Expression Quantitative Trait Loci in Human Brain Tissues

Jesse Raphael Gibbs

Department of Molecular Neuroscience and Reta Lila Weston Laboratories

Institute of Neurology

University College London

A thesis submitted for the degree of Doctor of Philosophy

August 2015

In partnership with:

Laboratory of Neurogenetics, National Institute on Aging,

National Institutes of Health, Bethesda, Maryland, USA

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

6 September 2015

Abstract

To what extent genetic variability influences gene expression in human primary tissues is a critical question in molecular genetics. Work investigating this phenomenon is not only interesting biologically, but also has the potential to provide mechanistic insight into traits, including disease. The past decade has seen tremendous progress in this field, and this thesis includes a description of work that spanned from the relatively early stages of this type of work, to current, more refined efforts.

This work sought to ask three questions: first, are eQTL detectable in brain tissues using whole genome methods; second, are eQTL measurably different in different parts of the brain; and third, does the investigation of eQTL in a particular neuronal cell type offer significant advantages over similar studies in tissue with a mixed cellular composition.

In the first part of this work, I present a pilot study aimed at assessing the feasibility of eQTL detection in brain tissue. This study showed that the use of genome wide genotyping and expression arrays revealed a number of significant eQTL, and that in general, when genetic variability was associated with expression, the genetic locus and the expressed transcript were physically close. This work was then expanded to assess eQTL in multiple brain regions, with an attempt to assess whether eQTL were measurably different between distinct brain regions. In this work, tissue from cerebral frontal cortex, cerebral temporal cortex, caudal pons, and cerebellum was used. The analysis showed that there are region-specific eQTL, but that many of the strongest eQTL were present in multiple tissues. Lastly, I show using

data from laser capture microdissected Purkinje cells that additional cell-type specific eQTL may be found that are not revealed when performing eQTL in heterogeneous tissue containing this cell type.

In summary this work initially revealed the feasibility of eQTL work in human brain, showed that eQTL were measurably different, but generally similar across varied brain tissues, and showed that there are likely several advantages in pursuing single cell type work in tandem with whole tissue efforts.

Acknowledgements

I would like to thank my graduate supervisors, John Hardy and Holger Hummerich, as well as Andrew Singleton and Mark Cookson for their guidance support and the opportunity to work on these projects.

These projects involved a great deal of work in design, procurement of samples, production of data and facilitation of computational analysis as such I would like to thank the myriad of extremely talented and generous individuals who made these works possible. Firstly, I would like to thank the individuals and their families for the donation of materials for medical research to the involved brain banks. The tissues repositories and affiliated staff: the University of Maryland's Robert Johnson and H Ronald Zielke; NIA Baltimore Longitudinal Study of Aging's Luigi Ferrucci and Johns Hopkins University's Juan Troncoso; as well as the National Institute of Aging Alzheimer Centers and Miami Brain Bank. The laboratories and their staffs involved: The Laboratory of Neurogenetics's; Andrew Singleton, Mark Cookson, Marcel van der Brug, Dena Hernandez, Bryan Traynor, Shiao-Lin Lai, Michael Nalls, Sampath Arepalli, Cynthia Crews, Allissa Dillman, Azad Kumar, Alexandra Beilina, Ravindran Kumaran, Ian P Rafferty, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, and Dan Longo; Cardiff University's Peter Holmans; University of Miami's Amanda J Myers; Translational Genomics Research Institute's, Jennifer A Webster, Dietrich Stephan, Eric M Reiman, Victoria L Zismann, Keta Joshipura, Matthew J Huentelman, Diane Hu-Lince, Keith D Coon, David W Craig, John V Pearson; Kronos Science Laboratory's Christopher B Heward. The computational facilities staff that made these analysis tractable; the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md.; and Translational Genome Research Institute's IBM System Cluster 1350, at the Arizona State University Campus. Finally, I would like to thank the Institute of Neurology, University College London, and the intramural research program National Institute on Aging, National Institutes of Health.

Table of Contents

Tables

Figures

Abbreviations

AD	Alzheimer’s disease
ALDH3A2	aldehyde dehydrogenase 3 family, member A2
ALS	Amyotrophic lateral sclerosis
ANOVA	Analysis of variance
APOA1	apolipoprotein A-I
APOCIII	apolipoprotein C-III
APOE	apolipoprotein E
AS	alternatively spliced
ASE	allele-specific expression
AVPR1A	arginine vasopressin receptor 1A
B3GTL	beta 1,3-galactosyltransferase-like
BAG5	BCL2-associated athanogene 5
biotin-UTP	biotin labelled analog of uridine triphosphate
BLSA	Baltimore Longitudinal Study of Aging
BRE	TFIIB recognition element
BRLMM	Bayesian Robust Linear Model with Mahalanobis distance classifier
C10orf85	chromosome 10 open reading frame 85
CALB1	calbindin 1, 28kDa
CCBL2	cysteine conjugate-beta lyase 2
CCZ1B	vacuolar protein trafficking and biogenesis associated homolog B (S. cerevisiae)
cDNA	complementary deoxyribonucleic acid
CEPH	Centre d'Etude du Polymorphisme Humain
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
CEU	Utah residents with ancestry from northern and western Europe
CFTR	cystic fibrosis transmembrane conductance regulator
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
Chr	chromosome

CHST7	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7
CHURC1	churchill domain containing 1
cM/Mb	centiMorgan per megabase
CNV	Copy-number variant
CpG	5’-cytosine-phosphodiester bond-guanine-3’
cQTL	clinical quantitative trait locus
CRBLM	cerebellum
CRE	cAMP response elements
CRISPR-Cas9	Clustered regularly interspaced palindromic repeat and Cas9
cRNA	complementary ribonucleic acid
CSF	cerebrospinal fluid
CTCF	CCCTC-binding factor
DACH	dachshund family transcription factor
DARC	Duffy blood group chemokine receptor
dbEST	database of Expressed Sequence Tags
dbGaP	database of Genotypes and Phenotypes
dbSNP	database of single nucleotide polymorphisms
DHS	ENCODE DNase I hypersensitivity site
DMD	dystrophin; Duchenne muscular dystrophy
DNA	deoxyribonucleic acid
DPE	downstream promoter element
DRG	dorsal root ganglia
DSB	Double-strand break
DSCR5	Down syndrome critical region gene 5
DTNBP1	dystrobrevin binding protein 1
EBV	Epstein–Barr virus
EIF5A	eukaryotic translation initiation factor 5A
EMBL-EBI	European Bioinformatics Institute
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait loci

ERV	endogenous retroviruses
esQTL	expression-specific QTL
EST	Expressed Sequence Tag
ETS	E twenty-six
FANTOM5	functional annotation of the mammalian genome 5
FCTX	cerebral frontal cortex
FDR	false discovery rate
FTLD	Frontotemporal lobar degeneration
fweR2fdr	family wise error rate to false discovery rate
GAK	cyclin G associated kinase
GATA1	GATA binding protein 1
GDF5	growth differentiation factor 5
GEO	Gene Expression Omnibus
GFAP	Glial fibrillary acidic protein
GIH	Gujarati Indians in Houston, Texas
GPNMB	glycoprotein transmembrane nmb
GRID2	glutamate receptor, ionotropic, delta 2
GTEx	The Genotype-Tissue Expression
GWAS	genome-wide association study
HBS1L	Hsp70 subfamily B suppressor 1-like (S. cerevisiae)
HCL	Hierarchical Clustering
HD	Huntington's disease
Hsp70	heat shock 70kDa protein 4
htSNP	haplotype tagging single nucleotide polymorphism
HTR2A	5-hydroxytryptamine (serotonin) receptor 2A, G protein-coupled
H-R	Hill-Robertson
HWE	Hardy–Weinberg equilibrium
IBD	identity-by-descent
IBS	identity-by-state
IGF2	insulin and insulin-like growth factor 2
IL8	interleukin 8

IL10	interleukin-10
InDel	insertion or deletion variant
INPP5E	inositol polyphosphate-5-phosphatase, 72 kDa
Inr	initiator element
IPDGC	International Parkinson's Disease Genomics Consortium
IPP	intracisternal A particle-promoted polypeptide
iPS	induced pluripotent stem
JHU	John Hopkins University
JPT	Japanese in Tokyo, Japan
Kb	kilo-bases
KCTD10	potassium channel tetramerization domain containing 10
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KIF1B	kinesin family member 1B
LCL	lymphoblastoid cell line
LCM	laser capture microdissection
LCT	lactase
LD	linkage disequilibrium
LNG	Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, USA
LRRK2	leucine-rich repeat kinase 2
LWK	Luhya in Webuye, Kenya
MAF	minor allele frequency
MAOA	monoamine oxidase A
MAPT	microtubule-associated protein tau
Mb	mega-bases
MBP	Myelin basic protein
MCM6	maintenance deficient 6 homologue
MDS	multidimensional scaling
methQTL	methylation QTL
MEX	Mexican ancestry in Los Angeles, California
mg	milligram

miRNA	micro ribonucleic acid
MKK	Maasai in Kinyawa, Kenya
mL	millilitre
MOBP	myelin-associated oligodendrocyte basic protein
mRNA	messenger ribonucleic acid
NABEC	North American Brain Expression Consortium
NAPRT1	nicotinate phosphoribosyltransferase domain containing 1
NCBI	National Center for Biotechnology Information, National Institutes of Health, USA
NCRAD	National Cell Repository for Alzheimer's Disease
NFkB	nuclear factor kappa B
ng/ul	nanogram per microliter
NHGRI	National Human Genome Research Institute
NIA	National Institute on Aging, National Institutes of Health, USA
NPC	nuclear pore complex
NQO2	NAD(P)H dehydrogenase, quinone 2
NRF-1	nuclear respiratory factor 1
NUCKS1	nuclear casein kinase and cyclin-dependent kinase substrate 1
NUPL2	nucleoporin like 2
ORMDL3	orosomucoid 1-like 3 (S. cerevisiae)
PAX6	paired box 6
PCA	principle component analysis
PCP2	Purkinje cell protein 2
PD	Parkinson's disease
PDYN	prodynorphin
PEX6	peroxisomal biogenesis factor 6
PLEKHM1	pleckstrin homology domain containing family M (with RUN domain) member 1
PolyQ	Polyglutamine
PRDM9	PR domain containing 9
PMI	post-mortem interval

PONS	caudal pons
PSP	Progressive Supranuclear Palsy
PPAPDC1A	phosphatidic acid phosphatase type 2 domain containing 1A
pQTL	protein QTL
psQTL	protein-specific QTL
PTD004	Obg-like ATPase 1
PTGER4	prostaglandin E receptor 4
PVALB	Parvalbumin
PVT	polymorphic transcript variation
RAB7L1	RAB7 member RAS oncogene family-like 1
RefSeq	Reference Sequence Database
ReMOAT	Reannotation and Mapping of Oligonucleotide Arrays Technologies
RFLP	restriction fragment length polymorphisms
RLMM	robust linear model of Mahalanobis distance
RNA	ribonucleic acid
RNA Pol II	RNA polymerase II holoenzyme complex
RNAseq	RNA-sequencing
RPL14	ribosomal protein L14
RPL5	ribosomal protein L5
RPS23	ribosomal protein S23
RPS26	ribosomal protein S26
rQTL	ribosomal occupancy QTL
RTC	Regulatory Trait Concordance
RT-PCR	reverse transcription polymerase chain reaction
RYBP	RING1 and YY1 binding protein
SCA7	spinocerebellar ataxia type 7
SCG3	secretogranin III
S.D.	standard deviation
SLC6A4	solute carrier family 6 (neurotransmitter transporter), member 4
SLC25A38	solute carrier family 25, member 38

SLE	Systemic lupus erythematosus
siRNA	small interfering RNA
smallRNA	small ribonucleic acid
SNP	single nucleotide polymorphism
SP1	specificity protein 1
SQSTM1	sequestosome 1
sQTL	splicing QTL
STAT6	signal transducer and activator of transcription 6, interleukin-4 induced
TCTX	cerebral temporal cortex
TFBS	transcription factor binding site
TES	transcription end site
tRNA	transfer ribonucleic acid
TBP	TATA-binding protein
TE	transposable elements
TF	transcription factor
TSI	Toscani in Italia
TSS	transcription start site
UKBEC	United Kingdom Brain Expression Consortium
UMARY	University of Maryland
USA	United States of America
UTR	untranslated region
UV	ultra-violet
VNN1	vanin 1
WGACON	whole genome Alzheimer’s disease control
YRI	Yoruba in Ibadan, Nigeria
ZNF143	zinc finger protein 143
ZNF266	zinc finger protein 266
ZNF419	zinc finger protein 419
µm	micrometre

Publications

Publications listed here are those relevant to my thesis. My additional publications are listed, at the end of my thesis, in Chapter 10.

1. Myers AJ, **Gibbs JR**, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J. A survey of genetic human cortical gene expression. Nat Genet. 2007 Dec;39(12):1494-9. Epub 2007 Nov 4. PubMed PMID: 17982457.

2. Jakobsson M, Scholz SW, Scheet P, **Gibbs JR**, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008 Feb 21;451(7181):998-1003. doi: 10.1038/nature06742. PubMed PMID: 18288195.

3. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, **Gibbs JR**, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Gieger C, Houlden H, Steffens M, Okun MS, Racette BA, Cookson MR, Foote KD, Fernandez HH, Traynor BJ, Schreiber S, Arepalli S, Zonozi R, Gwinn K, van der Brug M, Lopez G, Chanock SJ, Schatzkin A, Park Y, Hollenbeck A, Gao J, Huang X, Wood NW, Lorenz D, Deuschl G, Chen H, Riess O, Hardy JA, Singleton AB, Gasser T. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009 Dec;41(12):1308-12. doi: 10.1038/ng.487. Epub 2009 Nov 15. PubMed PMID: 19915575.

4. Guerreiro RJ, Beck J, **Gibbs JR**, Santana I, Rossor MN, Schott JM, Nalls MA, Ribeiro H, Santiago B, Fox NC, Oliveira C, Collinge J, Mead S, Singleton A, Hardy J. Genetic variability in CLU and its association with Alzheimer's disease. PLoS One. 2010 Mar 3;5(3):e9510. doi: 10.1371/journal.pone.0009510. PubMed PMID: 20209083.

5. **Gibbs JR**, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010 May 13;6(5):e1000952. doi: 10.1371/journal.pgen.1000952. PubMed PMID: 20485568.

6. Traynor BJ, Nalls M, Lai SL, **Gibbs RJ**, Schymick JC, Arepalli S, Hernandez D, van der Brug MP, Johnson JO, Dillman A, Cookson M, Moglia C, Calvo A, Restagno G, Mora G, Chiò A. Kinesin-associated

protein 3 (KIFAP3) has no effect on survival in a population-based cohort of ALS patients. Proc Natl Acad Sci U S A. 2010 Jul 6;107(27):12335-8. doi: 10.1073/pnas.0914079107. Epub 2010 Jun 21. PubMed PMID: 20566859.

7. Carrasquillo MM, Nicholson AM, Finch N, **Gibbs JR**, Baker M, Rutherford NJ, Hunter TA, DeJesus-Hernandez M, Bisceglio GD, Mackenzie IR, Singleton A, Cookson MR, Crook JE, Dillman A, Hernandez D, Petersen RC, Graff-Radford NR, Younkin SG, Rademakers R. Genome-wide screen identifies rs646776 near sortilin as a regulator of progranulin levels in human plasma. Am J Hum Genet. 2010 Dec 10;87(6):890-7. doi: 10.1016/j.ajhg.2010.11.002. Epub 2010 Nov 18. PubMed PMID: 21087763.

8. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet. 2011 Feb 19;377(9766):641-9. doi: 10.1016/S0140-6736(10)62345-8. Epub 2011 Feb 1. PubMed PMID: 21292315.

9. Höglinger GU, Melhem NM, Dickson DW, Sleiman PM, Wang LS, Klei L, Rademakers R, de Silva R, Litvan I, Riley DE, van Swieten JC, Heutink P, Wszolek ZK, Uitti RJ, Vandrovcova J, Hurtig HI, Gross RG, Maetzler W, Goldwurm S, Tolosa E, Borroni B, Pastor P; PSP Genetics Study Group, Cantwell LB, Han MR, Dillman A, van der Brug MP, **Gibbs JR**, Cookson MR, Hernandez DG, Singleton AB, Farrer MJ, Yu CE, Golbe LI, Revesz T, Hardy J, Lees AJ, Devlin B, Hakonarson H, Müller U, Schellenberg GD. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. Nat Genet. 2011 Jun 19;43(7):699-705. doi: 10.1038/ng.859. PubMed PMID: 21685912.

10. Hernandez DG, Nalls MA, Moore M, Chong S, Dillman A, Trabzuni D, **Gibbs JR**, Ryten M, Arepalli S, Weale ME, Zonderman AB, Troncoso J, O'Brien R, Walker R, Smith C, Bandinelli S, Traynor BJ, Hardy J, Singleton AB, Cookson MR. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. Neurobiol Dis. 2012 Jul;47(1):20-8. doi: 10.1016/j.nbd.2012.03.020. Epub 2012 Mar 12. PubMed PMID: 22433082.

11. Holton P, Ryten M, Nalls M, Trabzuni D, Weale ME, Hernandez D, Crehan H, **Gibbs JR**, Mayeux R, Haines JL, Farrer LA, Pericak-Vance MA, Schellenberg GD; Alzheimer's Disease Genetics Consortium, Ramirez-Restrepo M, Engel A, Myers AJ, Corneveaux JJ, Huentelman MJ, Dillman A, Cookson MR, Reiman EM, Singleton A, Hardy J, Guerreiro R. Initial assessment of the pathogenic mechanisms of the recently identified Alzheimer risk Loci. Ann Hum Genet. 2013

Mar;77(2):85-105. doi: 10.1111/ahg.12000. Epub 2013 Jan 30. PubMed PMID: 23360175.

12. Trabzuni D, Wray S, Vandrovcova J, Ramasamy A, Walker R, Smith C, Luk C, **Gibbs JR**, Dillman A, Hernandez DG, Arepalli S, Singleton AB, Cookson MR, Pittman AM, de Silva R, Weale ME, Hardy J, Ryten M. MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies. Hum Mol Genet. 2012 Sep 15;21(18):4094-103. doi: 10.1093/hmg/dds238. Epub 2012 Jun 20. PubMed PMID: 22723018.

13. Kumar A, **Gibbs JR**, Beilina A, Dillman A, Kumaran R, Trabzuni D, Ryten M, Walker R, Smith C, Traynor BJ, Hardy J, Singleton AB, Cookson MR. Age-associated changes in gene expression in human brain and isolated neurons. Neurobiol Aging. 2013 Apr;34(4):1199-209. doi: 10.1016/j.neurobiolaging.2012.10.021. Epub 2012 Nov 21. PubMed PMID: 23177596.

14. Scharf JM, Yu D, Mathews CA, Neale BM, Stewart SE, Fagerness JA, Evans P, Gamazon E, Edlund CK, Service SK, Tikhomirov A, Osiecki L, Illmann C, Pluzhnikov A, Konkashbaev A, Davis LK, Han B, Crane J, Moorjani P, Crenshaw AT, Parkin MA, Reus VI, Lowe TL, Rangel-Lugo M, Chouinard S, Dion Y, Girard S, Cath DC, Smit JH, King RA, Fernandez TV, Leckman JF, Kidd KK, Kidd JR, Pakstis AJ, State MW, Herrera LD, Romero R, Fournier E, Sandor P, Barr CL, Phan N, Gross-Tsur V, Benarroch F, Pollak Y, Budman CL, Bruun RD, Erenberg G, Naarden AL, Lee PC, Weiss N, Kremeyer B, Berrío GB, Campbell DD, Cardona Silgado JC, Ochoa WC, Mesa Restrepo SC, Muller H, Valencia Duarte AV, Lyon GJ, Leppert M, Morgan J, Weiss R, Grados MA, Anderson K, Davarya S, Singer H, Walkup J, Jankovic J, Tischfield JA, Heiman GA, Gilbert DL, Hoekstra PJ, Robertson MM, Kurlan R, Liu C, **Gibbs JR**, Singleton A; North American Brain Expression Consortium, Hardy J; UK Human Brain Expression Database, Strengman E, Ophoff RA, Wagner M, Moessner R, Mirel DB, Posthuma D, Sabatti C, Eskin E, Conti DV, Knowles JA, Ruiz-Linares A, Rouleau GA, Purcell S, Heutink P, Oostra BA, McMahon WM, Freimer NB, Cox NJ, Pauls DL. Genome-wide association study of Tourette's syndrome. Mol Psychiatry. 2013 Jun;18(6):721-8. doi: 10.1038/mp.2012.69. Epub 2012 Aug 14. PubMed PMID: 22889924.

15. Stewart SE, Yu D, Scharf JM, Neale BM, Fagerness JA, Mathews CA, Arnold PD, Evans PD, Gamazon ER, Davis LK, Osiecki L, McGrath L, Haddad S, Crane J, Hezel D, Illman C, Mayerfeld C, Konkashbaev A, Liu C, Pluzhnikov A, Tikhomirov A, Edlund CK, Rauch SL, Moessner R, Falkai P, Maier W, Ruhrmann S, Grabe HJ, Lennertz L, Wagner M, Bellodi L, Cavallini MC, Richter MA, Cook EH Jr, Kennedy JL, Rosenberg D, Stein DJ, Hemmings SM, Lochner C, Azzam A, Chavira DA, Fournier E, Garrido H, Sheppard B, Umaña P, Murphy DL, Wendland JR, Veenstra-VanderWeele J, Denys D, Blom R, Deforce D, Van Nieuwerburgh F, Westenberg HG, Walitza S, Egberts K, Renner T, Miguel EC, Cappi C, Hounie AG, Conceição do Rosário M, Sampaio AS, Vallada H, Nicolini H, Lanzagorta N, Camarena B, Delorme R,

Leboyer M, Pato CN, Pato MT, Voyiaziakis E, Heutink P, Cath DC, Posthuma D, Smit JH, Samuels J, Bienvenu OJ, Cullen B, Fyer AJ, Grados MA, Greenberg BD, McCracken JT, Riddle MA, Wang Y, Coric V, Leckman JF, Bloch M, Pittenger C, Eapen V, Black DW, Ophoff RA, Strengman E, Cusi D, Turiel M, Frau F, Macciardi F, **Gibbs JR**, Cookson MR, Singleton A; North American Brain Expression Consortium, Hardy J; UK Brain Expression Database, Crenshaw AT, Parkin MA, Mirel DB, Conti DV, Purcell S, Nestadt G, Hanna GL, Jenike MA, Knowles JA, Cox N, Pauls DL. Genome-wide association study of obsessive-compulsive disorder. Mol Psychiatry. 2013 Jul;18(7):788-98. doi: 10.1038/mp.2012.85. Epub 2012 Aug 14. Erratum in: Mol Psychiatry. 2013 Jul;18(7):843. Davis, L K [added]. PubMed PMID: 22889921.

16. Anttila V, Winsvold BS, Gormley P, Kurth T, Bettella F, McMahon G, Kallela M, Malik R, de Vries B, Terwindt G, Medland SE, Todt U, McArdle WL, Quaye L, Koiranen M, Ikram MA, Lehtimäki T, Stam AH, Ligthart L, Wedenoja J, Dunham I, Neale BM, Palta P, Hamalainen E, Schürks M, Rose LM, Buring JE, Ridker PM, Steinberg S, Stefansson H, Jakobsson F, Lawlor DA, Evans DM, Ring SM, Färkkilä M, Artto V, Kaunisto MA, Freilinger T, Schoenen J, Frants RR, Pelzer N, Weller CM, Zielman R, Heath AC, Madden PA, Montgomery GW, Martin NG, Borck G, Göbel H, Heinze A, Heinze-Kuhn K, Williams FM, Hartikainen AL, Pouta A, van den Ende J, Uitterlinden AG, Hofman A, Amin N, Hottenga JJ, Vink JM, Heikkilä K, Alexander M, Muller-Myhsok B, Schreiber S, Meitinger T, Wichmann HE, Aromaa A, Eriksson JG, Traynor BJ, Trabzuni D, Rossin E, Lage K, Jacobs SB, **Gibbs JR**, Birney E, Kaprio J, Penninx BW, Boomsma DI, van Duijn C, Raitakari O, Jarvelin MR, Zwart JA, Cherkas L, Strachan DP, Kubisch C, Ferrari MD, van den Maagdenberg AM, Dichgans M, Wessman M, Smith GD, Stefansson K, Daly MJ, Nyholt DR, Chasman DI, Palotie A; North American Brain Expression Consortium; UK Brain Expression Consortium; International Headache Genetics Consortium. Genome-wide meta-analysis identifies new susceptibility loci for migraine. Nat Genet. 2013 Aug;45(8):912-7. doi: 10.1038/ng.2676. Epub 2013 Jun 23. PubMed PMID: 23793025.

17. Nalls MA, Saad M, Noyce AJ, Keller MF, Schrag A, Bestwick JP, Traynor BJ, **Gibbs JR**, Hernandez DG, Cookson MR, Morris HR, Williams N, Gasser T, Heutink P, Wood N, Hardy J, Martinez M, Singleton AB; International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2); North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC). Genetic comorbidities in Parkinson's disease. Hum Mol Genet. 2014 Feb 1;23(3):831-41. doi: 10.1093/hmg/ddt465. Epub 2013 Sep 20. PubMed PMID: 24057672.

18. Beilina A, Rudenko IN, Kaganovich A, Civiero L, Chau H, Kalia SK, Kalia LV, Lobbestael E, Chia R, Ndukwe K, Ding J, Nalls MA; International Parkinson’s Disease Genomics Consortium; North American Brain Expression Consortium, Olszewski M, Hauser DN,

Kumaran R, Lozano AM, Baekelandt V, Greene LE, Taymans JM, Greggio E, Cookson MR. Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease. Proc Natl Acad Sci U S A. 2014 Feb 18;111(7):2626-31. doi: 10.1073/pnas.1318306111. Epub 2014 Feb 7. PubMed PMID: 24510904.

19. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, DeStefano AL, Kara E, Bras J, Sharma M, Schulte C, Keller MF, Arepalli S, Letson C, Edsall C, Stefansson H, Liu X, Pliner H, Lee JH, Cheng R; International Parkinson's Disease Genomics Consortium (IPDGC); Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI); 23andMe; GenePD; NeuroGenetics Research Consortium (NGRC); Hussman Institute of Human Genomics (HIHG); Ashkenazi Jewish Dataset Investigator; Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE); North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC); Greek Parkinson's Disease Consortium; Alzheimer Genetic Analysis Group, Ikram MA, Ioannidis JP, Hadjigeorgiou GM, Bis JC, Martinez M, Perlmutter JS, Goate A, Marder K, Fiske B, Sutherland M, Xiromerisiou G, Myers RH, Clark LN, Stefansson K, Hardy JA, Heutink P, Chen H, Wood NW, Houlden H, Payami H, Brice A, Scott WK, Gasser T, Bertram L, Eriksson N, Foroud T, Singleton AB. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet. 2014 Sep;46(9):989-93. doi: 10.1038/ng.3043. Epub 2014 Jul 27. PubMed PMID: 25064009.

1: Introduction

An essential challenge in the post-genome era is to understand the effects of genetic variation found within the genome. With the widespread application of highly parallel SNP (single nucleotide polymorphism) genotyping arrays much of the recent effort in human genetics has focused on defining the role of genetic variation in disease and physical traits. A smaller subset of work, however, has attempted to examine the more proximal effects of genetic variation, particularly their effects on mRNA (messenger ribonucleic acid) and protein levels. This has the potential to inform on several levels: first, it is a critical step toward understanding the pathobiological consequences of genetic variants linked to clinical phenotypes; second, it affords the opportunity to form inferences regarding relationships between genes based on patterns of co-regulation; and third, it provides a more complete view of multiple levels of regulation of gene expression than that provided by the traditional reductionist method.

The genetic code is largely fixed across human populations and, with rare exceptions, absolutely fixed within an individual. However, there is substantial variability in gene expression between individuals and across tissues. Much of the inter-individual differences will be embedded in genetic variation at the sequence level. However, changes in expression at the individual and tissue level will also reflect responses to external stimuli and this is likely to be mediated in part through epigenetic variation. Previously the relationship between genetic and epigenetic influences on gene expression is one that has been largely and necessarily confined to observations at single loci and transcripts in individual cell systems or tissues. The advent of genome-scale technologies provides unprecedented opportunities to expand upon these experiments. The integration of genetic and expression data promises to provide general observations regarding the relationship between genetic variation and expression. Beyond these observations, these data can be readily mined to unravel the network of effects associated with genomic variants. This may reveal some of the rather cryptic intermediate events that occur between DNA (deoxyribonucleic acid) variant and phenotype.

1.1: Research Aims

With the arrival of high-density SNP chips combined with the maturation of expression microarray platforms it is now feasible to capture most of the known common genetic variation as well as the expression profiles for well-known mRNA transcripts in the human genome for a large number of individuals. The mapping of these effects where genetic variation in a particular region of the genome is linked or associated with a change in the expression of a particular mRNA transcript is commonly referred to as an expression quantitative trait locus or eQTL. The study of expression quantitative traits is very similar to other quantitative traits, such as clinical measures, but here the trait or phenotype of interest is the expression of mRNA transcripts. These expression traits may also be referred to as molecular, intermediate, or endo-phenotypes as they are internal phenotypes that may lead to an external phenotype. The study of eQTL is the integration of genetic variation and gene expression variation by correlation, where the abundance of an mRNA transcript increases or decreases in relation to genotype (Figure 1.1).

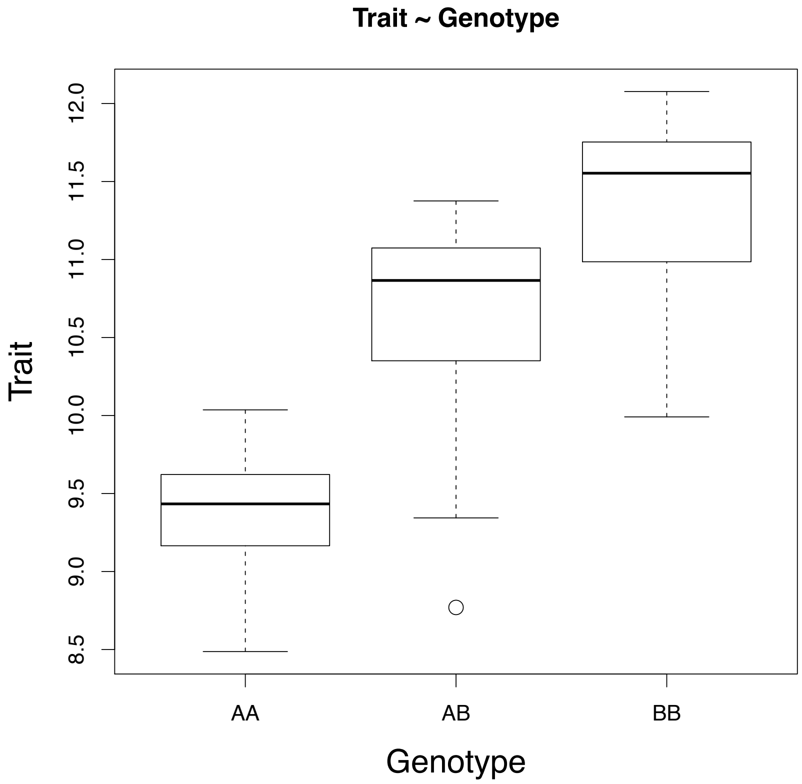


Figure 1.1: Example plot of the linear relationship between genotype and gene expression for an eQTL. The plot depicts this linear relationship, where the abundance of an mRNA transcript increases with the dosage of the B allele. In this example, the correlation is positive, but a negative correlation is also possible where mRNA's transcript abundance decreases with the dosage of the B allele.

The expectation of identifying eQTL is that we can begin to provide an additional layer of functional information onto genetic variation within the human genome as well as to understand the general characteristics of eQTL. The utility of such a resource is that many of the expression quantitative trait loci may overlap with regions of the genome associated with clinical traits or disease phenotypes. In recent years, hundreds (and now thousands) of genome wide association studies have been published many with robust and replicated findings (Hindorff *et al.* 2009). What is apparent from these many genome wide association studies is that large effect disease loci that tag protein-coding changes such as complement factor H for age related macular degeneration (Klein *et al.* 2005) and *APOE* for Alzheimer's disease (Rogaeva 2002) will not be the norm. Many of these disease-associated loci may confer moderate to small risk through changes in gene expression.

Towards this end, my thesis focuses on eQTL studies within brain tissues using whole-genome SNP genotyping and mRNA expression microarray data. The first study was conceptually a ‘pilot’ project for the detection of eQTL within a mix of cortical tissues from elderly neurologically normal individuals. The second study expands upon the ‘pilot‘ study, again using neurologically normal individuals but with improved analysis techniques and multiple brain tissues from each individual. The third study describes eQTL in a single neuronal cell type from human brain, using a subset of the second study’s subject cohort and makes use of additionally refined analysis methods. My thesis also includes a chapter describing the integration of eQTL and disease risk loci identified by genome-wide association studies (GWAS). The bulk of my thesis and central project is within the second cohort focusing on identifying eQTL within distinct human brain regions. However, the first ‘pilot’ cohort is seminal in my understanding of how to do this kind of work in a primary human tissue and was the first study, using whole genome and transcriptome data, to show that it is possible to see such effects in human brain tissue. The third study, of eQTL in a single neuronal cell type, is critical to beginning to understand eQTL in the context of a specific cell type and within heterogeneous tissues.

1.2: Human Genome

1.2.1: Human Genome and Genetic Variation

After the completion of the Human Genome Project, much research has focused on understanding variation within the human genome, among individuals, and among populations. A primary goal of this work is to understand how the patterns of variation can be applied towards localizing loci associated with complex traits, such as disease, within humans. The work to begin to achieve this understanding has accelerated rapidly since the draft human genome was sequenced. This effort requires not only cataloguing variation within the human genome but also the development of methods to assay and analyse this catalogue in order to better understand patterns within the variation. Of course, much of this work has also centred on how to maximize the efficiency and effectiveness of the data for generation and analysis.

Early studies of genetic variation between individuals and populations had shown that much of this variation is between individuals and to a lesser extent between populations. One of these very early studies was based on allele frequencies at 15 protein loci and found that 85% of genetic diversity is between subjects from within the same population (R C Lewontin 1972). This estimate was re-affirmed more than twenty years later in a much larger cohort using DNA markers. In this later study, 1,109 subjects were genotyped at 109 DNA markers, where the markers included microsatellites and restriction fragment length polymorphisms (RFLP). Microsatellites are short tandem repeats, containing two to six base pairs of repeating sequence. RFLPs are genetic markers captured by segmenting DNA using restriction enzymes and then separated according to their length by gel electrophoresis. The results showed that the within population differences accounted for 84.4% of the genetic variation and approximately 10% of the variation is accounted for by continental population differences (Barbujani *et al.* 1997). Additionally, when studying a larger more diverse set of populations the majority of genetic variation was again found to be primarily between individuals and not between populations (Rosenberg *et al.* 2002). In the Rosenberg study, which was based on microsatellites genotyped in 52 populations, it was found that ~94% of the genetic variation is among individuals within a population and ~4% of the variation is between major population groups.

The rate of introduction and the change in frequency of new combinations of alleles in a genome is determined by recombination. Recombination results in the selection for or against new haplotypes and this in turn may also lead to the selection of alleles that modify recombination rates (Otto and Lenormand 2002; Coop and Przeworski 2007). Two roles of recombination in mammals and other organisms are that it helps in homology recognition resulting in synapsis early in meiosis and then later provides the tension required for the correct chromatids to be pulled together binding through crossing over of the non-sister chromatids (Roeder 1997; Coop and Przeworski 2007). Synapsis is the pairing of two homologous chromosomes. A chromatid is a new copy of a replicated chromosome. Meiotic pairing, synapsis, and recombination occur during prophase I of meiosis. In prophase I, sister chromatids are brought into close reach with one another and sister chromatid cohesion is imparted by a meiosis-specific cohesion complex (Zickler and Kleckner 1999; Petronczki, Siomos and Nasmyth 2003). Double-strand breaks initiate meiotic recombination and Holliday junctions are formed during repair. These junctions are resolved as a gene conversion with or without a crossover. Where in non-crossover conversions the resolution does not include the exchange of flanking variants as opposed to a crossover conversion where flanking variants are exchanged (Allers and Lichten 2001; de Massy 2003; Coop and Przeworski 2007). Large crossover rates increase the genetic diversity within humans and divergence with other species (Hellmann *et al.* 2003, 2005). Recombination rates vary within in humans and there are differences between humans and other species (Lynn, Ashley and Hassold 2004). Recombination ensures the proper segregation of chromosomes during meiosis and breaks up genetic linkage between loci resulting in increased diversity.

Recombination can be measured indirectly using genotypes within families to create maps inferring the recombination events from the parents. These recombination maps are referred to as genetic maps, which consider the polymorphic loci in a linear fashion along a chromosome and their interlocus interval lengths. Previously, restriction fragment length polymorphisms (RFLPs) were used as the genomic variants for construction of these genetic maps (Botstein *et al.* 1980) and later microsatellites (Litt and Luty 1989; Weber and May 1989). A Morgan is a unit of length used to denote the linear distance in genomic maps of recombination fractions and measures the relationship between pairs of “marker” loci or variants in the genome. These genetic maps estimate the segregation of the alleles for a pair of markers determining if the pair is linked together forming a linkage group. For mapping purposes, Morgans are typically denoted as centiMorgans. The closer together two loci are the more unlikely it is that a double recombination event exists between them so the loci are linked. As the distance between loci increases, the possibility of a recombination event occurring between them increases as well. The Centre d'Etude du Polymorphisme humain (CEPH) was established to maintain a collection of a common set of pedigrees containing enough families to help facilitate the construction of linkage maps. This collection of samples allowed laboratories developing genotyping markers to work with the same samples and return this information to a public repository so that genome-wide linkage maps could be constructed (Dausset *et al.* 1990).

The larger chromosomes have more meiotic exchange and thus more recombination events. Variation in recombination can be driven by specific factors for specific chromosomes (Laurie and Hultén 1985; Lynn, Ashley and Hassold 2004). Recombination rates have also been found to be significantly correlated with GC content in the human genome (Kong *et al.* 2002; Lynn, Ashley and Hassold 2004). Recombination hotspots occur in small regions typically less than 1 to 2 Kb (kilobases) and separate long regions of cold spots typically 200 Kb; recombination hotspots are not randomly distributed in the genome (Jeffreys, Kauppi and Neumann 2001; May *et al.* 2002; Lynn, Ashley and Hassold 2004). There is an inverse relationship between linkage disequilibrium (LD) and recombination rates (Nordborg and Tavaré 2002). In 2004, McVean *et al.* published a study estimating recombination rates based on genotypes. They used European and African population genotypes and found variation in local recombination rates. They found that 50% of recombination events occur in less than 10% of the genome and occur on average every 200 Kb or less. These results suggested that recombination hotspots are a common feature of the human genome and typically occur outside of genes (McVean *et al.* 2004). In a similar study using LD patterns based on genome-wide genotyping, it was suggested that there are likely more the 25,000 hotspots in the human genome. The use of genome-wide genotyping allowed for fine-scale estimates of recombination. Additionally, it was found that these hotspots occur approximately every 50 Kb and that 80% of crossover events happen in 10 to 20% of the human genome (Myers *et al.* 2005). In 2005, Hellman *et al.* published a study examining diversity in the human genome and divergence with chimpanzees under the hypothesis that variation has two main drivers, mutation rates and natural selection. This study was undertaken based on previous observations that both within species diversity and between species divergence increase with recombination rates. The observation that both diversity and divergence increases with recombination rate suggests that there is a link with recombination and mutation directly or through another factor. By studying the human and chimpanzee genomes, they found that GC and CpG content, simple-repeats, and chromosomal distance from centromeres and telomeres also predict diversity and divergence. They suggest that changes in recombination rates since the split with our common chimpanzee ancestor are a better explanation for diversity within species. Their basis for this conclusion was the observation that recombination rates appear to have rapidly changed during human evolution (Hellmann *et al.* 2005). An earlier study of diversity and divergence in humans, also from Hellmann *et al.*, was undertaken to examine previous findings that regions with low recombination often have lower diversity within species but not lower divergence between species. They regenerated maps of recombination within related species and found that, between chimpanzees and baboons, regions with low recombination have less divergence and diversity for the two non-human primates (Hellmann *et al.* 2003). In 2011, Wegmann *et al.* published a study based on admixed human subjects to generate a recombination map of relative rates, which allowed for observing ancestry switch points. They used genotypes from African-American and African-Caribbean subjects, and found several thousand recombination events when compared to maps from non-admixed populations. Their results showed a fine-scale difference in recombination between populations suggesting that admixture does not have a large effect on recombination rates in humans (Wegmann *et al.* 2011).

The Hill-Robertson (H-R) effect suggests that in finite populations for two linked loci that selection at one locus reduces the effectiveness of selection at the other one (Hill and Robertson 1966; Felsenstein 1974). Intragenic H-R effects may be reduced advantageously by introns and may predict selection effectiveness differences between genes with different exon-intron structures (Comeron, Williford and Kliman 2008). Populations without recombination will accumulate deleterious mutations more rapidly, this is known as Mueller's Ratchet (Muller 1964). Both the H-R effect and Mueller's Ratchet suggest that the advantage of recombination is that it increases the rate of adaptation in a species by breaking up negative linkage disequilibrium (LD) that may result from selection and genetic drift (Felsenstein 1974; Barton and Otto 2005;

Keightley and Otto 2006; Comeron, Williford and Kliman 2008). Negative linkage disequilibrium is the association between beneficial and deleterious alleles at different loci occurring more often than expected by chance (Keightley and Otto 2006). In 2011, Udeba and Wilkins put forth a model to account for observed recombination mechanisms. Their model results suggested a Red Queen dynamic based on an intragenomic conflict model. The Red Queen hypothesis, put forth by van Valen in 1973, suggested that organisms must constantly adapt, evolve, and proliferate not just for reproductive advantage but also to survive against other evolving organism in environments that are also changing. The Udeba and Wilkins model accounts for evolutionary dynamics of hotspot turnover and the non-random targeting of recombination mechanisms. Chromosomal regions where crossover events occur more often are recombination hotspots and these regions are often small. Double-strand breaks (DSB) initiate recombination and requires involvement of the DSB repair mechanism, which may result in homologous chromosome exchange (crossover). This exchange would then result in a biased gene conversion. However, a biased gene conversion results in a transmission advantage preventing recombination and therefore the hotspot becomes transient. The persistence of hotspots over time when they should be self-destructive is known as the recombination hotspot paradox. They suggest that their intragenomic conflict model accounts for this. In their model fertility selection drives *trans* (distal) modifiers to maintain crossover rates which is in conflict with *cis*-acting (proximal) targets promoting their own transmission, thus the intragenomic conflict, leading to the Red Queen dynamics (Ubeda and Wilkins 2011). More recently, Lesecque *et al.* published a study of the evolution of human recombination hotspots and PR domain containing 9 (*PRDM9*) DNA-binding domain target motifs in the human genome. The PRDM9 protein is a zinc finger protein with a sequence-specific DNA binding domain that determines the location of recombination hotspots in humans (Baudat *et al.* 2010; Myers *et al.* 2010). The Lesecque *et al.* study examined the evolution of hotspots and PRDM9 target motifs by comparing the genomes of modern human and Denisovan to identify hotspot turnover in recent human evolution. This study found that even though Denisovans and modern humans share similar PRDM9 target motifs their recombination hotspots did not overlap. They also found that modern human hotspots are young, becoming active shortly before the split with Denisovans but long after divergence from chimpanzee ancestor. Their findings suggest that the loss of existing human hotspots, through biased gene conversion, should occur in the next three million years and this depletion would decrease fitness thereby favouring new PRDM9 alleles binding different motifs which supports the Red Queen hypothesis of recombination hotspot evolution (Lesecque *et al.* 2014).

As technologies and understanding continually improved, it became possible to start increasing the density of variation being genotyped, and to better comprehend the characteristics and patterns of variation in the human genome. Single Nucleotide Polymorphisms (SNPs) are the most common form of sequence variation in the human genome and thus it was a necessity to densely map these genetic variants. At the time, it was infeasible to whole genome sequence large cohorts of subjects in multiple populations, which would allow direct identification of most of the genetic variation within and across populations. Consequently, an alternative was pursued to catalogue common variants in the genome. This set of common genome variants could then be genotyped in any cohort of interest with the benefit of knowledge of population allele frequency and linkage disequilibrium. The underlying hypothesis behind this work was that based on the genotypes of common variants it should be possible to detect trait association signal at a genome-wide level and then proceed to search for causative variants within the localized region (Collins, Guyer and Charkravarti 1997). This approach also allows for a genome-wide search space so that researchers do not have to know putative functional information of all variants beforehand. Thus, candidate gene or region selections were not required a priori. This type of resource would provide a haplotype map of the human genome. Two of the key and related characteristics of genomic variation making this type of map possible are linkage disequilibrium (LD) and haplotypes. LD is the non-random association between markers of genetic variation in a genome and a haplotype is a set of co-segregating alleles on a chromosome. By contrast, when markers of genetic variation are independent of each other, such that no association exists between their alleles (i.e. no LD), these markers are considered to be in linkage equilibrium. Recombination, as discussed earlier, is the pairing of homologous chromosomes during meiosis where by sections of genetic material are copied from one chromosome to the other by breakage and re-joining; LD arises because of a lack of recombination between sites. Basing the catalogue on common SNPs the patterns of LD within this map should primarily reflect historical recombination and demographic events because the common SNPs will typically be older than rare SNPs (Chakravarti 1999). In 1999, Kruglyak published a study describing population simulations involving LD to guide the design of dense genotyping platforms. The designs of these platforms were based on using whole-genome LD to map common disease loci. In this study, LD between common variants was considered based on simulations of both general and isolated human populations. The results showed that the useful LD for mapping is not likely to extend beyond 3 Kb in the general population. These simulations suggested that approximately 500,000 SNPs would be required. This finding also held for isolated populations unless there was a significant founder bottleneck or the disease risk variant was not common (minor allele frequency less than 5%) (Kruglyak 1999). Another study was performed to find empirical evidence about the extent of LD in the human genome and whether it fits with the theory that was driving the design of possible whole-genome association studies based and SNP genotyping. This analysis was based on 38 variants with an allele frequency greater than 10%, under the assumption that these should be comparable to disease risk variants. These variants were from three regions on three different chromosomes, where previous variant mapping had already been performed. They genotyped these 38 variants in 1,600 subjects from four European populations. They found that the LD and allele frequencies among these populations were very similar and that there was an inverse relationship between LD and distance in general. Based on these finding they suggested that whole-genome genotyping scans for association studies would need variants spaced less than or equal to 5 Kb apart (Dunning *et al.* 2000).

Early cataloguing of common variation resulted in a map of 1.42 million single nucleotide polymorphisms released from multiple consortia, representing the most comprehensive map of human genome sequence variation at the time. The average density was one SNP every 1.9 kilobases (Kb) and 85% of exons were within 5 Kb of a SNP. This map was based on ethnically diverse populations (Sachidanandam *et al.* 2001). Using these denser maps of SNPs, a study was performed by Reich et al., to further characterize the size of LD blocks in three diverse populations. This study found that LD for Northern Europeans, around common SNPs, typically extended to 60 Kb but within an African population the typical size was reduced to 5 Kb. The authors suggested that these results may reflect a demographic event that occurred between 27-53 thousand years ago (Reich *et al.* 2001). LD is the correlation among proximal variants reflecting haplotypes descended from single, ancestral chromosomes. LD between genetic variants emerges as a result of selection or population history (population size, genetic drift, and population mixture), and decays because of recombination breaking up the ancestral haplotypes. Decay in LD is proportional to the number of generations since the ancestral haplotype formed. The presumed and simplest reason for the existence of long-range LD in a population is that a population under went a bottleneck, severe founder effect, or because of a lack of recombination. Such an effect can occur if a population is reduced so drastically in size that only a few ancestral haplotypes remained from which today's haplotype originated. LD in Europeans typically extends 60 Kb from common alleles whereas the Yoruban blocks are much smaller but primarily a subset of the blocks seen in Europeans. The smaller Yoruban haplotypes are within the longer European ones with very few specific to the Yorubans. This large difference in LD sizes while still being a subset suggests a population history event that occurred in Europeans after the divergence from ancestral Africans, likely a bottleneck or founder effect (Reich *et al.* 2001). Several other early studies demonstrated how LD and haplotypes confirm the feasibility of these cataloguing efforts based on variation from isolated regions of the human genome. Based on genotypes from a European population at 103 SNPs within 500 Kb on chromosome 5 it was observed that discrete haplotype blocks with limited diversity could be identified. These haplotype blocks were up to 100 Kb in size and typically contained between 2 and 4 haplotypes representing more than 90% of the genetic variation within the block, each block was flanked by sites of apparent recombination (Daly *et al.* 2001). In a similar study, based on common genetic variation from chromosome 21, it was also shown that haplotypes have a block like structure of limited diversity and suggested that 80% of the global human population can be characterized by three common haplotypes (Patil *et al.* 2001). It was also shown in another study that it was practical and possible to identify these common haplotypes based on fewer representative SNPs. These representative SNPs capture the pattern of LD for the adjacent markers in the haplotype, thereby tagging the haplotype (Figure 1.2). Thus, these representative SNPs were termed haplotype tag SNPs, ‘tag’ SNPs or htSNPs. This study of tag SNPs was based on 122 SNPs from nine genes covering 135 Kb of the genome in 384 European subjects. In identifying the htSNPs that capture the common haplotypes, it was found that the number of SNPs required to provide fine mapping in regions of high LD could be greatly reduced. Within this study the 122 SNPs could be reduced to 34 informative htSNPs (Johnson *et al.* 2001).

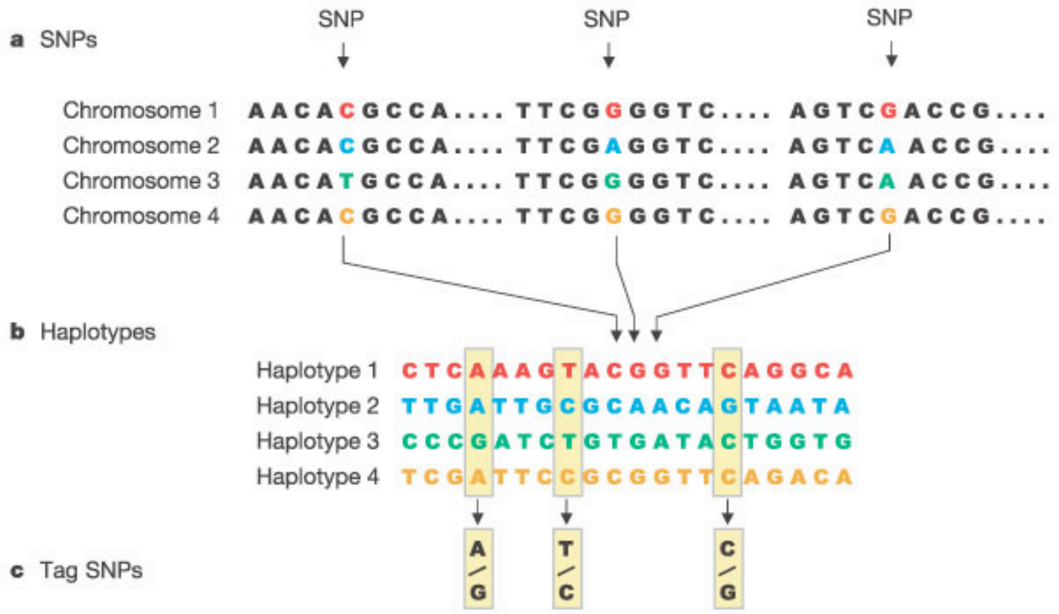

a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

Figure 1.2: Schematic of the relationship between SNPs, haplotypes, and haplotype tagging SNPs. a) SNPs present in DNA from four versions of the same small chromosomal region from different individuals. The majority of the DNA sequence in the region is identical; however, single nucleotide polymorphisms are present at three of the bases. Each of these SNPs is biallelic, where the first SNP's alleles are C and T and the 2nd and 3rd SNPs have G and A alleles. b) A haplotype is a set of co-segregating alleles in a chromosomal region. The three SNPs are within a larger region of variation, in this example the region has 20 SNPs, and four haplotypes are present in the population for this region. c) From the SNPs present, it is possible to identify a subset of three SNPs that tag (identify) these four haplotypes. In this example, for the three tag SNPs identified, if the combination A-T-C is present on a particular chromosome this pattern would match haplotype 1. This figure is reproduced from (International HapMap Consortium 2003).

Similar characteristics of haplotypes were also observed in another study over a larger portion of the genome and including subjects from different populations (Gabriel *et al.* 2002). This study also found that these regions could be parsed into haplotype blocks over large regions and containing only a few common haplotypes. This study was based on an analysis of 13 megabases (Mb) from 51 autosomal regions of the human genome and included subjects from Europe, Asia, and Africa that were successfully genotyped at 3,738 SNPs. These haplotypes also show some evidence of historical recombination events, with recombination events more frequent in African populations than in the European or Asian populations. There was also evidence that recombination rises more rapidly over a shorter genomic interval in African populations than European and Asian populations. The

African haplotype blocks averaged 9 Kb in size whereas European and Asian blocks were 18 Kb in size. The range of block sizes was also different between the populations, 1 - 94 Kb in the African population and 1 - 173 Kb for European and Asian populations. As seen in the previous studies, low haplotype diversity was observed with typically 3 to 5 common haplotypes that capture the majority of haplotypes in any region. These haplotypes could be identified based on 6 to 8 randomly chosen markers and additional common markers did not increase the identification of common haplotypes in regions with a low rate of historical recombination. The African population also had higher haplotype diversity averaging five haplotypes whereas the European and Asian populations averaged 4.2 and 3.5 respectively. Like the previous studies, these few common haplotypes still captured more than 90% of the diversity (Gabriel *et al.* 2002). The Gabriel *et al.* study found large regions containing both low and high rates of variation, as long as 100 Kb. They suggest that this arrangement is primarily driven by genealogical history with less than 25% being due to local mutation rate. This study provided a genome-wide estimate on the average correlation of variants (LD) as well as providing evidence that recombination hotspots are a general feature of the human genome and have a role in shaping genetic variation. Chromosomal regions inherited from shared ancestry and without recombination locks specific allele combinations in the population forming a haplotype. SNPs within a region of low recombination will track together in the population (Gabriel *et al.* 2002). In 2002, Dawson *et al.* published a study of LD patterns based on Centre d'Etude du Polymorphisme Humain (CEPH) families with replication in unrelated individuals from the United Kingdom. They found that the patterns of LD are highly variable across the genome. This study, based on 1,504 SNPs spaced on average every 15 Kb along chromosome 22 found large regions of almost complete LD interspersed with regions containing little to no LD. It was also observed that while LD decays with distance there is considerable variation in the size of each block. Some regions in almost complete LD spanned over 800 Kb while others regions smaller than 5 Kb contained almost no discernable LD. This study also observed a strong correlation between high LD and low recombination suggesting that recombination in humans is not random, that there are recombination hotspots, and that historical and contemporary recombination rates are similar (Dawson *et al.* 2002). In some rare instances LD can be quite strong over very large regions such as the ~2 megabase region on chromosome 17 flanking *MAPT* (Pittman *et al.* 2004). The *MAPT* region appears to completely lack recombination between the two major haplotypes present in the region. This large block of LD and considerable divergence between the haplotypes is apparently the result of a large inversion, ~900 Kb in size. The inversion impedes recombination between the haplotypes, with the inversion haplotype (H2) and the non-inversion haplotype (H1) showing no evidence for recombination within their study (Stefansson *et al.* 2005). This region is of particular interest as it is associated with several neurodegenerative diseases and is investigated throughout this thesis.

To formalize and facilitate the cataloguing of common genetic variation in the human genome the HapMap Project was formed. This created a public resource that characterizes common sequence variants (initially more than a million), their allele frequencies and the associations between them (i.e. their LD structure) based on four populations from Europe, Asia, and Africa

(International HapMap Consortium 2003, 2005). This resource has had three primary releases termed Phase 1, 2 and 3 since inception. Phase 1 was the initial release based on subjects from four populations and included 1.1 million common SNPs. The Phase 2 release expanded the density of genetic variants to 3.1 million SNPs common within the four HapMap Project populations (International HapMap Consortium *et al.* 2007). The Phase 3 release of the HapMap project expanded the number of populations surveyed from four populations to 11 global populations (Altshuler *et al.* 2010). In addition to expanding the number of populations genotyped for common SNPs, the Phase 3 release included copy number polymorphisms as well as rare variants identified by sequencing in select regions of the human genome. Copy number polymorphisms, more commonly referred to as copy number variants (CNVs), are structural variants where the number of copies of a portion of the genome is aberrant. The regions selected for sequencing are from the ENCyclopedia Of DNA Elements (ENCODE) regions. The ENCODE project was established to identify the functional elements within the human genome. The pilot phase of the ENCODE project focused on ~1% (30 megabases) of the human genome to specifically target functional elements including; genes, promoters, enhancers, transcription factor binding sites, DNase I hypersensitive sites, methylation sites, chromatin modifications, and multi-species conserved sequences (ENCODE Project Consortium 2004). The public release of these resources allowed for the rapid and continually improved design of assays which could genotype hundreds of thousands of informative tag SNPs allowing for effective genotyping in genome-wide association studies (GWAS) to be performed.

Inferring membership in a population based on genetic variability, resulting from the ability to densely genotype large sample cohorts, has allowed the study of populations based on genetics to move from theory to empirically driven findings (Pool *et al.* 2010). The method most commonly used to study population genetics is principal component analysis (PCA), which was proposed decades ago (Menozzi, Piazza and Cavalli-Sforza 1978) and renewed for use with studies based on high-density genotyping (Patterson, Price and Reich 2006; Price *et al.* 2006). The study from Monizzi *et al.* used multi-dimensional scaling (MDS) and PCA to spatially condense and show population structure based on genetic variation. They did so using 38 alleles from 10 loci in Europeans and Asian populations and found that the results map matched expectations from the hypothesis that early farming in Europe was a result of new migration rather than a technology diffusion into the population (Menozzi, Piazza and Cavalli-Sforza 1978). MDS is a method for information visualization, particularly for distance metrics, which aims to place each item in N-dimensional space such that the between-object distances are preserved as well as possible. PCA is a transformation that converts possibly correlated variables into linearly uncorrelated components, where the 1st component accounts for largest variance, 2nd the second most, etc. In a later study of spatial variation using PCA, it was found that the use of PCA with spatial data results in gradients that are a general sinusoidal mathematical artefact and therefore may not necessarily reflect specific migrations. However, the authors did find that using PCA does help correct for population structure in association studies (Novembre and Stephens 2008). Studies have also shown that beyond recent migration that linkage patterns can show additional historic information (Davison, Pritchard and Coop 2009), such as haplotype frequency and haplotype number with changes in population sizes (Lohmueller, Bustamante and Clark 2009). Additionally, it has been shown that using clusters of linked mutations can detect archaic population structures (Plagnol and Wall 2006). These studies suggested that in addition to haplotype patterns reflecting recent migration that small haplotype patterns reveal older gene flow and demographic events (Pool *et al.* 2010). Dense whole-genome genotyping also allows for further analysis of natural selection both negative and positive. Negative selection reduces genomic variation by removing variants, maintaining low frequencies for variants or by removing variants linked to damaging alleles (background selection) (Charlesworth, Morgan and Charlesworth 1993). Positive selection results in local reductions in diversity by “genetic hitchhiking”. Hitchhiking is when an advantageous variant’s frequency increases in population and neutral variants linked to the positive variant will be lost or become fixed along with the variant in the population, known as a “selective sweep”. The size of the selective sweep is affected by recombination rate and selection strength (Smith and Haigh 1974; Hudson and Kaplan 1988; Stephan, Song and Langley 2006). Spatial patterns of LD are produced by selective sweeps and represent hitchhiking signals that differ from stochastic patterns resulting from bottlenecks (Stephan, Song and Langley 2006; Jensen *et al.* 2007). These specific LD patterns may also reveal partial selective sweeps detected by the imbalance of haplotype homozygosity. Comparing haplotype homozygosity can also be used to detect selective sweeps that are population-specific (Sabeti *et al.* 2002, 2007; Voight *et al.* 2006).

The study of genetic variation in world populations has continued to increase in the number of populations considered and the amount of genetic variation that is assayable with expanding throughput and types of assays. Such a study of genetic variation in world populations was performed by Jakobsson *et al.*, and examined SNP, haplotype, and copy-number variation (CNV). This study was based on the genotyping of ~500,000 SNPs in 29 world populations. The analysis found that dense SNP genotypes allow for fine-scale inferences of population structure and that using haplotype analysis methods also revealed these same fine-scale inferences. The analysis of CNVs showed that they could also be used for detecting population structure but to a lesser degree. The results from the SNP, haplotype, and CNV analyses showed that increased LD patterns matched increases in geographic distance from Africa. The authors suggest that this LD increase may be expected based on serial founder effects of the out of Africa spread of human populations (Jakobsson *et al.* 2008). Another study, based on the dense genotyping of a large number of samples from multiple European populations, found that patterns of genetic variation also exist within spatially close populations. In this study, the authors considered variation from 3,000 Europeans based on genotypes at ~500,000 variants. They found that even though there are low levels of differentiation, within European populations, they did find correlation between genetic and geographic distances. These genetic correlations within European populations also aligned geographically to reveal a picture of European genetic diversity that matches European geographic maps when projected as a two dimension summary plot. Their results reinforce the idea that fine-scale population structure, based on genetic distances, needs to be accounted for when doing analysis for disease traits even within a population determined by genetic ancestry. Their results revealed a south-east to north-west axis within European populations based on genetic distance where haplotype diversity decreases from south to north (Novembre *et al.* 2008).

More recently it has become possible and feasible to directly sequence large cohorts of subjects using high-throughput short-read sequencing. It is now feasible to directly identify most of the genetic variation in a large number of subjects and this has been done in large public consortia such as the 1000 Genomes projects (1000 Genomes Project Consortium *et al.* 2010, 2012). These denser maps of genetic variation, in multiple populations, have also greatly improved our ability to impute genotypes, using the refined population haplotypes generated from these data. The ability to determine which haplotype an individual belongs to over short intervals, using a reference population(s) haplotypes, also makes it is possible to predict the genotypes within the interval, with a given probability, for variants that were not directly genotyped in the individual. This statistical inference of variants not originally genotyped from variants that were genotyped is an imputation of those unobserved genotypes through estimating the individual's haplotype and using LD to predict these genotypes. A study based on a cohort of ~72,000 parent-offspring pairs from Iceland (using imputation from whole-genome sequencing of 2,200 subjects), was undertaken to identify variants associated with recombination rate. Meiotic recombination yields new combinations of alleles contributing to genetic diversity and an individual's recombination counts vary in their gametes. They found 13 variants from eight regions associated with genome-wide recombination rate. Eight of 13 variants were previously unknown; three of these variants were male only, seven were female only, and three were for both. Two of these variants are low-frequency with large effects on recombination rates, with one of these increasing the male genetic map by 111 cM and the female map by 416 cM and is located in an intron (Kong *et al.* 2014). Another recent paper estimated the ages of rare variants, based on whole-genome sequencing of subjects in the 1000 Genomes Project. They found that the ages of rare variants are related to population histories and can be estimated by haplotype sharing patterns. Their analysis allows for estimating the age of each haplotype. Notably in considering haplotypes shared within and between populations, the ages of these haplotypes are consistent with known historical relationships among the populations. Their findings suggest that the age of haplotypes carrying variants that occur twice in populations, based on the populations represented in the 1000 Genomes Project, is 50 to 160 generations in Europe and Asia and 170 to 320 generations in Africa. They also note that haplotypes shared between continents (Europe and Asia) are much older, from 320 to 670 generations. When they considered the distribution of haplotypes containing these twice-occurring variants they suggest this pattern shows demography, recent bottlenecks, ancient splits, and modern mixture of populations. They also found that functional variants are younger than non-functional variants of the same frequency suggesting this is an effect of selection (Mathieson and McVean 2014).

A recent study of human population size and separation of populations was performed based on human genome sequences. The authors developed a method to analyse population separations that occurred less than 20,000 years ago. Their results suggested that genetic separation between non-African and African populations started long before 50,000 years ago. The analysis was also informative for more recent events including: population separations within Africa, Asia, and Europe; and bottlenecks in the early Americas (Schiffels and Durbin 2014). Another recent study considered whether selection is less effective at removing damaging mutations in Europeans than in Africans. They undertook the study to examine the hypothesis that since European populations have undergone size reductions since the split from West Africans that the removal of weakly deleterious mutations by natural selection would be less effective. Based on per-genome accumulations of nonsynonymous variants they found no evidence of higher amount in non-Africans. However, looking at more divergent populations they found that Denisovans did accumulate nonsynonymous mutations faster than both Neanderthals and modern humans (Do *et al.* 2015). Another study was recently performed considering recombination and its effects on the accumulation of damaging and disease associated mutations in humans. In non-recombining species, damaging mutations can accumulate potentially leading to the extinction of many asexual species. This study examined the accumulation of damaging mutations within chromosomes that have variable crossovers rates, based on 1,400 subjects. They found that recombination rates affect the distribution of damaging variants across the genome. Their results showed that exons in regions with low recombination rates are enriched for damaging variants, but this varies across populations with different demographic histories. Their results also suggest that new damaging mutations occurring in regions with higher recombination rates will more efficiently be removed by natural selection than mutations in regions with lower recombination rates. Regions of the human genome with lower recombination rates are enriched for conserved genes with essential functions such as cell cycle progression, mRNA processing, and DNA repair. The authors conclude that this co-enrichment of damaging variants and conserved genes with essential function likely affect human disease susceptibility (Hussin *et al.* 2015).

1.2.2: Gene Expression

The expressed human genome is also variable and this variation is likely to contribute to health related phenotypes, where variation of expression is likely to be an intermediate phenotype. With the ability to profile the relative mRNA transcript abundance (i.e. gene expression) for most of the known and predicted transcriptome in parallel, in a high throughput manner, we can begin to understand the patterns of expression variation. Understanding gene expression variation in multiple contexts is of importance and relevant to our understanding of molecular biology in general, and as it relates to health. The context of the variation is also important; gene expression varies not only in whether or not a gene is expressed in a particular cell or tissue type, but how much is expressed, how variable is this level of expression temporally and in relation to stimuli as well as to which particular transcripts and alternate splice forms are expressed. The last two decades have seen a great maturation in the technologies, assays, and methods that allow for the reproducible measurements of mRNA transcripts. These developments have allowed the field to begin to catalogue and better understand gene expression.

1.2.2.1: Regulation of Gene Expression

In eukaryotes, the basal or core promoter is required for transcription but by itself cannot result in high levels of expression (Wray *et al.* 2003). The core promoter is the genomic region near the transcription start site (TSS), typically +/- 40 base pairs, and is the site where the transcription machinery assembles (Figure 1.3A) (Yáñez-Cuna, Kvon and Stark 2013). This assembly includes transcription factors that are bound to sites they have affinity for in the promoter region, which may affect the specificity and frequency of transcription (Kuras and Struhl 1999; Lee and Young 2000; Lemon and Tjian 2000; Wray *et al.* 2003). The rate of transcription initiation is considered to be the primary point of control for regulation of gene expression in eukaryotes but other important mechanisms are also involved including: chromatin accessibility, DNA methylation, pre-mRNA splicing, mRNA stability, translation, post-translation modification, and degradation (Lemon and Tjian 2000; Wray *et al.* 2003). As transcription initiation is the primary control point of gene expression I will focus almost exclusively on transcription within my thesis. The assembly that initiates transcription is the RNA polymerase II holoenzyme complex (RNA Pol II). This complex is made up of approximately 12 proteins and is responsible for the transcription of genes in eukaryotes (Orphanides, Lagrange and Reinberg 1996; Lee and Young 2000; Wray *et al.* 2003). The RNA Pol II complex assembles at the core promoter, which is a sequence region in close proximity (5') to the transcription start site (TSS). While core promoter sequences differ for genes, common DNA elements that can be found in human core promoters include: CpG islands, the initiator element (Inr), the TATA-box, the TFIIB recognition element (BRE), and the downstream promoter element (DPE) (Sandelin *et al.* 2007; Yang *et al.* 2007).

CpG (5’-cytosine-phosphodiester bond-guanine-3’) islands are genomic regions enriched with CG dinucleotide content compared to the rest of the genome. CpG island promoters are most often associated with ubiquitously expressed genes, although they are also associated with tissue-specific genes including brain-specific genes (Schug *et al.* 2005; Gustincich *et al.* 2006). Promoters with CpG islands are the most common in the human genome, it is estimated that 72% to 76% of human promoters contain CpG islands (Saxonov, Berg and Brutlag 2006; Yang *et al.* 2007). Promoters with a TATA-box sequence motif are bound by the TATA-binding protein (TBP), which in combination with other TBP-associated factors brings the RNA Pol II complex to the DNA (Figure 1.3B) (Reinberg *et al.* 1998; Lee and Young 2000; Wray *et al.* 2003). TATA-box promoters are typically associated with genes that have tissue- or context-specific expression and only represent ~10% of human promoters, although ~24% of human promoters have TATA-like elements (Carninci *et al.* 2006; Ponjavic *et al.* 2006; Yang *et al.* 2007). Promoters with an initiator element (Inr) contain a consensus sequence motif that is distinct from the TATA-box motif, but both the Inr and TATA-box elements can be found together in some promoters and work together to recruit the transcription intiation complex. It has been estimated that ~46% of human promoters contain the mammalian consesus sequence Inr and that ~30% of promoters with an Inr element are TATA-less (Yang *et al.* 2007). The BRE (Transcription Factor II B) is a consensus sequence motif located upstream of the TATA-box and is present in 12% to 25% of human promoters and typically acts to increase or decrease transcription rates (Lagrange *et al.* 1998; Yang *et al.* 2007). The DPE is a consensus sequence motif found downstream of the TSS (typically 30 bases) in promoters that also have an Inr and are present in 12% to 25% of human promoters (Yang *et al.* 2007). Genes may have multiple core promoters and each of these promoters may initiate transcription at a different TSS. CpG islands are associated with promoters that have a broader distribution of TSSs while TATA-box promoters are associated with only one or a few consecutive nucleotides as TSSs (Carninci *et al.* 2005; Sandelin *et al.* 2007). The TSS does not have a specific sequence motif, as the translation start site does, but instead is determined by the second DNA contact point for the RNA Pol II complex and is approximately 30 bp downstream of the first RNA Pol II contact point (Wray *et al.* 2003). Many of the proteins that bind into the core promoter are ubiquitously expressed and known as general transcription factors; while others are known to have isoforms that are tissue-specific (Holstege *et al.* 1998; Wray *et al.* 2003). TATA-box promoters are bound by the TATA-binding protein (TBP) transcription factor while CpG island promoters are enriched for transcription factor or transcription factor family binding motifs including: E twenty-six (ETS), E2F, nuclear respiratory factor 1 (NRF-1), specificity protein 1 (SP1), cAMP response elements (CRE), and E-box (Rozenberg *et al.* 2008; Landolin *et al.* 2010).

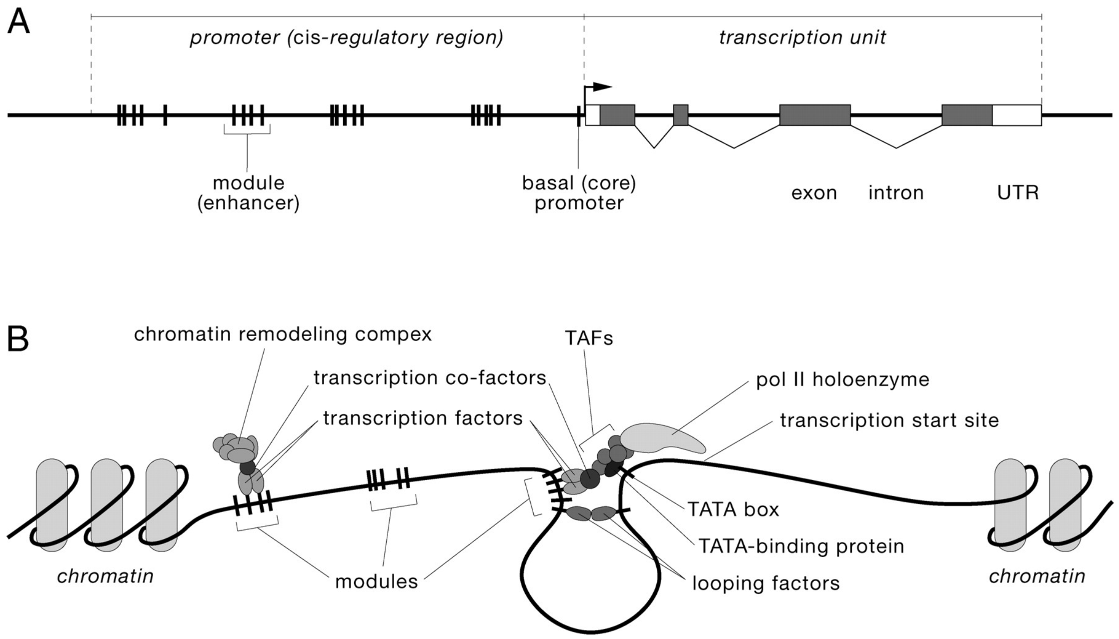


Figure 1.3: Simplified general schematic of the *cis*-regulatory region of a gene and a cartoon of this region at transcription initiation. A) General linear organization of a gene and its promoter region. The *cis*-regulatory region is located proximal (region of the left in this simplified schematic) to the transcription start site (TSS; black arrow in the centre) and the transcriptional unit is on the right. The core promoter is located near the TSS and transcript factor binding sites are interspersed within the regulatory region (indicated by vertical bars) and typically found in modular units (enhancers), which can be located both up and downstream of the TSS. B) Cartoon of a gene's promoter region during transcript initiation, for a gene with a TATA-box in the promoter. The chromatin is open so that the promoter region is available for interaction with regulatory proteins. RNA Pol II has assembled at the core promoter. Transcription factors are bound to binding sites and looping factors have brought some of these factors into proximity of the core promoter so they may interact with other regulatory factors. This figure is reproduced from (Wray *et al.* 2003).

Achieving increased levels of transcription beyond the low level possible with the complex forming at the core promoter typically requires other transcription factors bound to sites outside the core promoter. The presence of these other transcription factors and other cofactors in the nucleus can be temporal and differ among cell types (Lemon and Tjian 2000; Wray *et al.* 2003). For example, paired box 6 (PAX6) is a regulatory protein with temporal and abundance variation and is important during the development of neural tissues and the eye (Kammandel *et al.* 1999; Wray *et al.* 2003). Regulatory proteins affect transcription by influencing how often the RNA Pol II complex assembles onto the core promoter. This influence can be through protein-

protein interactions, where a transcription factor may interact to increase or decrease transcription rates by activation and repression domains (Torchia, Glass and Rosenfeld 1998; Wray *et al.* 2003). Studies of eukaryotic promoters suggest that there may typically be 10 to 50 binding sites for 5 to 15 different transcription factors in a typical promoter (Arnone and Davidson 1997; Wray *et al.* 2003). Most binding sites for transcription factors are 5 to 8 base pairs in length, but the “footprint” of the bound transcription factor on a segment of DNA is typically 10 to 20 base pairs. The 5 to 8 base pair binding motif may tolerate some polymorphic changes without losing functionality but the binding site motifs that a transcription factor binds can change in the presence of different binding partners (Wray *et al.* 2003). Transcription factors can also affect the binding of other factors by binding to sites such that they block the binding of another transcription factor at an adjacent site. The binding of other transcript factors at proximal sites can modulate the process, which allows gene regulation to be a dynamic and tuneable process (Jackson-Fisher *et al.* 1999; Kuras and Struhl 1999; Lee and Young 2000; Lemon and Tjian 2000; Wray *et al.* 2003). Transcription factors typically contain several functional domains including: DNA-binding, protein-protein interaction, intracellular trafficking, and ligand-binding domains (Abu-Shaar, Ryoo and Mann 1999; Carrión *et al.* 1999; Wray *et al.* 2003). The DNA-binding domains for most transcription factors are short motifs, typically five bp, and transcription factors may contain multiple DNA-binding domains. For these short motifs, a single amino acid change in the domain can alter its binding specificity (Treisman *et al.* 1989; Wray *et al.* 2003). The small size of these binding motifs also means the binding target motifs can occur often in the genome resulting in a lack of sequence specificity and many potential target sites. Because transcription factors may bind to multiple motifs and these sites may occur often in the genome, this means that many copies of the factor must be present in the nucleus for binding to occur at specific sites (Wray *et al.* 2003). Additionally, a transcription factor's specificity can be strongly modulated through cofactors or by post-translational modifications such as phosphorylation (Knoepfler and Kamps 1995; Berthelsen *et al.* 1998; Dröge and Müller-Hill 2001; Wray *et al.* 2003). Regulation of transcription can also be influenced by bound transcription factor(s) altering the chromatin structure through DNA methylation and histone modifications such as acetylation, where this remodelling can take place in a small timescales and in small regions such as a promoter or even within a promoter (Kadosh and Struhl 1998; Jones and Takai 2001; Richards and Elgin 2002; Wray *et al.* 2003). Before transcription the chromatin surrounding the core promoter and some of the transcript must be decondensed so that transcription factors can bind and recruit RNA Pol II to the core promoter (Reinberg *et al.* 1998; Wray *et al.* 2003). There are regulatory proteins, referred to as pioneer transcription factors that can initiate regulatory events in chromatin. These factors may bind cooperatively or sequentially, and open up local chromatin so that other factors can bind (Zaret and Carroll 2011). The forkhead box (FOX) proteins are an example of pioneer transcript factors. The FOX proteins mediate fine-tuning of spatial and temporal expression of genes during development and in adult tissue (Lam *et al.* 2013).

The size of *cis*-regulatory regions may vary from a few hundred bases to more than 100 kilobases (Kb). In some instances, the *cis*-regulatory region can be much further from the TSS. Such is the case for *Shh* locus in humans and mice where the *cis*-regulatory region is ~800 Kb from the TSS (Lettice *et al.* 2002; Wray *et al.* 2003). The position of *cis*-regulatory transcription factor binding sites may also vary relative to the TSS. These binding sites are typically within a few Kb upstream (5') of the core promoter, but they can also be found much further upstream as well, or in other instances within the 5' untranslated region (UTR), in introns, or downstream (3') of the gene and in rare instances in exons (Wray *et al.* 2003). A promoter's transcriptional yield is not simply based on which binding sites are present but also involves the sequence, relative position and orientation of the binding site as well as the expression of other transcription factors and cofactors. Thus interactions are complex and context dependent. Groups of transcription factor binding sites can operate as a unit or a functional module to yield a distinct aspect of a transcript's expression profile (Dynan 1989; Arnone & Davidson 1997; Wray *et al.* 2003). These functional modules may initiate transcription, increase transcription, mediate transcription signals, repress transcription, or modulate other functional modules; these modules are referred to as enhancers (Atchison 1988; Wray *et al.* 2003). These enhancers can be functionally related to the regulation of a transcript by way of DNA looping (Figure 1.3B). Regulatory proteins bound to DNA can affect the bending or looping of DNA, allowing other factors bound distally to be near each other for interaction, and this function may be necessary for the transcription factor to act as an activator or repressor (Fry and Farnham 1999; Scaffidi and Bianchi 2001; Wray *et al.* 2003). The looping of DNA allows for the interaction of these binding proteins even at distal sites (Simon *et al.* 1990; Neznanov, Umezawa and Oshima 1997; DiLeone, Russell and Kingsley 1998; Nielsen *et al.* 1998; Kammandel *et al.* 1999; Bamshad *et al.* 2002; Calhoun, Stathopoulos and

Levine 2002; Yuh *et al.* 2002; Wray *et al.* 2003). Binding sites that are not near the core promoter may interact with the preinitiation complex at the promoter through DNA looping or bending and as a result their distance and orientation is relatively independent of the target TSS (Wray *et al.* 2003; Yáñez-Cuna, Kvon and Stark 2013; Core *et al.* 2014; Shlyueva, Stampfel and Stark 2014). In general regulatory *cis*-architecture, the promoter is in the immediate proximity of the TSS and binds the preinitiation complex. Enhancers are typically less constrained in their genomic context and may act in a more cell-specific manner and help to bring specific transcription factors to the preinitiation complex at the promoter (Yáñez-Cuna, Kvon and Stark 2013; Shlyueva, Stampfel and Stark 2014). The interaction of transcription factors bound at distant sites through looping potentially allows for transcription at multiple genes to be affected. These distal interactions can be spatially restricted through insulators or bound elements likely involving chromatin modifications (Wolffe 1994; Bell and Felsenfeld 1999; Dillon and Sabbattini 2000; Wray *et al.* 2003). Typically, binding sites effect the expression of one gene, but instances of shared regulation also occur. Additionally, there are instances where genetic variation in shared regulatory regions affect the expression of multiple genes, for example: beta and gamma-globin (Metherall, Gillespie and Forget 1988; Grosveld *et al.* 1993); insulin and insulin-like growth factor 2 (*IGF2*) (Paquette *et al.* 1998); and apolipoprotein A-I (*APOA1*) and apolipoprotein C-III (*APOCIII*). In *APOA1* and *APOCII* the effect of variation in the shared regulatory region is tissue-specific; down-regulation of *APOA1* in colon and up-regulation of *APOCIII* in liver (Li *et al.* 1995; Naganawa *et al.* 1997; Wray *et al.* 2003).

Enhancers and repressors may be in introns and *cis*-regulatory elements may extend long distances both up and downstream of the transcribed sequence. Some evidence for this is that many noncoding regions show strong conservation and functional studies of these regions suggest they are regulatory elements generally containing sites for tissue-specific DNA-binding proteins (Kleinjan and van Heyningen 2005). In 2003, Nobrega *et al.* published a study where gene deserts were searched for conserved enhancers that modulate expression. In this study, the authors found conserved elements for human dachshund family transcription factor (*DACH*), which is flanked by two large gene deserts. They considered a 2.6 megabase (Mb) region where they found and validated enhancer elements for *DACH* in a 1.5 Mb region. These conserved elements were estimated be to approximately 1 billion years old. Based on mouse reporter assays they showed that these conserved elements were long-range enhancers driving expression. *DACH* is expressed in many tissues and involved in development of brain, limbs and sensory organs (Nobrega *et al.* 2003). Conversely, in a recent study, by Jacques *et al.*, considering the possible origins of primate-specific regulatory elements, it was found that transposable elements (TEs) appear to have contributed to many primate-specific regulatory elements. Transposable elements are DNA sequences that are mobile within a genome, where the location of the sequence can change or be duplicated to another location in a genome. This study used ENCODE DNase I hypersensitivity site (DHS) data from multiple cell types and found that 44% of these TEs were in open chromatin and this number increased to 63% when considering TEs in primate-specific regions. They also found that 80% of endogenous retroviruses (ERV), a specific subfamily of TEs, were also in open chromatin and that their derived sequences were activated in a cell-specific manner and were associated with nearby genes (Jacques, Jeyakani and Bourque 2013). In a recent study, from Heidari *et al.*, interaction maps of regulatory elements were constructed based on ENCODE data from human cells for 80% of DHSs, which included 99.7% of TSSs and 98% of enhancers. They found that cohesin, CCCTC-binding factor (*CTCF*), and zinc finger protein 143 (*ZNF143*) are proteins that are key contributors to the three-dimensional (3D) structure of chromatin and how distal chromatin state can affect transcription. When analysing these structural interactions, between cell types, they found that many enhancer-promoter interactions were cell-type specific. Additionally, they found that housekeeping genes are enriched for proximal events whereas distal events included genes involved in dynamic biological processes (Heidari *et al.* 2014). The 3D structure of chromatin that organizes the genome into functional regulatory compartments at a megabase scale, where regulatory elements such as promoters and enhancers can interact, are referred to as topologically associating domains (TADs). In 2012, Dixon *et al.* published a study of 3D genome organization and chromatin interaction. This study examined the 3D organization of genomes in human and mouse embryonic stem cells and differentiated cell types. They found that these megabase-sized local chromatin interacting domains were a pervasive structural feature of genome organization stable across different cell types and conserved across species. Additionally, they found that the boundaries of these domains were enriched for insulator binding protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons (Dixon *et al.* 2012). In another recent study, from Vierstra *et al.*, considering more than 1.3 million DHSs from 45 mouse tissue and cell lines, a comparison to human orthologous DHSs was performed. The authors found extensive *cis*-regulatory changes that appear to be mediated by turnover of transcription factor recognition elements during evolution. However, despite these pervasive changes to individual *cis*-regulatory regions within DHSs shared between mouse and human, 58.7% of transcription factor binding sites were conserved (Vierstra *et al.* 2014). In 2013, Sheffield *et al.* published a study considering how regulatory elements and transcription factors affect gene expression across cell types. The authors were able to develop a classifier based on 43 DHSs that could predict cell-type lineage. This study was based on 112 human samples for 72 cell types to analyse DHSs, promoters, CpG islands, and transcription factor motifs (Sheffield *et al.* 2013). In 2013, Xie *et al.* published a study of DNA methylation of TEs in human embryonic and adult tissues. TEs comprise a very large portion of the human genome and it is assumed that many of them are hypermethylated and inactive. They found that ~ 10% of TE families are hypomethylated in a tissue-specific manner. The regions containing hypomethylated TEs were proximal to genes that shared function important to the tissue and many showed enhancer activity. These findings suggest that TEs are responsible for setting up tissue-specific regulation and have tissue-specific epigenetic regulation (Xie *et al.* 2013).

The genetic basis of many adaptations are likely a result of variation in *cis*-regulatory sequences, *cis*-regulatory variants are more likely to affect certain kinds of traits than coding variants (Wray 2007). That *cis*-regulatory variation may intrinsically be able to affect certain traits centres on the flexibility that regulation is afforded as a dynamic process (Wray 2007). This flexibility allows for more context-dependent effects, whereas a coding variant would likely have a more static effect without context (Jacob and Monod 1961; Stern 2000; Wray 2007). It has been proposed that natural selection may be more efficient for *cis*-regulatory variants than coding variants (Stern 2000; Wray *et al.* 2003). The basis of this hypothesis has two parts: *cis*-regulatory variants are often co-dominant and *cis*-regulatory regions are modular (Wray 2007). In diploid organisms each allele is transcribed independently, based on allele-specific measures of transcript abundance (Ruvkun *et al.* 1991; Pastinen *et al.* 2004; Wittkopp, Haerum and Clark 2004; Ronald *et al.* 2005; Wray 2007). The independence of allele transcription allows for *cis*-regulatory variants to be co-dominant. Under co-dominance, heterozygote variants have fitness costs making these variants visible to selection before the allele frequencies reach a point were homozygotes emerge in the population. Therefore, if *cis*-regulatory variants are more co-dominant than coding variants, then natural selection will be more efficient in functional non-coding regions than coding regions (Ruvkun *et al.* 1991; Wray 2007). Additionally the modular nature that is often present in *cis*-regulatory regions suggests that a variant can occur in a portion of the region and possibly have some effect on a certain aspect of a transcript's expression profile but not likely a total effect on overall expression (Force *et al.* 1999; Stern 2000; Wray 2007). Conversely a nonsynonymous coding variant would very likely have an affect on protein function regardless of its expression context. This allows selection to operate more efficiently through a reduction in functional trade-offs especially in the context of gene expression in different tissues and cell types (Force *et al.* 1999; Stern 2000; Wray 2007). Sequence comparisons studies have shown that the number of conserved intergenic and coding nucleotides is similar. This similarity may suggest that the number of functional noncoding and protein-coding nucleotides is roughly equal, and much of this noncoding sequence may be phenotypically penetrant (Onyango *et al.* 2000; Frazer *et al.* 2001; Shabalina *et al.* 2001; Wray 2007). Conservation of promoter sequence would suggest that gene expression is modulated through stabilizing selection (Cavener 1992; Stone and Wray 2001; Wray 2007). It is also known that promoter sequences can rapidly diverge. For instance in a study of 20 regulatory regions between humans and rodents, it was found that a third of the binding sites that are functional in humans are probably not functional in rodents (Dermitzakis and Clark 2002; Wray 2007). Known *cis*-regulatory variants contribute to phenotypes that differ between closely related species such as the ability of human adults to digest lactose (lactose persistence), which is not present in other great apes. This human dietary adaption is thought to have evolved in the past 2,000 to 20,000 years (Swallow 2003; Bersaglieri *et al.* 2004; Wray 2007). Lactose persistence is linked to *cis*-regulatory genetic variation that increases the transcription of lactase (*LCT*). This *cis*-regulatory genetic variation for *LCT* is different between European and East African populations, but confers lactose persistence in both through the same regulatory landscape, which is located in an intron of maintenance deficient 6 homologue (*MCM6*) just upstream of *LCT* (Olds and Sibley 2003; Bersaglieri *et al.* 2004; Tishkoff *et al.* 2007; Wray 2007). A recent study of transcription factor binding in human and chimpanzees suggest that natural selection through regulatory genetic variation greatly affected transcription factor (TF) binding sites between the species 4 to 6 million years ago. This study found that on average transcription factor binding sites show weaker selection than protein coding nucleotides, but binding sites of several transcription factors show evidence of adaptation (Arbiza *et al.* 2013). In 2010, Kasowski *et al.* published a study of speciation and phenotypic diversity based on gene expression. This study considered transcription factor binding in humans and chimpanzees based on LCLs. They specifically looked at RNA Pol II and nuclear factor kappa B (NFkB) based on ChIP-seq data. ChIP-seq is a method that combines chromatin immunoprecipitation (ChIP) with DNA sequencing to identify binding sites of protein-DNA interactions. They found that within humans between 7% and 25% of binding regions for NFkB and RNA Pol II differed, and these differences were associated with SNPs or structural variants that were often correlated with changes in gene expression. Additionally, when comparing RNA Pol II binding they found extensive divergence in transcription factor binding between humans and chimpanzees for 32% of the binding regions (Kasowski *et al.* 2010). An example of a gene that shows expression differences between humans and chimpanzees as a result of regulatory variation is prodynorphin (*PDYN*). This gene is involved in the release of a neuropeptide linked with memory, emotional status, and pain perceptions. For *PDYN* there are functional *cis*-regulatory variants present in humans that are not present in chimpanzees that are linked to changes in gene expression (Wray 2007). Additionally, *PDYN* has been linked through expression analysis and genetic association to schizophrenia, bipolar disorder, and temporal lobe epilepsy; and show signs of positive selection in human evolution and balancing selection between populations (Peckys and Hurd 2001; Hurd 2002; Stögmann *et al.* 2002; Ventriglia *et al.* 2002; Rockman *et al.* 2005; Wray 2007). Other human behavioural and cognitive traits have also been linked to *cis*-regulatory variants, for the genes encoding: arginine vasopressin receptor 1A (*AVPR1A*) (Bachner-Melman *et al.* 2005; Hammock and Young 2005), 5-hydroxytryptamine (serotonin) receptor 2A, G protein-coupled (*HTR2A*) (Enoch *et al.* 1998), monoamine oxidase A (*MAOA*) (Caspi *et al.* 2002; Kim-Cohen *et al.* 2006), and solute carrier family 6 (neurotransmitter transporter), member 4 (*SLC6A4*) (Trefilov *et al.* 2000; Hariri *et al.* 2002; Bachner-Melman *et al.* 2005).

Genetic variants can affect regulatory binding sites in multiple ways; by creating or eliminating the site, modifying the site such that it becomes the target of a different transcription factor, or changing the spacing between sites (Belting, Shashikant and Ruddle 1998; Segal, Barnett and Crawford 1999; Trefilov *et al.* 2000; Rockman and Wray 2002; Wray 2007). Additionally, variants may have a *trans* effect. *Trans* effects could be through a *cis* effect on a transcription factor's expression levels. A *trans* variant may affect the DNA-binding domain of a transcription factor. A *trans* variant may also affect a protein-protein interaction domain of the transcription factor. These *trans* effects may impact the expression of many genes in many tissues (Dawson, Morris and Latchman 1996; Manzanares *et al.* 2000; Brickman *et al.* 2001; D'Elia *et al.* 2001; Wray 2007). Variants in *cis*-regulatory regions may alter transcription, but this alteration may not be carried through to the protein abundance as the regulation of gene and protein expression is a network of interacting genes and this network can modulate protein abundance typically through feedback loops (von Dassow *et al.* 2000; Milo *et al.* 2002; Wray 2007). In a study of translational control of messenger RNA (mRNA) by microRNAs (miRNA), it was shown that a single miRNA can repress the levels of hundreds of proteins. The repression effect is typically mild, and in addition to down regulation of mRNA may also affect its translation. Regulation of mRNAs by miRNAs may inhibit translation by inducing degradation. This study measured protein synthesis (pulsed stable isotope labelling with amino acids in cell culture, pSILAC) and mRNA expression (microarray) changes in response to transfecting in miRNA or lowering levels of endogenous miRNAs (Selbach *et al.* 2008). Another study of the impact of miRNA on expression was performed using quantitative mass spectrometry for protein measures. MicroRNA are an endogenous species of RNA typically 23 nucleotides in length that bind to target sites in mRNA and down regulate these targeted mRNAs. The authors perturbed their systems by adding miRNAs to cultured cells after deleting mir-223 in mouse neutrophils. Their data suggested that targeted binding sites are typically located in 3' UTRs and that hundreds of genes are repressed by individual miRNAs but to a relatively mild level. The down-regulated genes with the highest translational repression also displayed increased destabilization. They suggest miRNAs confer mild adjustments and act as a tuning of protein synthesis (Baek *et al.* 2008). In a study considering expression activity including: transcription, translation, and turnover, the authors used mRNA and protein levels from mammalian cells to study the correlations within these. They found better mRNA and protein level correlation than expected but the half-lives of the molecules were not correlated. Their findings may suggest that protein abundance is controlled during translation (Schwanhäusser *et al.* 2011). In a follow up to the above study, of the correlation of mRNA and protein abundance, it was found that systematic errors may underlie a substantial underestimate in the abundance of protein present per cell which suggested that 10% to 40% of protein variation is from mRNA expression. A re-analysis of the study estimated 10 fold more protein molecules per cell. Based on the corrected protein abundance levels mRNA expression accounted for 56% of the protein levels and that translation is 12% lower than previously estimated. This new study suggested that mRNA expression explained ~84% of protein level variation. These results suggest that transcription is the main driver in protein level variation and that translation, RNA degradation, and protein degradation are smaller contributors of the variation (Li, Bickel and Biggin 2014). In another recent study of protein levels, the regulation of protein expression during cellular differentiation was considered. Protein levels depend on transcription, translation, and degradation to determine steady-state level but less is known about how system-level perturbations impact on these levels. This study found that during differentiation that synthesis rate was the main determinant and that degradation rates were constant. They also found that synthesis and degradation rates are the reason that transcript and protein expression levels typically have poor correlation (Kristensen, Gsponer and Foster 2013). It has previously been shown that the average rate of transcription in mammalian cells is between 1.3 Kb and 4.3 Kb per minute and that the largest gene in the human genome, dystrophin (DMD; Duchenne muscular dystrophy), takes 16 hours for transcription (Tennyson, Klamut and Worton 1995; Ben-Ari *et al.* 2010; Maiuri *et al.* 2011). The median estimated half-life of mRNA is between 7 and 9 hours while the median half-life of proteins is estimated to be between 22 and 46 hours but can be as short as 45 minutes (Sharova *et al.* 2009; Eden *et al.* 2011; Schwanhäusser *et al.* 2011). In a study of orthologous protein and mRNA expression correlations across seven different species, it was found that even across diverse taxa protein abundances show higher correlation than the corresponding mRNAs. This study included mRNA and protein measures from: two bacteria, yeast, nematode, fly, human, and rice.

The authors were interested in this because while nematode and fly orthologous proteins were well correlated the corresponding mRNA was not. So while it is thought that mRNA transcription primarily determines protein levels and that post-transcription, translation and degradation play a lesser role these data suggest, there is likely strong selective pressure to maintain protein abundances even when mRNA abundances diverge (Laurent *et al.* 2010).

Regulatory adaptations are a significant part of phenotypic evolution (Wray 2007). There are phenotypic consequences to changes in the regulation of transcription. Sequence variants that result in a protein coding change may have multiple phenotypic effects (pleiotropy). For example a protein coding change in a transcription factor may alter its function or expression affecting its interaction with other regulatory proteins or binding site specificity and may have a regulatory effect on many genes. The modular organization of enhancers and promoters in *cis*-regulatory regions allows for discrete effects on expression that restricts the pleiotropy and permits these discrete effects to be modified by selection. Selection has increased efficiency in *cis*-regulatory regions allowing beneficial alleles to be fixed and eliminating deleterious ones because alleles in the *cis*-regulatory regions are likely to be codominant and visible to selection immediately as these variants may have fitness consequences as heterozygotes (Arnone and Davidson 1997; Stern 2000; Wray 2007). Altering the gene expression of functionally conserved proteins may largely account for evolution of form and these alterations in gene expression occur through *cis*-regulatory variation (Carroll 2008). Selection is active in promoter sequences, like it is for coding sequences, through negative or purifying selection, positive selection, overdominant selection, balancing selection, stabilizing selection, and compensatory selection (Guardiola *et al.* 1996; Cowell *et al.* 1998; Crawford, Segal and Barnett 1999; Romey *et al.* 1999, 2000; Hamblin and Di Rienzo 2000; Trefilov *et al.* 2000; Bamshad *et al.* 2002; Wray 2007). An example of compensatory selection in a promoter with an effect on expression and relevant to disease is at the cystic fibrosis transmembrane conductance regulator (*CFTR*) locus. A hypomorphic allele in the *CFTR* coding sequence causes cystic fibrosis but there are haplotypes where a second *cis*-variant modulates disease prognosis. This second *cis*-variant adds an additional binding site for the Sp1 transcription factor resulting in increased transcription and improved disease prognosis. The non-disease causing haplotypes never carry this second variant and therefore do not have the additional Sp1 binding site, suggesting a compensatory effect under positive selection (Romey *et al.* 1999, 2000; Wray 2007).

1.2.2.2: Variation in Gene Expression

Many early studies using parallel gene expression profiling arrays focused on the task of identifying which genes were expressed in which tissue. In 2000, Warrington *et al.* published a study looking at gene expression in 11 different human tissues (Warrington *et al.* 2000). These different tissues included both adult and fetal tissues assayed to capture expression levels for ~7,000 mRNA transcripts. In this study, they identified which transcripts are expressed in which tissue. To minimize individual variation, and presumably reduce costs, they used a pooling strategy. One of their research aims was to identify genes that are expressed in all tissues, as these would be likely candidates as 'housekeeping' genes. Their basis for this being that since these genes are expressed in all tested tissues from early fetal development through to adulthood these are likely required for cellular maintenance or 'housekeeping'. The tissues included in this study were: adult and fetal brain, adult and fetal kidney, adult and fetal lung, fetal liver, adult heart, adult pancreas, adult uterus, and adult testis. They identified 535 genes that are expressed in all 11 tissues, which they suggest are candidate housekeeping genes. Additionally, they found 400 genes that were expressed in fetal tissues but were absent from any adult tissue; 767 genes that were expressed in all four fetal tissues; and 695 that were expressed in all 7 adult tissues (Warrington *et al.* 2000). In a similar study of gene expression in multiple human tissues, Hsiao *et al.* surveyed 19 distinct tissue types from 59 samples. In addition to detecting which transcripts were expressed in which tissues they also considered the variation in expression levels between the tissues. They found 451 genes that were detected in all tissues. However, the variation in the expression levels of these genes between the tissues was such that they were able to detect tissue-specific signatures. Of the 451 ubiquitously expressed genes, which they also labelled 'housekeeping', 358 overlapped with the previous study. Here the tissue-specific signature of expression was based on statistical tests to identify highly expressed genes within a specific tissue. They identified genes that showed a tissue-specific signature: 618 in brain, 91 in kidney, 277 in liver, 75 in lung, 317 in muscle, 46 in prostate, and 101 in vulva. They also attempted to identify which genes were most variable within a tissue between individual subjects. There were some limitations based on their cohort and tissue sampling but the investigators were able to identify a small set of genes that appear variable within specific tissues. Their results suggest that kidney contained the highest variation in expression, but that brain, liver, lung, muscle, and, vulva also showed considerable gene expression variation between individuals (Hsiao *et al.* 2001). The recent functional annotation of the mammalian genome 5 (FANTOM5) study which mapped transcription start sites in 975 human and 399 mouse tissue and cell lines found that ‘housekeeping’ genes may be fewer than previously believed. Their findings suggest that many mammalian promoters include multiple TSSs with cell-type specific expression patterns. These cell-specific TSSs appear to have evolved at different rates, where as the promoters of broadly expressed genes show the most conservation (FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* 2014).

Additional studies continued surveying expression in human tissues. In 2002, Saito-Hisaminato *et al.* analysed gene expression in 25 adult and four fetal tissues. They found that many genes were highly expressed in only one or a few tissues and very few are exclusively expressed in a single tissue. They also found, based on gene profiles, that the tissues not only cluster well by distinct tissue but also by general tissue category such as: nerve, lymphoid, muscle, and adipose (Saito-Hisaminato *et al.* 2002). In 2003, Evans *et al.* investigated differences in distinct regions of the brain and similar brain tissues based on 13 normal human subjects in three regions. The cerebellar cortex and the cerebral cortex were highly divergent, and the anterior cingulate cortex and the dorsolateral prefrontal cortex were highly similar. This group found ~1,000 genes differentially expressed between cerebellum and cerebral cortex but very few between neocortical regions. Clustering revealed a 1st order branch between cerebellum and the neocortical regions, a 2nd order branch was between the replicating labs (3 total labs), and the 3rd order branch between individual subjects. The authors were unable to separate neocortical tissues based on their clustering approach. These data also revealed an important confound to consider when analysing microarray expression data, namely experimental batch and site effects. Of the reliably detected probes, 30% were detected in the cerebral cortices and 22% in cerebellum. They were unable to identify any differences in genes that were expressed in the cerebral cortices but 74 transcripts were specific to the cerebral cortical regions and 15 transcripts were specific to cerebellum (Evans *et al.* 2003).

As technology improved and costs began to decline, the feasibility of running more samples per tissue was reflected in a couple of the published studies. In 2005, Son *et al.* published a survey of human expression based on 30 subjects in 19 organs (158 total tissue samples). One particularly interesting aspect of this study from a methodological point of view was the inclusion within the analysis of their known covariates. They found that the covariate effects for age, sex, ethnicity, and post-mortem interval were smaller than that of tissue type. They also found that randomly sub-setting with as few as 100 genes can reproduce distinct tissue clustering: suggesting that differential expression of more than 90% of the genes is of biological origin. This study observed that tissues of a similar cellular composition and function cluster more closely together, but still clearly separate. They also observed a great deal of variation in transcript abundance levels in different tissues. Based on this heterogeneity of transcript expression they were able to identify a subset of tissues that were distinctive based on a set of characteristics. These characteristics included: higher expression variation, cluster categorization when considering the Gene Ontology term for high level energy consumption, and the number of tissue-specific genes expressed. The two brain tissues included in the study, cerebellum and cerebrum, were present in this subset of distinctive tissues based on these categories (Son *et al.* 2005). In addition, in 2005, Shyamsundar *et al.* published what at the time was a very large expression series surveying normal human tissue from 115 subjects in 35 different tissue types. This study again reaffirmed previous results showing tissue-specific gene expression and transcript abundance. This work reaffirmed that tissues cluster separately but also cluster in large part based on anatomical location, cellular composition and physiological function (Shyamsundar *et al.* 2005).

There were also some critical early studies that made use of highly parallel gene expression assays to survey gene expression variation using model organisms. In 2001, Jin *et al.* published a study in *Drosophila melanogaster*. Within this fly model, they found that gene expression was strongly affected by gender and very little by age. They also found that interactions between gender and genotype were present and affected up to 10% of the fly's expressed genome. This work concluded that there are genotypic contributions to transcriptional variation (Jin *et al.* 2001). In 2002, Oleksiak and colleagues surveyed population differences in teleost fish. In this study they found expression variation between populations of teleost fish but also variation in 18% of expressed genes between fish from the same population (Oleksiak, Churchill and Crawford 2002). A study of four natural population isolates in wine Yeast also identified that the genetic variation in these natural populations affected variation in gene expression on a genomic scale (Townsend, Cavalieri and Hartl 2003).

Early surveys of gene expression were also performed in mammalian model organisms and compared to gene expression in humans. In 2002, Su *et al.* published a study comparing gene expression in human and mouse tissues. This study used 46 human and 45 mouse tissues. Similar to the Warrington study of 2000 and the Hsiao work of 2001 they found that 6% of genes were expressed across all tissues again re-enforcing the possibility that these genes are housekeeping genes (Warrington *et al.* 2000; Hsiao *et al.* 2001; Su *et al.* 2002). Each individual tissue expressed 30-40% of the assayed genes, and 90% were expressed in at least one tissue. Based on an analysis of variance (ANOVA) they found that 78% of genes are differentially expressed in mice and 82% are differentially expressed in humans (Su *et al.* 2002). A study, in 2002, by Enard *et al.* compared the differences between human and non-human primates for gene and protein expression, as well as three mouse species (Enard et al. 2002). In primates they looked at both gene and protein expression in blood, liver, and brain from humans, chimpanzees, orangutans, and macaques. They found substantial variation between subjects of the same species and showed that for blood and liver that human and chimpanzee were more similar than chimpanzee and macaque. However, the chimpanzee brain cortex was more similar to that of the macaque than the human brain samples. Their data suggest an accelerated rate of change in gene expression in brain presumably associated with human evolution. This result was also supported in their protein work where they identified many species-specific expression patterns in both gene and protein expression, with this pattern being particularly pronounced in human brain (Enard et al. 2002). Additional studies have also compared gene expression of cell populations for human and other primates and found greater divergence in brain than in liver suggesting that approximately 10% of genes differed in expression. These studies also suggested the differential expression they detected is likely an underestimate based on the regions studied, assays used, and analysis methodologies that were based on large differences in expression (Enard *et al.* 2002; Cáceres *et al.* 2003; Khaitovich *et al.* 2004, 2005). In 2011, Cain *et al.* published a study of histone modifications, specifically H3K4me3, and gene expression between human and non-human primates in LCLs. They found that many H3K4me3 localizations are conserved in primates and enriched near transcript start sites. H3K4me3 is an epigenetic mark thought to promote expression. As expected, highly expressed genes are more likely to have histone modifications near TSSs than genes with lower expression levels. They also found that genes that are differentially expressed between primates also had differences in H3K4me3 marks near the TSS. Their estimates suggest that up to 7% of expression differences between primates are in some part related to H3K4me3 histone modifications (Cain *et al.* 2011). In a study of DNA methylation and gene expression between humans and chimpanzees in liver, heart, and kidney tissues it was found that methylation patterns between tissues are often conserved between these species. However, the authors also found a large number of gene expression differences between humans and chimpanzees that had corresponding differences in DNA methylation at the promoters of these genes. Based on their findings they estimate that 12-18% of gene expression differences between human and chimpanzees may be because of differences in DNA methylations in gene promoters (Pai *et al.* 2011).

In a study published, in 2007, by Storey *et al.* it was shown that variation in gene expression, like genetic variation, is also primarily between individuals within a population, and to a smaller degree between populations. Their study included individuals of European and African ancestry and analysed 5,194 genes expressed in lymphoblastoid cell lines (LCL). This study's results suggest that 83% of genes are differentially expressed between individuals and 17% between human populations (Storey *et al.* 2007). A recent study was undertaken to consider if local adaptations between human populations are driven by gene expression changes and not protein-coding changes. This study found that these local adaptations are 10 fold more likely to affect gene expression than amino acid changes. Additionally, they found that polygenic local adaptations show recent positive selection for (ultra-violet) UV radiation response, immune cell proliferation, and diabetes pathways. These results support the idea that gene expression changes have driven human adaptation (Fraser 2013).

Beyond which genes are expressed in which tissues and at what abundance, there are characteristics of these tissue-specific patterns that have also been identified. In 2004, Plotkin *et al.* reported a study considering synonymous codon usage in genes that are selectively expressed by tissue. For genes that have tissue-specific expression in humans, the authors determined how much these genes differ in their encoding of amino acids. The authors describe this difference as the distance between synonymous codon usage. They calculated this distance by computing the codon frequency per gene and using the frequencies for synonymous codons to compute how the genes differ in the encoding of each amino acid. They hypothesize that tissue-specific codon usage may be a mechanism of protein regulation and tissue differentiation through translation rate, modulation based on relative tRNA abundance, mRNA folding, and RNA transport. The investigators found that the codon usage distinguishes genes expressed in a tissue from those in another tissue, and that this codon usage in brains is selectively preserved throughout mammalian evolution based on human and mouse data (Plotkin, Robins and Levine 2004). In 2004, Yeo et al. published a study that examined splicing events in different human tissues. They found that splicing events were more prevalent in certain tissues than in others. These alternative splicing events were identified in human tissues based on the mining of cDNA sequences from Genbank and expressed sequence tags from dbEST from 16 human tissues. The largest splicing differences found were in brain, testis, and liver (Figure 1.4). The brain and testis had the highest level of exon skipping, while liver had the highest alternate 3' and 5' splice site usage (Yeo *et al.* 2004). In 2012, Barbosa-Morais *et al.* published a study comparing the transcriptomes of vertebrates, spanning 350 million years, considering splicing between vertebrate lineages. They found significant differences in splicing complexity with the highest complexity found within primates. The splicing profiles were more strongly related to species than to organ type, and that separation in the profiles occurred within the last six million years. They state that species-specific splicing in vertebrates is *cis* directed; although, there is a subset of pronounced splicing predicted to occur in *trans* (Barbosa-Morais *et al.* 2012).

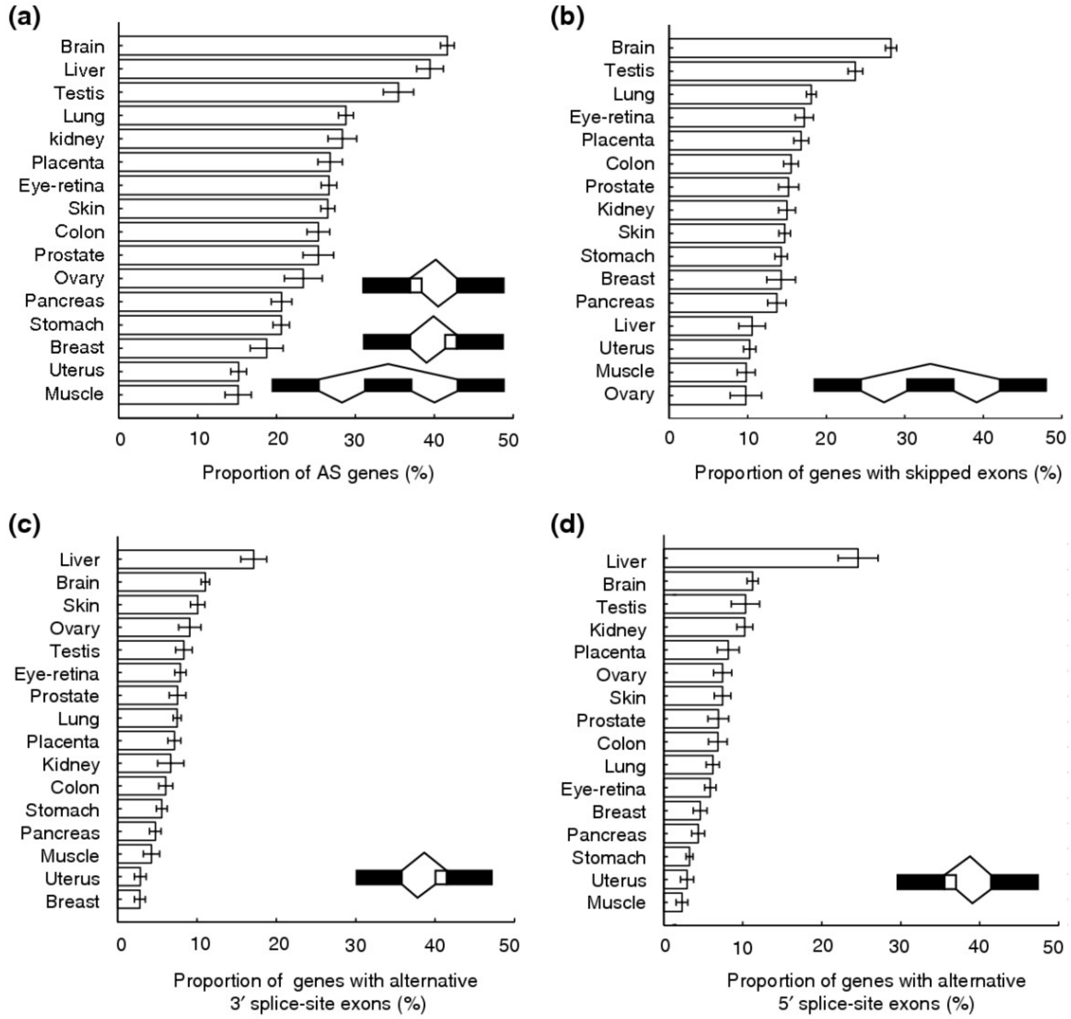


Figure 1.4: These plots show the proportion of alternative splicing (AS) present in 16 human tissues. The horizontal bars show the proportion of alternatively spliced genes (with estimated standard deviation) based on a random sampling of 20 expressed sequence tags (EST) from each gene, derived from a human tissue. The four plots represent different alternative splice type categories, with the splice types schematically represented in the plots. (a) Proportion of AS genes with skipped exons, alternative 3' exon splice sites, or alternative 5' exon splice sites. (b) Proportion of AS genes with skipped exons. (c) Proportion of AS genes with alternative 3' exon splice sites. (d) Proportion of AS genes with alternative 3' exon splice sites. This figure is reproduced from (Yeo *et al.* 2004).

In 2008, Wang *et al.* published a RNA-sequencing (RNAseq) based study on 15 human tissues examining mRNA isoform expression. The authors found that 92-94% of human genes undergo alternative splicing. Their analysis found that isoform variation of alternative splicing, cleavage, and polyadenylation occurred more often between tissues rather than between individuals. They also found that the more tissue-specific isoforms are correlated with conserved regulatory regions (Wang *et al.* 2008). In 2008, Pan et al. also published an RNAseq based study examining alternative splicing in human tissues. They identified novel splice junctions for 20% of genes, that have multiple exons, and that 95% of multi-exon genes have alternative splicing events in human tissues. Based on their results and previous studies results they estimate that between 6,000 and 10,000 genes are expressed in most tissues. Their results also found, as previously reported by Yeo *et al.*, that both brain and liver show the highest levels of alternative splicing (Pan *et al.* 2008). Based on the previous findings that splicing appears to be more prevalent in brain and testis, de la Grange *et al.* performed a study of splicing in 11 human tissues. They found that cerebellum, testis, and spleen showed the largest amount of differentially expressed alternative exons among the tissues tested. They also found that this variation was correlated with a larger number of splicing factors expressed at higher levels in these three tissues. For these three tissues they also found that a larger number of genes had higher expression as well (de la Grange *et al.* 2010). In 2014, Braunschwig *et al.* published a study exploring intron retention as a specific form of alternative splicing in mammals. While intron retention is thought to be less prevalent in mammals, it is the most common form of alternative splicing in plants and unicellular eukaryotes. This study found that intron retention is more frequent than previously thought, and serves as a mechanism for the suppression of inappropriately expressed transcripts as a form of transcriptome tuning. They formed this conclusion based on finding that as many as 75% of multi-exon genes show evidence of intron retention. This retention is correlated to *cis* features, but the retained intron leads to a reduction in expression of the transcript through nonsense-mediated decay, nuclear sequestration, turnover, local stalling of RNA Pol II, and reduced splicesomal components (Braunschweig *et al.* 2014).

These studies highlight that mRNA expression is often tissue-specific and variable across species, populations, and individuals. This variability includes not only where genes are expressed but at what abundance and in what form. In many of these studies, which considered the variability of expression within the same tissue and species, variation between individuals was consistently well supported. It is also important to note that even with very similar species and tissues of a similar cellular composition there is substantial variation. The expressed splice forms, mRNA transcripts, of genes also vary substantially between tissues in the same species. Several of the studies that included human brain tissues typically found that these tissues often stood out or are in the subset of outlier tissues when examining profiling metrics and characteristics. The divergence of expression in human brain tissues from that of other tissues may be a result of the complexity of the organ and the heterogeneity of its cellular composition, but this also likely reflects the biology underlying the evolution of the human brain and that gene expression may have driven many of the adaptations in this organ.

1.3: Foundations of Expression Quantitative Loci (eQTL)

Similar to the study of clinical traits and disease phenotypes prior to the HapMap Project (International HapMap Consortium 2005), and the arrival of high-density SNP chips, the study of eQTL was typically executed by linkage based studies in families and then afterwards using association studies in populations. Large-scale assessments of the role of genetic variability in the control of gene expression have also undergone rapid growth during the last decade. The bulk of this earlier work included: linkage based analysis of gene expression in CEPH (Centre d'Etude du Polymorphisme Humain; Utah residents with ancestry from northern and western Europe) lymphoblastoid cell lines and multiple tissues from rat and mouse crosses, and association based expression analyses in human lymphoblastoid cell lines (LCLs).

An initial proposal for combining genetic markers and gene expression data was from the field of plant biology and termed ‘genetical genomics’ (Jansen and Nap 2001). However, much earlier work already existed in understanding gene expression difference linked to genetic variants in complex organisms and resulting in a morphological phenotype. In fruit flies, it has been show that *cis*-regulatory variation can result in morphological phenotypes such as those related to abdominal pigmentation (Simpson, Woehl and Usui 1999; Wittkopp, True and Carroll 2002), wing pigmentation (Gompel *et al.* 2005; Prud’homme *et al.* 2006), the distribution of bristles (Simpson, Woehl and Usui 1999; Skaer and Simpson 2000; Sucena and Stern 2000), and larval denticle bands (Sucena and Stern 2000; Sucena *et al.* 2003). While specific phenotypes have shown strong association to genetic variation in gene expression, LD often confounds the ability to determine the precise region of sequence variation for these specific phenotypes (Karp *et al.* 2000; Beldade, Brakefield and Long 2002). More recently, Bickel *et al.* undertook a study in fly to identify causative genetic variants for cuticular pigmentation in flies. The fly genome contains high sequence variation and low LD, which lends itself to finding causative variants within a QTL. The authors focused on alleles segregating within the bric-a-brac locus, which has a large effect on cuticular pigmentation. They found that *cis*-regulation modulates transcription at the locus and that the variation has a cumulative effect through three functional regions: promoter, tissue-specific enhancer, and polycomb response element (Bickel, Kopp and Nuzhdin 2011). In stickleback fish, changes in skeletal morphology result from *cis*-regulatory variation. These genetic variations result in changes to the dorsal spines and pelvic girdle resulting in a loss or reduction in skeletal armour between different species of sticklebacks (Shang, Luo and Clayton 1997; Marcil *et al.* 2003; Cresko *et al.* 2004; Shapiro *et al.* 2004; Shapiro, Bell and Kingsley 2006).

Human phenotypic traits had also previously been linked to genetic variation in *cis*-regulation as well. Resistance to malaria results from cell-specific expression linked to *cis*-regulatory variants near Duffy blood group chemokine receptor (*DARC*), which is a receptor that binds interleukin 8 (*IL8*) (Horuk *et al.* 1993; Chaudhuri *et al.* 1994; Tournamille *et al.* 2004). *DARC* is expressed in multiple cell and tissue types. The expression of *DARC* in erythrocytes is the entry point of the malarial parasite Plasmodium vivax (Pogo and Chaudhuri 2000). However, there are *DARC* haplotypes that segregate between human populations that are resistant to infection from this parasite. This resistance results from the Duffy protein not being expressed in erythrocytes while still being expressed in other tissues. The cell-specific lack of expression in erythrocytes, for the infection-resistant haplotype, is the result of disrupting of a binding site for the transcription factor GATA binding protein 1 (*GATA1*) (Miller *et al.* 1976; Chaudhuri *et al.* 1995; Peiper *et al.* 1995; Tournamille *et al.* 1995; Iwamoto *et al.* 1996; Hadley and Peiper 1997). This example shows how a single *cis*-regulatory variant results in a phenotype with restricted pleiotropy and a significant fitness gain. It has also been shown in a studies of variation near *DARC* that this locus shows strong positive selection in geographic populations where malaria is prevalent (Hamblin and Di Rienzo 2000; Hamblin, Thompson and Di Rienzo 2002).

An early application of integrating genetics and gene expression in humans was published by Rockman and Wray in 2002. In this study, they published a survey of experimentally validated functional *cis*-regulatory variants. They carried out the study to understand if *cis*-regulatory variants may represent evolutionary changes in phenotypes. They authors looked at 140 polymorphisms possibly involved in the regulation of 107 genes in humans. They suggested that genetic variation contributes to variation in transcription rates and therefore to phenotypic variation. This conclusion was based on the observation that variation in functional *cis*-regulatory regions is widespread, can lead to large gene expression difference for 63% of the genes they assayed, and that on average humans are more heterozygous at *cis*-regulatory sites than at protein coding bases. This difference in heterozygosity suggests that *cis*-regulatory regions may store more heritable phenotypic variation and have higher substitution rates (Rockman and Wray 2002). A similar study, by Yan *et al.*, also found evidence of variation in gene expression in humans. In this study, it was observed that the difference in the expression of alleles from heterozygous individuals was more than 20%. This study examined allelic expression in 13 genes using 96 individuals from CEPH families. Allelic variation in mRNA expression was observed in 6 of the 13 genes. Three of the families, which were informative in displaying allelic variation, were fully consistent with Mendelian inheritance. Thus they concluded that *cis*-acting variation in gene expression is relatively common among normal individuals (Yan *et al.* 2002). In 2003, Bray *et al.* published a similar study of allele-specific expression. This study was based on expression in brain tissue from 60 subjects. The alleles selected for study were from common heterozygotes in 15 genes expressed in human brain. The brain tissues were frontal, parietal, and temporal cortex. Allele-specific expression differences were detected in 7 of the 15 genes examined. The detection of allele-specific expression was based on a difference of at least 20% in allele representation in at least one individual. The gene *DTNBP1*, encoding dystrobrevin binding protein 1, showed allele expression differences in multiple individuals, which exceeded 50% on average (Bray *et al.* 2003).

An early implementation of an eQTL study, on a genome-wide scale, was published by Schadt *et al.* in 2003. This study was performed with small sample sizes but for multiple species using microsatellite markers and expression data to perform linkage based analysis to identify eQTL in maize, murine, and human samples. Additionally, the authors were able to show, in the murine samples, an example of a clinical quantitative trait locus (cQTL) for obesity localizing together with an eQTL (Schadt *et al.* 2003). A 2003 study, from Cheung *et al.*, using human LCLs, was a key step in determining that genetics contributes to variation in gene expression in humans. This study used early arrays with pooled samples from 35 CEPH subjects to identify genes expressed with high variation. To validate their results they ran RT-PCR on five of the genes, with the highest variation, in 49 unrelated CEPH subjects; children from 5 CEPH families, and 10 pairs of monozygotic twins. These authors found that expression variation was higher in unrelated individuals than in siblings and that the expression variation was lowest in the monozygotic twins. Expression variation in the unrelated subjects was 3 to 11 times higher than in twins, while in siblings this variation was 2 to 5 times higher than in twins. This suggested that a component of expression variation is genetic and provided critical evidence for heritability in gene expression (Cheung *et al.* 2003). Linkage-based studies in additional CEPH families continued, making use of SNPs instead of microsatellite marker panels to capture genetic variation, and using regression methods to associate genetic variation with gene expression in addition to linkage analysis methods (Monks *et al.* 2004; Morley *et al.* 2004). The Monks et al. study, from 2004, showed expression heritability results in a cohort of 167 CEPH subjects. They found that 31% of genes were heritable (Figure 1.5). They also found that for genes with a linkage-based eQTL, that 75% of these had high heritability. Additionally, later studies suggest that the heritability of gene expression appears to affect 40-90% of genes, with median estimates of 15-35% (Monks *et al.* 2004; Dixon *et al.* 2007; Göring *et al.* 2007; Emilsson *et al.* 2008; Price *et al.* 2011; Grundberg *et al.* 2012).

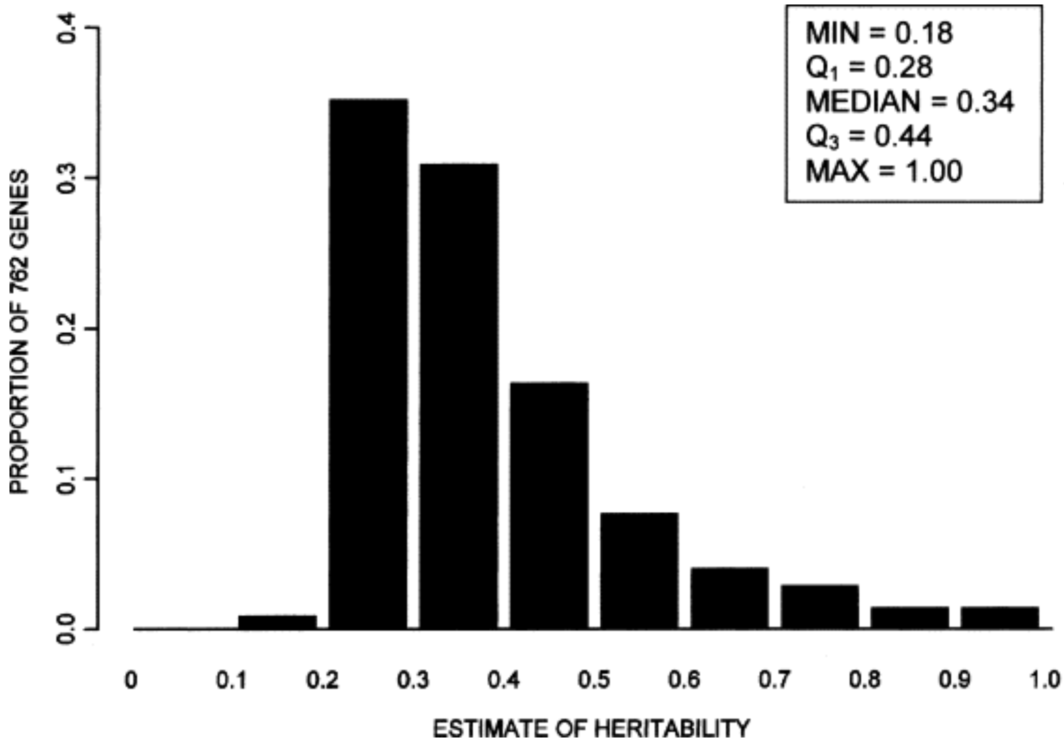


Figure 1.5: Histogram of heritability estimates for genes that are differentially expression and significantly heritable, based on a false-discovery rate of 5%. This figure is reproduced from (Monks *et al.* 2004).

After the early eQTL studies, which were typically family based, more population-oriented studies of eQTL began to be performed. In 2005, Stranger *et al.* published a seminal eQTL study based on expression in LCLs from 60 unrelated CEPH individuals from HapMap. Three hundred and seventy four of the 630 genes assayed, from ENCODE regions on chromosomes 20 and 21, were reliably detected and used for eQTL analysis. They ran *cis*- and *trans*-eQTL analysis making use of multiple association methods for comparison purposes. For this study, the authors considered a *cis* region as all variants that are 1 Mb proximal to the gene being tested, and *trans* as all variants outside (distal) of the *cis* region. This study also performed comparisons of different methods of multiple test correction techniques for their *cis* and *trans* results. In *cis,* they found between 10 and 40 significant eQTL depending on the test and correction method used, and very little evidence for reliable *trans*-

eQTL. This study also showed early methods for correcting for polymorphisms within the expression probe (Stranger *et al.* 2005). The polymorphism within expression probe is a critical assay artefact to correct for when performing eQTL analysis, when the expression levels are measured using a microarray. The artefact arises based on how the typical microarray chip works, and how the probes on the chips are designed. Typically, the probes are designed based on an N-mer sequence such that the unique transcript to be measured by the probe will hybridize to the matching unique transcript's sequence, when the transcript is present in the sample. The expression abundance, for the transcripts hybridized to the chip, can then be quantitatively measured based on the probe's relative intensity. However, the N-mer probe is typically designed from a transcript's sequence based on a reference genome and therefore variation from the reference sequence is not accounted for. This polymorphism within the probe can affect the hybridization of the transcript fragment to the chip resulting in a biased measurement of the transcript's abundance, usually a decrease in expression. This artefact can bias *cis*-eQTL analysis, as the variant within the assay probe may be in LD with the variant being tested against the transcript's expression level resulting in a false positive result. Cheung *et al.* also published a regional and genome-wide population based eQTL study in 2005. This study was performed using 57 unrelated CEU (CEPH; Utah residents with ancestry from northern and western Europe) individuals from the HapMap Project. The study used a regression based method and a dense SNP marker set, composed of 770,000 SNPs, to associate genetic variation with variation in gene expression for genes that had previously been found to have an eQTL in their earlier study using a linkage based analysis method (Cheung *et al.* 2005).

These early studies leveraged new technologies, and a growing base of reference genetic information to build the foundation for eQTL work. In combination, they showed the feasibility of this approach, and revealed critical insights into the genetic regulation of gene expression.

1.4: Rapid Expansion of the eQTL field of study

Continuing with a population-based association approach to eQTL, a study was conducted in all unrelated HapMap individuals, totalling 210 individuals from the four Phase II HapMap populations: CEU, YRI (Yoruba in Ibadan, Nigeria), JPT (Japanese in Tokyo, Japan) and CHB (Han Chinese in Beijing, China). Within this study, the genetic variation included both SNPs as well as CNVs (copy number variants) and the expression phenotypes of approximately 14,000 mRNA transcripts, from LCLs, were analysed. Based on an analysis of *cis*-eQTL, the authors identified 1061 genes with an eQTL where 86.3% were correlated with SNPs, 17.7% with CNVs, and 1.3% with both variant types (Stranger *et al.* 2007). Moffet *et al.* published a disease relevant study of eQTL in a genomic region associated with childhood asthma. A genome-wide association study (GWAS) was performed using the SNPs and transcripts within this region resulting in the finding that an eQTL for *ORMDL3* may contribute to the risk of childhood asthma (Moffatt *et al.* 2007). In a large linkage-based eQTL study, which included lymphocyte samples from 1240 subjects, the authors found that 85% of detected autosomal transcripts were heritable. Heritability is an estimate how much variation in a phenotype or trait is due to genetic variation between individuals in a population. The subjects included in this study were recruited as part of a heart study, which allowed the authors to integrate their eQTL results with linkage based QTL analysis of high-density lipoprotein cholesterol levels, in these subjects. Based on the integrated results they found that *cis*-regulatory variants for vanin 1 (*VNN1*) affect high-density lipoprotein cholesterol concentrations (Göring *et al.* 2007). In 2007, Libioulle *et al.* published a GWAS study for Crohn's disease, with replication, and identified two previously known loci and a novel locus associated with the disease. The novel locus was located within a 1.25 Mb gene desert on chromosome 5. They found that the Crohn's disease associated variants within this locus were also part of an eQTL for prostaglandin E receptor 4 (*PTGER4*), the gene most proximal to the GWAS locus (Libioulle *et al.* 2007). These studies and approaches have all demonstrated that genetic variability can be correlated with changes in gene expression (Gilad, Rifkin and Pritchard 2008; Cookson *et al.* 2009). However, most of these studies thus far have primarily used human LCL (lymphoblastoid cell lines) as the tissue to assay for gene expression phenotypes or have used non-human mammalian tissue (Hovatta *et al.* 2005, 2007; McClurg *et al.* 2007).

It is in the context of this described work that I began my doctoral research. During the period of time I have been working on my thesis, investigating eQTL in human brain tissues, the eQTL field has progressed a great deal, moving from an area of relatively sparse activity, to one that is central to our understanding of the biologic consequences of genetic variability and the interpretation of non-coding genetic variability associated with disease. This thesis describes the contributions I have made to the study of eQTL during this time and reflects the progress of this growing field, describing maturing methodological and analytical approaches. These contributions focus on the study of eQTL in human brain tissues following the progression from using a mix of brain tissues, four distinct brain tissue regions, and finally to a specific neuronal cell type in the brain, as well as a progression in the methods used to identify eQTL.

2: Identifying eQTL in Human Cortical Tissue

(Myers, Gibbs *et al.* 2007a)

Statement of Contribution to this Research:

I was involved in the conception and design of this study, including choice of expression platform, and selection of tissue. I performed data quality control, data analysis, and data interpretation. I co-drafted and edited the manuscript. Myers AJ and Hardy J were also involved in the conception, design, choice of platform, and tissue selection. Myers AJ, Hardy J, Webster JA and Holmans P and I drafted and edited the manuscript. Webster JA and Holmans P were also involved in the multiple test correction portion of the data analysis. I was not involved in the collection of the tissue samples or the generation of the genotype and mRNA expression data. Coordination of tissue collection was performed by Myers AJ and Hardy J. Genotyping of the samples was coordinated or performed by; Webster JA, Craig DW, Pearson JV, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, and Coon KD. Generation of mRNA expression data was coordinated or performed by; Myers AJ, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, and Nath P.

2.1: Introduction

We initiated an eQTL analysis in human cortical tissue using whole-genome genotyping and gene expression data. This study was one of the first whole-genome eQTL studies performed in human neurological tissues and one of the few early eQTL studies published based on a human tissue that was not LCL-based.

Our selection of brain tissue for this work was based on several factors. First, and primary of them was that our laboratory studies the genetic and etiologic basis of neurological disease; thus, this is our tissue of interest. Second, it has previously been shown that mRNA from post-mortem human brain can be utilized for the study of gene expression (Gilbert *et al.* 1981). Third, it has previously been shown that disease associated loci, including those related to neurological disease such as *APOE* and *MAPT*, are subject to distortions in allelic expression (Lambert *et al.* 1997; Bray *et al.* 2004; Myers *et al.* 2007b). At the time of inception of this work, we had begun investing considerable resources in identifying the genetic basis of complex disease using genome-wide genotyping. It was clear from early work that the disease-linked variants identified would not be amenable to traditional cell biology and modelling approaches, and likely that much of the immediate biologic effect of these alleles would be mediated through changes in expression. It has previously been shown that heritability of gene expression appears to affect between 40 and 90% of genes, with median estimates of between 15 and 35% (Monks *et al.* 2004; Dixon *et al.* 2007; Göring *et al.* 2007). Thus, we thought it was critical to produce a dataset that would allow us to mine the effects of genetic risk variants in a disease-relevant tissue.

Based on these priorities and rationale we embarked on the initial work described in this chapter. Using human brain samples from approximately 200 neuropathologically normal individuals, we assayed each sample at approximately 500,000 SNPs and 24,000 mRNA transcripts and carried out an association-based analysis between genotype and gene expression to determine if eQTL could be detected within human cortical tissue.

2.2: Materials and Methods

2.2.1: Subjects, National Cell Repository for Alzheimer's Disease (NCRAD) cohort

(Coordination of tissue collection was performed by Myers AJ and Hardy J.)

We wrote to all the National Institute on Aging Alzheimer Centres and the Miami Brain Bank requesting samples. We requested 1 gram of frozen human cortex, from neurologically-normal brain, and the following sample information: gender, race, age at death, consensus diagnosis, neuropathological diagnosis, Consortium to Establish a Registry for Alzheimer's Disease (CERAD) scores (Mack *et al.* 1992; Galasko *et al.* 1995: 199), Braak and Braak staging (Braak and Braak 1991), and cortical tissue region. We received 279 samples, which met the following criteria: first, they were self-defined as ethnically of European descent; second, they had no clinical history of stroke, cerebrovascular disease, Lewy bodies, or co-morbidity with any other known neurological disease; third, the donor had been assessed by a board certified neurologist and where available had a Braak and Braak score < 3 (43% of controls used for this study assessed) or a CERAD score indicating either sparse or no neuritic plaques (34% of the controls used for this study assessed); and fourth, they had an age at death greater than 65 years. Of the received samples, 201 were successfully assayed for genotype and expression data. After excluding samples that were ethnic outliers and samples that were possibly related 193, samples were used for analysis.

2.2.2: Assays

2.2.2.1: Genotyping

(Genotyping of the samples was coordinated or performed by: Webster JA, Craig DW, Pearson JV, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, and Coon KD. As this cohort is made up of controls used as part of a genome-wide association study of Alzheimer's Disease, the genotypes were assayed as described in that GWAS study (Coon *et al.* 2007).)

Sample DNA isolated from brain tissue was hybridized to the Affymetrix GeneChip Human Mapping 500K Array Set according to the manufacturer's protocols. Allele calls were determined using Affymetrix BRLMM Analysis Tool. The BRLMM algorithm is a modification, from Affymetrix, of the robust linear model of Mahalanobis distance (RLMM) algorithm developed for calling genotypes assayed on the Affymetrix array set. The BRLMM method includes a Bayesian step, which improves the estimates of clusters and variances for calling genotypes, and was developed for use with the Affymetrix 100K and 500K chips. The algorithm makes use of data from multiple chips and SNPs to train a classifier for calling genotypes. This RLMM algorithm was proposed to replace Affymetrix's initial Dynamic Model calling algorithm. For evaluation of the RLMM algorithm, the authors applied it to Affymetrix 100K SNP array data and compared the results to the existing Affymetrix Dynamic Model algorithm using genotypes from The HapMap Project for comparison (Di *et al.* 2005; Rabbee and Speed 2006).

The Affymetrix GeneChip 500K set is composed of two arrays each capturing approximately 250,000 SNPs. One array uses the *Nsp I* restriction enzyme and the other uses *Sty1* to capture 262,000 and 238,000 SNPs respectively. Each SNP is represented on the array by a set of 24 or 40 different 25-mer oligonucleotides. Each SNP is interrogated by a 6- or 10-probe quartet, and each probe quartet is made up of a perfect match and mismatch probe per allele. Based on the array's design and optimization criteria the arrays include a SNP every 5.8 kilobases, on average, providing ~65% coverage of genetic variation within the HapMap II CEU population, the SNP selection was not based on haplotype tagging variants (Barrett and Cardon 2006). The Affymetrix GeneChip genotyping arrays are a microarray-based platform for assaying genotypes based on hybridization (Figure 2.1). Matsuzaki *et al.*, described the chip platform in 2004, based on the 100K array set. The 100K array simultaneously assays approximately 116,000 SNP on an oligonucleotide array. They achieved call rates of 99% and reproducibility rates of 99.97%. Based on an analysis of trios, from the HapMap Project, the authors claim an accuracy of 99.7% in the resulting genotypes. The 100K array's design is based on including a marker approximately every 24 Kb in the genome, and almost 105,000 markers were common SNPs with a minor allele frequency greater than 5% (Matsuzaki *et al.* 2004).

Figure removed.

Third party copyright permission could not be obtained.

Figure 2.1: Cartoon of the Affymetrix GeneChip protocol. Total genomic DNA from a sample is digested with a restriction enzyme (Nsp I or Sty I; Nsp I is shown). The digested fragments are ligated to adaptors. The adaptor-ligated DNA fragments are then amplified, fragmented, labelled, and hybridized to the chip. This figure is reproduced from Affymetrix product literature.

2.2.2.2: mRNA Expression

(Generation of mRNA expression data was coordinated or performed by; Myers AJ, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, and Nath P.)

Sample RNA was reverse transcribed into complementary RNA (cRNA) and biotin-UTP labelled using the Illumina® TotalPrep™ RNA Amplification Kit from Ambion, Inc. (catalogue # L-1755), based on the Eberwine technique (Van Gelder *et al.* 1990). The cRNA was quantified by three replicate measurements using a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, Delaware, USA). The cRNA was then hybridized to Illumina HumanRef-8 version 1 Expression BeadChips using standard protocols. Six to eight chips (24-32 control samples) were run in parallel for each hybridization batch. Average detection scores across each expression chip were greater than 0.99.

The Illumina Sentrix HumanRef-8 v1.0 Expression BeadChip assays the expression levels of approximately 24,000 human Refseq transcripts. The Illumina Sentrix BeadChip arrays use 50-mer sequence probes designed to capture, through hybridization, specific transcripts based on the transcript's Refseq sequence. The Illumina expression arrays are a single colour system where hybridized transcripts are stained with streptavidin-Cy3 and quantitatively detected fluorescence emission. Each gene-specific probe contains an additional 29-mer address sequence for probe identification purposes and is then attached to a bead, this combination of bead and gene-specific oligonucleotide is refered to as a bead-type (Figure 2.2). Each bead on the array will have hundreds of thousands of these gene-specific probe and address oligomers attached and then the beads are assembled onto the array platform. The beads assemble spontaneously into more than 1.6 million etched microwell pits, allowing each bead-type to have more than 30-fold redundancy on the array on average. After the beads are assembled onto the array, a hybridization procedure is performed to map the array using the address portion of the bead-type, which also validates the hybridization performance of every bead on the array. Each Illumina BeadChip includes multiple separate arrays on the chip, where an individual sample is hybridized to an array and multiple samples are run per BeadChip, one per array. The Illumina HumanRef-8 BeadChip contains eight arrays allowing eight samples to be assayed per chip. Based on Illumina's technical documentation the arrays have a less than 0.017% false positive rate for differential expression between technical replicates. Illumina also suggests that the BeadChip expression measures correlate well with gene expression measured by quantitative real-time PCR. For comparison with quantitative real-time PCR measures, Illumina measured ratios of 20 genes from two human tissues and found a strong correlation ($R^2 = 0.9328$) between the measures from the Illumina BeadChip and quantitative PCR.

Figure removed.

Third party copyright permission could not be obtained.

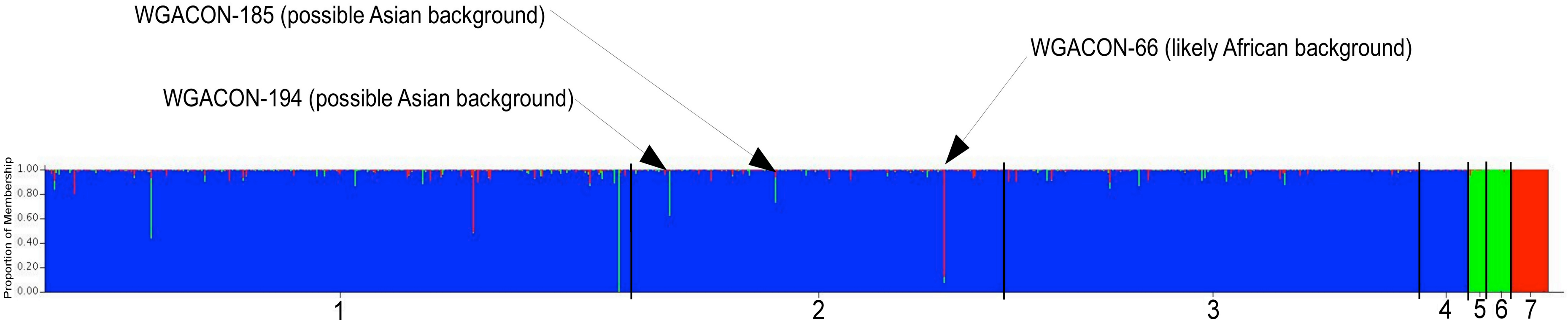
Figure 2.2: Cartoon of an Illumina expression bead-type. Each transcript is captured by a transcript-specific probe; this probe is a 50-mer sequence designed to match a unique portion of the transcript's reference sequence. The 50-mer probe is attached to a 29-mer address sequence so the probe can be identified. The probe and address oligonucleotide is then covalently attached to a bead. While this cartoon shows a single oligonucleotide attached to the bead, for simplicity, actually each bead has hundreds of thousands of these same gene-specific oligonucleotides attached. This figure is reproduced from Illumina product literature.

2.2.3: Data Analysis

2.2.3.1: Genotype Data

The following minimum genotype cut-off values were used during analysis: per sample call rate of at least 90%, per SNP call rate of at least 90%, per SNP minor allele frequency of at least 1% and non-significant (p-value > 0.05) for Hardy Weinberg Equilibrium test. The resulting sample genotyping call-rate had a mean of 97% and range 90%- 99%. Prior to analysis of the

366,140 SNPs the chromosome physical positions for each SNP were re-annotated from National Center for Biotechnology Information (NCBI) dbSNP based on Genome Build 36. Information about the ethnic structure of our cohort was obtained using the program *Structure* (Pritchard, Stephens and Donnelly 2000; Falush, Stephens and Pritchard 2003), and three ethnic outliers were removed (Figure 2.3). *Structure* was run using genotype data based on seven cohorts to examine ethnic bias within our series. The cohorts consisted of the control subjects used in this study, US Alzheimer's cases, US controls from the Coriell Cell Line Repository and samples from the four HapMap populations. All of these cohorts were run on either the Affymetrix 500K or the Illumina 550K genotyping platforms. Of the SNPs that overlapped between the Affymetrix and Illumina platforms, 2,035 were used in the analysis, where these variants were greater than 1 megabase apart, had a call rate greater than 98%, and a minor allele frequency greater than 10%. The SNPs were filtered on this basis to break up any LD that may exist between the remaining variants so the results would not be biased. The linkage model that *Structure* uses to detect population structure can be biased to overestimate divergence between ancestral populations and infer artificial admixture if SNPs in strong LD are included in the analysis (Falush, Stephens and Pritchard 2003). There were three ethnic outliers within the population used for this study. Two were of possible Asian descent (WGACON-185 and WGACON-194) and one of likely African descent (WGACON-66). We next examined the degree of relatedness among the samples within our cohort by using the pairwise identity-by-state (IBS) and identity-by-descent (IBD) analysis available in the PLINK analysis toolset (Purcell *et al.* 2007). The IBS/IBD analysis, using PLINK, estimates a genome-wide IBD measure between each pair of samples. This estimate, based on the sharing of alleles between each pair of samples, can identify individuals that appear more similar to each other than expected by chance (i.e. relatedness). Five samples were excluded based on having a high degree of relatedness to another subject in the cohort, likely subjects from the same family. Subjects removed based on being likely related to another subject with the cohort were: WGACON-2, WGACON-107, WGACON-149, WGACON-216 and WGACON-101.



1) 383 US Caucasian Non-Control Samples (Affy 500K)
2) 243 US Caucasian Controls Samples, our 201 is a subset in this group (Affy 500K)
3) 271 US Caucasian Controls from Coriell Cell Line Repositories (Illumina 550K)
4) 32 unrelated HapMap CEU samples (Illumina 550K)
5) 12 unrelated HapMap CHB samples (Illumina 550K)
6) 16 unrelated HapMap JPT samples (Illumina 550K)
7) 24 unrelated HapMap YRI samples (Illumina 550K)

Figure 2.3: Analysis of subject population membership based on genotype. Population membership was calculated using the program *Structure* (Pritchard, Stephens and Donnelly 2000; Falush, Stephens and Pritchard 2003) for 7 cohorts to examine ethnic bias within the sample series used in this study (cohort 2). In this instance, *Structure* was run assuming three populations (K = 3; African, Asian, and European). In the figure, each thin vertical line represents an individual subject where the proportion of an individual's membership in a population is colour coded. The colour codes are: blue for European, green for Asian, and red for African. The subjects used for eQTL analysis in this chapter are in the centre portion of the figure (cohort 2). As shown in the figure (for cohort 2) there were 3 population outliers detected, which were excluded from further analysis in the study. Two of the population outliers were of possible Asian extraction (WGACON-185 and WGACON-194) and one of likely African descent (WGACON-66). Figure reproduced from (Myers *et al.* 2007a).

2.2.3.2: mRNA expression data

All expression profiles were extracted and rank-invariant normalized (Schadt *et al.* 2001; Tseng *et al.* 2001; Workman *et al.* 2002) using the Illumina BeadStudio software. In rank invariant normalization, a subset of gene probes are identified and used to set the parameters for normalization. This subset of gene probes is selected based on whether or not the probe's rank changes across the samples, i.e. rank invariant. A gene's probe is rank invariant if the probe's expression intensity relative to other genes (rank) within a sample maintains the same relation across all samples. Prior to analysis of the mRNA transcripts, chromosome physical positions for each transcript's probes were re-annotated from NCBI's Entrez Gene based on Genome Build 36. Rank-invariant normalized expression data were log10 transformed and missing data was encoded as missing, not as a zero level of expression.

2.2.3.3: Selection of traits for analysis

Transcripts that were detected in less than 5% of the sample series were excluded from analysis, based on subjects that passed all genotype-based quality filters. Based on this detection criterion 14,078 mRNA transcripts were selected for expression quantitative trait loci analysis.

2.2.3.4: Expression Quantitative Trait Loci Analysis

For the eQTL analysis, the PLINK analysis toolset was used to perform a 1 degree of freedom allelic test of association. Briefly, the expression level of each transcript per sample was regressed on the number of minor alleles (0, 1 or 2) for the 366,140 SNPs that met the cut-off criteria to compute the effects of allele dosage on expression level. The analysis results were then separated into *cis* and *trans* associated SNP/transcript pair sets based on annotations. *Cis* SNPs were defined as those located within the genomic region that includes the gene encoding the transcript. The boundaries of the *cis*-genomic region were defined as 1 megabase (Mb) from the 5' end of the transcript and 1 Mb from the 3' end of the transcript. To account for missing expression data, results where there were not at least three expression data points for the minor homozygotes (BB genotype) were removed.

2.2.3.5: Corrections for multiple tests

(Webster JA and Holmans P were involved in the design and execution of the multiple test correction.)

The frequentist significance testing approach is one of the most common employed in genetics and most genetic researchers use p-values to show statistical significance within their results (Sham and Purcell 2014). Many variants are typically tested in genetic analyses leading to a higher test burden, especially under a genome-wide study. A significance threshold of a p-value less than 0.05 is typically used, resulting in a 5% chance of a false-positive result. However, with many tests being performed, assuming independence of tests, each test using this threshold still carries this 5% chance of a false positives so performing many tests increases the number of possible false positives. The traditional approach to correct for the multiple tests being performed is a Bonferroni correction. The Bonferroni correction adjusts the threshold of significance based on the number of independent tests being performed while maintaining a target family-wise error rate, typically 5%. However, for tests performed based on genetic variants many of these variants do not represent independent tests, as many of the variants are correlated through linkage disequilibrium. Another common approach for multiple test correction is to computationally generate an empirical distribution of smallest p-values based on repeated random swaps of the data. Random swapping breaks the trait and genotype relationship and then by comparing the p-value from the real data to the generated distribution an empirically adjusted p-value can be calculated (North, Curtis and Sham 2002).

To correct for the multiple tests performed in this study, an empirical distribution was created by randomly permuting the subject identifiers within the transcript expression data. One thousand of these permuted datasets were generated, the subject identifiers where swapped simultaneously for all transcripts within the dataset. These permuted dataset were then analysed in the same way as the actual data. For each dataset, the minimum p-value for each transcript/SNP association was recorded. A transcript-specific genome-wide empirical p-value was obtained for each transcript/SNP association by counting the number of simulated minimum p-values for that transcript, which were lower than the observed p-value and dividing by 1000. Transcript-specific genome-wide empirical p-values < 0.05 were considered significant for the transcript/SNP association. This provides an estimate of genome-wide significance that is robust to non-normality of transcript expression data and inter-SNP linkage disequilibrium because the repeated random swapping generates an empirical distribution as well as breaking the trait and genotype relationship.

Additionally, a study-wide correction for the number of transcripts tested in addition to the number of SNPs tested per transcript was performed. This additional correction was performed in two steps for those transcripts where an empirical transcript-specific genome-wide significant eQTL was identified. This correction was performed in two steps and only for a subset of the potentially significant eQTL because of the computational burden of running the increasing number of simulations. An additional round of the permutation procedure was repeated using 100,000 replicates, in order to obtain a more accurate estimate of the empirical p-value. Then one million permutations were performed to achieve a distribution sufficient to apply a Sidak correction to the transcript-specific genome-wide corrected empirical p-values. The Sidak single step p-value adjustment is given by the formula αcorr = $1 - (1 - \alpha)^C$ where α corr is the corrected p-value, α is the uncorrected p-value and C is the number of tests. This test is slightly less conservative than the commonly used Bonferroni correction. In our analysis the number of transcript tests was 14,078 such that a αcorr = 0.05 is approximately $\alpha = 3.64\text{x}10^{-6}$. Because of computational limitations, transcripts were selected for this test by the following criteria: significant transcript-specific empirical p-values less than 0.00001 and more than 15 associated SNPs. The four transcripts that met these criteria were: *KIF1B* (19 SNPs with transcript specific empirical p-values =0), *ZNF266* (18 SNPs), *RPL14* (17 SNPs), and *IPP* (27 SNPs).

2.2.3.6: Filtering for known Biological and Methodological Covariates

SNP and transcript-associated pairs that had a methodological covariate or biological covariate effect were removed from the result set. The methodological covariates included: day of expression hybridization, institute source of sample, post-mortem interval, and a covariate based on the total number of transcripts detected in each sample. The biological covariates included: gender, age at death, and cortical region. For the assessment of covariate effects we used a conservative approach. In the regression model, if any covariate term had an uncorrected p-value of less than 0.05 for a SNP-transcript pair, that SNP-transcript pair was excluded. This filtering step removed 52.2% and 23% of the *cis* and *trans* results respectively.

2.2.3.7: Polymorphism(s) in Assay Probes

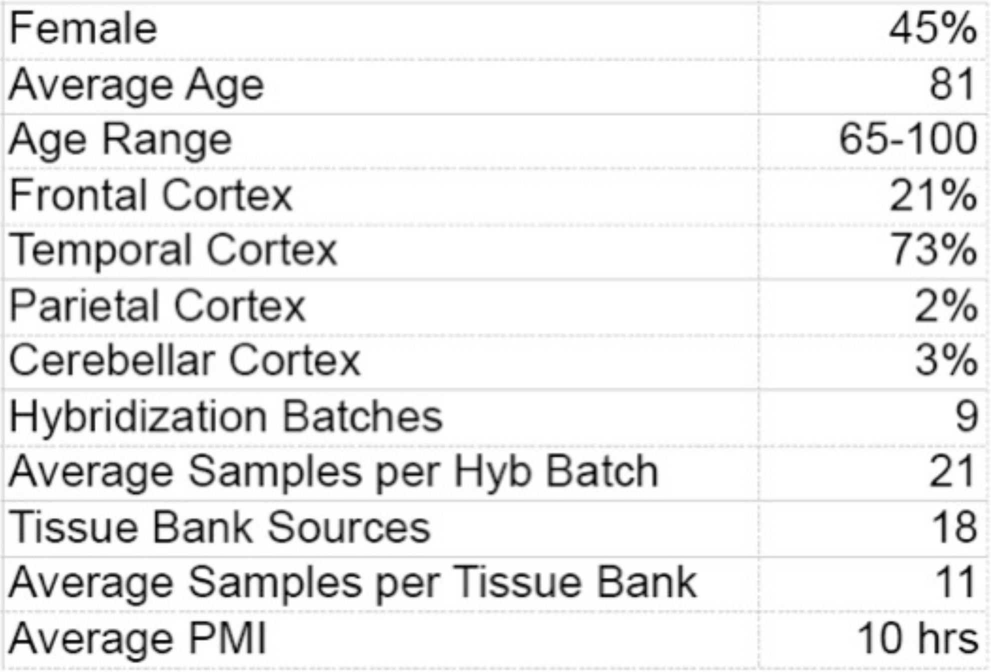
To account for any potential confounding effect of SNPs located within the transcript hybridization probes on the expression chips, any significant result where there was a variant in the transcript probe was removed. In order to do this we needed to determine whether known SNPs mapped to the transcript probes, thus creating a possible false positive through non-biologically relevant differential hybridization. We used the Hapmap II CEU genotyped data for this purpose. Hapmap II SNPs that were polymorphic in the CEU population were mapped with respect to the transcript probes. R-squared values between the significant SNP, within our screen, and the SNP within the probe from the HapMap II CEU dataset were then annotated within our results. For the transcript probes with a SNP located within the 50-mer hybridization probe it is presumed that a significant correlation between the probe's expression level and a *cis*-SNP is a false positive as a result of hybridization bias when the correlated SNP and the SNP in the probe are in LD. This filtering step removed 12.8% of the *cis* results. The potential confounds of polymorphisms within assay probe designs had previously been brought to light and we implemented a scan to remove the potential false positives from the study. In the concurrent period of this work, Alberts *et al.* published a study of this confound and its potential impact on eQTL studies. Their study showed that many mapped local eQTL in genetical genomics experiments do not reflect actual expression differences caused by sequence polymorphisms in *cis*-acting factors changing mRNA levels. Instead, they indicate hybridization differences caused by sequence polymorphisms in the mRNA region that is targeted by the microarray probes. Many such polymorphisms can be detected by a sensitive and novel statistical approach that takes the individual probe signals into account. Applying this approach to mouse and human eQTL data, they found that many local eQTL are falsely reported as "*cis*-acting" or "*cis*" and can be successfully detected and eliminated with this approach (Alberts *et al.* 2007).

2.2.4: Data and Biomaterial Access

Expression data and sample information have been deposited in NCBIs Gene Expression Omnibus (Edgar, Domrachev and Lash 2002) and are accessible through GEO Series accession number GSE8919 (Figures 9.1, *Appendix*). DNA from the samples, employed in this screen, is available upon request through the National Cell Repository for Alzheimer’s Disease (NCRAD), Indiana University, USA.

2.3: Results

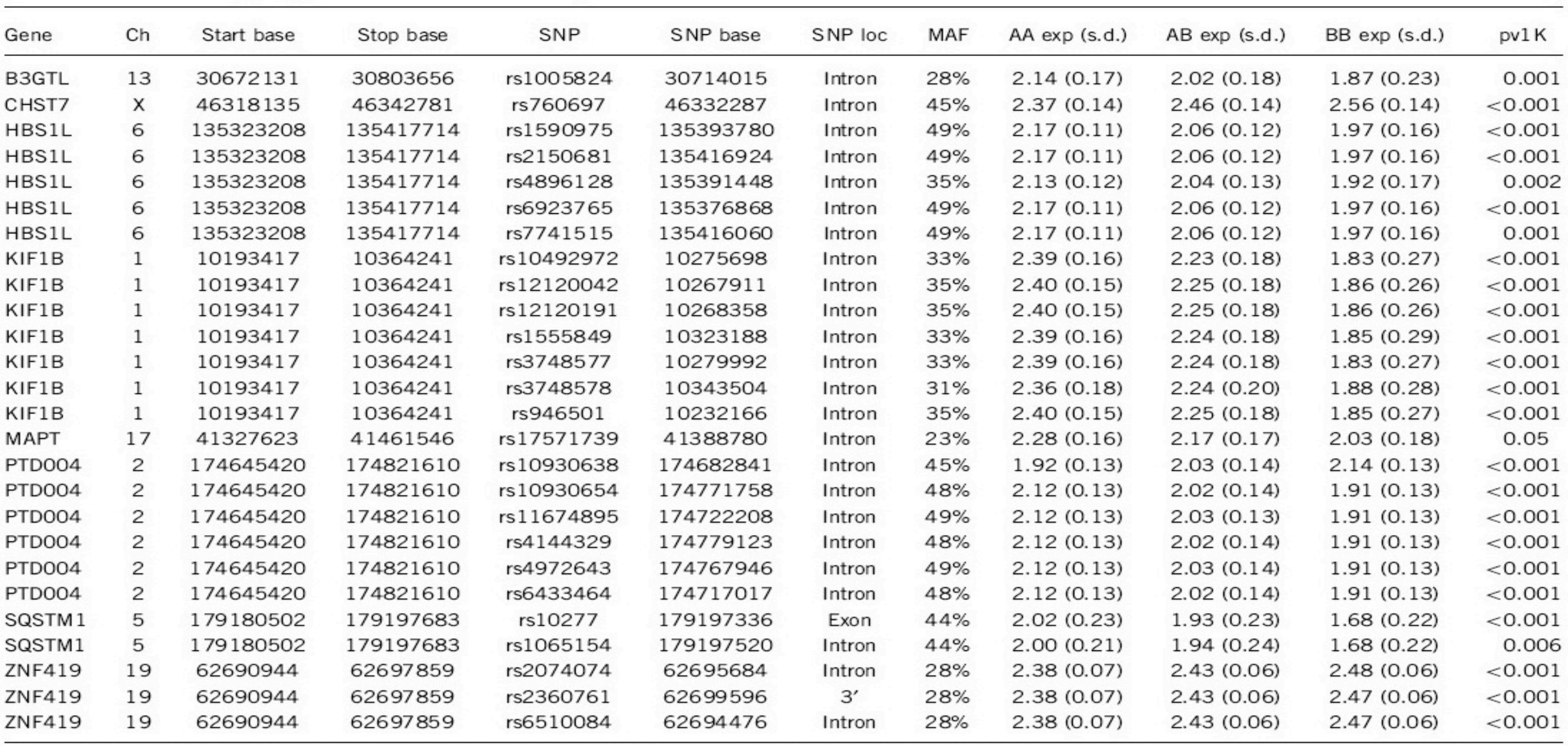
After quality control filtering 193 samples, 366,140 SNPs, and 14,078 mRNA transcripts were selected for the eQTL analysis. The analysis was performed by treating the expression profile of each transcript as the phenotype, i.e. a quantitative trait. A quantitative trait analysis was then performed on the genotype and expression data by linear regression to correlate allele dosage with expression. In addition, we corrected, by filtering, for several biological covariates and methodological covariates (Table 2.1).



Female	45%
Average Age	81
Age Range	65-100
Frontal Cortex	21%
Temporal Cortex	73%
Parietal Cortex	2%
Cerebellar Cortex	3%
Hybridization Batches	9
Average Samples per Hyb Batch	21
Tissue Bank Sources	18
Average Samples per Tissue Bank	11
Average PMI	10 hrs

Table 2.1: Summary of sample characteristics, which were used as covariates in the eQTL analysis. Table adapted from (Myers *et al.* 2007a).

Correlations between 366,140 SNPs and the expression of the 14,078 detected transcripts were assessed. In this analysis, after using a permutation based test correction and excluding results with a possible covariate effect, 852 transcripts were significantly correlated with genetic variation. An empirical p-value < 0.05, based on 1000 permutations, was used as a cut-off for a per trait significance of the SNP and transcript correlations, an eQTL. These significant association results were divided into *cis*-eQTL and *trans*-eQTL based on updated annotations. For this study, *cis* was defined as those associations that involved SNPs that are within the gene or within 1Mb flanking either the 5' or 3' end of the gene. The *trans* set are all the correlated SNP and transcript pairs that did not meet the *cis* criteria. *Trans* results are correlated pairs where the SNP and transcript were on different chromosomes or the SNP was greater than 1Mb from the transcript on the same chromosome. Of the 852 transcripts significantly correlated with genetic variants, 73 of these were correlated with one or more *cis* SNPs and 791 were correlated with one or more *trans* SNPs. While the total number of transcripts correlated in *trans* was greater than those in *cis,* calculating the proportions of significant possible *cis* and *trans* ratios revealed a significant enrichment for *cis* associations, with peak enrichment at approximately 20 Kb. The average distance, for *cis*-eQTL, between the SNP and the transcript is 55.4 Kb. On average the genetic variation accounts for 18.5% (*cis* mean is 22% and *trans* mean is 18.1%) of the expression variation for these 852 transcripts. Table 2.2 shows the results for the top eight transcripts with a *cis* eQTL.



Gene	Ch	Start base	Stop base	SNP	SNP base	SNP loc	MAF	AA exp (s.d.)	AB exp (s.d.)	BB exp (s.d.)	pv1K
B3GTL	13	30672131	30803656	rs1005824	30714015	Intron	28%	2.14 (0.17)	2.02 (0.18)	1.87 (0.23)	0.001
CHST7	X	46318135	46342781	rs760697	46332287	Intron	45%	2.37 (0.14)	2.46 (0.14)	2.56 (0.14)	<0.001
HBS1L	6	135323208	135417714	rs1590975	135393780	Intron	49%	2.17 (0.11)	2.06 (0.12)	1.97 (0.16)	<0.001
HBS1L	6	135323208	135417714	rs2150681	135416924	Intron	49%	2.17 (0.11)	2.06 (0.12)	1.97 (0.16)	<0.001
HBS1L	6	135323208	135417714	rs4896128	135391448	Intron	35%	2.13 (0.12)	2.04 (0.13)	1.92 (0.17)	0.002
HBS1L	6	135323208	135417714	rs6923765	135376868	Intron	49%	2.17 (0.11)	2.06 (0.12)	1.97 (0.16)	<0.001
HBS1L	6	135323208	135417714	rs7741515	135416060	Intron	49%	2.17 (0.11)	2.06 (0.12)	1.97 (0.16)	0.001
KIF1B	1	10193417	10364241	rs10492972	10275698	Intron	33%	2.39 (0.16)	2.23 (0.18)	1.83 (0.27)	<0.001
KIF1B	1	10193417	10364241	rs12120042	10267911	Intron	35%	2.40 (0.15)	2.25 (0.18)	1.86 (0.26)	<0.001
KIF1B	1	10193417	10364241	rs12120191	10268358	Intron	35%	2.40 (0.15)	2.25 (0.18)	1.86 (0.26)	<0.001
KIF1B	1	10193417	10364241	rs1555849	10323188	Intron	33%	2.39 (0.16)	2.24 (0.18)	1.85 (0.29)	<0.001
KIF1B	1	10193417	10364241	rs3748577	10279992	Intron	33%	2.39 (0.16)	2.24 (0.18)	1.83 (0.27)	<0.001
KIF1B	1	10193417	10364241	rs3748578	10343504	Intron	31%	2.36 (0.18)	2.24 (0.20)	1.88 (0.28)	<0.001
KIF1B	1	10193417	10364241	rs946501	10232166	Intron	35%	2.40 (0.15)	2.25 (0.18)	1.85 (0.27)	<0.001
MAPT	17	41327623	41461546	rs17571739	41388780	Intron	23%	2.28 (0.16)	2.17 (0.17)	2.03 (0.18)	0.05
PTD004	2	174645420	174821610	rs10930638	174682841	Intron	45%	1.92 (0.13)	2.03 (0.14)	2.14 (0.13)	<0.001
PTD004	2	174645420	174821610	rs10930654	174771758	Intron	48%	2.12 (0.13)	2.02 (0.14)	1.91 (0.13)	<0.001
PTD004	2	174645420	174821610	rs11674895	174722208	Intron	49%	2.12 (0.13)	2.03 (0.13)	1.91 (0.13)	<0.001
PTD004	2	174645420	174821610	rs4144329	174779123	Intron	48%	2.12 (0.13)	2.02 (0.14)	1.91 (0.13)	<0.001
PTD004	2	174645420	174821610	rs4972643	174767946	Intron	49%	2.12 (0.13)	2.03 (0.14)	1.91 (0.13)	<0.001
PTD004	2	174645420	174821610	rs6433464	174717017	Intron	48%	2.12 (0.13)	2.02 (0.14)	1.91 (0.13)	<0.001
SQSTM1	5	179180502	179197683	rs10277	179197336	Exon	44%	2.02 (0.23)	1.93 (0.23)	1.68 (0.22)	<0.001
SQSTM1	5	179180502	179197683	rs1065154	179197520	Intron	44%	2.00 (0.21)	1.94 (0.24)	1.68 (0.22)	0.006
ZNF419	19	62690944	62697859	rs2074074	62695684	Intron	28%	2.38 (0.07)	2.43 (0.06)	2.48 (0.06)	<0.001
ZNF419	19	62690944	62697859	rs2360761	62699596	3′	28%	2.38 (0.07)	2.43 (0.06)	2.47 (0.06)	<0.001
ZNF419	19	62690944	62697859	rs6510084	62694476	Intron	28%	2.38 (0.07)	2.43 (0.06)	2.47 (0.06)	<0.001

Table 2.2: Table shows the results for the top eight transcripts with a *cis*-eQTL, with annotation information for the SNP and Gene pairs. Table columns are: gene symbol (Gene), chromosomal (Ch), the gene's start and stop position on the chromosome (Start base and Stop base), variants dbSNP ID (SNP), the SNPs position on the chromosome (SNP base), the location of the SNP relative to the gene (SNP loc), the minor allele frequency for this SNP (MAF, based on these samples), genotype groups average expression (log10 of normalized intensity) with s.d. (AA, AB and BB Exp (s.d.)), and the empirical p-value based on 1,000 permutations (pv1K). Criteria for selection: was a *cis* test, no polymorphisms located within transcript probe, transcript specific empirical p-value based on 1,000 simulations < 0.05, gene expression detection rate within samples > 99%, SNP call rate within portion of sample used > 99%, number of minor homozygotes (BB genotype) > 3 and distance from SNP to gene < 3 kb. Genes are listed in alphabetical order. All genome positions are based on NCBI human genome build 39. Table adapted from (Myers *et al.* 2007a).

An analysis of the distances of SNPs correlated in *cis* with transcripts relative to the transcription start site (TSS) revealed a relatively symmetric distribution (Figure 2.4). This symmetry about the TSS is consistent with results shown for *cis*-eQTL in the HapMap lymphoblastoid cell line (LCL) samples (Stranger *et al.* 2007).

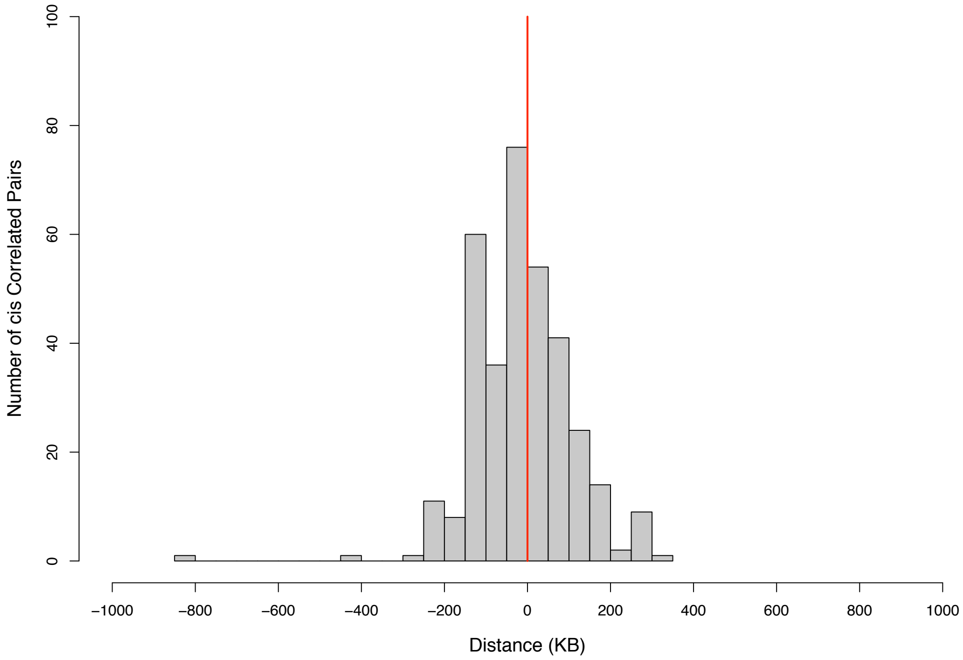


Figure 2.4: Distribution of *cis*-eQTL SNPs relative to the transcription start site (TSS).

It has previously been shown that *MAPT* expression is affected by the *MAPT* haplotype (Myers *et al.* 2007b). The cis-eQTL for *MAPT* in our results is consistent with this previous finding, alleles that occur on the major haplotype of *MAPT* (H1) are associated with higher *MAPT* transcript expression. It should be noted that a subset of the subjects included in our eQTL study were also used in the *MAPT* expression haplotype study. This provided an internal positive control that based on a genome- and transcriptome-wide analysis we can find effects that were previously seen in a candidate gene analysis of these samples, where that study was based on real-time PCR expression measures. It has also been shown that the *tau* (MAPT) protein expression levels in cerebrospinal fluid (CSF) vary with genetic variation at this locus (Figure 2.5) (Laws *et al.* 2007). In the Laws *et al.* study, their approach was to attempt to fine map the association of genetic variation on the H1c haplotype in the *MAPT* region with Alzheimer's disease (AD). Their approach made use of associating variants in this region with changes in *tau* protein levels in CSF, in subjects from Germany. Neurofibrillary tangles are present in AD pathology, which suggests a role for microtubule-associated protein *tau* (MAPT) in AD. Their analysis suggested that the AD locus could be narrowed to a region in close proximity of the SNP, rs242557, as this variant is correlated with CSF *tau* levels. The SNP, rs242557, is a haplotype tagging SNP for the H1/H2 haplotypes in the *MAPT* region and more specifically can be used to tag a sub-haplotype in the region, H1c, associated with AD (Laws *et al.* 2007).

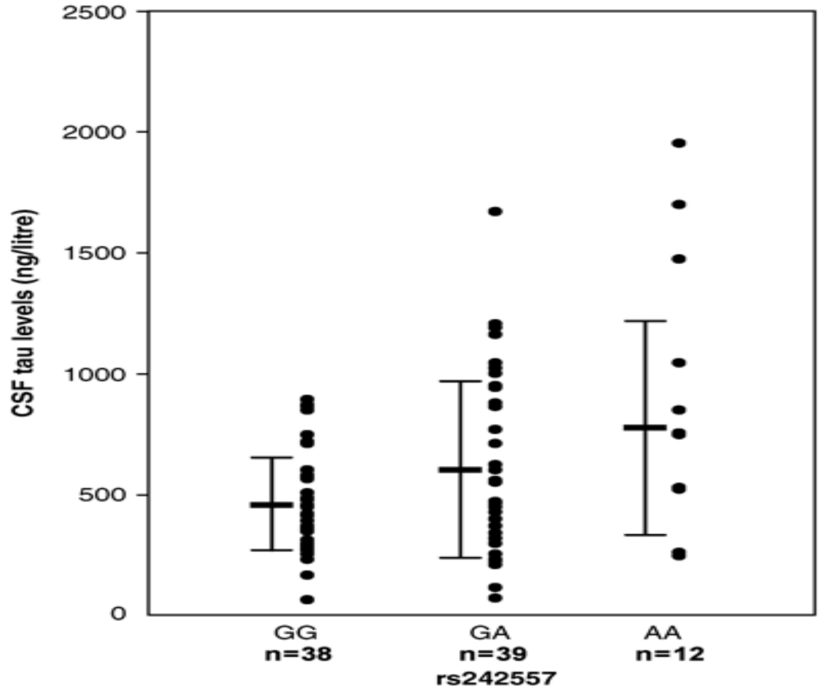


Figure 2.5: Plot showing the linear relationship between the genotypes of rs252557 and protein levels of *tau* in CSF. CSF levels per individual are shown according to the individual's genotype. Mean CSF levels and standard deviation of show next to the genotype groups. Linear regression of the allele dose with CSF levels results in a significant positive correlation with the A allele. Figure reproduced from (Laws *et al.* 2007).

While the *MAPT* locus did not contain the strongest eQTL found in this analysis, it is of importance because of its link to many neurodegenerative diseases, including tauopathies and Parkinson's disease (Simón-Sánchez *et al.* 2009; Höglinger *et al.* 2011). Also, as discussed in the introduction to my thesis, this region of the genome contains a large block of LD resulting from the presence of a genomic inversion that limits the recombination between the H1 and H2 haplotypes in the region. As such, we should see many *cis* SNPs correlated with the *MAPT* transcript; however, our initial analysis did not reveal this expected pattern of association. Upon further investigation of the *cis* signal it is apparent that there is association signal for the eQTL over the extended region of LD; however, much of the signal is just below the p-value cut-off threshold used and thus excluded from our significant results set (Figure 2.6).

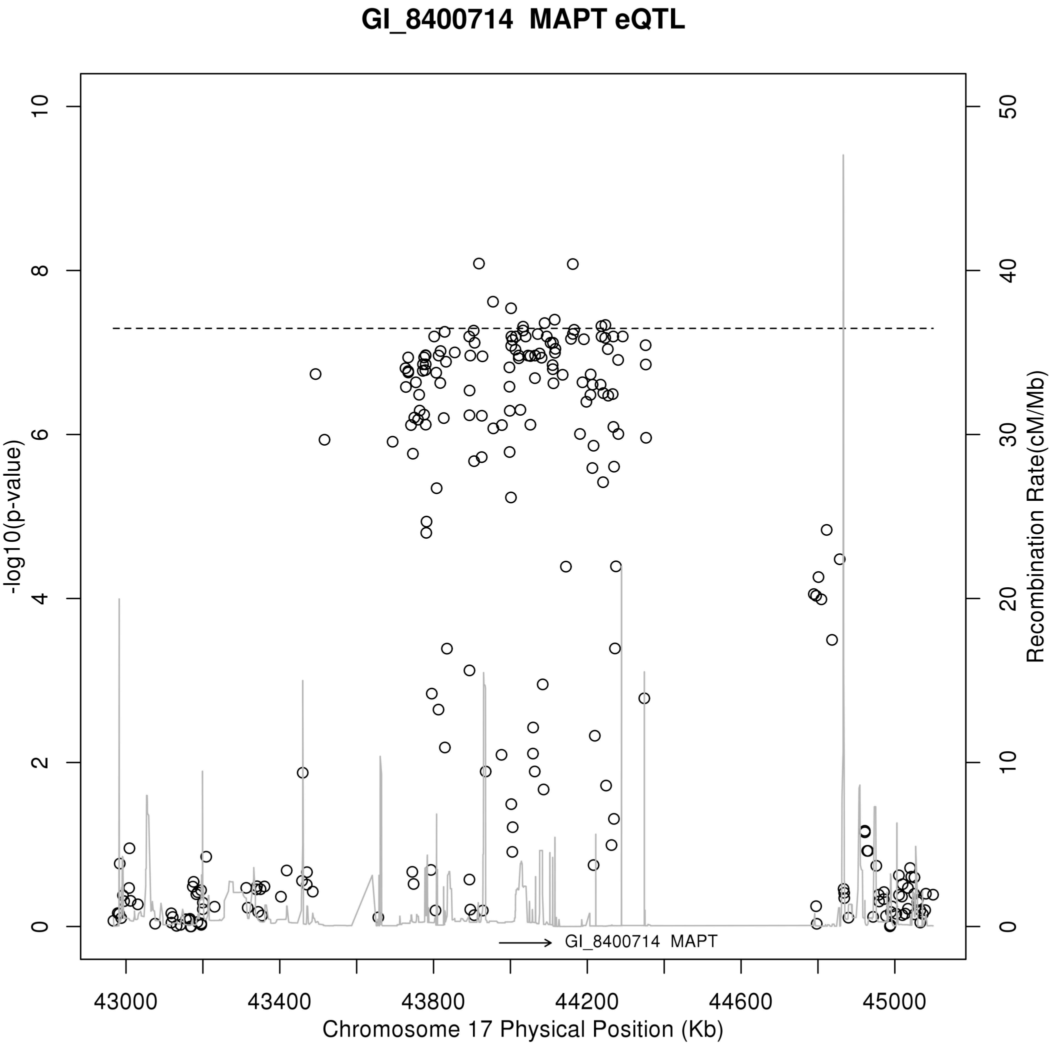


Figure 2.6: Manhattan plot showing this region's eQTL p-values for an mRNA transcript for the gene *MAPT*. Each point represents the p-value for a specific SNP along chromosome 17 that is *cis* to the *MAPT* transcript. Also included in the plot are the recombination rates (right axis) as a dark grey continuous line based on HapMap data. Threshold for significance is denoted by horizontal dashed line. The relative position of the gene is the labelled arrow centred near the bottom of the plot. The direction of the arrow is the gene's strand.

Another gene with an eQTL, that has been previously reported in population based eQTL studies in human lymphoblastoid cell lines, is ribosomal protein S26 (*RPS26)* (Cheung *et al.* 2005; Stranger *et al.* 2005). This gene is also found to have an eQTL in our analysis (Figure 2.7).

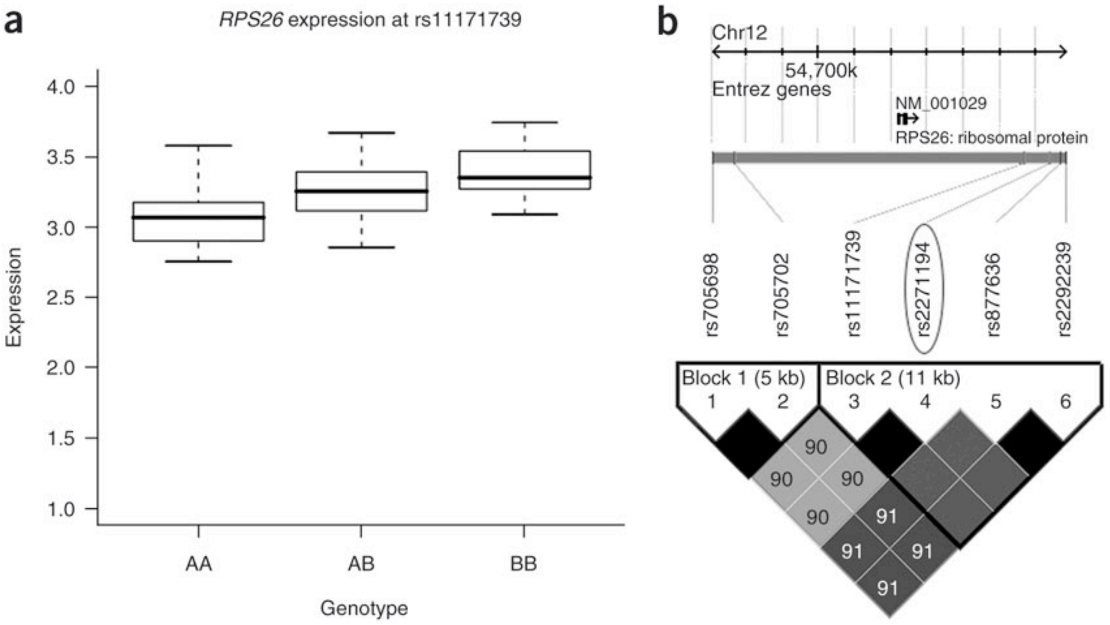


Figure 2.7: Shown is an eQTL for *RPS26* present in brain and also previously identified in eQTL analyses in LCLs. (a,b) Cheung *et al.* reported a significant association with the variant, rs2271194, which is in complete LD with two out of the six associated SNPs from our eQTL analysis in brain, including rs11171739, which was the SNP that gave the strongest association in brain. (a). Boxplot shows the linear relationship between genotype (x-axis is genotype allele dose), for rs11171739, and *RPS26* expression (y-axis is log10 normalized expression intensity) in brain. The boxplot shows expression summaries for the three genotype groups where top bar is maximum observation, lower bar is minimum observation, top of box is upper or third quartile, bottom of box is lower or first quartile, middle bar is median value. (b) LD plot, using the CEPH HapMap data, shows that the variants, significantly correlated with *RPS26* expression, rs11171739 in brain and rs2271194 (circled) in LCLs, are in complete LD. Complete LD between the two variants suggests it is likely that both screens are picking up the same association for *RPS26* expression. Haplotype block plots were created using Haploview (Barrett *et al.* 2005). Black boxes with no numbers indicate r^2 = 1. For r^2 values < 1, the r^2 value is shown as a percentage in white text in the box. Figure reproduced from (Myers *et al.* 2007a).

2.4: Discussion

In this study, we found that eQTL are detectable on a genome-wide scale in human brain tissues. Of cortically-expressed transcripts, the eQTL analysis performed suggests that 6% may have expression profiles that correlate with genotype. On average, genetic variation accounted for 18.5% of a gene expression variation for transcripts with a significant eQTL in brain. In 2007, Dixon *et al.* also published a large, for the time, eQTL study based on LCLs from 400 subjects. Their eQTL analysis included genotypes for 408,273 SNPs and 54,675 transcripts (from 20,599 genes). They first estimated the heritability of expression and found that 15,084 (28%) of the transcripts showed narrow-sense heritability. For these heritable transcripts, they performed an eQTL analysis and found that the peak SNP for each heritable expression trait on average accounted for 18% of the heritability. However, the more heritable the expression trait the more of the heritability was accounted for by the peak SNP from the eQTL analysis (Dixon *et al.* 2007). For the significant *cis*-eQTL identified, in our study of brain, on average the variant was located 55.4 Kb from the transcript. While these types of eQTL analysis have previously been performed in LCLs and for select genes in primary tissues, including brain, this was one of the first whole-genome eQTL studies in a human primary tissue. We found that while numerically more *trans* were detected in the analysis that the eQTL results are enriched for *cis*. Like previous studies, the *cis*-eQTL we identified are distributed symmetrically about the transcription start sites and the eQTL signals are generally larger closer to transcript start sites. This pattern likely reflects the LD present proximal to the gene and its regulatory regions and the decrease in LD with distance. In 2007, Stranger *et al.* published a study considering eQTL within and between populations from The HapMap Project. Their study was done using expression from LCLs including all 270 subjects from the HapMap Consortium. Their analysis included 2.2 million common SNPs, where variants were considered common if the MAF was at least 5% per population. They found 1,348 genes with a *cis*-eQTL and 180 with a *trans*-eQTL. Replication between at least two populations was found for 37% of the *cis* associations and 15% of the *trans* associations. Their results support an enrichment of functional *cis*-regulatory variants in the human genome (Stranger *et al.* 2007). In 2007, Bergen *et al.* published an eQTL study for genes commonly studied in cancer research in LCLs. They attempted to estimate how many genes may be affected by *cis*-variation using three separate analysis approaches. Their gene expression set included 697 genes from 30 LCLs combined with resequencing data from 552 genes, which resulted in 30 candidate genes with appropriate variation for a *cis*-eQTL analysis. They found significant *cis*-eQTL for eight of the genes tested. When they compared their results for 14 genes both with and without *cis*-eQTL signal to other literature sources they found 80% of genes with a *cis*-eQTL and 85% of genes without a *cis*-eQTL were concordant with previous studies. Based on their results and previous studies, they estimated that approximately 25% of genes have a significant *cis*-eQTL and that the eQTL signal accounts for approximately 30% of the gene's expression variation (Bergen *et al.* 2007). Based on these studies, there are some similarities in their findings when compared to the results from our study of eQTL in brain. The similarities include: distribution of *cis*-eQTL around the transcription start site, enrichment for *cis*- over *trans*-eQTL, and the average amount of expression variation, for transcripts with a significant eQTL, accounted for by the associated genetic variation. However, there are also some dissimilarities related to the proportion of genes identified with an eQTL and the rate of replication with other studies for the eQTL identified. For these dissimilarities, in our study of brain, we identified fewer transcripts proportionally than these other studies and fewer eQTL that have also been identified by other studies. These differences are likely related to the conservative approach we took in regards to excluding any result where a covariate had any indication of an effect on expression, and this study was one of the first to be performed genome-wide in a human primary tissue as opposed to LCLs.

This study does replicate an eQTL previously seen in gene specific studies on the expression of *MAPT*, a gene central to several neurodegenerative diseases. One of these studies, showing an eQTL for *MAPT* was performed in a subset of the samples included in this cohort (Myers *et al.* 2007b). This eQTL has also been seen previously for *MAPT* in a study related to progressive supranuclear palsy (Rademakers *et al.* 2005). Protein level changes of *tau* (MAPT) in relation to genotype, a protein QTL (pQTL), have also been previously reported in cerebrospinal fluid (Laws *et al.* 2007). Additionally the gene, *RPS26*, for which we have identified an eQTL, has also been reported in two other eQTL studies based on HapMap LCLs (Cheung *et al.* 2005; Stranger *et al.* 2005).

There are inherent limitations of this study. First, we used more than one brain region for the source of tissue; this means that the eQTL detected here may well be generalizable across brain tissues, and indeed across cell types, but that we may be missing other cell- or region-specific eQTL. Additionally, it is unlikely that including multiple brain tissues created false positives related to known variation in gene expression between brain tissues, given our conservative handling of covariates. It is instead likely that the approach taken to deal with covariate effects, and including the tissue region as a covariate, resulted in an increase of false negatives. For example, if a transcript did have a significant eQTL, where the genetic variation accounted for large amount of the expression variation, but at the same time the tissue region covariate had a very small but detectable effect for this eQTL, it would be excluded from our significant result set. In 2007, Hovatta *et al.* published an eQTL study based in inbred mice for region-specific brain expression. Their study design was to focus on specific brain regions whereas previous studies in mice typically used whole brain homogenates of large regions. However, the brain is heterogeneous and expression profiles do differ between regions. Their study included five brain regions from six inbred mouse strains. They did not find a large number of strain-specific genes based on gene expression but did find a large number of genes that showed region-specific expression profiles. However, for genes showing strain-specific expression profiles these were constant across brain regions. Based on an eQTL analysis, they found an enrichment of *cis*-regulators for strain-specific genes but for region-specific genes the enrichment was for *trans* elements. Their results suggest that many regulatory networks are tissue-specific and that this suggests that it is important to perform eQTL studies in tissues that are relevant to the phenotype of interest (Hovatta *et al.* 2007). A second limitation in our study was that the genotyping assay used was an early technology, and while it represented a significant advance over previous methods, the coverage of known common genetic variation was low compared to more recent genotyping platforms. More recent genotyping platforms typically base their design on variants that tag HapMap haplotypes thereby increasing the coverage of known genetic variability in the human genome captured during genotyping. Thirdly, this type of work captures expression in a cross-sectional manner, in post-mortem tissue; therefore, it is difficult to address other types of expression that may be influenced by genotype, such as (for example) induced expression, or sub-cellular localization of transcripts. Another limitation was the detection threshold selected in determining whether a transcript was well detected in our study and therefore suitable for eQTL analysis. In this study we defined a transcript as well detected if it was present in at least 5% of the samples. However, at the same time as part of the analysis, in identifying significant eQTL, a filter was applied post hoc requiring that all significant eQTL have at least three expression measurements for the minor homozygote to ensure that the allele dosage regression was not based on incomplete data. This means that many transcripts with a low detection rate were tested for an eQTL that would never be considered significant in our analysis design. This resulted in an unnecessary increase in our multiple test burden as well as computation time. Lastly, as with other eQTL methods of the time, this work relies on array-based expression analysis, and this method does not capture splicing events as effectively as (for example) sequence-based methods. These limitations notwithstanding, we believe this study provides an initial resource of eQTL in human cortical tissues that is of use for researchers investigating loci and gene models related to neurological disease. It is likely that further study of eQTL in primary tissues will add to our functional understanding of the effect of genetic variation in the human genome and serve as a resource in the study of complex traits associated with this variation.

3: Identifying eQTL in Distinct Human Brain Regions

(Gibbs *et al.* 2010)

Statement of Contribution to this Research:

I was involved in the conception and design of this study, including choice of genotyping platform, expression platform, and selection of tissue. I performed data quality control, data analysis, and data interpretation. I co-drafted and edited the manuscript. Cookson MR, Singleton AB, van der Brug MP, Hernandez DG, Traynor BJ, and Longo DL were also involved in the conception, design, choice of platform, and tissue selection. Cookson MR, Singleton AB, and I drafted and edited the manuscript. Nalls MA and Singleton AB contributed to the data analysis. I was not involved in the collection of the tissue samples or the generation of the genotype and mRNA expression data. Coordination and collection of the tissue was performed by: Traynor BJ, Troncoso J, Johnson R, Zielke HR, Lai SL, and Ferrucci L. Genotyping of the samples was coordinated or performed by: Hernandez DG, Traynor BJ, Arepalli S, Rafferty IP, and Lai SL. Generation of mRNA expression data was coordinated or performed by: van der Brug MP, Dillman A, and Cookson MR.

3.1: Introduction

Because of our interest in genomic regulation of expression and neurological disorders we embarked upon a series of experiments to provide a brain region-specific contextual framework for genetic regulation of gene expression. Based on the previous work (Chapter 2) it was apparent that a systematic analysis of eQTL in brain tissue was feasible. We embarked upon a set of integrated experiments designed to extend this previous work.

We obtained frozen brain tissue from the cerebellum (CRBLM), cerebral frontal cortex (FCTX), caudal pons (PONS), and cerebral temporal cortex (TCTX) from 150 subjects (a total of 600 tissue samples). We undertook two separate assays across this series, genome-wide genotyping more than 500,000 SNPs and mRNA expression profiling more than 24,000 transcripts in all four brain regions. Here I will discuss the results of these experiments, particularly in the context of integrated datasets to define expression quantitative trait loci (eQTL) and detailing differences and similarities across brain regions.

Based on the successful completion and publication of the preceding, ‘pilot’, study (Chapter 2) we were able to initiate a second study of eQTL within human brain tissues with an expanded study and refined analysis design. The expanded study design included collecting and assaying multiple brain tissues from each individual, making use of newer and improved assay types and applying lessons learned in the ‘pilot’ study to improve analysis design and efficiencies. We employed dense tagging-based, whole-genome, SNP genotyping thereby improving coverage of genetic variability, within the human genome, for cohorts of central European descent. Additionally, we used newer versions of gene expression arrays that capture mRNA transcripts for known human transcripts based on more recent RefSeq information. Improved analysis design and efficiency over the previous work were attained in three primary areas: improved covariates, the implementation of imputation, and exclusion of inappropriate tests. For covariates we had previously established that biological and technical covariates could have significant effects, and thus adjusted expression profiles for covariate information prior to eQTL analysis instead of filtering out possible significant eQTL where a covariate effect also exists. Using imputation to increase the density of genetic markers that are available for testing increases the power and ability to fine map the associations. Excluding transcripts that are not appropriate for inclusion, prior to eQTL analysis instead of after, reduced both the computational and multiple test burden.

This chapter describes this effort and the observations we made relating to improved analytical approaches, the improvement in results imparted by denser genotype coverage, and the comparison across distinct brain tissues from the same individuals.

3.2: Materials and Methods

3.2.1: Subjects, North American Brain Expression Consortium (NABEC)

(Coordination and collection of the tissue was performed by: Traynor BJ, Troncoso J, Johnson R, Zielke HR, Lai SL, and Ferrucci L.)

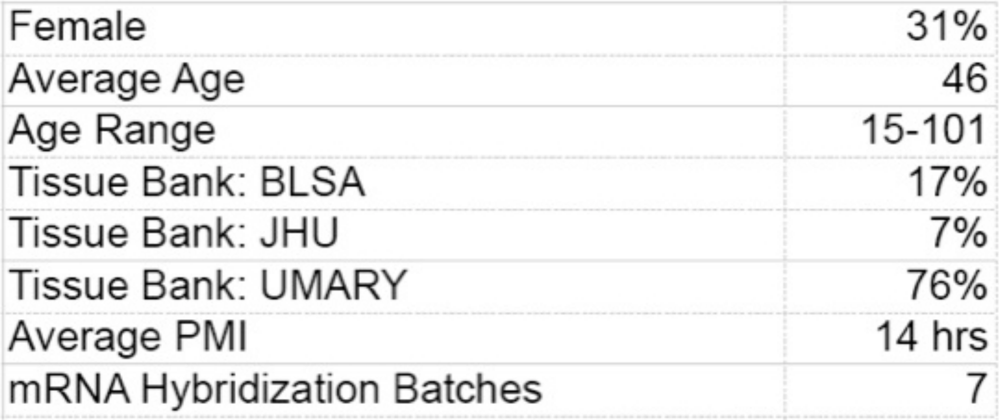
Frozen tissue samples from the cerebral frontal cortex, cerebral temporal cortex, cerebellum, and caudal pons were obtained from 150 subjects who had donated their brains for medical research. Approximately 100-200 mg aliquots of frozen tissue were sub-dissected from each of the 600 samples

(150 brains x four regions) resting on dry ice to avoid thawing. Separate pieces were cut for DNA extraction to be used in SNP genotyping assays and RNA extraction for expression assays. Genomic DNA for genotyping was extracted using the DNeasy Blood and Tissue Kit as per the manufacturer's instructions (Qiagen Inc., Valencia, California, USA). Total RNA was prepared using TRIzol (Invitrogen, Carlsbad, California, USA).

3.2.1.1: Subject Characteristics

One hundred and fourteen brains were sampled from the University of Maryland Brain Bank, Baltimore, Maryland, USA. Thirty-six brains were sampled from the Department of Neuropathology, Johns Hopkins University, Baltimore, either as routine autopsy cases (n = 10), or as part of the National Institute on Aging-sponsored Baltimore Longitudinal Study of Aging (BLSA, n = 26). All individuals were of non-Hispanic, Caucasian ethnicity, and none had a clinical history of neurological or cerebrovascular disease, or a diagnosis of cognitive impairment during life.

Summary statistics of the sample characteristics are shown in Table 3.1. The most common cause of death was accidental injury (n = 55 cases), followed by cardiovascular disease (n = 31), drug intoxication (n = 12), and pulmonary embolism (n = 3). Other causes of death included drowning (n = 3), respiratory disease (n = 2), compressional asphyxia (n = 1), suicide by hanging (n = 1), choking (n = 1), lightning strike (n = 1), liver disease (n = 1), mitral valve prolapse (n = 1), myocarditis (n = 1) and diabetic coma (n = 1). Cause of death was not available for the remaining 36 autopsies.



Female	31%
Average Age	46
Age Range	15-101
Tissue Bank: BLSA	17%
Tissue Bank: JHU	7%
Tissue Bank: UMARY	76%
Average PMI	14 hrs
mRNA Hybridization Batches	7

Table 3.1: Summary statistics of the subject characteristics

3.2.1.2: Sample Preparation

(Sample preparation from the tissue was performed by: Traynor BJ, Lai SL, Hernandez DG, and van der Brug MP.)

For each of the six hundred samples (150 brains x four regions), approximately 5 grams of frozen tissue was sub-dissected at either the University of Maryland Brain Bank or at the Department of Neuropathology, Johns Hopkins University, and sent on dry ice to the Laboratory of Neurogenetics (LNG), NIA. At LNG, 100-200mg aliquots of frozen tissue were sub-dissected from each sample. Samples were kept on dry ice to avoid thawing. Separate pieces were cut for DNA extraction to be used in SNP genotyping assays and RNA extraction for expression assays. Each tissue aliquot was stored at -80°C until use.

Genomic DNA extraction for genotyping was performed using the DNeasy Blood and Tissue Kit as per the manufacturer’s instructions (Qiagen Inc., Valencia, California). DNA concentration was determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, Delaware), and DNA extraction was repeated using a new tissue aliquot for samples with DNA concentration less than 50 ng/uL, or for samples where the 260nm/280nm wavelength absorption ratio was less than 1.7, indicative of significant protein contamination of the DNA sample.

For each of the 600 brain samples, total RNA was prepared from approximately 100 mg of tissue using a glass-Teflon homogenizer and 1 mL TRIzol (Invitrogen, Carlsbad, California, USA) according to the manufacturer’s instructions. RNA samples were re-suspended in RNAse free water to a final concentration of > 500 ng/uL.

3.2.2: Assays

3.2.2.1: SNP Genotyping

(Genotyping of the samples was coordinated or performed by: Hernandez DG, Traynor BJ, Arepalli S, Rafferty IP, and Lai SL.)

Genotyping was performed using DNA extracted from cerebellar tissue. SNP genotypes were assayed using Illumina Infinium HumanHap550 version 3 BeadChips (Illumina Inc., San Diego, California, USA) according to the manufacturer’s instructions. Genotype data was analysed using the Genotyping Analysis Module 3.2.32 within the BeadStudio software version 3.1.4 (Illumina Inc.). All 150 brain samples had an average call rate of 99.86% (range 97.72% - 99.95%).

The Illumina HumanHap550 chip assays genotypes for 561,466 SNPs across the genome. The Illumina Infinium genotyping platforms are based on the Sentrix bead arrays. The Sentrix arrays are a single base-resolution platform, which helps avoid some of the sequence complexity that may arise from calling genotypes based on oligonucleotide probe arrays. This system is based on allele-specific primer extension and includes a two-colour readout, one colour for each allele tested (Figure 3.1). Based on the manufacturer's comparisons to PCR-based genotyping assays the Illumina arrays had a call rate of 99.7%, reproducibility of 99.96%, and concordance rate of 99.97% (Gunderson *et al.* 2005). In a follow on paper describing the efficiency, accuracy and scalability of the Illumina whole-genome genotyping platform the authors compared the genotype reproducibility rates based on HapMap samples and found that concordance was above 99% (Steemers *et al.* 2006). This array is based on haplotype tagging SNPs and provides 87% coverage of known genetic variation in HapMap II CEU population (Li, Li and Guan 2008).

Figure 3.1: Cartoon of genotyping using the Illumina Infinium Sentrix arrays. A) Each variant is represented by a bead-type where fragmented sample DNA binds to a complementary probe sequence stopping one base before the allele being assay. B) Single-base extension incorporates one of four labelled nucleotides conferring allele-specificity, extending the probe with the correct base. C) Probes are laser excited causing the nucleotide label emits a colour signal for detection. Figure adapted from Illumina promotional material.

3.2.2.2: RNA Expression

(Generation of mRNA expression data was coordinated or performed by: van der Brug MP, Dillman A, and Cookson MR.)

Profiling of 22,184 mRNA transcripts was performed using HumanRef-8 version 2 Expression BeadChips (Illumina Inc.) in accordance with the manufacturer's protocol. Raw intensity values for each probe were transformed using the rank invariant normalization method (Workman et al., 2002; Schadt et al., 2001; Tseng et al., 2001) using the Gene Expression Module 3.2.7 within Illumina's BeadStudio software. The Illumina Sentrix

HumanRef-8 v2.0 Expression BeadChip assays the expression levels of approximately 24,000 human Refseq transcripts using 50-mer probes.

3.2.3: Data analysis

For each of the four brain regions, a regression analysis was performed on the expression intensities generated for mRNA transcript probes. Gender, age, post-mortem interval (PMI), tissue source, and hybridization batch were included as covariates in each of these analyses. Residuals from the regression analysis for each probe were then used as the quantitative trait for that probe in a genome-wide association analysis to identify quantitative trait loci. These analyses were performed using the *assoc* function within PLINK, which correlates allele dosage with change in the trait (Purcell et al., 2007). Each of the four tissue regions were analysed separately, and independent genome-wide association analyses were performed looking for quantitative trait loci associated with mRNA expression levels (expression QTL; eQTL). To correct for the large number of SNPs tested per trait, a genome-wide empirical p-value was computed for the asymptotic p-value for each SNP using 1,000 permutations of sample-label swapping. To correct for the number of traits being tested per tissue region, a false discovery rate (FDR) threshold was determined based on the empirical p-values using the *fweR2fdr* function of the *multtest* package in R (Pollard, Dudoit and van der Laan 2005). Empirical p-values were allowed to exceed this threshold if the linkage disequilibrium R^2 was greater than or equal to 0.7 with a SNP with empirical values within the FDR threshold. Keeping SNPs with a sub-significant empirical p-value, if they are in strong LD with a SNP significantly correlated with a transcript's expression, does not alter whether or not a significant eQTL is detected for a transcript, but does allow for a broader detection of the edges of the locus. The sequences of transcript probes with significant eQTL were examined for the presence of polymorphisms, with a MAF > 1%, using CEU HapMap data, and if present that eQTL was removed from the result set.

3.2.3.1: Genotyping data

The threshold call rate for inclusion of a sample in the analysis was 95%. Two samples initially had a call rate below this threshold, but were successfully re-genotyped using fresh DNA aliquots. Thus, all 150 brain samples had a call rate greater than 95%, and were included in the subsequent analyses for quality control of the subjects based on their genotypes (average call rate was 99.86%; range 97.72% - 99.95%, based on the *missing* procedure within the PLINK v1.04 software toolset (Purcell *et al.* 2007)).

The gender of the samples reported by the brain banks was compared against their genotypic gender using PLINK 's *check-sex* algorithm. The *check-sex* function determines a sample's genotypic gender based on heterozygosity across the X chromosome. Two samples with gender discrepancies were detected. One of these arose from a clerical error at the brain bank and was included in the analysis after correction of the clinical information, whereas the other sample (UMARY1496) was removed from subsequent analysis.

To confirm the ethnicity of the samples, Identity-By-State (IBS) clustering, principal components (PCA), and multidimensional scaling (MDS) analyses

(Price *et al.* 2006) were performed within PLINK using the genotypes from the brain samples that had been merged with data from the four HapMap I (International HapMap Consortium 2005) populations (n = 32 Caucasian (CEU), 12 Han Chinese, 16 Japanese and 24 Yoruban non-trio samples previously genotyped by Illumina and assayed on the Infinium HumanHap500 version genotyping chips). Outlier detection was based on a sample's distance, for the first and second principal components, being more than three standard deviations (S.D.) from the mean of the reported population group for those components. Two samples were outliers based on population and were excluded from further analysis (UMARY4545, UMARY927) (Figure 3.2). MDS is a method for information visualization, particularly for distance metrics. MDS aims to place each item in N-dimensional space such that the between-object distances are preserved as well as possible. PCA is a data transformation that converts possibly correlated variables into linearly uncorrelated components; where the first component accounts for largest variance, the second component the second most variance, and the *N*th component the least amount of variance. Genotype data of the samples was compared to identify cryptic relatedness using the Identity-By-Descent (IBD) procedure within PLINK. No samples were found to be from related individuals. The IBS/IBD analysis, using PLINK, estimates a genome-wide IBD measure between each pair of samples. This estimate, based on the sharing of alleles between each pair of samples, can identify individuals that appear more similar, to each other, than expected by chance.

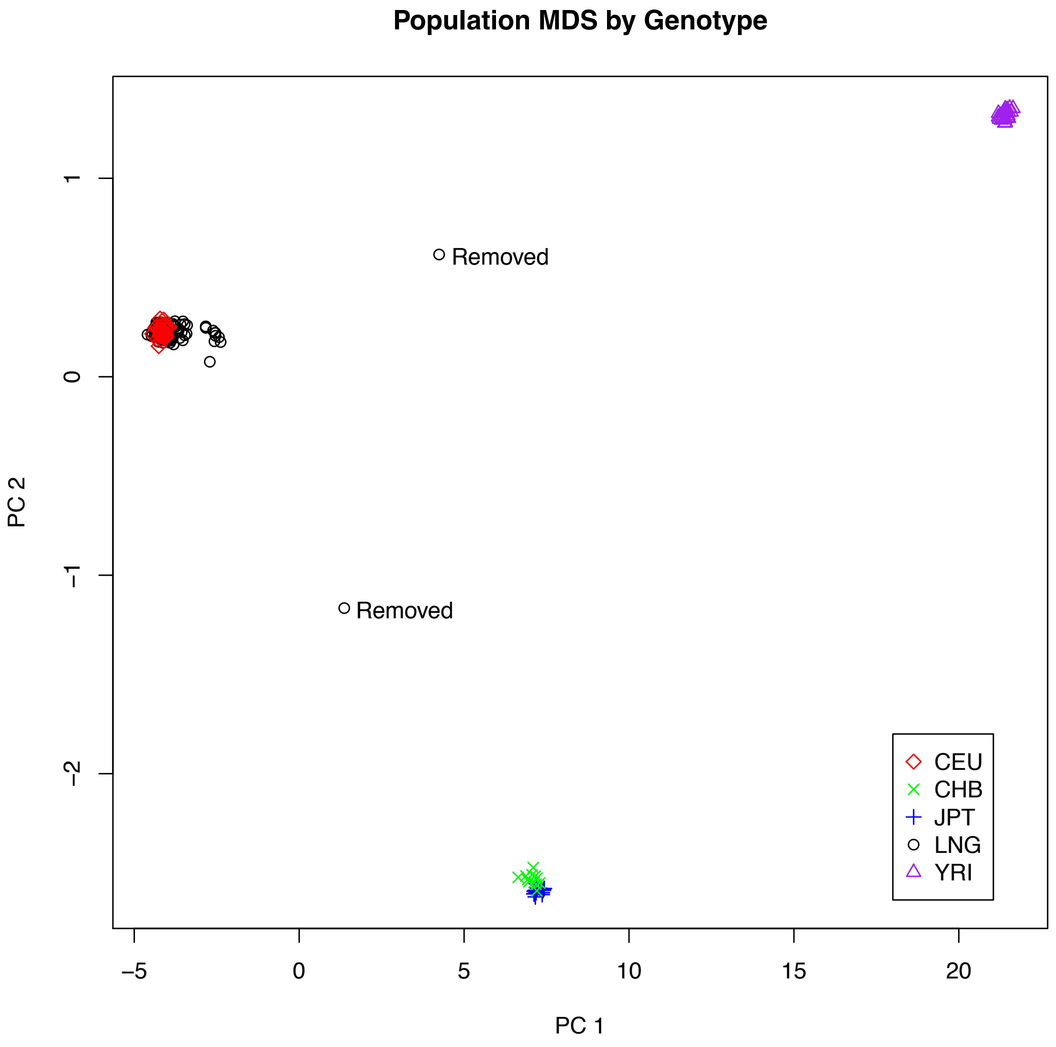
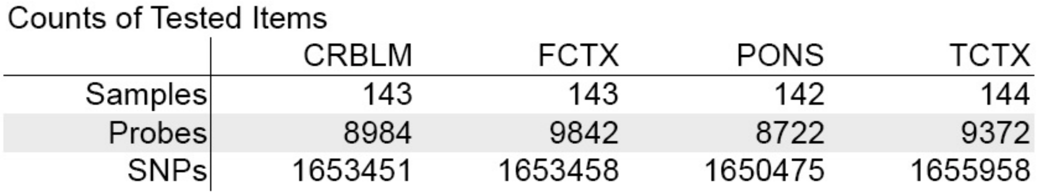


Figure 3.2: Population MDS plot based on genotype, from genome-wide IBS pairwise distances between the 150 samples used in this study (LNG, black circles) and HapMap samples (CEU, red diamonds; CHB, greens x's; JPT, blue +'s, and YRI, purple triangles). The plot shows that of the 150 samples, from the study, reported to be Caucasian individuals from the United States, two samples are ethnic outliers relative to the rest of the study cohort and the CEU population from HapMap (indicated by Removed labels). Outlier detection was based on 3 S.D. from the mean for the reported population group. Figure reproduced from (Gibbs *et al.* 2010).

Mach software version 1.0.16 (Li *et al.* 2009, 2010) and HapMap 2 CEU phase data (release 22) were used to impute genotypes for ~2.5 million SNPs. Imputed SNPs were excluded if the linkage disequilibrium r^2 values between imputed and known genotypes were less than 0.3, and if their posterior probability averages were less than 0.8 for the most likely imputed genotype. For each of the four tissue regions, SNPs were also excluded if: (a) the call rate was less than 95%, (b) Hardy-Weinberg equilibrium (HWE) p-value was less than 0.001, and (c) the SNP had less than 3 minor homozygotes present. On average, for the four tissues groups, ~1.6 million SNPs passed these quality threshold checks and were appropriate for use in the eQTL analyses. The exact numbers of SNPs used per tissue group are shown in Table 3.2.



Counts of Tested Items

	CRBLM	FCTX	PONS	TCTX
Samples	143	143	142	144
Probes	8984	9842	8722	9372
SNPs	1653451	1653458	1650475	1655958

Table 3.2 Summary counts of total subjects, mRNA transcript probes, and SNPs that were included for analysis per tissue region. Each column represents a tissue region: cerebellum (CRBLM), cerebral frontal cortex (FCTX), caudal pons (PONS), and cerebral temporal cortex (TCTX). Table adapted from (Gibbs *et al.* 2010).

3.2.3.2: RNA expression data

Raw intensity values for each probe were transformed using the rank invariant normalization method (Schadt *et al.* 2001; Tseng *et al.* 2001; Workman *et al.* 2002) for mRNA analysis. Individual samples that had an average detection score less than 0.99 were either discarded or re-run from a separate preparation. The following mRNA samples were excluded based on this metric: UMARY1668 (CRBLM), UMARY1909 (FCTX), UMARY4543 (PONS) and UMARY4782 (PONS). The following individuals were not run on mRNA expression arrays for any tissue region: BLSA1672, JHU1344 and JHU1361.

3.2.3.3: Clustering of Samples by Brain Region

Performing a Hierarchical Clustering (HCL) (Eisen *et al.* 1998) of the sample expression profiles using the TM4 MeV version 4.1.01 tool (Saeed *et al.* 2003), ‘Average Linkage clustering’ resulted in the samples separating by brain tissue region. Separation of samples into four distinct clusters matching the brain tissue region was clear (Figure 3.6A). For clustering all detected transcripts were used. The HCL samples trees were saved as Newick tree files and plotted again using the HyperTree tool (http://hypertree.sourceforge.net/). The Newick format is a standard format for representing visual data trees in a computer-readable format.

3.2.3.4: Selection of traits for analysis

Traits were excluded from analysis if they were detected in less than 95% of samples for each tissue region. For each tissue region and trait type the 95% threshold was determined using total number of analysable samples, for this pairing of region and trait. Only probes that were detected in 95% of all samples within a tissue type were used for further analysis. In total, 10,326 mRNA transcripts were analysed within at least one brain tissue region; 8,076 (78%) mRNA transcripts were analysed within all four brain tissue regions (Figure 3.3).

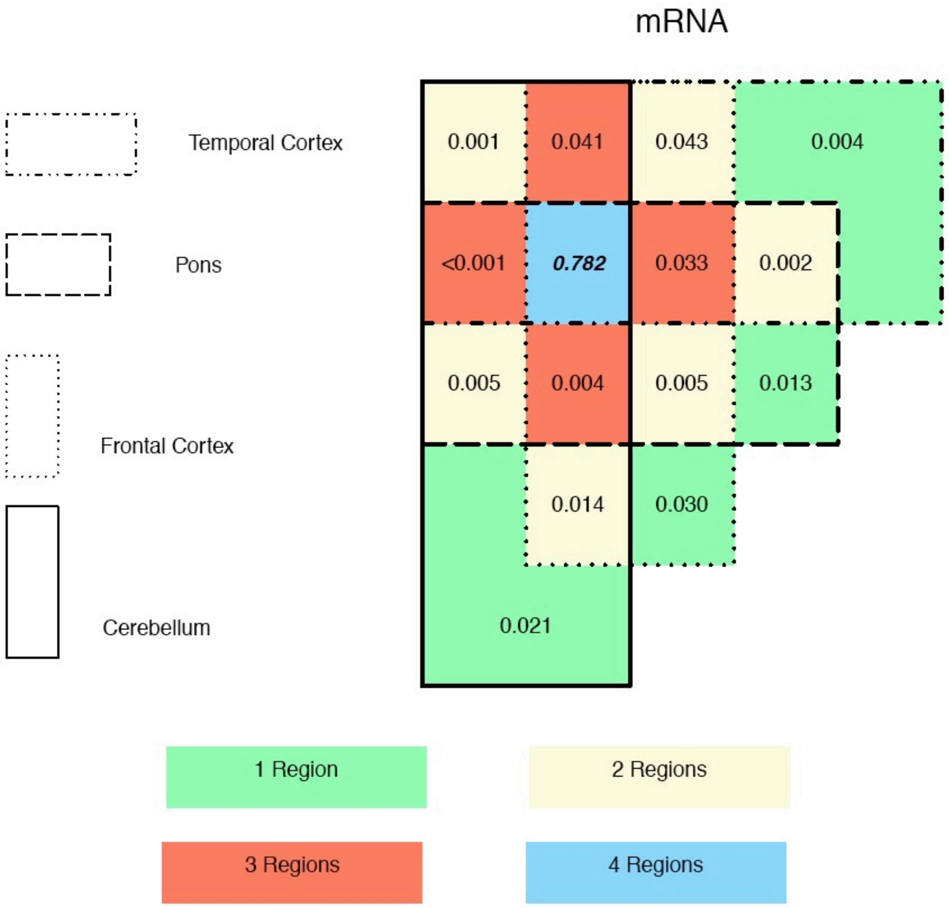


Figure 3.3: Stylized Venn diagrams showing the frequency overlap, between the four brain tissues, of the number of transcript probes that were detected in 95% of samples. The rectangles with different orientations and border, shown on the left legend represent the different tissue and the different squares represent overlapping frequencies between different tissues. The colour coded squares represent the number of tissues overlapping, where the central blue square in the diagram represent the number of probes reliably detected in all four tissues. Hence, the blue square in the diagram indicates that 78.2% of transcript probes were detected in all four tissue regions. Figure adapted from (Gibbs *et al.* 2010).

Using a 95% detection threshold, in the tissue sample series, is a more appropriate cut-off for inclusion in the eQTL analysis as it makes more appropriate use of the power of the series and reduces the number of traits analysed, which will then likely be discarded by multiple test correction and other trait selection criteria. In my previous thesis study, many initially suggestive *trans*-eQTL and some *cis* were false positives related to low detection rate of the transcript (Figure 3.4). These, in most instances, were filtered from the published result sets with additional test correction and filtering steps.

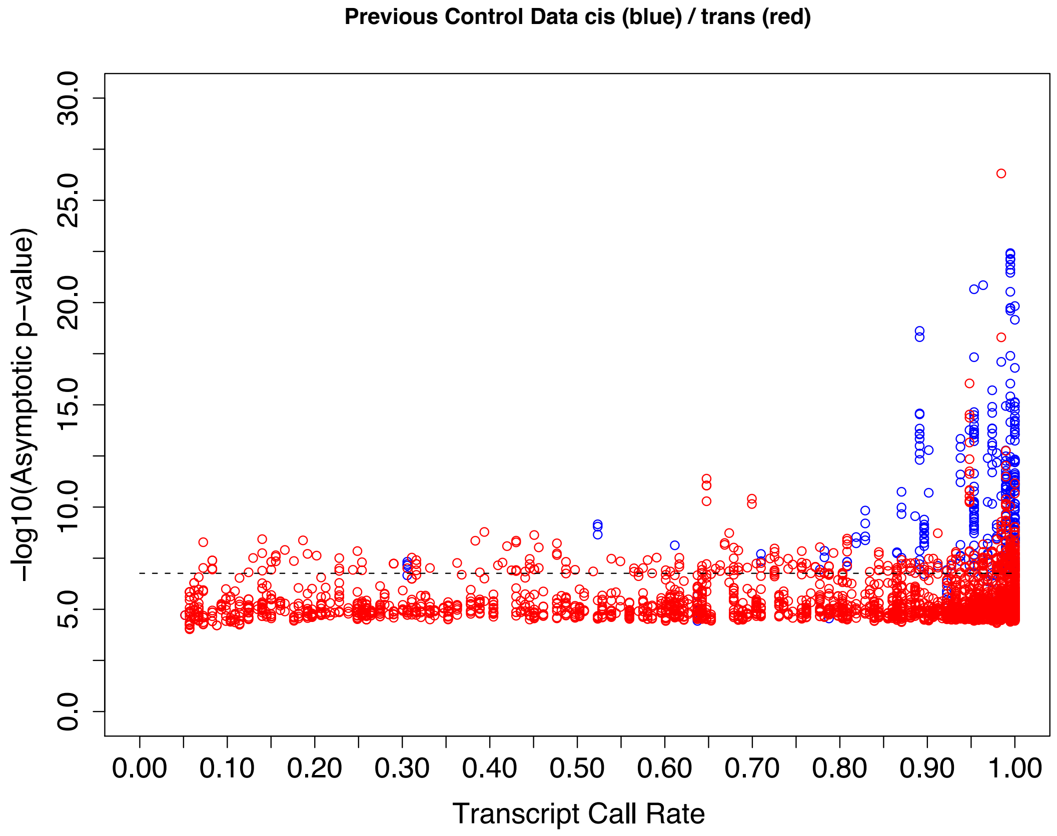


Figure 3.4: Plot showing transcript detection rate based on all subjects and suggestive or better eQTL signal from the mix cortical tissue eQTL analysis study (Chapter 2). In that study a transcript had to be detected in 5% of subjects to be included for analysis. In the later two studies this was increased to 95%. As the plot clearly demonstrates much of the *trans* signal (red) and some of the *cis* signal (blue) were for transcripts that were not well detected in the series. Even though many of these were removed by other corrections and filters, prior to publication for that study, it still dramatically increased computation time and test burden for transcripts that should not have been included for analysis.

3.2.3.5: Polymorphism(s) in Assay Probes

Sequence variants within the sequence of the probe used to assay individual traits may cause differential hybridization and inaccurate expression measurements. To exclude this confound, the sequences of probes with significant correlation to a trait were examined for the presence of known polymorphisms, with a MAF > 1%, using CEU HapMap data, and if present, that QTL was removed from the result set. Removal of eQTL, where the transcript's probe sequence contained a known polymorphism excluded 36 (10.1%) of transcripts from the CRBLM, 35 (9.5%) from FCTX, 30 (9.7%) from PONS, and 44 (10.1%) from TCTX. Of the above set of transcripts, excluded from the analysis results, seven were present in all four tissue regions results.

3.2.3.6: Correction for known Biological and Methodological Covariates

Prior to quantitative trait loci analysis, each trait was adjusted using the available biological and methodological covariates in an attempt to remove the influence of these potentially confounding affects. In R, each trait was regressed using the following model:

$$Y=\beta_0 + \beta_1X_1+\ldots\ \beta_nX_n + \varepsilon$$

In this model, Y is the trait profile ($\log_2$ normalized mRNA expression intensities) and $X_1 \ldots X_n$ represent the biological covariates Age and Gender and the methodological covariates post-mortem interval (PMI), which Brain Bank the samples was from, and which preparation / hybridization batch the sample was processed in. Within this model gender, tissue bank, and batch were treated as categorical covariates. After fitting each trait to the model the residuals from the model are kept and represent the trait in eQTL analyses. Thus, the expression variation attributable to gender, age, post-mortem interval, tissue bank and hybridization batch are removed prior to eQTL analysis. Histograms showing the proportion of mRNA traits that are potentially impacted by these covariates are shown in Figure 3.5. Covariates for hybridization batch and Brain Bank tissue source had the largest effect, but were also very colinear. Gender and post-mortem interval (PMI) covariates had the smallest effects. It is unknown whether the cause of death was confounding effect within our subject cohort. A cause of death covariate was not included in the covariate adjustment as the information was not complete for 24% of the subjects in the cohort. It has previously been shown that agonal state for conditions such as hypoxia and coma affect gene expression in post-mortem brain more than age, gender, and post-mortem interval, which are covariates that were available for analysis within our subjects (Tomita *et al.* 2004). It has also previously been reported by Li *et al.* that tissue pH (measure of acidity or basicity) from post-mortem human brain is indicative of agonal state. They found that subjects with a prolonged agonal state had a lower pH than subjects with brief deaths. Additionally, they found that samples with lower pH showed an increase in expression of transcription factors and genes encoding stress-response proteins, and decreased expression for energy metabolism and proteolytic activity related genes. The authors suggest that the functional specificity of these gene expression changes reflect a coordinated biological response in living cells and not random RNA degradation (Li *et al.* 2004).

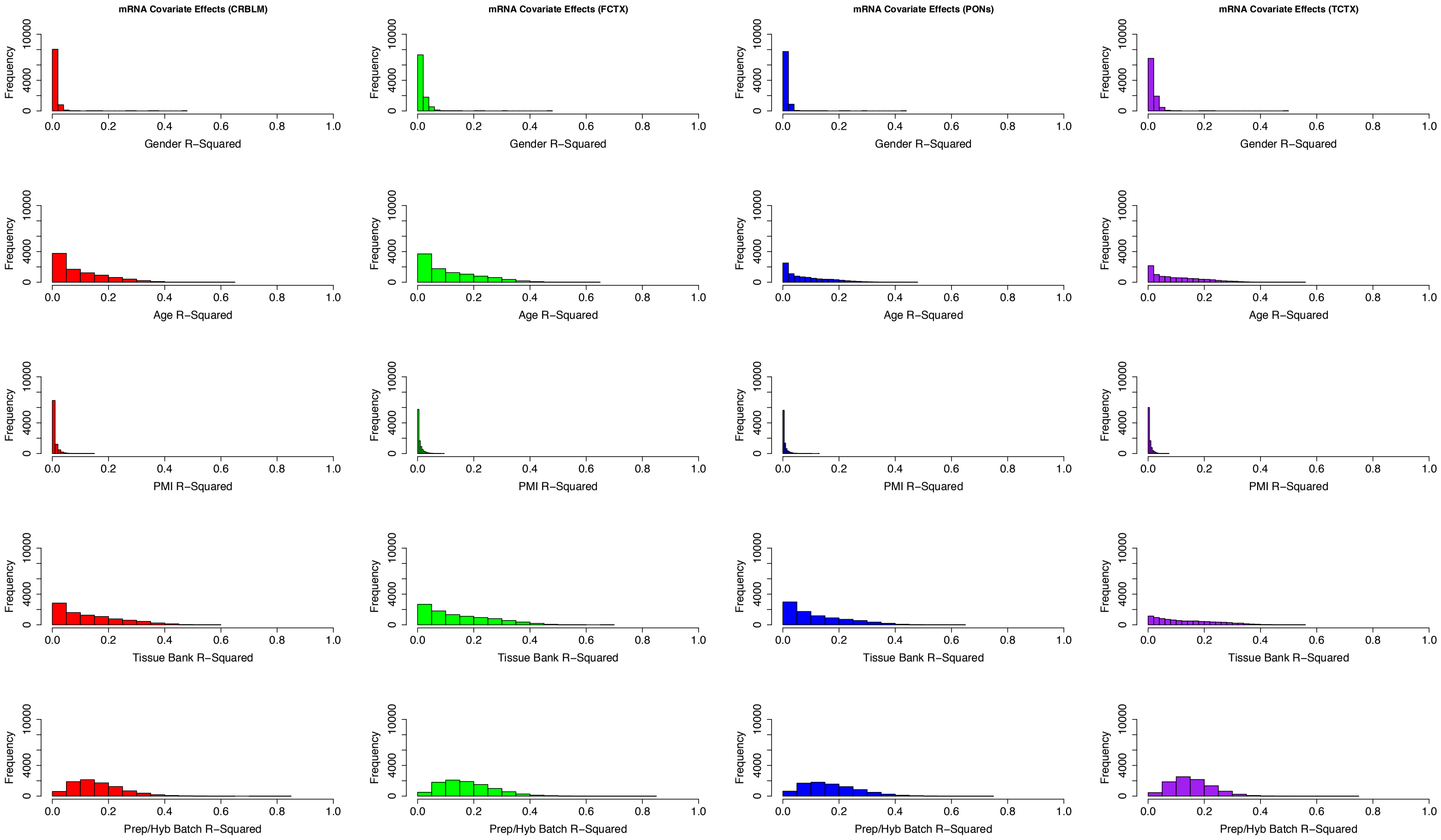


Figure 3.5: Histograms of potential covariate effects on mRNA expression levels, if the data had not been adjusted for these confounds prior to QTL analysis. Each row of sub-plots represents a single covariate and each column represents a brain tissue region: CRBLM (red), FCTX (green), PONS (blue) and TCTX (purple). The histograms show the number of probes (mRNA transcripts) on the y-axis and the R^2 values from the regression of the covariates with each probe. This figure is reproduced from (Gibbs *et al.* 2010).

3.2.3.7: Quantitative trait loci analysis

For each of the four brain regions, a regression analysis was performed on the residuals described in the preceding section for mRNA transcript expression levels. The trait residuals were then used as the quantitative phenotype for that probe in genome-wide association analysis looking for quantitative trait loci. These analyses were performed using the *assoc* function within PLINK, which correlates allele dosage with change in the trait. Each of the four tissue regions was analysed separately, and independent genome-wide association analyses were performed to identify expression quantitative trait loci (eQTL). The PLINK toolset quantitative trait association analysis fits data to the following model:

$$Y=\beta_0 + \beta_1 ADD+ \varepsilon$$

In this model, Y is the quantitative trait and ADD represent genotypes encoded as allele dosage.

3.2.3.8*:* Correction for multiple tests

To correct for the large number of SNPs tested per trait, a genome-wide empirical p-value was computed (North, Curtis and Sham 2002) for the asymptotic p-value for each SNP by using 1,000 permutations of swapping the sample labels of the traits, using the maxT permutation functionality provided within PLINK. A permutation based method using label-swapping of the traits is an appropriate method of test correction (Churchill and Doerge 1994) for these analyses as it is not dependent on these quantitative traits having a normal distribution and also allows the linkage disequilibrium of the genomic regions being tested against the traits to be maintained.

To correct for the number of traits being tested per tissue region, a false discovery rate (FDR) threshold was determined based on the empirical p-values using the *fwer2fdr*, family wise error rate to false discovery rate, function of the *multtest* package in R (Pollard, Dudoit and van der Laan 2005). The *multtest* package is an R library providing multiple methods for use to correct for multiple hypothesis testing. The FDR method finds the largest p-value that is substantially smaller than expected, based on the desired false discovery rate level, where this largest p-value and all p-values less than it can then be considered significant (Benjamini and Hochberg 1995). Empirical p-values were allowed to exceed this threshold if their linkage disequilibrium r^2 was greater than or equal to 0.7 with a SNP with empirical values within the FDR threshold.

3.2.3.9: Replicated eQTL

To identify eQTL that have previously been reported, we considered the results from studies within the Pritchard Lab eQTL Browser (http://eqtl.uchicago.edu/). Of our 282 mRNA transcripts with a *cis*-eQTL detected in at least one tissue, 149 (53%) of these may have also been seen in at least one or more previous eQTL studies. To avoid incomplete annotation information when comparing between studies, the search for overlap in the studies was based on the transcript's gene symbol. Searching based on gene symbol means that the 53% possible overlap with other studies is applicable for stating that a gene with an eQTL was shared between studies, but this does not imply that the same eQTL was seen in multiple studies. For instance different transcripts for the same gene or a different region of genetic variation may have been identified between the studies. For the 53 mRNA transcripts with a *cis*-eQTL that we detected in all four of brain regions, 37 of these have been identified in at least one or more eQTL studies. Overlap with findings in LCL HapMap samples (Stranger *et al.* 2007) is 76 transcripts, in cortical samples (Myers *et al.* 2007a) the overlap is 19 transcripts, and in liver (Schadt *et al.* 2008) the overlap is 108 transcripts. While it may be discouraging that the overlap with findings from the Myers *et al.* cortical samples is not larger, we believe that this is a function of the coverage of genetic variation included in the analysis and changes to the analysis design, such as adjusting for covariates prior to analysis instead of removing all results with a potential covariate effect. The current study includes genotypes from 550K (~1.65 million after imputation and selection) SNPs whereas the Myers *et al.* study included 366,140 SNPs; the Schadt study used 782,476 SNPs; and the Stranger study used ~2.2 million SNPs. So ~53% of our transcripts with a *cis*-eQTL may have also been seen in at least one other study; this does not take into consideration differences in the tissues, assay platforms, analysis methods and annotations used in these studies.

3.2.4: Data Access

The genotype and expression data are publicly available as NCBI dbGaP study accession phs000249.v1.p1 and NCBI GEO series accession GSE15745 (Edgar, Domrachev and Lash 2002; Barrett *et al.* 2007; Mailman *et al.* 2007) (Figures 9.2 and 9.3, *Appendix*).

3.3: Results

3.3.1: mRNA transcript levels differ between brain regions

To assess whether differences in mRNA expression were consistently different between brain regions a global comparison of these measures across tissues was performed. Unsupervised cluster analysis using these data demonstrated that the four brain regions have different expression profiles (Figure 3.6A). Expression pattern differences were most distinct between cerebellum, pons and cerebral cortical tissue, with frontal and temporal cortices clearly separating within the dataset. These data show that mRNA expression levels vary measurably and markedly between brain regions.

The next analysis was limited to the mRNA dataset of those probes where sufficient detection was observed in 95% of samples analysed in each tissue region. This provided data on a total of 10,326 probes against individual mRNA transcripts. The distribution of observed transcript abundance was plotted as a histogram for each tissue (Figure 3.6B). We next compared mRNA expression levels at individual loci directly between each possible pair of tissue regions (Figure 3.6B). In general levels of expression were quite similar between tissues. Measures within frontal and temporal cortices were consistently the most alike whereas cerebellar tissue provided the most distinct profile of the four regions.

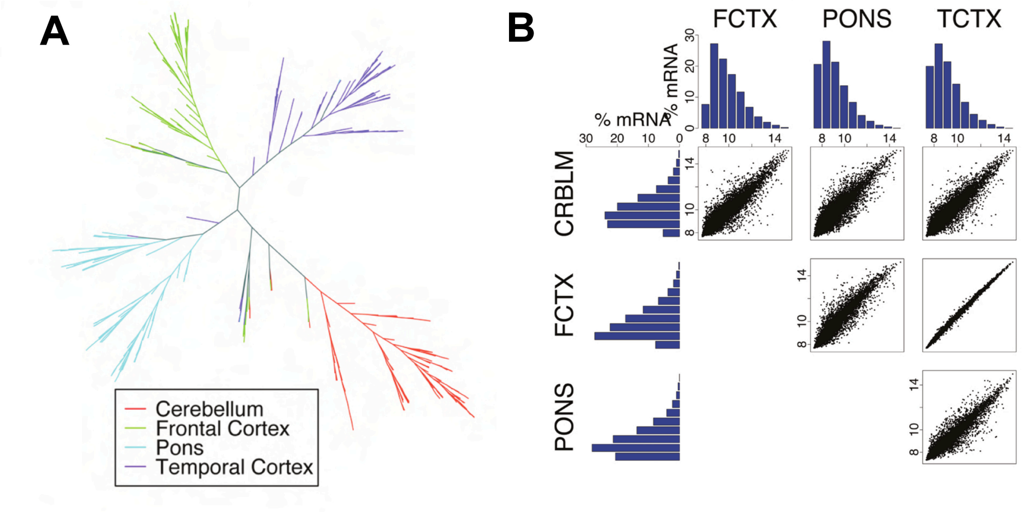


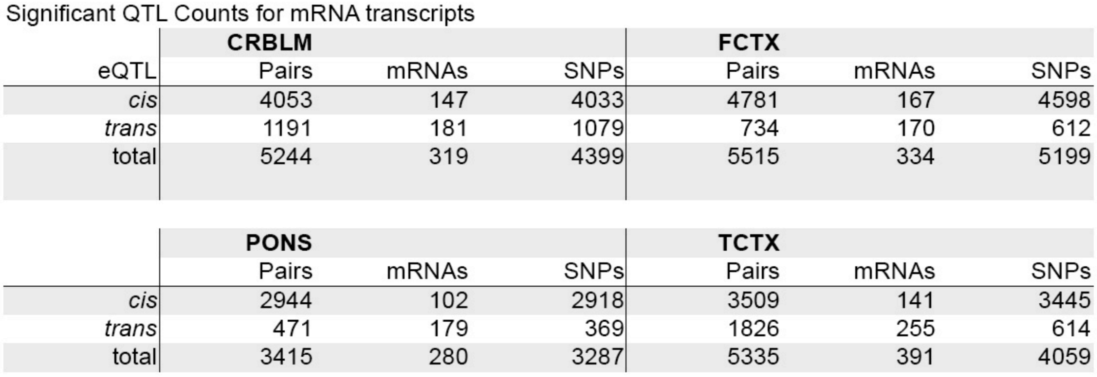
Figure 3.6: Analysis of mRNA measures across four human brain regions. (A), unsupervised cluster analysis of mRNA expression levels. Cluster branches from each brain region are colour coded accordingly and demonstrate consistent separation of cerebellum, pons and cerebral cortical samples; with separation of frontal and temporal cortex samples using mRNA transcript levels. (B) Tissue based pairwise comparisons of mRNA expression. The analyses in these figures used only transcripts that were well detected in each pairing of tissues compared. Histograms show the distribution of mRNA expression levels for each tissue, axes are log$_2$ normalized expression intensities and the % of transcripts at that expression level. Scatter plots are direct comparison of the level of each detected transcript in each tissue pair; axes are the log$_2$ normalized expression intensities. Notably frontal cortex (FCTX) and temporal cortex (TCTX) show the most similar patterns of expression; conversely, all comparisons against cerebellar (CRBLM) tissue show this tissue to have the most distinct patterns for all measures. Figure was adapted from (Gibbs *et al.* 2010).

3.3.2: Genotype influences mRNA expression

A primary aim underlying these experiments was to examine the extent of genetic control of expression within brain tissues. To investigate this process, we undertook a series of eQTL analyses. From the 537,411 genotyped SNPs that passed quality control filtering we imputed 2,545,178 SNPs. After additional quality and analysis specifications filtering 1,629,853 SNPs (average) were used for analysis. With these data, we then performed regression of allele dosage against each measure using expression of mRNA transcripts as the dependent variable and genotype as the independent variable and treating each tissue as a separate analysis. We corrected for number of tests per trait by permutation and for the number of traits using an

FDR-like measure. This yielded a necessarily conservative threshold for significance (Churchill and Doerge, 1994). Prior to analysis, each trait was adjusted using available biological and methodological covariates in an attempt to reduce the influence of systematic confounding effects. *Post hoc* we annotated significant eQTL as *cis* if the SNP was within 1 megabase (Mb) of the transcript being tested; all other SNP-dependent variable tests were designated as *trans*. Notably, because the designation of *cis*- and *trans*-eQTL tests was performed *post hoc*, there was no distinction in terms of level of statistical correction between these groups.

There were a large number of significant correlations detected between genetic variation and variation in the expression of mRNA transcripts, with significant eQTL detected in each of the four brain regions; ranging from 280 (3.2%) in the pons to 391 (4.2%) in the temporal cortex (Table 3.3). These eQTL accounted for between 18% and 77%, of corrected expression levels, of associated transcripts between individuals. On average the eQTL accounted for 28% of the expression variation of the associated transcript (*cis*-eQTL mean is 30% and *trans*-eQTL mean is 22%).



Significant QTL Counts for mRNA transcripts

	CRBLM			**FCTX**		
eQTL	Pairs	mRNAs	SNPs	Pairs	mRNAs	SNPs
cis	4053	147	4033	4781	167	4598
trans	1191	181	1079	734	170	612
total	5244	319	4399	5515	334	5199

	PONS			**TCTX**		
	Pairs	mRNAs	SNPs	Pairs	mRNAs	SNPs
cis	2944	102	2918	3509	141	3445
trans	471	179	369	1826	255	614
total	3415	280	3287	5335	391	4059

Table 3.3: Counts of significant eQTL by brain region; cerebellum (CRBLM), cerebral frontal cortex (FCTX), caudal pons (PONS), and cerebral temporal cortex (TCTX). Counts include total, *cis* and *trans* numbers for correlated pairs of mRNA transcripts and SNPs as well as unique number of mRNA transcript probes and SNPs. Table adapted from (Gibbs *et al.* 2010).

To assess the enrichment of detected *cis*-eQTL relative to those in *trans* we calculated the number of observed and possible *cis*- and *trans*-eQTL for mRNA expression levels. Based on a definition of *cis* at 1Mb, these data showed an enrichment of *cis*-eQTL relative to *trans*. The peak enrichment of *cis*-eQTL was observed at ~68 Kb for mRNA transcripts.

The abundance of *cis*-eQTL for mRNA expression prompted us to examine the distribution of *cis*-eQTL (Figure 3.7A-D). This revealed that both the number of significant eQTL and the strength of association between SNP and mRNA expression level were inversely correlated with physical distance between the genetic variation and the transcript start site (TSS) of the mRNA transcript in question. The average distance, for *cis*-eQTL, between the SNP and the transcript is 70 Kb. The most significant *cis*-eQTL tended to be present in all four tissues tested (Figure 3.7E). Of the transcripts with *cis*-eQTL that were significantly detected in at least one brain region, 53% have been previously reported, when intersecting by gene names instead of individual gene mRNA transcripts or assay probe IDs. This number increased to 70% when analysis is limited to those *cis*-eQTL detected in all four tissues (Myers et al., 2007a; Schadt et al., 2008; Stranger et al., 2007). Table 3.4 lists the top ten *cis*-eQTL transcripts found in this study.

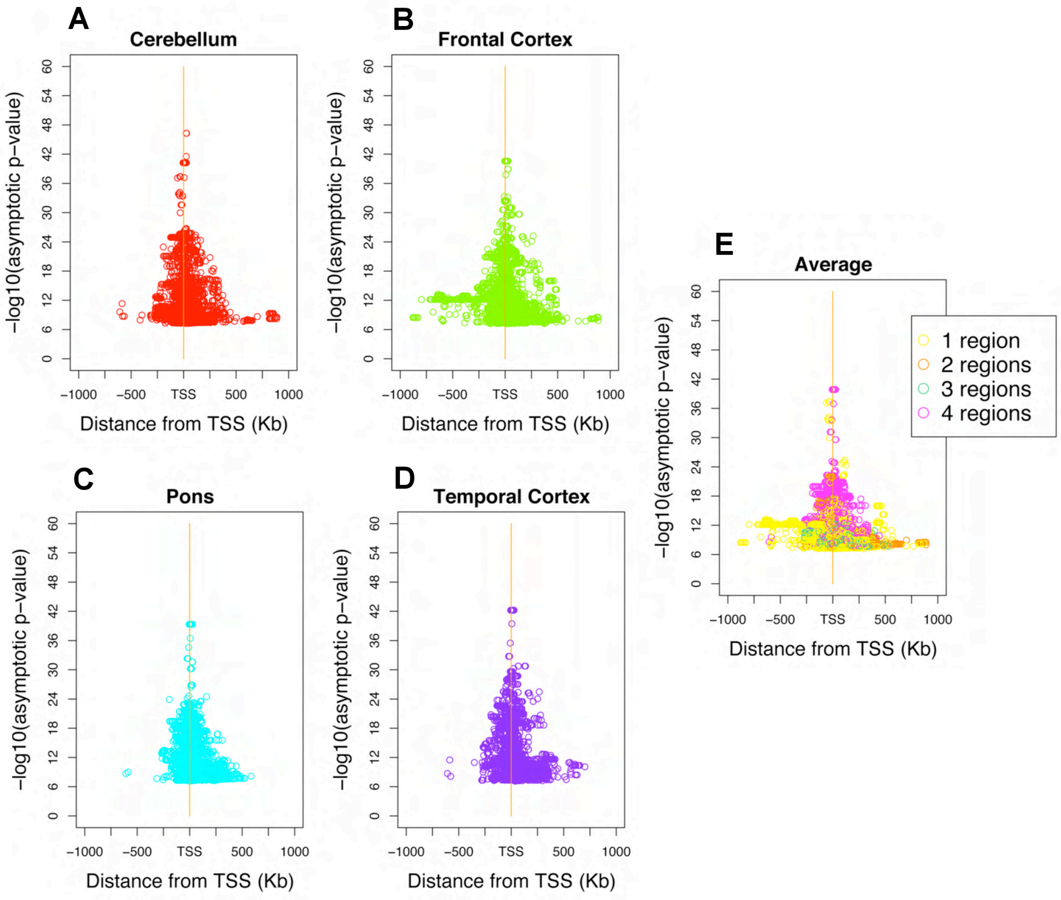


Figure 3.7: (A-D) Significant *cis*-eQTL p-values per tissue region relative to the transcription start site (TSS). (E) Average p-values for *cis*-eQTL across regions. The more significant *cis*-QTL tended to be both closer to the transcription start site and common across tissue regions tested. Figure adapted from (Gibbs *et al.* 2010).

3.3.3: Detected eQTL are consistent across brain regions

In order to compare detected eQTL between tissues, we selected every SNP-transcript pair that passed the defined threshold for significance in at least one tissue. We then compared R^2 values for each eQTL across tissues using ternary plots (Figure 3.8). The majority of significant *cis*-eQTL were shared across the four brain regions, while *trans* signals shared across tissues are almost complete absent.

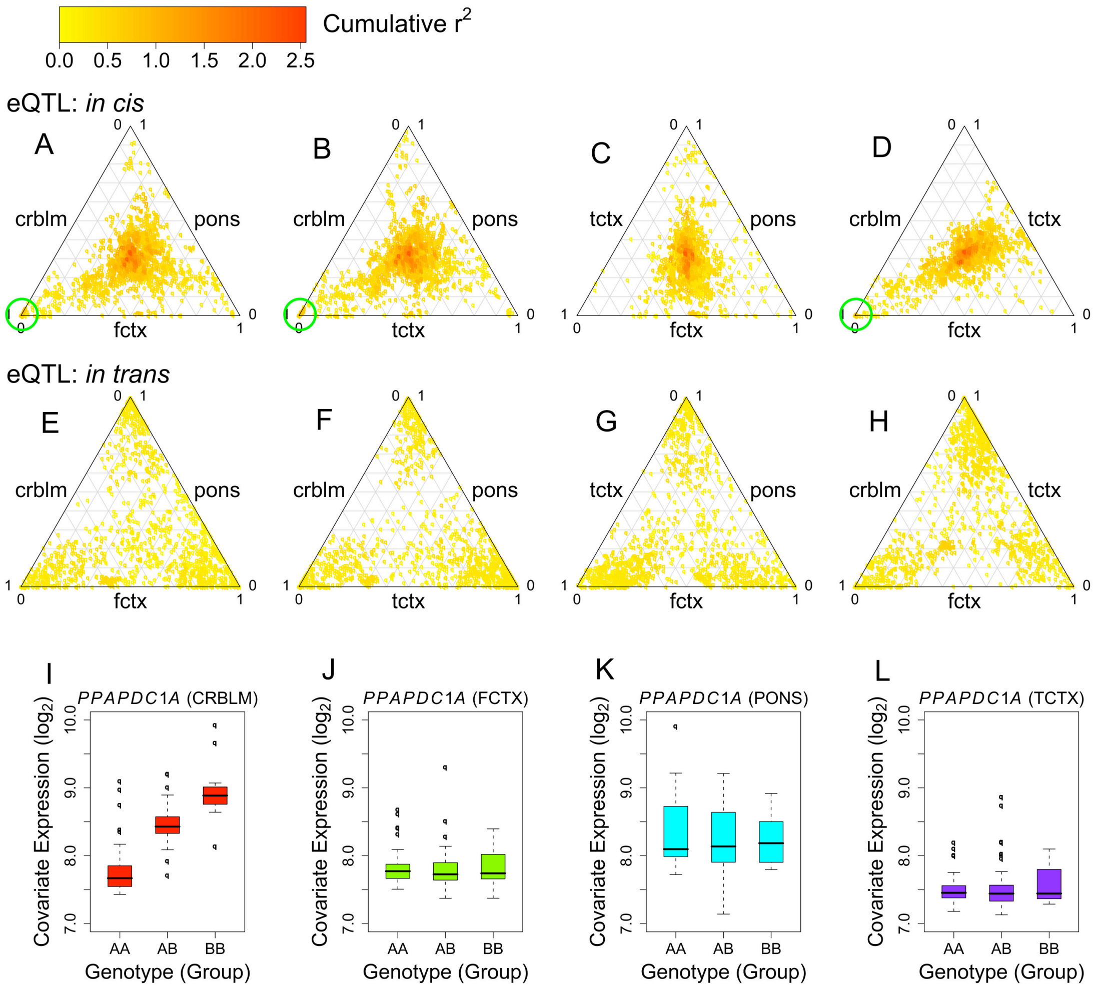


Figure 3.8: Comparison of eQTL across tissue regions. Any eQTL that passed our threshold for significance in at least one tissue was included in the Ternary plots. The colour of the points in the ternary plots reflects the cumulative R^2 value, from the correlations, for all tissues tested within each plot. Points toward the centre indicate an equal R^2 value across the three regions under investigation. Points toward the corner of a plot indicate a high R^2 in one of the three tissues; points toward the edges of the plot indicate a high R^2 in two of the three tissues. (A-H) Comparing eQTL in every three-way combination of the four tissues for *cis* (A-D) and *trans* (E-H). Notably the cumulative R^2 is generally higher for *cis* compared to *trans* loci. Green circles highlight a cluster of relatively high cumulative R^2 values driven primarily by the observed R^2 within cerebellar tissue. These points were revealed to be a *cis*-eQTL involving 20 SNPs and two neighbouring transcripts, *PPAPDC1A* and *C10orf85*. (Q-T) Boxplots show expression level plotted against genotype for one of these eQTL SNP-transcript pairs (SNP rs2182513 and *PPAPDC1A*) and illustrates that this is a tissue-specific eQTL limited to the cerebellum. *C10orf85* follows the same pattern with an eQTL present in cerebellum but not in the other three tissue. Figure adapted from (Gibbs *et al.* 2010).

These plots illustrate that the great majority of *cis*-eQTL with strong effect sizes were consistent across these tissues. We found tissue-specific eQTL to be less common for large effect sizes, but there were observable events.

However, for individual significant SNP and trait correlations, tissue-specific correlations were detectable (Figure 3.9). For example, while strong eQTL were found for churchill domain containing 1 (*CHURC1*) in all tissues and as reported previously in liver (Schadt *et al.* 2008), several *cis*-eQTL for phosphatidic acid phosphatase type 2 domain containing 1A (*PPAPDC1A*) were restricted to the cerebellum, despite reliable detection of the transcript in all four brain regions (Figures 3.10 and 3.11).

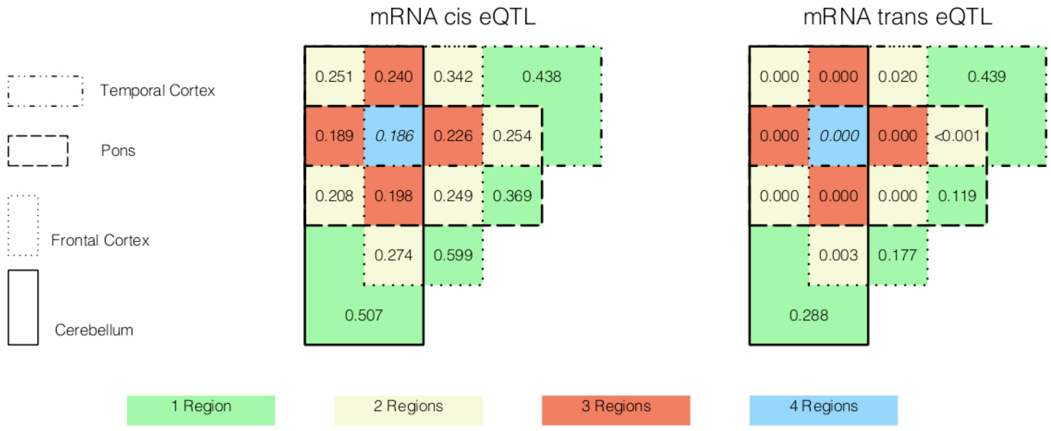


Figure 3.9: Stylized Venn diagrams showing the frequency overlap of the number of significant SNP and transcript correlations that are detected in the four brain tissues. The rectangles with different orientations and border, shown on the left legend represent the different tissue and the different squares represent overlapping frequencies between different tissues. The colour coded squares represent the number of tissues overlapping, where the central blue square in each Venn diagram represent the proportion of significant correlations detected in all four tissues. Hence, the blue square in the *cis* diagram indicates that 18.6% of the significant SNP and transcript correlations detected were present in all four of the tissue regions. Whereas the green squares show the proportion of significant SNP and transcript correlations only detected in that region of the brain.

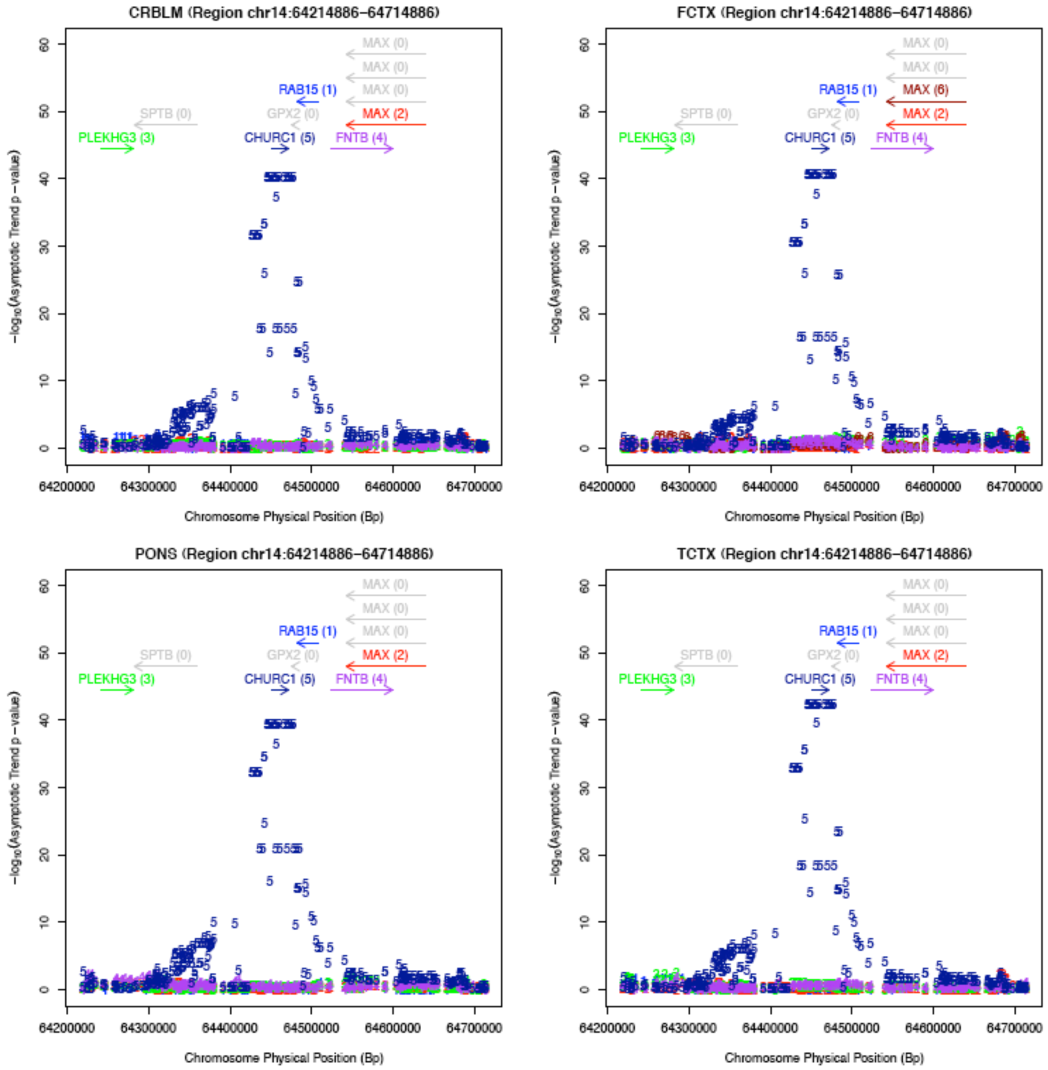


Figure 3.10: Manhattan plots, one for each brain tissue region, show the p-values from the correlations between SNPs and mRNA transcripts in a 500Kb region centred on *CHURC1*. A *cis*-eQTL for CHURC1 in this genomic region is present in all four brain tissues. A *cis*-QTL for *CHURC1* has also been reported within liver (Schadt *et al.* 2008). Within each plot the X-axis is the physical position along this region of the chromosome and the Y-axis is the –log$_{10}$(asymptotic p-values) for the correlations. The p-values are colour coded and numbered to match the annotated transcripts labelled in the top portion of the plots. Thus in cerebellum (CRBLM), cerebral frontal cortex (FCTX), caudal (PONS), and cerebral temporal cortex (TCTX) the dark blue '5's are p-values, for individual SNPs, correlated with expression levels of *CHURC1*. mRNA transcript annotations shown in grey are those where a probe is present on the expression platform but not detected in 95% of the tissues. Figure reproduced from (Gibbs *et al.* 2010).

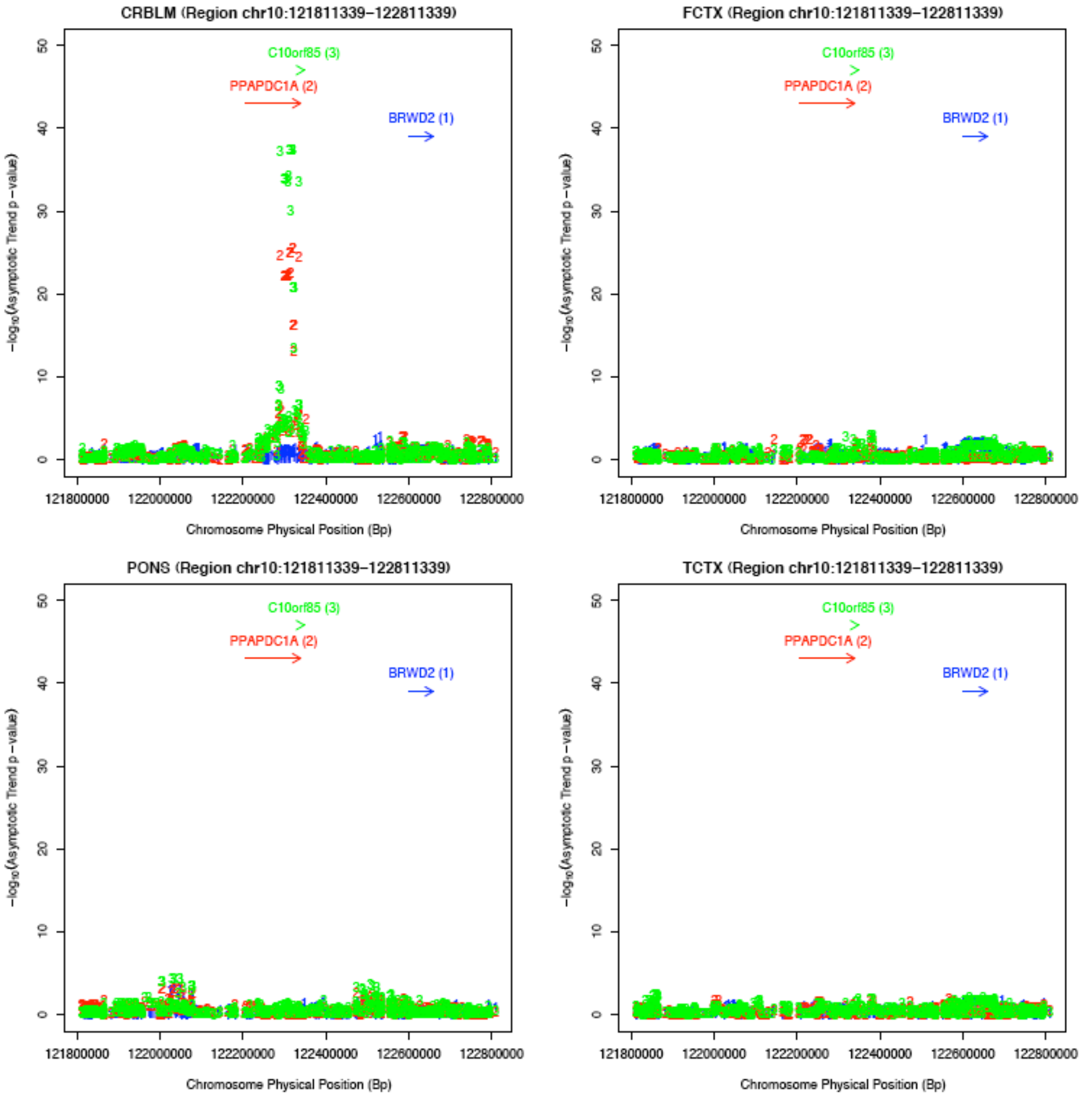


Figure 3.11: Manhattan plots show an example of a *cis*-eQTL for an mRNA that appears to be tissue-specific. The plots, one for each brain tissue region, show the p-values of correlations between SNPs and mRNA transcripts in a 1Mb region centred on *PPAPDC1*. Within each plot the X-axis is the physical position along this region of the chromosome and the Y-axis is the –log$_{10}$(asymptotic p-values) for the SNP and transcript correlations. The p-values are colour coded and numbered to match the annotated transcripts labelled in the top portion of the plots. Thus in cerebellum (CRBLM) the red '2's are p-values for SNPs correlated with the expression levels of *PPAPDC1A*. Also present at this same genomic locus is another tissue specific eQTL for the mRNA transcript *C10orf85*, shown as green '3's. Figure reproduced from (Gibbs *et al.* 2010).

IlmnID	Symbol	Chr	Strand	TSS	SNP	SNP_Loc	MAF	Dist_To_TSS	p-value (best)	BB_Effect_Direction	Sig_Tissues
ILMN_1719064	KCTD10	12	-	108399537	rs6663	KCTD10 UTR3'	0.24	28551	5.02E-047	increase	crblm, fctx, pons, tctx
ILMN_1732815	CHURC1	14	+	64450892	rs10151701	CHURC1 upstream	0.20	-2690	5.78E-043	increase	crblm, fctx, pons, tctx
ILMN_1665207	C10orf85	10	+	122347711	rs2420726	PPAPDC1A intron	0.40	-33074	3.97E-038	increase	crblm
ILMN_1723116	AMFR	16	-	55016945	rs2550309	AMFR intron	0.41	24702	6.60E-034	decrease	crblm, fctx, pons, tctx
ILMN_1711245	SCG3	15	+	49760841	rs2623258	TMOD2 UTR3'	0.36	128045	1.70E-031	increase	crblm, fctx, pons, tctx
ILMN_1683279	PEX6	6	-	43054865	rs2296805	GNMT intron	0.40	18129	5.72E-031	decrease	crblm, fctx, pons, tctx
ILMN_1811301	INPP5E	9	-	138454076	rs8413	INPP5E UTR3'	0.46	10944	8.38E-030	decrease	crblm, fctx, pons, tctx
ILMN_1774949	DSCR5	21	-	37367327	rs3787780	TTC3 intron	0.50	-17157	4.41E-029	increase	pons, tctx
ILMN_1772459	RPS23	5	-	81609990	rs226208	ATP6AP1L intron	0.34	-29210	4.81E-029	decrease	crblm, fctx, pons, tctx
ILMN_1710752	NAPRT1	8	-	144731635	rs10282929	TIGD5 missense	0.20	-21285	4.01E-028	decrease	crblm, fctx, pons, tctx

Symbol	Name
KCTD10	potassium channel tetramerisation domain containing 10 (KCTD10), mRNA.
CHURC1	churchill domain containing 1 (CHURC1), mRNA.
C10orf85	chromosome 10 open reading frame 85 (C10orf85), mRNA.
AMFR	autocrine motility factor receptor (AMFR), transcript variant 2, mRNA.
SCG3	secretogranin III (SCG3), mRNA.
PEX6	peroxisomal biogenesis factor 6 (PEX6), mRNA.
INPP5E	inositol polyphosphate-5-phosphatase, 72 kDa (INPP5E), mRNA.
DSCR5	Down syndrome critical region gene 5 (DSCR5), transcript variant 2, mRNA.
RPS23	ribosomal protein S23 (RPS23), mRNA.
NAPRT1	nicotinate phosphoribosyltransferase domain containing 1 (NAPRT1), mRNA.

Table 3.4: Table, top portion showing the top 10 transcripts based on significant *cis*-eQTL from all tissues, only the best SNP p-value is shown. Columns included are: the Illumina Probe ID, Hugo Gene Symbol, Chromosome the gene is located on, strand, chromosomal position of the transcript start site, SNP ID for best correlated SNP, location of the correlated SNP (relative to gene), minor allele frequency for the SNP, distance between the SNP and the transcription start site oriented by gene strand, p-value from the linear regression (from most significant tissue), the direction of the effect found in the regression where increase indicates an increase in transcript abundance with minor allele's dosage and which tissue regions this SNP and transcript pair were significant in. Tissues are cerebellum (crblm), cerebral frontal cortex (fctx), caudal pons (pons), and cerebral temporal cortex (tctx). Bottom portion shows gene names.

3.4: Discussion

The work I have described here and the public release of the data resulting from this effort, aim to facilitate an understanding of the initial consequences of common genetic variation on gene expression in brain. These data show clearly that, as expected, patterns of expression are measurably different across brain tissues. eQTL analyses reveal an abundance of significant eQTL that are predominantly *cis* in nature. Previous eQTL results, from the human cerebral cortex (Myers *et al.* 2007a), as well as eQTL results from HapMap lymphoblastoid cell lines (LCL) (Veyrieras et al., 2008) have suggested that SNPs proximal to genes, including SNPs upstream of the transcription start site (TSS), within the gene, and downstream of the transcription end site (TES) have a greater influence on gene expression than those further away. This is presumably because genetic variation around promoter elements, splice sites, and 3' UTRs affects transcription, splicing, and mRNA stability (Kwan *et al.* 2008) that results in an enrichment of *cis*- over *trans*-eQTL. The power to detect the signal in a local *cis*-variant is likely a direct effect whereas *trans,* if detectable, are likely indirect and underpowered for accurate detection in series of these sizes (Pastinen, Ge and Hudson 2006). As the effect from *trans*-variation is likely to occur through an indirect effect, this means there is likely another unaccounted-for intermediate effect in between the genetic variation and the trait being measured, which decreases the power of detection (Schliekelman 2008).

It is notable, particularly given the systematic differences in expression patterns among the four tissues, that many of the significant eQTL with the strongest signal were consistent across brain tissues. Tissue-specific eQTL are also observable in the current data set, suggesting that there are genetic effects on expression that are dependent on the tissue type used irrespective of expression levels of the mRNA. This emphasizes the importance of exploring eQTL in the context of a relevant tissue. The ability to analyse and compare both distinct and similar brain tissue regions, from the same subjects, in this study allowed for a more thorough comparison of eQTL in brain tissues. Whereas in the previous study, described in Chapter 2 of a single sample set based on a mix of cortical tissues, the only whole-genome eQTL results available for comparison were from HapMap lymphoblastoid cell lines (LCL), at that time. This study and data created the initial cohort that has since become known as the North American Brain Expression Consortium (NABEC).

The use of LCLs in the discovery and understanding of eQTL has been quite useful; however, there may be potential problems with using LCLs as a proxy of other tissues and cell types in eQTL analysis. When using LCLs there may be potential artefacts in expression levels. Expression levels may correlate with Epstein–Barr virus (EBV) load and growth rates, some genes may exhibit monoallelic expression and LCLs are generated from a blood cell lineage (Choy *et al.* 2008; Plagnol *et al.* 2008). In a 2009, a report from Dimas *et al.* suggested that tissue specificity for eQTL is widespread with only 20 to 30% of eQTL replicating across tissues (Dimas *et al.* 2009). An interpretation of the Dima *et al.* results, suggests that eQTL shared across tissues are closer to the TSS and have a larger effect size while tissue-specific eQTL are more widely dispersed in *cis* and have a smaller effect sizes (Figure 3.12) (Dimas *et al.* 2009; Montgomery and Dermitzakis 2011). More generally it has also been shown that eQTL shared across tissues typically falls somewhere between 40% and 80% depending on tissue similarity and analysis methods (Heinzen *et al.* 2008; Ding *et al.* 2010). In 2008, Emilsson *et al.* published a large study of expression and obesity traits in an Icelandic population from blood and adipose tissue. Their results showed correlations between obesity traits and gene expression in both tissues, but the results were much stronger in adipose than in blood, 50% and 10% respectively. The stronger correlations in adipose tissue with obesity traits were likely expected as this tissue is primarily composed of adipocytes and obesity can be characterized by the increase in size of adipocytes. They also identified eQTL based on segregation and linkage analysis. Of the eQTL they identified, 50% were shared by both tissues and were strongly *cis* rather than *trans* (Emilsson *et al.* 2008). In 2011, Innocenti *et al.* published a study considering the reproducibility of eQTL studies. This study based on human liver from 3 cohorts found that ~30% of SNP-expression correlations failed to replicate. They suggest that other factors associated with the tissue were confounders for replication including: drug exposure, clinical descriptors, and tissue ascertainment. They reiterated that the array's expression probe design can be a confounder, if polymorphisms within the assay probes are not accounted for. They found that controlling for these possible confounds increased the replication rate. They also found that the most replicable eQTL variants were those enriched at gene starts and stops. For 14 genes, they did additional validation and fine mapping confirming haplotype-specific in-vitro expression differences. Overall, their study potentially validated hundreds of eQTL in human liver. They suggest that many of these eQTL may be informative in indentifying and functionally characterizing the genetic contribrution to diseases and complex traits (Innocenti *et al.* 2011). Innocenti *et al.* reference two examples of previously characterized mechanistic links to disease and complex traits that also intersected with replicated eQTL within their study: warfarin drug response and vitamin K epoxide reductase complex subunit 1 (*VKORC1*) expression (Rieder *et al.* 2005), and sortilin 1 (*SORT1*) expression correlations with lipid levels and heart disease (Kathiresan *et al.* 2008). Additionally, Innocenti *et al.* suggest that their results may also support the hypotheses that NOD2 expression levels are linked to leprosy risk (Zhang *et al.* 2009), and that C2orf43 gene expression levels are linked to prostate cancer risk (Takata *et al.* 2010).

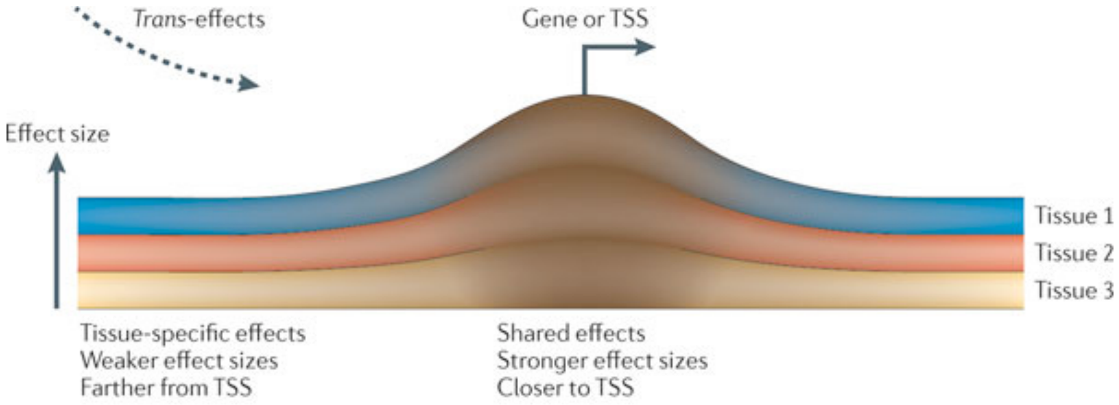


Figure 3.12: Cartoon from Montgomery and Dermitzakis review showing larger and shared effects for eQTL are closer to the transcription start site (TSS) and that weaker and tissue-specific effects tend to be further from the TSS and that *trans* effects may help to understand gene and regulatory networks. Figure reproduced from (Montgomery and Dermitzakis 2011).

Other studies, performed during the late 2000s, also began to consider other forms of expression or regulatory traits and their association with genetic variation. In 2009, Wei Zhang *et al.* published a study examining alternative splicing (AS). This study made use of Affymetrix exon expression arrays in

176 LCLs from HapMap CEU and Yoruban subjects. They identified local and distant genetic variants associated with transcript isoform variation between the two populations and found a substantial fraction (8%) of transcripts with isoform variation were associated with genetic variation (Zhang *et al.* 2009b). In 2009, Kun Zhang *et al.* published a study of allele-specific expression (ASE), which was based on four cell lines from two subjects for the Personal Genome Project. They found that between 11% and 22% of heterozygous mRNA SNPs showed allele-specific expression per cell line, and that between 4% and 8% of these were tissue-specific. When analyses were expanded to include two pairs of siblings, they found that ASE was more similar among the siblings than in the unrelated subjects. In their results, genetic variation accounted for more variation in allelic ratios of expression than tissue type or growth conditions. Based on expression of alleles by strand, they suggest that allelic ratios are primarily *cis*-regulated on the sense strand (Zhang *et al.* 2009a). In 2010, Dandan Zhang *et al.* published a study of CpG site methylation QTL (methQTL) in human cerebellum from ~150 subjects. They found that 9% of CpG sites that displayed large variation between subjects were also correlated with *cis* genetic variants. They also found that ~1% of methQTL were also eQTL, where both DNA methylation and gene expression were also correlated (Zhang *et al.* 2010). It should be noted that DNA methylation at CpG sites has also been generated in all four tissues for the NABEC subjects described in this chapter. I have not included a description of this data, analysis, or results because this data is part of another student's thesis on DNA methylation in human tissues. I will briefly describe the methylation results here, as they are similar to those found by the Zhang *et al.* study, which were also done using human cerebellar tissue. I identified

methQTL in the NABEC tissues following almost the exact same analysis as used for the identification of eQTL in this chapter. For methQTL, we found between 4% and 5% of CpG sites methylation levels were significantly correlated with genetic variation. The genetic variants, for the methQTL, accounted for between 18% and 88% of the CpG methylation level, at individual CpG sites. I then performed an analysis of the intersection of eQTL and methQTL to determine if these traits were independent of each other. I found that while 4.8% of the genetic loci intersected (2.6% of methQTL and 8.2% of eQTL) their effects on gene expression levels were independent.

While we were able to detect tissue-specific eQTL in the four brain tissues, the majority of the eQTL detected appear to have a high degree of sharing between the tissues. Even though we selected tissues that are both diverse and similar, where cerebellum was the most distinct and the cerebral cortical regions were similar, these tissues are heterogeneous in cellular composition. This heterogeneous cellular composition makes it difficult to determine if the degree of eQTL sharing we see between brain regions is a result of eQTL being similar between different types of neurons or different types of glia, or do they appear shared because specific neuronal and glia cell types have cell-specific eQTL but the presence of enough of these cells in bulk tissue regions is allowing us to detect the majority of the cell-specific signal but without a way to deconvolute the signal. The possibility may also exist that with glia and granule cells being a common class of cell types in the brain that the eQTL we are detecting represent shared eQTL among glial cell types and shared eQTL among granule cell types. With this in mind we undertook the next study in my thesis, identifying eQTL in samples enriched for a specific neuronal cell type.

4: Identifying eQTL in a specific Human Neuronal cell type

Statement of Contribution to this Research:

I was involved in the conception and design of this study, including choice of genotyping platform, expression platform, and selection of tissue. I performed data quality control, data analysis, and data interpretation. Cookson MR and Singleton AB were also involved in the conception, design, choice of platform, and tissue and cell type selection. I was not involved in the collection of the tissue samples or the generation of the genotype and mRNA expression data. Coordination and collection of the tissue was performed by Traynor BJ. Genotyping of the samples was coordinated or performed by: Hernandez DG, Traynor BJ, Arepalli S, Rafferty IP, and Lai SL. Laser capture microdissection was performed by: Kumar A, Beilina A, and Kumaran R. Generation of mRNA expression data was performed by: Dillman A, Kumaran R, and Kumar A.

4.1: Introduction

As shown in the previous two chapters, genetic diversity contributes to variation in gene expression in human brain. However, one difficulty with examining gene expression in tissues with heterogeneous cellular composition is that different cells have different gene expression patterns. For example, in the brain there are many types of neurons as well as different glia and other cell types. In my previously described eQTL studies, the tissue samples were based on human brain regions that contain a heterogeneous cell mix. The heterogeneity of the tissues adds a degree of ambiguity in characterizing the eQTL identified. In 2011, Price *et al.* published a study of *cis* and *trans* heritability of gene expression based on 722 familial Icelandic subjects from blood and adipose tissue. They found that the heritability of gene expression for 37% of transcripts in blood and 24% transcripts in adipose was attributable to *cis*-regulatory variants. They also found that gene expression correlations between the tissues were also due to heritability of *cis*-regulatory loci, but this was not the case for *trans*-regulation for the two tissues. They repeated a similar analysis in unrelated individuals and found similar results. They suggest this means that tissues with heterogeneous cell types will be more effected by *cis*-regulation than tissues of homogenous cell types (Price *et al.* 2011). In order to understand whether changes in the cellular composition of the brain influenced the previous observations, we repeated the analyses in neurons isolated by laser capture microdissection (LCM). In 1999, Luo *et al.* published a study using LCM and microarrays to examine differential expression between adjacent large and small neurons from dorsal root ganglia (DRG) in rats. They found that they could cleanly capture adjacent large and small neurons and identified 40 transcripts differentially expressed, where 26 were preferentially expressed in small neurons and 14 in large neurons (Luo *et al.* 1999). In 2012, Friedrich *et al.* published a study of gene expression in Purkinje cells from mice examining Polyglutamine (PolyQ) diseases: spinocerebellar ataxia type 7 (SCA7) and Huntington's disease (HD). These diseases share a cerebellar degenerative phenotype of progressive selective cell loss and formation of protein aggregates. In this study, they used laser capture microdissection (LCM) to compare gene expression in Purkinje cells from transgenic PolyQ mouse models using microarrays. They used real-time PCR for validation of their results. They found a similar reduced expression of mRNA in their mouse models where decreases in aldolase C and phospholipase C beta3 increased the vulnerability of Purkinje cells to excitotoxic events. Additionally, they found that the decrease in mRNA expression, in their mouse models, was facilitated by the *Pcp2* promoter (Friedrich *et al.* 2012). The Purkinje cell is also the specific neuronal cell chosen for our study. This cell type was chosen because Purkinje cells are a large and distinctive neuronal cell type found in the cerebellar cortex, a tissue already collected for the NABEC cohort. Additionally, while the dendrites of Purkinje cells branch very profusely they do so in a flattened almost two-dimensional layer. These aspects of the Purkinje cells make them easily identifiable, using simple rapid staining in frozen sections, and their flattened structure makes them amenable for capture with LCM. Thus we performed a *cis*-eQTL analysis in a subset (N = 85) of the North American Brain Expression Consortium (NABEC) subjects, where laser capture microdissection was used to isolate Purkinje neurons from the cerebellum. Additionally, for this subset of NABEC subjects, eQTL analysis was repeated using data from the bulk cerebellum and cerebral frontal cortex samples.

4.2: Materials and Methods

4.2.1: Subjects

(Coordination and collection of the tissue was performed by Traynor BJ.)

This study was composed of 100 neurologically normal Caucasian subjects from the United States; these subjects are from the North American Brain Expression Consortium (NABEC). Tissue from the cerebellum and cerebral frontal cortex were previously obtained for all subjects. After filtering subjects based on quality control steps, based on data from all assays for the cell and tissue groups, 85 subjects remained and were used in the analysis.

Frozen tissue samples of the cerebral frontal cortex and cerebellum were obtained from each of 388 subjects who had donated their brains for medical research, making up the current NABEC cohort. Of these, 100 subjects were chosen from the University of Maryland Brain Bank and Baltimore Longitudinal Study of Aging collections within the NABEC cohort. All individuals were of non-Hispanic Caucasian ethnicity, none of the subjects had a clinical history of neurological or cerebrovascular disease, or a diagnosis of cognitive impairment during life. The average age at time of death was 38.6 years of age (range, 16 – 101 years). Of the 100 subjects, 33% were female. The average post-mortem interval was 14.5 hours (range, 4 – 28 hours).

4.2.2: Assays

4.2.2.1: SNP Genotyping

(Genotyping of the samples was coordinated or performed by: Hernandez DG, Traynor BJ, Arepalli S, Rafferty IP, and Lai SL.)

Genomic DNA extraction for genotyping was performed using the DNeasy Blood and Tissue Kit as per the manufacturer's instructions (Qiagen Inc., Valencia, California, USA). DNA concentration was determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE), and DNA extraction was repeated using a new tissue aliquot for samples with DNA concentration less than 50ng/ul, or where the 260nm/280nm wavelength absorption ratio was less than 1.7, indicative of significant protein contamination of the DNA sample. SNP genotyping was performed using DNA extracted from cerebellar tissue for each subject using Infinium HumanHap550 version 3 BeadChips (Illumina Inc., San Diego, California, USA) according to the manufacturer's instructions. Genotype data was analysed using the Genotyping Analysis Module 3.2.32 within the BeadStudio software version 3.1.4 (Illumina Inc.).

4.2.2.2: mRNA Expression

(Laser capture microdissection was performed by Kumar A, Beilina A, and Kumaran R. Generation of mRNA expression data was performed by Dillman A, Kumaran R, and Kumar A (Kumar *et al.* 2013).)

Laser capture microdissection (LCM) is a sample extraction technique that allows for specific sub-selection of a tissue sample to be dissected out (Figure 4.1). Typically, this sub-selection is for extracting a particular cell or group of cells from a tissue sample in order to obtain a pure population of cells to assay. Emmert-Buck et al. described a LCM method*,* in 1996. Their method starts with a transparent film being placed over a tissue sample. The sample is then microscopically viewed and the region or cells of interest are then selected (Figure 4.2A). An infrared laser then applies a short duration focus pulse that thermally adheres the transfer film to the selected region of the tissue (Figure 4.2B). The film, with the selected tissue, can then be extracted from the larger tissue section (Figure 4.2C) (Emmert-Buck *et al.* 1996). This LCM method offered advantages of other microdissection techniques of the time. These advantages included: no manual microdissection, one step transfers, the transferred tissue on film retains original morphology, laser focus size allows for targeting of single cell. A possible drawback is that the localized heating of the film has some direct absorption by the underlying tissue, as enough energy must be applied to the film to raise the temperature to fusion point (Emmert-Buck *et al.* 1996).

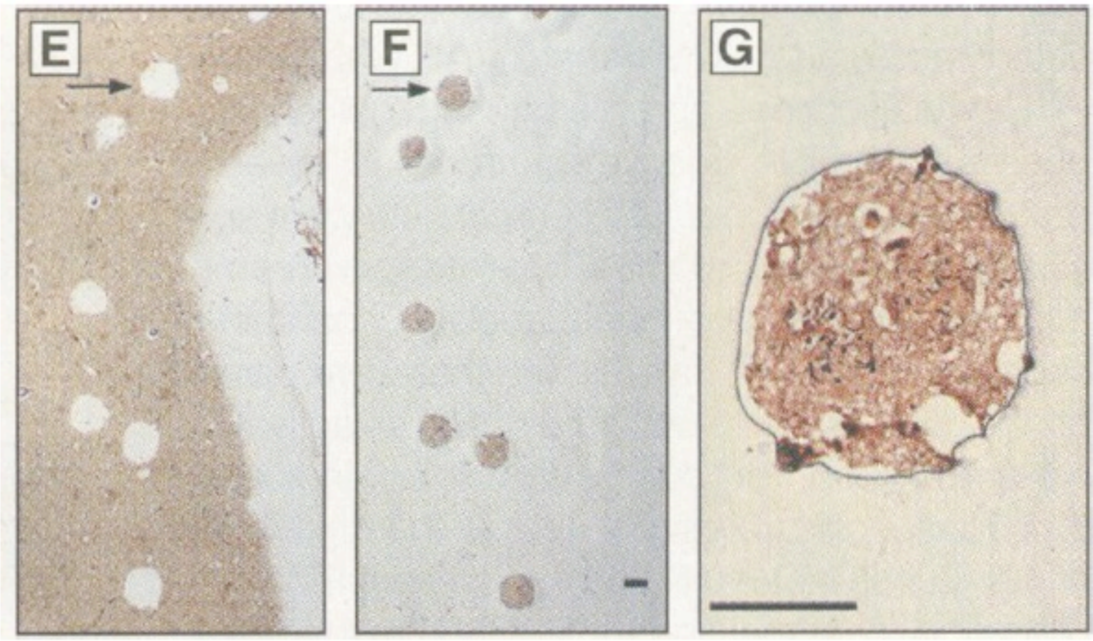


Figure 4.1: Example of LCM extraction of Alzheimer's disease plaques from a section of frontal cortex. E) Image of the frontal cortex section with the LCM selected regions already removed; the white circular shapes one of which is indicated by the black arrow. F) Image of the extracted regions, selected from the larger tissue sample, on the transfer film. The extracted regions are the darker circular shapes one of which is highlighted by a black arrow. Since the extracted tissue is still on the transfer film the pattern of extracted tissue aligns to their removal points on the previous image. G) A zoomed image (scale bar, 50 µm) of one of the extracted plaques, where the neurofibrillary tangle is the darker staining. The figure images are adapted and reproduced from (Emmert-Buck *et al.* 1996).

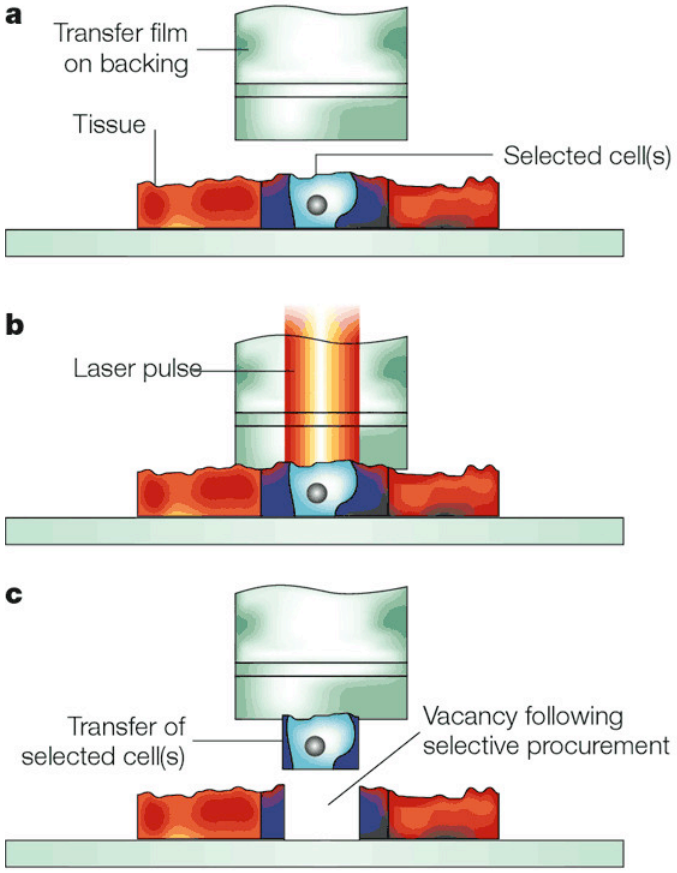

a
Transfer film on backing
Tissue
Selected cell(s)
b
Laser pulse
c
Transfer of selected cell(s)
Vacancy following selective procurement

Figure 4.2: A cartoon of LCM extraction process. a) The transfer film is placed over the tissue and a sub-section is selected. b) A laser pulse thermally bounds the selected sub-section of tissue to the transfer file. c) The sub-selection tissue can then be extracted fro the larger tissue section. This figure is reproduced from (Liotta and Petricoin 2000).

For bulk tissue, total RNA was extracted using Trizol, biotinylated and amplified using the Illumina® TotalPrep-96 RNA Amplification Kit. For laser-capture microdissection (LCM) tissue was immersed in Shandon M-1 embedding matrix (Thermo Electron Corporation, Rockford, IL) and stored at -80°C until use. Cryostat sections (7–8 μm thick) were cut and stained with Cresyl Violet (Ambion, Austin, TX). Laser capture microdissection was performed with ArcturusXT microdissection system (Arcturus, Mountain View, CA). The cell bodies of between 70 and 150 Purkinje cells were captured per subject. The excised cells were selected from the slide surface and captured on LCM Macro Caps. High-quality cellular total RNA was recovered from the collected cells using PicoPureTM RNA isolation kit (Arcturus) and treated with

RNase-free DNase (Qiagen, Valencia, California, USA). The quality of the RNA was analysed using an Agilent 2100 bioanalyzer (Agilent, Foster City, California, USA). Two rounds of amplification were carried out with the Ambion MessageAmp II aRNA kit. It should be noted that the LCM prepared samples underwent two rounds of amplification whereas the bulk tissue samples underwent one round of amplification, and this may introduce some bias to the expression measurements. However, it is unlikely that this bias, if present, would affect the type of primary analyses being performed in this study since the analyses are within tissue and then results compared across tissues. Amplified RNA from either bulk tissue extracts or LCM Purkinje cells were hybridized onto Illumina HumanHT-12 v3 Expression BeadChip (Illumina). The Illumina HumanHT-12 v3 Expression BeadChips assay the expression levels of approximately 49,000 human Refseq transcripts using 50-mer probes. The Illumina HumanHT-12 v3 Expression BeadChip is constructed based on the Illumina Sentrix bead arrays, like the HumanRef-8 v1.0 and v2.0 assays described in Chapters 2.2.2.2 and 3.2.2.2 respectively. However, the HumanHT-12 platform contains more arrays per chip, where 12 samples can be processed per chip and contains more bead-types per array assaying 49,000 RNA transcripts.

4.2.3: Data Analysis

4.2.3.1: Genotype data

Genotype based filtering included subject and SNP filtering based on call rate, expected subject gender, relatedness among subjects and population outliers when combined with HapMap3 genotypes. Genotype-based metrics for filtering were acquired using the PLINK toolset (Purcell *et al.* 2007) and R (R Core Team 2012).

The threshold call rate for inclusion of the sample in analysis was 95%. The SNP call rate was computed using the *missing* command within the PLINK v1.07 software toolset. The gender of the samples reported to NABEC by the brain banks was compared against their genotypic gender using PLINK 's *check-sex* algorithm. The *check-sex* algorithm determines a sample's genotypic gender based on heterozygosity across the X chromosome. Genotype data of the samples were compared for cryptic relatedness using the Identity-By-Descent (IBD) procedure within PLINK. No subjects were excluded based on call rate, gender or relatedness.

To confirm the ethnicity of the samples, Identity-By-State (IBS) clustering and multidimensional scaling analyses were performed within PLINK. The ethnicity check was run using the genotypes from the NABEC samples and genotypes from the HapMap3 populations. Outlier detection was based on a sample's distance in the first and second principal components being more than three standard deviations (S.D.) from the mean of the study cohort for those components. Three subjects were identified as outliers based on population and excluded from further analysis. Population structure for the first two principal components after excluding NABEC population outliers is shown in Figure 4.3.

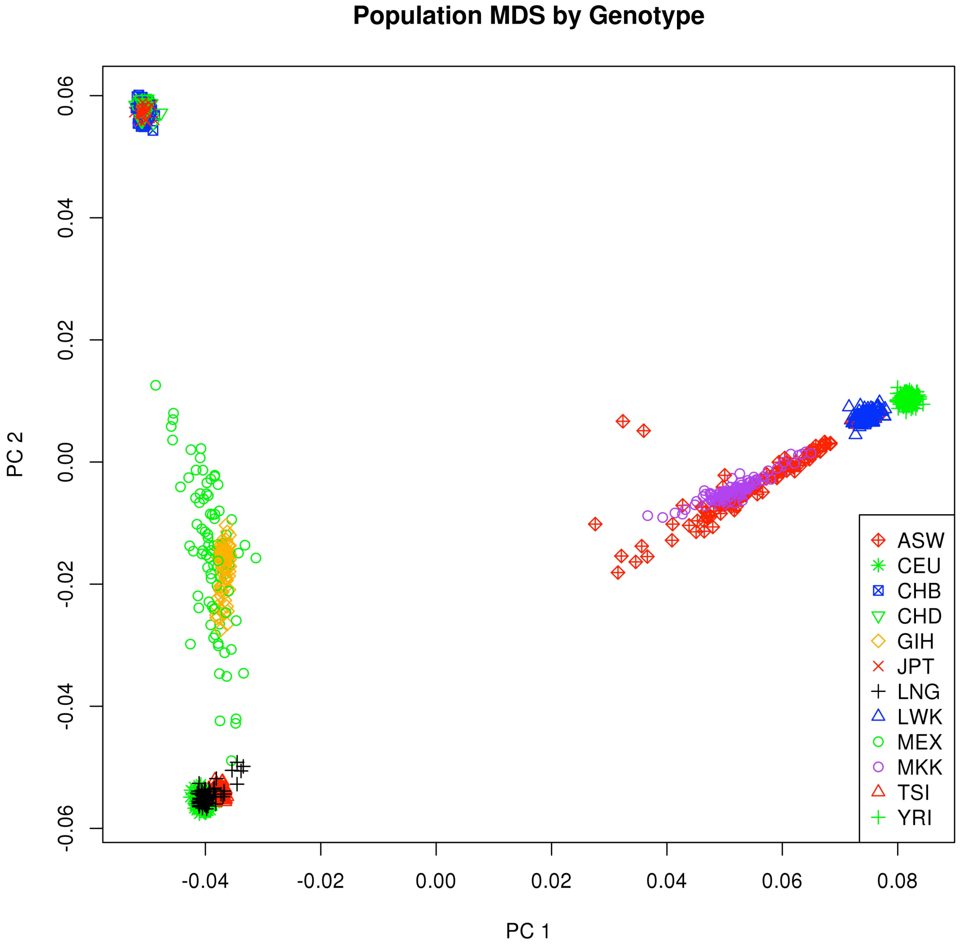


Figure 4.3: Population MDS plot, based on genotype, from genome wide Identity-By-State pairwise distances between the subjects used in this study (LNG, black +'s) and HapMap III population samples. The plot shows that all of the post quality control screened subjects used in this study match their reported population ethnicity of Caucasians of European decent. Outlier detection was based on 3 S.D. from the mean for the reported population group. There are no population outliers as this plot was re-generated post removal of outlier subjects.

A two-step imputation process was performed excluding genotyped SNPs where SNP and subject call rate was less than 95%, MAF was less than 1% and Hardy-Weinberg equilibrium (HWE) p-value was less than 0.000001. Mach (Li *et al.* 2010; Howie *et al.* 2012) and MiniMac (Howie *et al.* 2012) were used to impute genotypes for ~38.9 million autosomal SNPs based on European reference haplotypes from the 1000 Genomes Phase1 v2.20101123 data (1000 Genomes Project Consortium *et al.* 2012). Imputed SNPs were excluded if their MAF was less than 0.035 or the r^2 was less than

0.3 between known and imputed genotypes. This process resulted in ~6.4 million autosomal SNPs available for eQTL analysis. The MAF threshold of 0.035 is an estimate that establishes a lower bound for the smallest allele frequency testable in this sample series for eQTL analysis. This lower bound was determined based on the MAF cutoff at which it becomes less likely that the imputed allele dosages would have at least 3 minor homozygotes present in this study.

4.2.3.2: mRNA expression data

Transcript expression data was cubic spline normalized (Workman *et al.* 2002) and exported using the Illumina GenomeStudio Gene Expression module. Cubic spline normalization removes curvature that may exist in the data as a result of non-linear relationships between samples or groups of samples. The Illumina HumanHT-12 probes were re-annotated using the ReMOAT tool (Barbosa-Morais *et al.* 2010) to identify probes that may have design issues. The ReMOAT annotation tool performs a re-alignment of the Illumina probes then re-annotates the probes based on multiple public genomic and transcriptome resources and then scores the probes quality. I excluded all probes that were annotated as 'Bad' or 'No Match' by the ReMOAT tool. Probes are labelled as bad if the probe aligns to repeat sequences, intergenic or intronic regions, more than one transcript from different genes, or contains more than three mismatches to target sequence. Probes are labelled as no match if they do not align to any transcript or genomic region. Filtering based on the ReMOAT quality score fields resulted in 14067 of the 48803 probes being excluded from analysis. Expression-based filtering included removal of subjects if any of their sample expression profiles were outliers based on their mean normalized intensity profile or their overall detection rate in any of the sample groups. A subject's overall detection rate was computed per sample type as the fraction of detected (expressed) probes from the total probe set that passed the ReMOAT probe quality filter. Outlier detection was based on a sample's distance for their mean expression profile and transcript detection rate being more than three standard deviations (S.D.) from the mean of the study cohort for those measures. Four subjects where excluded as outliers based on their average expression level and detection rate; two were Purkinje cell samples (Figure 4.4), one that was an outlier in both cerebellum and frontal cortex, and an additional one that was also an outlier in cerebellum. Additionally, eight of the subjects were not assayed for either cerebellum or cerebral frontal cortex.

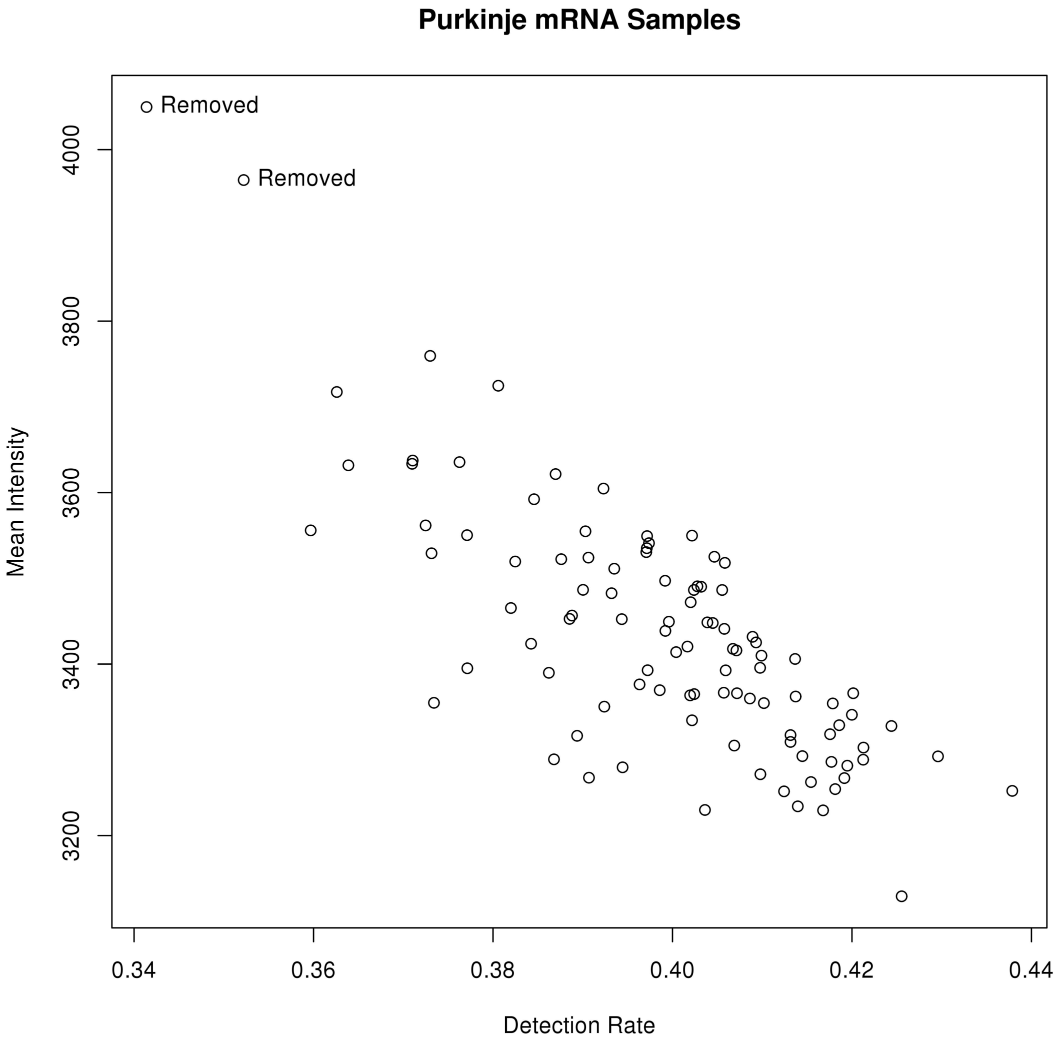


Figure 4.4: Scatterplot visualizing the quality control step applied for detecting poor quality mRNA samples based on their overall profile being identified as an outlier. Here the data for the Purkinje cell samples is shown. The two metrics tested for outlier detection are the subjects overall detection rate (fraction of expressed probes), based on all QCed probes, and the average intensity over on all QCed probes. Outlier detection was based on 3 S.D. from the mean of the expression intensity or detection rate for the study cohort. As shown above two subjects were found to be outliers and excluded from the eQTL analysis in all groups.

4.2.3.3: Polymorphism(s) in Assay Probes

Sequence variants within the sequence interval of the probe design used to assay individual transcripts may cause differential hybridization and inaccurate expression measurement. To correct for potential hybridization bias resulting from polymorphisms within the mRNA 50-mer probe, I identified all probes where this artefact may affect the analysis and excluded these probes. A probe containing a polymorphism was only considered to have an effect and therefore excluded if it contained a variant that had a MAF greater than or equal to 0.03529. This MAF is the same estimated lower bound appropriate for eQTL analysis in a cohort of this size. We used variants and their frequencies based on the European subjects from the 1000 Genomes Phase1 v2.20101123 (1000 Genomes Project Consortium *et al.* 2012) data to identify the probe set for exclusion from further analysis. This removed 1,938 probes from the overall chip content in addition to those previously identified with design issues by the ReMOAT tool.

4.2.3.4: Selection of traits for analysis

Expression probes were considered reliably detected within an individual sample if the Illumina Detection p-value was <= 0.01. An expression probe was selected for eQTL analysis if the probe was reliably detected for 95% of the QC filtered subjects within a tissue sample group and free from probe design issues. In total 10,850 mRNA transcripts were analysed within at least one sample group: 7,044 mRNA transcripts were present within all three groups; 8,025 in Purkinje cell, 9,869 in cerebellum, and 9,983 in cerebral frontal cortex.

4.2.3.5: Correction for known Biological and Methodological Covariates

The selected probe expression profiles were then adjusted using known covariates for subject age, gender, post-mortem interval, principal components 1 through 12 based on identity-by-state pairwise distances within this cohort and the HapMap3 populations representing any possible population substructure in the cohort (Price *et al.* 2006), and the mRNA sample preparation/hybridization batch. The expression profiles were then $\log_2$ transformed and covariates were stepwise fitted in R (R Core Team 2012) against the following model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

In this model, $\beta_0 \ldots \beta_n$ represent the continuous and categorical covariates. The residuals of this model fit for each probe were then standardized to a z-score and used as the quantitative trait for the eQTL analysis. A z-score (or standard score) is a measure of how many standard deviations a data point is from the mean of the data set.

4.2.3.6: Expression quantitative trait loci analysis

eQTL analysis was then performed using the standardized residuals for every selected and adjusted trait in both brain tissue regions and Purkinje cells using mach2qtl (Li *et al.* 2009) to regress the trait with the allele dosage probabilities. For each trait analysed, only SNPs that are *cis* to the trait and passed our imputation quality control and SNP MAF threshold were considered in the analysis. For these analyses, *cis* is defined as the genomic region that contains the trait (the gene encoding the transcript), where the boundaries of the genomic region are +/- 1Mb from the mRNA transcript start or end site.

4.2.3.7: Correction for multiple tests

To correct for the large number of tests performed in the eQTL analysis of these three sample groups I applied a Bonferroni correction based on the total number of estimated independent tests performed per sample group. The approximate number of independent tests being performed can accurately be estimated by considering the number of LD blocks and inter block variants being tested (Duggal *et al.* 2008). To estimate the number of independent variants I reduced the (1000 Genomes Project Consortium *et al.* 2012) European subjects genotype data to the SNP set that was included in the eQTL analysis for our cohort. This genotype set was then LD pruned to estimate the number of SNPs representing the amount of independent genetic variation represented within the imputed genotype set used in the analysis (Nicodemus *et al.* 2005). Then for each eQTL analysis group, the total number of tests performed was computed by summing all *cis* tests per mRNA transcript probe. Each independent SNP may be tested in *cis* against multiple transcript probes and each of these is a separate test included in the sum of *cis* tests. Based on the total approximate independent test counts per analysis group a Bonferroni cut-off was determined. The number of approximate independent tests performed per analysis group differs based on the number of transcript probes that were reliably detected and selected for analysis within that group (as discussed in Section 4.2.3.4). For the Purkinje cell samples a total of 2,758,709 independent tests (36,710,361 total actual) were performed and the threshold for significance was a p-value <= 1.81x10^{-8}. In the cerebellar tissue samples a total of 3,375,842 independent tests (45,054,510 total actual) were performed with a threshold of significance of p-value <= 1.48x10^{-8}. In the cerebral frontal cortex tissue samples a total of

3,436,351 independent tests (45,853,437 total actual) were performed with a threshold of significance of p-value <= 1.46x10^{-8}.

To estimate a threshold of suggestive eQTL signal a Benjamini & Hochberg (1995) false discovery rate was computed (Benjamini and Hochberg 1995). This cut-off was computed based on all *cis* tests performed per analysis group regardless of independence of the SNPs. For the Purkinje cell group this threshold for suggestive eQTL signal is a p-value <= 1.52x10^{-6}, in cerebellum p-value <= 1.88x10^{-5} and in cerebral frontal cortex the threshold is a p-value <= 1.55X10^{-5}.

4.2.3.8: Data Access

The genotype and expression data for this study is publically available as NCBI's dbGaP study accession phs000249.v1.p1 and NCBI's GEO series accession GSE37205 (Edgar, Domrachev and Lash 2002; Barrett *et al.* 2007; Mailman *et al.* 2007) (Figures 9.2 and 9.4, *Appendix*).

4.3: Results

4.3.1: Expression in single and heterogeneous neuronal cell type populations

Prior to eQTL analysis a general comparison of expression levels between Purkinje cells, cerebellum and cerebral frontal cortex was performed. Within Purkinje cells 8,025 mRNA transcript probes were well detected while 9,869 and 9,983 mRNA transcripts were well detected in cerebellum and cerebral frontal cortex respectively. Here well detected is defined as detected within at least 95% of the post quality control screened samples. In total 10,850 mRNA transcripts are well detected in at least one group and 65% of these are well detected in all three groups (groups are Purkinje cell, cerebellum and cerebral frontal cortex sample sets). Of this total population of detectable mRNA transcripts, 74% were well detected in Purkinje cells while 91% and 92% of this total transcript population were well detected in cerebellum and cerebral frontal cortex respectively (Figure 4.5). It is plausible that the significant decrease in well-detected transcripts within the Purkinje cell group may be expected, as it is a single neuronal cell type whereas the other two groups comprise tissues of mixed cell types. It is important to note however that the Purkinje cell group was isolated by laser capture micro-dissection, thus it may be possible that some portion of this decrease in well detected mRNA transcript probes is an artefact of the sample isolation and preparation method. The use of LCM is labour intensive and involves a time sensitive protocol. It has been shown that RNA quality can be affected during LCM primarily by humidity or the presence of water but also by cell count and tissue staining (Clément-Ziza *et al.* 2008; Ordway *et al.* 2009).

Simply examining those transcripts detected in each tissue, I did not find an excess of the transcripts detected in Purkinje cells in cerebellar tissue (90%) when compared to those found in the frontal cortex (91%). This likely reflects the fact that Purkinje cells only represent a fraction of the total cell population in cerebellum and therefore transcripts that are exclusively Purkinje-specific would not constitute enough of a signal within the total cerebellum to be detected. Additionally for this study it was the cell body of the Purkinje cells that were captured for analysis, these cell bodies are located within the Purkinje cell layer of the cerebellar cortex, so not only do Purkinje cells represent a fraction of the total cell population in the cerebellum but also their cell bodies are not distributed throughtout the tissue. The Purkinje cell layer is a narrow region of the cerebellar cortex located between the molecular layer, which contains the dendrites of the Purkenje cells, and the granule cell layer. This observation, that cell-specific transcripts will likely not be well detected in whole tissue (and implicitly that the transcripts detected in whole tissue are therefore likely to be quite ubiquitously expressed in brain) is quite generalizable. This is reflected in the relatively small percentages of mRNA transcripts that are well detected only within a single tissue or cell type: 4.6% of transcripts were well detected in Purkinje cell only, 6.1% in cerebellum only, and 6.6% in frontal cortex only. As further reinforcement of this possibility, 65% of the mRNA transcript probes are well detected in all three groups and an additional 18% are shared between the mixed cell type samples of cerebellum and frontal cortex (Figure 4.5).

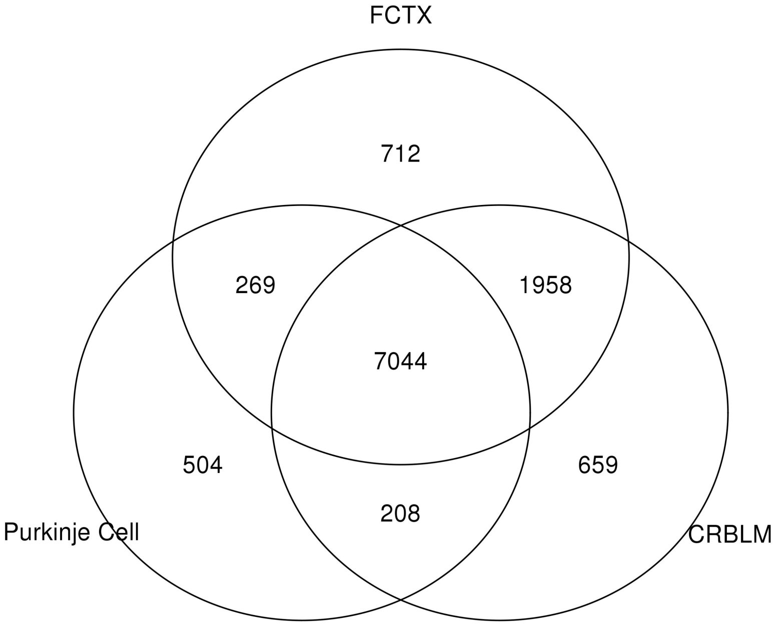


Figure 4.5: Venn diagram showing set intersections for well-detected mRNA transcript probes between the two brain tissue regions and specific cell type.

Next I assessed how similar the expression profiles of the mRNA transcripts are within the tissues. Considering the population of mRNA transcripts well detected in all three sample groups the overall expression profile of Purkinje cell was not significantly more similar to either of these two bulk tissue regions, the R^2 was 0.55 and 0.57 when comparing with cerebellum and cerebral frontal cortex respectively. Additionally the Purkinje cell data are enriched for genes such as *CALB1*, *PCP2* (also known as Purkinje cell-specific protein L7) and *GRID2*. These genes are known to be specifically expressed by Purkinje cells (Figure 4.6) (Oberdick, Levinthal and Levinthal 1988; Zhang, Zhang and Oberdick 2002; Rong, Wang and Morgan 2004). These three Purkinje specific markers are well detected in both the Purkinje cell and cerebellum samples but are not well detected in frontal cortex samples; where well detected means that transcript is present in 95% of the samples. In considering the average expression levels of these Purkinje specific markers both *CALB1* and *PCP2* are more highly expressed in the Purkinje cell samples than in the cerebellum samples whereas *GRID2* is similar in both (Figure 4.7). It should be noted that while LCM enriches for a specific cell type, and markers of Purkinje cells were more highly expressed in LCM compared to bulk tissue, the separation is imperfect and tightly associated cells such as glia and granule cells may also be captured with the Purkinje cell bodies and contribute some signal in the LCM Purkinje cell samples. Myelin basic protein (*MBP*) is a known oligodendroglial marker (Friedrich *et al.* 2012); there are three transcript probes for this gene on the Illumina platform that were detected in 92%, 6% and 36% of our Purkinje cell samples. Glial fibrillary acidic protein (*GFAP*) is a known astroglial marker (Friedrich *et al.* 2012); there is one transcript probe for this gene on the used expression platform, which was detected in all of our Purkinje cell samples. The expression of *GFAP* is likely an indication of Bergmann glial cells also being captured with the LCM samples as they are found in the Purkinje layer of the cerebellum and are known to express *GFAP*. Parvalbumin (*PVALB*) is another neuronal marker for Purkinje cells (Friedrich *et al.* 2012); the expression array includes one probe for this gene, which was detected in all of the Purkinje cell samples.

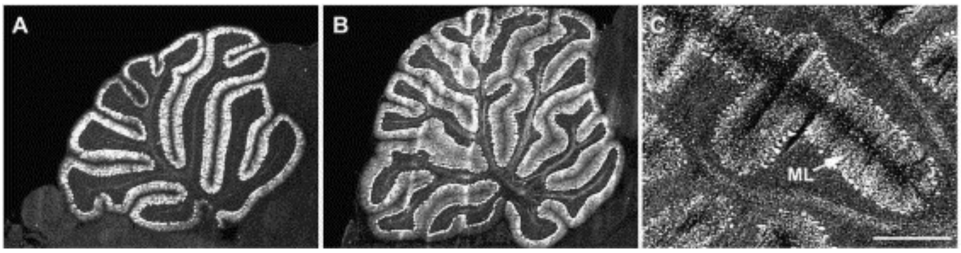


Figure 4.6: In situ hybridization showing the localization of L7 mRNA in Purkinje cell dendrites from A) mouse, B) rat, and C) human (bar in panel = 1.25 mm) This figure is reproduced from (Zhang, Zhang and Oberdick 2002).

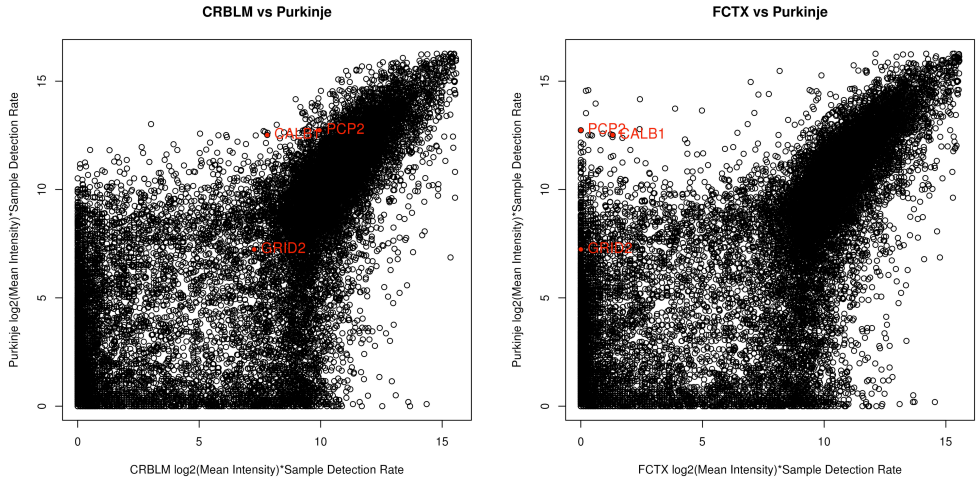


Figure 4.7: Scatterplots showing comparisons between bulk tissue regions and specific neuronal cell type. Each point represents the average expression over all QCed subjects in a tissue region and cell type for all transcript probes that are well detected in at least one analysis group. As each transcript probe is not well detected in all groups the average expression level is scaled by the detection rate within the group. On the left is the comparison between cerebellum and Purkinje cell and the right plot is between the frontal cortex and Purkinje cell. In both plots some Purkinje specific gene markers are highlighted in red: *GRID2*, *CALB1* and *PCP2*. Plot on the right shows that the Purkinje specific markers are close to zero for abundance in frontal cortex as they should be, as Purkinje cells are not present in that tissue region. In the plot on the left all three markers are detected, as Purkinje cells are present in the cerebellum; however, as should be expected the expression levels for *PCP2* and *CALB1* are higher in Purkinje cells.

4.3.2: Genotype effects mRNA expression

Much like our previous eQTL study in our larger four brain region study I was able to identify eQTL in Purkinje cells, cerebellum and cerebral frontal cortex samples even with the reduced sample size (~60% of previous subject cohort size). As with other eQTL studies the strength of the signal was evenly distributed around the transcription start site (Figure 4.8), with stronger signal typically closer to the TSS. The average distance between the transcript start site and the significantly correlated SNPs is 56 kilobases (Kb) and per analysis group: 36 Kb for Purkinje cell, 56 Kb for cerebellum, and 59 Kb for cerebral frontal cortex. It is unknown whether the tighter distribution for Purkinje cell is because this is a single cell type while the other two are heterogeneous tissues. For the heterogeneous tissues, it may be possible that the distribution is broader because we are detecting more eQTL related to enhancers modulating expression for the multitude of neuronal and non-neuronal cell types present in the bulk tissue. For instance, if many of the eQTL signals in the core promoters are shared across cell types and most of this signal is more proximal to the transcription start site (TSS), this will draw the average distance closer to the TSS with a more narrow distribution of distances. While if the eQTL signal found in cell-specific enhancers is located further from the TSS, both up and downstream of the TSS, this will elongate the average distance from the TSS. When detecting these eQTL for cell-specific enhancer(s) in a single type of cell this will also broaden the tail of the distribution; however, in a bulk tissue of heterogeneous cell types if the signals from multiple cell-specific enhancers are detected this will broaden the overall distribution of the distances. Within the Purkinje cell group 10 (0.1%) traits (mRNA transcript probes) were found to have a significant correlation with *cis* SNPs, or 472 trait/SNP pairs. Significant *cis*-eQTL were also identified, as expected, in our heterogeneous tissue regions, 64 (0.6%) traits or 2,565 trait/SNP pairs in cerebellum and 61 traits (0.6%) or 2,090 trait/SNP pairs in cerebral frontal cortex.

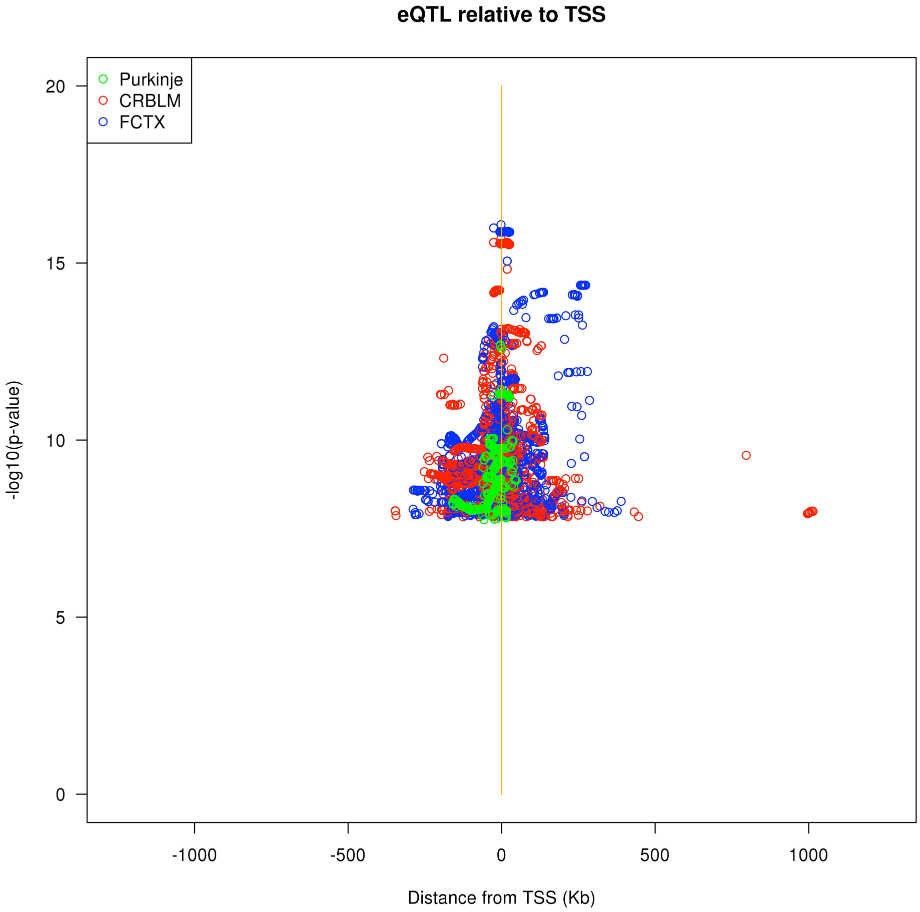


Figure 4.8: Distribution of all significantly correlated SNPs relative to the transcription start site (adjusted for strand) of the mRNA transcript the SNP is correlated with. The X-axis is the physical distance between a SNP and the TSS of the transcript and the Y-axis is –log10(p-value) from the regression test between the SNP's allele dosage and the transcript's expression levels. The analysis tissue groups are colour coded as follows, Purkinje cell (green), cerebellum (CRBLM, red), and cerebral frontal cortex (FCTX, blue). Results are concordant with previous eQTL studies being fairly evenly distributed about the transcription start site with the eQTL signal being stronger closer to the gene.

4.3.3: Many eQTL appear to be cell- and tissue-region specific

In contrast to our previous study, many of the significant eQTL detected appear to be cell-type and tissue-region specific. A large portion of this is likely accounted for by the previous study including two biologically and ultrastructurally related regions (cerebral frontal and temporal cortex). In that study, it was shown that both the cerebral frontal and temporal cortex had high similarity in their overall expression profile and their eQTL (Figures 3.6B and 3.8D). Of the 10 mRNA expression probes or traits with a significant eQTL in the Purkinje cell analysis, five of these were only significant within the Purkinje cell and not significant in the cerebellum or cerebral frontal cortex samples (Figure 4.9). For the other five mRNA expression probes, with significant eQTL in Purkinje cell, four of these also had significant signal in both cerebellum and cerebral frontal cortex, and 1 was shared between only the Purkinje cell and cerebellum groups. Table 4.1 lists these 10 transcripts with their best eQTL p-value from the Purkinje cell analysis. The single mRNA transcript probe shared between Purkinje cell and cerebellum is for peroxisomal biogenesis factor 6 (*PEX6)*. Mutations in *PEX6* are linked to Zellweger syndrome, a severe neonatal neurodegenerative disorder with a hallmark of delayed cerebellar development (Volpe and Adams 1972). For the heterogeneous tissues the significant eQTL identified within the cerebellum and cerebral frontal cortex also appear to be specific to their tissue regions with 35 of 64 and 33 of 61 respectively significant only within their tissue region. As may be expected the heterogeneous tissue types shared more in their intersection than with the Purkinje cell group.

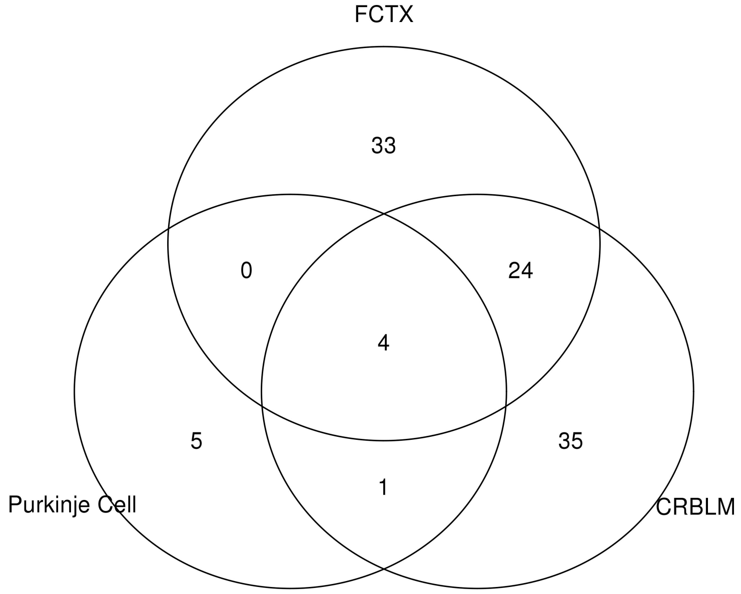


Figure 4.9: Venn diagram showing set intersections for mRNA transcript probes, with a significant eQTL, detected between the two brain tissue regions and Purkinje cells. As shown, between 50% and 55% of transcripts, with a significant eQTL, are cell or tissue specific.

It is important to note that while based on significant eQTL results, the abundance of cell and region-specific eQTL signal does not mean that some suggestive signal is completely absent for the other cell and tissue groups. To further investigate the similarity in the eQTL signals I plotted the regression effect (correlation coefficient) for each of the two-way comparisons based on the union of all significant trait and SNP pairings (Figure 4.10). The effect sizes are computed by mach2qtl (Li *et al.* 2009), used to identify the correlations by regression between allele dose probabilities and expression levels. As shown in these plots, many trait and SNP signals are of similar effect size but not necessarily statistically significant within each of the analysis groups. However, there are very strong signals that are specific to the Purkinje cell type, or one of the other two tissue regions.

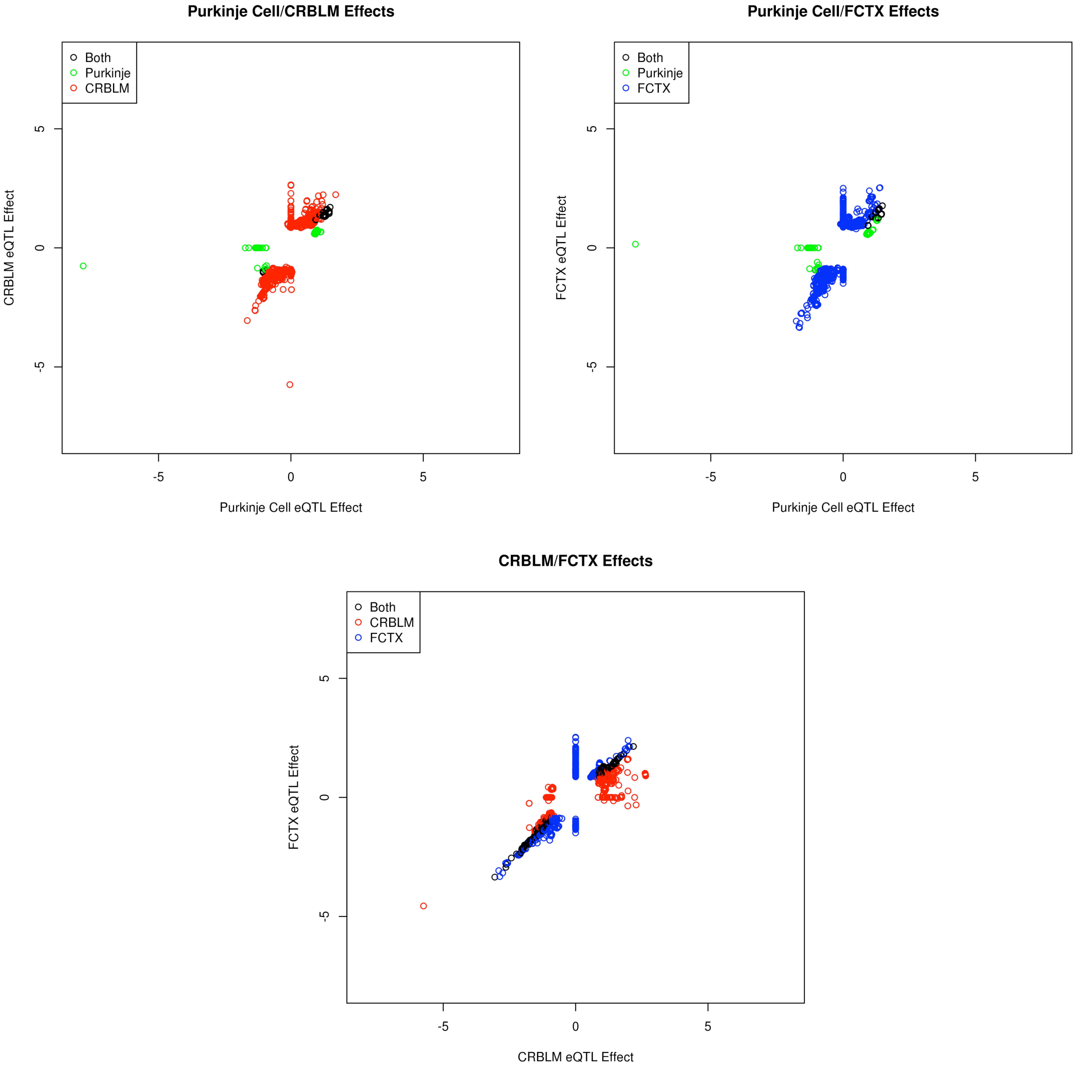


Figure 4.10: Scatter plots of the effect sizes compared between the different analysis groups for all SNP and traits pairs that were significant in one or both of the comparison tissue groups plotted. The effect sizes (regression correlation coefficients) were computed by mach2qtl (Li *et al.* 2009) to identify the correlations between allele dose and expression levels. If the SNP and trait pairing was significant in only one group it is colour coded by group: Purkinje cell (green), cerebellum (CRBLM, red) and cerebral frontal cortex (FCTX, blue). If the pairing was significant in both comparison groups it is colour coded black. As expected most significant signals follow positive or negative correlations in both comparison groups, i.e. signal should be in top right or bottom left quadrants of the plots. Also, completely vertical and horizontal lines of signal are primarily artificial in cases where a correlation is significant in one analysis group but was not tested in another analysis group because the transcript was not well detected in that analysis group, in these instances the effect size is set to zero for the missing correlation.

To further elucidate the differences between group specific and shared eQTL signal I have included some of the mRNA transcript probe specific results using regional Manhattan plots to show the associated SNP's position relative to the mRNA transcript the expression probe captures. Included are examples of eQTL that appear to be cell-specific, but with suggestive signal in the other tissues, shared but borderline sub-significant in one of the tissues, and significant in all three groups. Of the five apparent Purkinje cell specific eQTL one of these is because the mRNA transcript was well detected in Purkinje cells but not in the cerebellum or cerebral frontal cortex samples, and therefore did not meet the quality control thresholds for inclusion in their analysis groups. This eQTL is for a transcript of *CCZ1B*, encoding CCZ1 vacuolar protein trafficking and biogenesis associated homolog B. The regional Manhattan plot is shown in Figure 4.11.

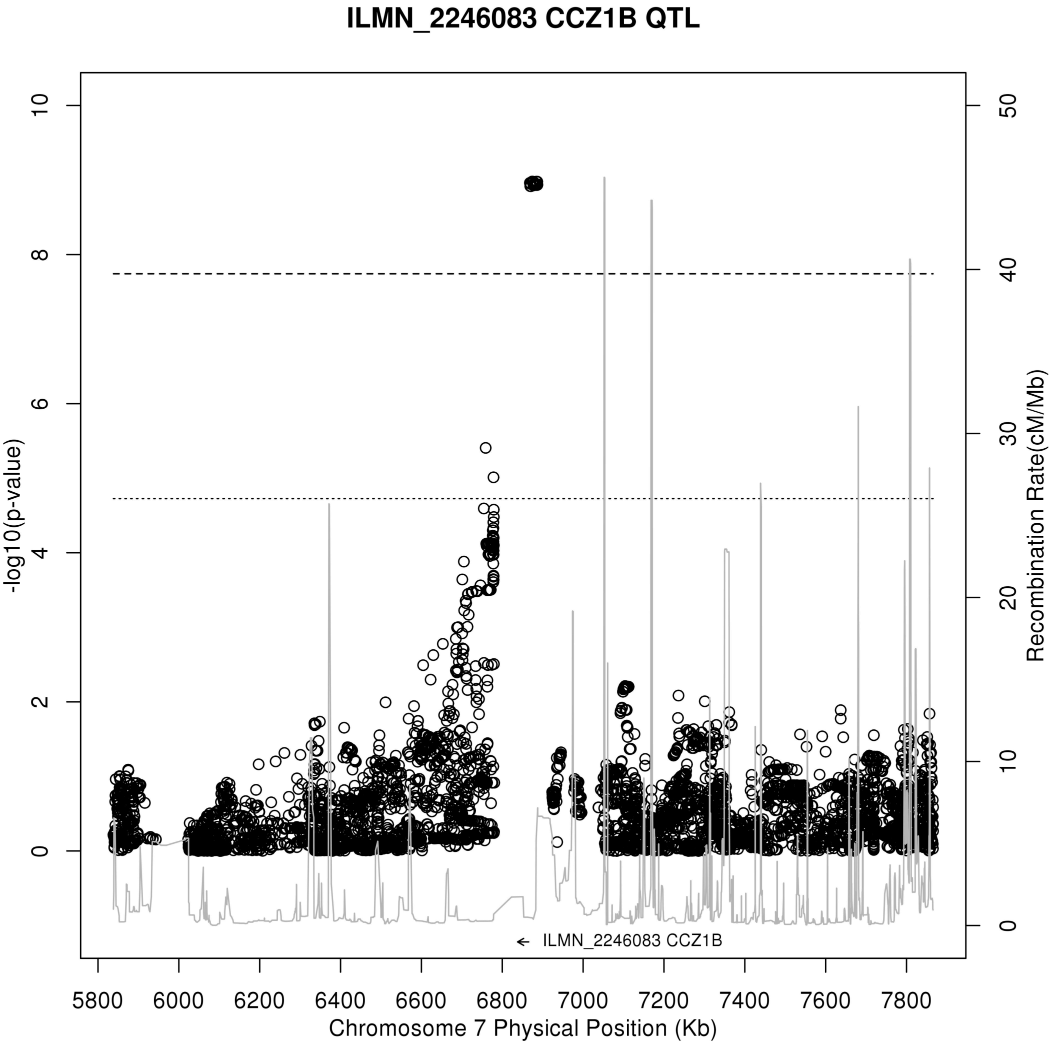


Figure 4.11: Manhattan plot showing this region's eQTL p-values for an mRNA transcript for the gene *CCZ1B*. Each point represents the p-value for a specific SNP, along chromosome 7, that is *cis* to the *CCZ1B* transcript. Here only the Purkinje cell eQTL is present and significant, there is no data for cerebellum and cerebral frontal cortex as this transcript is not well detected in those tissues, and so does not meet criteria for inclusion in analysis of those tissues. Also included in the plot are the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. Threshold for significance is denoted by horizontal dashed line and the suggestive signal threshold is denoted by the horizontal dotted line. The relative position of the gene is the labelled arrow centred near the bottom of the plot. The direction of the arrow is the gene's strand.

An example of an mRNA transcript with a significant eQTL only within Purkinje cells but also with suggestive eQTL in cerebellum and cerebral frontal cortex is for a transcript of *ALDH3A2*, encoding aldehyde dehydrogenase 3 family member A2. As shown in the regional Manhattan plot for this eQTL while a significant signal is not seen intersecting with that of the

Purkinje cell signal, there is clearly a suggestive peak for this transcript in cerebellum and cerebral frontal cortex over the same genomic interval (Figure 4.12).

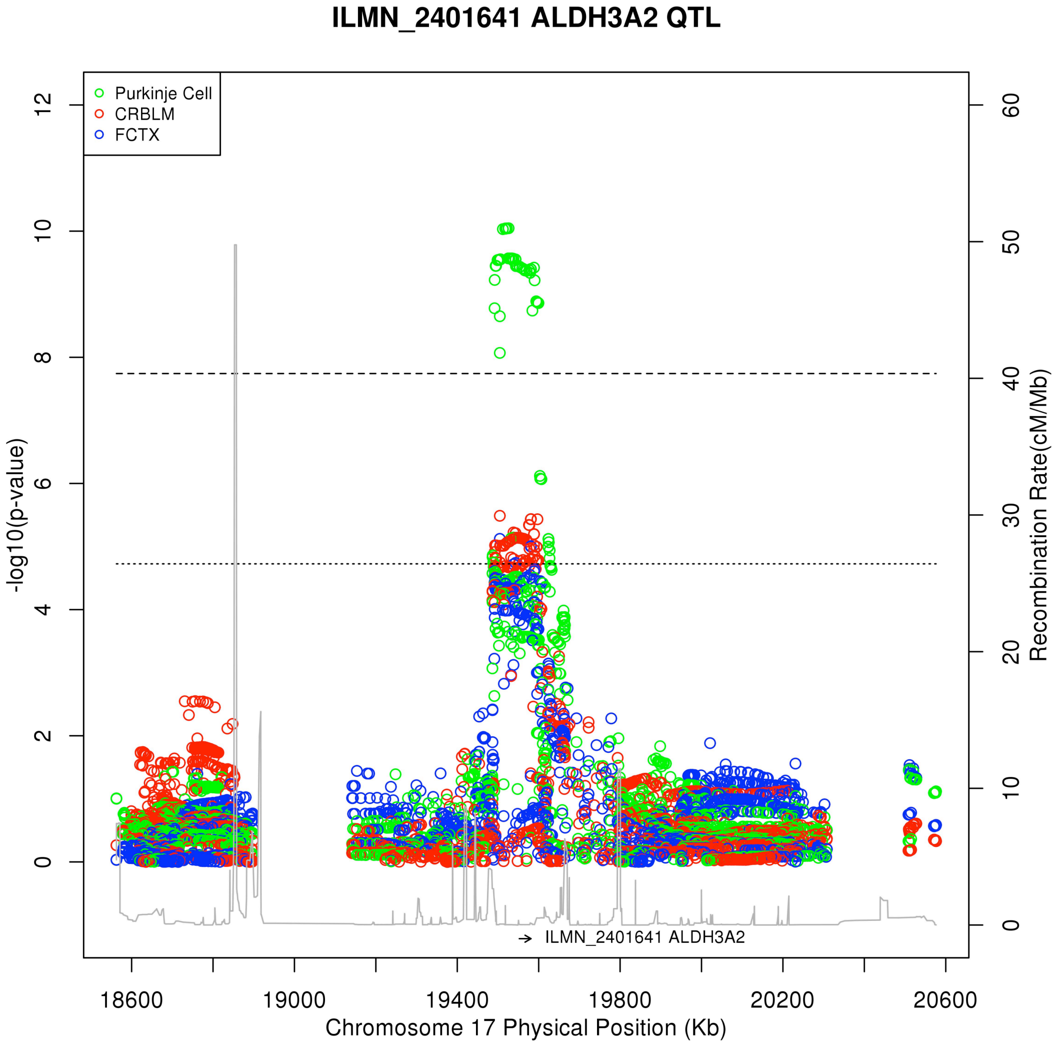


Figure 4.12: Manhattan plot showing this region's eQTL p-values for an mRNA transcript for the gene *ALDH3A2*. Each point represents the p-value for a specific SNP, along chromosome 17, that is *cis* to the *ALDH3A2* transcript. Here only the Purkinje cell eQTL is significant; however, there is some suggestive signal also present in both cerebellum (CRBLM, red) and cerebral frontal cortex (FCTX, blue). Also included in the plot are the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. Threshold for significance is denoted by horizontal dashed line and the suggestive signal threshold is denoted by the horizontal dotted line. The relative position of the gene is the labelled arrow centred near the bottom of the plot. The direction of the arrow is the gene's strand.

As an example of an eQTL that has shared signal across analysis groups, but within one of the groups the signal was borderline sub-significant, is for a transcript of *PEX6*. This transcript was mentioned earlier as the eQTL transcript probe that was found to be significant in both the Purkinje cell and cerebellum but not within the cerebral frontal cortex. Upon further inspection of the eQTL signal for this probe, while not reaching statistical significance within the current analysis, the signal is close to the predetermined threshold for significance. This is clearly shown in the regional Manhattan plot for *PEX6* (Figure 4.13).

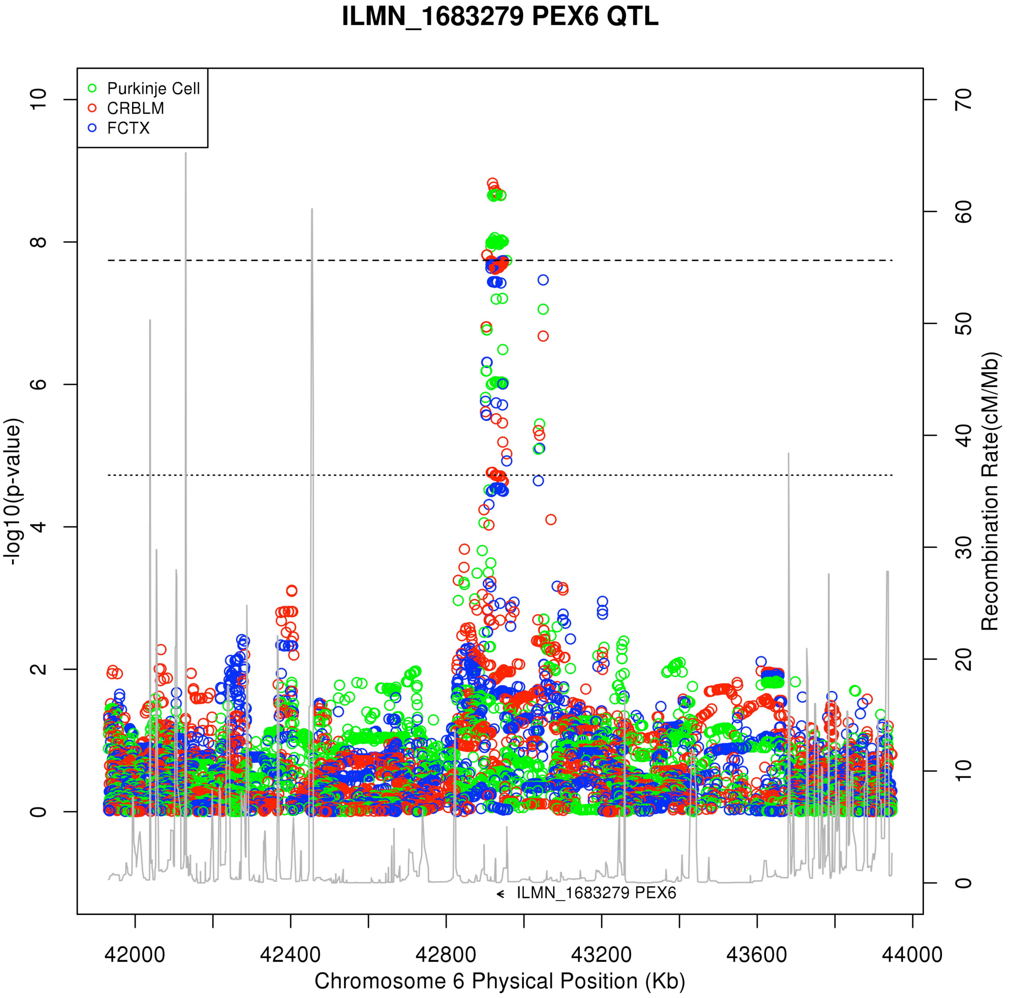


Figure 4.13: Manhattan plot showing this region's eQTL p-values for an mRNA transcript for the gene *PEX6*. Each point represents the p-value for a specific SNP, along chromosome 6, that is *cis* to the *PEX6* transcript. Here both the Purkinje cell (green) and cerebellum (CRBLM, red) eQTL are significant; however signal for cerebral frontal cortex (FCTX, blue) is also but borderline sub-significant. Also included in the plot are the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. Threshold for significance is denoted by horizontal dashed line and the suggestive signal threshold is denoted by the horizontal dotted line. The relative position of the gene is the labelled arrow centred near the bottom of the plot. The direction of the arrow is the gene's strand.

Lastly an example of an eQTL with a significant signal across Purkinje cells, cerebellum and frontal cortex is shown (Figure 4.14). This eQTL is for an mRNA transcript for *CHURC1*. This eQTL is also one of the strongest and most consistent eQTL detected across studies. The regional Manhattan plot for the *CHURC1* eQTL from this analysis is shown below and is clearly significant in each of the three sample groups used in this analysis. Considering not just the *CHURC1* transcript, with a significant eQTL, but all of the individual SNPs making up the eQTL there was very high overlap between the tissues suggesting this is the same eQTL in all three analysis groups. All of the SNPs identified, for the *CHURC1* eQTL, from the Purkinje cell analysis were also significant in cerebellum and cerebral frontal cortex, and 98% of the significant SNPs from the cerebral frontal cortex eQTL were significant in cerebellum. In addition to the strong eQTL seen for *CHURC1* in NABEC related studies previous studies have also identified eQTL for *CHURC1* in lymphoblastoid cell lines from the HapMap populations (Stranger *et al.* 2007; Veyrieras *et al.* 2008) and in human liver (Schadt *et al.* 2008).

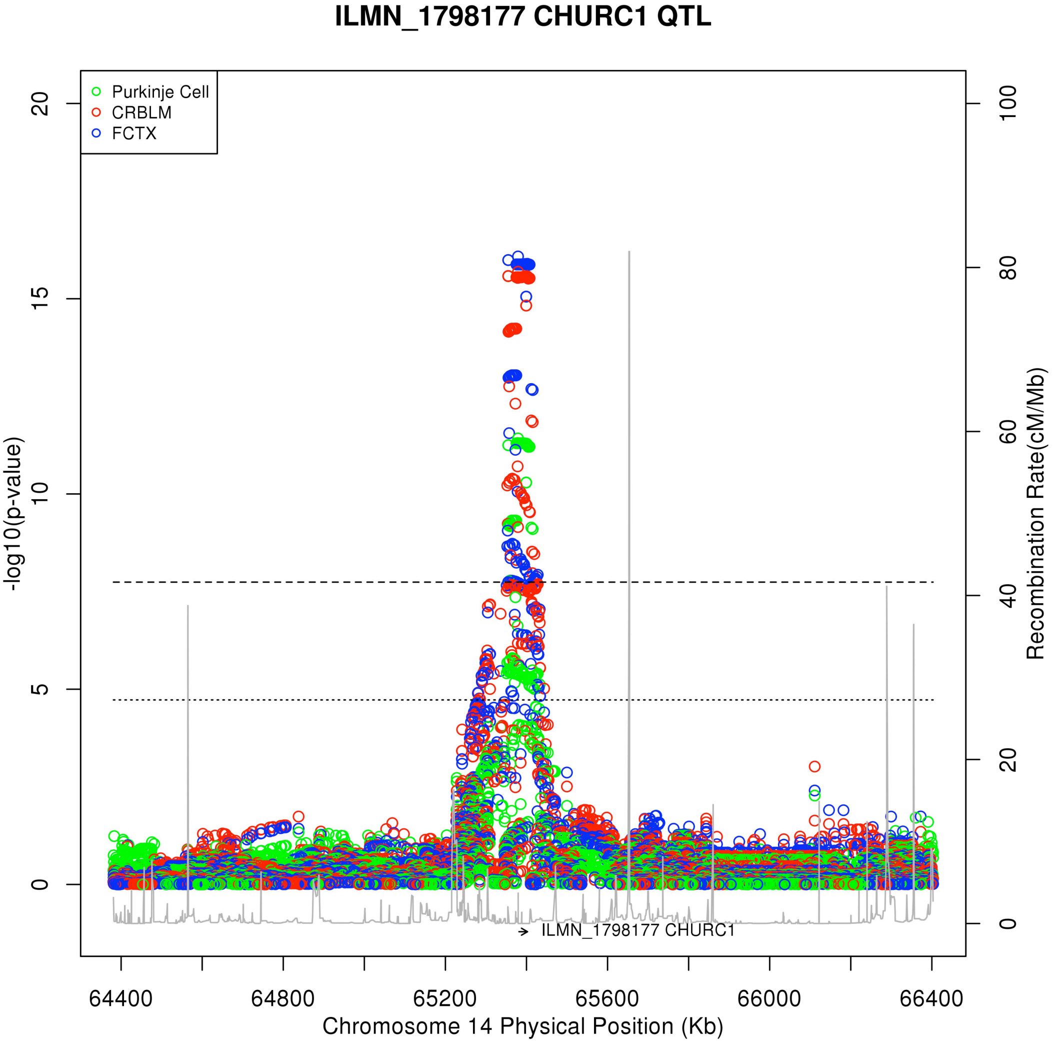


Figure 4.14: Manhattan plot showing this region's eQTL p-values for an mRNA transcript for the gene *CHURC1*. Each point represents the p-value for a specific SNP, along chromosome 14, that is *cis* to the *CHURC1* transcript. Here the eQTL is clearly significant in all groups: Purkinje cell (green), cerebellum (CRBLM, red) and cerebral frontal cortex (FCTX, blue). Also included in the plot are the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. Threshold for significance is denoted by horizontal dashed line and the suggestive signal threshold is denoted by the horizontal dotted line. The relative position of the gene is the labelled arrow centred near the bottom of the plot. The direction of the arrow is the gene's strand.

To get a broader picture of the overlap of eQTL, I generated a larger pool of associated loci using a less restrictive correction, recognizing this will invariably increase the number of false positives. When considering suggestive or better eQTL, based on a FDR correction per group with a mean cut-off of p-value <= 1.19x10^{-5} (1.88x10^{-5} in Purkinje cells, 1.55x10^{-5} in cerebellum, and 1.52x10^{-6} in cerebral frontal cortex) additional signal that is both shared and specific is possibly revealed. The proportion of shared and specific transcripts with an eQTL remains approximately 50% shared and 50% specific for Purkinje cells, while the proportions of transcripts with a region-specific eQTL increases for the other two groups from 55% to 74% in cerebellum and from 54% to 68% for frontal cortex. It is important to note that using a less restrictive test correction results in increasing the sensitivity to detect an eQTL but also results in a decrease in a specificity of the detection. This means that more false negatives may be recovered but at the same time more false positives will be introduced, beyond the 5% originally allowed for. It is also important to clarify that these shared and non-shared counts of eQTL are based on set overlaps of the unique mRNA transcript probes with eQTL signal. However, many more of the genetic variants making up the eQTL may also be specific to Purkinje cells or the other two tissue regions. While this is difficult to assess on a whole-genome basis, it will likely be useful for individual loci of interest to perform finer scale analyses on the existing data, and likely additional functional characterization.

ILMN_ID	RefSeq	Symbol	Chr	Strand	TSS	SNP	SNP Loc	MAF	Dist_to_TSS	P-VALUE	Effect_Direction	Sig_Tissues
ILMN_1719064	NM_031954	KCTD10	12	-	109915155	rs9943689	KCTD10 intron	0.22	24816	2.05E-013	increase	crblm, fctx
ILMN_1798177	NM_145165	CHURC1	14	+	65381078	rs7144811	CHURC1 upstream	0.2	-1931	3.78E-012	increase	crblm, fctx
ILMN_2401641	NM_000382	ALDH3A2	17	+	19552063	rs6587200	LOC101060602 intron	0.43	-25938	9.00E-011	increase	
ILMN_1778371	NM_001008662	CCBL2	1	-	89458643	rs7549363	CCBL2 downstream	0.43	89878	1.04E-010	decrease	crblm, fctx
ILMN_2087080	NM_000969	RPL5	1	+	93297593	rs139806239	RPL5 upstream	0.04	-33016	7.51E-010	decrease	
ILMN_2246083	NM_198097	CCZ1B	7	-	6865926	rs4724860	CCZ1B upstream	0.25	-8347	1.05E-009	decrease	
ILMN_1794522	NM_001970	EIF5A	17	+	7210855	rs1054378	EIF5A UTR3'	0.41	4681	1.85E-009	increase	crblm, fctx
ILMN_1683279	NM_000287	PEX6	6	-	42946981	rs6940814	PEX6 downstream	0.41	22049	2.16E-009	decrease	crblm
ILMN_1712918	NM_000904	NQO2	6	+	3000066	rs138616686	NQO2 exon	0.27	3904	5.87E-009	decrease	
ILMN_1721842	NM_012234	RYBP	3	-	72495774	rs12496352	RYBP intron	0.16	41619	1.26E-008	increase	

Symbol	Name
KCTD10	potassium channel tetramerization domain containing 10
CHURC1	churchill domain containing 1
ALDH3A2	aldehyde dehydrogenase 3 family, member A2
CCBL2	cysteine conjugate-beta lyase 2
RPL5	ribosomal protein L5
CCZ1B	CCZ1 vacuolar protein trafficking and biogenesis associated homolog B
EIF5A	eukaryotic translation initiation factor 5A
PEX6	peroxisomal biogenesis factor 6
NQO2	NAD(P)H dehydrogenase, quinone 2
RYBP	RING1 and YY1 binding protein

Table 4.1: Table, top portion showing the top 10 transcripts based on significant *cis*-eQTL from Purkinje cell analysis, only the best SNP p-value is shown. Columns included are: the Illumina Probe ID, RefSeq transcript ID, Hugo Gene Symbol, Chromosome the gene is located on, strand, chromosomal position of the transcript start site, SNP Id for best correlated SNP, location of the correlated SNP (relative to gene), minor allele frequency for the SNP, distance between the SNP and the transcription start site oriented by gene strand, p-value from the linear regression, the direction of the effect found in the regression where increase indicates an increase in transcript abundance with minor allele's dosage, and which of the other two tissue regions this SNP and transcript pair were significant in. Tissues are cerebellum (crblm) and cerebral frontal cortex (fctx). Bottom portion shows gene names.

4.4: Discussion

The current work shows that cell-type specific eQTL signals can be identified. This reinforces the idea that undertaking eQTL analysis in tissues and specific cell types relevant to biological, cellular and molecular function is important for understanding how particular mRNA transcripts vary with genetic variation within the context of a specific cell-type. Even with the modest sample size used within this study, both shared and tissue-specific eQTL could be identified and the resolution of these eQTL should improve with much larger subject cohorts. While performing eQTL analyses in heterogeneous tissue regions still remains relevant, the application of this work in specific cell types should elucidate much more of the effect of genetic variation on gene expression in a functionally specific context. This is apparent in these results in the set overlap between Purkinje cells and the cerebellum. While Purkinje cells are one of the largest and most distinct neurons in the brain and only found in the cerebellum it is the cerebellar granule cells that are the most numerous neurons in this tissue. The cerebellar granule cell is one of the smallest and most densely packed neurons in the brain. The cerebellum also contains other interneuron cell types such as Golgi cells, basket cells, stellate cell, and Bergman glial cells. It is possible that analysis of a large enough cohort of cerebellum samples would capture all the signal that is detectable with a smaller cohort of Purkinje cell samples; however, there are likely two limitations to this possibility: first, the signal may be too small to detect within the inherent noise of the assay; second, eQTL have been detected with contrasting directions of effect when comparing tissues, and presumably this extends to cell types, thus a signal may be masked in a heterogeneous tissue sample. Purkinje cells only represent a fraction of the total cell population in cerebellum and therefore transcripts that are exclusively Purkinje-specific would not constitute enough of a signal within the total cerebellum to be detected. Lastly, and perhaps more importantly the use of single cell types provides resolution as to which cell types are also contributing or possible reducing the relative signal between the functional neuronal cell type contexts. For example, consider a situation where a particular mRNA transcript has a significant eQTL that is found in both Purkinje cells and another cerebellar cell type such as granule cells. If in this instance separate portions of the loci, such as different haplotype blocks across the promoter, are more important in one cell context but not in the other, it may not be possible to observe the net eQTL signal identified using bulk tissue.

In 2009, Lee *et al.* published a study of tissue-specific expression from fibroblasts, LCLs, induced pluripotent stem (iPS) cells, and differentiated iPS cells. They suggest that using iPS cells helps reduce in-vitro experimental noise that allows for the detection of tissue-specific *cis*-regulatory variants that effect gene expression. They also found that allele specific expression (ASE) is both genotype- and cell-dependent, but that the majority of genotype effects are detectable and consistent across cell types, except for genes on the X chromosome (Lee *et al.* 2009). Studies using multi-tissue samples often focus on across tissue replication so tissue-specific signals may be harder to account for especially across studies but also within studies by not accounting for differences in effect sizes, allelic direction, or additional effects from non-linked variants on the same gene (Fu *et al.* 2012). In the 2011 Nica *et al.*

report, they performed an analysis based on twins from the MuTHER study and measured gene expression in LCLs, skin, and fat to identify *cis*-eQTL. In line with other studies, of the time, they found that 4.7% of genes have an eQTL. Based on method refinement for comparing across-tissue eQTL signals, they found that 30% of their eQTL are shared between all three tissues and that 29% are tissue-specific. However, for the shared eQTL between 10% and 20% have significant differences in magnitude of effect (Nica *et al.* 2011). In 2012, Fu *et al.* published a multi-tissue *cis*-eQTL study based on 85 subjects with gene expression from liver, two adipose tissues, and muscle that were compared to eQTL results from more than 1,200 blood samples. They examined the possible tissue-dependent eQTL based on four categories: specific regulation, alternative regulation, different effect size, and opposite direction of effect (Figure 4.15). The authors consider specific regulation as *cis*-regulatory genetic variation that correlates with gene expression but only in one tissue. Under specific regulation, SNP X is associated with gene A's expression in Tissue 1 but not in Tissue 2; 33% of the eQTL variants they identified fit this category. They consider alternative regulation as *cis*-regulatory genetic variation that is correlated with gene expression in multiple tissues but that each independent variant (locus) is associated with a specific tissue. Under alternative regulation, SNP X is associated with gene A's expression in Tissue 1 but not Tissue 2 while SNP Y is also associated with gene A's expression but in Tissue 2 and not in Tissue 1; 14% of the eQTL variants they identified fit this category. The authors also found that 48% of the eQTL variants they identified were correlated with gene expression in multiple tissues but the effect sizes were different in magnitude, but in the same direction, between the tissues. Additionally, they found that

4% of eQTL variants that were associated with gene expression in multiple tissues had an effect that was in the opposite direction for the same variant in different tissues (Fu *et al.* 2012).

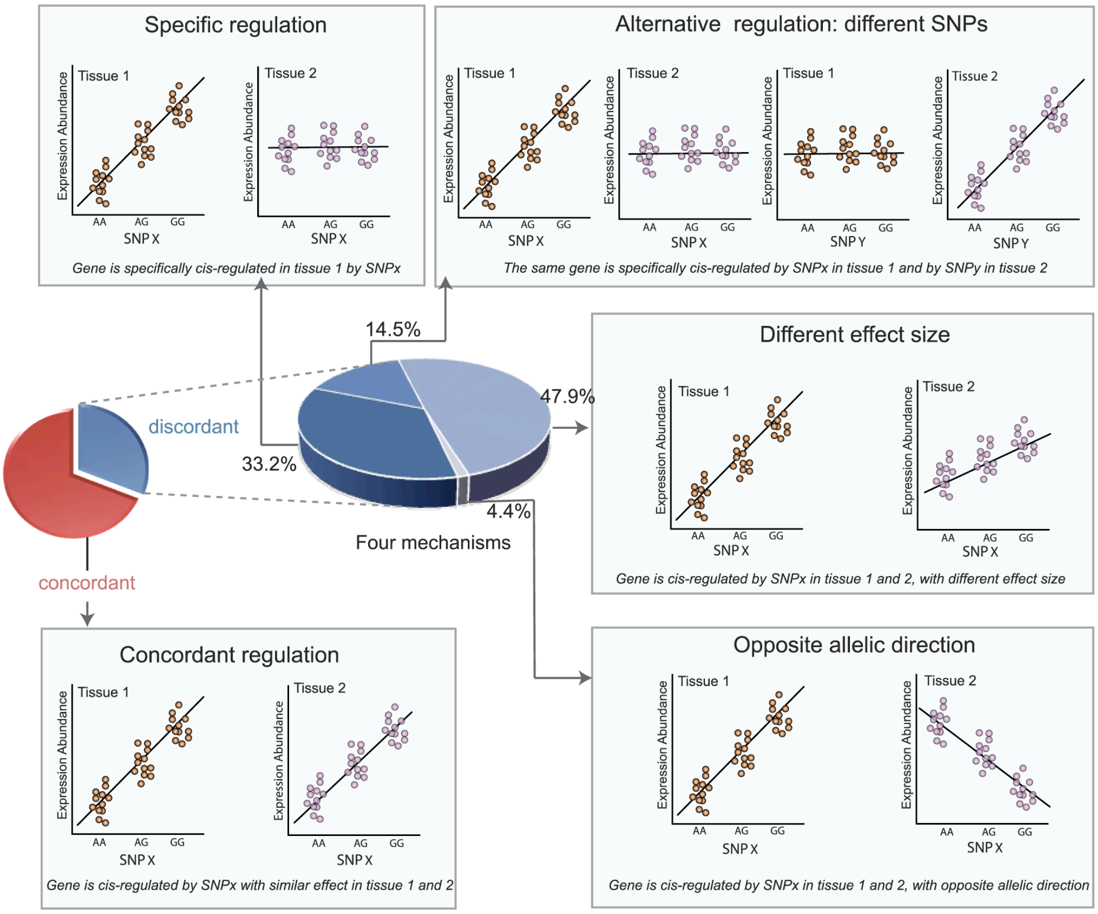

Specific regulation
Tissue 1
Tissue 2
Expression Abundance
AA AG GG
SNP X
Gene is specifically cis-regulated in tissue 1 by SNPx
Alternative regulation: different SNPs
SNP Y
The same gene is specifically cis-regulated by SNPx in tissue 1 and by SNPy in tissue 2
14.5%
47.9%
discordant
33.2%
4.4%
Four mechanisms
concordant
Different effect size
Gene is cis-regulated by SNPx in tissue 1 and 2, with different effect size
Concordant regulation
Gene is cis-regulated by SNPx with similar effect in tissue 1 and 2
Opposite allelic direction
Gene is cis-regulated by SNPx in tissue 1 and 2, with opposite allelic direction

Figure 4.15: Representation of tissue specific eQTL differences found. The pie charts, left centre, show the proportion of concordant and disconcordant eQTL identified between tissues. The disconcordant portion is shown as another pie chart, centre of figure, showing the proportion of disconcordance by the four types considered: specific regulation (top left), alternative regulation (top right), different effect size (centre right), and opposite allele direction (bottom right). This figure is reproduced from (Fu *et al.* 2012).

While identifying eQTL based on samples enriched for Purkinje cells was informative in the detection of cell-type specific eQTL there still remains much to be done in eQTL of brain tissues and cell-types. The comparison of Purkinje cells to heterogeneous brain tissues did identify both shared and cell-specific eQTL but this is not as informative as comparing eQTL from multiple specific cell types from the brain. For instance, being able to identify cell-specific eQTL in Purkinje cells, cerebellar granule cells, and Bergmann glial cells for comparison with the cerebellum would be more informative in understanding cell specific eQTL in brain. Using LCM to extract these specific types of cells from whole tissue is probably not appropriate in order to perform this kind of analysis. Purkinje cells, cerebellar granule cells, and Bergmann glial are all found in close proximity in the same layer of the cerebellum. While it is likely that using LCM can enrich the extracted sample for a particular cell type, as we have done in this chapter, it is probably not sufficient when comparing between cell types found packed so closely together. Using LCM may be appropriate for comparison of larger neurons that are not found in within the same tissue regions, such as Purkinje cells from the cerebellum and pyramidal neurons from the cerebral cortex, although these samples would still contain some portion of the smaller cell types located in close proximity such as glia and granule cells. There are additional methods for isolating cell types, but these also have similar limitations to LCM, such as: Translating Ribosome Affinity Purification (TRAP), Immunopanning (PAN), and Fluorescence Activated Cell Sorting (FACS) (Okaty, Sugino and Nelson 2011). In the future a more feasible approach for getting cell populations enriched for specific cell types for analysis may be to differentiate iPS cells into the cell type(s) of interest for study. However, currently there are many limitations in the protocols for differentiating iPS cells into the desired cell type, including heterogeneity of cell types, variability of iPS cell clones, consistently generating a large number of cells, and notably that the *in vitro* context may not be a true representation of the *in vivo* context (Santostefano *et al.* 2015). Even as protocols improve to overcome these iPS cell limitations it may not be practical to generate enough of these iPS cells lines such that it is possible to adequately represent common genetic variation so that a transcriptome-wide analysis could be performed. Using iPS would still be appropriate in a candidate based approach where you would be able to determine the number of subjects and the genotypes such that you could pick the subjects that are representative of the genetic variation for the gene or the *cis*-regulatory region of interest.

5: Application to disease GWAS loci

5.1: Introduction

The integration of eQTL and genome-wide association study (GWAS) loci has become a more common occurrence, where the drive is to begin to formulate possible hypotheses on how these loci confer risk for disease when the loci intersect. In studies examining the intersection of eQTL and GWAS results it has been suggested that eQTL are enriched at GWAS loci (Verlaan *et al.* 2009; Nica *et al.* 2010; Nicolae *et al.* 2010). Many studies suggesting how disease risk may arise from changes in expression associated with genetic variants have been published: for Crohn's disease (Libioulle *et al.* 2007), childhood asthma (Cantero-Recasens *et al.* 2010), lupus (Nica *et al.* 2010; Sakurai *et al.* 2013), Type 2 diabetes (Zhong *et al.* 2010), osteoarthritis (Syddall *et al.* 2013), and drug response for rheumatoid arthritis (Cui *et al.* 2013). Of course the appropriate study, replication, and validation of these hypotheses must take place. In some instances these are just intersections between an expression quantitative trait and disease risk locus, such as the one from a replicated GWAS locus for Crohn's disease. Examination of the Crohn's disease locus, considered a 1.25 megabase gene desert (large region lacking in genes) on chromosome 5 where the risk variants are also part of a *cis*-eQTL, for the nearest gene in the region, prostaglandin receptor EP4 (*PTGER4*) (Libioulle *et al.* 2007). Other studies have made efforts to look beyond the intersection of disease risk loci and eQTL by integrating existing public annotations. In 2010, Zhong *et al.* published a study integrating eQTL (in liver, and two adipose tissues), GWAS variants and pathway information from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000). The inclusion of the KEGG information made it possible to elucidate which pathways may play a role in disease. Based on pathways enriched for eQTL genes whose variants also intersected disease associated risk variants they identified 16 pathways that may play a role in Type 2 diabetes (Zhong *et al.* 2010). Other studies have considered how other expression related QTL might intersect with disease risk. In 2009, Fraiser and Xie published a study examining transcript isoform variation, which they labelled polymorphic transcript variation (PTV). This study was based on gene expression in human B cells from two populations. They found that tens of thousands of exons showed variation in expression levels that were heritable, and correlated with *cis*-variants (splicing QTL, sQTL). The *cis*-variants correlated with PTV were enriched for risk variants associated with four autoimmune diseases: Crohn's disease (CD), Type 1 diabetes (T1D), rheumatoid arthritis (RA), and ankylosing spondylitis (AS). B cells are known to have a role in autoimmune diseases. They suggest that for eight of the common risk variants of immune disease, that PTV may be the risk mechanism (Fraser and Xie 2009). In 2010, Nica *et al.* published a study on the integration eQTL and GWAS loci. They developed an algorithm for the integration of disease- and expression-associated loci that accounts for local LD structure in order to identify GWAS signals that may also be *cis*-eQTL, which they called Regulatory Trait Concordance (RTC). Applying their RTC method, they found an enrichment of *cis*-eQTL among GWAS SNPs. Their method confirmed prior eQTL disease mediated effects for *ORMDL3* and asthma, *C8orf13* and lupus, and *SLC22A25* and Crohn’s disease (Nica *et al.* 2010).

Beyond the intersection of eQTL and disease risk loci, other studies have begun to focus on possible mechanisms in identifying the risk variant and how the risk for disease is conferred through gene expression. In a follow-up of the *ORMDL3* eQTL associated with childhood asthma (Moffatt *et al.* 2007), it was found that the changes in *ORMDL3* expression result in a change of inflammatory response. This change in inflammatory response is by way of altered endoplasmic reticulum-mediated calcium signalling leading to an unfolded-protein response inducing inflammation (Cantero-Recasens *et al.* 2010). In a 2013 study, by Syddal *et al.*, of how osteoarthritis risk may be modulated by expression, the authors focused on a variant in the 5' UTR of growth differentiation factor (*GDF5*), a gene that is associated with increased risk of osteoarthritis in Europeans and Asian populations. The risk susceptibility is through decreased expression of *GDF5*. They found four *trans*-acting factors binding to the 5'UTR of *GDF5,* three of which repress expression via the osteoarthritis risk allele for a *cis*-regulatory variant in the binding site (Figure 5.1) (Syddall *et al.* 2013).

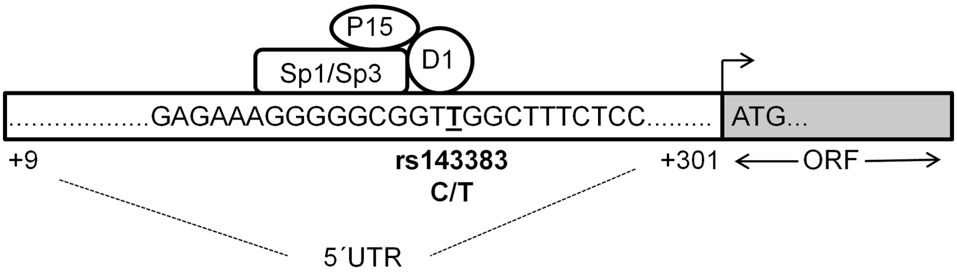


Figure 5.1: Schematic showing the allele in the *cis*-regulatory variant that changes a binding site in the 5' UTR of *GDF5*, for the factors *Sp1*, *Sp3*, *DEAF-1*, and *P15* complex, which results in repressing *GDF5* expression. This figure is reproduced from (Syddall *et al.* 2013).

In 2013, Sakurai *et al.* published a study to identify the potential causal risk variant for Systemic lupus erythematosus (SLE). In patients with SLE both mRNA and protein levels of interleukin-10 (*IL-10*) are elevated in sera. SLE is associated with a risk variant 9.2 Kb upstream of *IL10.* They found that preferential binding of the transcription factor Elk-1*,* at the *IL10* risk allele, increases expression of *IL-10* in SLE patients for both mRNA and protein (Figure 5.2). Sera levels of IL-10 are elevated in SLE patients and correlated with disease activity. Expression levels of both phosphorylated Elk-1 and IL-10 were elevated in SLE patient's cells (Sakurai *et al.* 2013).

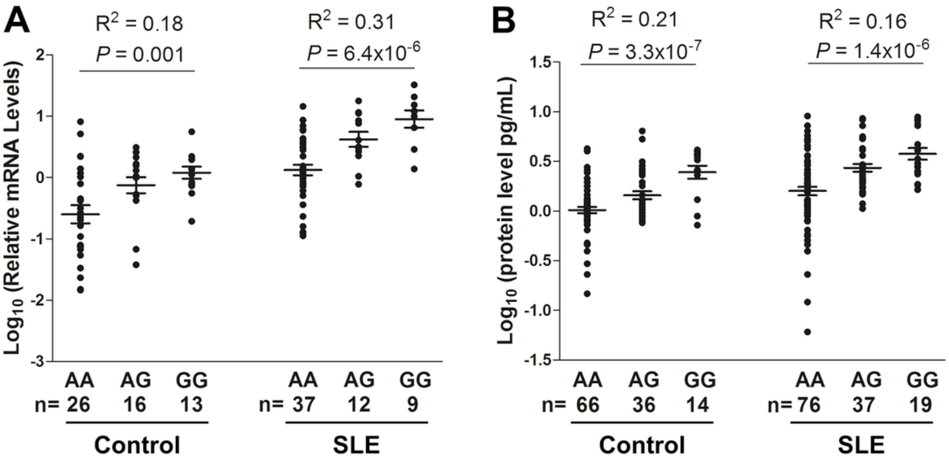


Figure 5.2: Correlation plots of *IL-10* mRNA expression levels (A) and protein levels (B), for controls and SLE patients, by genotype at the SLE risk variant, rs3122605. The risk allele is correlated with increased mRNA expression (A) and protein levels (B), of IL-10, by allele dose for the G allele in both patients and controls. Figure reproduced from (Sakurai *et al.* 2013).

In 2013, Cui *et al.* published a study of drug response for anti-TNF therapies (etanercept, infliximab and adalimumab) in patients with rheumatoid arthritis. The performed a meta-GWAS to identify a genetic basis for why some patients failed to have adequate drug response. It was found that a variant associated with change in disease activity, for etanercept patients but not the other two therapies, is predicted to disrupt a transcription factor binding-site in *CD84*, an immune-related gene. The allele associated with drug response also correlated with higher expression of *CD84* in blood. *CD84* expression correlates with disease activity score. Additionally, in a replication analysis, which included multiple ethnicities, the drug response for etanercept was significant only for European ancestry patients (Cui *et al.* 2013). These studies provide examples for the formation of a disease aetiology hypothesis where disease risk loci and eQTL intersect and in some instances begin to provide evidence related to these hypotheses.

5.2: Examples in Neurological diseases

An important reason for the execution of the work presented in my thesis and the creation of this data resource is to allow us and others to understand how loci associated with neurological diseases, we work on, may have an effect on gene expression in human brain. To this end, we have referred to these NABEC eQTL data in the study of GWAS related to neurological diseases. These studies include Progressive Supranuclear Palsy (PSP) at loci near the *SLC25A38*/*MOBP* and the *MAPT* H1/H2 inversion polymorphism region (Höglinger *et al.* 2011) (Figure 5.3a-c). These data have also been used to investigate loci associated with Tourette's Syndrome (Scharf *et al.* 2012), Obsessive-Compulsive Disorder (Stewart *et al.* 2012), amyotrophic lateral sclerosis (ALS) (Traynor *et al.* 2010), frontotemporal lobar degeneration (FTLD) (Carrasquillo *et al.* 2010) and Alzheimer's disease (Guerreiro *et al.* 2010; Holton *et al.* 2013).

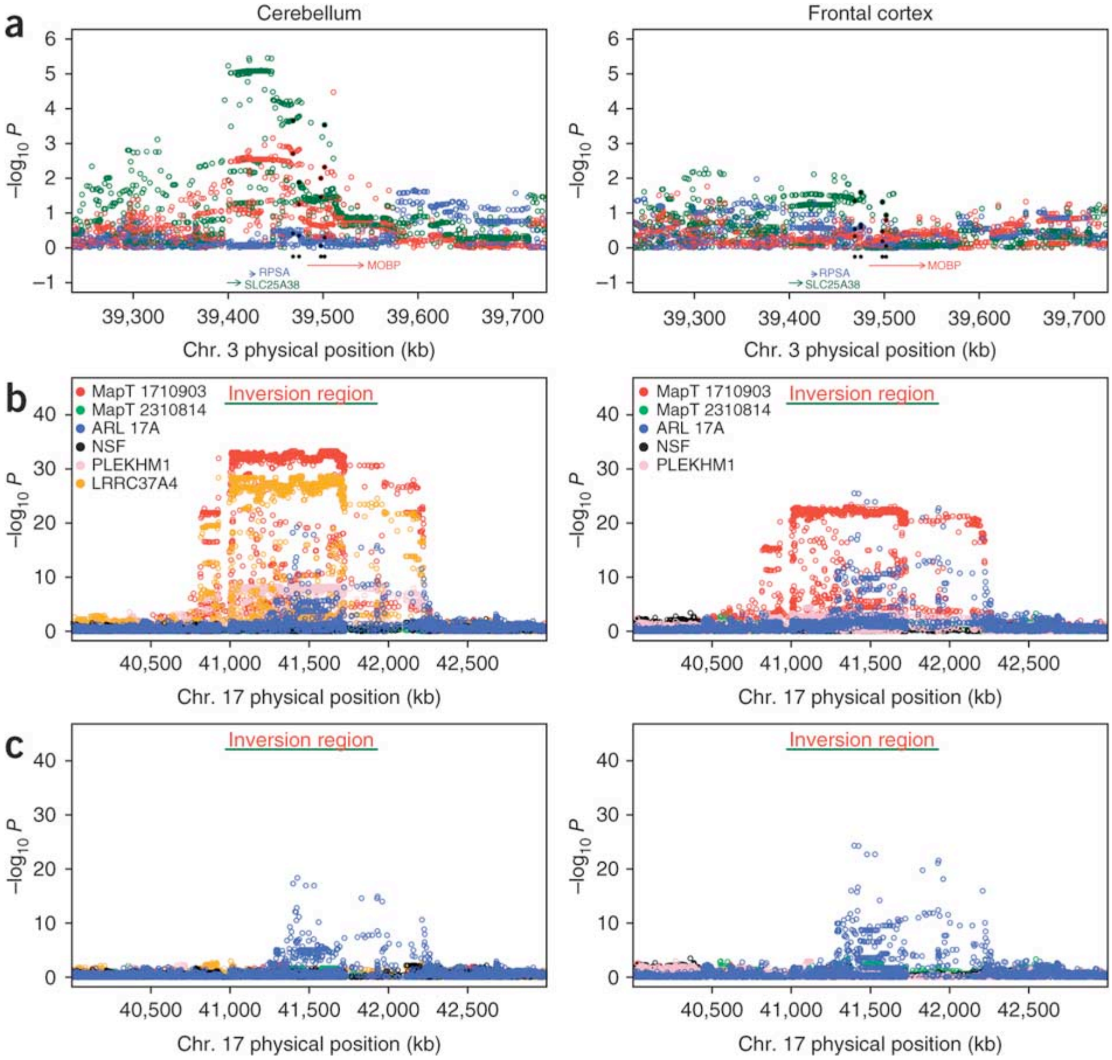


Figure 5.3: Manhattan plots for PSP associated regions that also have significant eQTL. (a) eQTL results for the *SLC25A38-MOBP* region, on chromosome 3, shown are the eQTL p-values per SNP for three transcripts; *MOBP* is red, *RPSA* is blue, and *SLC25A38* is green. (b) eQTL results for the H1/H2 inversion polymorphism region near *MAPT,* on chromosome 17; eQTL p-values colour coded by transcript probe provided in legend. (c) eQTL results for the for the H1/H2 inversion polymorphism region, on chromosome 17 near *MAPT*, controlling for H1/H2 by conditional analysis based on variant that tags the haplotypes. After controlling for the haplotype only an eQTL for ARL17A remains. Plots in panel (c) use the same colour legend as panel (b). Each plot in the left panel are results from the cerebellum and plots in the right panel are from cerebral frontal cortex. The colour of each data point is colour coded per transcript and represents the p-value for a SNP allele dose correlation with a transcript's expression level. Each SNP is tested against multiple cis transcripts. This figure is reproduced from (Höglinger *et al.* 2011)

While the proximity of an eQTL to a GWAS loci does not imply that the disease risk and expression effect are the same loci, in a few clear instances we have shown that the eQTL effect signal intersects with a GWAS signal, and at the very least is a potential candidate for the biologic effect. An example of a significant eQTL overlapping with a significant locus is one from a meta-GWAS for Migraine (Anttila *et al.* 2013) that included eQTL data based on combined NABEC and UK Brain Expression Consortium (UKBEC) data in cerebellum and cerebral frontal cortex. In this study the eQTL overlapped with the GWAS loci with moderate LD between the eQTL and GWAS loci, but one locus was in perfect LD. The eQTL in perfect LD with the GWAS locus is for a transcript from the gene *STAT6*, signal transducer and activator of transcription 6, interleukin-4 induced. STAT6 phosphorylation has been shown to lead to prostaglandin release as a result of astrocyte response to oxidative stress (Park *et al.* 2012), and in macrophages has been shown that transcription factor activation signals are transduced by STAT family members (Lawrence and Natoli 2011). Additionally, we have made great use of the NABEC/UKBEC eQTL data within our GWAS studies of Parkinson's disease (PD). Beginning with an early GWAS of PD (Simón-Sánchez *et al.* 2009) we were able to intersect the signals of our disease risk alleles with those of an eQTL for a transcript of *MAPT* showing an increase in *MAPT* abundance with risk allele dosage (Figure 5.4).

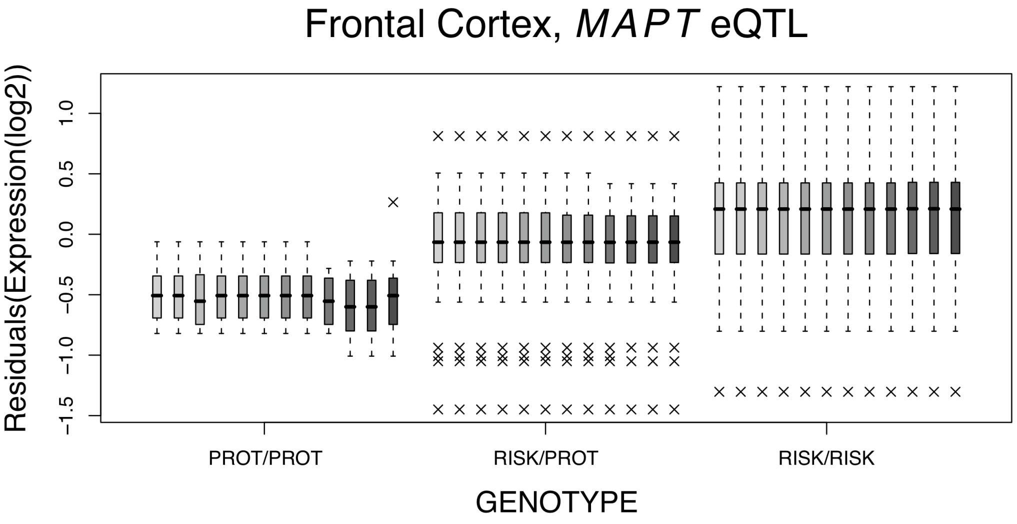


Figure 5.4: Shown are boxplots for several SNPs associated with Parkinson's Disease (Simón-Sánchez *et al.* 2009) that also are significant eQTL for a *MAPT* transcript in the cerebral frontal cortex. The figure shows an increase in the abundance of the *MAPT* transcript with dosage of the risk allele.

It has since been shown that our original finding of an eQTL for the *MAPT* transcript is likely an artefact of a polymorphism within the Illumina 50-mer probe for this transcript. While I have consistently screened our probes for this type of artefact, I have done so using allele frequencies within populations of European descent, where available. Unfortunately, the variant within the mRNA probe design is for a 2 base pair InDel present on the H2 haplotype for this region. The earlier screenings performed for this type of artefact were based on HapMap SNP data for the European population and this InDel variant was not present in that dataset, at that time, with an allele frequency in the European population. This type of screening has vastly improved with the release of the 1000 Genomes genotypes (1000 Genomes Project Consortium *et al.* 2012) based on sequencing data, where it is now possible to screen from this resource, in the appropriate population, at the appropriate allele frequency that includes both single nucleotide and InDel variants.

In a more recent analysis of PD risk loci based on results from the International Parkinson's Disease Genomics Consortium (IPDGC) we have continued to make use of the NABEC/UKBEC eQTL data showing overlap between disease risk and eQTL at multiple risk loci (Nalls *et al.* 2011, 2013, 2014). Based on the most significant variants for the 26 loci associated with PD, from the most recent meta-GWAS (Nalls *et al.* 2014), the NABEC/UKBEC eQTL results from cerebellum and cerebral frontal cortex were scanned based on these PD risk variants to see if they were also part of eQTL. In performing the scan I limited the search space to just the *cis* possible eQTL tests that would involve these PD risk variants and only the transcripts well detected (detected in 95% of samples) in the NABEC/UKBEC series that were within 1 Mb of these risk variants. Reducing the eQTL search space in this simple manner, only including eQTL tests which included the PD risk variants, also reduces the multiple test burden therefore increasing the sensitivity to detect signal, but of course this also reduces specificity. Using this PD risk and eQTL intersection search scheme results in approximately 360 independent eQTL tests per tissue, yielding a significance threshold of $1.4\text{x}10^{-4}$ based on a Bonferroni multiple test correction to maintain a 5% false positive rate. This search identified three PD risk loci that were also correlated with expression changes for five genes (six transcripts) in one or both brain tissues. Two of these eQTL are for pleckstrin homology domain containing family M (with RUN domain) member 1 (*PLEKHM1*) and leucine rich repeat containing 37, member A4 (pseudogene) (*LRRC37A4)* in the *MAPT* region of chromosome 17. Both, *PLEKHM1* and *LRRC37A4*, are correlated with the same PD risk variant rs17649553 and significant only in the cerebellum; these two eQTL were also correlated with the H1/H2 *MAPT* haplotype, denoting the large inversion polymorphism present in that genomic region. Three other genes with an eQTL, present in human brain, intersect with two PD risk loci. The first of these risk loci is on chromosome 1, in the region of nuclear casein kinase and cyclin-dependent kinase substrate 1 (*NUCKS1*) and RAB7 member RAS oncogene family-like 1 (*RAB7L1*), where the most significant (p-value = 1.36x10^{-13}) risk variant from the PD meta-GWAS is rs823118 (Nalls *et al.* 2014). This PD risk variant is located 1.35 Kb downstream of the 3'UTR of *RAB7L1* and 4.2 Kb upstream of the 5'UTR of *NUCKS1* and is correlated with expression changes in both genes. While the eQTL for *NUCKS1* shows significant signal in both cerebellum and cerebral frontal cortex, the *RAB7L1* eQTL is only significant in cerebellum (Figure 5.5).

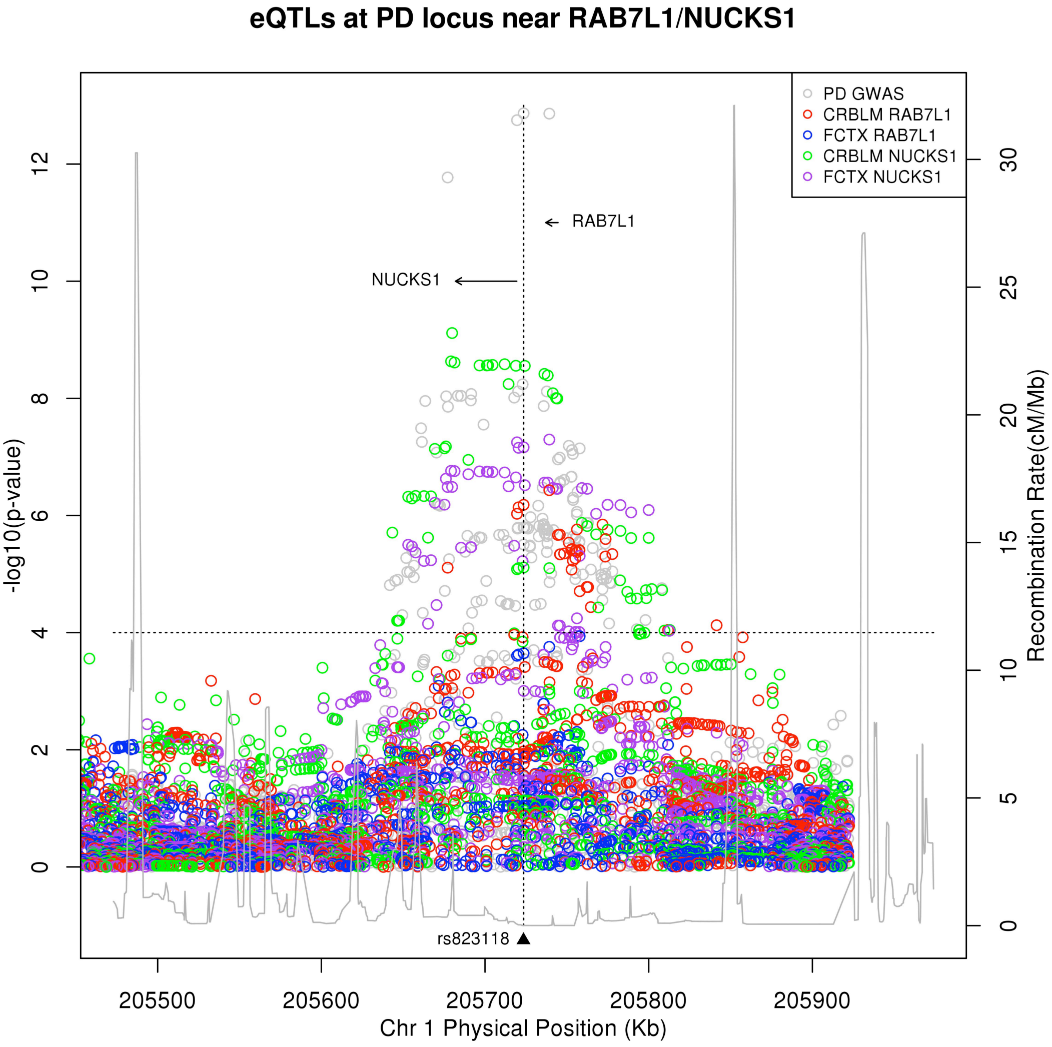


Figure 5.5: Manhattan plot of region, on chromosome 1, across the GWAS associated PD risk locus spanning *NUCKS1* and *RAB7L1*. The plot shows association p-values representing the significance of correlation between a variant's allele dosage and the transcript's expression level. The points are colour coded by transcript and tissue region in red, blue, green, and purple while the grey data points represent the p-values from the meta-GWAS for PD. The plot also includes the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. The threshold for significance is denoted by horizontal dotted line. The relative positions of the genes are the labelled arrows centred in the upper portion of the plot. The direction of the arrow is the gene's strand. The most significant PD risk variant, rs823118, is labelled near the bottom of the plot, at its chromosomal position, and the vertical dashed line.

It has recently been shown that the protein RAB7L1 interacts with the protein leucine-rich repeat kinase 2 (LRRK2) (MacLeod *et al.* 2013; Beilina *et al.* 2014), *LRRK2* is a gene harbouring Mendelian mutations that cause PD (Paisán-Ruíz *et al.* 2004; Nichols *et al.* 2005) as well as a risk locus associated with sporadic PD (Nalls *et al.* 2014). *LRRK2* encodes a multidomain protein with GTPase (enzymes that hydrolyze guanosine triphosphate) and kinase activities, and has been shown to be involved in macroautophagy-lysosomal protein degradation and Golgi apparatus integrity (MacLeod *et al.* 2006; Heo, Kim and Seol 2010; Dodson *et al.* 2012; Stafa *et al.* 2012). Current findings, from MacLeod *et al.*, suggest that the retromer and lysosomal pathways are important in PD pathogenesis. They show that a deficiency of *RAB7L1* gene expression, in rodent primary neurons and its ortholog in fly dopamine neurons, recapitulates some of the degenerative phenotype observed in expression models for a familial PD mutation in *LRRK*2. However, they also find that overexpression of *RAB7L1* rescued the mutant *LRRK2* phenotype. Additionally, both the *RAB7L1* PD risk and *LRRK2* PD linked mutation resulted in endosomal and Golgi sorting defects; and affected the *VPS35* component of the retromer complex. They suggest the potential causal risk variant may be a variant in LD with a PD risk variant that results in the alternative splicing of *RAB7L1,* where the risk allele leads to increased skipping of exon 2 (MacLeod *et al.* 2013). In another study, from Beilina *et al.*, it was shown that RAB7L1 is a protein-binding partner of LRRK2. This binding interaction was identified using an unbiased screen from protein-protein interaction arrays. Additionally, the genes BCL2-associated athanogene 5 (*BAG5*) and cyclin G associated kinase (*GAK*) encode proteins that are also part of this protein complex; *GAK* is also a PD associated locus. These protein interactions were validated in cell lines and in mouse brain. The authors' experiments suggest that the proteins encoded by *LRRK2*, *RAB7L1*, *GAK*, *BAG5*, and heat shock 70kDa protein 4 (*Hsp70*) form a single protein complex. Examining the cellular localization of these proteins the authors found that both RAB7L1 and GAK are largely vesicular in neuron localization, which suggest that RAB7L1 directs Lrrk2 to *trans*-Golgi network derived vesicles. Based on these results, the authors suggest that these proteins form a complex that promotes clearance of Golgi-derived vesicles through the autophagy-lysosome system (Beilina *et al.* 2014).

The second PD risk locus that intersects with an eQTL, in the brain, is on chromosome 7 and located in an intron of the glycoprotein transmembrane nmb gene (*GPNMB*), where the most significant (p-value = 2.37×10^{-12}) risk variant from the PD meta-GWAS is rs199347 (Nalls *et al.* 2014). This PD risk variant is correlated with changes in gene expression for two transcripts of nucleoporin like 2 (*NUPL2*); rs199347 is located in the intron of *GPNMB* and is 53.1 Kb downstream of the 3'UTR of *NUPL2*. Both *NUPL2* transcripts have a significant eQTL in both cerebellum and cerebral frontal cortex (Figure 5.6). Functionally, a hypothesis may be formed that the effects of an expression change in *NUPL2* may have an affect on the LRRK2 protein binding complex described by Beilina *et al.*, in reference to the RAB7L1 and LRRK2 protein-protein interaction. *NUPL2*, previously known as *hCG1*, is part of the nuclear pore complex (NPC) and required for export of mRNA from the nucleus to the cytoplasm. Based on a previous finding, from Kendirgi *et al.*, NUPL2 protein may be required for the export of *Hsp70* mRNA from the nucleus to the cytoplasm. This finding was based on experiments using a small interfering RNA (siRNA) knockdown of *hCG1* (*NUPL2*) mRNA that resulted in decreased Hsp70 protein levels, under heat shock conditions in HeLa cells. The decreased protein levels resulted from *Hsp70* mRNA not being exported from the nucleus to the cytoplasm because of the reduction in NUPL2 protein levels (Kendirgi *et al.* 2005). It is known that Hsp70 protein is important in removing clathrin from vesicles, including at the *trans*-Golgi. One way to connect *NUPL2* and *LRRK2*, a known PD gene, and other gene candidates from PD GWAS: *BAG3/5*, *Hsp70*, *GAK*, and *RAB7L1* is to assume that there is a simultaneous protein complex at the *trans*-Golgi that removes clathrin. Increasing protein levels of any one of these would promote protein complex formation and, hence, function. In this model, more NUPL2 protein would mean more cytosolic Hsp70 protein resulting in more LRRK2 protein complex function, exacerbated by stimulation. Of course, this hypothesis would require many experiments to support its claims but they are testable ideas. Knockdown of *NUPL2* could diminish LRRK2 relocalization to the *trans*-Golgi. Additionally, this hypothesis requires a few assumptions, including that promoting the formation of the LRRK2 protein complex alters the removal of clathrin from vesicles or the clearance of Golgi-derived vesicles through the autophagy-lysosome pathway and that this alteration in function is deleterious and involved in the disease pathway of PD. (Note: Hypotheses of how *NUPL2* may be involved in PD through HSP70 and LRRK2 are based on personal communications with Mark Cookson and Andrew Singleton).

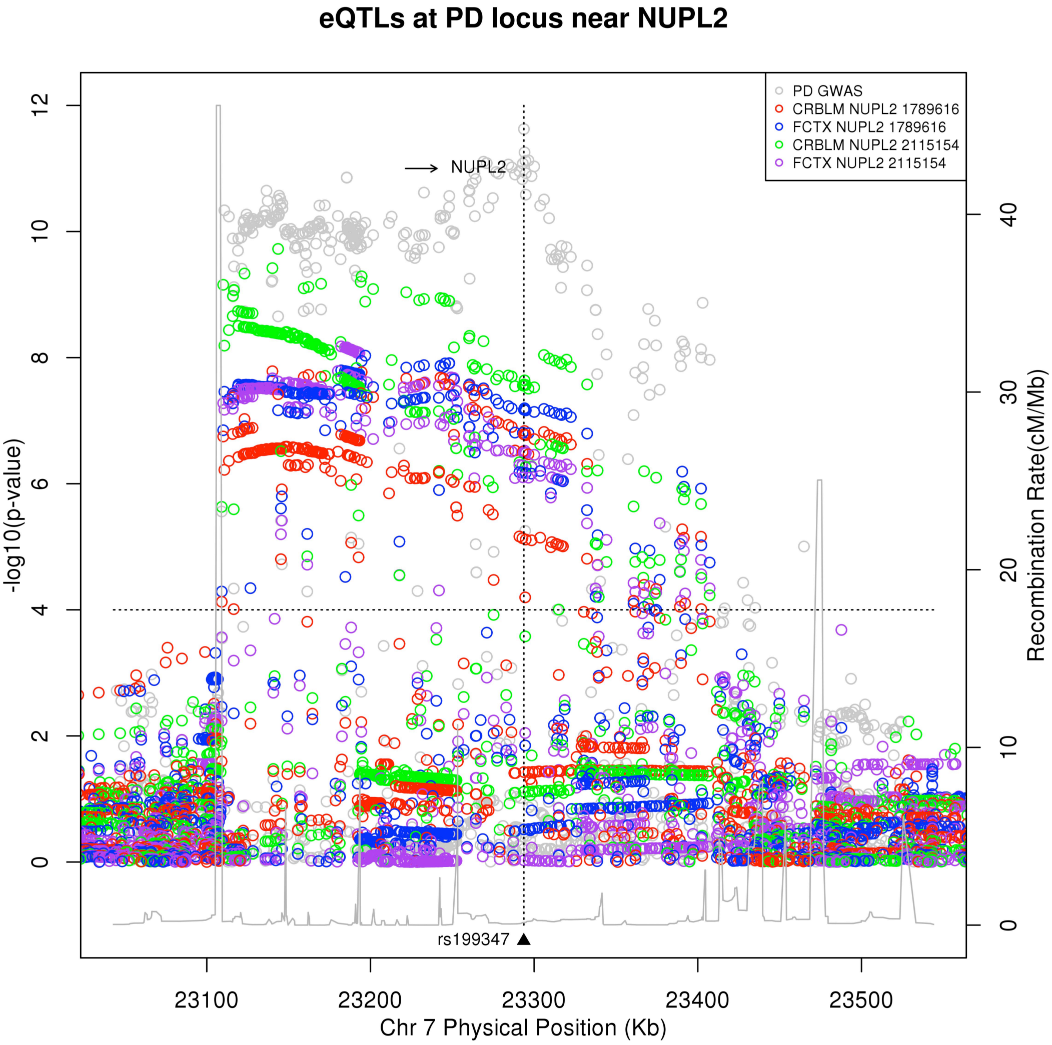


Figure 5.6: Manhattan plot of region, on chromosome 7, across the GWAS associated PD risk locus spanning *NUPL2*. The plot shows association p-values representing the significance of correlation between a variant's allele dosage and the transcript's expression level. The points are colour coded by transcript and tissue region in red, blue, green, and purple while the grey data points represent the p-values from the meta-GWAS for PD. The plot also includes the recombination rates (right axis) as a dark grey continuous line based on HapMap III data. The threshold for significance is denoted by horizontal dotted line. The relative positions of the gene are the labelled arrow centred in the upper portion of the plot. The direction of the arrow is the gene's strand. The most significant PD risk variant, rs199347, is labelled near the bottom of the plot, at its chromosomal position, and by the vertical dotted line.

5.2: Other GWAS loci

To access the prevalence of other disease- and complex trait-associated variants with eQTL variants in brain, a search of variants from the NHGRI-EBI GWAS catalogue (Welter *et al.* 2014) was performed against the

NABEC/UKBEC eQTL results from cerebellum and cerebral frontal cortex. The NHGRI-EBI GWAS catalogue is a repository provided by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI), which maintains a manually curated, and quality controlled collection of published genome-wide association studies. For this search, the catalogue was downloaded on 4 April 2015 and included 15,653 variants for 1,276 diseases or traits from 2,140 published studies. Based on the *cis*-tested variants from the NABEC/UKBEC eQTL analysis, 12,888 of the NHGRI-EBI GWAS catalogue variants were tested against one or more of 9,374 transcript expression probes in one or both brain tissues. This testable set results in approximately 119,000 independent eQTL tests per tissue yielding a significance threshold of $4.7x10^{-7}$ when applying Bonferroni based multiple test correction to maintain a 5% false positive rate. At this threshold of significance, 320 (2.5%) variants are associated with 189 (2.0%) expression traits for 201 (15.8%) disease or complex traits from 247 (11.5%) studies in the NHGRI-EBI GWAS catalogue. These variants were then annotated so the distribution of variant types relative to transcription, could be considered, based on the following annotations: exon, intron, UTR, intergenic, upstream, and downstream. Upstream and downstream regions included variants that lie within 5 Kb of the 5' and 3' UTR respectively for these annotation purposes. For the variants that were both GWAS risk variants and part of an eQTL, when compared to all GWAS risk variants that did not show a significant eQTL, there were shifts within the distributions of variant types. For the variants that were both GWAS risk and eQTL variants there are increases, when compared to GWAS risk variants that were not eQTL variants, in the percentage of variants located upstream (+7.2%), exon

(+1.1%), 3’ UTR (+0.7%) and downstream (+8.8%) (Figure 5.7). Conversely, this same variant set also showed a decrease in the number of variants present in introns (-12.7%) and intergenic (-5.3%) regions while the 5’ UTR remained relatively unchanged. When considering which of these regions may be functionally active regulatory regions, the variants were annotated based on ENCODE DNase I hypersensitivity sites (DHS) and transcription factor binding site (TFBS) clusters (ENCODE Project Consortium 2012). These two annotations should denote most of the active sites where transcription factors bind based on the diverse set of cell and tissue types used in the generation of these data. The DHS annotation data includes ~2.9 million DHSs based on 125 cell types including differentiated primary cells (56.8%), immortalized primary cells (12.8%), malignancy-derived cell lines (24.0%) and multipotent and pluripotent progenitor cells (6.4%) (Thurman *et al.* 2012). The TFBS annotation data is based on 161 transcription factors and 91 cell types (Gerstein *et al.* 2012; Wang *et al.* 2012). It is estimated that between 8.5% and 19.4% of the human genome is covered by a DHS footprint or TFBS motif and that 94.4% of transcription factor occupancy sites are within DHS footprints (ENCODE Project Consortium 2012; Thurman *et al.* 2012; Kellis *et al.* 2014). For GWAS risk variants that were also part of eQTL, these are enriched for both DHS and TFBS clusters with both having ~10% increases in regulatory active regions when compared to GWAS risk variants that do not show a significant eQTL signal (Figure 5.8).

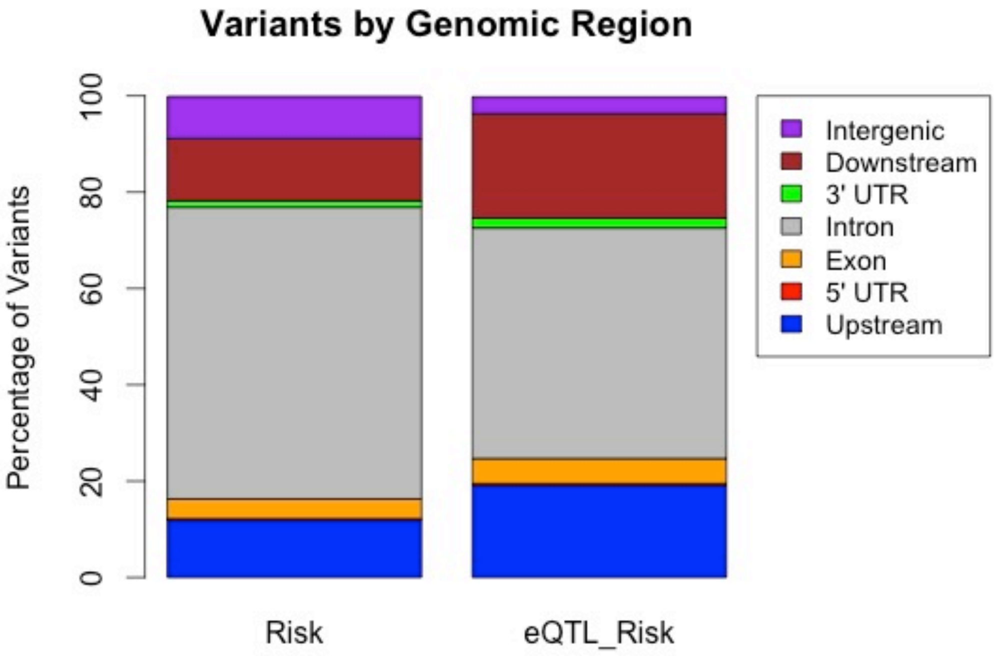


Figure 5.7: Bar graphs showing the distribution of variants by their genomic location relative to transcription for all NHGRI-EBI GWAS risk variants that were tested with and without a significant a *cis*-eQTL in the combined NABEC and UKBEC analysis in cerebellum and cerebral frontal cortex. The left column (Risk) represents all risk variants tested and the right column (eQTL_Risk) represents all risk variants that are also significant as a *cis*-eQTL variant.

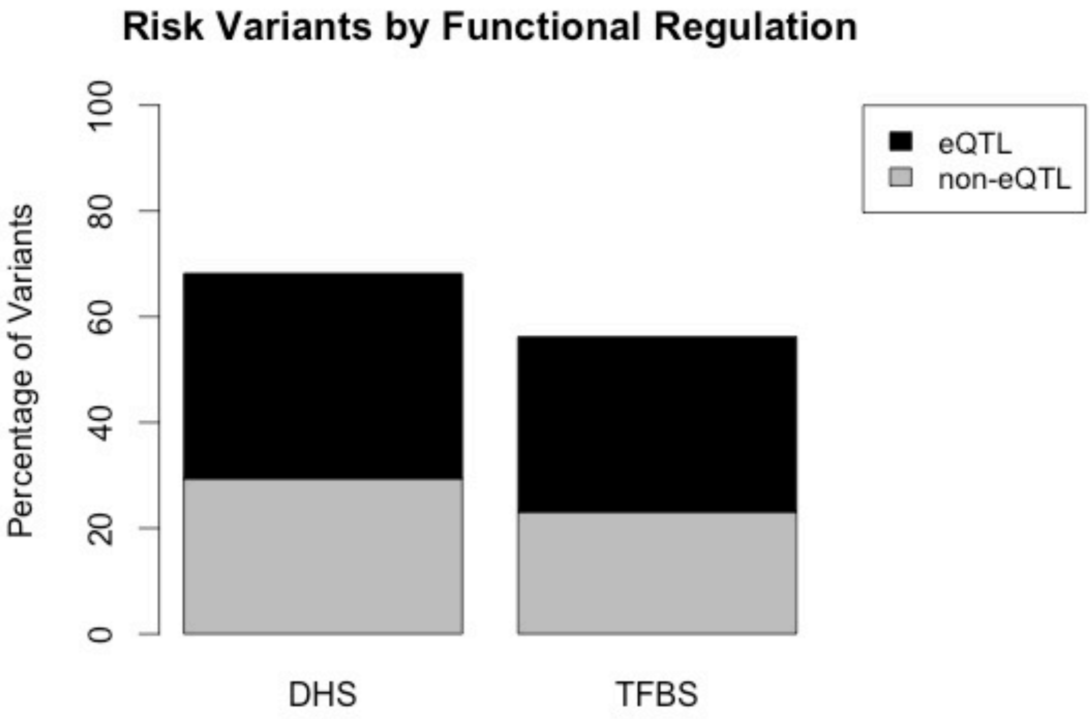


Figure 5.8: Bar graphs showing the distribution of variants by there genomic location relative to functionally active regulatory regions based on ENCODE v3 annotations for all NHGRI-EBI GWAS risk variants that were tested with and without a significant cis-eQTL in the combined NABEC and UKBEC analysis of cerebellum and cerebral frontal cortex. In the bars for both DHS (DNase I hypersensitivity sites) and TFBS (transcription factor binding sites) the grey colour represents risk variants that are not significant as an cis-eQTL variant and the black colour represents all risk variants that are also significant as a *cis*-eQTL variant.

5.3: Discussion

In this chapter I have highlighted applications of eQTL to disease risk loci performed by others: Crohn's disease (Libioulle *et al.* 2007; Nica *et al.* 2010), autoimmune diseases (Fraser and Xie 2009), Type 2 diabetes (Zhong *et al.* 2010), asthma (Cantero-Recasens *et al.* 2010; Nica *et al.* 2010), and lupus (Nica *et al.* 2010). I have also highlighted a few studies where not only were eQTL being applied to disease risk variants but some work has been done to begin to understand disease mechanisms: inflammatory response in childhood asthma (Cantero-Recasens *et al.* 2010), osteoarthritis risk altering a regulatory binding site (Syddall *et al.* 2013), Systemic lupus erythematosus and *IL10* (Sakurai *et al.* 2013), and response to anti-TNF therapies (Sakurai *et al.* 2013). These studies have made use of eQTL information to begin to understand the underlying molecular mechanisms linking a genetic locus to disease or complex trait resulting from changes in expression. I have also highlighted several studies where we have made use of the NABEC and UKBEC brain eQTL data to investigate neurological diseases: Progressive Supranuclear Palsy (Höglinger *et al.* 2011), Tourette's Syndrome (Scharf *et al.* 2012), Obsessive-Compulsive Disorder (Stewart *et al.* 2012), amyotrophic lateral sclerosis (ALS) (Traynor *et al.* 2010), frontotemporal lobar degeneration (FTLD) (Carrasquillo *et al.* 2010), Alzheimer's disease (Guerreiro *et al.* 2010; Holton *et al.* 2013), Migraine (Anttila *et al.* 2013), and Parkinson's disease (Simón-Sánchez *et al.* 2009; Nalls *et al.* 2011, 2014). Additionally, I have used the NABEC and UKBEC brain eQTL variants to investigate all disease risk variants maintained by the NHGRI-EBI GWAS catalogue. Scanning the brain eQTL results based on more than 15,000 risk variants for more that 1,200 diseases or traits, 15% of these diseases or complex traits had an associated risk variant that was also an eQTL variant. In evaluating the risk variants that were also eQTL variants compared to all risk variants considered, the variants that were both risk and eQTL variants show an increase for being located within 5 Kb up and downstream of a transcript, and a decrease in intronic and intergenic variants. Additionally, the variants that were both risk and eQTL variants showed an increase in being located in regions of DHS and TFBS occupancy compared to the variants that were not significant as an eQTL. This enrichment of GWAS risk variants that are also eQTL variants in active regulatory elements may support the hypothesis that the functional effect of the risk is mediated through changes in expression.

Many of the above applications of eQTL in the analysis of risk variants are fairly rudimentary, typically based on the actual risk variant being present in an eQTL variant set or in LD with a variant in the eQTL variant set. However, methods are being developed to integrate disease risk and eQTL variants beyond a simple intersection of the genetic loci. In 2014, Corradin *et al.* published a method for defining enhancer-gene interactions in relation to the multiple enhancer variant hypothesis for common traits. Under this hypothesis a set of variants in LD affects multiple enhancers resulting in a cooperative affect on gene expression. To provide evidence of this model they considered GWAS results from six common autoimmune disorders based on data from HapMap B lymphoblasts. They found some evidence to support this hypothesis, but the effects on gene expression were modest (Corradin *et al.* 2014). In 2012, Schaub *et al.* published a method using ENCODE functional data to investigate disease associated variants. The method's purpose was to identify a functional variant as the actual risk variant for the disease association signal. They found a significant enrichment of variants within regulatory elements that are also associated with diseases. The strength of this enrichment increases when multiple functional sources and higher confidence disease associated variants are used (Schaub *et al.* 2012). In 2011, Lappalainen *et al.* published a study proposing a hypothesis suggesting that *cis*-eQTL may modulate the penetrance of deleterious protein coding variants. The hypothesis is that heterozygous deleterious coding variants and heterozygous *cis*-regulatory variants have an epistatic interaction, where if the deleterious coding allele is on the higher expressing haplotype of the *cis*-regulatory allele the penetrance of the coding variant increases. They analysed genotype and mRNA expression data generated from both sequencing and arrays from CEU and Yoruban subjects from the 1000 Genomes project to begin to test this hypothesis. They found an underrepresentation of functional coding variants on the higher expressed regulatory haplotypes suggesting purifying selection against these deleterious coding alleles, through their regulatory backgrounds. They also found that the allele frequency distributions of the eQTL alleles might support their suggestion of purifying selection. They found that LD between eQTL variants and nonsynonymous coding variants was stronger than LD between eQTL variants and synonymous coding variants. Additionally, they found that eQTL signals that intersect with GWAS signals show an enrichment for this type of putative epistatic interaction, suggesting that rare coding variants may attain higher penetrance through *cis*-regulatory eQTL (Lappalainen *et al.* 2011). In 2014, Pickrell published a report describing the integration of annotation information applied to GWAS data for 18 human traits. The model included 450 genomic annotations to model if these elements are enriched or depleted for loci associated with GWAS traits. They found that between 2% and 20% of trait associated SNPs influence protein sequence. They also found that repressed chromatin was depleted for several traits and that cell-type specific DNase-I hypersensitive sites were enriched for several traits. The integration of their annotation model as weights into GWAS analysis increased the number of high-confidence associations by ~5% (Pickrell 2014).

It is important to re-iterate that when using eQTL data to inform on a possible disease mechanism, when these loci intersect, that this information is only sufficient for forming hypotheses about disease mechanism(s). The appropriate study designs still need to be formulated and carried out to test these hypotheses. It is also important to note that while the large effect *cis*-eQTL typically replicate across many studies and tissue types, often many of these have not been validated using a different assay type to quantify the mRNA expression measures such as quantitative real-time PCR. The presence of the eQTL at a disease risk locus does not mean that the disease risk is conferred through expression changes in that transcript. Conversely the absence of an eQTL does not mean that the disease risk is not mediated through changes in gene expression. As eQTL can be tissue-specific, an eQTL may be present in one tissue while being absent in another that may be the disease relevant tissue. Additionally, it is not uncommon for multiple transcripts, from different genes, to be associated with the same *cis* genetic locus. Both of these instances occur at the PD risk locus near *RAB7L1*, one of the PD loci I discussed earlier in this chapter. At this PD locus there is a significant eQTL for *RAB7L1* in cerebellum but a significant eQTL for this transcript is not present in cerebral frontal cortex. This PD locus also has a significant eQTL in both cerebellum and cerebral frontal cortex for a transcript of *NUCKS1*. It is highly suggestive that *RAB7L1* may be the relevant gene for PD risk at this locus on the basis that RAB7L1 is a protein binding partner of LRRK2, a known Mendelian and risk gene for PD. The empirical evidence still needs to be gathered to link the genetic risk near *RAB7L1* to the gene *RAB7L1* and likewise evidence needs to be gathered to confirm the *RAB7L1* eQTL in a disease relevant cell type, such as dopaminergic neurons, as well as evidence to evaluate whether or not the disease risk is mediated through changes in expression of this gene.

6: Conclusions

As I have shown in these studies that make up my thesis, eQTL are manifested in human brain tissues. I have also demonstrated how these eQTL can begin to be applied in our understanding of disease risk variants. Many of the eQTL signals I have identified are shared between distinct tissue regions and even with an individual cell type, however there are also many eQTL that appear to be distinct among tissue regions and cell types. I have shown an example of a gene, *CHURC1*, which displays very strong eQTL signal across multiple neuronal tissues as well as in studies on non-neuronal tissue performed by others. I have also provided examples of eQTL that appear to be specific to a human brain region (*PPAPDC1A*) or distinct to an individual neuronal cell type (*CCZ1B*). Given our understanding of the regulation of gene expression it is not unrealistic to expect that eQTL may be identifiable for many RNA transcripts in many cell types. The identification of these eQTL would require a cohort size with sufficient power so that both common and less common variants could be included in the eQTL analysis as well as expression measures from a diverse set of tissue and cell types. The genetic variation from these eQTL are likely have a direct *cis*-regulatory affect on mRNA expression by altering the promoters or enhancers, splice site enhancers or suppressors, or miRNA binding sites within the 3’ UTR. With a sufficient increase in power it may also become possible to start reliably detecting *trans* effects. *Trans*-regulatory effects are biologically feasible through genetic variation that affects the expression of regulatory proteins, changes to the DNA binding domain within regulatory proteins affecting which promoters and enhancers they bind to, or genetic variation in a regulatory protein’s protein-protein interaction domain possibly affecting how or when regulatory proteins assemble at a gene's promoter and enhancer regions. However, this does not imply that all identified eQTL would be of the same importance within each cellular context as the relative effect within individual cell types may be much stronger or weaker. A possible example of this is the significant eQTL I have identified for a mRNA transcript from the gene *ALDH3A2,* which has a much stronger signal within Purkinje cells than the suggestive signal observed in the heterogeneous cell populations from the cerebellum or cerebral frontal cortex. I have also shown a Purkinje cell-specific example, *CCZ1B,* which has a significant eQTL in Purkinje cells but none at all within the bulk tissue regions. However, it is also the case that the same mRNA transcript may have significant eQTL in multiple cell types or tissue regions but that the regions of genetic variation exists in distinct or overlapping blocks of variation and therefore does not necessarily represent the same eQTL signal. It is feasible in this scenario that portions of this variation may be more impactful within certain specific cellular contexts than in others, and this in turn may reflect the cell-type specific expression of regulatory factors such as transcription binding factors. Some studies have made efforts at better resolving the variation within a locus that appears to be cell and tissue-dependent, but these studies have typically included highly differentiated tissue types such as blood and brain (Heinzen *et al.* 2008; Dimas *et al.* 2009; Kwan *et al.* 2009; Fu *et al.* 2012; Hernandez *et al.* 2012). This reinforces the importance of undertaking eQTL studies ultimately within individual cell types. Performing eQTL studies within specific cell types informs us as to the effect of genetic variation on gene expression within the context of that cell type.

My thesis shows that ultimately performing these types of studies in all individual cell types would provide a more granular resolution of eQTL. This does not discount the applicability of performing these types of studies in heterogeneous tissues as well, which is currently a more feasible approach. Each layer of functional knowledge we develop is informative to the next. This is not only important for our general knowledge and understanding of biology but creates a basis for hypothesis generation when applied to phenotypes affecting cellular and molecular processes related to human health and disease. As a researcher in the field of neurodegenerative disease and disorders, I have performed and continue to perform these types of analysis in human brain tissues so that we will have a functional foundation that provides clues into the aetiology of the diseases that we study. A large public consortium is currently underway to greatly add to the already public eQTL datasets and should open many more avenues for the discovery and characterization of eQTL as well as the application of this information in understanding disease. The Genotype-Tissue Expression (GTEx) project is a large public resource database established so that investigators can study the relationship between genetic variation and gene expression in many human tissues (GTEx Consortium 2013). The GTEx project aims to have genotype and expression data for 47 tissues from 900 subjects by the end of 2015. The four-tissue eQTL project described in Chapter 3 was an early dataset included in an early GTEx repository and browser for software pilot testing, which was available to access through the NCBI eQTL browser.

In viewing this doctoral work, the efforts here reflect the evolution of the modern eQTL field. Our early work centred on the feasibility of eQTL in human brain tissue, and showed clearly that there was much to discover in this regard. We moved on to more refined experiments, which attempted to look at varied tissues, in a larger set of samples, and aimed to answer questions regarding the importance of examining cell type specific QTL. While this work has mirrored the considerable progress of the field there is still much to be done. These projects have generated vast amounts of data in relation to expressed transcripts in human neurological tissues and will no doubt provide many data mining opportunities in the years to come. Not all of those mining analysis would be appropriate for inclusion within an individual graduate research thesis. As such I would consider the completion of my graduate thesis of expression quantitative traits in human neurological tissue to be logically complete after the detection and characterization of the eQTL in human brain. I believe the primary detection of these eQTL is fairly complete, in the existing cohort, and described in my thesis and that the bulk of the work still remaining is within the characterization of these eQTL. Additionally much work remains to be done in using more integrative approaches in applying eQTL, and other functional, information towards understanding disease risk variants. The work of characterization should contain much of the following future work.

7: Future Work

Moving beyond the simple QTL analysis, there are several spaces where additional work would be useful. It is likely to be informative to expand analysis from genotypes to haplotypes. In addition to providing haplotype block-specific eQTL, this will allow further investigation of the regional structure of the *cis* eQTL signals along the lines of work performed by Veyrieras and colleagues in 2008. This has the benefit of providing greater resolution regarding the location of the probable quantitative trait nucleotide(s), and thus insight into the functional basis of the genotype trait relationship. It may also be more biologically appropriate to alter the local boundries defining the *cis* region for eQTL analysis from a fixed distance, typically +/- 1Mb proximal to the transcript, and instead define the *cis* boundries based on the topologically associating domain (TAD) that the transcript is found in. It would be informative to investigate whether these expression signals are recent targets of natural selection and if present are they enriched in genes based on cellular function. Signatures of selection should be expected for variants from eQTL based on previous studies that assessed selection in *cis*-regulatory regions, eQTL, and beneficial adaptive traits in humans conferred by eQTL. ENCODE analyses of primate-specific *cis*-regulatory elements found that these elements display evidence of negative selection (ENCODE Project Consortium 2012). It has also been suggested that *cis*-eQTL may modulate the penetrance of deleterious protein coding changes through purifying (negative) selection (Lappalainen *et al.* 2011). While it has also been suggested that cis variants associated with changes in expression are enriched for variants showing signatures of recent positive selection (Kudaravalli *et al.* 2009). Two well studied instances of recent beneficial adapations in humans are malaria resistance and lactose persistence both or which are eQTL and show signatures of recent positive selection (Hamblin and Di Rienzo 2000; Hamblin, Thompson and Di Rienzo 2002; Olds and Sibley 2003; Bersaglieri *et al.* 2004; Tishkoff *et al.* 2007). In 2009, Flint and Mackay published a review of comparing QTL of quantitative phenotypes in mice, flies, and humans. They reported the existence of eQTL for a large number of loci but with moderate effects in all three species. However, based on studies done to that point they found very few homologous QTL (Flint and Mackay 2009). One area that has been considerably challenging is the reliable detection of *trans*-eQTL: these have generally proven difficult to replicate and it may be necessary to take several approaches in this regard, likely including anchoring observations in biologically plausible events (Dimas *et al.* 2008). Such an effort may also require the integration of other reference data such as that generated by ENCODE from a diverse set of tissues and cell types (~125) while recognizing that these cell types, or the conditions from which they were generated such as cancer cell lines (24%), may not be appropriate proxies for complete information. While these approaches will be useful it is likely that *trans* analysis will remain difficult until cohort sizes are large enough such that sufficient power is attained to reliably detect signal for intermediate effects. For instance, Westra *et al.* published a meta-analysis of *trans*-eQTL in peripheral blood, with a discovery cohort that included 5,311 subjects and a replication cohort of 2,775 subjects, in which they were able to replicate trans-eQTL for 103 loci. However, this analysis, while transcriptome-wide for gene expression measures, limited the genetic variation included for analysis to only 4,542 SNPs that had been previously implicated in complex traits or disease. The authors reported replication rates of 52% and 79% (based on individual SNPs not loci) when the replication was performed in two cohorts with gene expression measures also from blood samples. The replication rates were much smaller when when considering smaller eQTL cohorts from non-blood gene expression measures, between 2% and 7%. Additionally the authors note that 95% of the trans-eQTL identified in their discovery analysis accounted for less than 3% of the gene expression variance and that the replication cohorts they had access to lack sufficient power for replication of these particular loci (Westra *et al.* 2013). In the same vein as including external reference data, it will also be useful to expand our current models from pairwise Trait~Genotype to more integrative models Trait~Genotype (or Haplotype) with additional data resources such as miRNA expression and CpG methylation as covariates. Where DNA methylation measures for CpG sites that are cis-regulatory or miRNA expression levels for miRNA with putative binding sites within the mRNA 3’ UTR (or the entire transcript) could be used as covariates within the regression models (Zhang and Su 2008; Younger, Pertsemlidis and Corey 2009).

Ultimately, it would of course be extremely informative to include protein levels, and protein modifications in this effort, as this is often the next most proximal readout, to (in the example of disease) clinical phenotype. In 2010, Garge *et al.* published a study identifying protein QTL (pQTL) based on 24 LCLs from HapMap subjects for 544 proteins. They found 24 proteins (15 genes) where genetic variation accounted for more than 50% of the variation in protein abundance. Of the 24 proteins, 19 were associated with *cis*-variation and 4 of these were nonsynonymous coding variants that resulted in altered migration patterns on 2D gels (Garge *et al.* 2010). In 2013, Wu *et al.* published results re-affirming that mRNA expression levels are not a perfect proxy measure of protein abundance. They suggest non-correlations of these levels may be because of post-translational processes. In studies of mRNA expression and protein abundance the correlation of these measures has typically been modest. This study was based on ~6000 genes from 95 LCLs from HapMap subjects with protein levels measured by quantitative mass spectrometry. They found that protein levels vary between individuals and populations and that protein levels are heritable. They identified *cis*-pQTL, some of which did not have a previously identified eQTL based on eQTL analysis of mRNA expression from the same HapMap LCLs. A pQTL without a corresponding eQTL, where genetic variation is correlated with variation in protein abundance but not variation in mRNA expression would suggest that the effect of the genetic variation is post-transcriptional. They also found that sets of proteins involved in similar biology processes are well correlated between individuals, suggesting that these processes are tightly regulated (Wu *et al.* 2013). In 2014, Hause *et al.* published a study of pQTL based on 68 Yoruban Hapmap LCLs considering protein levels for 441 protein isoforms of transcription factors and signalling genes. They identified 12 *cis*-pQTL and 160 *trans*-pQTL. Two thirds of the eQTL from the same cohort and transcript set were also pQTL but this study also found that many pQTL did not have a corresponding eQTL. They found a significant enrichment in 5’ and 3’ UTRs and depletion in introns for eQTL. They also found 5’ and 3’ UTR enrichment for pQTL variants, but also saw enrichment for coding variants. They suggest that these pQTL variant enrichments may be involved in protein stability or miRNA-mediated regulation of mRNA translation. Their results also suggest there might be buffering of eQTL and that pQTL may contribute to phenotypic diversity (Hause *et al.* 2014). In 2015, Battle *et al.* published a study of eQTL, ribosomal occupancy QTL (rQTL), and pQTL based on 75 LCLs from HapMap Yoruban subjects. This study combined pre-existing genotype and RNAseq based expression measures with ribosomal profiling and protein abundance from quantitative high-resolution mass spectrometry. In general, they found consistencies between QTL types. They found that on average the expression variation identified for eQTL was attenuated or buffered at the protein level. They also identified pQTL where eQTL were not present. Additionally, they were able to identify expression- and protein-specific QTL (esQTL and psQTL). They found that when eQTL and pQTL are discordant that the ribosomal data typically tracked with the RNA. They suggest this means that psQTL are not capturing signal related to transcription or translation but possibly rates of protein degradation. They find that psQTL are enriched for UTR and coding variants, but the enrichment in coding variants is larger for nonsynonymous variants than other coding variants (Battle *et al.* 2015).

Another particularly exciting extension of the current work lies in the application of sequence-based assays of gene expression (RNAseq) within brain tissues. While array based work has many positives, the application of high-throughput sequencing, particularly with the development of improving analytical approaches, offers more potential insight. RNAseq offers several advantages: first it provides absolute knowledge regarding genotype, capturing rare and common variants; second it removes probe-based design limitations of microarrays; third, it allows for detection and analysis of low abundance RNA species for analysis, such as *LRRK2*; fourth, it allows assay of exon usage; and fifth, it can also directly reveal expression differences, in the context of allele-specific expression, by allowing the observation of an imbalance of alleles expressed within any informative transcript. Removing the probe-based limitation will allow for analysis of all detectable splice forms and whether or not eQTL are more associated with general gene expression or expression of gene splice forms as well as to be able to more clearly identify allelic expression. A probe-free design also removes the artefact of polymorphisms within the probe. In addition to detecting low abundance mRNA, protocols are now available for more general sequencing of RNA, including strand-specific and total RNA sequencing. There are many hurdles to overcome in RNAseq work and the sample preparation and data analysis are more challenging than with array-based work. Our laboratory has performed some preliminary transcriptome sequencing on ~60 smallRNA tissue samples, ~75 mRNA tissue samples and now approximately 300 total RNA tissue samples, from the NABEC cohort described in this thesis, where these subjects also have targeted deep resequencing and exome sequencing data. This work remains in the early stages; however, it promises to reveal further understanding into the genetic basis of expression. In 2008, Sultan *et al.* published a very early deep sequencing based study of gene expression in human embryonic kidney and B cells. They found that 66% of polyA transcripts mapped to known genes and 34% to unannotated genomic regions. Their findings suggested that RNAseq based measures of gene expression can detect 25% more expressed genes than microarrays. Additionally, with RNAseq based methods splicing events are measureable and they found that exon skipping is the most common form of alternative splicing within their analysis and results (Sultan *et al.* 2008). In 2009, Tang *et al.* published a study describing a single-cell digital expression assay for performing mRNAseq. They suggest their method allows the detection of more expression events than trying to do so with other protocols requiring more material from more cells. They ran their assay on a single mouse blastomere and detected the expression of 75% more genes, between 8% and 19% of the genes expressed multiple isoforms in the same cell. Additionally, they detected more than 1,700 novel splice junctions (Tang *et al.* 2009). In 2010, Pickrell *et al.* published their RNAseq study based on 69 LCLs from HapMap Nigerians subjects. This study used a pooled RNAseq approach to survey the transcriptional landscape and found extensive use of unannotated UTRs as well as 100 new coding exons. For the genes, with these new exons, 4.6% also had a *cis*-eQTL. For 90% of the eQTL, they identified, they found that the variants affecting expression are within 15 Kb of the gene. They also found that 88% of eQTL have reads from the higher expressing haplotype suggesting that many eQTL are allele-specific and result from modulation of *cis*-regulatory elements. Of note, based on heterozygotes, the ratio of reads from the fraction of the higher expression haplotypes correlates with the strength of the eQTL. Additionally, they found that genetic variation correlated with splicing (sQTL) and these were enriched in or near consensus splice sites (Pickrell *et al.* 2010). In 2010, Montgomery *et al.* also published an RNAseq study based on 60 LCLs from HapMap CEU subjects. Their results suggest that sequencing allows for improved dynamic range and better quantification of alternative and high expression transcripts, which should allow for improved eQTL detection. They found more eQTL using RNAseq than expression arrays, when comparing to results generated from the same samples previously. However, some of these differences were related to the array's saturation at the higher end and splicing complexity that are not well captured by the array's probe designs. The RNAseq based data also allowed for the detection of QTL for long noncoding RNAs as well as allele-specific expression (ASE). Their results suggest that rare ASE may be markers of rare eQTL variants. They also identified QTL for splicing (sQTL). For the sQTL identified, 41% of these were exon skipping, 17% alternative acceptor, 13% multiple exon skipping, 6% alternative donor, 5% exclusive exons, and 5% were for retained introns (Montgomery *et al.* 2010).

Another area I would like to pursue in relation to eQTL work in human brain is working with specific cell types, as I discussed in chapter 4. Following a similar course as Lee *et al.* described, in 2009, in using iPS cells to generate samples enriched for specific cell types which should already represent that natural genetic variation present in the subjects (Lee *et al.* 2009). Using iPS cells may also allow for the analysis of eQTL in neuronal cells types from patients with neurodegenerative diseases. Although, this would be depenedent on being able to overcome some of the limitations that can occur in iPS cell work such as generating the desired cell type without heterogeneity or a method that can account for the heterogeneity in the cell population. Currently it is not informative, from a disease perspective, to work with brain tissues of patients who died from neurodegenerative diseases such as Parkinson's and Alzheimer's disease. For neurodegenerative disease there is considerable cell loss in the brain, so assaying gene expression in these disease tissues is not informative for understanding the disease process. The ability to differentiate iPS cells into neuronal cells types and study them in a pre- or early-disease state would be informative. More recently, it has become possible to more easily edit genomic sequence using CRISPR-Cas9. Clustered regularly interspaced palindromic repeat and Cas9 (CRISPR-Cas9) allows targeted editing and manipulation of the genome using the Cas9 enzyme mechanism with a bacterially RNA-guided system (Barrangou 2014; Doudna and Charpentier 2014). I could foresee using the iPS cell based work to more narrowly map eQTL of interest in cells types of interest, such as Dopaminergic neurons generated from controls and PD patients, and in combination with other functional data identify the putative cis-regulatory risk variants. Using CRISPR-Cas9 these putative variants could be manipulated to identify causal risk variants and understand their effect in the cells and tissues relevant to disease. Although, currently I believe that CRISPR-Cas9 may need more refinement, as it is my understanding that beyond the targeted edits that sometimes other off target edits, such as InDels, can also occur.

8. References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.

1000 Genomes Project Consortium, Abecasis GR, Auton A *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.

Abu-Shaar M, Ryoo HD, Mann RS. Control of the nuclear localization of Extradenticle by competing nuclear import and export signals. *Genes Dev* 1999;**13**:935–45.

Alberts R, Terpstra P, Li Y *et al.* Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE* 2007;**2**:e622.

Allers T, Lichten M. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 2001;**106**:47–57.

Altshuler DM, Gibbs RA, Peltonen L *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;**467**:52–8.

Anttila V, Winsvold BS, Gormley P *et al.* Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nat Genet* 2013;**45**:912–7.

Arbiza L, Gronau I, Aksoy BA *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* 2013;**45**:723–9.

Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Dev Camb Engl* 1997;**124**:1851–64.

Bachner-Melman R, Dina C, Zohar AH *et al.* AVPR1a and SLC6A4 gene polymorphisms are associated with creative dance performance. *PLoS Genet* 2005;**1**:e42.

Baek D, Villén J, Shin C *et al.* The impact of microRNAs on protein output. *Nature* 2008;**455**:64–71.

Bamshad MJ, Mummidi S, Gonzalez E *et al.* A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. *Proc Natl Acad Sci U S A* 2002;**99**:10539–44.

Barbosa-Morais NL, Dunning MJ, Samarajiwa SA *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 2010;**38**:e17.

Barbosa-Morais NL, Irimia M, Pan Q *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* 2012;**338**:1587–93.

Barbujani G, Magagni A, Minch E *et al.* An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* 1997;**94**:4516–9.

Barrangou R. RNA events. Cas9 targeting and the CRISPR revolution. *Science* 2014;**344**:707–8.

Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet* 2006;**38**:659–62.

Barrett JC, Fry B, Maller J *et al.* Haploview: analysis and visualization of LD and haplotype maps. *Bioinforma Oxf Engl* 2005;**21**:263–5.

Barrett T, Troup DB, Wilhite SE *et al.* NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 2007;**35**:D760–5.

Barton NH, Otto SP. Evolution of recombination due to random drift. *Genetics* 2005;**169**:2353–70.

Battle A, Khan Z, Wang SH *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 2015;**347**:664–7.

Baudat F, Buard J, Grey C *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 2010;**327**:836–40.

Beilina A, Rudenko IN, Kaganovich A *et al.* Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease. *Proc Natl Acad Sci U S A* 2014;**111**:2626–31.

Beldade P, Brakefield PM, Long AD. Contribution of Distal-less to quantitative variation in butterfly eyespots. *Nature* 2002;**415**:315–8.

Bell AC, Felsenfeld G. Stopped at the border: boundaries and insulators. *Curr Opin Genet Dev* 1999;**9**:191–8.

Belting HG, Shashikant CS, Ruddle FH. Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proc Natl Acad Sci U S A* 1998;**95**:2355–60.

Ben-Ari Y ’ara, Brody Y, Kinor N *et al.* The life of an mRNA in space and time. *J Cell Sci* 2010;**123**:1761–74.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser* 1995;**B**:289–300.

Bergen A, Baccarelli A, McDaniel T *et al.* cis sequence effects on gene expression. *BMC Genomics* 2007;**8**:296.

Bersaglieri T, Sabeti PC, Patterson N *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004;**74**:1111–20.

Berthelsen J, Zappavigna V, Ferretti E *et al.* The novel homeoprotein Prep1 modulates Pbx-Hox protein cooperativity. *EMBO J* 1998;**17**:1434–45.

Bickel RD, Kopp A, Nuzhdin SV. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. *PLoS Genet* 2011;**7**:e1001275.

Botstein D, White RL, Skolnick M *et al.* Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980;**32**:314–31.

Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol (Berl)* 1991;**82**:239–59.

Braunschweig U, Barbosa-Morais NL, Pan Q *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* 2014;**24**:1774–86.

Bray NJ, Buckland PR, Owen MJ *et al.* Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet* 2003;**113**:149–53.

Bray NJ, Jehu L, Moskvina V *et al.* Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Hum Mol Genet* 2004;**13**:2885–92.

Brickman JM, Clements M, Tyrell R *et al.* Molecular effects of novel mutations in Hesx1/HESX1 associated with human pituitary disorders. *Dev Camb Engl* 2001;**128**:5189–99.

Cáceres M, Lachuer J, Zapala MA *et al.* Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* 2003;**100**:13030–5.

Cain CE, Blekhman R, Marioni JC *et al.* Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* 2011;**187**:1225–34.

Calhoun VC, Stathopoulos A, Levine M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* 2002;**99**:9243–7.

Cantero-Recasens G, Fandos C, Rubio-Moscardo F *et al.* The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. *Hum Mol Genet* 2010;**19**:111–21.

Carninci P, Kasukawa T, Katayama S *et al.* The transcriptional landscape of the mammalian genome. *Science* 2005;**309**:1559–63.

Carninci P, Sandelin A, Lenhard B *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;**38**:626–35.

Carrasquillo MM, Nicholson AM, Finch N *et al.* Genome-wide screen identifies rs646776 near sortilin as a regulator of progranulin levels in human plasma. *Am J Hum Genet* 2010;**87**:890–7.

Carrión AM, Link WA, Ledo F *et al.* DREAM is a Ca2+-regulated transcriptional repressor. *Nature* 1999;**398**:80–4.

Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 2008;**134**:25–36.

Caspi A, McClay J, Moffitt TE *et al.* Role of genotype in the cycle of violence in maltreated children. *Science* 2002;**297**:851–4.

Cavener DR. Transgenic animal studies on the evolution of genetic regulatory circuitries. *BioEssays News Rev Mol Cell Dev Biol* 1992;**14**:237–44.

Chakravarti A. Population genetics--making sense out of sequence. *Nat Genet* 1999;**21**:56–60.

Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics* 1993;**134**:1289–303.

Chaudhuri A, Polyakova J, Zbrzezna V *et al.* The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. *Blood* 1995;**85**:615–21.

Chaudhuri A, Zbrzezna V, Polyakova J *et al.* Expression of the Duffy antigen in K562 cells. Evidence that it is the human erythrocyte chemokine receptor. *J Biol Chem* 1994;**269**:7835–8.

Cheung VG, Conlin LK, Weber TM *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 2003;**33**:422–5.

Cheung VG, Spielman RS, Ewens KG *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005;**437**:1365–9.

Choy E, Yelensky R, Bonakdar S *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* 2008;**4**:e1000287.

Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics* 1994;**138**:963–71.

Clément-Ziza M, Munnich A, Lyonnet S *et al.* Stabilization of RNA during laser capture microdissection by performing experiments under argon atmosphere or using ethanol as a solvent in staining solutions. *RNA N Y N* 2008;**14**:2698–704.

Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997;**278**:1580–1.

Comeron JM, Williford A, Kliman RM. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* 2008;**100**:19–31.

Cookson W, Liang L, Abecasis G *et al.* Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009;**10**:184–94.

Coon KD, Myers AJ, Craig DW *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* 2007;**68**:613–8.

Coop G, Przeworski M. An evolutionary view of human recombination. *Nat Rev Genet* 2007;**8**:23–34.

Core LJ, Martins AL, Danko CG *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* 2014;**46**:1311–20.

Corradin O, Saiakhova A, Akhtar-Zaidi B *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* 2014;**24**:1–13.

Cowell LG, Kepler TB, Janitz M *et al.* The distribution of variation in regulatory gene segments, as present in MHC class II promoters. *Genome Res* 1998;**8**:124–34.

Crawford DL, Segal JA, Barnett JL. Evolutionary analysis of TATA-less proximal promoter function. *Mol Biol Evol* 1999;**16**:194–207.

Cresko WA, Amores A, Wilson C *et al.* Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc Natl Acad Sci U S A* 2004;**101**:6050–5.

Cui J, Stahl EA, Saevarsdottir S *et al.* Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet* 2013;**9**:e1003394.

Daly MJ, Rioux JD, Schaffner SF *et al.* High-resolution haplotype structure in the human genome. *Nat Genet* 2001;**29**:229–32.

von Dassow G, Meir E, Munro EM *et al.* The segment polarity network is a robust developmental module. *Nature* 2000;**406**:188–92.

Dausset J, Cann H, Cohen D *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 1990;**6**:575–7.

Davison D, Pritchard JK, Coop G. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor Popul Biol* 2009;**75**:331–45.

Dawson E, Abecasis GR, Bumpstead S *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 2002;**418**:544–8.

Dawson SJ, Morris PJ, Latchman DS. A single amino acid change converts an inhibitory transcription factor into an activator. *J Biol Chem* 1996;**271**:11631–3.

D'Elia AV, Tell G, Paron I *et al.* Missense mutations of human homeoboxes: A review. *Hum Mutat* 2001;**18**:361–74.

Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 2002;**19**:1114–21.

DiLeone RJ, Russell LB, Kingsley DM. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. *Genetics* 1998;**148**:401–8.

Dillon N, Sabbattini P. Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. *BioEssays News Rev Mol Cell Dev Biol* 2000;**22**:657–65.

Dimas AS, Deutsch S, Stranger BE *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 2009;**325**:1246–50.

Dimas AS, Stranger BE, Beazley C *et al.* Modifier effects between regulatory and protein-coding variation. *PLoS Genet* 2008;**4**:e1000244.

Di X, Matsuzaki H, Webster TA *et al.* Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinforma Oxf Engl* 2005;**21**:1958–63.

Dixon AL, Liang L, Moffatt MF *et al.* A genome-wide association study of global gene expression. *Nat Genet* 2007;**39**:1202–7.

Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.

Dodson MW, Zhang T, Jiang C *et al.* Roles of the Drosophila LRRK2 homolog in Rab7-dependent lysosomal positioning. *Hum Mol Genet* 2012;**21**:1350–63.

Do R, Balick D, Li H *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* 2015;**47**:126–31.

Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 2014;**346**:1258096.

Dröge P, Müller-Hill B. High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *BioEssays News Rev Mol Cell Dev Biol* 2001;**23**:179–83.

Duggal P, Gillanders EM, Holmes TN *et al.* Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* 2008;**9**:516.

Dunning AM, Durocher F, Healey CS *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* 2000;**67**:1544–54.

Eden E, Geva-Zatorsky N, Issaeva I *et al.* Proteome half-life dynamics in living human cells. *Science* 2011;**331**:764–8.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.

Eisen MB, Spellman PT, Brown PO *et al.* Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;**95**:14863–8.

Emilsson V, Thorleifsson G, Zhang B *et al.* Genetics of gene expression and its effect on disease. *Nature* 2008.

Emmert-Buck MR, Bonner RF, Smith PD *et al.* Laser capture microdissection. *Science* 1996;**274**:998–1001.

Enard W, Khaitovich P, Klose J *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* 2002a;**296**:340–3.

Enard W, Khaitovich P, Klose J *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* 2002b;**296**:340–3.

ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.

Enoch MA, Kaye WH, Rotondo A *et al.* 5-HT2A promoter polymorphism -1438G/A, anorexia nervosa, and obsessive-compulsive disorder. *Lancet Lond Engl* 1998;**351**:1785–6.

Evans SJ, Choudary PV, Vawter MP *et al.* DNA microarray analysis of functionally discrete human brain regions reveals divergent transcriptional profiles. *Neurobiol Dis* 2003;**14**:240–50.

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;**164**:1567–87.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H *et al.* A promoter-level mammalian expression atlas. *Nature* 2014;**507**:462–70.

Felsenstein J. The evolutionary advantage of recombination. *Genetics* 1974;**78**:737–56.

Flint J, Mackay TFC. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res* 2009;**19**:723–33.

Force A, Lynch M, Pickett FB *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 1999;**151**:1531–45.

Fraser HB. Gene expression drives local adaptation in humans. *Genome Res* 2013;**23**:1089–96.

Fraser HB, Xie X. Common polymorphic transcript variation in human disease. *Genome Res* 2009;**19**:567–75.

Frazer KA, Sheehan JB, Stokowski RP *et al.* Evolutionarily conserved sequences on human chromosome 21. *Genome Res* 2001;**11**:1651–9.

Friedrich B, Euler P, Ziegler R *et al.* Comparative analyses of Purkinje cell gene expression profiles reveal shared molecular abnormalities in models of different polyglutamine diseases. *Brain Res* 2012;**1481**:37–48.

Fry CJ, Farnham PJ. Context-dependent transcriptional regulation. *J Biol Chem* 1999;**274**:29583–6.

Fu J, Wolfs MGM, Deelen P *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* 2012;**8**:e1002431.

Gabriel SB, Schaffner SF, Nguyen H *et al.* The structure of haplotype blocks in the human genome. *Science* 2002;**296**:2225–9.

Galasko D, Edland SD, Morris JC *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part XI. Clinical milestones in patients with Alzheimer's disease followed over 3 years. *Neurology* 1995;**45**:1451–5.

Garge N, Pan H, Rowland MD *et al.* Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol Cell Proteomics MCP* 2010;**9**:1383–99.

Gerstein MB, Kundaje A, Hariharan M *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;**489**:91–100.

Gibbs JR, van der Brug MP, Hernandez DG *et al.* Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010;**6**:e1000952.

Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet TIG* 2008, DOI: S0168-9525(08)00177-7.

Gilbert JM, Brown BA, Strocchi P *et al.* The preparation of biologically active messenger RNA from human postmortem brain tissue. *J Neurochem* 1981;**36**:976–84.

Gompel N, Prud'homme B, Wittkopp PJ *et al.* Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* 2005;**433**:481–7.

Göring HHH, Curran JE, Johnson MP *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007;**39**:1208–16.

de la Grange P, Gratadou L, Delord M *et al.* Splicing factor and exon profiling across human tissues. *Nucleic Acids Res* 2010;**38**:2825–38.

Grosveld F, Antoniou M, Berry M *et al.* The regulation of human globin gene switching. *Philos Trans R Soc Lond B Biol Sci* 1993;**339**:183–91.

Grundberg E, Small KS, Hedman AK *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* 2012;**44**:1084–9.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5.

Guardiola J, Maffei A, Lauster R *et al.* Functional significance of polymorphism among MHC class II gene promoters. *Tissue Antigens* 1996;**48**:615–25.

Guerreiro RJ, Beck J, Gibbs JR *et al.* Genetic variability in CLU and its association with Alzheimer's disease. *PloS One* 2010;**5**:e9510.

Gunderson KL, Steemers FJ, Lee G *et al.* A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 2005;**37**:549–54.

Gustincich S, Sandelin A, Plessy C *et al.* The complexity of the mammalian transcriptome. *J Physiol* 2006;**575**:321–32.

Hadley TJ, Peiper SC. From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. *Blood* 1997;**89**:3077–91.

Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 2000;**66**:1669–79.

Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* 2002;**70**:369–83.

Hammock EAD, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 2005;**308**:1630–4.

Hariri AR, Mattay VS, Tessitore A *et al.* Serotonin transporter genetic variation and the response of the human amygdala. *Science* 2002;**297**:400–3.

Hause RJ, Stark AL, Antao NN *et al.* Identification and Validation of Genetic Variants that Influence Transcription Factor and Cell Signaling Protein Levels. *Am J Hum Genet* 2014, DOI: 10.1016/j.ajhg.2014.07.005.

Heidari N, Phanstiel DH, He C *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res* 2014;**24**:1905–17.

Heinzen EL, Ge D, Cronin KD *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* 2008;**6**:e1.

Hellmann I, Ebersberger I, Ptak SE *et al.* A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* 2003;**72**:1527–35.

Hellmann I, Prüfer K, Ji H *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res* 2005;**15**:1222–31.

Heo HY, Kim K-S, Seol W. Coordinate Regulation of Neurite Outgrowth by LRRK2 and Its Interactor, Rab5. *Exp Neurobiol* 2010;**19**:97–105.

Hernandez DG, Nalls MA, Moore M *et al.* Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol Dis* 2012;**47**:20–8.

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res* 1966;**8**:269–94.

Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;**106**:9362–7.

Höglinger GU, Melhem NM, Dickson DW *et al.* Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat Genet* 2011;**43**:699–705.

Holstege FC, Jennings EG, Wyrick JJ *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 1998;**95**:717–28.

Holton P, Ryten M, Nalls M *et al.* Initial assessment of the pathogenic mechanisms of the recently identified Alzheimer risk Loci. *Ann Hum Genet* 2013;**77**:85–105.

Horuk R, Chitnis CE, Darbonne WC *et al.* A receptor for the malarial parasite Plasmodium vivax: the erythrocyte chemokine receptor. *Science* 1993;**261**:1182–4.

Hovatta I, Tennant RS, Helton R *et al.* Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature* 2005;**438**:662–6.

Hovatta I, Zapala MA, Broide RS *et al.* DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biol* 2007;**8**:R25.

Howie B, Fuchsberger C, Stephens M *et al.* Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;**44**:955–9.

Hsiao LL, Dangond F, Yoshida T *et al.* A compendium of gene expression in normal human tissues. *Physiol Genomics* 2001;**7**:97–104.

Hudson RR, Kaplan NL. The coalescent process in models with selection and recombination. *Genetics* 1988;**120**:831–40.

Hurd YL. Subjects with major depression or bipolar disorder show reduction of prodynorphin mRNA expression in discrete nuclei of the amygdaloid complex. *Mol Psychiatry* 2002;**7**:75–81.

Hussin JG, Hodgkinson A, Idaghdour Y *et al.* Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet* 2015, DOI: 10.1038/ng.3216.

Innocenti F, Cooper GM, Stanaway IB *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait Loci in primary human liver tissue. *PLoS Genet* 2011;**7**:e1002078.

International HapMap Consortium. The International HapMap Project. *Nature* 2003;**426**:789–96.

International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.

International HapMap Consortium, Frazer KA, Ballinger DG *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–61.

Iwamoto S, Li J, Sugimoto N *et al.* Characterization of the Duffy gene promoter: evidence for tissue-specific abolishment of expression in Fy(a-b-) of black individuals. *Biochem Biophys Res Commun* 1996;**222**:852–9.

Jackson-Fisher AJ, Chitikila C, Mitra M *et al.* A role for TBP dimerization in preventing unregulated gene expression. *Mol Cell* 1999;**3**:717–27.

Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961;**3**:318–56.

Jacques P-É, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* 2013;**9**:e1003504.

Jakobsson M, Scholz SW, Scheet P *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008;**451**:998–1003.

Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet TIG* 2001;**17**:388–91.

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001;**29**:217–22.

Jensen JD, Thornton KR, Bustamante CD *et al.* On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* 2007;**176**:2371–9.

Jin W, Riley RM, Wolfinger RD *et al.* The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat Genet* 2001;**29**:389–95.

Johnson GC, Esposito L, Barratt BJ *et al.* Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;**29**:233–7.

Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. *Science* 2001;**293**:1068–70.

Kadosh D, Struhl K. Targeted recruitment of the Sin3-Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin in vivo. *Mol Cell Biol* 1998;**18**:5121–7.

Kammandel B, Chowdhury K, Stoykova A *et al.* Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. *Dev Biol* 1999;**205**:79–97.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.

Karp CL, Grupe A, Schadt E *et al.* Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol* 2000;**1**:221–6.

Kasowski M, Grubert F, Heffelfinger C *et al.* Variation in transcription factor binding among humans. *Science* 2010;**328**:232–5.

Kathiresan S, Melander O, Guiducci C *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008;**40**:189–97.

Keightley PD, Otto SP. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 2006;**443**:89–92.

Kellis M, Wold B, Snyder MP *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 2014;**111**:6131–8.

Kendirgi F, Rexer DJ, Alcázar-Román AR *et al.* Interaction between the shuttling mRNA export factor Gle1 and the nucleoporin hCG1: a conserved mechanism in the export of Hsp70 mRNA. *Mol Biol Cell* 2005;**16**:4304–15.

Khaitovich P, Hellmann I, Enard W *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 2005;**309**:1850–4.

Khaitovich P, Muetzel B, She X *et al.* Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* 2004;**14**:1462–73.

Kim-Cohen J, Caspi A, Taylor A *et al.* MAOA, maltreatment, and gene-environment interaction predicting children's mental health: new evidence and a meta-analysis. *Mol Psychiatry* 2006;**11**:903–13.

Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* 2005;**76**:8–32.

Klein RJ, Zeiss C, Chew EY *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;**308**:385–9.

Knoepfler PS, Kamps MP. The pentapeptide motif of Hox proteins is required for cooperative DNA binding with Pbx1, physically contacts Pbx1, and enhances DNA binding by Pbx1. *Mol Cell Biol* 1995;**15**:5811–9.

Kong A, Gudbjartsson DF, Sainz J *et al.* A high-resolution recombination map of the human genome. *Nat Genet* 2002;**31**:241–7.

Kong A, Thorleifsson G, Frigge ML *et al.* Common and low-frequency variants associated with genome-wide recombination rate. *Nat Genet* 2014;**46**:11–6.

Kristensen AR, Gsponer J, Foster LJ. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Mol Syst Biol* 2013;**9**:689.

Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999;**22**:139–44.

Kudaravalli S, Veyrieras J-B, Stranger BE *et al.* Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol* 2009;**26**:649–58.

Kumar A, Gibbs JR, Beilina A *et al.* Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol Aging* 2013;**34**:1199–209.

Kuras L, Struhl K. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature* 1999;**399**:609–13.

Kwan T, Benovoy D, Dias C *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* 2008;**40**:225–31.

Kwan T, Grundberg E, Koka V *et al.* Tissue effect on genetic control of transcript isoform variation. *PLoS Genet* 2009;**5**:e1000608.

Lagrange T, Kapanidis AN, Tang H *et al.* New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* 1998;**12**:34–44.

Lambert JC, Pérez-Tur J, Dupire MJ *et al.* Distortion of allelic expression of apolipoprotein E in Alzheimer's disease. *Hum Mol Genet* 1997;**6**:2151–4.

Lam EW-F, Brosens JJ, Gomes AR *et al.* Forkhead box proteins: tuning forks for transcriptional harmony. *Nat Rev Cancer* 2013;**13**:482–95.

Landolin JM, Johnson DS, Trinklein ND *et al.* Sequence features that drive human promoter function and tissue specificity. *Genome Res* 2010;**20**:890–8.

Lappalainen T, Montgomery SB, Nica AC *et al.* Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet* 2011;**89**:459–63.

Laurent JM, Vogel C, Kwon T *et al.* Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 2010;**10**:4209–12.

Laurie DA, Hultén MA. Further studies on chiasma distribution and interference in the human male. *Ann Hum Genet* 1985;**49**:203–14.

Lawrence T, Natoli G. Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nat Rev Immunol* 2011;**11**:750–61.

Laws SM, Friedrich P, Diehl-Schmid J *et al.* Fine mapping of the MAPT locus using quantitative trait analysis identifies possible causal variants in Alzheimer's disease. *Mol Psychiatry* 2007;**12**:510–7.

Lee J-H, Park I-H, Gao Y *et al.* A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* 2009;**5**:e1000718.

Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 2000;**34**:77–137.

Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;**14**:2551–69.

Lesecque Y, Glémin S, Lartillot N *et al.* The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet* 2014;**10**:e1004790.

Lettice LA, Horikoshi T, Heaney SJH *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* 2002;**99**:7548–53.

Libioulle C, Louis E, Hansoul S *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 2007;**3**:e58.

Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2014;**2**:e270.

Li JZ, Vawter MP, Walsh DM *et al.* Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum Mol Genet* 2004;**13**:609–16.

Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet EJHG* 2008.

Liotta L, Petricoin E. Molecular profiling of human cancer. *Nat Rev Genet* 2000;**1**:48–56.

Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 1989;**44**:397–401.

Li WW, Dammerman MM, Smith JD *et al.* Common genetic variation in the promoter of the human apo CIII gene abolishes regulation by insulin and may contribute to hypertriglyceridemia. *J Clin Invest* 1995;**96**:2601–5.

Li Y, Willer CJ, Ding J *et al.* MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010;**34**:816–34.

Li Y, Willer C, Sanna S *et al.* Genotype imputation. *Annu Rev Genomics Hum Genet* 2009;**10**:387–406.

Lohmueller KE, Bustamante CD, Clark AG. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 2009;**182**:217–31.

Luo L, Salunga RC, Guo H *et al.* Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat Med* 1999;**5**:117–22.

Lynn A, Ashley T, Hassold T. Variation in human meiotic recombination. *Annu Rev Genomics Hum Genet* 2004;**5**:317–49.

Mack WJ, Freed DM, Williams BW *et al.* Boston Naming Test: shortened versions for use in Alzheimer's disease. *J Gerontol* 1992;**47**:P154–8.

MacLeod DA, Rhinn H, Kuwahara T *et al.* RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk. *Neuron* 2013;**77**:425–39.

MacLeod D, Dowman J, Hammond R *et al.* The familial Parkinsonism gene LRRK2 regulates neurite process morphology. *Neuron* 2006;**52**:587–93.

Mailman MD, Feolo M, Jin Y *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;**39**:1181–6.

Maiuri P, Knezevich A, De Marco A *et al.* Fast transcription rates of RNA polymerase II in human cells. *EMBO Rep* 2011;**12**:1280–5.

Manzanares M, Wada H, Itasaki N *et al.* Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. *Nature* 2000;**408**:854–7.

Marcil A, Dumontier E, Chamberland M *et al.* Pitx1 and Pitx2 are required for development of hindlimb buds. *Dev Camb Engl* 2003;**130**:45–55.

de Massy B. Distribution of meiotic recombination sites. *Trends Genet TIG* 2003;**19**:514–22.

Mathieson I, McVean G. Demography and the age of rare variants. *PLoS Genet* 2014;**10**:e1004528.

Matsuzaki H, Dong S, Loi H *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 2004;**1**:109–11.

May CA, Shone AC, Kalaydjieva L *et al.* Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat Genet* 2002;**31**:272–5.

McClurg P, Janes J, Wu C *et al.* Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics* 2007;**176**:675–83.

McVean GAT, Myers SR, Hunt S *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 2004;**304**:581–4.

Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science* 1978;**201**:786–92.

Metherall JE, Gillespie FP, Forget BG. Analyses of linked beta-globin genes suggest that nondeletion forms of hereditary persistence of fetal hemoglobin are bona fide switching mutants. *Am J Hum Genet* 1988;**42**:476–81.

Miller LH, Mason SJ, Clyde DF *et al.* The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* 1976;**295**:302–4.

Milo R, Shen-Orr S, Itzkovitz S *et al.* Network motifs: simple building blocks of complex networks. *Science* 2002;**298**:824–7.

Moffatt MF, Kabesch M, Liang L *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007a;**448**:470–3.

Moffatt MF, Kabesch M, Liang L *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007b;**448**:470–3.

Monks SA, Leonardson A, Zhu H *et al.* Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* 2004;**75**:1094–105.

Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. *Nat Rev Genet* 2011;**12**:277–82.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010;**464**:773–7.

Morley M, Molony CM, Weber TM *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004;**430**:743–7.

Muller HJ. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutat Res* 1964;**106**:2–9.

Myers AJ, Gibbs JR, Webster JA *et al.* A survey of genetic human cortical gene expression. *Nat Genet* 2007a;**39**:1494–9.

Myers AJ, Pittman AM, Zhao AS *et al.* The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol Dis* 2007b;**25**:561–70.

Myers S, Bottolo L, Freeman C *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science* 2005;**310**:321–4.

Myers S, Bowden R, Tumian A *et al.* Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 2010;**327**:876–9.

Naganawa S, Ginsberg HN, Glickman RM *et al.* Intestinal transcription and synthesis of apolipoprotein AI is regulated by five natural polymorphisms upstream of the apolipoprotein CIII gene. *J Clin Invest* 1997;**99**:1958–65.

Nalls MA, Pankratz N, Lill CM *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson’s disease. *Nat Genet* 2014;**46**:989–93.

Nalls MA, Plagnol V, Hernandez DG *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson’s disease: a meta-analysis of genome-wide association studies. *Lancet* 2011;**377**:641–9.

Nalls MA, Saad M, Noyce AJ *et al.* Genetic comorbidities in Parkinson’s disease. *Hum Mol Genet* 2013, DOI: 10.1093/hmg/ddt465.

Neznanov N, Umezawa A, Oshima RG. A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J Biol Chem* 1997;**272**:27549–57.

Nica AC, Montgomery SB, Dimas AS *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* 2010;**6**:e1000895.

Nica AC, Parts L, Glass D *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genet* 2011;**7**:e1002003.

Nichols WC, Pankratz N, Hernandez D *et al.* Genetic screening for a single common LRRK2 mutation in familial Parkinson’s disease. *Lancet Lond Engl* 2005;**365**:410–2.

Nicodemus KK, Liu W, Chase GA *et al.* Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genet* 2005;**6 Suppl 1**:S78.

Nicolae DL, Gamazon E, Zhang W *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;**6**:e1000888.

Nielsen LB, Kahn D, Duell T *et al.* Apolipoprotein B gene expression in a series of human apolipoprotein B transgenic mice generated with recA-assisted restriction endonuclease cleavage-modified bacterial artificial chromosomes. An intestine-specific enhancer element is located between 54 and 62 kilobases 5’ to the structural gene. *J Biol Chem* 1998;**273**:21800–7.

Nobrega MA, Ovcharenko I, Afzal V *et al.* Scanning human gene deserts for long-range enhancers. *Science* 2003;**302**:413.

Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. *Trends Genet TIG* 2002;**18**:83–90.

North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. *Am J Hum Genet* 2002;**71**:439–41.

Novembre J, Johnson T, Bryc K *et al.* Genes mirror geography within Europe. *Nature* 2008;**456**:98–101.

Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008;**40**:646–9.

Oberdick J, Levinthal F, Levinthal C. A Purkinje cell differentiation marker shows a partial DNA sequence homology to the cellular sis/PDGF2 gene. *Neuron* 1988;**1**:367–76.

Okaty BW, Sugino K, Nelson SB. A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PloS One* 2011;**6**:e16493.

Olds LC, Sibley E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum Mol Genet* 2003;**12**:2333–40.

Oleksiak MF, Churchill GA, Crawford DL. Variation in gene expression within and among natural populations. *Nat Genet* 2002;**32**:261–6.

Onyango P, Miller W, Lehoczky J *et al.* Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res* 2000;**10**:1697–710.

Ordway GA, Szebeni A, Duffourc MM *et al.* Gene expression analyses of neurons, astrocytes, and oligodendrocytes isolated by laser capture microdissection from human brain: detrimental effects of laboratory humidity. *J Neurosci Res* 2009;**87**:2430–8.

Orphanides G, Lagrange T, Reinberg D. The general transcription factors of RNA polymerase II. *Genes Dev* 1996;**10**:2657–83.

Otto SP, Lenormand T. Resolving the paradox of sex and recombination. *Nat Rev Genet* 2002;**3**:252–61.

Pai AA, Bell JT, Marioni JC *et al.* A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet* 2011;**7**:e1001316.

Paisán-Ruíz C, Jain S, Evans EW *et al.* Cloning of the gene containing mutations that cause PARK8-linked Parkinson’s disease. *Neuron* 2004;**44**:595–600.

Pan Q, Shai O, Lee LJ *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008, DOI: ng.259.

Paquette J, Giannoukakis N, Polychronakos C *et al.* The INS 5’ variable number of tandem repeats is associated with IGF2 expression in humans. *J Biol Chem* 1998;**273**:14158–64.

Park SJ, Lee JH, Kim HY *et al.* Astrocytes, but not microglia, rapidly sense H_2O_2via STAT6 phosphorylation, resulting in cyclooxygenase-2 expression and prostaglandin release. *J Immunol Baltim Md 1950* 2012;**188**:5132–41.

Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. *Hum Mol Genet* 2006;**15 Spec No 1**:R9–16.

Pastinen T, Sladek R, Gurd S *et al.* A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics* 2004;**16**:184–93.

Patil N, Berno AJ, Hinds DA *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;**294**:1719–23.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.

Peckys D, Hurd YL. Prodynorphin and kappa opioid receptor mRNA expression in the cingulate and prefrontal cortices of subjects diagnosed with schizophrenia or affective disorders. *Brain Res Bull* 2001;**55**:619–24.

Peiper SC, Wang ZX, Neote K *et al.* The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. *J Exp Med* 1995;**181**:1311–7.

Petronczki M, Siomos MF, Nasmyth K. Un ménage à quatre: the molecular biology of chromosome segregation in meiosis. *Cell* 2003;**112**:423–40.

Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* 2014;**94**:559–73.

Pickrell JK, Marioni JC, Pai AA *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010, DOI: 10.1038/nature08872.

Pittman AM, Myers AJ, Duckworth J *et al.* The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum Mol Genet* 2004;**13**:1267–74.

Plagnol V, Uz E, Wallace C *et al.* Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PloS One* 2008;**3**:e2966.

Plagnol V, Wall JD. Possible ancestral structure in human populations. *PLoS Genet* 2006;**2**:e105.

Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. *Proc Natl Acad Sci U S A* 2004;**101**:12588–91.

Pogo AO, Chaudhuri A. The Duffy protein: a malarial and chemokine receptor. *Semin Hematol* 2000;**37**:122–9.

Pollard KS, Dudoit S, Laan MJ van der. Multiple Testing Procedures: the multtest Package and Applications to Genomics. In: Gentleman R, Carey VJ, Huber W, et al. (eds.). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer New York, 2005, 249–71.

Ponjavic J, Lenhard B, Kai C *et al.* Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol* 2006;**7**:R78.

Pool JE, Hellmann I, Jensen JD *et al.* Population genetic inference from genomic sequence variation. *Genome Res* 2010;**20**:291–300.

Price AL, Helgason A, Thorleifsson G *et al.* Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLoS Genet* 2011;**7**:e1001317.

Price AL, Patterson NJ, Plenge RM *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.

Prud'homme B, Gompel N, Rokas A *et al.* Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 2006;**440**:1050–3.

Purcell S, Neale B, Todd-Brown K *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 2007;**81**:559–75.

Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinforma Oxf Engl* 2006;**22**:7–12.

Rademakers R, Melquist S, Cruts M *et al.* High-density SNP haplotyping suggests altered regulation of tau gene expression in progressive supranuclear palsy. *Hum Mol Genet* 2005;**14**:3281–92.

R C Lewontin. EVOL BIOL. *Evol Biol* 1972;**6**:381–98.

R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2012.

Reich DE, Cargill M, Bolk S *et al.* Linkage disequilibrium in the human genome. *Nature* 2001;**411**:199–204.

Reinberg D, Orphanides G, Ebright R *et al.* The RNA polymerase II general transcription factors: past, present, and future. *Cold Spring Harb Symp Quant Biol* 1998;**63**:83–103.

Richards EJ, Elgin SCR. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* 2002;**108**:489–500.

Rieder MJ, Reiner AP, Gage BF *et al.* Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med* 2005;**352**:2285–93.

Rockman MV, Hahn MW, Soranzo N *et al.* Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol* 2005;**3**:e387.

Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol* 2002;**19**:1991–2004.

Roeder GS. Meiotic chromosomes: it takes two to tango. *Genes Dev* 1997;**11**:2600–21.

Rogaeva E. The solved and unsolved mysteries of the genetics of early-onset Alzheimer’s disease. *Neuromolecular Med* 2002;**2**:1–10.

Romey MC, Guittard C, Chazalette JP *et al.* Complex allele [-102T>A+S549R(T>G)] is associated with milder forms of cystic fibrosis than allele S549R(T>G) alone. *Hum Genet* 1999;**105**:145–50.

Romey MC, Pallares-Ruiz N, Mange A *et al.* A naturally occurring sequence variation that creates a YY1 element is associated with increased cystic fibrosis transmembrane conductance regulator gene expression. *J Biol Chem* 2000;**275**:3561–7.

Ronald J, Brem RB, Whittle J *et al.* Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet* 2005;**1**:e25.

Rong Y, Wang T, Morgan JI. Identification of candidate Purkinje cell-specific markers by gene expression profiling in wild-type and pcd(3J) mice. *Brain Res Mol Brain Res* 2004;**132**:128–45.

Rosenberg NA, Pritchard JK, Weber JL *et al.* Genetic structure of human populations. *Science* 2002;**298**:2381–5.

Rozenberg JM, Shlyakhtenko A, Glass K *et al.* All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC Genomics* 2008;**9**:67.

Ruvkun G, Wightman B, Bürglin T *et al.* Dominant gain-of-function mutations that lead to misregulation of the C. elegans heterochronic gene lin-14, and the evolutionary implications of dominant mutations in pattern-formation genes. *Dev Camb Engl Suppl* 1991;**1**:47–54.

Sabeti PC, Reich DE, Higgins JM *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002;**419**:832–7.

Sabeti PC, Varilly P, Fry B *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007;**449**:913–8.

Sachidanandam R, Weissman D, Schmidt SC *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;**409**:928–33.

Saeed AI, Sharov V, White J *et al.* TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* 2003;**34**:374–8.

Saito-Hisaminato A, Katagiri T, Kakiuchi S *et al.* Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res Int J Rapid Publ Rep Genes Genomes* 2002;**9**:35–45.

Sakurai D, Zhao J, Deng Y *et al.* Preferential binding to Elk-1 by SLE-associated IL10 risk allele upregulates IL10 expression. *PLoS Genet* 2013;**9**:e1003870.

Sandelin A, Carninci P, Lenhard B *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 2007;**8**:424–36.

Santostefano KE, Hamazaki T, Biel NM *et al.* A practical guide to induced pluripotent stem cell research using patient samples. *Lab Investig J Tech Methods Pathol* 2015;**95**:4–13.

Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 2006;**103**:1412–7.

Scaffidi P, Bianchi ME. Spatially precise DNA bending is an essential activity of the sox2 transcription factor. *J Biol Chem* 2001;**276**:47296–302.

Schadt EE, Li C, Ellis B *et al.* Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl* 2001;**Suppl 37**:120–5.

Schadt EE, Molony C, Chudin E *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008;**6**:e107.

Schadt EE, Monks SA, Drake TA *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;**422**:297–302.

Scharf JM, Yu D, Mathews CA *et al.* Genome-wide association study of Tourette's syndrome. *Mol Psychiatry* 2012, DOI: 10.1038/mp.2012.69.

Schaub MA, Boyle AP, Kundaje A *et al.* Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;**22**:1748–59.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 2014, DOI: 10.1038/ng.3015.

Schliekelman P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* 2008;**178**:2201–16.

Schug J, Schuller W-P, Kappen C *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 2005;**6**:R33.

Schwanhäusser B, Busse D, Li N *et al.* Global quantification of mammalian gene expression control. *Nature* 2011;**473**:337–42.

Segal JA, Barnett JL, Crawford DL. Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. *J Mol Evol* 1999;**49**:736–49.

Selbach M, Schwanhäusser B, Thierfelder N *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* 2008, DOI: nature07228.

Shabalina SA, Ogurtsov AY, Kondrashov VA *et al.* Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet TIG* 2001;**17**:373–6.

Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014;**15**:335–46.

Shang J, Luo Y, Clayton DA. Backfoot is a novel homeobox gene expressed in the mesenchyme of developing hind limb. *Dev Dyn Off Publ Am Assoc Anat* 1997;**209**:242–53.

Shapiro MD, Bell MA, Kingsley DM. Parallel genetic origins of pelvic reduction in vertebrates. *Proc Natl Acad Sci U S A* 2006;**103**:13753–8.

Shapiro MD, Marks ME, Peichel CL *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 2004;**428**:717–23.

Sharova LV, Sharov AA, Nedorezov T *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res Int J Rapid Publ Rep Genes Genomes* 2009;**16**:45–58.

Sheffield NC, Thurman RE, Song L *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* 2013;**23**:777–88.

Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 2014;**15**:272–86.

Shyamsundar R, Kim YH, Higgins JP *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biol* 2005;**6**:R22.

Simon J, Peifer M, Bender W *et al.* Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. *EMBO J* 1990;**9**:3945–56.

Simón-Sánchez J, Schulte C, Bras JM *et al.* Genome-wide association study reveals genetic risk underlying Parkinson’s disease. *Nat Genet* 2009, DOI: 10.1038/ng.487.

Simpson P, Woehl R, Usui K. The development and evolution of bristle patterns in Diptera. *Dev Camb Engl* 1999;**126**:1349–64.

Skaer N, Simpson P. Genetic analysis of bristle loss in hybrids between Drosophila melanogaster and D. simulans provides evidence for divergence of cis-regulatory sequences in the achaete-scute gene complex. *Dev Biol* 2000;**221**:148–67.

Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res* 1974;**23**:23–35.

Son CG, Bilke S, Davis S *et al.* Database of mRNA gene expression profiles of multiple human organs. *Genome Res* 2005;**15**:443–50.

Stafa K, Trancikova A, Webber PJ *et al.* GTPase activity and neuronal toxicity of Parkinson's disease-associated LRRK2 is regulated by ArfGAP1. *PLoS Genet* 2012;**8**:e1002526.

Steemers FJ, Chang W, Lee G *et al.* Whole-genome genotyping with the single-base extension assay. *Nat Methods* 2006;**3**:31–3.

Stefansson H, Helgason A, Thorleifsson G *et al.* A common inversion under selection in Europeans. *Nat Genet* 2005;**37**:129–37.

Stephan W, Song YS, Langley CH. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* 2006;**172**:2647–63.

Stern DL. Evolutionary developmental biology and the problem of variation. *Evol Int J Org Evol* 2000;**54**:1079–91.

Stewart SE, Yu D, Scharf JM *et al.* Genome-wide association study of obsessive-compulsive disorder. *Mol Psychiatry* 2012, DOI: 10.1038/mp.2012.85.

Stögmann E, Zimprich A, Baumgartner C *et al.* A functional polymorphism in the prodynorphin gene promotor is associated with temporal lobe epilepsy. *Ann Neurol* 2002;**51**:260–3.

Stone JR, Wray GA. Rapid evolution of cis-regulatory sequences via local point mutations. *Mol Biol Evol* 2001;**18**:1764–70.

Storey JD, Madeoy J, Strout JL *et al.* Gene-expression variation within and among human populations. *Am J Hum Genet* 2007;**80**:502–9.

Stranger BE, Forrest MS, Clark AG *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet* 2005;**1**:e78.

Stranger BE, Nica AC, Forrest MS *et al.* Population genomics of human gene expression. *Nat Genet* 2007;**39**:1217–24.

Su AI, Cooke MP, Ching KA *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 2002;**99**:4465–70.

Sucena E, Delon I, Jones I *et al.* Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. *Nature* 2003;**424**:935–8.

Sucena E, Stern DL. Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc Natl Acad Sci U S A* 2000;**97**:4530–4.

Sultan M, Schulz MH, Richard H *et al.* A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science* 2008, DOI: 1160342.

Swallow DM. Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 2003;**37**:197–219.

Syddall CM, Reynard LN, Young DA *et al.* The identification of trans-acting factors that regulate the expression of GDF5 via the osteoarthritis susceptibility SNP rs143383. *PLoS Genet* 2013;**9**:e1003557.

Takata R, Akamatsu S, Kubo M *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat Genet* 2010;**42**:751–4.

Tang F, Barbacioru C, Wang Y *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**:377–82.

Tennyson CN, Klamut HJ, Worton RG. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet* 1995;**9**:184–90.

Thurman RE, Rynes E, Humbert R *et al.* The accessible chromatin landscape of the human genome. *Nature* 2012;**489**:75–82.

Tishkoff SA, Reed FA, Ranciaro A *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 2007;**39**:31–40.

Tomita H, Vawter MP, Walsh DM *et al.* Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. *Biol Psychiatry* 2004;**55**:346–52.

Torchia J, Glass C, Rosenfeld MG. Co-activators and co-repressors in the integration of transcriptional responses. *Curr Opin Cell Biol* 1998;**10**:373–83.

Tournamille C, Blancher A, Le Van Kim C *et al.* Sequence, evolution and ligand binding properties of mammalian Duffy antigen/receptor for chemokines. *Immunogenetics* 2004;**55**:682–94.

Tournamille C, Colin Y, Cartron JP *et al.* Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* 1995;**10**:224–8.

Townsend JP, Cavalieri D, Hartl DL. Population genetic variation in genome-wide gene expression. *Mol Biol Evol* 2003;**20**:955–63.

Traynor BJ, Nalls M, Lai S-L *et al.* Kinesin-associated protein 3 (KIFAP3) has no effect on survival in a population-based cohort of ALS patients. *Proc Natl Acad Sci U S A* 2010;**107**:12335–8.

Trefilov A, Berard J, Krawczak M *et al.* Natal dispersal in rhesus macaques is related to serotonin transporter gene promoter variation. *Behav Genet* 2000;**30**:295–301.

Treisman J, Gönczy P, Vashishtha M *et al.* A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* 1989;**59**:553–62.

Tseng GC, Oh MK, Rohlin L *et al.* Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 2001;**29**:2549–57.

Ubeda F, Wilkins JF. The Red Queen theory of recombination hotspots. *J Evol Biol* 2011;**24**:541–53.

Van Gelder RN, von Zastrow ME, Yool A *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci U S A* 1990;**87**:1663–7.

Ventriglia M, Bocchio Chiavetto L, Bonvicini C *et al.* Allelic variation in the human prodynorphin gene promoter and schizophrenia. *Neuropsychobiology* 2002;**46**:17–21.

Verlaan DJ, Ge B, Grundberg E *et al.* Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res* 2009;**19**:118–27.

Veyrieras J-B, Kudaravalli S, Kim SY *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008;**4**:e1000214.

Vierstra J, Rynes E, Sandstrom R *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* 2014;**346**:1007–12.

Voight BF, Kudaravalli S, Wen X *et al.* A map of recent positive selection in the human genome. *PLoS Biol* 2006;**4**:e72.

Volpe JJ, Adams RD. Cerebro-hepato-renal syndrome of Zellweger: an inherited disorder of neuronal migration. *Acta Neuropathol (Berl)* 1972;**20**:175–98.

Wang ET, Sandberg R, Luo S *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;**456**:470–6.

Wang J, Zhuang J, Iyer S *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012;**22**:1798–812.

Warrington JA, Nair A, Mahadevappa M *et al.* Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2000;**2**:143–7.

Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989;**44**:388–96.

Wegmann D, Kessner DE, Veeramah KR *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet* 2011, DOI: 10.1038/ng.894.

Welter D, MacArthur J, Morales J *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:D1001–6.

Westra H-J, Peters MJ, Esko T *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013;**45**:1238–43.

Wittkopp PJ, Haerum BK, Clark AG. Evolutionary changes in cis and trans gene regulation. *Nature* 2004;**430**:85–8.

Wittkopp PJ, True JR, Carroll SB. Reciprocal functions of the Drosophila yellow and ebony proteins in the development and evolution of pigment patterns. *Dev Camb Engl* 2002;**129**:1849–58.

Wolffe AP. Gene regulation. Insulating chromatin. *Curr Biol CB* 1994;**4**:85–7.

Workman C, Jensen LJ, Jarmer H *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 2002;**3**:research0048.

Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 2007;**8**:206–16.

Wray GA, Hahn MW, Abouheif E *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 2003;**20**:1377–419.

Wu L, Candille SI, Choi Y *et al.* Variation and genetic control of protein abundance in humans. *Nature* 2013;**499**:79–82.

Xie M, Hong C, Zhang B *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* 2013;**45**:836–41.

Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. *Trends Genet TIG* 2013;**29**:11–22.

Yang C, Bolotin E, Jiang T *et al.* Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* 2007;**389**:52–65.

Yan H, Yuan W, Velculescu VE *et al.* Allelic variation in human gene expression. *Science* 2002;**297**:1143.

Yeo G, Holste D, Kreiman G *et al.* Variation in alternative splicing across human tissues. *Genome Biol* 2004;**5**:R74.

Younger ST, Pertsemlidis A, Corey DR. Predicting potential miRNA target sites within gene promoters. *Bioorg Med Chem Lett* 2009;**19**:3791–4.

Yuh C-H, Brown CT, Livi CB *et al.* Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol* 2002;**246**:148–61.

Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* 2011;**25**:2227–41.

Zhang D, Cheng L, Badner JA *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet* 2010;**86**:411–9.

Zhang F-R, Huang W, Chen S-M *et al.* Genomewide association study of leprosy. *N Engl J Med* 2009a;**361**:2609–18.

Zhang K, Li JB, Gao Y *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 2009b;**6**:613–8.

Zhang R, Su B. MicroRNA regulation and the variability of human cortical gene expression. *Nucleic Acids Res* 2008, DOI: gkn431.

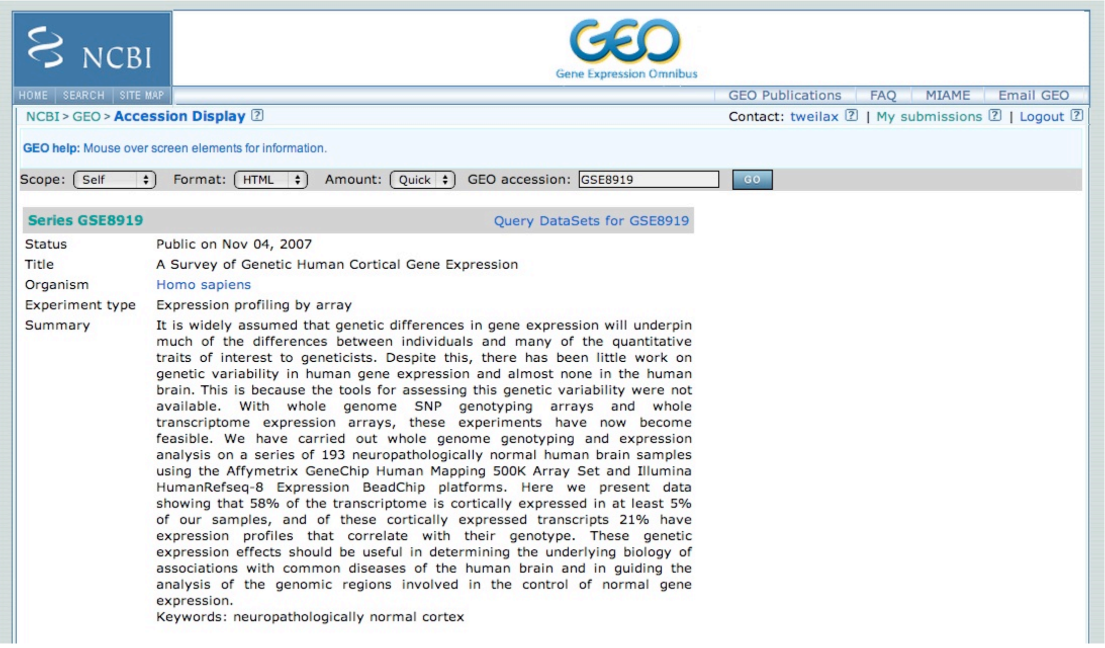
Zhang W, Duan S, Bleibel WK *et al.* Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum Genet* 2009c;**125**:81–93.

Zhang X, Zhang H, Oberdick J. Conservation of the developmentally regulated dendritic localization of a Purkinje cell-specific mRNA that encodes a G-protein modulator: comparison of rodent and human Pcp2(L7) gene structure and expression. *Brain Res Mol Brain Res* 2002;**105**:1–10.

Zhong H, Yang X, Kaplan LM *et al.* Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* 2010;**86**:581–91.

Zickler D, Kleckner N. Meiotic chromosomes: integrating structure and function. *Annu Rev Genet* 1999;**33**:603–754.

9. Appendix

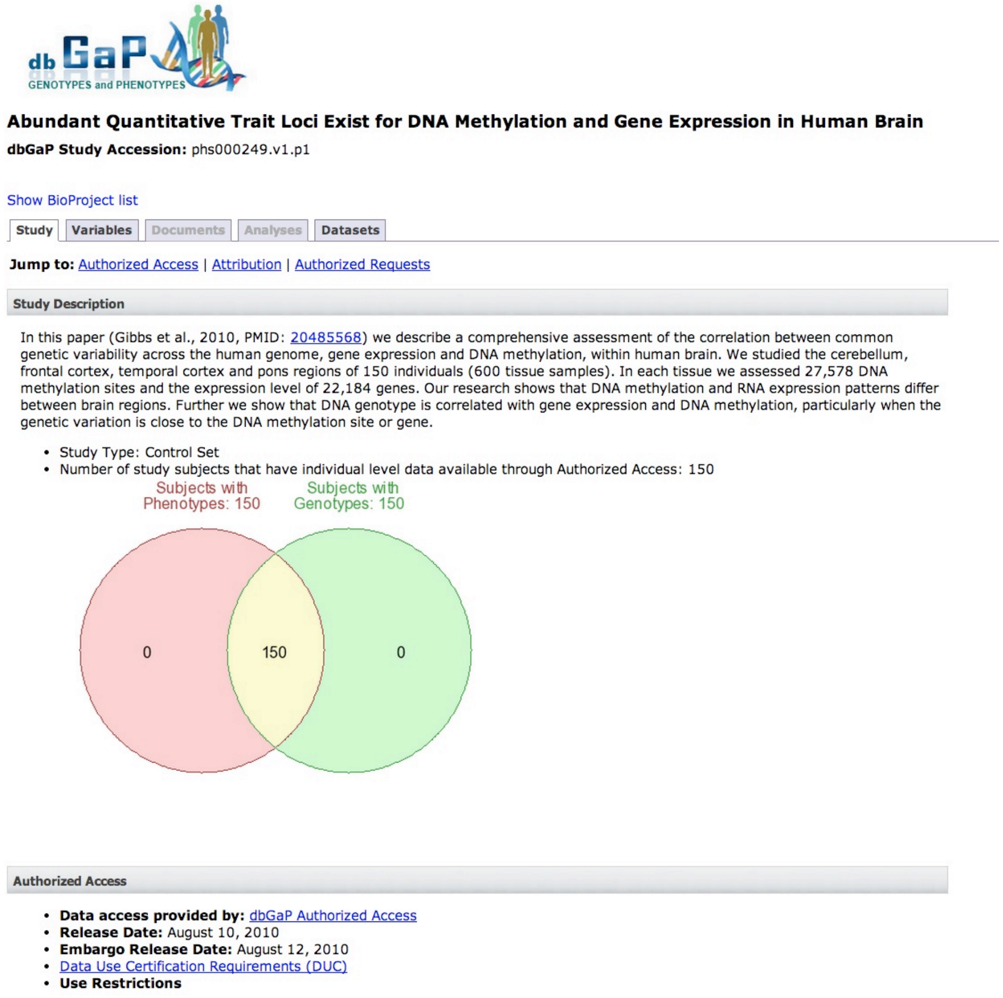

NCBI
HOME | SEARCH | SITE MAP
GEO
Gene Expression Omnibus
GEO Publications | FAQ | MIAME | Email GEO
NCBI > GEO > Accession Display
Contact: tweilax | My submissions | Logout
GEO help: Mouse over screen elements for information.
GO
Query DataSets for GSE8919

Scope: Self Format: HTML Amount: Quick GEO accession: GSE8919

Series GSE8919

Status	Public on Nov 04, 2007
Title	A Survey of Genetic Human Cortical Gene Expression
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	It is widely assumed that genetic differences in gene expression will underpin much of the differences between individuals and many of the quantitative traits of interest to geneticists. Despite this, there has been little work on genetic variability in human gene expression and almost none in the human brain. This is because the tools for assessing this genetic variability were not available. With whole genome SNP genotyping arrays and whole transcriptome expression arrays, these experiments have now become feasible. We have carried out whole genome genotyping and expression analysis on a series of 193 neuropathologically normal human brain samples using the Affymetrix GeneChip Human Mapping 500K Array Set and Illumina HumanRefseq-8 Expression BeadChip platforms. Here we present data showing that 58% of the transcriptome is cortically expressed in at least 5% of our samples, and of these cortically expressed transcripts 21% have expression profiles that correlate with their genotype. These genetic expression effects should be useful in determining the underlying biology of associations with common diseases of the human brain and in guiding the analysis of the genomic regions involved in the control of normal gene expression. Keywords: neuropathologically normal cortex

Figure 9.1: Figure is a screen capture for the GEO study page for the 'pilot' study cohort used in the eQTL analysis of mixed cortical tissues described in Chapter 2. GEO (Gene Expression Omnibus) is a public repository at the National Institutes of Health in the USA that allows public access to gene expression data used in studies (Edgar, Domrachev and Lash 2002; Barrett *et al.* 2007). The is for series accession GSE8919.



Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain

dbGaP Study Accession: phs000249.v1.p1

Jump to: Authorized Access | Attribution | Authorized Requests

Study Description

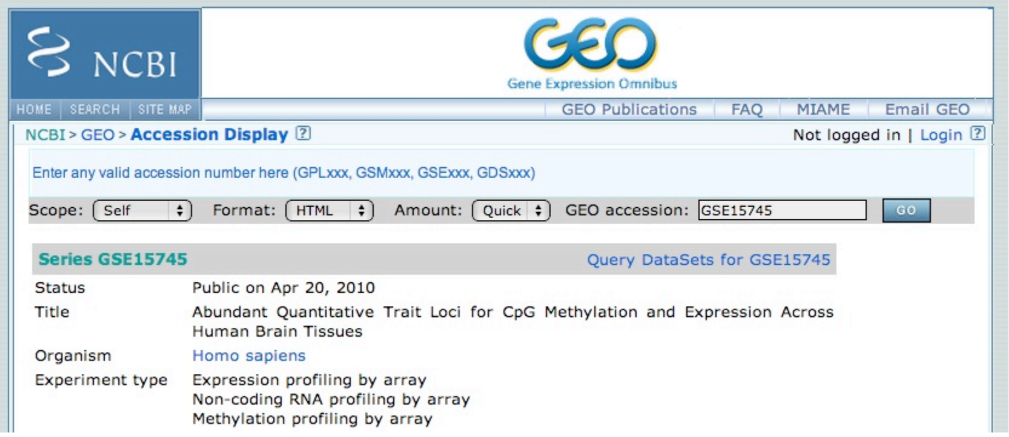
In this paper (Gibbs et al., 2010, PMID: 20485568) we describe a comprehensive assessment of the correlation between common genetic variability across the human genome, gene expression and DNA methylation, within human brain. We studied the cerebellum, frontal cortex, temporal cortex and pons regions of 150 individuals (600 tissue samples). In each tissue we assessed 27,578 DNA methylation sites and the expression level of 22,184 genes. Our research shows that DNA methylation and RNA expression patterns differ between brain regions. Further we show that DNA genotype is correlated with gene expression and DNA methylation, particularly when the genetic variation is close to the DNA methylation site or gene.

- Study Type: Control Set
- Number of study subjects that have individual level data available through Authorized Access: 150

Authorized Access

- **Data access provided by:** dbGaP Authorized Access
- **Release Date:** August 10, 2010
- **Embargo Release Date:** August 12, 2010
- Data Use Certification Requirements (DUC)
- **Use Restrictions**

Figure 9.2: Figure is a screen capture of the dbGaP study page for the initial NABEC cohort used in the eQTL analysis described in Chapter 3. dbGaP (data bases of Genotypes and Phenotyes) is a public repository at the National Institutes of Health in the USA that allows public access with appropriate approval to genotypes data used in studies (Mailman *et al.* 2007). This is for study accession phs000249.v1.p1.



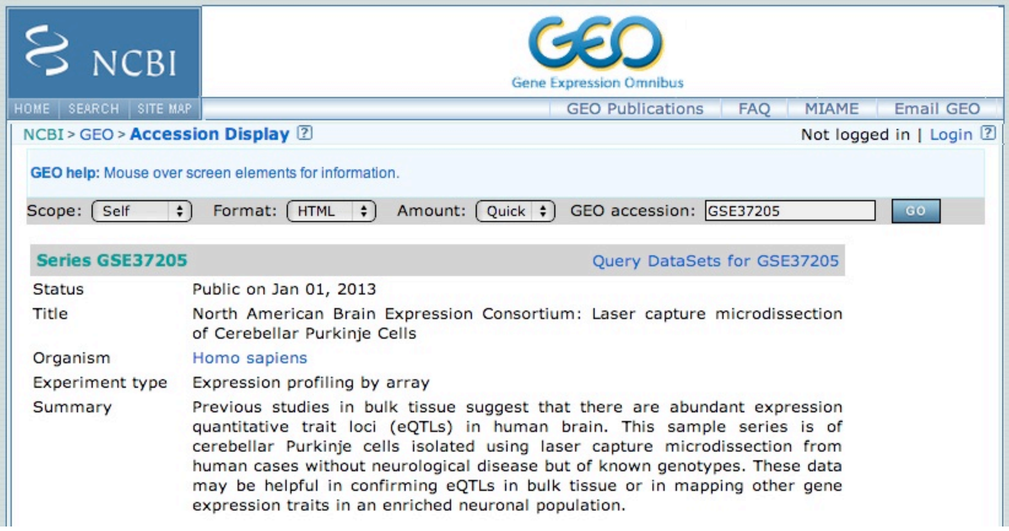
Enter any valid accession number here (GPLxxx, GSMxxx, GSExxx, GDSxxx)

GEO accession: GSE15745

Series GSE15745

Status	Public on Apr 20, 2010
Title	Abundant Quantitative Trait Loci for CpG Methylation and Expression Across Human Brain Tissues
Organism	Homo sapiens
Experiment type	Expression profiling by array Non-coding RNA profiling by array Methylation profiling by array

Figure 9.3: Figure is a screen capture for the GEO study page for the initial NABEC cohort used in the eQTL analysis described in Chapter 3. GEO (Gene Expression Omnibus) is a public repository at the National Institutes of Health in the USA that allows public access to gene expression data used in studies (Edgar, Domrachev and Lash 2002; Barrett *et al.* 2007). This is for series accession GSE15745.



GEO help: Mouse over screen elements for information.

GEO accession: GSE37205

Series GSE37205

Status	Public on Jan 01, 2013
Title	North American Brain Expression Consortium: Laser capture microdissection of Cerebellar Purkinje Cells
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Previous studies in bulk tissue suggest that there are abundant expression quantitative trait loci (eQTLs) in human brain. This sample series is of cerebellar Purkinje cells isolated using laser capture microdissection from human cases without neurological disease but of known genotypes. These data may be helpful in confirming eQTLs in bulk tissue or in mapping other gene expression traits in an enriched neuronal population.

Figure 9.4: Figure is a screen capture for the GEO study page for the Purkinje cell data from the NABEC cohort used in the eQTL analysis described in Chapter 4. GEO (Gene Expression Omnibus) is a public repository at the National Institutes of Health in the USA that allows public access to gene expression data used in studies (Edgar, Domrachev and Lash 2002; Barrett *et al.* 2007). This is for series accession GSE37205.

10. Additional Publications

1. Fung HC, Xiromerisiou G, **Gibbs JR**, Wu YR, Eerola J, Gourbali V, Hellström O, Chen CM, Duckworth J, Papadimitriou A, Tienari PJ, Hadjigeorgiou GM, Hardy J, Singleton AB. Association of tau haplotype-tagging polymorphisms with Parkinson's disease in diverse ethnic Parkinson's disease cohorts. Neurodegener Dis. 2006;3(6):327-33. PubMed PMID: 17192721.

2. Paisán-Ruíz C, Evans EW, Jain S, Xiromerisiou G, **Gibbs JR**, Eerola J, Gourbali V, Hellström O, Duckworth J, Papadimitriou A, Tienari PJ, Hadjigeorgiou GM, Singleton AB. Testing association between LRRK2 and Parkinson's disease and investigating linkage disequilibrium. J Med Genet. 2006 Feb;43(2):e9. PubMed PMID: 16467219.

3. **Gibbs JR**, Singleton A. Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS Genet. 2006 Oct 6;2(10):e150. Review. PubMed PMID: 17029559.

4. Fung HC, Scholz S, Matarin M, Simón-Sánchez J, Hernandez D, Britton A, **Gibbs JR**, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodríguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2006 Nov;5(11):911-6. PubMed PMID: 17052657.

5. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, **Gibbs JR**, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. 2007 Jan 1;16(1):1-14. Epub 2006 Nov 20. PubMed PMID: 17116639.

6. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, **Gibbs JR**, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, Crews C, Bruijn L, Rothstein J, Mora G, Restagno G, Chiò A, Singleton A, Hardy J, Traynor BJ. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2007 Apr;6(4):322-8. PubMed PMID: 17362836.

7. Matarín M, Brown WM, Scholz S, Simón-Sánchez J, Fung HC, Hernandez D, **Gibbs JR**, De Vrieze FW, Crews C, Britton A, Langefeld CD, Brott TG, Brown RD Jr, Worrall BB, Frankel M, Silliman S, Case LD, Singleton A, Hardy JA, Rich SS, Meschia JF. A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Lancet Neurol. 2007 May;6(5):414-20. PubMed PMID: 17434096.

8. Simon-Sanchez J, Scholz S, Matarin Mdel M, Fung HC, Hernandez D, **Gibbs JR**, Britton A, Hardy J, Singleton A. Genomewide SNP assay reveals mutations underlying Parkinson disease. Hum Mutat. 2008 Feb;29(2):315-22. PubMed PMID: 17994548.

9. Camargos S, Scholz S, Simón-Sánchez J, Paisán-Ruiz C, Lewis P, Hernandez D, Ding J, **Gibbs JR**, Cookson MR, Bras J, Guerreiro R, Oliveira CR, Lees A, Hardy J, Cardoso F, Singleton AB. DYT16, a novel young-onset dystonia-parkinsonism disorder: identification of a segregating mutation in the stress-response protein PRKRA. Lancet Neurol. 2008 Mar;7(3):207-15. doi: 10.1016/S1474-4422(08)70022-X. Epub 2008 Feb 1. PubMed PMID: 18243799.

10. Paisán-Ruíz C, Nath P, Washecka N, **Gibbs JR**, Singleton AB. Comprehensive analysis of LRRK2 in publicly available Parkinson's disease cases and neurologically normal controls. Hum Mutat. 2008 Apr;29(4):485-90. doi: 10.1002/humu.20668. PubMed PMID: 18213618.

11. Matarin M, Simon-Sanchez J, Fung HC, Scholz S, **Gibbs JR**, Hernandez DG, Crews C, Britton A, De Vrieze FW, Brott TG, Brown RD Jr, Worrall BB, Silliman S, Case LD, Hardy JA, Rich SS, Meschia JF, Singleton AB. Structural genomic variation in ischemic stroke. Neurogenetics. 2008 May;9(2):101-8. doi: 10.1007/s10048-008-0119-3. Epub 2008 Feb 21. PubMed PMID: 18288507.

12. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, Rafferty I, Lauretani F, Murray A, **Gibbs JR**, Paolisso G, Rafiq S, Simon-Sanchez J, Lango H, Scholz S, Weedon MN, Arepalli S, Rice N, Washecka N, Hurst A, Britton A, Henley W, van de Leemput J, Li R, Newman AB, Tranah G, Harris T, Panicker V, Dayan C, Bennett A, McCarthy MI, Ruokonen A, Jarvelin MR, Guralnik J, Bandinelli S, Frayling TM, Singleton A, Ferrucci L. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008 May 9;4(5):e1000072. doi: 10.1371/journal.pgen.1000072. PubMed PMID: 18464913.

13. van der Brug MP, Blackinton J, Chandran J, Hao LY, Lal A, Mazan-Mamczarz K, Martindale J, Xie C, Ahmad R, Thomas KJ, Beilina A, **Gibbs JR**, Ding J, Myers AJ, Zhan M, Cai H, Bonini NM, Gorospe M, Cookson MR. RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. Proc Natl Acad Sci U S A. 2008 Jul 22;105(29):10244-9. doi: 10.1073/pnas.0708518105. Epub 2008 Jul 14. PubMed PMID: 18626009.

14. Liu W, Ding J, **Gibbs JR**, Wang SJ, Hardy J, Singleton A. A simple and efficient algorithm for genome-wide homozygosity analysis in disease. Mol Syst Biol. 2009;5:304. doi: 10.1038/msb.2009.53. Epub 2009 Sep 15. PubMed PMID: 19756043.

15. Nalls MA, Simon-Sanchez J, **Gibbs JR**, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet. 2009 Mar;5(3):e1000415. doi: 10.1371/journal.pgen.1000415. Epub 2009 Mar 13. PubMed PMID: 19282984.

16. Webster JA, **Gibbs JR**, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, McCorquodale DS 3rd, Cuello C, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV; NACC-Neuropathology Group, Heward CB, Reiman EM, Stephan D, Hardy J, Myers AJ. Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet. 2009 Apr;84(4):445-58. doi: 10.1016/j.ajhg.2009.03.011. PubMed PMID: 19361613.

17. Chiò A, Schymick JC, Restagno G, Scholz SW, Lombardo F, Lai SL, Mora G, Fung HC, Britton A, Arepalli S, **Gibbs JR**, Nalls M, Berger S, Kwee LC, Oddone EZ, Ding J, Crews C, Rafferty I, Washecka N, Hernandez D, Ferrucci L, Bandinelli S, Guralnik J, Macciardi F, Torri F, Lupoli S, Chanock SJ, Thomas G, Hunter DJ, Gieger C, Wichmann HE, Calvo A, Mutani R, Battistini S, Giannini F, Caponnetto C, Mancardi GL, La Bella V, Valentino F, Monsurrò MR, Tedeschi G, Marinou K, Sabatelli M, Conte A, Mandrioli J, Sola P, Salvi F, Bartolomei I, Siciliano G, Carlesi C, Orrell RW, Talbot K, Simmons Z, Connor J, Pioro EP, Dunkley T, Stephan DA, Kasperaviciute D, Fisher EM, Jabonka S, Sendtner M, Beck M, Bruijn L, Rothstein J, Schmidt S, Singleton A, Hardy J, Traynor BJ. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Hum Mol Genet. 2009 Apr 15;18(8):1524-32. doi: 10.1093/hmg/ddp059. Epub 2009 Feb 4. PubMed PMID: 19193627.

18. Scholz SW, Houlden H, Schulte C, Sharma M, Li A, Berg D, Melchers A, Paudel R, **Gibbs JR**, Simon-Sanchez J, Paisan-Ruiz C, Bras J, Ding J, Chen H, Traynor BJ, Arepalli S, Zonozi RR, Revesz T, Holton J, Wood N, Lees A, Oertel W, Wüllner U, Goldwurm S, Pellecchia MT, Illig T, Riess O, Fernandez HH, Rodriguez RL, Okun MS, Poewe W, Wenning GK, Hardy JA, Singleton AB, Del Sorbo F, Schneider S, Bhatia KP, Gasser T. SNCA variants are associated with increased risk for multiple system atrophy. Ann Neurol. 2009 May;65(5):610-4. doi: 10.1002/ana.21685. Erratum in: Ann Neurol. 2010 Feb;67(2):277. Del Sorbo, Francesca [added]; Schneider, Susanne [added]; Bhatia, Kailash P [added]. PubMed PMID: 19475667.

19. Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, **Gibbs JR**, Launer L, Hardy J, Singleton AB. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics. 2009 Jul;10(3):183-90. doi: 10.1007/s10048-009-0182-4. Epub 2009 Mar 7. PubMed PMID: 19271249.

20. International Stroke Genetics Consortium; Wellcome Trust Case-Control Consortium 2. Failure to validate association between 12p13

variants and ischemic stroke. N Engl J Med. 2010 Apr 22;362(16):1547-50.

21. Laaksovirta H, Peuralinna T, Schymick JC, Scholz SW, Lai SL, Myllykangas L, Sulkava R, Jansson L, Hernandez DG, **Gibbs JR**, Nalls MA, Heckerman D, Tienari PJ, Traynor BJ. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. Lancet Neurol. 2010 Oct;9(10):978-85. doi: 10.1016/S1474-4422(10)70184-8. PubMed PMID: 20801718.

22. Johnson JO, **Gibbs JR**, Van Maldergem L, Houlden H, Singleton AB. Exome sequencing in Brown-Vialetto-van Laere syndrome. Am J Hum Genet. 2010 Oct 8;87(4):567-9; author reply 569-70. doi: 10.1016/j.ajhg.2010.05.021. PubMed PMID: 20920669.

23. Singleton AB, **Gibbs JR**. Another locus, a new method. Brain. 2010 Dec;133(Pt 12):3492-3. doi: 10.1093/brain/awq331. PubMed PMID: 21126992.

24. Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, **Gibbs JR**, Brunetti M, Gronka S, Wuu J, Ding J, McCluskey L, Martinez-Lage M, Falcone D, Hernandez DG, Arepalli S, Chong S, Schymick JC, Rothstein J, Landi F, Wang YD, Calvo A, Mora G, Sabatelli M, Monsurrò MR, Battistini S, Salvi F, Spataro R, Sola P, Borghero G; ITALSGEN Consortium, Galassi G, Scholz SW, Taylor JP, Restagno G, Chiò A, Traynor BJ. Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron. 2010 Dec 9;68(5):857-64. doi: 10.1016/j.neuron.2010.11.036. Erratum in: Neuron. 2011 Jan 27;69(2):397. PubMed PMID: 21145000.

25. Hernandez DG, Nalls MA, **Gibbs JR**, Arepalli S, van der Brug M, Chong S, Moore M, Longo DL, Cookson MR, Traynor BJ, Singleton AB. Distinct DNA methylation changes highly correlated with chronological age in the human brain. Hum Mol Genet. 2011 Mar 15;20(6):1164-72. doi: 10.1093/hmg/ddq561. Epub 2011 Jan 7. PubMed PMID: 21216877.

26. International Parkinson Disease Genomics Consortium. A two-stage meta-analysis identifies several new loci for Parkinson’s disease. PLoS Genet 2011 Jun;7(6):e1002142.

27. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, **Gibbs JR**, Harries LW, Chong S, Moore M, Weedon MN, Guralnik JM, Bandinelli S, Murray A, Ferrucci L, Singleton AB, Melzer D, Frayling TM. Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum Mol Genet. 2011 Oct 15;20(20):4082-92. doi: 10.1093/hmg/ddr328. Epub 2011 Jul 28. PubMed PMID: 21798870.

28. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, **Gibbs JR**, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L, Kalimo H, Paetau A, Abramzon Y, Remes AM, Kaganovich A,

Scholz SW, Duckworth J, Ding J, Harmer DW, Hernandez DG, Johnson JO, Mok K, Ryten M, Trabzuni D, Guerreiro RJ, Orrell RW, Neal J, Murray A, Pearson J, Jansen IE, Sondervan D, Seelaar H, Blake D, Young K, Halliwell N, Callister JB, Toulson G, Richardson A, Gerhard A, Snowden J, Mann D, Neary D, Nalls MA, Peuralinna T, Jansson L, Isoviita VM, Kaivorinne AL, Hölttä-Vuori M, Ikonen E, Sulkava R, Benatar M, Wuu J, Chiò A, Restagno G, Borghero G, Sabatelli M; ITALSGEN Consortium, Heckerman D, Rogaeva E, Zinman L, Rothstein JD, Sendtner M, Drepper C, Eichler EE, Alkan C, Abdullaev Z, Pack SD, Dutra A, Pak E, Hardy J, Singleton A, Williams NM, Heutink P, Pickering-Brown S, Morris HR, Tienari PJ, Traynor BJ. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011 Oct 20;72(2):257-68. doi: 10.1016/j.neuron.2011.09.010. Epub 2011 Sep 21. PubMed PMID: 21944779.

29. Guerreiro RJ, Lohmann E, Kinsella E, Brás JM, Luu N, Gurunlian N, Dursun B, Bilgic B, Santana I, Hanagasi H, Gurvit H, **Gibbs JR**, Oliveira C, Emre M, Singleton A. Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. Neurobiol Aging. 2012 May;33(5):1008.e17-23. doi: 10.1016/j.neurobiolaging.2011.10.009. Epub 2011 Dec 6. PubMed PMID: 22153900.

30. Landouré G, Sullivan JM, Johnson JO, Munns CH, Shi Y, Diallo O, **Gibbs JR**, Gaudet R, Ludlow CL, Fischbeck KH, Traynor BJ, Burnett BG, Sumner CJ. Exome sequencing identifies a novel TRPV4 mutation in a CMT2C family. Neurology. 2012 Jul 10;79(2):192-4. doi: 10.1212/WNL.0b013e31825f04b2. Epub 2012 Jun 6. PubMed PMID: 22675077.

31. Simón-Sánchez J, Kilarski LL, Nalls MA, Martinez M, Schulte C, Holmans P; International Parkinson's Disease Genomics Consortium; Wellcome Trust Case Control Consortium, Gasser T, Hardy J, Singleton AB, Wood NW, Brice A, Heutink P, Williams N, Morris HR. Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinson's disease. PLoS One. 2012;7(3):e28787.

32. Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide BM, Schjeide LM, Meissner E, Zauft U, Allen NC, Liu T, Schilling M, Anderson KJ, Beecham G, Berg D, Biernacka JM, Brice A, DeStefano AL, Do CB, Eriksson N, Factor SA, Farrer MJ, Foroud T, Gasser T, Hamza T, Hardy JA, Heutink P, Hill-Burns EM, Klein C, Latourelle JC, Maraganore DM, Martin ER, Martinez M, Myers RH, Nalls MA, Pankratz N, Payami H, Satake W, Scott WK, Sharma M, Singleton AB, Stefansson K, Toda T, Tung JY, Vance J, Wood NW, Zabetian CP; 23andMe Genetic Epidemiology of Parkinson's Disease Consortium; International Parkinson's Disease Genomics Consortium; Parkinson's Disease GWAS Consortium; Wellcome Trust Case Control Consortium 2), Young P, Tanzi RE, Khoury MJ, Zipp F, Lehrach H, Ioannidis JP, Bertram L. Comprehensive research synopsis and systematic meta-

analyses in Parkinson's disease genetics: The PDGene database. PLoS Genet. 2012;8(3):e1002548.

33. Sailer A, Scholz SW, **Gibbs JR**, Tucci A, Johnson JO, Wood NW, Plagnol V, Hummerich H, Ding J, Hernandez D, Hardy J, Federoff HJ, Traynor BJ, Singleton AB, Houlden H. Exome sequencing in an SCA14 family demonstrates its utility in diagnosing heterogeneous diseases. Neurology. 2012 Jul 10;79(2):127-31. doi: 10.1212/WNL.0b013e31825f048e. Epub 2012 Jun 6. PubMed PMID: 22675081.

34. Johnson JO, **Gibbs JR**, Megarbane A, Urtizberea JA, Hernandez DG, Foley AR, Arepalli S, Pandraud A, Simón-Sánchez J, Clayton P, Reilly MM, Muntoni F, Abramzon Y, Houlden H, Singleton AB. Exome sequencing reveals riboflavin transporter mutations as a cause of motor neuron disease. Brain. 2012 Sep;135(Pt 9):2875-82. doi: 10.1093/brain/aws161. Epub 2012 Jun 26. PubMed PMID: 22740598.

35. Keller MF, Saad M, Bras J, Bettella F, Nicolaou N, Simón-Sánchez J, Mittag F, Büchel F, Sharma M, **Gibbs JR**, Schulte C, Moskvina V, Durr A, Holmans P, Kilarski LL, Guerreiro R, Hernandez DG, Brice A, Ylikotila P, Stefánsson H, Majamaa K, Morris HR, Williams N, Gasser T, Heutink P, Wood NW, Hardy J, Martinez M, Singleton AB, Nalls MA; International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. Hum Mol Genet. 2012 Nov 15;21(22):4996-5009. doi: 10.1093/hmg/dds335. Epub 2012 Aug 13. Erratum in: Hum Mol Genet. 2013 Jul 15;22(14):2973. Hum Mol Genet. 2013 Apr 15;22(8):1696. PubMed PMID: 22892372.

36. Mittag F, Büchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, Simón-Sánchez J, Nalls MA, Keller M, Hernandez DG, **Gibbs JR**, Lesage S, Brice A, Heutink P, Martinez M, Wood NW, Hardy J, Singleton AB, Zell A, Gasser T, Sharma M; International Parkinson’s Disease Genomics Consortium. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum Mutat. 2012 Dec;33(12):1708-18. doi: 10.1002/humu.22161. Epub 2012 Aug 3. PubMed PMID: 22777693.

37. Guerreiro RJ, Lohmann E, Brás JM, **Gibbs JR**, Rohrer JD, Gurunlian N, Dursun B, Bilgic B, Hanagasi H, Gurvit H, Emre M, Singleton A, Hardy J. Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia-like syndrome without bone involvement. JAMA Neurol. 2013 Jan;70(1):78-84. doi: 10.1001/jamaneurol.2013.579. PubMed PMID: 23318515.

38. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JS, Younkin S, Hazrati L, Collinge J, Pocock J, Lashley T, Williams J, Lambert JC, Amouyel P, Goate A, Rademakers R, Morgan K, Powell J, St George-Hyslop P, Singleton A,

Hardy J; Alzheimer Genetic Analysis Group. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013 Jan 10;368(2):117-27.

39. Hammer MB, Eleuch-Fayache G, Schottlaender LV, Nehdi H, **Gibbs JR**, Arepalli SK, Chong SB, Hernandez DG, Sailer A, Liu G, Mistry PK, Cai H, Shrader G, Sassi C, Bouhlal Y, Houlden H, Hentati F, Amouri R, Singleton AB. Mutations in GBA2 cause autosomal-recessive cerebellar ataxia with spasticity. Am J Hum Genet. 2013 Feb 7;92(2):245-51. doi: 10.1016/j.ajhg.2012.12.012. Epub 2013 Jan 17. PubMed PMID: 23332917.

40. Holmans P, Moskvina V, Jones L, Sharma M; International Parkinson's Disease Genomics Consortium, Vedernikov A, Buchel F, Saad M, Bras JM, Bettella F, Nicolaou N, Simón-Sánchez J, Mittag F, **Gibbs JR**, Schulte C, Durr A, Guerreiro R, Hernandez D, Brice A, Stefánsson H, Majamaa K, Gasser T, Heutink P, Wood NW, Martinez M, Singleton AB, Nalls MA, Hardy J, Morris HR, Williams NM. A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson's disease. Hum Mol Genet. 2013 Mar 1;22(5):1039-49. doi: 10.1093/hmg/dds492. Epub 2012 Dec 7. Erratum in: Hum Mol Genet. 2014 Jan 15;23(2):562. Sadd, Mohamad [corrected to Saad, Mohamad]. PubMed PMID: 23223016.

41. Hammer MB, Eleuch-Fayache G, **Gibbs JR**, Arepalli SK, Chong SB, Sassi C, Bouhlal Y, Hentati F, Amouri R, Singleton AB. Exome sequencing: an efficient diagnostic tool for complex neurodegenerative disorders. Eur J Neurol. 2013 Mar;20(3):486-92. doi: 10.1111/j.1468-1331.2012.03883.x. Epub 2012 Oct 9. PubMed PMID: 23043354.

42. Ramasamy A, Trabzuni D, **Gibbs JR**, Dillman A, Hernandez DG, Arepalli S, Walker R, Smith C, Ilori GP, Shabalin AA, Li Y, Singleton AB, Cookson MR; NABEC, Hardy J; UKBEC, Ryten M, Weale ME. Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. Nucleic Acids Res. 2013 Apr;41(7):e88. doi: 10.1093/nar/gkt069. Epub 2013 Feb 21. PubMed PMID: 23435227.

43. Dillman AA, Hauser DN, **Gibbs JR**, Nalls MA, McCoy MK, Rudenko IN, Galter D, Cookson MR. mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. Nat Neurosci. 2013 Apr;16(4):499-506. doi: 10.1038/nn.3332. Epub 2013 Feb 17. PubMed PMID: 23416452.

44. Wood AR, Perry JR, Tanaka T, Hernandez DG, Zheng HF, Melzer D, **Gibbs JR**, Nalls MA, Weedon MN, Spector TD, Richards JB, Bandinelli S, Ferrucci L, Singleton AB, Frayling TM. Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One. 2013 May 16;8(5):e64343. doi: 10.1371/journal.pone.0064343. Print 2013. PubMed PMID: 23696881.

45. Pichler I, Del Greco M F, Gögele M, Lill CM, Bertram L, Do CB, Eriksson N, Foroud T, Myers RH; PD GWAS Consortium, Nalls M, Keller MF; International Parkinson's Disease Genomics Consortium; Wellcome Trust Case Control Consortium 2, Benyamin B, Whitfield JB; Genetics of Iron Status Consortium, Pramstaller PP, Hicks AA, Thompson JR, Minelli C. Serum iron levels and the risk of Parkinson disease: a mendelian randomization study. PLoS Med. 2013;10(6):e1001462. doi: 10.1371/journal.pmed.1001462. Epub 2013 Jun 4. Erratum in: PLoS Med. 2013 Jun;10(6).

46. Trabzuni D, Ramasamy A, Imran S, Walker R, Smith C, Weale ME, Hardy J, Ryten M; North American Brain Expression Consortium. Widespread sex differences in gene expression and splicing in the adult human brain. Nat Commun. 2013;4:2771. doi: 10.1038/ncomms3771. PubMed PMID: 24264146.

47. Klebe S, Golmard JL, Nalls MA, Saad M, Singleton AB, Bras JM, Hardy J, Simon-Sanchez J, Heutink P, Kuhlenbäumer G, Charfi R, Klein C, Hagenah J, Gasser T, Wurster I, Lesage S, Lorenz D, Deuschl G, Durif F, Pollak P, Damier P, Tison F, Durr A, Amouyel P, Lambert JC, Tzourio C, Maubaret C, Charbonnier-Beaupel F, Tahiri K, Vidailhet M, Martinez M, Brice A, Corvol JC; French Parkinson's Disease Genetics Study Group; International Parkinson's Disease Genomics Consortium (IPDGC). The Val158Met COMT polymorphism is a modifier of the age at onset in Parkinson's disease with a sexual dimorphism. J Neurol Neurosurg Psychiatry. 2013 Jun;84(6):666-73.

48. Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA, Ng MC, Adeyemo AA, Allison MA, Bielak LF, Chen G, Graff M, Irvin MR, Rhie SK, Li G, Liu Y, Liu Y, Lu Y, Nalls MA, Sun YV, Wojczynski MK, Yanek LR, Aldrich MC, Ademola A, Amos CI, Bandera EV, Bock CH, Britton A, Broeckel U, Cai Q, Caporaso NE, Carlson CS, Carpten J, Casey G, Chen WM, Chen F, Chen YD, Chiang CW, Coetzee GA, Demerath E, Deming-Halverson SL, Driver RW, Dubbert P, Feitosa MF, Feng Y, Freedman BI, Gillanders EM, Gottesman O, Guo X, Haritunians T, Harris T, Harris CC, Hennis AJ, Hernandez DG, McNeill LH, Howard TD, Howard BV, Howard VJ, Johnson KC, Kang SJ, Keating BJ, Kolb S, Kuller LH, Kutlar A, Langefeld CD, Lettre G, Lohman K, Lotay V, Lyon H, Manson JE, Maixner W, Meng YA, Monroe KR, Morhason-Bello I, Murphy AB, Mychaleckyj JC, Nadukuru R, Nathanson KL, Nayak U, N'diaye A, Nemesure B, Wu SY, Leske MC, Neslund-Dudas C, Neuhouser M, Nyante S, Ochs-Balcom H, Ogunniyi A, Ogundiran TO, Ojengbede O, Olopade OI, Palmer JR, Ruiz-Narvaez EA, Palmer ND, Press MF, Rampersaud E, Rasmussen-Torvik LJ, Rodriguez-Gil JL, Salako B, Schadt EE, Schwartz AG, Shriner DA, Siscovick D, Smith SB, Wassertheil-Smoller S, Speliotes EK, Spitz MR, Sucheston L, Taylor H, Tayo BO, Tucker MA, Van Den Berg DJ, Edwards DR, Wang Z, Wiencke JK, Winkler TW, Witte JS, Wrensch M, Wu X, Yang JJ, Levin AM, Young TR, Zakai NA, Cushman M, Zanetti KA, Zhao JH, Zhao W, Zheng Y, Zhou J, Ziegler RG, Zmuda JM, Fernandes JK, Gilkeson GS, Kamen DL, Hunt KJ, Spruill IJ, Ambrosone CB, Ambs S, Arnett DK, Atwood L, Becker DM, Berndt SI,

Bernstein L, Blot WJ, Borecki IB, Bottinger EP, Bowden DW, Burke G, Chanock SJ, Cooper RS, Ding J, Duggan D, Evans MK, Fox C, Garvey WT, Bradfield JP, Hakonarson H, Grant SF, Hsing A, Chu L, Hu JJ, Huo D, Ingles SA, John EM, Jordan JM, Kabagambe EK, Kardia SL, Kittles RA, Goodman PJ, Klein EA, Kolonel LN, Le Marchand L, Liu S, McKnight B, Millikan RC, Mosley TH, Padhukasahasram B, Williams LK, Patel SR, Peters U, Pettaway CA, Peyser PA, Psaty BM, Redline S, Rotimi CN, Rybicki BA, Sale MM, Schreiner PJ, Signorello LB, Singleton AB, Stanford JL, Strom SS, Thun MJ, Vitolins M, Zheng W, Moore JH, Williams SM, Ketkar S, Zhu X, Zonderman AB; NABEC Consortium; UKBEC Consortium; BioBank Japan Project; AGEN Consortium, Kooperberg C, Papanicolaou GJ, Henderson BE, Reiner AP, Hirschhorn JN, Loos RJ, North KE, Haiman CA. A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. Nat Genet. 2013 Jun;45(6):690-6. doi: 10.1038/ng.2608. Epub 2013 Apr 14. PubMed PMID: 23583978.

49. Salehi B, Preuss N, van der Veen JW, Shen J, Neumeister A, Drevets WC, Hodgkinson C, Goldman D, Wendland JR, Singleton A, **Gibbs JR**, Cookson MR, Hasler G. Age-modulated association between prefrontal NAA and the BDNF gene. Int J Neuropsychopharmacol. 2013 Jul;16(6):1185-93. doi: 10.1017/S1461145712001204. Epub 2012 Dec 20. PubMed PMID: 23253771.

50. Moskvina V, Harold D, Russo G, Vedernikov A, Sharma M, Saad M, Holmans P, Bras JM, Bettella F, Keller MF, Nicolaou N, Simón-Sánchez J, **Gibbs JR**, Schulte C, Durr A, Guerreiro R, Hernandez D, Brice A, Stefánsson H, Majamaa K, Gasser T, Heutink P, Wood N, Martinez M, Singleton AB, Nalls MA, Hardy J, Owen MJ, O'Donovan MC, Williams J, Morris HR, Williams NM; IPDGC and GERAD Investigators. Analysis of genome-wide association studies of Alzheimer disease and of Parkinson disease to determine if these 2 diseases share a common genetic risk. JAMA Neurol. 2013 Oct;70(10):1268-76. PubMed PMID: 23921447.

51. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S, Jeng AT, Cooper B, Skorupa T, Carrell D, Levitch D, Hsu S, Choi J, Ryten M; UK Brain Expression Consortium, Hardy J, Ryten M, Trabzuni D, Weale ME, Ramasamy A, Smith C, Sassi C, Bras J, **Gibbs JR**, Hernandez DG, Lupton MK, Powell J, Forabosco P, Ridge PG, Corcoran CD, Tschanz JT, Norton MC, Munger RG, Schmutz C, Leary M, Demirci FY, Bamne MN, Wang X, Lopez OL, Ganguli M, Medway C, Turton J, Lord J, Braae A, Barber I, Brown K; Alzheimer's Research UK Consortium, Passmore P, Craig D, Johnston J, McGuinness B, Todd S, Heun R, Kölsch H, Kehoe PG, Hooper NM, Vardy ER, Mann DM, Pickering-Brown S, Brown K, Kalsheker N, Lowe J, Morgan K, David Smith A, Wilcock G, Warden D, Holmes C, Pastor P, Lorenzo-Betancor O, Brkanac Z, Scott E, Topol E, Morgan K, Rogaeva E, Singleton AB, Hardy J, Kamboh MI, St George-Hyslop P, Cairns N, Morris JC, Kauwe JS, Goate AM. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's

disease. Nature. 2014 Jan 23;505(7484):550-4. doi: 10.1038/nature12825. Epub 2013 Dec 11. PubMed PMID: 24336208.

52. Johnson JO, Pioro EP, Boehringer A, Chia R, Feit H, Renton AE, Pliner HA, Abramzon Y, Marangi G, Winborn BJ, **Gibbs JR**, Nalls MA, Morgan S, Shoai M, Hardy J, Pittman A, Orrell RW, Malaspina A, Sidle KC, Fratta P, Harms MB, Baloh RH, Pestronk A, Weihl CC, Rogaeva E, Zinman L, Drory VE, Borghero G, Mora G, Calvo A, Rothstein JD; ITALSGEN Consortium, Drepper C, Sendtner M, Singleton AB, Taylor JP, Cookson MR, Restagno G, Sabatelli M, Bowser R, Chiò A, Traynor BJ. Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. Nat Neurosci. 2014 May;17(5):664-6. doi: 10.1038/nn.3688. Epub 2014 Mar 30. PubMed PMID: 24686783.

53. Sassi C, Guerreiro R, **Gibbs R**, Ding J, Lupton MK, Troakes C, Lunnon K, Al-Sarraj S, Brown KS, Medway C, Lord J, Turton J, Mann D, Snowden J, Neary D, Harris J, Bras J; ARUK Consortium, Morgan K, Powell JF, Singleton A, Hardy J. Exome sequencing identifies 2 novel presenilin 1 mutations (p.L166V and p.S230R) in British early-onset Alzheimer's disease. Neurobiol Aging. 2014 Oct;35(10):2422.e13-6. doi: 10.1016/j.neurobiolaging.2014.04.026. Epub 2014 May 2. PubMed PMID: 24880964.

54. Benitez BA, Jin SC, Guerreiro R, Graham R, Lord J, Harold D, Sims R, Lambert JC, **Gibbs JR**, Bras J, Sassi C, Harari O, Bertelsen S, Lupton MK, Powell J, Bellenguez C, Brown K, Medway C, Haddick PC, van der Brug MP, Bhangale T, Ortmann W, Behrens T, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD, Haines JL, Turton J, Braae A, Barber I, Fagan AM, Holtzman DM, Morris JC; 3C Study Group; EADI consortium; Alzheimer's Disease Genetic Consortium (ADGC); Alzheimer's Disease Neuroimaging Initiative (ADNI); GERAD Consortium, Williams J, Kauwe JS, Amouyel P, Morgan K, Singleton A, Hardy J, Goate AM, Cruchaga C. Missense variant in TREML2 protects against Alzheimer's disease. Neurobiol Aging. 2014 Jun;35(6):1510.e19-26. doi: 10.1016/j.neurobiolaging.2013.12.010. Epub 2013 Dec 21. PubMed PMID: 24439484.

55. Dong J, Gao J, Nalls M, Gao X, Huang X, Han J, Singleton AB, Chen H; International Parkinson's Disease Genomics Consortium (IPDGC). Susceptibility loci for pigmentation and melanoma in relation to Parkinson's disease. Neurobiol Aging. 2014 Jun;35(6):1512.e5-10. doi: 10.1016/j.neurobiolaging.2013.12.020. Epub 2013 Dec 27. PubMed PMID: 24439955.

56. Mencacci NE, Isaias IU, Reich MM, Ganos C, Plagnol V, Polke JM, Bras J, Hersheson J, Stamelou M, Pittman AM, Noyce AJ, Mok KY, Opladen T, Kunstmann E, Hodecker S, Münchau A, Volkmann J, Samnick S, Sidle K, Nanji T, Sweeney MG, Houlden H, Batla A, Zecchinelli AL, Pezzoli G, Marotta G, Lees A, Alegria P, Krack P, Cormier-Dequaire F, Lesage S, Brice A, Heutink P, Gasser T, Lubbe SJ, Morris HR, Taba P, Koks S, Majounie E, **Raphael Gibbs J**, Singleton A, Hardy J, Klebe S, Bhatia KP, Wood NW; International

Parkinson’s Disease Genomics Consortium and UCL-exomes consortium. Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. Brain. 2014 Sep;137(Pt 9):2480-92. doi: 10.1093/brain/awu179. Epub 2014 Jul 2. PubMed PMID: 24993959.

57. Johnson JO, Glynn SM, **Gibbs JR**, Nalls MA, Sabatelli M, Restagno G, Drory VE, Chiò A, Rogaeva E, Traynor BJ. Mutations in the CHCHD10 gene are a common cause of familial amyotrophic lateral sclerosis. Brain. 2014 Dec;137(Pt 12):e311. doi: 10.1093/brain/awu265. Epub 2014 Sep 26. PubMed PMID: 25261972.

58. Johnson JO, Stevanin G, van de Leemput J, Hernandez DG, Arepalli S, Forlani S, Zonozi R, **Gibbs JR**, Brice A, Durr A, Singleton AB. A 7.5-Mb duplication at chromosome 11q21-11q22.3 is associated with a novel spastic ataxia syndrome. Mov Disord. 2015 Feb;30(2):262-6. doi: 10.1002/mds.26059. Epub 2014 Dec 27. PubMed PMID: 25545641.

59. Sassi C, Guerreiro R, **Gibbs R**, Ding J, Lupton MK, Troakes C, Al-Sarraj S, Niblock M, Gallo JM, Adnan J, Killick R, Brown KS, Medway C, Lord J, Turton J, Bras J; Alzheimer's Research UK Consortium, Morgan K, Powell JF, Singleton A, Hardy J. Investigating the role of rare coding variability in Mendelian dementia genes (APP, PSEN1, PSEN2, GRN, MAPT, and PRNP) in late-onset Alzheimer's disease. Neurobiol Aging. 2014 Dec;35(12):2881.e1-6. doi: 10.1016/j.neurobiolaging.2014.06.002. Epub 2014 Jun 16. PubMed PMID: 25104557.

60. Nalls MA, Bras J, Hernandez DG, Keller MF, Majounie E, Renton AE, Saad M, Jansen I, Guerreiro R, Lubbe S, Plagnol V, **Gibbs JR**, Schulte C, Pankratz N, Sutherland M, Bertram L, Lill CM, DeStefano AL, Faroud T, Eriksson N, Tung JY, Edsall C, Nichols N, Brooks J, Arepalli S, Pliner H, Letson C, Heutink P, Martinez M, Gasser T, Traynor BJ, Wood N, Hardy J, Singleton AB; International Parkinson's Disease Genomics Consortium (IPDGC); Parkinson's Disease meta-analysis consortium. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiol Aging. 2015 Mar;36(3):1605.e7-12. doi: 10.1016/j.neurobiolaging.2014.07.028. Epub 2014 Aug 4. PubMed PMID: 25444595.

61. Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, De T; UK Brain Expression Consortium; North American Brain Expression Consortium, Coin L, de Silva R, Cookson MR, Singleton AB, Hardy J, Ryten M, Weale ME. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014 Oct;17(10):1418-28. doi: 10.1038/nn.3801. Epub 2014 Aug 31. PubMed PMID: 25174004.

62. Coffee and Caffeine Genetics Consortium, Cornelis MC, Byrne EM, Esko T, Nalls MA, Ganna A, Paynter N, Monda KL, Amin N, Fischer K, Renstrom F, Ngwa JS, Huikari V, Cavadino A, Nolte IM, Teumer A, Yu K, Marques-Vidal P, Rawal R, Manichaikul A, Wojczynski MK, Vink JM, Zhao JH, Burlutsky G, Lahti J, Mikkilä V, Lemaitre RN, Eriksson J,

Musani SK, Tanaka T, Geller F, Luan J, Hui J, Mägi R, Dimitriou M, Garcia ME, Ho WK, Wright MJ, Rose LM, Magnusson PK, Pedersen NL, Couper D, Oostra BA, Hofman A, Ikram MA, Tiemeier HW, Uitterlinden AG, van Rooij FJ, Barroso I, Johansson I, Xue L, Kaakinen M, Milani L, Power C, Snieder H, Stolk RP, Baumeister SE, Biffar R, Gu F, Bastardot F, Kutalik Z, Jacobs DR Jr, Forouhi NG, Mihailov E, Lind L, Lindgren C, Michaëlsson K, Morris A, Jensen M, Khaw KT, Luben RN, Wang JJ, Männistö S, Perälä MM, Kähönen M, Lehtimäki T, Viikari J, Mozaffarian D, Mukamal K, Psaty BM, Döring A, Heath AC, Montgomery GW, Dahmen N, Carithers T, Tucker KL, Ferrucci L, Boyd HA, Melbye M, Treur JL, Mellström D, Hottenga JJ, Prokopenko I, Tönjes A, Deloukas P, Kanoni S, Lorentzon M, Houston DK, Liu Y, Danesh J, Rasheed A, Mason MA, Zonderman AB, Franke L, Kristal BS; International Parkinson's Disease Genomics Consortium (IPDGC); North American Brain Expression Consortium (NABEC); UK Brain Expression Consortium (UKBEC), Karjalainen J, Reed DR, Westra HJ, Evans MK, Saleheen D, Harris TB, Dedoussis G, Curhan G, Stumvoll M, Beilby J, Pasquale LR, Feenstra B, Bandinelli S, Ordovas JM, Chan AT, Peters U, Ohlsson C, Gieger C, Martin NG, Waldenberger M, Siscovick DS, Raitakari O, Eriksson JG, Mitchell P, Hunter DJ, Kraft P, Rimm EB, Boomsma DI, Borecki IB, Loos RJ, Wareham NJ, Vollenweider P, Caporaso N, Grabe HJ, Neuhouser ML, Wolffenbuttel BH, Hu FB, Hyppönen E, Järvelin MR, Cupples LA, Franks PW, Ridker PM, van Duijn CM, Heiss G, Metspalu A, North KE, Ingelsson E, Nettleton JA, van Dam RM, Chasman DI. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. Mol Psychiatry. 2015 May;20(5):647-56. doi: 10.1038/mp.2014.107. Epub 2014 Oct 7. PubMed PMID: 25288136.

63. Wood AR, Tuke MA, Nalls M, Hernandez D, **Gibbs JR**, Lin H, Xu CS, Li Q, Shen J, Jun G, Almeida M, Tanaka T, Perry JR, Gaulton K, Rivas M, Pearson R, Curran JE, Johnson MP, Göring HH, Duggirala R, Blangero J, Mccarthy MI, Bandinelli S, Murray A, Weedon MN, Singleton A, Melzer D, Ferrucci L, Frayling TM. Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes. Hum Mol Genet. 2015 Mar 1;24(5):1504-12. doi: 10.1093/hmg/ddu560. Epub 2014 Nov 6. PubMed PMID: 25378555.

64. Renton AE, Pliner HA, Provenzano C, Evoli A, Ricciardi R, Nalls MA, Marangi G, Abramzon Y, Arepalli S, Chong S, Hernandez DG, Johnson JO, Bartoccioni E, Scuderi F, Maestri M, **Gibbs JR**, Errichiello E, Chiò A, Restagno G, Sabatelli M, Macek M, Scholz SW, Corse A, Chaudhry V, Benatar M, Barohn RJ, McVey A, Pasnoor M, Dimachkie MM, Rowin J, Kissel J, Freimer M, Kaminski HJ, Sanders DB, Lipscomb B, Massey JM, Chopra M, Howard JF Jr, Koopman WJ, Nicolle MW, Pascuzzi RM, Pestronk A, Wulf C, Florence J, Blackmore D, Soloway A, Siddiqi Z, Muppidi S, Wolfe G, Richman D, Mezei MM, Jiwa T, Oger J, Drachman DB, Traynor BJ. A genome-wide association study of myasthenia gravis. JAMA Neurol. 2015 Apr;72(4):396-404. doi: 10.1001/jamaneurol.2014.4103. PubMed PMID: 25643325.

65. Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivières S, Jahanshad N, Toro R, Wittfeld K, Abramovic L, Andersson M, Aribisala BS, Armstrong NJ, Bernard M, Bohlken MM, Boks MP, Bralten J, Brown AA, Chakravarty MM, Chen Q, Ching CR, Cuellar-Partida G, den Braber A, Giddaluru S, Goldman AL, Grimm O, Guadalupe T, Hass J, Woldehawariat G, Holmes AJ, Hoogman M, Janowitz D, Jia T, Kim S, Klein M, Kraemer B, Lee PH, Olde Loohuis LM, Luciano M, Macare C, Mather KA, Mattheisen M, Milaneschi Y, Nho K, Papmeyer M, Ramasamy A, Risacher SL, Roiz-Santiañez R, Rose EJ, Salami A, Sämann PG, Schmaal L, Schork AJ, Shin J, Strike LT, Teumer A, van Donkelaar MM, van Eijk KR, Walters RK, Westlye LT, Whelan CD, Winkler AM, Zwiers MP, Alhusaini S, Athanasiu L, Ehrlich S, Hakobjan MM, Hartberg CB, Haukvik UK, Heister AJ, Hoehn D, Kasperaviciute D, Liewald DC, Lopez LM, Makkinje RR, Matarin M, Naber MA, McKay DR, Needham M, Nugent AC, Pütz B, Royle NA, Shen L, Sprooten E, Trabzuni D, van der Marel SS, van Hulzen KJ, Walton E, Wolf C, Almasy L, Ames D, Arepalli S, Assareh AA, Bastin ME, Brodaty H, Bulayeva KB, Carless MA, Cichon S, Corvin A, Curran JE, Czisch M, de Zubicaray GI, Dillman A, Duggirala R, Dyer TD, Erk S, Fedko IO, Ferrucci L, Foroud TM, Fox PT, Fukunaga M, **Gibbs JR**, Göring HH, Green RC, Guelfi S, Hansell NK, Hartman CA, Hegenscheid K, Heinz A, Hernandez DG, Heslenfeld DJ, Hoekstra PJ, Holsboer F, Homuth G, Hottenga JJ, Ikeda M, Jack CR Jr, Jenkinson M, Johnson R, Kanai R, Keil M, Kent JW Jr, Kochunov P, Kwok JB, Lawrie SM, Liu X, Longo DL, McMahon KL, Meisenzahl E, Melle I, Mohnke S, Montgomery GW, Mostert JC, Mühleisen TW, Nalls MA, Nichols TE, Nilsson LG, Nöthen MM, Ohi K, Olvera RL, Perez-Iglesias R, Pike GB, Potkin SG, Reinvang I, Reppermund S, Rietschel M, Romanczuk-Seiferth N, Rosen GD, Rujescu D, Schnell K, Schofield PR, Smith C, Steen VM, Sussmann JE, Thalamuthu A, Toga AW, Traynor BJ, Troncoso J, Turner JA, Valdés Hernández MC, van 't Ent D, van der Brug M, van der Wee NJ, van Tol MJ, Veltman DJ, Wassink TH, Westman E, Zielke RH, Zonderman AB, Ashbrook DG, Hager R, Lu L, McMahon FJ, Morris DW, Williams RW, Brunner HG, Buckner RL, Buitelaar JK, Cahn W, Calhoun VD, Cavalleri GL, Crespo-Facorro B, Dale AM, Davies GE, Delanty N, Depondt C, Djurovic S, Drevets WC, Espeseth T, Gollub RL, Ho BC, Hoffmann W, Hosten N, Kahn RS, Le Hellard S, Meyer-Lindenberg A, Müller-Myhsok B, Nauck M, Nyberg L, Pandolfo M, Penninx BW, Roffman JL, Sisodiya SM, Smoller JW, van Bokhoven H, van Haren NE, Völzke H, Walter H, Weiner MW, Wen W, White T, Agartz I, Andreassen OA, Blangero J, Boomsma DI, Brouwer RM, Cannon DM, Cookson MR, de Geus EJ, Deary IJ, Donohoe G, Fernández G, Fisher SE, Francks C, Glahn DC, Grabe HJ, Gruber O, Hardy J, Hashimoto R, Hulshoff Pol HE, Jönsson EG, Kloszewska I, Lovestone S, Mattay VS, Mecocci P, McDonald C, McIntosh AM, Ophoff RA, Paus T, Pausova Z, Ryten M, Sachdev PS, Saykin AJ, Simmons A, Singleton A, Soininen H, Wardlaw JM, Weale ME, Weinberger DR, Adams HH, Launer LJ, Seiler S, Schmidt R, Chauhan G, Satizabal CL, Becker JT, Yanek L, van der Lee SJ, Ebling M, Fischl B, Longstreth WT Jr, Greve D, Schmidt H, Nyquist P, Vinke LN, van Duijn CM, Xue L, Mazoyer B, Bis JC, Gudnason V, Seshadri S, Ikram MA; Alzheimer’s Disease Neuroimaging Initiative; CHARGE

Consortium; EPIGEN; IMAGEN; SYS, Martin NG, Wright MJ, Schumann G, Franke B, Thompson PM, Medland SE. Common genetic variants influence human subcortical brain structures. Nature. 2015 Apr 9;520(7546):224-9. doi: 10.1038/nature14101. Epub 2015 Jan 21. PubMed PMID: 25607358.

66. Zukosky K, Meilleur K, Traynor BJ, Dastgir J, Medne L, Devoto M, Collins J, Rooney J, Zou Y, Yang ML, **Gibbs JR**, Meier M, Stetefeld J, Finkel RS, Schessl J, Elman L, Felice K, Ferguson TA, Ceyhan-Birsoy O, Beggs AH, Tennekoon G, Johnson JO, Bönnemann CG. Association of a Novel ACTA1 Mutation With a Dominant Progressive Scapuloperoneal Myopathy in an Extended Family. JAMA Neurol. 2015 Jun;72(6):689-98. doi: 10.1001/jamaneurol.2015.37. PubMed PMID: 25938801.

67. Jansen IE, Bras JM, Lesage S, Schulte C, **Gibbs JR**, Nalls MA, Brice A, Wood NW, Morris H, Hardy JA, Singleton AB, Gasser T, Heutink P, Sharma M; IPDGC. CHCHD2 and Parkinson's disease. Lancet Neurol. 2015 Jul;14(7):678-9. doi: 10.1016/S1474-4422(15)00094-0. PubMed PMID: 26067110.

68. Nalls MA, McLean CY, Rick J, Eberly S, Hutten SJ, Gwinn K, Sutherland M, Martinez M, Heutink P, Williams NM, Hardy J, Gasser T, Brice A, Price TR, Nicolas A, Keller MF, Molony C, **Gibbs JR**, Chen-Plotkin A, Suh E, Letson C, Fiandaca MS, Mapstone M, Federoff HJ, Noyce AJ, Morris H, Van Deerlin VM, Weintraub D, Zabetian C, Hernandez DG, Lesage S, Mullins M, Conley ED, Northover CA, Frasier M, Marek K, Day-Williams AG, Stone DJ, Ioannidis JP, Singleton AB; Parkinson's Disease Biomarkers Program and Parkinson's Progression Marker Initiative investigators*. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. Lancet Neurol. 2015 Aug 10. pii: S1474-4422(15)00178-7. doi: 10.1016/S1474-4422(15)00178-7. [Epub ahead of print] PubMed PMID: 26271532.