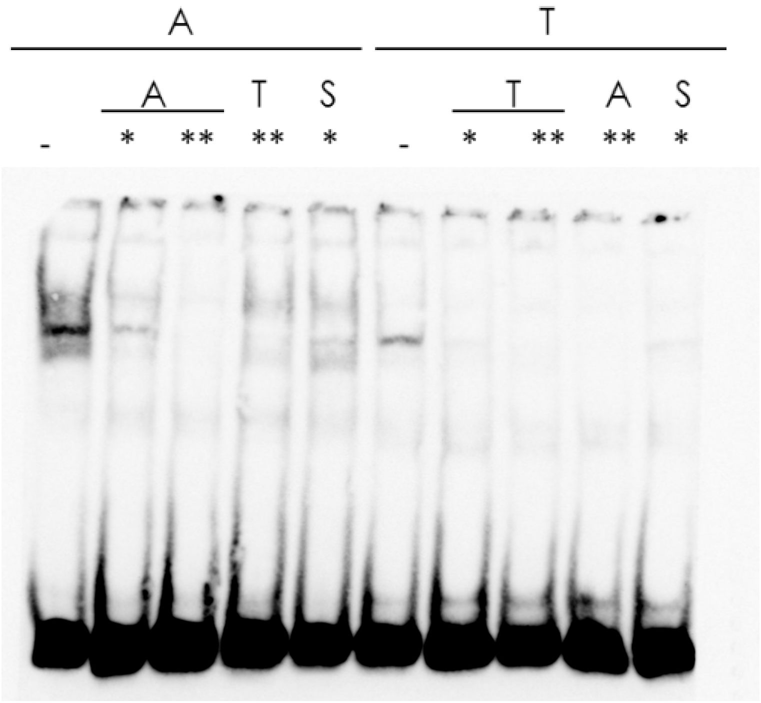


Figure D.1 EMSA for SH-SY5Y nuclear extract and the rs11855415 alleles. The gel image displays the binding of SH-SY5Y nuclear extract to probes containing the rs11855415 SNP A versus T alleles. Midway band indicates the protein:DNA complex with the unbound DNA probe at the bottom of each lane. The presence of a competitor is denoted above each lane: -, no competitor; S, scrambled competitor; and *, 10-fold and **, 100-fold excess of competitor respectively

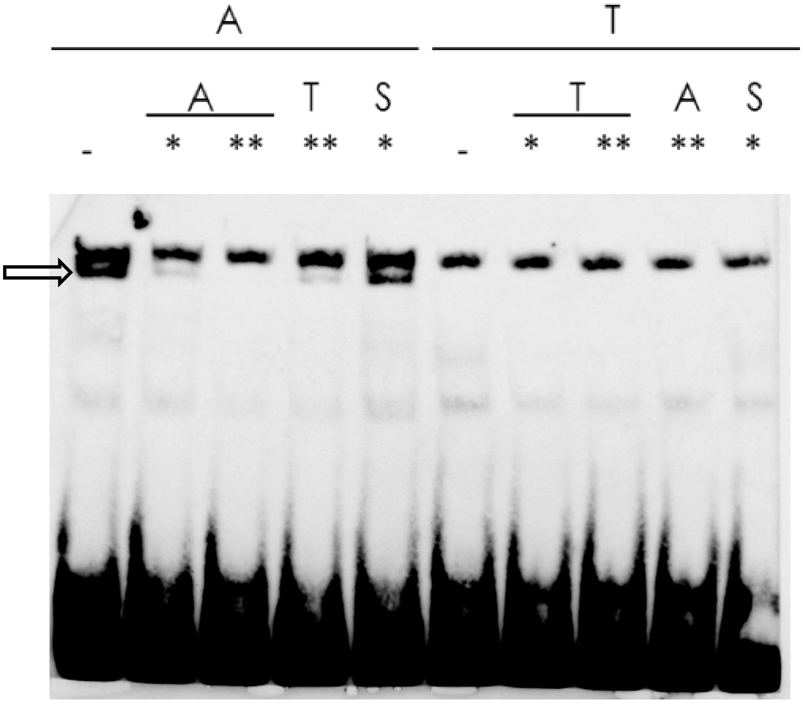


Figure D.2 EMSA for HEK293 nuclear extract and the rs11855415 alleles. The gel image displays the binding of HEK293 nuclear extract to probes containing the rs11855415 SNP A versus T alleles. Arrow indicates the protein:DNA complex with the unbound DNA probe at the bottom of each lane. The presence of a competitor is denoted above each lane: -, no competitor; S, scrambled competitor; and *, 10-fold and **, 100-fold excess of competitor respectively

Appendix E Luciferase Assay

E.1 Lipofectamine P3000 transfection protocol

1 µl of the Renilla vector pRL-TK (20ng/µl) was added to 1 µl of DNA plasmid (80ng/µl), 0.2 µl of P3000 reagent and 2.8 µl of Opti-MEM (Invitrogen) to give a 5 µl volume which was allowed to incubate for up to 10 minutes. To this, 0.3 µl Lipofectamine 3000 and 4.7 µl Opti-MEM were added to give a total volume of 10 µl which was then added to each well. Note for the GFP positive control 100ng was added to each well instead of Renilla and DNA.

E.2 Luciferase construct table

Table E.1 Luciferase assay reporter plasmids. The 6 pairs of pGL4.10 luciferase plasmids were designed to include the 1707-1839bp PCSK6 secondary promoter region containing rs11855415 SNP A and T alleles and VNTR 6,9 and 10 alleles in both sense and antisense strand directions

VNTR	*rs11855415*	*Strand Direction*	*Size (bp)*
6x33bp	A	Sense Antisense	1707
6x33bp	T	Sense Antisense	1707
9x33bp	A	Sense Antisense	1806
9x33bp	T	Sense Antisense	1806
10x33bp	A	Sense Antisense	1839
10x33bp	T	Sense Antisense	1839

E.3 Luciferase assay plasmids

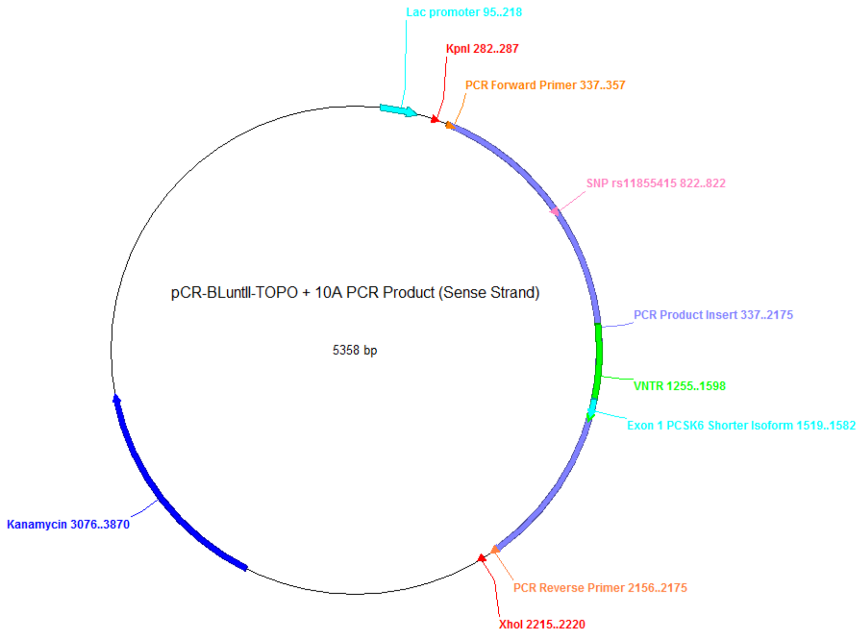

Lac promoter 95..218
KpnI 282..287
PCR Forward Primer 337..357
SNP rs11855415 822..822
pCR-BLuntII-TOPO + 10A PCR Product (Sense Strand)
5358 bp
PCR Product Insert 337..2175
VNTR 1255..1598
Exon 1 PCSK6 Shorter Isoform 1519..1582
Kanamycin 3076..3870
PCR Reverse Primer 2156..2175
XhoI 2215..2220

Figure E.1 Plasmid map of the pCR-BluntII-TOPO entry vectorcomplete with inserted PCR product containing the VNTR 10 and rs11855415 A alleles. The PCR product here has been inserted in a direction corresponding to the sense strand direction. This was the entry-level vector used for all future cloning. For primers used to produce PCR product insert see Appendix A.

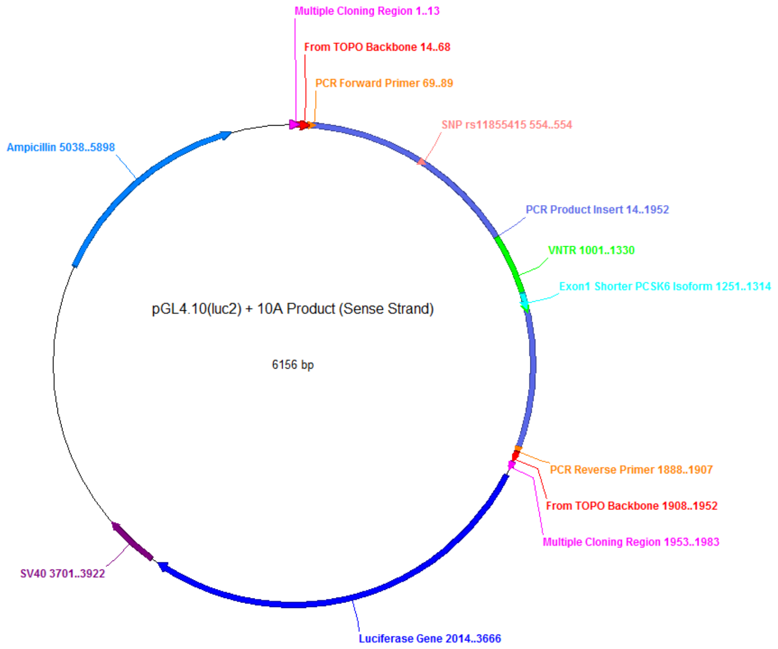

Multiple Cloning Region 1..13
From TOPO Backbone 14..68
PCR Forward Primer 69..89
SNP rs11855415 554..554
Ampicillin 5038..5898
PCR Product Insert 14..1952
VNTR 1001..1330
Exon1 Shorter PCSK6 Isoform 1251..1314
pGL4.10(luc2) + 10A Product (Sense Strand)
6156 bp
PCR Reverse Primer 1888..1907
From TOPO Backbone 1908..1952
Multiple Cloning Region 1953..1983
SV40 3701..3922
Luciferase Gene 2014..3666

Figure E.2 Plasmid map of the pGL4.10 luciferase reporter vectorcomplete with inserted PCR product containing the VNTR 6 and rs11855415 A alleles. The PCR product in Figure X was excised from the pCR-BluntII-TOPO vector by KpnI and XhoI restriction enzymes and cloned in to this pGL4.10 luciferase reporter plasmid (Promega). In this plasmid the PCR product was inserted in a direction corresponding to the sense strand direction.

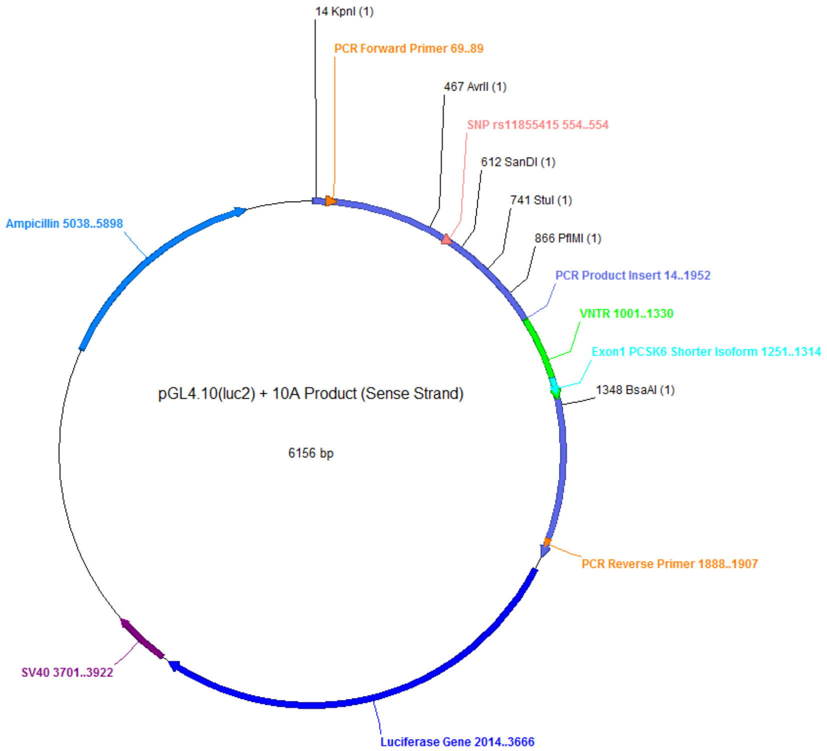

14 KpnI (1)
PCR Forward Primer 69..89
467 AvrII (1)
SNP rs11855415 554..554
612 SanDI (1)
741 StuI (1)
866 PflMI (1)
PCR Product Insert 14..1952
VNTR 1001..1330
Exon1 PCSK6 Shorter Isoform 1251..1314
1348 BsaAI (1)
Ampicillin 5038..5898
pGL4.10(luc2) + 10A Product (Sense Strand)
6156 bp
PCR Reverse Primer 1888..1907
SV40 3701..3922
Luciferase Gene 2014..3666

Figure E.3 Plasmid map of the pGL4.10 luciferase reporter 10A vectorcomplete with inserted PCR product containing the VNTR 10 and rs11855415 A alleles. The 10A pGL4.10 vector was used as the template vector from which increasingly-sized segments were removed to define the minimal promoter capable of driving transcription in a sense strand direction (see Methods section 5.3.5). The relevant restriction enzyme sites are noted.

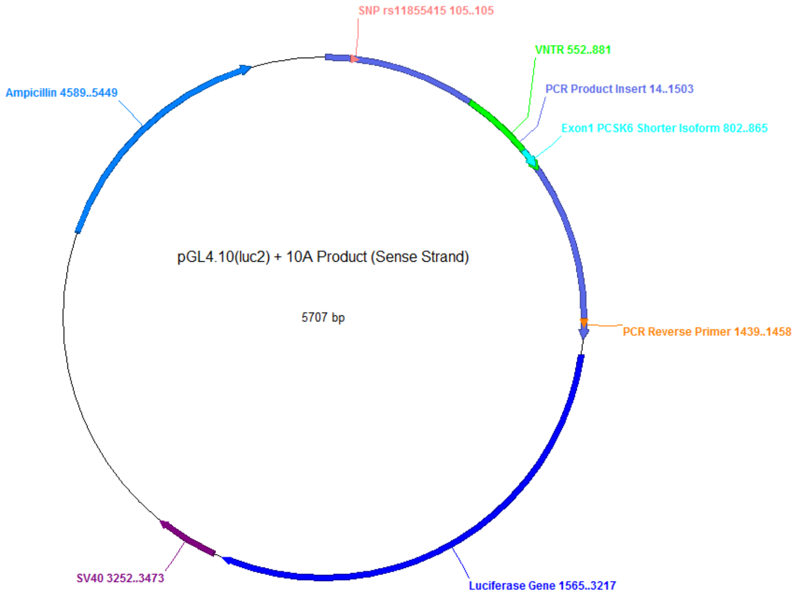

SNP rs11855415 105..105
VNTR 552..881
Ampicillin 4589..5449
PCR Product Insert 14..1503
Exon1 PCSK6 Shorter Isoform 802..865
pGL4.10(luc2) + 10A Product (Sense Strand)
5707 bp
PCR Reverse Primer 1439..1458
SV40 3252..3473
Luciferase Gene 1565..3217

Figure E.4 Minimal Promoter construct used for allelic variation assays.This plasmid map represents the 3rd bar from Figure 5.12. Following the minimal promoter assay, this was the plasmid with the shortest sequence that would include the rs11855415 SNP. The plasmid was subsequently used when disrupting the predicted transcription factor bind site centred at allele A of the SNP rs11855415 (Figure 5.14).

E.4 Luciferase assay antisense strand results

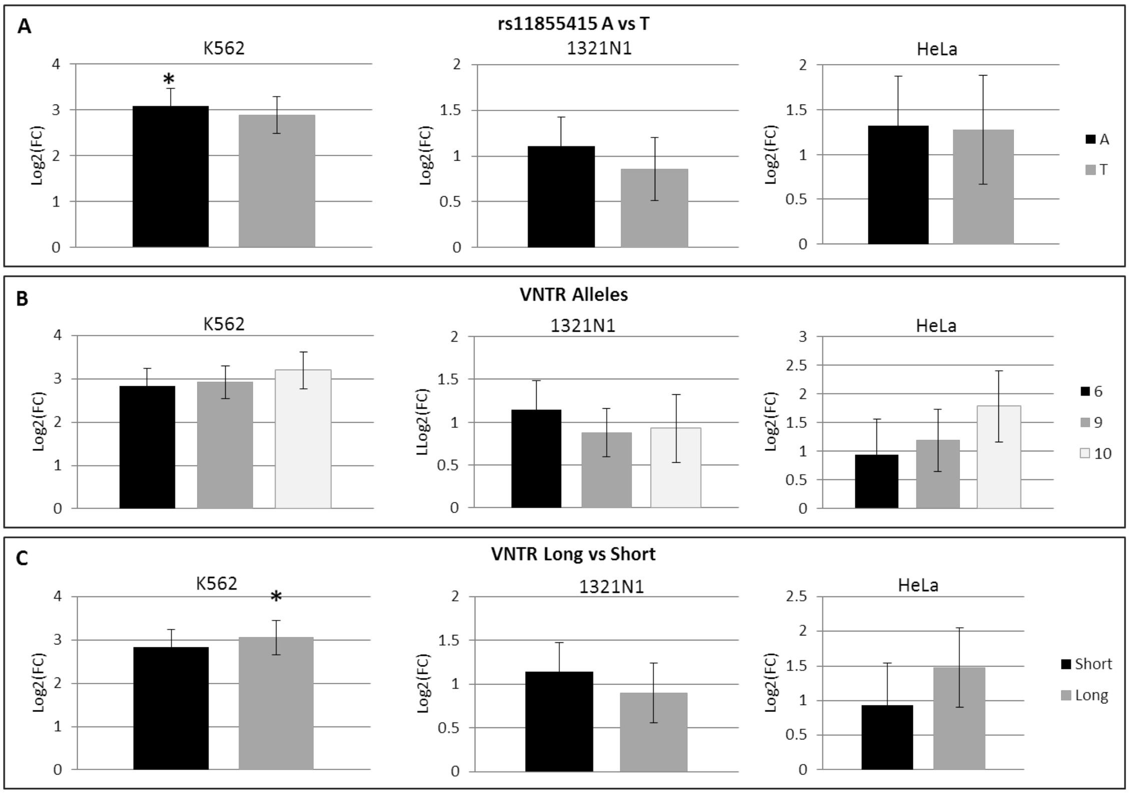


Figure E.5 Antisense luciferase-expressing construct results which tested the effect of allelic variation for (A) the rs11855415 SNP and (B,C) the VNTR in the antisense strand direction. Allelic differences in promoter activity were measured with luciferase constructs transfected in K562, 1321N1 and HeLa cell lines for the (B) VNTR 6,9 and 10 alleles (C) VNTR alleles classified as short (allele 6) and long (alleles 9 and 10) according to the association analysis performed by Arning *et al.*, (2013). Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation. Data are representative of at least 3 independent experiments performed in triplicate. Luciferase expression was measured relative to the empty pGL4 vector following renilla normalisation and log transformed ($\log_2$ fold change). Bars represent Mean(SD) (N = 3). *P-value of less than 0.05 was considered significant.

E.5 Sequencing data for mutagenesis of rs11855415 TFBS

The sequence below confirms the rs11855415 T allele (red) and the surrounding sequence underlined which is to be replaced by ‘nonsense’ sequence via mutagenesis. The plasmid is based on the 10A AvrII (-783 bp) construct from the Minimal Promoter assay (see Figure 5.11) with site-directed mutagenesis used to disrupt the 5 base pairs centred at rs11855415 entirely. The accompanying chromatogram indicates the resulting Sanger sequencing.

```
>sanger sequencing of the 10A plasmid
CCCAGTCAGACATTTCTCTGGCTACTGGCCGCTAGGCAACAGAGTGAGACCCTAGGTCTAAATACACACA
TACACACATACACACACACACACATAAAGTCAAACAGAGAGTGTAATAAAAATGAAGTCTCTCTTCCATT
CCAGTCCAGCCTCCTCCCATGAGGCTGCCTGTCTGCTGGGTCCCTGGGAATTCTTCCTGCTCTCTTTTTC
CACACAGTGCACACCACTGACCCTGAGTGTGGAGCAGCCCCTCTTGTTTGTCCCACCAAAACTATATCTG
AAAATAAAAATCCAGTTTT
```

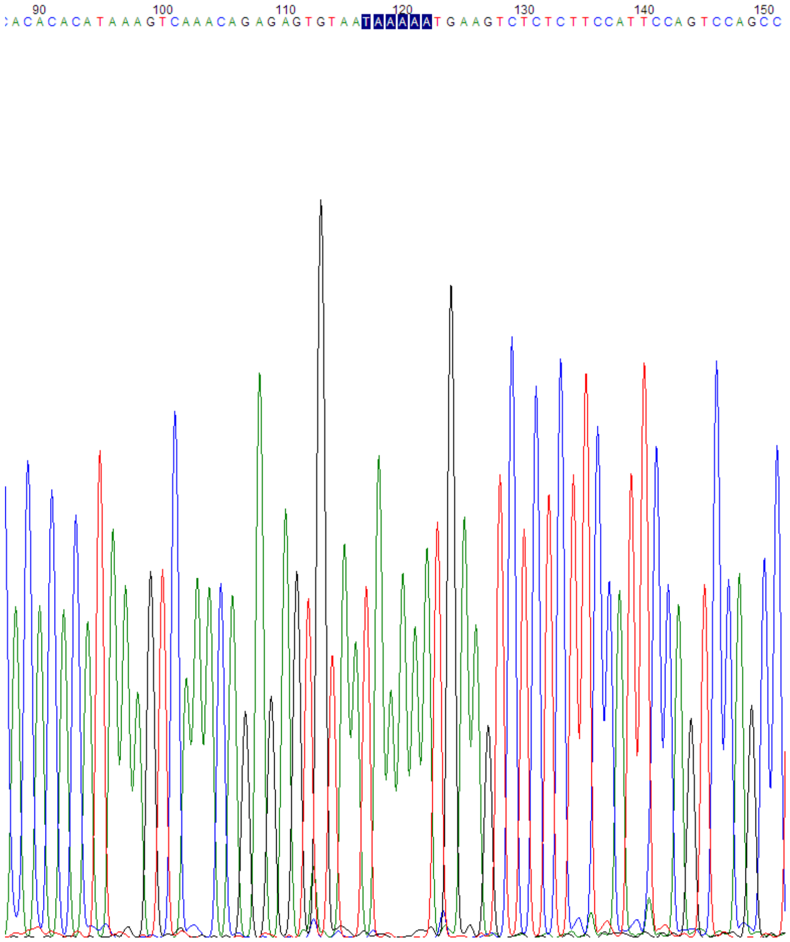


Figure E.6 Chromatogram confirming the sequence of the 10A plasmid as used in the rs11855415 TFBS mutagenesis luciferase assay experiment (see Section 5.4.5)

The sequence below confirms the rs11855415 SNP and the surrounding sequence has been replaced with 'nonsense' sequence (GCTAGC) via mutagenesis (see Appendix A.5 for primers used). The plasmid is based on the 10A AvrII (-783 bp) construct from the Minimal Promoter assay. The accompanying chromatogram indicates the resulting Sanger sequencing. Note Transfac v2014.4 did not return any known TFBSs for the 10 base pair sequence surrounding the inserted 'GCTAGC' sequence using the 'vertebrate non redundant min FP profile.

```
>sanger sequencing of the 10A plasmid post-mutagenesis
CCCGGTCATACATTTCTCTGGCTAACTGGCCGCTAGGCAACAGAGTGAGACCCTGTCTAAATACACACAT
ACACACATACACACACACACACATAAAGTCAAACAGAGAGTGTAAGCTAGCTGAAGTCTCTCTTCCATTC
CAGTCCAGCCTCCTCCCATGAGGCTGCCTGTCTGCTGGGTCCCTGGGAATTCTTCCTGCTCTCTTTTTCC
ACACAGTGCACACCACTGACCCTGAGTGTGGAGCA
```

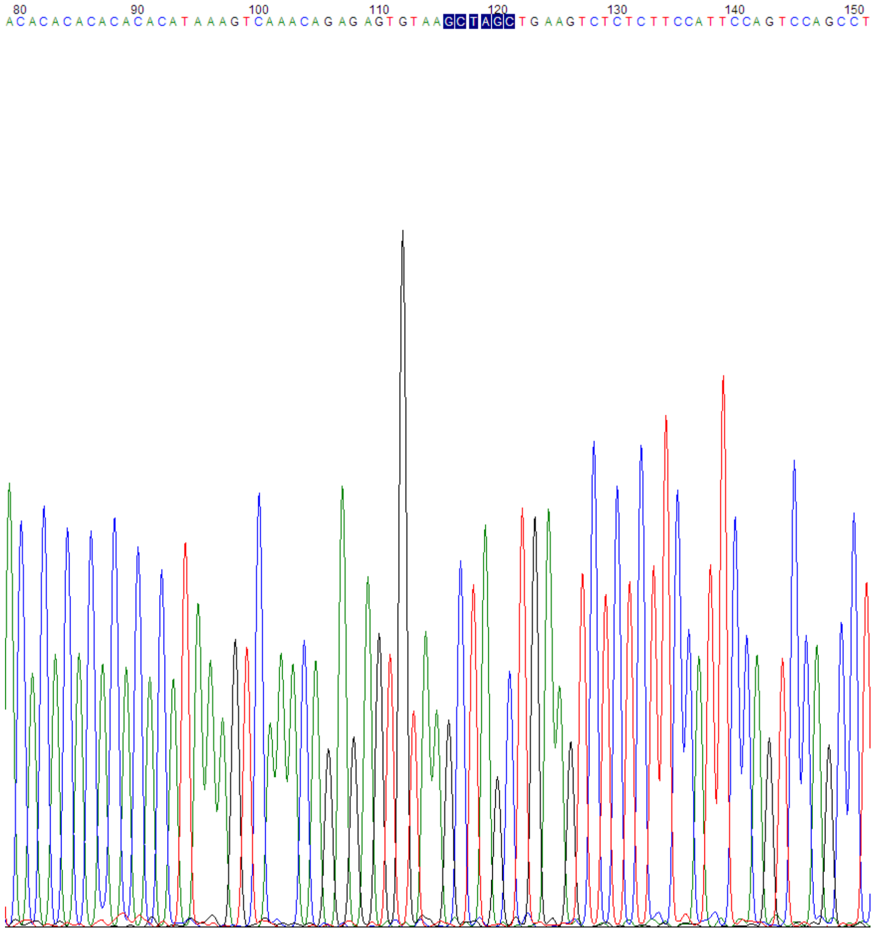


Figure E.7 Chromatogram confirming the sequence of the 10A plasmid following mutagenesis as used in the rs11855415 TFBS mutagenesis luciferase assay experiment (see Section 5.4.5)

Appendix F RNA-Seq

F.1 Source and shell script developed to download RNA-Seq files

```
#
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCa
ltechRnaSeq/
$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCa
ltechRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCa
ltechRnaSeq/$i; done;


$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCs
hlLongRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCs
hlLongRnaSeq/$i; done;


$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCs
hlShortRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCs
hlShortRnaSeq/$i; done;


$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGi
sRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGi
sRnaSeq/$i; done;


$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHa
ibRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
```

```
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHa
ibRnaSeq/$i; done;


$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSy
dhRnaSeq/
$ grep ".bam" index.html | grep -v ".bai" | awk -F "\"" '{ print $2
}' > names.txt
$ for i in `cat names.txt`; do echo $i; wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSy
dhRnaSeq/$i; done;
```

F.2 Script for creating BAM files

```
echo $1
module add samtools
module add bwa

if [ -e $1.bam ]
then
        echo "$1.bam exist"
        exit
fi

if [ -e $1.sam ]
then
        samtools view -Sbh $1.sam > $1.bam
        samtools sort -@ 5 $1.bam $1.sorted
        mv $1.sorted.bam $1.bam
        samtools index $1.bam
        rm $1.sam
else
        if [ -e $1.fastq.gz ]
        then
                bwa mem -t 10 ../ref/chr15.txt $1.fastq.gz > $1.sam
        else
                bwa mem -t 10 ../ref/chr15.txt $1.fastq.tgz >
$1.sam
        fi
        samtools view -Sbh $1.sam > $1.bam
        samtools sort -@ 5 $1.bam $1.sorted
        mv $1.sorted.bam $1.bam
        samtools index $1.bam
        rm $1.sam
fi
```

F.3 Script for alignment with SAMtools

```
echo $1
if [ ! -d $1 ]; then
        mkdir $1
else
        rm $1/*
fi


## 1
echo data_1
samtools view -b $1.bam CM000677.2:101300000-101490000 | samtools
view - | grep -w "CM000677.2" | awk ' { if ( $4 >= 101489374-
length($10) && $4 <= 101489984+length($10) ) { print $0 } } ' >
$1/dashat.txt

## 2
echo data_2
samtools view -b $1.bam CM000677.2:101300000-101490000 | samtools
view - | grep -w "CM000677.2" | awk ' { if ( $4 >= 101398404-
length($10) && $4 <= 101398576+length($10) ) { print $0 } } ' >
$1/dashat.txt

## 3b
echo data_3b
samtools view -b $1.bam CM000677.2:101300000-101490000 | samtools
view - | grep -w "CM000677.2" | awk ' { if ( $4 >= 101873769-
length($10) && $4 <= 101873844+length($10) ) { print $0 } } ' >
$1/dashat.txt

## 4
echo data_4
samtools view -b $1.bam CM000677.2:101300000-101490000 | samtools
view - | grep -w "CM000677.2" | awk ' { if ( $4 >= 101324850-
length($10) && $4 <= 101325046+length($10) ) { print $0 } } ' >
$1/dashat.txt

## 5
echo data_5
samtools view -b $1.bam CM000677.2:101300000-101490000 | samtools
view - | grep -w "CM000677.2" | awk ' { if ( $4 >= 101303928-
length($10) && $4 <= 101305355+length($10) ) { print $0 } } ' >
$1d/dashat.txt
```

Appendix G Rev-ChIP Supplemental Materials and Methods

The following biotinylated probes (MWG Eurofins) were used as probe baits in the Rev-ChIP assay (reverse complement sequences not shown)

rs11855415 T allele	5' - CGTAGAAAGTGTAATAATAATGAAGTCT
rs11855415 A allele	5' - CGTAGAAAGTGTAATAAAAATGAAGTCT

Mascot database searching

Charge state deconvolution and deisotoping were not performed. All MS/MS samples were analysed using Mascot (Matrix Science, London, UK; version 2.5.1). Mascot was set up to search the NCBInr_20150331 database (selected for Homo sapiens, unknown version, 304051 entries) assuming the digestion enzyme trypsin. Mascot was searched with a fragment ion mass tolerance of 0.100 Da and a parent ion tolerance of 20 PPM. O+18 of pyrrolysine and iodoacetamide derivative of cysteine were specified in Mascot as fixed modifications. Oxidation of methionine was specified in Mascot as a variable modification.

Criteria for protein identification

Scaffold (version Scaffold_4.4.3, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 99.5% probability to achieve an FDR less than 1.0% by the Peptide Prophet algorithm (Keller *et al*., 2002). Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii *et al*., 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins were annotated with GO terms from NCBI (downloaded May 16, 2015) (Ashburner *et al*., 2000).

Table G.1 Full mass spectrometry output for Rev-ChIP assay rs1185545 A vs T alleles using hNSC cell line nuclear extract

(1) Data is presented according to proteins binding to the A allele:

MS/MS Identified Proteins	*Accession Number*	*Mol. Weight*	*rs11855415 allele*		*Control**
			$A^{\dagger}$	*$T^{\dagger}$*	
unnamed protein product	gi|189069149 (+2)	34 kDa	5	0	0
microtubule-associated protein 1B, isoform CRA_a	gi|119616102 (+4)	257 kDa	3	0	0
SRY (sex determining region Y)-box 5, isoform CRA_d	gi|119616895 (+14)	64 kDa	3	0	0
histone H1x	gi|5174449	22 kDa	3	2	1
hCG2016250, isoform CRA_a	gi|119618532 (+9)	21 kDa	3	1	2
nucleophosmin isoform 1	gi|10835063 (+13)	33 kDa	2	1	1
myristoylated alanine-rich C-kinase substrate	gi|153070260 (+2)	32 kDa	3	2	1
hCG1640785, isoform CRA_a	gi|119569329 (+1)	14 kDa	6	2	1
ribosomal protein S10, isoform CRA_a	gi|119624187 (+2)	20 kDa	5	4	2
ubiquitin associated protein 2-like, isoform CRA_a	gi|119573598 (+22)	114 kDa	5	4	1
translation elongation factor 1 alpha 1-like 14	gi|15277711 (+14)	43 kDa	5	4	3
ceruloplasmin (ferroxidase), isoform CRA_b	gi|119599289 (+9)	123 kDa	5	4	1
transcription factor SOX-3	gi|30061556 (+1)	45 kDa	5	4	1
YTH domain family protein 2 isoform 1	gi|116812575 (+3)	62 kDa	6	5	1
filamin-A isoform 1	gi|116063573 (+13)	280 kDa	6	5	0
YTH domain family protein 3	gi|116235460 (+7)	64 kDa	8	5	1
unnamed protein product	gi|194376170 (+1)	26 kDa	8	6	0
unnamed protein product	gi|194387670 (+1)	19 kDa	8	6	0
MYL6 protein	gi|113812151 (+4)	16 kDa	14	8	4
musashi homolog 2 (Drosophila), isoform CRA_a	gi|119614912 (+11)	37 kDa	9	8	3
unnamed protein product	gi|158255914 (+3)	42 kDa	13	11	2
RNA-binding protein Musashi homolog 1	gi|4505255	39 kDa	16	11	4
histone H1.5	gi|4885381	23 kDa	22	14	3
liver histone H1e	gi|126035028 (+3)	22 kDa	27	15	4

PRO2619	gi|11493459 (+22)	57 kDa	19	17	7
keratin 10 isoform CRA_b	gi|119581085 (+4)	63 kDa	22	19	16
vimentin	gi|340219 (+1)	54 kDa	35	21	13
tropomyosin alpha-1 chain isoform 4	gi|63252900	33 kDa	27	21	6
actin, beta, partial	gi|14250401	41 kDa	49	41	15
myosin-10 isoform 2	gi|367460087 (+3)	229 kDa	83	63	19
poly [ADP-ribose] polymerase 1	gi|156523968 (+2)	113 kDa	108	81	3
myosin-9	gi|12667788	227 kDa	143	112	38

(2) Data is presented according to proteins binding to the T allele:

MS/MS Identified Proteins	*Accession Number*	*Mol. Weight*	*rs11855415 allele*		*Control**
			A†	*T*†	
signal recognition particle 14kDa (homologous Alu RNA binding protein), isoform CRA_b	gi|119612797 (+1)	17 kDa	0	2	0
ribosomal protein S16, isoform CRA_b	gi|119577297 (+2)	16 kDa	0	2	1
CUG triplet repeat, RNA binding protein 1, isoform CRA_e	gi|119588316 (+12)	31 kDa	0	2	1
hnRNP-E1	gi|460771 (+1)	38 kDa	0	6	4
poly(rC)-binding protein 2 isoform b	gi|14141166 (+5)	38 kDa	0	8	4
DAZ-associated protein 1 isoform b	gi|25470886 (+9)	43 kDa	2	4	3
unnamed protein product	gi|189066545 (+3)	29 kDa	2	5	2
A0=heterogeneous nuclear ribonucleoprotein [human, placenta, Peptide, 305 aa]	gi|1911429 (+1)	31 kDa	2	8	5
ribosomal protein L23a, isoform CRA_a	gi|119571516 (+3)	22 kDa	3	4	1
caldesmon 1, isoform CRA_a	gi|119604232 (+14)	94 kDa	3	7	1
RNA-binding protein 4 isoform 1	gi|93277122	40 kDa	5	9	2
Heterogeneous nuclear ribonucleoprotein U-like 1	gi|12803479 (+5)	96 kDa	6	8	5
replication protein A 70 kDa DNA-binding subunit	gi|4506583	68 kDa	6	14	5
myosin regulatory light chain 12B	gi|15809016 (+2)	20 kDa	8	11	3

tubulin beta-2B chain [Mus musculus]	gi|21746161	50 kDa	8	12	8
TDP43	gi|130750552 (+2)	45 kDa	8	12	5
ribosomal protein S18, isoform CRA_c	gi|119624101 (+1)	15 kDa	9	11	3
tubulin beta-5 chain [Mus musculus]	gi|7106439 (+1)	50 kDa	9	12	9
keratin, type II cytoskeletal 2 epidermal	gi|47132620	65 kDa	9	15	11
drebrin 1, isoform CRA_a	gi|119605395 (+5)	76 kDa	11	12	1
unnamed protein product	gi|194387362 (+3)	35 kDa	11	18	9
heterogeneous nuclear ribonucleoprotein D-like isoform a	gi|14110407 (+7)	46 kDa	13	18	9
heterogeneous nuclear ribonucleoprotein M isoform a	gi|14141152 (+1)	78 kDa	16	39	24
heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa), isoform CRA_f	gi|119626284	30 kDa	17	21	12

*Unrelated biotinylated molecule

† Indicates a quantative value for the normalised value of total spectra

Table G.2 Full mass spectrometry output for Rev-ChIP assay rs1185545 A vs T alleles using 1321N1 cell line nuclear extract.

(1) Data is presented according to preferential binding to the A allele

MS/MS Identified Proteins	*Accession Number*	*Mol. Weight*	*rs11855415 allele*		*Control**
			A†	*T*†	
microtubule-associated protein 4	gi|187383 (+31)	121 kDa	9	0	0
forkhead box C2 (MFH-1, mesenchyme forkhead 1), isoform CRA_b	gi|119615822 (+1)	49 kDa	7	0	0
heat shock-related 70 kDa protein 2	gi|13676857 (+6)	70 kDa	4	0	0
general transcription factor II, i, isoform CRA_a	gi|119590000 (+23)	42 kDa	3	0	1
actinin, alpha 4, isoform CRA_c	gi|119577215 (+7)	104 kDa	3	0	0
dermcidin preproprotein	gi|16751921 (+1)	11 kDa	3	0	0
unnamed protein product	gi|193786488 (+5)	83 kDa	3	0	0
transcriptional activator protein Pur-beta	gi|15147219	33 kDa	1	0	1
unnamed protein product	gi|158259911 (+11)	52 kDa	9	1	0
beta-polymerase	gi|190156 (+14)	38 kDa	7	1	0
zinc-finger homeodomain protein 4	gi|109638254 (+1)	397 kDa	7	1	0
SET translocation (myeloid leukemia-associated), isoform CRA_c	gi|119608226 (+7)	29 kDa	7	1	0
leucine-zipper protein FKSG13	gi|11034809 (+3)	43 kDa	6	1	2
60S ribosomal protein L13 isoform 1	gi|15431295 (+1)	24 kDa	3	1	3
60S ribosomal protein L14	gi|78000181	23 kDa	3	1	2
ribosomal protein L18, isoform CRA_b	gi|119572744 (+5)	19 kDa	3	1	2
unnamed protein product	gi|189065517 (+8)	80 kDa	3	1	1
A kinase (PRKA) anchor protein 8	gi|119604880 (+3)	76 kDa	3	1	1
protein transport protein Sec61 subunit beta	gi|5803165 (+1)	10 kDa	3	1	1
40S ribosomal protein S14	gi|5032051	16 kDa	3	1	1
transcription factor 7-like 2 (T-cell specific, HMG-box), isoform CRA_b	gi|119569892 (+62)	59 kDa	3	1	0
60S acidic ribosomal protein P2	gi|4506671 (+1)	12 kDa	3	1	3
60S ribosomal protein L11 [Mus musculus]	gi|13385408 (+4)	20 kDa	3	1	2
Ribosomal protein S6	gi|15342049 (+5)	29 kDa	3	1	2

ribosomal protein L9, isoform CRA_a	gi|119613332 (+2)	15 kDa	3	1	1
X-ray repair cross-complementing protein 5	gi|10863945 (+1)	83 kDa	15	3	0
polynucleotide kinase 3'-phosphatase, isoform CRA_c	gi|119572952 (+3)	65 kDa	13	3	0
lamin A/C, isoform CRA_a	gi|119573381 (+7)	78 kDa	7	3	1
chromodomain helicase DNA binding protein 4, isoform CRA_a	gi|119609183 (+11)	218 kDa	7	3	0
ras-related protein Rab-1B	gi|13569962 (+1)	22 kDa	4	3	2
unnamed protein product	gi|158259071 (+1)	40 kDa	4	3	2
hCG23783, isoform CRA_a	gi|119621875 (+5)	23 kDa	4	3	2
high mobility group AT-hook 1, isoform CRA_b	gi|119624168 (+5)	34 kDa	4	3	1
replication factor C (activator 1) 1, 145kDa, isoform CRA_a	gi|119613328 (+5)	115 kDa	4	3	0
ribosomal protein S5, isoform CRA_b	gi|119592989 (+3)	22 kDa	4	3	1
lactotransferrin	gi|119585171 (+31)	78 kDa	4	3	1
Chain A, Cyclophilin B Complexed With [d-(Cholinylester)ser8]-Cyclosporin	gi|1310882 (+5)	20 kDa	4	3	0
ligase III, DNA, ATP-dependent	gi|19550955 (+5)	96 kDa	22	4	0
ribosomal protein, partial	gi|337518 (+1)	22 kDa	10	4	2
unnamed protein product	gi|158255940 (+5)	61 kDa	7	4	1
DNA repair protein XRCC1	gi|190684675 (+3)	69 kDa	7	4	0
glyceraldehyde-3-phosphate dehydrogenase	gi|31645 (+4)	36 kDa	6	4	1
unnamed protein product	gi|193786502 (+3)	46 kDa	13	6	7
Cart-1	gi|1098654 (+1)	37 kDa	9	6	0
filamin-A isoform 1	gi|116063573 (+7)	280 kDa	9	6	1
unnamed protein product	gi|194387670 (+1)	19 kDa	7	6	0
N-methylpurine-DNA	gi|14336679 (+5)	33 kDa	10	7	0
histone 1, H1t	gi|119575933 (+4)	24 kDa	10	7	0
78 kDa glucose-regulated protein precursor	gi|16507237 (+2)	72 kDa	9	7	1
hCG1640785, isoform CRA_a	gi|119569329 (+1)	14 kDa	9	7	2
CCAAT/enhancer-binding protein beta	gi|28872796 (+2)	36 kDa	12	9	1
X-ray repair cross-complementing protein 6	gi|4503841 (+1)	70 kDa	22	10	1
extracellular matrix protein 1	gi|1488324 (+2)	61 kDa	13	10	4

YTH domain family protein 3	gi|116235460 (+6)	64 kDa	12	10	3
collagen, type VI, alpha 3, isoform CRA_c	gi|119591511 (+5)	321 kDa	21	12	2
keratin, type II cytoskeletal 5	gi|119395754 (+4)	62 kDa	10	12	4
ribosomal protein S18, isoform CRA_c	gi|119624101 (+1)	15 kDa	15	13	5
elongation factor 1-alpha 1	gi|4503471 (+8)	50 kDa	16	15	4
heterogeneous nuclear ribonucleoprotein U isoform b	gi|14141161 (+4)	89 kDa	22	17	16
histone H1.5	gi|4885381	23 kDa	25	20	6
cytokeratin 9	gi|435476 (+1)	62 kDa	37	22	20
liver histone H1e	gi|126035028 (+4)	22 kDa	37	25	8
PRO2619	gi|11493459 (+22)	57 kDa	27	25	11
actin, beta, partial	gi|14250401 (+9)	41 kDa	36	26	13
neuroblast differentiation-associated protein AHNAK isoform 1	gi|61743954 (+4)	629 kDa	39	28	5
poly [ADP-ribose] polymerase 1	gi|156523968 (+2)	113 kDa	156	134	10

(2) Data is presented according to preferential binding to the A allele

			rs11855415 allele		
MS/MS Identified Proteins	*Accession Number*	*Mol. Weight*	*A*†	*T*†	*Control**
Chain A, Core Of The Alu Domain Of The Mammalian Srp	gi|11513832 (+1)	10 kDa	0	3	0
DNA topoisomerase 1	gi|11225260 (+15)	91 kDa	0	3	2
ribosomal protein S19, partial	gi|16924231 (+1)	17 kDa	0	3	1
Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	gi|15079888 (+2)	23 kDa	0	3	0
cold inducible RNA binding protein, isoform CRA_b	gi|119589927 (+1)	20 kDa	0	4	2
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a-like 1	gi|16741295 (+3)	106 kDa	0	7	2
unnamed protein product	gi|189066545 (+2)	29 kDa	0	12	3
ribosomal protein S4, X-linked, isoform CRA_a	gi|119592221 (+4)	43 kDa	1	3	3
KIAA0185	gi|1136430 (+2)	210 kDa	1	3	3
ubiquitin associated protein 2-like, isoform CRA_f	gi|119573603 (+18)	105 kDa	1	3	2

PSIP1 protein	gi|116283688 (+11)	30 kDa	1	3	1
ribosomal protein S16, isoform CRA_b	gi|119577297 (+2)	16 kDa	1	3	2
replication protein A3, 14kDa, isoform CRA_a	gi|119614009 (+2)	9 kDa	1	4	1
SUB1 homolog (S. cerevisiae)	gi|16307067 (+2)	14 kDa	1	6	2
DNA topoisomerase 3-alpha	gi|10835218 (+4)	112 kDa	1	7	2
RNA binding protein fox-1 homolog 2 isoform 5	gi|133925803 (+5)	47 kDa	1	16	2
hCG2016250, isoform CRA_c	gi|119618534 (+7)	29 kDa	3	4	3
ribosomal protein S10, isoform CRA_a	gi|119624187 (+2)	20 kDa	3	4	2
MYL6 protein	gi|113812151 (+6)	16 kDa	3	6	4
Similar to RIKEN cDNA 3930401K13 gene, partial	gi|13277568 (+6)	57 kDa	3	6	0
replication protein A 70 kDa DNA-binding subunit	gi|4506583	68 kDa	3	28	7
myosin regulatory light chain 12B	gi|15809016 (+3)	20 kDa	4	6	4
signal recognition particle 14 kDa protein [Pongo abelii]	gi|197099116	15 kDa	4	6	1
tropomyosin (227 AA)	gi|825723 (+2)	27 kDa	4	7	1
Chain B, Structure Of The Hsddb1-Hsddb2 Complex	gi|221046722 (+3)	49 kDa	4	7	0
DAZ-associated protein 1 isoform b	gi|25470886 (+4)	43 kDa	4	9	5
keratin, type I cytoskeletal 13 isoform a	gi|131412225 (+2)	50 kDa	4	12	4
heterogeneous nuclear ribonucleoprotein A0	gi|5803036	31 kDa	4	13	5
poly(rC)-binding protein 2 isoform b	gi|14141166 (+5)	38 kDa	4	17	5
RecQ protein-like (DNA helicase Q1-like)	gi|12654453 (+5)	73 kDa	6	7	0
unnamed protein product	gi|194379372 (+2)	52 kDa	6	7	5
Heterogeneous nuclear ribonucleoprotein U-like 1	gi|12803479 (+5)	96 kDa	6	9	4
ceruloplasmin (ferroxidase), isoform CRA_b	gi|119599289 (+6)	123 kDa	6	9	2
hnRNP-E1	gi|460771 (+1)	38 kDa	6	19	7
heterogeneous nuclear ribonucleoprotein D-like isoform a	gi|14110407 (+7)	46 kDa	7	25	7
musashi homolog 2 (Drosophila), isoform CRA_a	gi|119614912 (+10)	37 kDa	9	10	1
gamma-interferon-inducible protein 16 isoform 2	gi|112789562 (+3)	82 kDa	9	12	0
YTH domain family protein 2 isoform 1	gi|116812575 (+3)	62 kDa	10	13	2

KIAA1398 protein	gi|14133249 (+2)	170 kDa	12	13	12
damage-specific DNA binding protein 1, 127kDa, isoform CRA_d	gi|119594342 (+14)	128 kDa	12	16	0
Annexin A2	gi|16306978 (+6)	39 kDa	12	17	1
nucleolysin TIA-1 isoform p40 isoform 2	gi|188219591 (+3)	43 kDa	13	16	6
nucleolysin TIAR isoform 1	gi|4507499 (+1)	42 kDa	16	22	6
unnamed protein product	gi|194387362 (+3)	35 kDa	16	32	10
heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa), isoform CRA_f	gi|119626284	30 kDa	18	31	13
Chain A, Human Mitochondrial Single-Stranded Dna Binding Protein	gi|2624694	15 kDa	21	23	4
keratin 10 (epidermolytic hyperkeratosis; keratosis palmaris et plantaris), isoform CRA_b	gi|119581085 (+4)	63 kDa	21	25	14
epidermal cytokeratin 2	gi|181402 (+1)	66 kDa	22	26	12
helicase-like transcription factor	gi|21071052 (+3)	114 kDa	34	36	3
keratin 1	gi|11935049 (+4)	66 kDa	42	44	21
vimentin	gi|62414289	54 kDa	42	51	41

Appendix H Association analysis

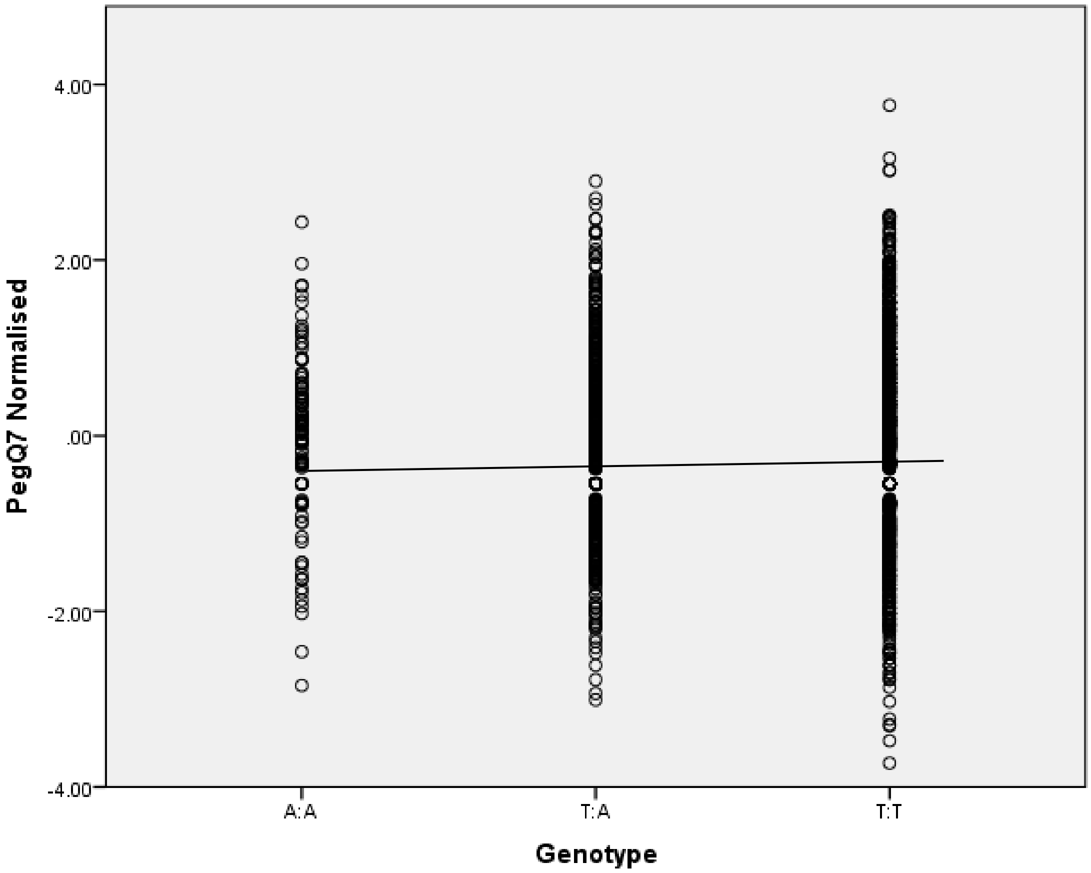


Figure H.1 Plot of the normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals from the Unaffected subgroup.

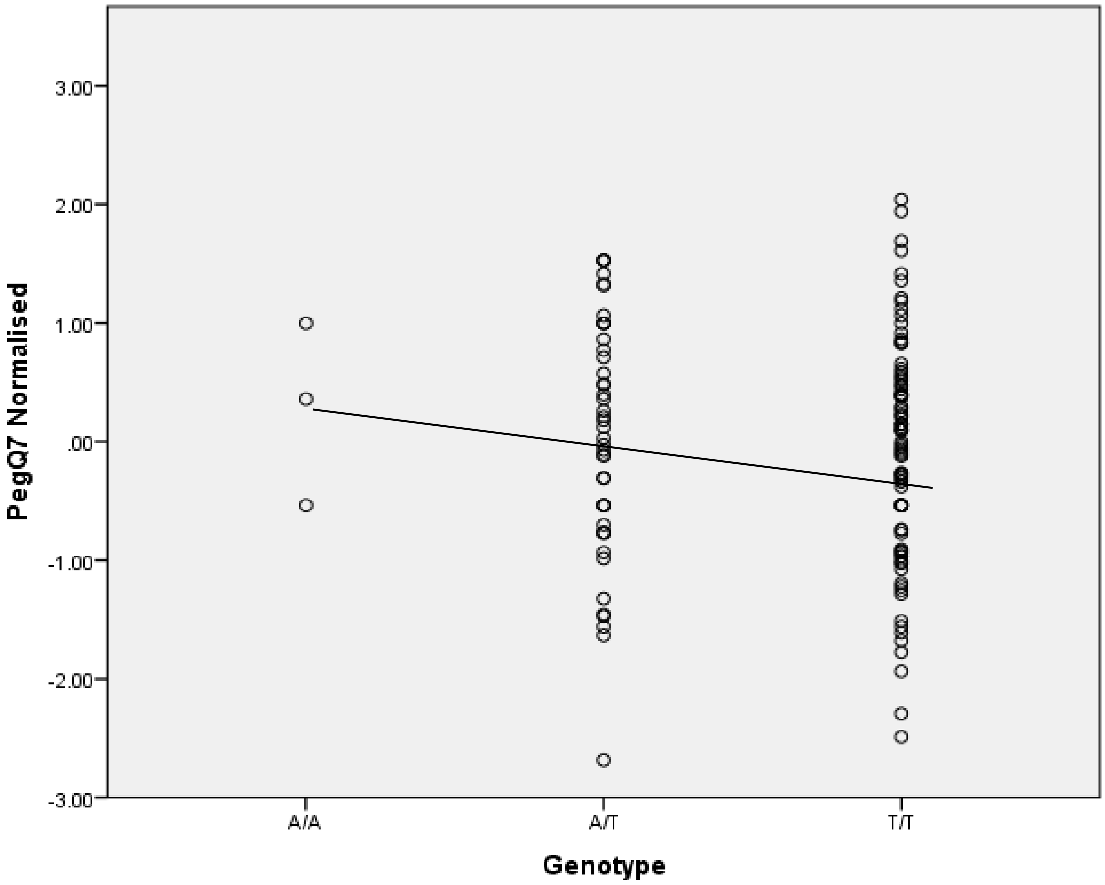


Figure H.2 Plot of normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals of the RD subgroup.With a minor allele frequency of 0.15 for this subgroup (N=183), there are a minimal number of A/A genotyped individuals available.

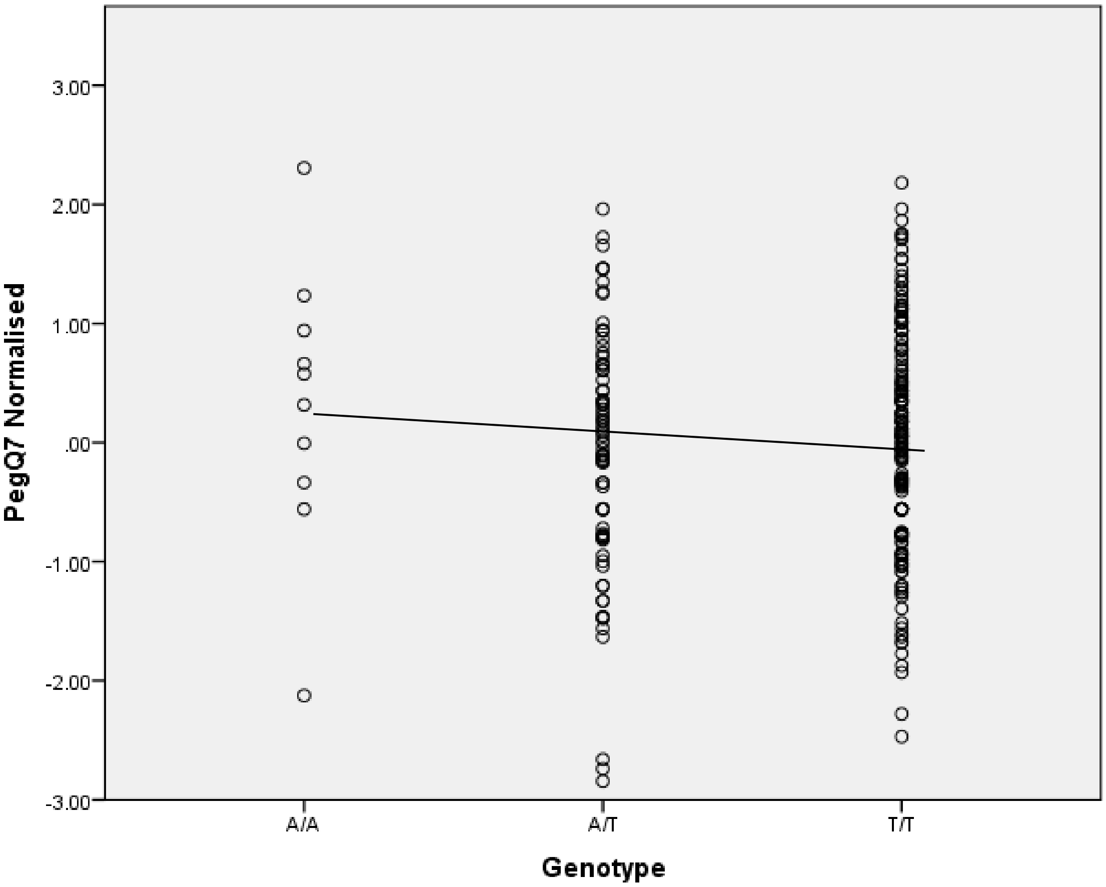


Figure H.3 Plot of normalised PegQ7 (mean=0, SD=1) distribution (y-axis) for each genotype of rs11855415 (x-axis) in individuals of the Affected subgroup.

Appendix I Genetic variant genotyping

For the genotyping of both the rs11855415 SNP and the rs10523972 VNTR, gDNA was extracted from all relevant cell lines using the QIAamp DNA mini kit according to the manufacturer's instructions. rs11855415 SNP genotyping was performed using the TaqMan assay (Life Technologies) which consisted of 5 µl GTX Express, 0.25 µl 20 X TaqMan Gene Expression Assay (rs11855415), 2.75 µl H_20 and 40 ng of gDNA in a 10 µl total reaction volume. Conditions used were 20 s at 95 °C followed by 40 cycles of 15 s at 95 °C and 60 s at 60 °C. For the VNTR genotyping, VNTR alleles were determined visually by 2.0% agarose gel electrophoresis of the PCR product (1 µl 6 X Orange-G dye added to 5 µl PCR product) following a PCR with the following conditions: 60 s at 95 °C followed by 40 cycles of 10 s at 95 °C, 20 s at 60 °C, 20 s at 72 °C. A primer pair flanking the VNTR was used: 5'-ACAGGGCTCGGTTCATTAAG and 5'-TCGGAATGTGGCTGTAACTG. PCR product size was dependent on VNTR allele - VNTR 10 allele corresponds to 516 bp, a 9 allele is 483 bp, 8 allele is 450 bp and a 6 allele corresponds to 384 bp. PCR product clean-up was performed (2 µl of ExoSAP-IT for every 5 µl PCR product, 37 °C 15min then 80 °C 15min) before confirming all sequences by Sanger sequencing (DNA Sequencing and Services, Dundee).

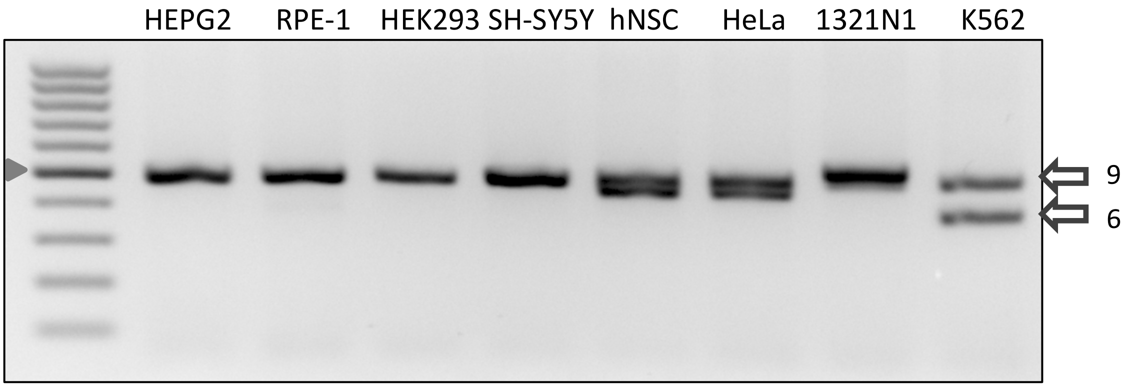


Figure I.1 Genotyping of cell lines for the VNTR identified in the secondary promoter of PCSK6. Prior to use as a potential model for functional analysis, various cell lines cell lines were genotyped and the resulting product sequenced. See Table I.1 for a list of resulting genotypes. The grey triangle indicates 500 bp on the Gene O'Ruler 100 bp ladder. Arrows indicate 9 and 6 copies of the 33 bp tandem repeat, the K562 genotype 9/6 has been highlighted. In this figure a VNTR 10 allele corresponds to 516 bp, a 9 allele is 483 bp, 8 allele is 450 bp and a 6 allele corresponds to 384 bp.

Table I.1 Genotyping of cell lines for the rs11855415 SNP and the VNTR identified in the secondary promoter of PCSK6

	Tissue	*rs11855415*	*VNTR*
CHP212	Brain	TT	9/9
LAN5	Brain	TT	9/6
LMR32	Brain	AT	9/9
KELLY	Brain	TT	9/6
M17	Brain *	TT	9/9
MRC5V2	Lung **	AT	9/9
NT2	Testis ***	AT	9/9
SKNF1	Brain *	AT	-
SKNMC	Brain †	TT	6/6
SKNAS	Brain *	TT	9/9
293T	Kidney	TT	9/9
SH-SY5Y	Brain *	TT	9/9
K562	Bone marrow	TT	9/6
RPE-1	Retina	TT	9/9
HeLa	Cervix	AT	9/8
hNSC	Brain	AT	9/8
HEPG2	Liver	AT	9/9
1321N1	Brain	AA	10/10 ††
HEK293	Kidney	TT	9/9

Rows in grey indicate cell lines previously genotyped by our collaborator Dr William Brandler and a dash (–) indicates an ambiguous VNTR genotype where genotyping was unsuccessful
* derived from metastatic site: bone marrow ** foetal from human *** derived from metastatic site: lung
† derived from metastatic site: supra-orbital area †† At this resolution a 10/9 VNTR cannot be discounted.
Note: none of the cell lines employed in this project are known to possess rearrangements on chromosome 15 according to the European Collection of Cell Cultures (ECACC).

References

ADES, C. & RAMIRES, E. N. 2002. Asymmetry of leg use during prey handling in the spider *Scytodes globula* (Scytodidae). *Journal of Insect Behavior,* 15**,** 563-570.

AGARWAL, V., BELL, G. W., NAM, J. W. & BARTEL, D. P. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *Elife,* 4.

AJAWATANAWONG, P., ATKINSON, G. C., WATSON-HAIGH, N. S., MACKENZIE, B. & BALDAUF, S. L. 2012. SeqFIRE: a web application for automated extraction of indel regions and conserved blocks from protein multiple sequence alignments. *Nucleic Acids Res,* 40**,** W340-7.

ALLENDORF, F. W. & LUIKART, G. 2007. *Conservation and the genetics of populations,* Malden, MA, Blackwell Pub.

ALLENDORF, F. W., LUIKART, G. & AITKEN, S. N. 2013. *Conservation and the genetics of populations,* Hoboken, John Wiley & Sons.

ALLISON, P. D. 2002. *Missing data,* Thousand Oaks, Calif., Sage Publications.

AMMERMAN, A. J. & CAVALLI-SFORZA, L. L. 1984. *The neolithic transition and the genetics of populations in Europe,* Princeton, N.J., Princeton University Press.

ANDERSON, E. D., MOLLOY, S. S., JEAN, F., FEI, H., SHIMAMURA, S. & THOMAS, G. 2002. The ordered and compartment-specific autoproteolytic removal of the furin intramolecular chaperone is required for enzyme activation. *Journal of Biological Chemistry,* 277**,** 12879-12890.

ANDREW, R. J., TOMMASI, L. & FORD, N. 2000. Motor Control by Vision and the Evolution of Cerebral Lateralization. *Brain and Language,* 73**,** 220-235.

ANDRIC, M., SOLODKIN, A., BUCCINO, G., GOLDIN-MEADOW, S., RIZZOLATTI, G. & SMALL, S. L. 2013. Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia,* 51**,** 1619-1629.

ANFORA, G., RIGOSI, E., FRASNELLI, E., RUGA, V., TRONA, F. & VALLORTIGARA, G. 2011. Lateralization in the invertebrate brain: left-right asymmetry of olfaction in bumble bee, *Bombus terrestris*. *PLoS One,* 6**,** e18903.

ANNETT, M. 1970. A classification of hand preference by association analysis. *Br J Psychol,* 61**,** 303-21.

ANNETT, M. 1992. Five tests of hand skill. *Cortex,* 28**,** 583-600.

ANNETT, M. 1994. Handedness as a continuous variable with dextral shift: sex, generation, and family handedness in subgroups of left- and right-handers. *Behav Genet,* 24**,** 51-63.

ANNETT, M. 1998. Handedness and cerebral dominance: the right shift theory. *J Neuropsychiatry Clin Neurosci,* 10**,** 459-69.

ANNETT, M. & KILSHAW, D. 1984. Lateral preference and skill in dyslexics: implications of the right shift theory. *J Child Psychol Psychiatry,* 25**,** 357-77.

ARMOUR, J. A., DAVISON, A. & MCMANUS, I. 2014. Genome-wide association study of handedness excludes simple genetic models. *Heredity,* 112**,** 221-225.

ARNING, L., OCKLENBURG, S., SCHULZ, S., NESS, V., GERDING, W. M., HENGSTLER, J. G., FALKENSTEIN, M., EPPLEN, J. T., GUNTURKUN, O. & BESTE, C. 2013. VNTR Polymorphism Is Associated with Degree of Handedness but Not Direction of Handedness. *PLoS One,* 8**,** e67251.

ARNOLD, S. J. & ROBERTSON, E. J. 2009. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol,* 10**,** 91-103.

ARTENSTEIN, A. W. & OPAL, S. M. 2011. Proprotein Convertases in Health and Disease. *New England Journal of Medicine,* 365**,** 2507-2518.

ASAI, T., SUGIMORI, E. & TANNO, Y. 2011. A psychometric approach to the relationship between hand-foot preference and auditory hallucinations in the general population: Atypical cerebral lateralization may cause an abnormal sense of agency. *Psychiatry Research,* 189**,** 220-227.

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., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

BACIADONNA, L., ZUCCA, P. & TOMMASI, L. 2010. Posture in ovo as a precursor of footedness in ostriches (*Struthio camelus*). *Behavioural Processes,* 83**,** 130-133.

BACKWELL, P. R. Y., MATSUMASA, M., DOUBLE, M., ROBERTS, A., MURAI, M., KEOGH, J. S. & JENNIONS, M. D. 2007. What are the consequences of being left-clawed in a predominantly right-clawed fiddler crab? *Proceedings of the Royal Society B: Biological Sciences,* 274**,** 2723-2729.

BADANO, J. L., MITSUMA, N., BEALES, P. L. & KATSANIS, N. 2006. The ciliopathies: an emerging class of human genetic disorders. *Annu Rev Genomics Hum Genet,* 7**,** 125-48.

BARRANTES-VIDAL, N., GÓMEZ-DE-REGIL, L., NAVARRO, B., VICENS-VILANOVA, J., OBIOLS, J. & KWAPIL, T. 2013. Psychotic-like symptoms and positive schizotypy are associated with mixed and ambiguous handedness in an adolescent community sample. *Psychiatry Research,* 206**,** 188-194.

BASSI, D. E., ZHANG, J., CENNA, J., LITWIN, S., CUKIERMAN, E. & KLEIN-SZANTO, A. J. 2010. Proprotein convertase inhibition results in decreased skin cell proliferation, tumorigenesis, and metastasis. *Neoplasia,* 12**,** 516-26.

BATISTA, P. J. & CHANG, H. Y. 2013. Long noncoding RNAs: Cellular address codes in development and disease. *Cell,* 152**,** 1298-1307.

BAUER, D. E., KAMRAN, S. C., LESSARD, S., XU, J., FUJIWARA, Y., LIN, C., SHAO, Z., CANVER, M. C., SMITH, E. C., PINELLO, L., SABO, P. J., VIERSTRA, J., VOIT, R. A., YUAN, G. C., PORTEUS, M. H., STAMATOYANNOPOULOS, J. A., LETTRE, G. & ORKIN, S. H. 2013. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science,* 342**,** 253-7.

BAUER, R. H. 1993. Lateralization of neural control for vocalization by the frog (*Rana Pipiens*). *Psychobiology,* 21**,** 243-248.

BENDTSEN, J. D., NIELSEN, H., VON HEIJNE, G. & BRUNAK, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol,* 340**,** 783-95.

BERNSTEIN, B. E., BIRNEY, E., DUNHAM, I., GREEN, E. D., GUNTER, C. & SNYDER, M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

BEST, M. & DEMB, J. B. 1999. Normal planum temporale asymmetry in dyslexics with a magnocellular pathway deficit. *Neuroreport,* 10**,** 607-12.

BIANCO, I. H. & WILSON, S. W. 2009. The habenular nuclei: a conserved asymmetric relay station in the vertebrate brain. *Philos Trans R Soc Lond B Biol Sci,* 364**,** 1005-20.

BISAZZA, A., CANTALUPO, C., CAPOCCHIANO, M. & VALLORTIGARA, G. 2000. Population lateralisation and social behaviour: A study with 16 species of fish. *Laterality,* 5**,** 269-284.

BISAZZA, A., ROGERS, L. J. & VALLORTIGARA, G. 1998. The origins of cerebral asymmetry: a review of evidence of behavioural and brain lateralization in fishes, reptiles and amphibians. *Neuroscience & Biobehavioral Reviews,* 22**,** 411-426.

BLOSS, C. S., DELIS, D. C., SALMON, D. P. & BONDI, M. W. 2010. APOE genotype is associated with left-handedness and visuospatial skills in children. *Neurobiol Aging,* 31**,** 787-95.

BODMER, W. & BONILLA, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet,* 40**,** 695-701.

BOROD, J. C., CARON, H. S. & KOFF, E. 1984. Left-handers and right-handers compared on performance and preference measures of lateral dominance. *Br J Psychol,* 75 (Pt 2)**,** 177-86.

BOYD, A., GOLDING, J., MACLEOD, J., LAWLOR, D. A., FRASER, A., HENDERSON, J., MOLLOY, L., NESS, A., RING, S. & DAVEY SMITH, G. 2013. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol,* 42**,** 111-27.

BRADSHAW, J. L. & ROGERS, L. J. 1993. *The evolution of lateral asymmetries, language, tool use, and intellect,* San Diego, Academic Press.

BRAINSPAN 2011. BrainSpan: Atlas of the Developing Human Brain [Internet].

BRANDLER, W. M., MORRIS, A. P., EVANS, D. M., SCERRI, T. S., KEMP, J. P., TIMPSON, N. J., ST POURCAIN, B., SMITH, G. D., RING, S. M., STEIN, J., MONACO, A. P., TALCOTT, J. B., FISHER, S. E., WEBBER, C. & PARACCHINI, S. 2013. Common variants in left/right asymmetry genes and pathways are associated with relative hand skill. *PLoS Genet,* 9**,** e1003751.

BRANDLER, W. M. & PARACCHINI, S. 2014. The genetic relationship between handedness and neurodevelopmental disorders. *Trends in molecular medicine,* 20**,** 83-90.

BROWN, C. & MAGAT, M. 2011. The evolution of lateralized foot use in parrots: a phylogenetic approach. *Behavioral Ecology***,** arr114.

BRYDEN, M. P., ROY, E. A., MCMANUS, I. C. & BULMAN-FLEMING, M. B. 1997. On the genetics and measurement of human handedness. *Laterality,* 2**,** 317-36.

BULMAN-FLEMING, M. B., BRYDEN, M. P. & ROGERS, T. T. 1997. Mouse paw preference: effects of variations in testing protocol. *Behavioural Brain Research,* 86**,** 79-87.

BURGESS, D. & FREELING, M. 2014. The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. *The Plant Cell,* 26**,** 946-961.

BUSH, W. S. & MOORE, J. H. 2012. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology,* 8**,** e1002822.

BYRNE, R. W. & BYRNE, J. M. 1991. Hand preferences in the skilled gathering tasks of mountain gorillas (*Gorilla G Beringei*). *Cortex,* 27**,** 521-546.

C., M. I. 2002. *Right Hand, Left Hand,* London, Weidenfeld & Nicolson.

CABILI, M. N., TRAPNELL, C., GOFF, L., KOZIOL, M., TAZON-VEGA, B., REGEV, A. & RINN, J. L. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev,* 25**,** 1915-27.

CANNING, C., CRAIN, D., EATON, T. S., NUESSLY, K., FRIEDLAENDER, A., HURST, T., PARKS, S., WARE, C., WILEY, D. & WEINRICH, M. 2011. Population-level lateralized feeding behaviour in North Atlantic humpback whales, *Megaptera novaeangliae*. *Animal Behaviour,* 82**,** 901-909.

CANTALUPO, C. & HOPKINS, W. D. 2001. Asymmetric Broca's area in great apes. *Nature,* 414**,** 505.

CARTHARIUS, K., FRECH, K., GROTE, K., KLOCKE, B., HALTMEIER, M., KLINGENHOFF, A., FRISCH, M., BAYERLEIN, M. & WERNER, T. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics,* 21**,** 2933-42.

CASEY, M. B. & MARTINO, C. M. 2000. Asymmetrical hatching behaviors influence the development of postnatal laterality in domestic chicks (*Gallus gallus*). *Dev Psychobiol,* 37**,** 13-24.

CASHMORE, L., UOMINI, N. & CHAPELAIN, A. 2008. The evolution of handedness in humans and great apes: a review and current issues. *J Anthropol Sci,* 86**,** 7-35.

CATTS, H. W., ADLOF, S. M., HOGAN, T. & WEISMER, S. E. 2005. Are Specific Language Impairment and Dyslexia Distinct Disorders? *Journal of speech, language, and hearing research : JSLHR,* 48**,** 1378-1396.

CEMAZAR, M., HRELJAC, I., SERSA, G. & FILIPIC, M. 2010. Construction of EGFP Expressing HepG2 Cell Line Using Electroporation. *In:* DÖSSEL, O. & SCHLEGEL, W. (eds.) *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany.* Springer Berlin Heidelberg.

CHANDRASEKAR, G., VESTERLUND, L., HULTENBY, K., TAPIA-PÁEZ, I. & KERE, J. 2013. The Zebrafish Orthologue of the Dyslexia Candidate Gene *DYX1C1* Is Essential for Cilia Growth and Function. *PLoS ONE,* 8**,** e63123.

CHAPELAIN, A. S., HOGERVORST, E., MBONZO, P. & HOPKINS, W. D. 2011. Hand Preferences for Bimanual Coordination in 77 Bonobos (*Pan paniscus*): Replication and Extension. *International Journal of Primatology,* 32**,** 491-510.

CHENG, M., WATSON, P. H., PATERSON, J. A., SEIDAH, N., CHRETIEN, M. & SHIU, R. P. C. 1997. Pro-protein convertase gene expression in human breast cancer. *International Journal of Cancer,* 71**,** 966-971.

CHERN, T.-M., VAN NIMWEGEN, E., KAI, C., KAWAI, J., CARNINCI, P., HAYASHIZAKI, Y. & ZAVOLAN, M. 2006. A Simple Physical Model Predicts Small Exon Length Variations. *PLoS Genet,* 2**,** e45.

CHEUNG, Y. F., KAN, Z., GARRETT-ENGELE, P., GALL, I., MURDOCH, H., BAILLIE, G. S., CAMARGO, L. M., JOHNSON, J. M., HOUSLAY, M. D. & CASTLE, J. C. 2007. PDE4B5, a novel, super-short, brain-specific cAMP phosphodiesterase-4 variant whose isoform-specifying N-terminal region is identical to that of cAMP phosphodiesterase-4D6 (PDE4D6). *J Pharmacol Exp Ther,* 322**,** 600-9.

CHIRON, C., JAMBAQUE, I., NABBOUT, R., LOUNES, R., SYROTA, A. & DULAC, O. 1997. The right brain hemisphere is dominant in human infants. *Brain,* 120 (Pt 6)**,** 1057-65.

CHO, H. S. & LEAHY, D. J. 2002. Structure of the extracellular region of HER3 reveals an interdomain tether. *Science,* 297**,** 1330-3.

CHURCHMAN, L. S. & WEISSMAN, J. S. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature,* 469**,** 368-73.

CONARD, N. J. & RICHTER, J. 2011. *Neanderthal Lifeways, Subsistence and Technology: One Hundred Fifty Years of Neanderthal Study*, Springer Science & Business Media.

CONCHA, M. L., BURDINE, R. D., RUSSELL, C., SCHIER, A. F. & WILSON, S. W. 2000. A Nodal Signaling Pathway Regulates the Laterality of Neuroanatomical Asymmetries in the Zebrafish Forebrain. *Neuron,* 28**,** 399-409.

CONSORTIUM, G. T. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet,* 45**,** 580-5.

CONSTAM, D. B. & ROBERTSON, E. J. 2000. SPC4/PACE4 regulates a TGFbeta signaling network during axis formation. *Genes Dev,* 14**,** 1146-55.

COOKSON, W., LIANG, L., ABECASIS, G., MOFFATT, M. & LATHROP, M. 2009. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics,* 10**,** 184-194.

CORBALLIS, M. C. 1997. The genetics and evolution of handedness. *Psychol Rev,* 104**,** 714-27.

CORBALLIS, M. C. 2003. From mouth to hand: gesture, speech, and the evolution of right-handedness. *Behav Brain Sci,* 26**,** 199-208; discussion 208-60.

CORBALLIS, M. C. 2009. The evolution and genetics of cerebral asymmetry. *Philos Trans R Soc Lond B Biol Sci,* 364**,** 867-79.

CORBETTA, D., WILLIAMS, J. & SNAPP-CHILDS, W. 2006. Plasticity in the development of handedness: evidence from normal development and early asymmetric brain injury. *Dev Psychobiol,* 48**,** 460-71.

CORE, L. J., WATERFALL, J. J. & LIS, J. T. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science,* 322**,** 1845-8.

COREY, D. M., HURLEY, M. M. & FOUNDAS, A. L. 2001. Right and left handedness defined: a multivariate approach using hand preference and hand performance measures. *Neuropsychiatry Neuropsychol Behav Neurol,* 14**,** 144-52.

CORP, N. & BYRNE, R. W. 2004. Sex difference in chimpanzee handedness. *American Journal of Physical Anthropology,* 123**,** 62-68.

CORRADIN, O. & SCACHERI, P. C. 2014. Enhancer variants: evaluating functions in common disease. *Genome Med,* 6**,** 85.

CSERMELY, D. 2004. Lateralisation in birds of prey: adaptive and phylogenetic considerations. *Behav Processes,* 67**,** 511-20.

D'ANJOU, F., ROUTHIER, S., PERREAULT, J. P., LATIL, A., BONNEL, D., FOURNIER, I., SALZET, M. & DAY, R. 2011. Molecular Validation of PACE4 as a Target in Prostate Cancer. *Transl Oncol,* 4**,** 157-72.

DA CRUZ E SILVA, E. F., FOX, C. A., OUIMET, C. C., GUSTAFSON, E., WATSON, S. J. & GREENGARD, P. 1995. Differential expression of protein phosphatase 1 isoforms in mammalian brain. *J Neurosci,* 15**,** 3375-89.

DAVIDSON, R. J. 2004. Well-being and affective style: neural substrates and biobehavioural correlates. *Philos Trans R Soc Lond B Biol Sci,* 359**,** 1395-411.

DE CASTRO, E., SIGRIST, C. J., GATTIKER, A., BULLIARD, V., LANGENDIJK-GENEVAUX, P. S., GASTEIGER, E., BAIROCH, A. & HULO, N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res,* 34**,** W362-5.

DEFRIES, J. C. 1989. Gender ratios in children with reading disability and their affected relatives: a commentary. *J Learn Disabil,* 22**,** 544-5.

DEININGER, P. 2011. Alu elements: know the SINEs. *Genome Biol,* 12**,** 236.

DELIC, S., LOTTMANN, N., JETSCHKE, K., REIFENBERGER, G. & RIEMENSCHNEIDER, M. J. 2012. Identification and functional validation of CDH11, PCSK6 and SH3GL3 as novel glioma invasion-associated candidate genes. *Neuropathol Appl Neurobiol,* 38**,** 201-12.

DENNIS, M. Y., PARACCHINI, S., SCERRI, T. S., PROKUNINA-OLSSON, L., KNIGHT, J. C., WADE-MARTINS, R., COGGILL, P., BECK, S., GREEN, E. D. & MONACO, A. P. 2009. A Common Variant Associated with Dyslexia Reduces Expression of the KIAA0319 Gene. *PLoS Genetics,* 5**,** e1000436.

DERRIEN, T., JOHNSON, R., BUSSOTTI, G., TANZER, A., DJEBALI, S., TILGNER, H., GUERNEC, G., MARTIN, D., MERKEL, A., KNOWLES, D. G., LAGARDE, J., VEERAVALLI, L., RUAN, X., RUAN, Y., LASSMANN, T., CARNINCI, P., BROWN, J. B., LIPOVICH, L., GONZALEZ, J. M., THOMAS, M., DAVIS, C. A., SHIEKHATTAR, R., GINGERAS, T. R., HUBBARD, T. J., NOTREDAME, C., HARROW, J. & GUIGO, R. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

DIJKMANS, T. F., VAN HOOIJDONK, L. W., FITZSIMONS, C. P. & VREUGDENHIL, E. 2010. The doublecortin gene family and disorders of neuronal structure. *Cent Nerv Syst Agents Med Chem,* 10**,** 32-46.

DIMAS, A. S., DEUTSCH, S., STRANGER, B. E., MONTGOMERY, S. B., BOREL, C., ATTAR-COHEN, H., INGLE, C., BEAZLEY, C., ARCELUS, M. G., SEKOWSKA, M., GAGNEBIN, M., NISBETT, J., DELOUKAS, P., DERMITZAKIS, E. T. & ANTONARAKIS, S. E. 2009. Common regulatory variation impacts gene expression in a cell type dependent manner. *Science (New York, N.Y.),* 325**,** 1246-1250.

DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., XUE, C., MARINOV, G. K., KHATUN, J., WILLIAMS, B. A., ZALESKI, C., ROZOWSKY, J., RODER, M., KOKOCINSKI, F., ABDELHAMID, R. F., ALIOTO, T., ANTOSHECHKIN, I., BAER, M. T., BAR, N. S., BATUT, P., BELL, K., BELL, I., CHAKRABORTTY, S., CHEN, X., CHRAST, J., CURADO, J., DERRIEN, T., DRENKOW, J., DUMAIS, E., DUMAIS, J., DUTTAGUPTA, R., FALCONNET, E., FASTUCA, M., FEJES-TOTH, K., FERREIRA, P., FOISSAC, S., FULLWOOD, M. J., GAO, H., GONZALEZ, D., GORDON, A., GUNAWARDENA, H., HOWALD, C., JHA, S., JOHNSON, R., KAPRANOV, P., KING, B., KINGSWOOD, C., LUO, O. J., PARK, E., PERSAUD, K., PREALL, J. B., RIBECA, P., RISK, B., ROBYR, D., SAMMETH, M., SCHAFFER, L., SEE, L. H., SHAHAB, A., SKANCKE, J., SUZUKI, A. M., TAKAHASHI, H., TILGNER, H., TROUT, D., WALTERS, N., WANG, H., WROBEL, J., YU, Y., RUAN, X., HAYASHIZAKI, Y., HARROW, J., GERSTEIN, M., HUBBARD, T., REYMOND, A., ANTONARAKIS, S. E., HANNON, G., GIDDINGS, M. C., RUAN, Y., WOLD, B., CARNINCI, P., GUIGO, R. & GINGERAS, T. R. 2012. Landscape of transcription in human cells. *Nature,* 489**,** 101-8.

DOOLITTLE, W. F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences,* 110**,** 5294-5300.

DOOLITTLE, W. F. & SAPIENZA, C. 1980. SELFISH GENES, THE PHENOTYPE PARADIGM AND GENOME EVOLUTION. *Nature,* 284**,** 601-603.

DRAGOVIC, M. & HAMMOND, G. 2005. Handedness in schizophrenia: a quantitative review of evidence. *Acta Psychiatrica Scandinavica,* 111**,** 410-419.

DRIESEN, N. & RAZ, N. 1995. The influence of sex, age, and handedness on corpus callosum morphology: A meta-analysis. *Psychobiology,* 23**,** 240-247.

DUARA, R., KUSHCH, A., GROSS-GLENN, K., BARKER, W. W., JALLAD, B., PASCAL, S., LOEWENSTEIN, D. A., SHELDON, J., RABIN, M., LEVIN, B. & ET AL. 1991. Neuroanatomic differences between dyslexic and normal readers on magnetic resonance imaging scans. *Arch Neurol,* 48**,** 410-6.

ECKERT, M. A., LEONARD, C. M., RICHARDS, T. L., AYLWARD, E. H., THOMSON, J. & BERNINGER, V. W. 2003. Anatomical correlates of dyslexia: frontal and cerebellar findings. *Brain,* 126**,** 482-94.

EDWARDS, STACEY L., BEESLEY, J., FRENCH, JULIET D. & DUNNING, ALISON M. 2013. Beyond GWASs: Illuminating the Dark Road from Association to Function. *American Journal of Human Genetics,* 93**,** 779-797.

EGLINTON, E. & ANNETT, M. 1994. Handedness and dyslexia: a meta-analysis. *Percept Mot Skills,* 79**,** 1611-6.

EHRET, G. 1987. Left hemisphere advantage in the mouse brain for recognizing ultrasonic communication calls. *Nature,* 325**,** 249-251.

EICHLER, E. E., NICKERSON, D. A., ALTSHULER, D., BOWCOCK, A. M., BROOKS, L. D., CARTER, N. P., CHURCH, D. M., FELSENFELD, A., GUYER, M., LEE, C., LUPSKI, J. R., MULLIKIN, J. C., PRITCHARD, J. K., SEBAT, J., SHERRY, S. T., SMITH, D., VALLE, D. & WATERSTON, R. H. 2007. Completing the map of human genetic variation. *Nature,* 447**,** 161-5.

ELIAS, L. J. & BRYDEN, M. P. 1998. Footedness is a better predictor of language lateralisation than handedness. *Laterality,* 3**,** 41-51.

FABBRO, F., LIBERA, L. & TAVANO, A. 2002. A callosal transfer deficit in children with developmental language disorder. *Neuropsychologia,* 40**,** 1541-6.

FABRE-THORPE, M., FAGOT, J., LORINCZ, E., LEVESQUE, F. & VAUCLAIR, J. 1993. Laterality in cats: paw preference and performance in a visuomotor activity. *Cortex,* 29**,** 15-24.

FAGOT, J. & VAUCLAIR, J. 1991. Manual laterality in nonhuman primates: a distinction between handedness and manual specialization. *Psychological bulletin,* 109**,** 76.

FAIRFAX, B. P., MAKINO, S., RADHAKRISHNAN, J., PLANT, K., LESLIE, S., DILTHEY, A., ELLIS, P., LANGFORD, C., VANNBERG, F. O. & KNIGHT, J. C. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature genetics,* 44**,** 502-510.

FAURIE, C. & RAYMOND, M. 2004. Handedness frequency over more than ten thousand years. *Proceedings of the Royal Society B: Biological Sciences,* 271**,** S43-S45.

FAURIE, C. & RAYMOND, M. 2005. Handedness, homicide and negative frequency-dependent selection. *Proc Biol Sci,* 272**,** 25-8.

FENG, H., QIN, Z. & ZHANG, X. 2013. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer letters,* 340**,** 179-191.

FIDALGO, M., BARRALES, R. R., IBEAS, J. I. & JIMENEZ, J. 2006. Adaptive evolution by mutations in the FLO11 gene. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 11228-11233.

FIELD, L. L., SHUMANSKY, K., RYAN, J., TRUONG, D., SWIERGALA, E. & KAPLAN, B. J. 2013. Dense-map genome scan for dyslexia supports loci at 4q13, 16p12, 17q22; suggests novel locus at 7q36. *Genes, Brain and Behavior,* 12**,** 56-69.

FISKERSTRAND, C. E., LOVEJOY, E. A. & QUINN, J. P. 1999. An intronic polymorphic domain often associated with susceptibility to affective disorders has allele dependent differential enhancer activity in embryonic stem cells. *FEBS Letters,* 458**,** 171-174.

FLAMES, N., LONG, J. E., GARRATT, A. N., FISCHER, T. M., GASSMANN, M., BIRCHMEIER, C., LAI, C., RUBENSTEIN, J. L. & MARIN, O. 2004. Short- and long-range attraction of cortical GABAergic interneurons by neuregulin-1. *Neuron,* 44**,** 251-61.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BILLIS, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FITZGERALD, S., GIL, L., GIRÓN, C. G., GORDON, L., HOURLIER, T., HUNT, S., JOHNSON, N., JUETTEMANN, T., KÄHÄRI, A. K., KEENAN, S., KULESHA, E., MARTIN, F. J., MAUREL, T., MCLAREN, W. M., MURPHY, D. N., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RUFFIER, M., SHEPPARD, D., TAYLOR, K., THORMANN, A., TREVANION, S. J., VULLO, A., WILDER, S. P., WILSON, M., ZADISSA, A., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., HARROW, J., HERRERO, J., HUBBARD, T. J. P., KINSELLA, R., MUFFATO, M., PARKER, A., SPUDICH, G., YATES, A., ZERBINO, D. R. & SEARLE, S. M. J. 2014. Ensembl 2014. *Nucleic Acids Research,* 42**,** D749-D755.

FLIEGAUF, M., BENZING, T. & OMRAN, H. 2007. When cilia go bad: cilia defects and ciliopathies. *Nat Rev Mol Cell Biol,* 8**,** 880-93.

FORRESTER, G. S., QUARESMINI, C., LEAVENS, D. A., MARESCHAL, D. & THOMAS, M. S. 2013. Human handedness: an inherited evolutionary trait. *Behavioural brain research,* 237**,** 200-206.

FRANCKS, C., DELISI, L. E., FISHER, S. E., LAVAL, S. H., RUE, J. E., STEIN, J. F. & MONACO, A. P. 2003a. Confirmatory evidence for linkage of relative hand skill to 2p12-q11. *Am J Hum Genet,* 72**,** 499-502.

FRANCKS, C., DELISI, L. E., SHAW, S. H., FISHER, S. E., RICHARDSON, A. J., STEIN, J. F. & MONACO, A. P. 2003b. Parent-of-origin effects on handedness and schizophrenia susceptibility on chromosome 2p12-q11. *Hum Mol Genet,* 12**,** 3225-30.

FRANCKS, C., FISHER, S. E., MACPHIE, I. L., RICHARDSON, A. J., MARLOW, A. J., STEIN, J. F. & MONACO, A. P. 2002. A genomewide linkage screen for relative hand skill in sibling pairs. *Am J Hum Genet,* 70**,** 800-5.

FRANCKS, C., FISHER, S. E., MARLOW, A. J., MACPHIE, I. L., TAYLOR, K. E., RICHARDSON, A. J., STEIN, J. F. & MONACO, A. P. 2003c. Familial and genetic effects on motor coordination, laterality, and reading-related cognition. *Am J Psychiatry,* 160**,** 1970-7.

FRANCKS, C., MAEGAWA, S., LAUREN, J., ABRAHAMS, B. S., VELAYOS-BAEZA, A., MEDLAND, S. E., COLELLA, S., GROSZER, M., MCAULEY, E. Z., CAFFREY, T. M., TIMMUSK, T., PRUUNSILD, P., KOPPEL, I., LIND, P. A., MATSUMOTO-ITABA, N., NICOD, J., XIONG, L., JOOBER, R., ENARD, W., KRINSKY, B., NANBA, E., RICHARDSON, A. J., RILEY, B. P., MARTIN, N. G., STRITTMATTER, S. M., MOLLER, H. J., RUJESCU, D., ST CLAIR, D., MUGLIA, P., ROOS, J. L., FISHER, S. E., WADE-MARTINS, R., ROULEAU, G. A., STEIN, J. F., KARAYIORGOU, M., GESCHWIND, D. H., RAGOUSSIS, J., KENDLER, K. S., AIRAKSINEN, M. S., OSHIMURA, M., DELISI, L. E. & MONACO, A. P. 2007. LRRTM1 on chromosome 2p12 is a maternally suppressed gene that is associated paternally with handedness and schizophrenia. *Mol Psychiatry,* 12**,** 1129-39, 1057.

FRASNELLI, E., VALLORTIGARA, G. & ROGERS, L. J. 2012. Left–right asymmetries of behaviour and nervous system in invertebrates. *Neuroscience & Biobehavioral Reviews,* 36**,** 1273-1291.

FRAYER, D. W., LOZANO, M., BERMUDEZ DE CASTRO, J. M., CARBONELL, E., ARSUAGA, J. L., RADOVCIC, J., FIORE, I. & BONDIOLI, L. 2012. More than 500,000 years of right-handedness in Europe. *Laterality,* 17**,** 51-69.

FRAZIER, T. W., KESHAVAN, M. S., MINSHEW, N. J. & HARDAN, A. Y. 2012. A two-year longitudinal MRI study of the corpus callosum in autism. *J Autism Dev Disord,* 42**,** 2312-22.

FRIEDERICI, A. D. 2006. The neural basis of language development and its impairment. *Neuron,* 52**,** 941-52.

FRISCHKNECHT, R. & SEIDENBECHER, C. I. 2012. Brevican: A key proteoglycan in the perisynaptic extracellular matrix of the brain. *The International Journal of Biochemistry & Cell Biology,* 44**,** 1051-1054.

FUKE, S., SUO, S., TAKAHASHI, N., KOIKE, H., SASAGAWA, N. & ISHIURA, S. 2001. The VNTR polymorphism of the human dopamine transporter (DAT1) gene affects gene expression. *Pharmacogenomics J,* 1**,** 152-156.

GALABURDA, A. M. 1989. Ordinary and extraordinary brain development: Anatomical variation in developmental dyslexia. *Ann Dyslexia,* 39**,** 65-80.

GALABURDA, A. M., LEMAY, M., KEMPER, T. L. & GESCHWIND, N. 1978. Right-left asymmetrics in the brain. *Science,* 199**,** 852-6.

GALABURDA, A. M., SHERMAN, G. F., ROSEN, G. D., ABOITIZ, F. & GESCHWIND, N. 1985. Developmental dyslexia: four consecutive patients with cortical anomalies. *Ann Neurol,* 18**,** 222-33.

GEMAYEL, R., CHO, J., BOEYNAEMS, S. & VERSTREPEN, K. J. 2012. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes,* 3**,** 461-480.

GEMAYEL, R., VINCES, M. D., LEGENDRE, M. & VERSTREPEN, K. J. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet,* 44**,** 445-77.

GENOMES PROJECT, C., ABECASIS, G. R., AUTON, A., BROOKS, L. D., DEPRISTO, M. A., DURBIN, R. M., HANDSAKER, R. E., KANG, H. M., MARTH, G. T. & MCVEAN, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K.-K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J., LIAN, J., MONAHAN, H., O/'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M., WEISSMAN, S. M. & SNYDER, M. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature,* 489**,** 91-100.

GHIRLANDA, S. & VALLORTIGARA, G. 2004. The evolution of brain lateralization: a game-theoretical analysis of population structure. *Proc Biol Sci,* 271**,** 853-7.

GIBSON, G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet,* 13**,** 135-145.

GIBSON, K. R. 1993. The evolution of lateral asymmetries, language, tool-use, and intellect. By John Bradshaw and Lesley Rogers. San Diego: Academic Press, 1992 ISBN 0-12-124560-8. xiii + 463 pp. $72 (cloth). *American Journal of Physical Anthropology,* 92**,** 123-124.

GILGER, J. W., PENNINGTON, B. F., GREEN, P., SMITH, S. M. & SMITH, S. D. 1992. Reading disability, immune disorders and non-right-handedness: twin and family studies of their relations. *Neuropsychologia,* 30**,** 209-27.

GILJOV, A., KARENINA, K., INGRAM, J. & MALASHICHEV, Y. 2015. Parallel Emergence of True Handedness in the Evolution of Marsupials and Placentals. *Curr Biol*.

GILJOV, A., KARENINA, K. & MALASHICHEV, Y. 2013. Forelimb preferences in quadrupedal marsupials and their implications for laterality evolution in mammals. *BMC Evol Biol,* 13**,** 61.

GOEZ, H. & ZELNIK, N. 2008. Handedness in patients with developmental coordination disorder. *J Child Neurol,* 23**,** 151-4.

GONG, C. & MAQUAT, L. E. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature,* 470**,** 284-8.

GONG, J., LIU, W., ZHANG, J., MIAO, X. & GUO, A. Y. 2015. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res,* 43**,** D181-6.

GRANDE, C. & PATEL, N. H. 2009. Nodal signalling is involved in left-right asymmetry in snails. *Nature,* 457**,** 1007-1011.

GRAUR, D., ZHENG, Y., PRICE, N., AZEVEDO, R. B., ZUFALL, R. A. & ELHAIK, E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol,* 5**,** 578-90.

GRAVELEY, B. R., BROOKS, A. N., CARLSON, J. W., DUFF, M. O., LANDOLIN, J. M., YANG, L., ARTIERI, C. G., VAN BAREN, M. J., BOLEY, N., BOOTH, B. W., BROWN, J. B., CHERBAS, L., DAVIS, C. A., DOBIN, A., LI, R., LIN, W., MALONE, J. H., MATTIUZZO, N. R., MILLER, D., STURGILL, D., TUCH, B. B., ZALESKI, C., ZHANG, D., BLANCHETTE, M., DUDOIT, S., EADS, B., GREEN, R. E., HAMMONDS, A., JIANG, L., KAPRANOV, P., LANGTON, L., PERRIMON, N., SANDLER, J. E., WAN, K. H., WILLINGHAM, A., ZHANG, Y., ZOU, Y., ANDREWS, J., BICKEL, P. J., BRENNER, S. E., BRENT, M. R., CHERBAS, P., GINGERAS, T. R., HOSKINS, R. A., KAUFMAN, T. C., OLIVER, B. & CELNIKER, S. E. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature,* 471**,** 473-479.

GREENE, C. S., PENROD, N. M., WILLIAMS, S. M. & MOORE, J. H. 2009. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One,* 4**,** e5639.

GRIMMER, M. R. & WEISS, W. A. 2006. Childhood tumors of the nervous system as disorders of normal development. *Curr Opin Pediatr,* 18**,** 634-8.

GU, X., SHIN, B. H., AKBARALI, Y., WEISS, A., BOLTAX, J., OETTGEN, P. & LIBERMANN, T. A. 2001. Tel-2 is a novel transcriptional repressor related to the Ets factor Tel/ETV-6. *J Biol Chem,* 276**,** 9421-36.

GUHANIYOGI, J. & BREWER, G. 2001. Regulation of mRNA stability in mammalian cells. *Gene,* 265**,** 11-23.

GÜNTÜRKÜN, O., KESCH, S. & DELIUS, J. D. 1988. Absence of footedness in domestic pigeons. *Animal Behaviour,* 36**,** 602-604.

GUPTA, R. A., SHAH, N., WANG, K. C., KIM, J., HORLINGS, H. M., WONG, D. J., TSAI, M. C., HUNG, T., ARGANI, P., RINN, J. L., WANG, Y., BRZOSKA, P., KONG, B., LI, R., WEST, R. B., VAN DE VIJVER, M. J., SUKUMAR, S. & CHANG, H. Y. 2010. Long non-coding RNA

HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature,* 464**,** 1071-6.

GUTTMAN, M., DONAGHEY, J., CAREY, B. W., GARBER, M., GRENIER, J. K., MUNSON, G., YOUNG, G., LUCAS, A. B., ACH, R., BRUHN, L., YANG, X., AMIT, I., MEISSNER, A., REGEV, A., RINN, J. L., ROOT, D. E. & LANDER, E. S. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature,* 477**,** 295-300.

GUVEN, M., ELALMIS, D. D., BINOKAY, S. & TAN, U. 2003. Population-level right-paw preference in rats assessed by a new computerized food-reaching test. *International Journal of Neuroscience,* 113**,** 1675-1689.

HALLE, F., GAHR, M. & KREUTZER, M. 2003. Effects of unilateral lesions of HVC on song patterns of male domesticated canaries. *Journal of Neurobiology,* 56**,** 303-314.

HAMPSON, E. & SANKAR, J. S. 2012. Hand preference in humans is associated with testosterone levels and androgen receptor gene polymorphism. *Neuropsychologia,* 50**,** 2018-25.

HAN, L., VICKERS, K. C., SAMUELS, D. C. & GUO, Y. 2015. Alternative applications for distinct RNA sequencing strategies. *Brief Bioinform,* 16**,** 629-39.

HARRIS, L. J. 1992. Left-handedness. *Handbook of Neuropsychology,* 6**,** 145-208.

HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S., BARNES, I., BIGNELL, A., BOYCHENKO, V., HUNT, T., KAY, M., MUKHERJEE, G., RAJAN, J., DESPACIO-REYES, G., SAUNDERS, G., STEWARD, C., HARTE, R., LIN, M., HOWALD, C., TANZER, A., DERRIEN, T., CHRAST, J., WALTERS, N., BALASUBRAMANIAN, S., PEI, B., TRESS, M., RODRIGUEZ, J. M., EZKURDIA, I., VAN BAREN, J., BRENT, M., HAUSSLER, D., KELLIS, M., VALENCIA, A., REYMOND, A., GERSTEIN, M., GUIGO, R. & HUBBARD, T. J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.

HAUDRY, A., PLATTS, A. E., VELLO, E., HOEN, D. R., LECLERCQ, M., WILLIAMSON, R. J., FORCZEK, E., JOLY-LOPEZ, Z., STEFFEN, J. G., HAZZOURI, K. M., DEWAR, K., STINCHCOMBE, J. R., SCHOEN, D. J., WANG, X., SCHMUTZ, J., TOWN, C. D., EDGER, P. P., PIRES, J. C., SCHUMAKER, K. S., JARVIS, D. E., MANDAKOVA, T., LYSAK, M. A., VAN DEN BERGH, E., SCHRANZ, M. E., HARRISON, P. M., MOSES, A. M., BUREAU, T. E., WRIGHT, S. I. & BLANCHETTE, M. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet,* 45**,** 891-898.

HAWRYLYCZ, M. J., LEIN, E. S., GUILLOZET-BONGAARTS, A. L., SHEN, E. H., NG, L., MILLER, J. A., VAN DE LAGEMAAT, L. N., SMITH, K. A., EBBERT, A., RILEY, Z. L., ABAJIAN, C., BECKMANN, C. F., BERNARD, A., BERTAGNOLLI, D., BOE, A. F., CARTAGENA, P. M., CHAKRAVARTY, M. M., CHAPIN, M., CHONG, J., DALLEY, R. A., DALY, B. D., DANG, C., DATTA, S., DEE, N., DOLBEARE, T. A., FABER, V., FENG, D., FOWLER, D. R., GOLDY, J., GREGOR, B. W., HARADON, Z., HAYNOR, D. R., HOHMANN, J. G., HORVATH, S., HOWARD, R. E., JEROMIN, A., JOCHIM, J. M., KINNUNEN, M., LAU, C., LAZARZ, E. T., LEE, C., LEMON, T. A., LI, L., LI, Y., MORRIS, J. A., OVERLY, C. C., PARKER, P. D., PARRY, S. E., REDING, M., ROYALL, J. J., SCHULKIN, J., SEQUEIRA, P. A., SLAUGHTERBECK, C. R., SMITH, S. C., SODT, A. J., SUNKIN, S. M., SWANSON, B. E., VAWTER, M. P., WILLIAMS, D., WOHNOUTKA, P., ZIELKE, H. R., GESCHWIND, D. H., HOF, P. R., SMITH, S. M., KOCH, C., GRANT, S. G. & JONES, A. R. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature,* 489**,** 391-9.

HEDRICK, P. W. 2005. *Genetics of populations,* Boston, Jones and Bartlett Publishers.

HEDRICK, P. W. 2011. *Genetics of populations,* Sudbury, Mass., Jones and Bartlett Publishers.

HENDERSON, S., SUGDEN, D. & BARNETT, A. 2007. Movement Assessment Battery for Children 2. Kit and Manual. London: Harcourt Assessment/Pearson.

HEPPER, P. G., SHAHIDULLAH, S. & WHITE, R. 1991. Handedness in the human fetus. *Neuropsychologia,* 29**,** 1107-11.

HEPPER, P. G., WELLS, D. L. & LYNCH, C. 2005. Prenatal thumb sucking is related to postnatal handedness. *Neuropsychologia,* 43**,** 313-5.

HIRNSTEIN, M. & HUGDAHL, K. 2014. Excess of non-right-handedness in schizophrenia: meta-analysis of gender effects and potential biases in handedness assessment. *The British Journal of Psychiatry,* 205**,** 260-267.

HIROKAWA, N., TANAKA, Y., OKADA, Y. & TAKEDA, S. 2006. Nodal Flow and the Generation of Left-Right Asymmetry. *Cell,* 125**,** 33-45.

HIROTA, K., MIYOSHI, T., KUGOU, K., HOFFMAN, C. S., SHIBATA, T. & OHTA, K. 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature,* 456**,** 130-4.

HOOK, M. 2004. The Evolution of Lateralized Motor Functions. *In:* ROGERS, L. & KAPLAN, G. (eds.) *Comparative Vertebrate Cognition.* Springer US.

HOOSAIN, R. 1990. Left handedness and handedness switch amongst the Chinese. *Cortex,* 26**,** 451-4.

HOPKINS, W. D. 2006. Comparative and familial analysis of handedness in great apes. *Psychological bulletin,* 132**,** 538.

HOPKINS, W. D. & CANTALUPO, C. 2005. Individual and setting differences in the hand preferences of chimpanzees (*Pan troglodytes*): A critical analysis and some alternative explanations. *Laterality,* 10**,** 65-80.

HOPKINS, W. D., PHILLIPS, K. A., BANIA, A., CALCUTT, S. E., GARDNER, M., RUSSELL, J., SCHAEFFER, J., LONSDORF, E. V., ROSS, S. R. & SCHAPIRO, S. J. 2011. Hand preferences for coordinated bimanual actions in 777 great apes: implications for the evolution of handedness in hominins. *Journal of human evolution,* 60**,** 605-611.

HOPKINS, W. D., STOINSKI, T. S., LUKAS, K. E., ROSS, S. R. & WESLEY, M. J. 2003. Comparative assessment of handedness for a coordinated bimanual task in chimpanzees (*Pan troglodytes*), gorillas (*Gorilla gorilla*) and orangutans (*Pongo pygmaeus*). *J Comp Psychol,* 117**,** 302-8.

HRANILOVIC, D., STEFULJ, J., SCHWAB, S., BORRMANN-HASSENBACH, M., ALBUS, M., JERNEJ, B. & WILDENAUER, D. 2004. Serotonin transporter promoter and intron 2 polymorphisms: relationship between allelic variants and gene expression. *Biological Psychiatry,* 55**,** 1090-1094.

HU, H. Y., HE, L. & KHAITOVICH, P. 2014. Deep sequencing reveals a novel class of bidirectional promoters associated with neuronal genes. *BMC Genomics,* 15**,** 457.

HUANG, D. Y., LIN, Y. T., JAN, P. S., HWANG, Y. C., LIANG, S. T., PENG, Y., HUANG, C. Y., WU, H. C. & LIN, C. T. 2008. Transcription factor SOX-5 enhances nasopharyngeal carcinoma progression by down-regulating SPARC gene expression. *The Journal of Pathology,* 214**,** 445-455.

HUNT, E. R., SHEA-WHELLER, T., ALBERY, G. F., BRIDGER, T. H., GUMN, M. & FRANKS, N. R. 2014. Ants show a leftward turning bias when exploring unknown nest sites. *Biology Letters,* 10.

HUR, E.-M. & ZHOU, F.-Q. 2010. GSK3 signaling in neural development. *Nature reviews. Neuroscience,* 11**,** 539-551.

HYND, G. W., HALL, J., NOVEY, E. S., ELIOPULOS, D., BLACK, K., GONZALEZ, J. J., EDMONDS, J. E., RICCIO, C. & COHEN, M. 1995. Dyslexia and corpus callosum morphology. *Arch Neurol,* 52**,** 32-8.

HYND, G. W., SEMRUD-CLIKEMAN, M., LORYS, A. R., NOVEY, E. S. & ELIOPULOS, D. 1990. Brain morphology in developmental dyslexia and attention deficit disorder/hyperactivity. *Arch Neurol,* 47**,** 919-26.

ILLINGWORTH, S. & BISHOP, D. V. 2009. Atypical cerebral lateralisation in adults with compensated developmental dyslexia demonstrated using functional transcranial Doppler ultrasound. *Brain Lang,* 111**,** 61-5.

ITANI, J., TOKUDA, K., FURUYA, Y., KANO, K. & SHIN, Y. 1963. The social construction of natural troops of Japanese monkeys in takasakiyama. *Primates,* 4**,** 1-42.

IVLIEV, A. E., T HOEN, P. A. C., VAN ROON-MOM, W. M. C., PETERS, D. J. M. & SERGEEVA, M. G. 2012. Exploring the Transcriptome of Ciliated Cells Using In Silico Dissection of Human Tissues. *PLoS ONE,* 7**,** e35618.

IYENGAR, B. R., CHOUDHARY, A., SARANGDHAR, M. A., VENKATESH, K. V., GADGIL, C. J. & PILLAI, B. 2014. Non-coding RNA interact to regulate neuronal development and function. *Frontiers in Cellular Neuroscience,* 8**,** 47.

JIN, T. Q., ITO, Y., LUAN, X. H., DANGARIA, S., WALKER, C., ALLEN, M., KULKARNI, A., GIBSON, C., BRAATZ, R., LIAO, X. B. & DIEKWISCH, T. G. H. 2009. Elongated Polyproline Motifs Facilitate Enamel Evolution through Matrix Subunit Compaction. *Plos Biology,* 7**,** 10.

JOHNSON, A. D., HANDSAKER, R. E., PULIT, S. L., NIZZARI, M. M., O'DONNELL, C. J. & DE BAKKER, P. I. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics,* 24**,** 2938-9.

JOHNSON, J. M., CASTLE, J., GARRETT-ENGELE, P., KAN, Z., LOERCH, P. M., ARMOUR, C. D., SANTOS, R., SCHADT, E. E., STOUGHTON, R. & SHOEMAKER, D. D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science,* 302**,** 2141-4.

JOHNSON, R., RICHTER, N., JAUCH, R., GAUGHWIN, P. M., ZUCCATO, C., CATTANEO, E. & STANTON, L. W. 2010. The Human Accelerated Region 1 noncoding RNA is repressed by REST in Huntington's disease. *Physiol Genomics*.

KANHERE, A., VIIRI, K., ARAUJO, C. C., RASAIYAAH, J., BOUWMAN, R. D., WHYTE, W. A., PEREIRA, C. F., BROOKES, E., WALKER, K., BELL, G. W., POMBO, A., FISHER, A. G., YOUNG, R. A. & JENNER, R. G. 2010. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell,* 38**,** 675-88.

KANITZ, A., GYPAS, F., GRUBER, A. J., GRUBER, A. R., MARTIN, G. & ZAVOLAN, M. 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol,* 16**,** 150.

KATAYAMA, S., TOMARU, Y., KASUKAWA, T., WAKI, K., NAKANISHI, M., NAKAMURA, M., NISHIDA, H., YAP, C. C., SUZUKI, M., KAWAI, J., SUZUKI, H., CARNINCI, P., HAYASHIZAKI, Y., WELLS, C., FRITH, M., RAVASI, T., PANG, K. C., HALLINAN, J., MATTICK, J., HUME, D. A., LIPOVICH, L., BATALOV, S., ENGSTROM, P. G., MIZUNO, Y., FAGHIHI, M. A., SANDELIN, A., CHALK, A. M., MOTTAGUI-TABAR, S., LIANG, Z., LENHARD, B. & WAHLESTEDT, C. 2005. Antisense transcription in the mammalian transcriptome. *Science,* 309**,** 1564-6.

KELLER, A., NESVIZHSKII, A. I., KOLKER, E. & AEBERSOLD, R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem,* 74**,** 5383-92.

KIGHT, S. L., STEELMAN, L., COFFEY, G., LUCENTE, J. & CASTILLO, M. 2008. Evidence of population-level lateralized behaviour in giant water bugs, *Belostoma flumineum Say* (Heteroptera: Belostomatidae): T-maze turning is left biased. *Behavioural Processes,* 79**,** 66-69.

KIMURA, K., WAKAMATSU, A., SUZUKI, Y., OTA, T., NISHIKAWA, T., YAMASHITA, R., YAMAMOTO, J., SEKINE, M., TSURITANI, K., WAKAGURI, H., ISHII, S., SUGIYAMA, T., SAITO, K., ISONO, Y., IRIE, R., KUSHIDA, N., YONEYAMA, T., OTSUKA, R., KANDA, K., YOKOI, T., KONDO, H., WAGATSUMA, M., MURAKAWA, K., ISHIDA, S., ISHIBASHI, T., TAKAHASHI-FUJII, A., TANASE, T., NAGAI, K., KIKUCHI, H., NAKAI, K., ISOGAI, T. & SUGANO, S. 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res,* 16**,** 55-65.

KIPPNER, L. E., KIM, J., GIBSON, G. & KEMP, M. L. 2014. Single cell transcriptional analysis reveals novel innate immune cell types. *PeerJ,* 2**,** e452.

KLAR, A. A single locus, RGHT, specifies preference for hand utilization in humans. Cold Spring Harbor Symposia on Quantitative Biology, 1996. Cold Spring Harbor Laboratory Press, 59-65.

KLAR, A. J. 2003. Human handedness and scalp hair-whorl direction develop from a common genetic mechanism. *Genetics,* 165**,** 269-76.

KNECHT, S., DRAGER, B., DEPPE, M., BOBE, L., LOHMANN, H., FLOEL, A., RINGELSTEIN, E. B. & HENNINGSEN, H. 2000. Handedness and hemispheric language dominance in healthy humans. *Brain,* 123**,** 2512-2518.

KOUFAKI, A., & PAPADATOU-PASTOU, M. Μετα-ανάλυση: Δυσλεξία και προτίμηση χεριού (in Greek) [Meta-analysis: dyslexia and hand preference]. Proceedings of the 2nd Hellenic Conference of Educational Sciences, May 27-30, 2010 2010 Athens.

KUMAR, V., WESTRA, H.-J., KARJALAINEN, J., ZHERNAKOVA, D. V., ESKO, T., HRDLICKOVA, B., ALMEIDA, R., ZHERNAKOVA, A., REINMAA, E., VÕSA, U., HOFKER, M. H., FEHRMANN, R. S. N., FU, J., WITHOFF, S., METSPALU, A., FRANKE, L. & WIJMENGA, C. 2013. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genet,* 9**,** e1003201.

KWAN, K. Y., LAM, M. M., KRSNIK, Z., KAWASAWA, Y. I., LEFEBVRE, V. & SESTAN, N. 2008. SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc Natl Acad Sci U S A,* 105**,** 16021-6.

LAI, C. S., FISHER, S. E., HURST, J. A., VARGHA-KHADEM, F. & MONACO, A. P. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature,* 413**,** 519-23.

LASKA, M. 1996. Manual laterality in spider monkeys (*Ateles geoffroyi*) solving visually and tactually guided food-reaching tasks. *Cortex,* 32**,** 717-726.

LAVAL, S. H., DANN, J. C., BUTLER, R. J., LOFTUS, J., RUE, J., LEASK, S. J., BASS, N., COMAZZI, M., VITA, A., NANKO, S., SHAW, S., PETERSON, P., SHIELDS, G., SMITH, A. B., STEWART, J., DELISI, L. E. & CROW, T. J. 1998. Evidence for linkage to psychosis and cerebral asymmetry (relative hand skill) on the X chromosome. *Am J Med Genet,* 81**,** 420-7.

LAYNE, J. D. J. 2013. Novel insights into the function and regulation of group X secretory phospholipase A2. *Theses and Dissertations--Nutritional Sciences. Paper 10*.

LEASK, S. J. & CROW, T. J. 2001. Word acquisition reflects lateralization of hand skill. *Trends Cogn Sci,* 5**,** 513-516.

LEDUC, R., MOLLOY, S. S., THORNE, B. A. & THOMAS, G. 1992. ACTIVATION OF HUMAN FURIN PRECURSOR PROCESSING ENDOPROTEASE OCCURS BY AN INTRAMOLECULAR AUTOPROTEOLYTIC CLEAVAGE. *Journal of Biological Chemistry,* 267**,** 14304-14308.

LEGENDRE, M., POCHET, N., PAK, T. & VERSTREPEN, K. J. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research,* 17**,** 1787-1796.

LEONARD, C. M., LOMBARDINO, L. J., WALSH, K., ECKERT, M. A., MOCKLER, J. L., ROWE, L. A., WILLIAMS, S. & DEBOSE, C. B. 2002. Anatomical risk factors that distinguish dyslexia from SLI predict reading skill in normal children. *J Commun Disord,* 35**,** 501-31.

LEVIN, M. 2005. Left-right asymmetry in embryonic development: a comprehensive review. *Mech Dev,* 122**,** 3-25.

LI, L., WANG, X., STOLC, V., LI, X., ZHANG, D., SU, N., TONGPRASIT, W., LI, S., CHENG, Z., WANG, J. & DENG, X. W. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet,* 38**,** 124-9.

LI, Y. C., KOROL, A. B., FAHIMA, T., BEILES, A. & NEVO, E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology,* 11**,** 2453-2465.

LING, K. H., HEWITT, C. A., BEISSBARTH, T., HYDE, L., CHEAH, P. S., SMYTH, G. K., TAN, S. S., HAHN, C. N., THOMAS, T., THOMAS, P. Q. & SCOTT, H. S. 2011. Spatiotemporal regulation of multiple overlapping sense and novel natural antisense transcripts at the Nrgn and Camk2n1 gene loci during mouse cerebral corticogenesis. *Cereb Cortex,* 21**,** 683-97.

LIPOVICH, L., DACHET, F., CAI, J., BAGLA, S., BALAN, K., JIA, H. & LOEB, J. A. 2012. Activity-Dependent Human Brain Coding/Noncoding Gene Regulatory Networks. *Genetics,* 192**,** 1133-1148.

LLORENTE, M., RIBA, D., PALOU, L., CARRASCO, L., MOSQUERA, M., COLELL, M. & FELIU, O. 2011. Population-Level Right-Handedness for a Coordinated Bimanual Task in Naturalistic Housed Chimpanzees: Replication and Extension in 114 Animals From Zambia and Spain. *American Journal of Primatology,* 73**,** 281-290.

LORENZ, R., BERNHART, S. H., HONER ZU SIEDERDISSEN, C., TAFER, H., FLAMM, C., STADLER, P. F. & HOFACKER, I. L. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol,* 6**,** 26.

LOVESTONE, S., KILLICK, R., DI FORTI, M. & MURRAY, R. 2007. Schizophrenia as a GSK-3 dysregulation disorder. *Trends Neurosci,* 30**,** 142-9.

LOZANO-RUIZ, M., BERMÚDEZ DE CASTRO, J. M., MARTINÓN-TORRES, M. & SARMIENTO, S. 2004. Cutmarks on fossil human anterior teeth of the Sima de los Huesos Site (Atapuerca, Spain). *Journal of Archaeological Science,* 31**,** 1127-1135.

MACKAY, T. F. 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics,* 15**,** 22-33.

MACNEILAGE, P. F., ROGERS, L. J. & VALLORTIGARA, G. 2009. Origins of the left & right brain. *Sci Am,* 301**,** 60-7.

MACNEILAGE, P. F., STUDDERT-KENNEDY, M. G. & LINDBLOM, B. 1987. Primate handedness reconsidered. *Behavioral and Brain Sciences,* 10**,** 247-263.

MAISOG, J. M., EINBINDER, E. R., FLOWERS, D. L., TURKELTAUB, P. E. & EDEN, G. F. 2008. A Meta-analysis of Functional Neuroimaging Studies of Dyslexia. *Annals of the New York Academy of Sciences,* 1145**,** 237-259.

MANDAL, M. K., BULMAN-FLEMING, M. B. & TIWARI, G. 2000. *Side bias : a neuropsychological perspective,* Dordrecht ; Boston, Kluwer Academic Publishers.

MARSCHIK, P. B., EINSPIELER, C., STROHMEIER, A., PLIENEGGER, J., GARZAROLLI, B. & PRECHTL, H. F. R. 2008. From the reaching behavior at 5 months of age to hand preference at preschool age. *Developmental Psychobiology,* 50**,** 511-518.

MASSINEN, S., HOKKANEN, M.-E., MATSSON, H., TAMMIMIES, K., TAPIA-PÁEZ, I., DAHLSTRÖM-HEUSER, V., KUJA-PANULA, J., BURGHOORN, J., JEPPSSON, K. E., SWOBODA, P., PEYRARD-JANVID, M., TOFTGÅRD, R., CASTRÉN, E. & KERE, J. 2011. Increased Expression of the Dyslexia Candidate Gene DCDC2 Affects Length and Signaling of Primary Cilia in Neurons. *PLoS ONE,* 6**,** e20580.

MATYS, V., KEL-MARGOULIS, O. V., FRICKE, E., LIEBICH, I., LAND, S., BARRE-DIRRIE, A., REUTER, I., CHEKMENEV, D., KRULL, M., HORNISCHER, K., VOSS, N., STEGMAIER, P., LEWICKI-POTAPOV, B., SAXEL, H., KEL, A. E. & WINGENDER, E. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res,* 34**,** D108-10.

MAZOYER, B., ZAGO, L., JOBARD, G., CRIVELLO, F., JOLIOT, M., PERCHEY, G., MELLET, E., PETIT, L. & TZOURIO-MAZOYER, N. 2014. Gaussian mixture modeling of hemispheric lateralization for language in a large sample of healthy individuals balanced for handedness. *PLoS One,* 9**,** e101165.

MCGREW, W. C. & MARCHANT, L. F. 1997. On the other hand: Current issues in and meta-analysis of the behavioral laterality of hand function in nonhuman primates. *American Journal of Physical Anthropology,* 104**,** 201-232.

MCKEEVER, W. F. 2004. An X-linked three allele model of hand preference and hand posture for writing. *Laterality,* 9**,** 149-73.

MCMANUS, I. 1985a. Handedness, language dominance and aphasia: a genetic model. *Psychological medicine. Monograph supplement,* 8**,** 3-40.

MCMANUS, I. C. 1985b. Right- and left-hand skill: failure of the right shift model. *The British journal of psychology,* 76**,** Pt 1/.

MCMANUS, I. C., DAVISON, A. & ARMOUR, J. A. L. 2013. Multilocus genetic models of handedness closely resemble single-locus models in explaining family data and are compatible with genome-wide association studies. *Annals of the New York Academy of Sciences,* 1288**,** 48-58.

MCMANUS, I. C., MARTIN, N., STUBBINGS, G. F., CHUNG, E. M. & MITCHISON, H. M. 2004. Handedness and situs inversus in primary ciliary dyskinesia. *Proc Biol Sci,* 271**,** 2579-82.

MEDLAND, S., LINDGREN, C., MAGI, R., NEALE, B., ALBRECHT, E., ESKO, T., EVANS, D., HOTTENGA, J., IKRAM, M. & MANGINO, M. Meta-analysis of GWAS for handedness: results from the ENGAGE consortium. American Society of Human Genetics, Meeting Abstract, 2009a.

MEDLAND, S. E., DUFFY, D. L., SPURDLE, A. B., WRIGHT, M. J., GEFFEN, G. M., MONTGOMERY, G. W. & MARTIN, N. G. 2005. Opposite effects of androgen receptor CAG repeat length on increased risk of left-handedness in males and females. *Behav Genet,* 35**,** 735-44.

MEDLAND, S. E., DUFFY, D. L., WRIGHT, M. J., GEFFEN, G. M., HAY, D. A., LEVY, F., VAN-BEIJSTERVELDT, C. E., WILLEMSEN, G., TOWNSEND, G. C., WHITE, V., HEWITT, A. W., MACKEY, D. A., BAILEY, J. M., SLUTSKE, W. S., NYHOLT, D. R., TRELOAR, S. A., MARTIN, N. G. & BOOMSMA, D. I. 2009b. Genetic influences on handedness: data from 25,732 Australian and Dutch twin families. *Neuropsychologia,* 47**,** 330-7.

MEGUERDITCHIAN, A. & VAUCLAIR, J. 2006. Baboons communicate with their right hand. *Behavioural Brain Research,* 171**,** 170-174.

MEGUERDITCHIAN, A., VAUCLAIR, J. & HOPKINS, W. D. 2013. On the origins of human handedness and language: A comparative review of hand preferences for bimanual coordinated actions and gestural communication in nonhuman primates. *Developmental Psychobiology,* 55**,** 637-650.

MELAMUD, E. & MOULT, J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Research,* 37**,** 4873-4886.

MERCER, T. R. & MATTICK, J. S. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol,* 20**,** 300-7.

MEUNIER, H., FAGARD, J., MAUGARD, A., BRISEÑO, M., FIZET, J., CANTELOUP, C., DEFOLIE, C. & VAUCLAIR, J. 2013. Patterns of hemispheric specialization for a communicative gesture in different primate species. *Developmental Psychobiology,* 55**,** 662-671.

MICHEL, G. F. 1981. Right-handedness: a consequence of infant supine head-orientation preference? *Science,* 212**,** 685-7.

MICHEL, G. F. & GOODWIN, R. 1979. Intrauterine birth position predicts newborn supine head position preferences. *Infant Behavior and Development,* 2**,** 29-38.

MICHEL, G. F., TYLER, A. N., FERRE, C. & SHEU, C. F. 2006. The manifestation of infant hand-use preferences when reaching for objects during the seven-to thirteen-month age period. *Dev Psychobiol,* 48**,** 436-43.

MIRZA, A. H., KAUR, S., BRORSSON, C. A. & POCIOT, F. 2014. Effects of GWAS-Associated Genetic Variants on lncRNAs within IBD and T1D Candidate Loci. *PLoS ONE,* 9**,** e105723.

MITCHELL, A., CHANG, H. Y., DAUGHERTY, L., FRASER, M., HUNTER, S., LOPEZ, R., MCANULLA, C., MCMENAMIN, C., NUKA, G., PESSEAT, S., SANGRADOR-VEGAS, A., SCHEREMETJEW, M., RATO, C., YONG, S. Y., BATEMAN, A., PUNTA, M., ATTWOOD, T. K., SIGRIST, C. J., REDASCHI, N., RIVOIRE, C., XENARIOS, I., KAHN, D., GUYOT, D., BORK, P., LETUNIC, I., GOUGH, J., OATES, M., HAFT, D., HUANG, H., NATALE, D. A., WU, C. H., ORENGO, C., SILLITOE, I., MI, H., THOMAS, P. D. & FINN, R. D. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res,* 43**,** D213-21.

MIZUTA, I., TAKAFUJI, K., ANDO, Y., SATAKE, W., KANAGAWA, M., KOBAYASHI, K., NAGAMORI, S., SHINOHARA, T., ITO, C., YAMAMOTO, M., HATTORI, N., MURATA, M., KANAI, Y., MURAYAMA, S., NAKAGAWA, M. & TODA, T. 2013. YY1 binds to alpha-synuclein 3'-flanking region SNP and stimulates antisense noncoding RNA expression. J Hum Genet, 58, 711-9.

MOFFATT, M. F., KABESCH, M., LIANG, L., DIXON, A. L., STRACHAN, D., HEATH, S., DEPNER, M., VON BERG, A., BUFE, A., RIETSCHEL, E., HEINZMANN, A., SIMMA, B., FRISCHER, T., WILLIS-OWEN, S. A. G., WONG, K. C. C., ILLIG, T., VOGELBERG, C., WEILAND, S. K., VON MUTIUS, E., ABECASIS, G. R., FARRALL, M., GUT, I. G., LATHROP, G. M. & COOKSON, W. O. C. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature,* 448**,** 470-473.

MORAIS, V. A., VERSTREKEN, P., ROETHIG, A., SMET, J., SNELLINX, A., VANBRABANT, M., HADDAD, D., FREZZA, C., MANDEMAKERS, W., VOGT-WEISENHORN, D., VAN COSTER, R., WURST, W., SCORRANO, L. & DE STROOPER, B. 2009. Parkinson's disease mutations in PINK1 result in decreased Complex I activity and deficient synaptic function. *EMBO Mol Med,* 1**,** 99-111.

MORGAN, M. J. & CORBALLIS, M. C. 1978. On the biological basis of human laterality: II. The mechanisms of inheritance. *Behavioral and Brain Sciences,* 1**,** 270-277.

MORRISSY, A. S., GRIFFITH, M. & MARRA, M. A. 2011. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res,* 21**,** 1203-12.

NAGASAKI, H., ARITA, M., NISHIZAWA, T., SUWA, M. & GOTOH, O. 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene,* 364**,** 53-62.

NESVIZHSKII, A. I., KELLER, A., KOLKER, E. & AEBERSOLD, R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem,* 75**,** 4646-58.

NICHOLLS, M. E., CHAPMAN, H. L., LOETSCHER, T. & GRIMSHAW, G. M. 2010. The relationship between hand preference, hand performance, and general cognitive ability. *J Int Neuropsychol Soc,* 16**,** 585-92.

NICHOLLS, M. E., ORR, C. A. & LINDELL, A. K. 2005. Magical ideation and its relation to lateral preference. *Laterality,* 10**,** 503-15.

O'RAHILLY, R. & MULLER, F. 2010. Developmental stages in human embryos: revised and new measurements. *Cells Tissues Organs,* 192**,** 73-84.

OCKLENBURG, S., BESTE, C. & ARNING, L. 2014. Handedness genetics: considering the phenotype. *Frontiers in Psychology,* 5**,** 1300.

OCKLENBURG, S. & GUNTURKUN, O. 2012. Hemispheric asymmetries: the comparative view. *Front Psychol,* 3**,** 5.

ODDY, H. C. & LOBSTEIN, T. J. 1972. Hand and eye dominance in schizophrenia. *Br J Psychiatry,* 120**,** 331-2.

OLDFIELD, R. C. 1971. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia,* 9**,** 97-113.

OLSON, R., WISE, B., CONNERS, F., RACK, J. & FULKER, D. 1989. Specific deficits in component reading and language skills: genetic and environmental influences. *J Learn Disabil,* 22**,** 339-48.

ONG, S. E. & MANN, M. 2006. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc,* 1**,** 2650-60.

PAL, L. R., YU, C.-H., MOUNT, S. M. & MOULT, J. 2015. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics,* 16**,** S4-S4.

PAL, S., GUPTA, R., KIM, H., WICKRAMASINGHE, P., BAUBET, V., SHOWE, L. C., DAHMANE, N. & DAVULURI, R. V. 2011. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res,* 21**,** 1260-72.

PALMER, A. R. 2002. Chimpanzee right-handedness reconsidered: Evaluating the evidence with funnel plots. *American Journal of Physical Anthropology,* 118**,** 191-199.

PALMER, A. R. 2012. Developmental origins of normal and anomalous random right-left asymmetry: lateral inhibition versus developmental error in a threshold trait. *Contributions to Zoology,* 81**,** 111-124.

PANG, K. C., FRITH, M. C. & MATTICK, J. S. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet,* 22**,** 1-5.

PAPADATOU-PASTOU, M., MARTIN, M., MUNAFO, M. R. & JONES, G. V. 2008. Sex differences in left-handedness: a meta-analysis of 144 studies. *Psychol Bull,* 134**,** 677-99.

PAPADEMETRIOU, E., SHEU, C. F. & MICHEL, G. F. 2005. A meta-analysis of primate hand preferences, particularly for reaching. *J Comp Psychol,* 119**,** 33-48.

PARKHOMCHUK, D., BORODINA, T., AMSTISLAVSKIY, V., BANARU, M., HALLEN, L., KROBITSCH, S., LEHRACH, H. & SOLDATOV, A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res,* 37**,** e123.

PARNELL, R. J. 2001. Hand preference for food processing in wild western lowland gorillas (*Gorilla gorilla gorilla*). *J Comp Psychol,* 115**,** 365-75.

PATTERSON, N., RICHTER, D. J., GNERRE, S., LANDER, E. S. & REICH, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature,* 441**,** 1103-8.

PAUL, L. K. 2011. Developmental malformation of the corpus callosum: a review of typical callosal development and examples of developmental disorders with callosal involvement. *Journal of neurodevelopmental disorders,* 3**,** 3-27.

PELECHANO, V. & STEINMETZ, L. M. 2013. Gene regulation by antisense transcription. *Nat Rev Genet,* 14**,** 880-893.

PENNINGTON, B. F. & BISHOP, D. V. 2009. Relations among speech, language, and reading disorders. *Annu Rev Psychol,* 60**,** 283-306.

PENNINGTON, B. F., SMITH, S. D., KIMBERLING, W. J., GREEN, P. A. & HAITH, M. M. 1987. Left-handedness and immune disorders in familial dyslexics. *Arch Neurol,* 44**,** 634-9.

PERELLE, I. B. & EHRMAN, L. 1994. An international study of human handedness: the data. *Behav Genet,* 24**,** 217-27.

PERKINS, D. N., PAPPIN, D. J., CREASY, D. M. & COTTRELL, J. S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20**,** 3551-67.

PETRILL, S. A., DEATER-DECKARD, K., THOMPSON, L. A., SCHATSCHNEIDER, C., DETHORNE, L. S. & VANDENBERGH, D. J. 2007. Longitudinal genetic analysis of early reading: The Western Reserve Reading Project. *Read Writ,* 20**,** 127-146.

PHILLIPSON, L. 1997. EDGE MODIFICATION AS AN INDICATOR OF FUNCTION AND HANDEDNESS OF ACHEULIAN HANDAXES FROM KARIANDUSI, KENYA. *Lithic Technology,* 22**,** 171-183.

PILS, B. & SCHULTZ, J. 2004. Inactive enzyme-homologues find new function in regulatory processes. *J Mol Biol,* 340**,** 399-404.

PIONTKIVSKA, H., YANG, M. Q., LARKIN, D. M., LEWIN, H. A., REECY, J. & ELNITSKI, L. 2009. Cross-species mapping of bidirectional promoters enables prediction of unannotated 5' UTRs and identification of species-specific transcripts. *BMC Genomics,* 10**,** 189.

POLLARD, K. S., SALAMA, S. R., LAMBERT, N., LAMBOT, M. A., COPPENS, S., PEDERSEN, J. S., KATZMAN, S., KING, B., ONODERA, C., SIEPEL, A., KERN, A. D., DEHAY, C., IGEL, H., ARES, M., JR., VANDERHAEGHEN, P. & HAUSSLER, D. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature,* 443**,** 167-72.

POLLICK, A. S. & DE WAAL, F. B. 2007. Ape gestures and language evolution. *Proc Natl Acad Sci U S A,* 104**,** 8184-9.

PONJAVIC, J., OLIVER, P. L., LUNTER, G. & PONTING, C. P. 2009. Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. *PLoS Genet,* 5**,** e1000617.

PONTING, C. P., OLIVER, P. L. & REIK, W. 2009. Evolution and functions of long noncoding RNAs. *Cell,* 136**,** 629-41.

PRICHARD, E., PROPPER, R. E. & CHRISTMAN, S. D. 2013. Degree of Handedness, but not Direction, is a Systematic Predictor of Cognitive Performance. *Front Psychol,* 4**,** 9.

PROPPER, R., CHRISTMAN, S. & PHANEUF, K. 2005. A mixed-handed advantage in episodic memory: A possible role of interhemispheric interaction. *Memory & Cognition,* 33**,** 751-757.

PROVINS, K. A. 1997. Handedness and speech: a critical reappraisal of the role of genetic and environmental factors in the cerebral lateralization of function. *Psychol Rev,* 104**,** 554-71.

PROVINS, K. A., MILNER, A. D. & KERR, P. 1982. ASYMMETRY OF MANUAL PREFERENCE AND PERFORMANCE. *Perceptual and Motor Skills,* 54**,** 179-194.

PRUITT, K. D., BROWN, G. R., HIATT, S. M., THIBAUD-NISSEN, F., ASTASHYN, A., ERMOLAEVA, O., FARRELL, C. M., HART, J., LANDRUM, M. J., MCGARVEY, K. M., MURPHY, M. R., O'LEARY, N. A., PUJAR, S., RAJPUT, B., RANGWALA, S. H., RIDDICK, L. D., SHKEDA, A., SUN, H., TAMEZ, P., TULLY, R. E., WALLIN, C., WEBB, D., WEBER, J., WU, W., DICUCCIO, M., KITTS, P., MAGLOTT, D. R., MURPHY, T. D. & OSTELL, J. M. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res,* 42**,** D756-63.

PRUUNSILD, P., KAZANTSEVA, A., AID, T., PALM, K. & TIMMUSK, T. 2007. Dissecting the human BDNF locus: bidirectional transcription, complex splicing, and multiple promoters. *Genomics,* 90**,** 397-406.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

RAMOS, A. D., DIAZ, A., NELLORE, A., DELGADO, R. N., PARK, K. Y., GONZALES-ROYBAL, G., OLDHAM, M. C., SONG, J. S. & LIM, D. A. 2013. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell,* 12**,** 616-28.

RAMSKÖLD, D., KAVAK, E. & SANDBERG, R. 2012. How to Analyze Gene Expression Using RNA-Sequencing Data. *In:* WANG, J., TAN, A. C. & TIAN, T. (eds.) *Next Generation Microarray Bioinformatics.* Humana Press.

RAMSKOLD, D., WANG, E. T., BURGE, C. B. & SANDBERG, R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol,* 5**,** e1000598.

RAMUS, F. 2014. Neuroimaging sheds new light on the phonological deficit in dyslexia. *Trends Cogn Sci,* 18**,** 274-5.

RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc,* 8**,** 2281-308.

RANDLER, C. 2007. Foot preferences during resting in wildfowl and waders. *Laterality,* 12**,** 191-197.

RASMUSSEN, T. & MILNER, B. 1977. The role of early left-brain injury in determining lateralization of cerebral speech functions. *Ann N Y Acad Sci,* 299**,** 355-69.

RAYMOND, M., PONTIER, D., DUFOUR, A.-B. & MOLLER, A. P. 1996. *Frequency-Dependent Maintenance of Left Handedness in Humans*.

REICH, D. E. & LANDER, E. S. 2001. On the allelic spectrum of human disease. *Trends Genet,* 17**,** 502-10.

REVZIN, A. 1989. Gel electrophoresis assays for DNA-protein interactions. *Biotechniques,* 7**,** 346-55.

RHIE, S. K., COETZEE, S. G., NOUSHMEHR, H., YAN, C., KIM, J. M., HAIMAN, C. A. & COETZEE, G. A. 2013. Comprehensive functional annotation of seventy-one breast cancer risk Loci. *PLoS One,* 8**,** e63925.

RICHLAN, F., KRONBICHLER, M. & WIMMER, H. 2009. Functional abnormalities in the dyslexic brain: a quantitative meta-analysis of neuroimaging studies. *Hum Brain Mapp,* 30**,** 3299-308.

RINN, J. L. & CHANG, H. Y. 2012. Genome regulation by long noncoding RNAs. *Annual review of biochemistry,* 81**,** 10.1146/annurev-biochem-051410-092902.

RIVAS, M. A., BEAUDOIN, M., GARDET, A., STEVENS, C., SHARMA, Y., ZHANG, C. K., BOUCHER, G., RIPKE, S., ELLINGHAUS, D. & BURTT, N. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics,* 43**,** 1066-1073.

ROBINS, A. & ROGERS, L. J. 2002. Limb preference and skeletal asymmetry in the cane toad, *Bufo marinus* (Anura : Bufonidae). *Laterality,* 7**,** 261-275.

ROGERS, L. J. 2008. Development and function of lateralization in the avian brain. *Brain Res Bull,* 76**,** 235-44.

ROGERS, L. J. & ANDREW, R. J. 2002. *Comparative vertebrate lateralization,* Cambridge ; New York, Cambridge University Press.

ROGERS, L. J., VALLORTIGARA, G. & ANDREW, R. J. 2013. *Divided brains: the biology and behaviour of brain asymmetries*, Cambridge University Press.

ROGERS, L. J. & WORKMAN, L. 1993. Footedness in birds. *Animal Behaviour,* 45**,** 409-411.

RORDEN, C., DAVIS, B., GEORGE, M. S., BORCKARDT, J. & FRIDRIKSSON, J. 2008. Broca's area is crucial for visual discrimination of speech but not non-speech oral movements. *Brain stimulation,* 1**,** 383-385.

SAINT PIERRE, A. & GÉNIN, E. 2014. How important are rare variants in common disease? *Briefings in Functional Genomics*.

SAKAI, M., HISHII, T., TAKEDA, S. & KOHSHIMA, S. 2006. Laterality of flipper rubbing behaviour in wild bottlenose dolphins (*Tursiops aduncus*): Caused by asymmetry of eye use? *Behavioural Brain Research,* 170**,** 204-210.

SALTA, E. & DE STROOPER, B. 2012. Non-coding RNAs with essential roles in neurodegenerative disorders. *Lancet Neurol,* 11**,** 189-200.

SANNA, S., JACKSON, A. U., NAGARAJA, R., WILLER, C. J., CHEN, W.-M., BONNYCASTLE, L. L., SHEN, H., TIMPSON, N., LETTRE, G. & USALA, G. 2008. Common variants in the GDF5-

UQCC region are associated with variation in human height. *Nature genetics,* 40**,** 198-203.

SATZ, P. & GREEN, M. F. 1999. Atypical handedness in schizophrenia: Some methodological and theoretical issues. *Schizophrenia Bulletin,* 25**,** 63-78.

SAVITZ, J., VAN DER MERWE, L., SOLMS, M. & RAMESAR, R. 2007. Lateralization of hand skill in bipolar affective disorder. *Genes Brain Behav,* 6**,** 698-705.

SCERRI, T. S., BRANDLER, W. M., PARACCHINI, S., MORRIS, A. P., RING, S. M., RICHARDSON, A. J., TALCOTT, J. B., STEIN, J. & MONACO, A. P. 2011a. PCSK6 is associated with handedness in individuals with dyslexia. *Human Molecular Genetics,* 20**,** 608-614.

SCERRI, T. S., MORRIS, A. P., BUCKINGHAM, L. L., NEWBURY, D. F., MILLER, L. L., MONACO, A. P., BISHOP, D. V. & PARACCHINI, S. 2011b. DCDC2, KIAA0319 and CMIP are associated with reading-related traits. *Biol Psychiatry,* 70**,** 237-45.

SCHIER, A. F. & SHEN, M. M. 2000. Nodal signalling in vertebrate development. *Nature,* 403**,** 385-389.

SCHIFFMAN, J., PESTLE, S., MEDNICK, S., EKSTROM, M., SORENSEN, H. & MEDNICK, S. 2005. Childhood laterality and adult schizophrenia spectrum disorders: a prospective investigation. *Schizophr Res,* 72**,** 151-60.

SCHUMACHER, J., ANTHONI, H., DAHDOUH, F., KONIG, I. R., HILLMER, A. M., KLUCK, N., MANTHEY, M., PLUME, E., WARNKE, A., REMSCHMIDT, H., HULSMANN, J., CICHON, S., LINDGREN, C. M., PROPPING, P., ZUCCHELLI, M., ZIEGLER, A., PEYRARD-JANVID, M., SCHULTE-KORNE, G., NOTHEN, M. M. & KERE, J. 2006. Strong genetic evidence of DCDC2 as a susceptibility gene for dyslexia. *Am J Hum Genet,* 78**,** 52-62.

SEITZ, A., GOUREVITCH, D., ZHANG, X. M., CLARK, L., CHEN, P., KRAGOL, M., LEVENKOVA, N., RUX, J., SAMULEWICZ, S. & HEBER-KATZ, E. 2005. Sense and antisense transcripts of the apolipoprotein E gene in normal and ApoE knockout mice, their expression after spinal cord injury and corresponding human transcripts. *Hum Mol Genet,* 14**,** 2661-70.

SHEN, M. M. 2007. Nodal signaling: developmental roles and regulation. *Development,* 134**,** 1023-34.

SICOTTE, N. L., WOODS, R. P. & MAZZIOTTA, J. C. 1999. Handedness in twins: a meta-analysis. *Laterality,* 4**,** 265-86.

SIEG, A. E., ZANDONA, E., IZZO, V. M., PALADINO, F. V. & SPOTILA, J. R. 2010. Population level "flipperedness" in the eastern Pacific leatherback turtle. *Behav Brain Res,* 206**,** 135-8.

SIGOVA, A. A., MULLEN, A. C., MOLINIE, B., GUPTA, S., ORLANDO, D. A., GUENTHER, M. G., ALMADA, A. E., LIN, C., SHARP, P. A., GIALLOURAKIS, C. C. & YOUNG, R. A. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A,* 110**,** 2876-81.

SILLÉ, F. C. M., THOMAS, R., SMITH, M. T., CONDE, L. & SKIBOLA, C. F. 2012. Post-GWAS Functional Characterization of Susceptibility Variants for Chronic Lymphocytic Leukemia. *PLoS ONE,* 7**,** e29632.

SMIGIELSKI, E. M., SIROTKIN, K., WARD, M. & SHERRY, S. T. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res,* 28**,** 352-5.

SNOWLING, M., BISHOP, D. V. & STOTHARD, S. E. 2000. Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry,* 41**,** 587-600.

SOMMER, I., ALEMAN, A., RAMSEY, N., BOUMA, A. & KAHN, R. 2001. Handedness, language lateralisation and anatomical asymmetry in schizophrenia - Meta-analysis. *British Journal of Psychiatry,* 178**,** 344-351.

SOTOZAKI, H. & PARLOW, S. 2006. Interhemispheric communication involving multiple tasks: A study of children with dyslexia. *Brain Lang,* 98**,** 89-101.

STANCHER, G., CLARA, E., REGOLIN, L. & VALLORTIGARA, G. 2006. Lateralized righting behavior in the tortoise (*Testudo hermanni*). *Behav Brain Res,* 173**,** 315-9.

STEENHUIS, R. E., BRYDEN, M. P., SCHWARTZ, M. & LAWSON, S. 1990. Reliability of hand preference items and factors. *J Clin Exp Neuropsychol,* 12**,** 921-30.

STOCK, J. T., SHIRLEY, M. K., SARRINGHAUS, L. A., DAVIES, T. G. & SHAW, C. N. 2013. Skeletal evidence for variable patterns of handedness in chimpanzees, human hunter-gatherers, and recent British populations. *Ann N Y Acad Sci,* 1288**,** 86-99.

STRÖCKENS, F., GÜNTÜRKÜN, O. & OCKLENBURG, S. 2013. Limb preferences in non-human vertebrates. *Laterality: Asymmetries of Body, Brain and Cognition,* 18**,** 536-575.

SU, W. Y., LI, J. T., CUI, Y., HONG, J., DU, W., WANG, Y. C., LIN, Y. W., XIONG, H., WANG, J. L., KONG, X., GAO, Q. Y., WEI, L. P. & FANG, J. Y. 2012. Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. *Cell Res,* 22**,** 1374-89.

SUN, T. & WALSH, C. A. 2006. Molecular approaches to brain asymmetry and handedness. *Nat Rev Neurosci,* 7**,** 655-62.

TAN, U. & TAN, M. 1999. Incidences of asymmetries for the palmar grasp reflex in neonates and hand preference in adults. *Neuroreport,* 10**,** 3253-6.

TAPLEY, S. M. & BRYDEN, M. P. 1985. A group test for the assessment of performance between the hands. *Neuropsychologia,* 23**,** 215-221.

TAYLOR, P. J., DALTON, R. & FLEMINGER, J. J. 1980. Handedness in schizophrenia. *Br J Psychiatry,* 136**,** 375-83.

TEAM, A. S. 2001. ALSPAC–the avon longitudinal study of parents and children. *Paediatric and perinatal epidemiology,* 15**,** 74-87.

TEN DONKELAAR, H. J. & LAMMENS, M. 2009. Development of the human cerebellum and its disorders. *Clin Perinatol,* 36**,** 513-30.

TENG, E. L., LEE, P. H., YANG, K. S. & CHANG, P. C. 1976. Handedness in a Chinese population: biological, social, and pathological factors. *Science,* 193**,** 1148-50.

THOMAS, G. 2002. Furin at the cutting edge: From protein traffic to embryogenesis and disease. *Nature Reviews Molecular Cell Biology,* 3**,** 753-766.

THOMAS, M. 2006. Yerkes, Hamilton and the experimental study of the ape mind: from evolutionary psychiatry to eugenic politics. *Stud Hist Philos Biol Biomed Sci,* 37**,** 273-94.

THOMPSON, J. R., GÖGELE, M., WEICHENBERGER, C. X., MODENESE, M., ATTIA, J., BARRETT, J. H., BOEHNKE, M., DE GRANDI, A., DOMINGUES, F. S., HICKS, A. A., MARRONI, F., PATTARO, C., RUGGERI, F., BORSANI, G., CASARI, G., PARMIGIANI, G., PASTORE, A., PFEUFER, A., SCHWIENBACHER, C., TALIUN, D., CONSORTIUM, C., FOX, C. S., PRAMSTALLER, P. P. & MINELLI, C. 2013. SNP Prioritization Using a Bayesian Probability of Association. *Genetic Epidemiology,* 37**,** 214-221.

TIAN, D., SUN, S. & LEE, J. T. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell,* 143**,** 390-403.

TONNESSEN, F. E., LOKKEN, A., HOIEN, T. & LUNDBERG, I. 1993. Dyslexia, left-handedness, and immune disorders. *Arch Neurol,* 50**,** 411-6.

TRAN, U. S., STIEGER, S. & VORACEK, M. 2015. Mixed-footedness is a more relevant predictor of schizotypy than mixed-handedness. *Psychiatry Research,* 225**,** 446-451.

TRESS, M. L., MARTELLI, P. L., FRANKISH, A., REEVES, G. A., WESSELINK, J. J., YEATS, C., OLASON, P. I., ALBRECHT, M., HEGYI, H., GIORGETTI, A., RAIMONDO, D., LAGARDE, J., LASKOWSKI, R. A., LOPEZ, G., SADOWSKI, M. I., WATSON, J. D., FARISELLI, P., ROSSI, I., NAGY, A., KAI, W., STORLING, Z., ORSINI, M., ASSENOV, Y., BLANKENBURG, H., HUTHMACHER, C., RAMIREZ, F., SCHLICKER, A., DENOEUD, F., JONES, P., KERRIEN, S., ORCHARD, S., ANTONARAKIS, S. E., REYMOND, A., BIRNEY, E., BRUNAK, S., CASADIO, R., GUIGO, R., HARROW, J., HERMJAKOB, H., JONES, D. T., LENGAUER, T., ORENGO, C. A., PATTHY, L., THORNTON, J. M., TRAMONTANO, A. & VALENCIA, A. 2007. The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci U S A,* 104**,** 5495-500.

TREVARTHEN, C. 1996. Lateral Asymmetries in Infancy: Implications for the Development of the Hemispheres. *Neuroscience & Biobehavioral Reviews,* 20**,** 571-586.

TRINKLEIN, N. D., ALDRED, S. F., HARTMAN, S. J., SCHROEDER, D. I., OTILLAR, R. P. & MYERS, R. M. 2004. An Abundance of Bidirectional Promoters in the Human Genome. *Genome Research,* 14**,** 62-66.

TSUJI, A., HINE, C., TAMAI, Y., YONEMOTO, K., MORI, K., YOSHIDA, S., BANDO, M., SAKAI, E., MORI, K., AKAMATSU, T. & MATSUDA, Y. 1997. Genomic organization and alternative splicing of human PACE4 (SPC4), kexin-like processing endoprotease. *J Biochem,* 122**,** 438-52.

TURSKY, M. L., BECK, D., THOMS, J. A., HUANG, Y., KUMARI, A., UNNIKRISHNAN, A., KNEZEVIC, K., EVANS, K., RICHARDS, L. A., LEE, E., MORRIS, J., GOLDBERG, L., IZRAELI, S., WONG, J. W., OLIVIER, J., LOCK, R. B., MACKENZIE, K. L. & PIMANDA, J. E. 2015. Overexpression of ERG in cord blood progenitors promotes expansion and recapitulates molecular signatures of high ERG leukemias. *Leukemia,* 29**,** 819-27.

UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, A., KAMPF, C., SJOSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P. H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLEN, M., VON HEIJNE, G., NIELSEN, J. & PONTEN, F. 2015. Proteomics. Tissue-based map of the human proteome. *Science,* 347**,** 1260419.

UNIPROT, C. 2015. UniProt: a hub for protein information. *Nucleic Acids Res,* 43**,** D204-12.

USDIN, K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res,* 18**,** 1011-9.

VAN AGTMAEL, T., FORREST, S. M. & WILLIAMSON, R. 2002. Parametric and non-parametric linkage analysis of several candidate regions for genes for human handedness. *Eur J Hum Genet,* 10**,** 623-30.

VAUCLAIR, J., MEGUERDITCHIAN, A. & HOPKINS, W. D. 2005. Hand preferences for unimanual and coordinated bimanual tasks in baboons (*Papio anubis*). *Cognitive Brain Research,* 25**,** 210-216.

VERNES, S. C., SPITERI, E., NICOD, J., GROSZER, M., TAYLOR, J. M., DAVIES, K. E., GESCHWIND, D. H. & FISHER, S. E. 2007. High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *Am J Hum Genet,* 81**,** 1232-50.

VERSACE, E. & VALLORTIGARA, G. 2015. Forelimb preferences in human beings and other species: multiple models for testing hypotheses on lateralization. *Front Psychol,* 6**,** 233.

VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. & YANG, J. 2012. Five years of GWAS discovery. *The American Journal of Human Genetics,* 90**,** 7-24.

VOGLER, A. J., KEYS, C. E., ALLENDER, C., BAILEY, I., GIRARD, J., PEARSON, T., SMITH, K. L., WAGNER, D. M. & KEIM, P. 2007. Mutations, mutation rates, and evolution at the hypervariable VNTR loci of *Yersinia pestis*. *Mutat Res,* 616**,** 145-58.

VON PLESSEN, K., LUNDERVOLD, A., DUTA, N., HEIERVANG, E., KLAUSCHEN, F., SMIEVOLL, A. I., ERSLAND, L. & HUGDAHL, K. 2002. Less developed corpus callosum in dyslexic subjects-a structural MRI study. *Neuropsychologia,* 40**,** 1035-44.

WADSWORTH, S. J., OLSON, R. K. & DEFRIES, J. C. 2010. Differential Genetic Etiology of Reading Difficulties as a Function of IQ: An Update. *Behavior Genetics,* 40**,** 751-758.

WAHL, O. F. 1976. Handedness in schizophrenia. *Percept Mot Skills,* 42**,** 944-6.

WANG, L., WANG, S. & LI, W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics,* 28**,** 2184-2185.

WANG, Y., PARAMASIVAM, M., THOMAS, A., BAI, J., KAMINEN-AHOLA, N., KERE, J., VOSKUIL, J., ROSEN, G. D., GALABURDA, A. M. & LOTURCO, J. J. 2006. DYX1C1 functions in neuronal migration in developing neocortex. *Neuroscience,* 143**,** 515-22.

WARD, J. P., MILLIKEN, G. W., DODSON, D. L., STAFFORD, D. K. & WALLACE, M. 1990. Handedness as a Function of Sex and Age in a Large Population of Lemur. *Journal of Comparative Psychology,* 104**,** 167-173.

WARD, L. D. & KELLIS, M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol,* 30**,** 1095-106.

WARREN, D. M., STERN, M., DUGGIRALA, R., DYER, T. D. & ALMASY, L. 2006. Heritability and linkage analysis of hand, foot, and eye preference in Mexican Americans. *Laterality,* 11**,** 508-24.

WASYLYK, B., WASYLYK, C., FLORES, P., BEGUE, A., LEPRINCE, D. & STEHELIN, D. 1990. The c-ets proto-oncogenes encode transcription factors that cooperate with c-Fos and c-Jun for transcriptional activation. *Nature,* 346**,** 191-3.

WATERS, N. S. & DENENBERG, V. H. 1994. Analysis of two measures of paw preference in a large population of inbred mice. *Behav Brain Res,* 63**,** 195-204.

WATSON, G. S., PUSAKULICH, R. L., WARD, J. P. & HERMANN, B. 1998. Handedness, footedness, and language laterality: evidence from Wada testing. *Laterality,* 3**,** 323-30.

WEI, W., PELECHANO, V., JARVELIN, A. I. & STEINMETZ, L. M. 2011. Functional consequences of bidirectional promoters. *Trends Genet,* 27**,** 267-76.

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.

WERNER, A. & SAYER, J. A. 2009. Naturally occurring antisense RNA: function and mechanisms of action. *Curr Opin Nephrol Hypertens,* 18**,** 343-9.

WESTFALL, J. E., JASPER, J. D. & CHRISTMAN, S. 2012. Inaction inertia, the sunk cost effect, and handedness: avoiding the losses of past decisions. *Brain Cogn,* 80**,** 192-200.

WILLEMS, R. M., DER HAEGEN, L. V., FISHER, S. E. & FRANCKS, C. 2014. On the other hand: including left-handers in cognitive neuroscience and neurogenetics. *Nat Rev Neurosci,* 15**,** 193-201.

WILLEMS, R. M., ÖZYÜREK, A. & HAGOORT, P. 2007. When language meets action: The neural integration of gesture and speech. *Cerebral Cortex,* 17**,** 2322-2333.

WITELSON, S. F. 1985. The brain connection: the corpus callosum is larger in left-handers. *Science,* 229**,** 665-8.

WRANA, J. L., ATTISANO, L., WIESER, R., VENTURA, F. & MASSAGUE, J. 1994. Mechanism of activation of the TGF-beta receptor. *Nature,* 370**,** 341-7.

WRAY, G. A. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet,* 8**,** 206-16.

XIE, C., YUAN, J., LI, H., LI, M., ZHAO, G., BU, D., ZHU, W., WU, W., CHEN, R. & ZHAO, Y. 2014. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res,* 42**,** D98-103.

XU, Z., WEI, W., GAGNEUR, J., CLAUDER-MUNSTER, S., SMOLIK, M., HUBER, W. & STEINMETZ, L. M. 2011. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol,* 7**,** 468.

YAHAGI, S. & KASAI, T. 1999. Motor evoked potentials induced by motor imagery reveal a functional asymmetry of cortical motor control in left- and right-handed human subjects. *Neurosci Lett,* 276**,** 185-8.

YANG, M. Q., KOEHLY, L. M. & ELNITSKI, L. L. 2007. Comprehensive Annotation of Bidirectional Promoters Identifies Co-Regulation among Breast and Ovarian Cancer Genes. *PLoS Comput Biol,* 3**,** e72.

YAO, B. & JIN, P. 2014. Unlocking epigenetic codes in neurogenesis. *Genes Dev,* 28**,** 1253-71.

ZUCCA, P., BACIADONNA, L., MASCI, S. & MARISCOLI, M. 2011. Illness as a source of variation of laterality in lions (*Panthera leo*). *Laterality,* 16**,** 356-366.

ZUCCA, P., PALLADINI, A., BACIADONNA, L. & SCARAVELLI, D. 2010. Handedness in the echolocating Schreiber's Long-Fingered Bat (*Miniopterus schreibersii*). *Behavioural Processes,* 84**,** 693-695.