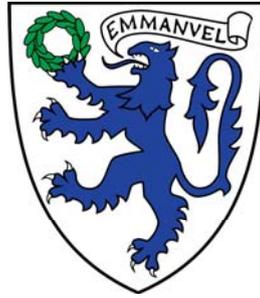




**The contribution of rare variants to risk
of schizophrenia and
neurodevelopmental disorders**



Tarjinder Singh

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

January 2017

Declaration

I hereby declare that I carried out the work described in this Thesis between September 2012 and August 2016 under the supervision of Dr. Jeffrey C. Barrett at the Wellcome Trust Sanger Institute. The contents of this Thesis has not been submitted in whole or in part for any other degree or qualification at the University of Cambridge, or any other University. This Thesis does not exceed the specified length limit, and is formatted according to the requirements set by the Biology Degree Committee and the Board of Graduate Studies.

Tarjinder Singh
January 2017

Acknowledgements

I would first like to thank my supervisor Jeff for giving me the opportunity to write my thesis in his group and for his support and encouragement in the past four years. I arrived in Cambridge in 2012 with plans to complete a one-year MPhil degree in the genetics of blood traits before attending medical school the following year. Little did I know that I would be fortunate enough to stay at the Sanger and explore an area of genetics and medicine that I had never previously encountered. Now, four years later, I am more motivated than ever to delve deeper into the world of statistical genetics and use its advances to understand the fundamental causes of mental illnesses. This formative experience would not be possible if not for the patience, mentorship, and guidance that Jeff has shown me. I sincerely hope we keep in touch, and perhaps find an opportunity to work together again in the future.

The research in this Thesis requires the coordinated efforts of numerous collaborators in the UK and around the world. I thank the many clinicians and scientists in the UK10K consortium who designed the initial study that laid the foundation for this Thesis. In particular, I would like to thank Mike Owen, Mick O'Donovan, Dave Curtis, and Matthew Hurles for productive discussions, advice, and contributions to this work. I want to express my gratitude to the Wellcome Trust and Williams College for their generous financial support. Most importantly, I want to thank the tens of thousands of patients and participants who enrolled in the studies described in this Thesis, without whom none of this work is possible and for which I am indescribably grateful.

On a more personal note, I want to extend my gratitude to the members of the Barrett team, past and present, for engaging team meetings, memorable retreats, and entertaining lunch discussions. I also want to thank my friends at Emmanuel College for all the enjoyable Formal dinners, pub crawls, and European travels that have made life in Cambridge so entertaining. In particular, I want to thank Albert and Uttara for their unwavering friendship, support, and honesty. I am grateful to the PhD students at the Sanger Institute, especially the Class of 2012, for all the enjoyable times we've had, from the scenic drives through the Peak District to the adventures in Warsaw. Lastly, I want to thank my parents and sister for their patience, encouragement, love and support throughout the years.

Abstract

In recent years, whole-exome sequencing has successfully identified genes in which rare variants confer substantial risk for neurodevelopmental disorders, such as autism spectrum disorders and intellectual disability. In many of these studies, the same gene is implicated in a wide variety of diagnoses and presentations. Despite a number of rare variant studies in schizophrenia, no gene has been significantly implicated using rare coding variants. In this Thesis, I compiled the largest rare variant data set in schizophrenia to date, and meta-analysed the whole-exome sequences of 1,077 trios, 4,268 cases, and 9,343 matched controls. With these data, I identified a genome-wide significant association between rare loss-of-function (LoF) variants in *SETDIA* and risk for schizophrenia. I additionally found that *SETDIA* is substantially depleted of LoF variants in the general population, and that LoF variants in this gene increased risk for a range of neurodevelopmental disorders. Combined, our results implicate epigenetic regulation, specifically histone modification, as a mechanism in the pathogenesis of schizophrenia, and suggest that rare risk alleles may potentially be shared between schizophrenia and other neurodevelopmental disorders.

To better understand if *SETDIA* finding can be generalized to a larger number of rare schizophrenia risk variants, I jointly analysed the trio and case-control exome data with array-based copy number variant calls from 6,882 cases and 11,255 controls. I found that individuals with schizophrenia carried a significantly higher burden of rare damaging variants in 3,488 “highly constrained” genes with a near-complete depletion of truncating variants. Rare variant enrichment analyses demonstrated that the rare schizophrenia risk variants were most strongly enriched in autism risk genes, and genes diagnostic of severe developmental disorders. I further showed that schizophrenia patients with intellectual disability had a greater enrichment of rare damaging variants in highly constrained genes, but that a weaker but significant enrichment existed throughout the larger schizophrenia population. Combined, these results demonstrate that schizophrenia risk loci of large effect across a range of variant types implicate a common set of genes shared with broader neurodevelopmental disorders, suggesting a path forward in identifying additional risk genes in psychiatric disorders and further supporting a neurodevelopmental etiology to the pathogenesis of schizophrenia.

Table of contents

List of figures	xiii
List of tables	xvii
1 Introduction	1
1.1 Schizophrenia	1
1.1.1 Diagnostic criteria and clinical heterogeneity in presentation	2
1.1.2 Disease management and prognosis	3
1.1.3 Epidemiology and global burden of disease	4
1.1.4 Environmental risk factors	6
1.2 The genetic architecture of schizophrenia	7
1.2.1 Family studies find substantial genetic component to risk	7
1.2.2 Genome-wide association studies implicate common polygenic variation	8
1.2.3 Recurrent copy number events confer substantial risk	10
1.2.4 Shared common risk variants across psychiatric disorders	12
1.3 Whole-exome sequencing as a means of studying rare variants	12
1.3.1 Common study designs for sequencing studies	14
1.4 Early results from sequencing in schizophrenia	16
1.5 Biological insights from genetic studies of schizophrenia	17
1.6 Goals of this Thesis	19
2 A protocol for the quality control of whole-exome sequencing data sets	21
2.1 Challenges behind the production and analysis of sequencing data	21
2.1.1 Publication note and contributions	23
2.2 Materials and methods	23
2.2.1 Sample collections	23
2.3 Sequence data production	26

2.3.1	Sample preparation	26
2.3.2	Alignment and BAM processing	26
2.3.3	Variant calling	27
2.4	Variant calling and quality control across capture and batch	27
2.4.1	Adjusting for differences between capture and batch	27
2.5	Sample-level quality control for case-control analysis	29
2.5.1	Sample-level QC in the UK10K-INTERVAL case-control data set	29
2.5.2	Sample-level QC in the Finnish and Swedish case-control data sets	31
2.6	Variant filtering in case-control data sets	32
2.6.1	Variant filtering in the UK10K-INTERVAL data set	32
2.6.2	Variant filtering in the Finnish and Swedish data sets	36
2.7	Comparison of population genetics metrics across data sets	36
2.8	Systematic annotation of coding variants	38
2.9	Evaluating the effectiveness of existing <i>in silico</i> predictors of pathogenicity	39
2.9.1	The interpretation of protein-coding consequences	39
2.9.2	A description of existing annotation tools	40
2.9.3	Strategy for evaluating variant annotation tools	41
2.9.4	Preparation of annotation files	43
2.9.5	Classifiers display variable performance depending on test data	43
2.9.6	A comparison of annotation approach with other whole-exome sequencing studies	45
2.10	A meta-analysis of published schizophrenia parent-proband trio studies	48
2.11	Gene-specific mutation rates based on GENCODE transcripts	49
2.12	Discussion	51
2.13	Consortia	53
2.13.1	UK10K consortium	53
2.13.2	DDD Study	54
2.13.3	Swedish Schizophrenia Study	54
2.13.4	INTERVAL study	54
2.13.5	Sequencing Initiative Suomi project	54
3	SETDIA is associated with schizophrenia and neurodevelopmental disorders	57
3.1	Introduction	57
3.1.1	Motivation behind rare variant analyses in psychiatric disorders	57
3.1.2	Early studies of rare variants in psychiatric disorders	58
3.1.3	Emerging results from sequencing studies of neurodevelopmental disorders	59

3.1.4	Goal and aims	60
3.1.5	Publication note and contributions	61
3.2	Materials and methods	61
3.2.1	Gene-based analysis in the case-control data set	61
3.2.2	Meta-analysis of <i>de novo</i> mutations and case-control burden	62
3.2.3	Frequentist method of meta-analysis using Fisher's method	62
3.2.4	Bayesian modeling of <i>de novo</i> and case-control variants using TADA	63
3.2.5	Validation of variants of interest	64
3.2.6	Functional consequence of the exon 16 splice acceptor deletion	64
3.2.7	Phenotype clustering in DDD probands	65
3.3	Results	65
3.3.1	Study design	65
3.3.2	LoF variants in <i>SETDIA</i> are associated with schizophrenia	66
3.3.3	Robustness of the <i>SETDIA</i> association	69
3.3.4	<i>SETDIA</i> is associated with severe developmental disorders	78
3.3.5	Power calculations to show co-morbid cognitive impairment in schizophrenia <i>SETDIA</i> carriers	82
3.3.6	<i>De novo</i> burden in neurodevelopmental disorders	84
3.4	Discussion	86
4	Schizophrenia risk genes are shared with neurodevelopmental disorders	91
4.1	Introduction	91
4.1.1	Early evidence for a neurodevelopmental etiology to schizophrenia	91
4.1.2	Sharing of rare variants between autism spectrum disorders and intellectual disability	92
4.1.3	Individual loci increasing risk for schizophrenia and neurodevelopmental disorders	93
4.1.4	Genes with near-complete depletion of protein-truncating variants	94
4.1.5	Aims and goals	95
4.1.6	Publication note and contributions	96
4.2	Methods	96
4.2.1	Sample collections	96
4.2.2	Rare variant gene set enrichment analyses	97
4.2.3	Combined joint analysis	99
4.2.4	Description of gene sets	100
4.2.5	Conditional analyses	102
4.2.6	Rare variants and cognition in schizophrenia	103

4.3	Results	104
4.3.1	Study design	104
4.3.2	Selection of allele frequency thresholds and consequence severity	106
4.3.3	Robustness of enrichment analyses	109
4.3.4	Rare, damaging schizophrenia variants are concentrated in constrained genes	110
4.3.5	Comparing the enrichment in constrained genes across neurodevelopmental disorders	112
4.3.6	Schizophrenia risk genes are shared with other neurodevelopmental disorders	115
4.3.7	Schizophrenia rare variants are associated with intellectual disability	117
4.4	Discussion	125
5	Discussion and future directions	127
5.1	Summary of findings	127
5.2	Limitations of results described in this Thesis	128
5.2.1	Limitations in the interpretation of protein-coding consequences	128
5.2.2	Insufficient standardisation of clinical data	130
5.2.3	Limitations in the definition of the constrained gene list	131
5.2.4	Interpretation and generalisability of gene set results	132
5.3	Future directions	133
5.3.1	Whole-genome sequencing at the population scale	133
5.3.2	Specificity of shared risk alleles for individual psychiatric disorders	136
5.3.3	<i>In vitro</i> and <i>in vivo</i> modeling of risk genes for neurodevelopmental disorders	137
5.4	Concluding remarks	139
	References	141

List of figures

1.1	Risk variants for schizophrenia.	11
2.1	Density plots of sequence coverage in the UK10K, INTERVAL, and DDD datasets.	28
2.2	Principal components analysis of UK and Finnish samples in the UK10K schizophrenia dataset.	30
2.3	The evaluation of different variant filtering thresholds using rare DDD inherited variants and Mendelian inconsistent variants as a testing set.	33
2.4	Variant metrics in the UK10K and INTERVAL datasets after each variant filtering step.	35
2.5	Variant counts summarised according to variant class and sequencing batch in the UK10K, INTERVAL, Finnish, and Swedish datasets.	37
2.6	Distributions of TiTv and frameshift-inframe ratios in the UK10K, INTERVAL, Finnish, and Swedish datasets.	38
2.7	ROC curve evaluating the performance of missense classifiers on UniProt pathogenic and benign variants.	44
2.8	ROC curve evaluating the performance of missense classifiers on pathogenic <i>de novo</i> mutations and benign variants from UniProt.	46
2.9	ROC curve evaluating the performance of missense classifiers on pathogenic <i>de novo</i> mutations and ExAC missense variants with MAF > 1%.	47
2.10	Correlation between mutation rates generated using GENCODE and RefSeq transcript databases.	52
2.11	The ratio of the damaging missense mutation rate to the missense mutation rate of each GENCODE coding gene.	52
3.1	Study design for the schizophrenia exome meta-analysis.	66
3.2	Manhattan plot of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.	67

3.3	QQ plots of the rare variant association analysis of LoF variants in 4,264 cases and 9,343 controls.	68
3.4	Manhattan plot of the meta-analysis of <i>de novo</i> mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.	70
3.5	QQ plot of the meta-analysis of <i>de novo</i> mutations and case-control variants in 1,077 trios, 4,264 cases and 9,343 controls.	71
3.6	The genomic position and coding consequences of 16 <i>SETD1A</i> LoF variants observed in the schizophrenia exome meta-analysis, the DDD study, and the SiSU project.	71
3.7	Results from the minigene experiment assessing the impact of the exon 16 splice acceptor site variant.	75
3.8	The robustness of the <i>SETD1A</i> result across reasonable parameters in the TADA model.	76
3.9	<i>De novo</i> microdeletion of a single copy of <i>SETD1A</i> identified in the DDD study.	81
3.10	Sample size curves for detecting an increased risk of pre-morbid cognitive impairment in schizophrenia <i>SETD1A</i> LoF carriers.	83
3.11	A comparison of genome-wide <i>de novo</i> mutation rates in probands with autism, developmental disorders, schizophrenia, and controls.	85
3.12	Mendelian disorders of epigenetic machinery at histone H3.	87
3.13	SET1/COMPASS complex	88
4.1	The overlap between autism risk genes and dominant developmental disorder genes.	94
4.2	Analysis workflow.	105
4.3	Q-Q plots of <i>P</i> -values from enrichment tests of 1,766 gene sets.	107
4.4	The use of frequency and size cut-offs in CNV gene sets enrichment tests to reduce genomic inflation.	108
4.5	Q-Q plots of <i>P</i> -values from enrichment tests of random gene sets.	110
4.6	Non-random sampling of genes in the 1,766 gene sets resulted in non-null enrichment of disruptive variants.	111
4.7	Enrichment of schizophrenia rare variants in constrained genes.	112
4.8	Enrichment of <i>de novo</i> mutations in genes with near-complete depletion of truncating variants across schizophrenia and neurodevelopmental disorders.	113
4.9	Enrichment of <i>de novo</i> mutations in genes ordered and grouped by genic constraint across schizophrenia and neurodevelopmental disorders.	114

4.10	Enrichment of case-control SNVs in genes ordered and grouped by genic constraint.	115
4.11	Enrichment of rare variants in constrained genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.	122
4.12	Enrichment of rare variants in diagnostic developmental disorder genes between schizophrenia (SCZ) individuals with ID, schizophrenia individuals without ID, and matched controls.	124
5.1	Risk variants for schizophrenia, with <i>SETD1A</i> included.	129
5.2	Distribution of overlap coefficients with the constrained gene set.	134
5.3	Heatmap of overlap coefficients calculated between $FDR < 5\%$ gene sets. .	135

List of tables

2.1	Description of samples collections included as cases in the UK10K schizophrenia analysis.	24
2.2	Description of samples collections included as controls in the UK10K schizophrenia analysis.	25
2.3	Description and summary of statistical tools developed to predict the pathogenicity of coding variants.	42
2.4	Published studies identifying <i>de novo</i> mutations in schizophrenia parent-proband trios using whole-exome sequencing.	50
3.1	Meta-analysis results for 1,077 trios, 4,264 cases and 9,343 controls. Only <i>SETDIA</i> reached exome-wide significance.	70
3.2	Results from statistical tests associating disruptive variants in <i>SETDIA</i> to schizophrenia and developmental delay.	72
3.3	TADA results using the hyperparameters in the De Rubeis <i>et al.</i> autism meta-analysis. Only <i>SETDIA</i> has a q -value < 0.01	77
3.4	Burden tests associating disruptive variants in <i>SETDIA</i> to schizophrenia and developmental delay.	78
3.5	Phenotypes of individuals in the schizophrenia exome meta-analysis who carry LoF variants in <i>SETDIA</i>	79
3.6	Phenotypes of individuals in the DDD study and SiSU project who carry LoF variants in <i>SETDIA</i>	80
4.1	Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR $< 1\%$	116
4.2	Gene sets enriched for rare coding variants conferring risk for schizophrenia at FDR $< 5\%$	119
4.3	Results from enrichment analyses of FDR $< 5\%$ gene sets, conditional on brain-expressed and ExAC constrained genes.	121

4.4 Phenotypes of schizophrenia individuals with cognitive information carrying
LoF variants in developmental disorder genes. 123