Title
Informing Genetic Models of Autism via Transcriptional Network Analysis in Brain and Blood

Author
Luo, Rui

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Informing Genetic Models of Autism via Transcriptional Network Analysis in Brain and Blood

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Rui Luo

2014

ABSTRACT OF THE DISSERTATION

Informing Genetic Models of Autism via Transcriptional Network Analysis in Brain and Blood

by

Rui Luo

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2014

Professor Daniel Geschwind, Chair

Autism Spectrum Disorders (ASDs) are a group of heritable neruodevlopmental disorders. Both common and rare genetic variants are known to play a role in ASDs. However the functional impact of genetic variants remains largely unexplored. In this study, we conducted transcriptome profiling analysis to uncover the expression alterations that are associated with autism. The transcriptome profiling also aids us exploring the regulatory patterns of genetic variants, and better understanding the genetic models of autism. Since brain tissue is not accessible on a large scale, we profiled mRNAs of lymphoblast cell lines (LCLs) from three independent cohorts to determine whether we could detect a reproducible blood gene expression pattern associated with ASD. RNA from a total of 978 patients, and 651 controls, including 607 unaffected siblings analyzed for differential expression. Although few genes were consistently

differentially expressed between ASD and controls, we did find five (*CMKOR1*, *DKFZP564O0823, PITPNC1, PRKCB1* and *VIM*) that were differentially expressed in at least two cohorts LCLs and previously published brain samples. Similarly, using LCL gene expression to classify subjects by disease status performed only slightly above chance. Using weighted gene co-expression network analysis (WGCNA), we were able to identify a module correlated with ASD in both AGRE and NIMH cohorts that overlapped with genes previously found to be mis-expressed in post mortem brain from ASD cases. eQTL analysis identified SNPs that were associated with LCL gene expression, including several in AHI1, a Joubert Syndrome gene dysregulated in ASD brain and lymphoblasts. Four of the 23 SNPs that were significantly correlated with the expression level of AHI1 reside in the same haplotype block previously associated with ASD, suggesting that risk for ASD is mediated via AHI1 transcript levels. Overall, we found a weak, but consistent signal in LCLs further suggesting that peripheral lymphoblast gene expression may be useful for studying ASD.

Rare variants including Copy Number Variants (CNVs) and Single Nucleotide Variants (SNVs) are found to play an important role to the etiology of ASD together with common variants. We next interrogated gene expression in lymphoblasts from 244 families with discordant siblings in the Simons Simplex Collection in order to identify potentially pathogenic variation. Our results reveal that the overall frequency of significantly mis-expressed genes (which we refer to here as outliers) identified in probands and unaffected siblings do not differ. However, in probands, but not their unaffected siblings, the group of outlier genes is significantly enriched in neural-related pathways including neuropeptide signaling, synaptogenesis and cell adhesion. We demonstrate that outlier genes cluster within the most pathogenic CNVs (rare de novo CNVs) and can be used to prioritize rare CNVs of potentially

unknown significance. Several non-recurrent CNVs with significant gene expression alterations are identified, including deletions on chromosome 3q27, 3p13 and 3p26, and duplications at 2p15, suggesting these as potential novel ASDs loci. In addition, we identify distinct pathways disrupted in 16p11.2 microdeletions, microduplications and 7q11.23 duplications, and show that specific genes within the 16p CNV interval correlate with differences in head circumference, an ASDs relevant phenotype. This study provides evidence that pathogenic structural variants have functional impact on transcriptome alterations in ASDs at a genome-wide level, and demonstrates the utility of this approach for prioritization of genes for subsequent functional analysis.

Genetic studies have identified dozens of ASDs susceptibility genes, yet the interaction between ASD risk genes are pooly understood. In the aim of identify the molecular mechanisms and potential convergening pathways of ASD risk genes, the last chapter of my research utilizes transcriptome profiling to answer two questions: 1) do these genetic loci converge on specific laminar expression patterns, and 2) where does the phenotypic specificity of ASDs arise, given its genetic overlap with intellectual disability (ID)? To answer these, we mapped ASDs and ID risk genes to non-human primate and human brain transcriptome. We found ASDs genes are enriched in superficial cortical layers and glutamatergic projection neurons at the circuit level. Furthermore, we show that the patterns of ASDs and ID risk genes are distinct, providing a novel biological framework for investigating the pathophysiology of ASDs. In this chapter, we demonstrated the importance of understanding ASD gene interaction with systems biology method.

The dissertation of Rui Luo is approved.

Steve Horvath

Matteo Pellegrini

Stanley Nelson

Daniel Geschwind, Committee Chair

University of California, Los Angeles

2014

To my parents, Kangjin Luo and Jianqing Lu; my husband, Chaochao Cai

for their unconditional love and support

To my little son, Yuwei Cai

your accomplishment will be greater

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to start by thanking my mentor, Dr. Daniel Geschwind, for his continuous support over the past five years. I am very grateful for his mentorship; he is very kind and helpful. He gives me the freedom to do independent research, taught me how to think critically and independently, as well as how to make oral presentations. I owe my growth as a scientist and development as a person to working with him. In life, he has been very supportive in encouraging me to think and pursue my dreams.

I am deeply grateful for my committee members, Drs. Steve Horvath, Stanley Nelson and Matteo Pellegrini for their time and efforts. I really appreciate their warm welcome to my intrusions. Their suggestions, support and enthusiasm have moved me forward along the path. In particular, I would like to thank Dr. Steve Horvath, who provided me the chance to do a rotation in his lab and taught me statistical skills. Without his suggestions on my research work, I cannot be so productive.

My PhD life could not been so wonderful without members of the Geschwind lab. In particular, I would like to give a special thank you to Dr. Mike Oldham, Dr. Irina Voineagu, Dr. Jeremy Miller and Dr. T. Grant Belgard, for their guidance and support. Dr. Mike Oldham and Dr. Jeremy Miller's great accomplishments in the Geschwind Lab have inspired me to join the lab and continue my research work through my PhD study. Dr. Irina Voineagu has been very supportive and taught me how to conduct bench work step by step. I also want to thanks Dr. Jennifer K Lowe, who helped manage the autism data sets and organized autism-related projects. I would also like to thank my fellow graduate students: Yuan Tian, Neel Parikshak, Donar Werling and Jamee Bomar for all the help and inspiring discussion. I want to thank the research

staffs in our lab, especially Jing Ou and Kun Gao, who provides great aids to process the samples for me. Lastly, I would like to thank our lab managers Lauren Kawaguchi and Jenifer Sakai, for all your help on the lab routine, which makes my life in Geschwind lab much easier.

I am also thankful for the ACCESS and Human Genetics program for admitting me into this great program and the excellent curriculum.

I am extremely grateful for my unbelievably supportive family and friends. I want to thank my parents and my husband, for their unconditional love and support, for believing in me and encouraging me to reach up for the sky. I want to thank my son, who provides me endless happiness and teaches me how to balance work and life.

Chapter 1 is a draft manuscript of my work on "Analyzing expresion profiling of Lymphoblast cell lines to detect autism associated transcriptome alterations", authored by Rui Luo, Jennifer K. Lowe, Steve Horvath, Audrey Thurm and Daniel Geschwind. The authors would like to thank Donna Werling, Yuan Tian for their suggestions. We also want to thank our lab managers Lauren Kawaguchi and Jenifer Sakai for their assistance.

Chapter 2 is adapted from a published paper "Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders", authored by Rui Luo, Stephan J. Sanders, Yuan Tian, Irina Voineagu, Ni Huang, Su H. Chu, Lambertus Klei, Chaochao Cai, Jing Ou, Jennifer K. Lowe, Matthew E. Hurles, Bernie Devlin, Matthew W. State, and Daniel H. Geschwind. The authors would like to acknowledge the support from Autism Speaks to provide the funding support.

Chapter 3 is adapted from three papers. The first paper is "Transcriptional architecture of

the primate neocortex" authored by Amy Bernard, Laura S. Lubbers, Keith Q. Tanis, Rui Luo, Alexei A. Podtelezhnikov, Eva M. Finney, Mollie M.E. McWhorter, Kyle Serikawa, Tracy Lemon, Rebecca Morgan, Catherine Copeland, Kimberly Smith, Vivian Cullen, Jeremy Davis-Turak, Chang-Kyu Lee, Susan M. Sunkin, Andrey P. Loboda, David M. Levine, David J. Stone, Michael J. Hawrylycz, Christopher J. Roberts, Allan R. Jones, Daniel H. Geschwind and Ed S. Lein. In this paper, Amy Bernard is the first author. She designed and conducted the experiment. Ed S. Lein is the corresponding author who wrote the manuscript. I conducted the majority part of statistical analyses including ANVOA, WGCNA in this paper. I also made Figure 3 and Figure 5 in this paper. The second manuscript is in preparation: "Spatiotemporal dynamics of the postnatal developing primate transcriptome" authored by Amy Bernard, Rui Luo, Jeremy Miller, Trygve Bakken, Daniel H. Geschwind and Ed S. Lein. In this paper, Amy Bernard and I are the co-first authors. She designed and conducted the experiment. Ed S. Lein is the corresponding author. I led all the statistical analyses of this project, wrote the first draft of the Results section of the manuscript and made all the figures. Jeremy Miller and Trygve Bakken wrote the introduction and discussion of this manuscript. Ed S. Lein and Daniel Geschwind are corresponding authors, who provided importants advises. The last paper is "Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism" authored by Neelroop N. Parikshak, Rui Luo, Alice Zhang, Hyejung Won, Jennifer K. Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H. Geschwind. In this paper, Neelroop N. Parikshak is the first author, who led the project and wrote the paper. I was in charge of the laminar enrichment section, and made Figure 5 and Figure 6. Steve Horvath provided important advises for the statistical analysis for this project. Daniel H. Geschwind is the corresponding author who supervised this project and helped on the manuscript writing. I would like to thank

the Allen Brain Institution for providing the data sets for analysis. To ensure the use of these papers in my thesis to meet the copyright requirements of UCLA, endorsement letters from the first authors and senior authors of these papers are attached at the end of the main part of this thesis.

This dissertation was supported by R08741 M09R10124 grant (to D.H.G., M.S., B.D.) from the Simons Foundation and a Pilot Grant (D.H.G.) 20104829 from the Simons Foundation, the NIMH (ACE Network grant-5R01 MH081754-04 to D.H.G. (PI) and MS), and Dennis Weatherstone pre-doctoral fellowship from Autism Speaks (to R.L).

VITA

2008	Bachelor of Science, Molecular Cell Biology University of Science and Technology of China, Hefei, China
2009-2010	Teaching Assistant Department of Molecular Cellular and Developmental Biology, University of California, Los Angeles
2008-2013	Graduate Student Researcher Geschwind Lab Human Genetics, Department of Medicine University of California, Los Angeles

PUBLICATIONS

Parikshak, N. N., **Luo, R**., Zhang, A., Won, H., Lowe, J. K., Chandran, V., et al. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*, 155(5), 1008–1021. doi:10.1016/j.cell.2013.10.031 PMCID: PMC3107252

Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. *American Journal of Human Genetics*. 2012. Epub 2012/06/26. doi: 10.1016/j.ajhg.2012.05.011. PubMed PMID: 22726847.

Langfelder P, **Luo R**, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS Comput Biol*. 2011;7(1):e1001057. PMCID: 3024255.

Sanders SJ, Ercan-Sencicek AG, Hus V*, **Luo R***, Murtha MT, Moreno-De-Luca D, et al. Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron*. 2011;70(5):863-85.

Bernard A*, Lubbers LS*, Tanis KQ*, **Luo R**, Podtelezhnikov AA, Finney EM, et al. Transcriptional architecture of the primate neocortex. *Neuron*. 2011;73(6):1083-99.

Cai C, Langfelder P, Fuller TF, Oldham MC, **Luo R**, van den Berg LH, et al. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*. 2010;11:589. PMCID: 3091510.

Celestino-Soper PB, Violante S, Crawford EL, **Luo R**, Lionel AC, Delaby E, et al. A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism.

Proceedings of the National Academy of Sciences of the United States of America. 2012;109(21):7974-81. Epub 2012/05/09. doi: 10.1073/pnas.1120210109. PubMed PMID: 22566635; PubMed Central PMCID: PMC3361440.

Chen H, **Luo R**, Gong S, Matta SG, Sharp BM. Protection Genes in Nucleus Accumbens Shell Affect Vulnerability to Nicotine Self-Administration across Isogenic Strains of Adolescent Rat. *PloS one*. 2014;9(1):e86214. doi: 10.1371/journal.pone.0086214. PubMed PMID: 24465966; PubMed Central PMCID: PMC3899218.

Introduction

ASD Background and Features

Autism Spectrum Disorders (ASDs) are a group of neurodevelopmental disorders that include autism, pervasive developmental disorder not otherwise specified (PDD-NOS), and Asperger's syndrome [1]. Based on the Diagnostic and Statistical Manual of Mental Disorder (DSM-V), a child is diagnosed with ASDs if he or she meets the following criteria: A) Persistent deficits in social communication and social interaction across multiple contexts, B) Restricted, repetitive patterns of behavior, interests, or activities. Severity is based on social communication impairments and restricted, repetitive patterns of behavior [2]. Additional features often comorbid with ASDs include sensory and motor abnormalities, ADHD, epilepsy, and developmental regression [1, 4]. Those with ASDs can range from being mentally disabled to having above average intelligence [5]. Currently, there is an increase in prevalence of ASDs, with estimated 1 out of 88 children as ASDs (CDC, ADDM network, 2012). Multiple sociocultural factors, including age at diagnosis, changing diagnostic criteria and broader inclusion rates but not biological factors, would contribute to this increased prevalence [6, 7].

Genetic studies of ASD

Both family and twin studies indicate that ASDs are a highly heritable neruosychatric disorder. The monozygotic twins have a much higher concordance rate (50%-90%) compared to dizygotic twins (0%-30%) [8, 9]. Interestingly, the risk is 3-fold in second born male siblings versus females, supporting models of reduced penetrance in females [8, 10]. Current studies have reported a variety of genetic causes that account for roughly 20% of ASD cases. Recent exome sequencing studies indicate that the number of ASD-implicated genes is between 200 and 1000

[11-16]. Studies suggest a mixed genetic model of ASD, indicating that common variants as SNPs as well as rare variants as CNVs and SNVs, are playing a role together to cause the heterogeneity of ASD [17-19].

The contribution of common genetic variation to ASDs has been evaluated by genome-wide association (GWA) studies, which compare the frequency of single nucleotide polymorphisms (SNPs) in cases and controls. Three major GWA studies of ASDs have recently been completed: Wang et al, conducted a GWA study with about 2000 multiplex families from Autism Genetic Resource Exchange (AGRE), and found SNP: rs4307059 associated with ASDs at genome-wide significance, located in an intergenic region between cadherin 9 (*CDH9*) and cadherin 10 (*CDH10*)[20]. Weiss et al. utilized 1031 multiplex autism families for GWA and showed genetic association reaching the genome-wide significance threshold for SNP: rs10513025 located 80kb upstream of semaforin 5A, between *SEMA5A* and the bitter taste receptor *TASR1*[21]. Anney et al. identified SNP: rs4141463 located in an intron of *MACROD2* using 1558 ASDs families [22]. None of the above studies replicated each other's findings, indicating the small effect sizes of common alleles in ASD. This may due to the relatively small sample size in each study comparing to other GWA stuidies. This also suggests that each of the common variants has a relatively minor effect size to disease, and many common variants, are necessary to lead the disease phenotype in each case.

The effects of rare variants in ASD are evaluated by measuring the frequency of rare copy number variation (CNV) and single nucleotide variants (SNV) in cases and controls. Several studies have identified CNVs that are related to ASD [22-24]. Two studies [22, 23] found that *de novo* CNVs occur more frequently in ASD cases than controls. Although none of

the individual CNVs were proven to be causal, these studies highlighted the fact that *de novo* mutations could contribute a significant proportion of the genetic abnormalities in ASDs. The most frequent chromosomal aberrations observed in ASDs are the maternal duplication of chromosome 15q11-13 and a 600kb microdeletion/ microduplication at 16p11.2, each occurring in approximately 1% of sporadic ASD cases. In addition, several rare CNVs have been found in cased but not in controls. Although the association of the rare CNVs with ASDs is difficult to reach genome-wide significance due to their low frequency, the genes spanned by these CNVs are relevant candidates for further evaluation by functional studies and targeted gene resequencing. Here are the list of genes shown in recurrent CNVs and have been indicated to be ASD risk genes by functional studies: *A2BP1*, *ANKRD11*, *C16orf72*, *CDH13*, *CDH18*, *DDX53*, *DLGAP2* [25, 26], *DPP6*, *DPYD*, *FHIT*, *FLJ16237*, *NLGN4*, *NRXN1*, *SHANK2*, *SHANK3*, *SLC4A10*, *SYNGAP1*, *USP7* [12, 15].

With the advances of next-generation sequencing techniques, four independent exome-sequncing studies have been conducted [11, 13, 14, 16]. In one study [16], there is a significant increase in the number of non-synonymous and nonsense *de novo* SNVs in cases compared to unaffected sibs when looking across all genes [OR of 1.93 (all non-synonymous to silent SNVs); OR of 4.03 (nonsense/splice site to silent SNVs)] and brain-expressed genes only [OR of 2.22 (all non-synonymous to silent SNVs); OR of 5.65 (nonsense/splice site mutations to silent SNVs)]. The other study reports a two-fold increase in frame shift, splice site, and nonsense *de novo* mutations in cases versus controls [11]. By combing SNVs of frame shift, splice site or nonsense *de novo* variants in cases across all four studies, five high-priority genes were identified that are recurrent in ASDs: *DYRK1A, POGZ, SCN2A, KATNAL2* and *CHD8*. Based on the genetic findings, these genes are worthy of downstream functional analysis.

Transcriptome studies of ASDs

Our studies of autism genetics indicate the fact that genetic variants may cause disease phenotype via affecting gene expression. Thus it becomes necessary to integrate the genetic findings with transcriptomic data. Ttranscriptom studies of ASD can also potentially identify the molecular mechnisam and uncover convergening signaling pathways of ASDs.

With the hope of identifying gene expressions that are altered in ASD cases, several studies have analyzed genome-wide expression profiles of ASD cases using readily available peripheral tissues such as lymphoblast cell lines (LCLs) [27-29] and blood [30-33]. Several pathways have been identified that are associated with ASD in each study, including: steroid biosynthesis, oxidative stress and ubiquitination. However, results shown little convergence in terms of the dysregulated pathways. The reasons for the non-overlapping could be following: (1) they used a relatively small sample size (smaller than 100), 2) each study utilized different study design and criteria to define dysregulated genes, (2) EBV transformation increased the variability of gene expression in LCLs.

As a brain disorder, the ideal tissue to study expression patterns associated with ASDs is the postmortem brain. A recent study conducted by Voineagu et al. [34] examined trascriptome profiles from three different brain regions (frontal cortex, temporal cortex and cerebellum) of 19 autism cases and 17 controls. By differential expression analysis of cortex samples, 444 genes

were identified as differentially expressed. Pathway analysis detected an up-regulation of genes involved in immune response, while a down-regulation of genes functioning at the synapse.

By using a network-based approach, Voineagu et al. found two modules that are related with ASDs status. Each module contains a list of genes that are co-expressed cross samples. The module (M12) with the strongest ASDs correlation consisted of genes down-regulated in ASDs and is enriched in genes with synaptic function and vesicular transport. Interestingly, this module showed enrichment for ASDs genetic association signal as measured in the GWA study by Wang et al[20], as genes in M12 has a significantly lower p-value distribution compared to genome background. This initial exploration of autism brain transcriptome illustrates the possibility to find the converging pathways for ASDs in a disease-relevant tissue.

References

1. Bill BR, Geschwind DH: Genetic advances in autism: heterogeneity and convergence on shared pathways. *Current Opinion in Genetics & Development* 2009, 19(3):271-278.

2. Geschwind DH: Autism: many genes, common pathways? *Cell* 2008, 135(3):391-395.

3. Gottesman, II, Gould TD: The endophenotype concept in psychiatry: etymology and strategic intentions. *The American Journal of Psychiatry* 2003, 160(4):636-645.

4. Tuchman R, Rapin I: Epilepsy in autism. *Lancet Neurology* 2002, 1(6):352-358.

5. Chakrabarti S, Fombonne E: Pervasive developmental disorders in preschool children: confirmation of high prevalence. *The American Journal of Psychiatry* 2005, 162(6):1133-1141.

6. Hertz-Picciotto I, Delwiche L: The rise in autism and the role of age at diagnosis. *Epidemiology* 2009, 20(1):84-90.

7. King M, Bearman P: Diagnostic change and the increased prevalence of autism. *International Journal of Epidemiology* 2009, 38(5):1224-1234.

8. Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J, Fedele A, Collins J, Smith K *et al*: Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of General Psychiatry* 2011, 68(11):1095-1102.

9. Rosenberg RE, Law JK, Yenokyan G, McGready J, Kaufmann WE, Law PA: Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of Pediatrics & Adolescent Medicine* 2009, 163(10):907-914.

10. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K, Law P, Qiu S, Lord C, Sebat J *et al*: A unified genetic theory for sporadic and inherited autism. *Proceedings of the National Academy of Sciences of the United States of America* 2007, 104(31):12831-12836.

11. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A *et al*: De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012, 74(2):285-299.

12. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K *et al*: Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 2011, 70(5):886-897.

13. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V *et al*: Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012, 485(7397):242-245.

14. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD *et al*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012, 485(7397):246-250.

15. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA *et al*: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011, 70(5):863-885.

16. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012, 485(7397):237-241.

17. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M *et al*: Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genetics* 2011, 7(11):e1002334.

18. Leblond CS, Heinrich J, Delorme R, Proepper C, Betancur C, Huguet G, Konyukh M, Chaste P, Ey E, Rastam M *et al*: Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. *PLoS Genetics* 2012, 8(2):e1002521.

19. State MW, Levitt P: The conundrums of understanding genetic risks for autism spectrum disorders. *Nature Neuroscience* 2011, 14(12):1499-1506.

20. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM *et al*: Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009, 459(7246):528-533.

21. Weiss LA, Arking DE, Gene Discovery Project of Johns H, the Autism C, Daly MJ, Chakravarti A: A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 2009, 461(7265):802-808.

22. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT *et al*: A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics* 2010, 19(20):4072-4082.

23. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, Hill RS, Mukaddes NM, Balkhy S, Gascon G, Hashmi A *et al*: Identifying autism loci and genes by tracing recent shared ancestry. *Science* 2008, 321(5886):218-223.

24. Cook EH, Jr., Scherer SW: Copy-number variations associated with neuropsychiatric conditions. *Nature* 2008, 455(7215):919-923.

25. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y *et al*: Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics* 2008, 82(2):477-488.

26. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS *et al*: Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010, 466(7304):368-372.

27. Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, Sigman M, Steindler C, Pellegrini S, Schanen NC, Warren ST *et al*: Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics* 2007, 16(14):1682-1698.

28. Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME, Luu T, Lai Y, Lee NH: Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Research : Official Journal of the International Society for Autism Research* 2009, 2(2):78-97.

29. Hu VW, Nguyen A, Kim KS, Steinberg ME, Sarachana T, Scully MA, Soldin SJ, Luu T, Lee NH: Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis. *PLoS One* 2009, 4(6):e5775.

30. Gregg JP, Lit L, Baron CA, Hertz-Picciotto I, Walker W, Davis RA, Croen LA, Ozonoff S, Hansen R, Pessah IN *et al*: Gene expression changes in children with autism. *Genomics* 2008, 91(1):22-29.

31. Kong SW, Shimizu-Motohashi Y, Campbell MG, Lee IH, Collins CD, Brewster SJ, Holm IA, Rappaport L, Kohane IS, Kunkel LM: Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. *Neurogenetics* 2013, 14(2):143-152.

32. Kong SW, Collins CD, Shimizu-Motohashi Y, Holm IA, Campbell MG, Lee IH, Brewster SJ, Hanson E, Harris HK, Lowe KR *et al*: Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* 2012, 7(12):e49475.

33. Glatt SJ, Tsuang MT, Winn M, Chandler SD, Collins M, Lopez L, Weinfeld M, Carter C, Schork N, Pierce K *et al*: Blood-based gene expression signatures of infants and toddlers with autism. *Journal of the American Academy of Child and Adolescent Psychiatry* 2012, 51(9):934-944 e932.

34. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011, 474(7351):380-384.

CHAPTER 1: Large-scale lymphoblast gene expression profiling of ASD and unaffected controls reveals weak shared signals across cohorts

1.1 Abstract

Autism Spectrum Disorders (ASDs) are a group of heritable neuro-developmental disorders, which result from multiple genetic and environment factors. Since brain tissue is not accessible on a large scale, we profiled mRNAs of lymphoblast cell lines (LCLs) from three independent cohorts to determine whether we could detect a reproducible blood gene expression pattern associated with ASD. RNA from a total of 978 patients and 651 controls, including 607 unaffected siblings, was analyzed for differential expression. Although few genes were consistently differentially expressed between ASD and controls, we did find five (*CMKOR1*, *DKFZP564O0823*, *PITPNC1*, *PRKCB1* and *VIM*) that were differentially expressed in two LCLs studies and previously published brain samples. Similarly, using LCL gene expression to classify subjects by disease status performed only slightly above chance. Using weighted gene co-expression network analysis (WGCNA), we were able to identify a module correlated with ASD in both AGRE and NIMH cohorts that overlapped with genes previously found to be mis-expressed in post-mortem brain from ASD cases. eQTL analysis identified SNPs that were associated with LCL gene expression, including several in *AHI1*, a Joubert Syndrome gene

dysregulated in ASD brain and lymphoblasts.. Four of the 23 SNPs that were significantly correlated with the expression level of *AHI1* reside in the same haplotype block previously associated with ASD, suggesting that risk for ASD is mediated via AHI1 transcript levels. Overall, we found a weak, but consistent signal in LCLs further suggesting that peripheral lymphoblast gene expression may be useful for studying ASD.

1.2 Introduction

Autism spectrum disorders (ASDs) are a group of neuro-developmental disorders that are characterized by two core domains: deficits in social interaction, as well as restricted repetitive behaviors [1, 2] . Both family [3, 4] and twin studies [5] indicate ASDs are highly heritable neuropsychiatric disorders. The contribution of common and rare genetic variants to ASDs has been examined by different methods, including linkage analysis, genome-wide association studies (GWAS), copy number variation and exome-sequencing studies [6-14]. Heritability analysis indicates that 40%-60% of ASD is explained by common genetic variation [15]. CNV and exome sequencing analyses have identified rare variants that alter dozens of protein-coding genes in ASD. However, none of them individually accounts for more than 1% of ASD cases [16], although combined, rare variants are predicted to account for at least 15% of ASD [6, 17]. These results support an extremely heterogeneous genetic architecture for ASD, leading to its conceptualization as the ASDs [18].

Despite the heterogeneity of ASD, studies suggest ASD converges on a few specific biological pathways. A recent study shows that ASD risk genes tightly coexpressed in modules that implicate distinct biological functions during human cortical development [19]. At a circuit level, ASD genes are enriched in superficial cortical layers and glutamatergic projection neurons

[19]. Transcriptomic analysis, the comprehensive study of genes and their functions, offers an approach to study ASDs because of the ability to measure global gene expression changes. Since ASDs are neurodevelopmental disorders, brain tissue is the primary choice for functional analysis. The first transcriptomic study of autism identified about 30 differentially expressed genes in cerebellum and highlights the glutamate receptor as related to ASDs [20]. The most comprehensive study (Voineagu et al. [21]) to date examines transcriptome profiles from three different brain regions in post mortem brain tissue (frontal cortex, temporal cortex and cerebellum) and identified over 400 differentially expressed genes between patients with ASD and controls. Pathway analyses detected an up-regulation of genes involved in immune response and a down-regulation of genes involved in synaptic function. By using a network-based approach, Voineagu et al. [21] found a module (M12) of co-expressed genes enriched in genes with synaptic function and vesicular transport and down-regulated in ASD. This work illustrates the utility of gene expression profiling for understanding the pathophysiology of ASDs.

Since post-mortem tissue is hard to access on a large scale, several studies have analyzed genome-wide expression profiles of ASD cases using more readily available peripheral tissues such as lymphoblast cell lines (LCLs) [22-26] and blood [27-31]. An early study by Nishimura et al. [22] compared mRNA expression profiles in LCLs from males with autism due to a fragile X mutation (FMR1-FM) or a 15q11–q13 duplication (dup(15q)), and non-autistic controls. They identified 68 genes that are dysregulated in common between autism with FMR1-FM and dup(15q), as well as a potential molecular link between FMR1-FM and dup(15q), the cytoplasmic FMR1 interacting protein 1 (CYFIP1) [22]. Another study by Luo et al [32] used gene expression to annotate the pathogenicity of rare ASD-associated mutations and found distinct patterns of transcriptional dysregulation in several recurrent CNVs, including

(del)16p11.2 and (dup)7q11.23. There are several studies using LCLs which compare sporadic autistic cases with controls [22-26]. A few studies have been done by Hu et al in this field. In comparing gene expression profiles of LCLs from monozygotic twins discordant for autism severity, they identified 44 genes as differentially expressed in ASDs [25]. A comparison between sib pairs discordant for autism diagnosis identified 45 differentially expressed genes [24]. These genes are found to be involved in neural-development and steroid biosynthesis. Another study by Hu et al conducted expression profiling of 116 lymphoblastoid cell lines (LCL) from individuals with sporadic autism who are divided into three phenotypic subgroups according to severity scores [23]. Recently, Seno et al [26] used gene and miRNA expression profiling using LCL-derived total RNA to evaluate possible transcripts and networks of molecules involved in ASD. They identified several novel genes and miRNAs dysegulated in ASD compared with controls, including *HEY1*, *SOX9*, *miR-486* and *miR-181b*.

Many questions relating to transcriptional profiling of peripheral tissues remain, most prominently whether there is a common "ASD" expression signal in ASD and whether this reflects changes occurring in the brain. Over the five ASD transcriptome profiling studies, several pathways have been identified that are associated with ASDs in each study, including: neuronal development and steroid biosynthesis [24], circadian rhythms [23] and nature killer cytotoxicity [27]. However, differentially expressed genes and pathways identified in each study show little overlap. A recent review (Voineagu [33]) found 21 genes that were differentially expressed in both peripheral tissues and brain. This indicates the possibility of finding genes with shared expression alterations between brain and easily accessible peripheral tissues.

In an attempt to overcome issues of power and comparability, we conducted a rigorous expression profiling analysis of lymphoblast cell lines derived from more than 1000 samples

collected from three independent cohorts: AGRE (http://agre.autismspeaks.org), NIMH (http://www.nimh.nih.gov/index.shtml) and SSC (http://sfari.org/resources/simons-simplex-collection), the largest such gene expression analysis in any neuropsychiatric disease. We processed each cohort with the same criteria and an identical pipeline to make results comparable. Both standard differential expression and network-based methods were utilized to detect genes or gene clusters that are dysregulated in ASD across independent datasets. In addition, we compared our findings with differentially expressed genes identified in ASD post mortem brains to identify dysregulated genes shared between LCLs and brain tissue [21].

We also leveraged this large dataset to derive a comprehensive eQTL map, so as to further assess the function of ASD-associated genetic variation. Overall, our study provides evidence that peripheral tissues may be useful for studying ASD. However, compared with brain tissue, the signal from generalized expression changes in lymphoblasts is relatively weak, consistent with ASD's genetic heterogeneity.

1.3 Several differentially expressed (DEX) genes in LCLs overlap with DEX genes in autistic brain

We profiled the whole-genome mRNA of lymphoblast cell lines (LCLs) from three independent autism cohorts using Illumina microarrays. After pre-processing steps (Methods), 627 samples (283 cases and 344 controls) from 333 multiplex families in the AGRE cohort, 142 samples (99 cases and 43 controls) in the NIMH cohort, and 409 samples (221 cases and 188 controls) from 241 simplex families in the SSC cohort remained for downstream analysis (Table 1).

We first conducted differential expression analysis in each cohort (Methods) and

identified 417 differentially expressed (DEX) genes in AGRE, 680 in NIMH, and only 4 in SSC (p-value < 0.01) (Figure 1, Table S2). The proportion of genes identified in each cohort remained roughly similar regardless of the statistical cutoff used in these three cohorts (Figure 1). To evaluate whether the small number of DEX genes identified in SSC is due to large *de novo* CNVs in this cohort [34], we reanalyzed these data by removing the 54 samples with at least one large *de novo* CNV, but did not find any DEX genes (p-value < 0.01). This indicates that the autistic cases in SSC are less likely to share genes that are differentially expressed. Instead, individual autistic cases may exhibit specific mis-expressed genes potentially caused by rare genetic variants [32]. We then examined overlap between the DEX gene lists from each cohort. Fourteen genes were differentially expressed (p-value < 0.01) in the AGRE and NIMH cohorts, although this overlap did not reach significance (hypergeometric test; p-value = 0.558). The shared DEX genes include *TMPRSS3,* which is involved in sodium channel regulator activity; *MRPL41,* functioning in apoptosis and cell cycle; and *DKFZP564O0823*, which is also dysregulated in post mortem ASD brain [21].

To carry out a comprehensive comparison between LCLs and brain gene expression, we compiled a list of LCLs DEX genes reported in seven studies [22-25, 32] together with DEX genes (p-value <0.01) identified from this study. By comparing the comprehensive DEX gene list with the 444 DEX genes identified in autistic brain [35], we identified 56 genes as differentially expressed in both tissues (Table 2). Interestingly, there were five genes (*CMKOR1, DKFZP564O082, PITPNC1, PRKCB1* and *VIM*) which were altered in more than one study of peripheral blood or LCLs and brain.

To assess the convergence of DEX genes at the pathway level, we ran DAVID GO for the DEX

genes (p-value < 0.01) from AGRE and NIMH (Methods), but not the SSC, since there were too few genes in the latter. Mitochondrial related pathways were identified in both cohorts (in AGRE: Mitochondrial substrate/solute carrier (p-value = 1.1e-02); in NIMH: mitochondrial ribosome (p-value = 1.6e-02)). Interestingly, this pathway has been reported previously to be associated with ASD [36-39].

1.4 Prediction of autism using LCLs gene expression signatures

Several studies have assessed the potential to distinguish ASD cases from controls using either SNP or expression data [29, 40]. To test whether genes differentially expressed in LCLs could be used as molecular diagnostic biomarkers for ASDs, we utilized the DEX gene lists from each cohort to build prediction models (we call DEX_prediction model) with two powerful prediction methods: Random Forest (RF) and Support Vector Machines (SVM) (Methods). By using the DEX genes to classify samples from the cohort in which the DEX genes were identified, the Area Under Curve (AUC) in AGRE was 0.67, in NIMH was 0.76, while in SSC 0.69, all of which were slightly higher than the background AUC value (Figure 3), indicating a slightly higher classification power to distinguish cases from controls compared to background. However, when applying the classifier derived from one cohort to the other two independent cohorts, none of the DEX-predictions performed better than chance (Figure 3).

1.5 Network analysis identifies a neural-related module that is associated with ASDs in AGRE and NIMH cohorts

Weighted gene co-expression network analysis (WGCNA) is a systems biology method for leveraging the correlation patterns among genes across microarray samples which are used for finding clusters (modules) of highly correlated genes that correspond to shared biological

function [41]. This method has been successfully applied in various biological contexts including cancer , mouse genetics [41-43] as well as ASD brain [35]. To find discrete clusters of co-expressed genes showing transcriptional differences between autistic cases and controls, we built co-expression networks in each cohort (Methods). The expression levels of each module were summarized using the first principal component (the module eigengene (ME)). MEs were used to assess whether modules are related to clinical phenotypes, experimental variables or confounders. Here we assessed ME relationship to disease status, age, or sex as well as batch effects. In the NIMH cohort, five additional measured confounders were also included (Methods). In AGRE, we detected two modules: brown module (AGRE_Mbrown, correlation = 0.094, p-value = 0.02) and lightgreen module (AGRE_Mlightgreen, correlation = 0.11, p-value = 0.006) that were ASD-correlated with a significant nominal p-value (p-value < 0.05). In the NIMH cohort, the modules magenta (NIMH_Mmagenta, correlation = 0.16, p-value = 0.04) and purple (NIMH_Mpurple, correlation = -0.16, p-value = 0.04) were significantly correlated with ASD, but none of the confounders were correlated with disease. In the SSC, none of the modules were related to ASD, consistent with the differential expression analysis results in SSC cohort. Even in AGRE, the correlation value between the brown and lightgreen modules and ASD was relatively minor and was not significant after Bonferroni correction.

We next examined network and module reproducibility and their trait relationships across data sets. Supplemental Figure 2 shows that most of the modules are preserved (z summary > 2) at the network level, indicating that the modules detected represent groups of genes with robust co-expression patterns.

To explore whether the module –trait relationships were preserved, we focused on the

four modules (AGRE_Mbrown, AGRE_Mlightgreen, NIMH_Mmagenta, NIMH_Mpurple) that show significant correlation with ASD (Methods). We detected that AGRE_Mbrown module was the only module that was significantly correlated with disease status in both AGRE (correlation = 0.094, p-value = 0.02) and NIMH (correlation = 0.32, p-value = 3.5e-05), but not SSC (correlation = 0.0015, p-value = 0.97). Interestingly, this module overlapped significantly with M12 – an ASD-associated module identified in ASD post mortem brain co-expression network analyses (hypergeometric p-value = 5.5e-03). Furthermore, GO enrichment show genes in AGRE_Mbrown were enriched in neural-related pathways, such as neurological system processes (p-value = 4.2e-09), neuroactive ligand-receptor interaction (p-value = 3.7e-06) and cell-cell adhesion (p-value = 1.9e-04). One of the hub genes (genes with highest connectivity in a module) in AGRE_Mbrown was *KCNJ10*, a major player in astrocyte-mediated regulation of [K(+)](o) in brain, which harbors recurrent mutations in ASD [44].

1.6 eQTL identifies SNPs that are associated with gene expression patterns

To explore the potential genetic cause for the expression alteration in ASD versus controls, we conducted expression quantitative trait loci (eQTL) analysis. In our three cohorts, AGRE contained the largest number of samples with both expression and genotyping data. So a genome-wide eQTL analysis was conducted in the AGRE cohort using efficient mixed-model association (EMMAX) [45], which utilizes a kinship matrix to control for the population and family structure (Methods). After Bonferroni correction (p-value < 4.5e-12), 8813 *cis*-eQTL and 878 *trans*-eQTL remained significant.

Among the significant eQTLs, SNPs identified in gene *AHI1* were of particular interest, since mutations in *AHI1* cause specific forms of Joubert syndrome [46-48], and *AHI1* was down-

regulated in cases in both AGRE and ASD brain data sets [35]. Common variants in *AHI1* have been previously associated with schizophrenia and ASD [47, 48] and a high proportion of Joubert patients have an ASD. In our eQTL analysis, 23 SNPs were associated with the expression level of *AHI1* (*cis*-eQTL) (Figure 4). Previous work identified a linkage disequilibrium block associated with ASD [47], which contains four of the 23 SNPs influencing *AHI1* transcript levels (rs6931735, rs2064430, rs7772681 and rs13208164).

1.7 Discussion

Our results show that there are a few genes that were differentially expressed in LCLs in at least two independent ASD cohorts or previously published brain samples. Given the genetic heterogeneity observed in ASD, the lack of a strong gene expression signal shared across all patients and cohorts is unsurprising. Similarly, a gene expression-based classifier performed only slightly above chance within individual cohorts and consistent with chance across cohorts. Using WGCNA, a powerful network-based method, we were able to identify a module correlated with ASD in both the AGRE and NIMH cohorts which overlapped with genes previously found to be mis-expressed in post mortem brain from ASD cases. This suggests that there are shared, albeit weak signals identifiable in peripheral tissue that may reflect changes occurring in the brain, but that sophisticated analytic techniques may be needed to identify them more robustly.

eQTL analysis identified multiple SNPs associated with ASD risk genes, including *AHI1*, whose ASD association is supported by multiple lines of evidence [46-48]. The eQTL analysis links the previously identified genetic association signals in ASD and schizophrenia with transcriptome changes via the identification of significant eQTL within this gene. These data suggest that *AHI1* common variants exert their functional effects on ASD susceptibility [47] via

modulating *AHI1* transcript levels. This is consistent with recessive mutations in *AHI1* leading to a syndromic form of ASD and the significant reduction of *AHI1* transcript levels observed in ASD brain [21].

There are several issues for previous LCLs transcriptome studies, which may result in little convergence: (1) they used a relatively small sample size (total number of samples is smaller than 100); (2) Instead of a network-based method, they analyzed individual genes using t-tests, in which multiple testing problems are very conspicuous and results are easily contaminated by false positive discovery; (3) each analysis has a different study design, which makes the results less comparable. The three cohorts analyzed here are by far the largest datasets used for study of the transcriptome in ASD. To avoid the confounder of different microarray platforms and analyses, all of the samples were run on the Illumina platform and analyzed using the same pipeline. Thus, the confounding factors which may have limited consistency in results between previous studies cannot account for the lack of overlap observed here. The lack of significant overlap in the differentially expressed genes between these cohorts is likely due to multiple factors. Firstly, these cohorts are heterogeneous and were recruited with different objectives. The SSC consists of simplex families with the goal of enriching for rare, non-inherited genetic variation [6, 7], so common shared expression changes might not be expected in SSC, consistent with our observations. AGRE is a multiplex cohort recruited to identify heritable contributions [49] and the NIMH cohort is recruited without regard to family structure and contains single incidence and multiplex families. The non-significant overlap of DEX gene lists with this large sample size indicates the difficulty in finding convergening molecular alterations for sporadic ASD cases in LCLs. One interesting observation for the differential expression analysis is that we identified more differentially expressed genes in NIMH than in

AGRE and SSC. One possible reason is the different controls involved. In NIMH, the typically developing children were collected, while the other two cohorts used as controls unaffected siblings, which potentially carry some autistic features and likely reduce our power to detect the transcriptome changes related to autism. In AGRE, 50% of samples are gender matched while this is not the case for SSC cohort. This may bring additional heterogenicity for SSC cohort.

Secondly, LCLs are a peripheral tissue, not brain and furthermore represent a clone of EBV-transformed B cells. The transformation process can cause transcript and epigenetic alterations [50], thus adding another layer of variability. Only 60% of the genes expressed in the brain are also expressed in LCLs or blood [51]. This limits our power to detect brain-specific genes in peripheral tissues, especially when the genetic alterations present across samples are highly variable, as is the case here. This is in contrast to studying major gene forms of ASD where a shared LCL expression profile has been observed [22]. Thirdly, given the genetic heterogeneity of ASD, finding mutation- or patient-specific alterations may be a more powerful approach, as we have previously shown [32]. Outlier analysis, which detects mis-expression from the mean in specific individuals, has been shown to be able to detect ASD related pathways as well as the functional impact of rare CNVs [32]. Meanwhile, a systems biology method is preferred over a simple t-test to uncover the molecular etiology of ASD considering its complexity. The work by Parikshak et al [19] demonstrates the power of integrating multi-level evidence including genetic variation, transcriptome profiling and protein-protein networks using a systems biology method to detect converging pathways for ASD. Our results also show that by applying WGCNA, we were able to identify a module enriched in neural-related genes that are correlated with ASD.

Although few genes were consistently differentially expressed between ASD and controls, we did find five genes (*CMKOR1, DKFZP564O0823, PITPNC1, PRKCB1* and *VIM*) that were differentially expressed in LCLs from at least two cohorts and previously published brain gene expression studies. Among them, the *PRKCB1* gene was found to be associated with ASD [52]. Thus, several forms of evidence, none conclusive, support *PRKCB1* as a candidate gene for ASD. The other four genes have not been associated with ASD or studied extensively in other brain disorders.

Biomarkers that would facilitate earlier autism spectrum disorder diagnosis are crucial, thus several studies have assessed the potential to identify a classifier that can distinguish autism cases from controls using either SNP or expression data [29, 40]. A recent study reported a classifier with 70% accuracy if an individual has an autism spectrum disorder using 237 single-nucleotide polymorphisms (SNPs) [40]. However, further analyses show that ancestral origins is a potential confounding factors [53], so careful study design and reproducibility test in independent samples are critical to define any universal classifiers.

Considering its easily accessible nature, one aim is to find biomarkers in peripheral tissues for ASD diagnosis. In our study, we were unable to find a universal classifier that can predict disease status in multiple cohorts. Although previous studies have reported potential biomarkers for ASD in LCLs or blood, none have been replicated independently [22-31]. This indicates the difficulty of finding general biomarkers in peripheral tissues for ASD. Based on this, we suggest that identifying subgroups with more homogeneous etiologies may be the most productive way to conduct classifier analysis in ASD. Multi-level information (including genetics, transcriptomics and proteins) and more advanced statistical methods may also be needed to build a predictive model. This is supported by the current analyses, as with a network-

based approach, we were able to identify one module: AGRE_Mbrown that is ASD related (nominal p-value in AGRE and NIMH cohorts). Although the relationship between LCL gene expression and ASD is small compared to the correlation in brain tissue between M12 and ASD affection status [21], this is not unexpected considering we are using peripheral tissues. In the SSC, none of the modules correlated with ASD, and the AGRE_Mbrown was not preserved in SSC either. This suggests as expected that rare genetic variants including CNVs and SNVs [6, 7] make it difficult to detect any group difference between ASD and non-ASD samples in this cohort. These data call for future independent validation of the AGRE_Mbrown module in other cohorts to replicate its ASD association in AGRE and NIMH.

To take full advantage of this large data set, we conducted eQTL analysis to find any regulatory variants related to gene expression and ASD. Using a stringent statistical cutoff (Methods), we found only a small number of significant eQTLs. In our study, we detected more significant *cis*-eQTLs than *trans*-eQTL. This agrees with the previous observation that the large proportion of the intraspecific differences in transcript level is due to *cis*-effects on the gene [54-59]. Remarkably, the expression of *AHI1*, which is down-regulated in both AGRE LCLs and ASD post-mortem brain [35], is shown to be regulated by 23 of *cis* SNPs. One of the *cis* SNPs, rs6931735, has a high correlation (R^2=0.88) with SNP rs11154801, which is significantly associated with schizophrenia at the genome-wide level [48]. Also, this SNP is in the LD block identified to be associated with autism [47]. The convergence of the genetic association with ASD and the eQTL within AHI1 is intriguing and provides the first link between this association and disease mechanism.

In summary, we present the largest and most comprehensive genome-wide expression profiling study in ASDs. With three independent datasets, we are able to evaluate the

convergence of DEX gene list and the reproducibility of our predictions. Only weak signals are seen shared across cohorts, reflecting the genetic heterogeneity of the disorder. However, we do identify a handful of genes and a co-expression module shared by LCLs and brain tissue. Based on these results and others [21-26, 32], we are cautiously optimistic that LCLs may be useful for studying ASD. However, careful attention needs to be paid to study designs and objectives. So far, peripheral tissues appear most powerful for studying the impact of individual genetic variants, rather than shared, generalizable patterns [22, 32]. Whether whole blood would be more consistent than LCLs remains to be determined, since whole blood includes multiple cell types and is not EBV transformed. High throughput, efficient transformation of peripheral somatic cells to neural progenitors and neurons would likely provide the most optimal tissue for biomarker discovery. Until such tissue is available, LCLs remain a viable alternative.

1.8 Materials and Methods

Individuals and lymphoblast cell lines (LCLs) analyzed in this study

AGRE cohort was created by merging two large collections. The first collection contained 311 cases and 203 controls from 200 multiplex families; the second one contained 305 cases and 190 controls from 202 multiplex families. Multiplex families are defined as families with more than one affected sibling in a family while simplex families only contain one affected sibling. For Simon Simplex Collection cohort: we analyzed individuals from Simon Simplex Collection (SSC) in two stages. In the first stage, we collected 386 individuals from 196 simplex families (190 matched sib pairs plus 5 siblings and 1 proband). In the second stage, we prioritized 53 samples with *de novo* CNVs (42 probands, 8 siblings and 3 mothers who carry 16p11.2 events) [7]. Phenotype information can be found at Simons Foundation Autism

Research Initiative (SFARI) database and is shown in supplemental Table S1. This study was approved by the Institutional Review Boards at all participating institutions, including UCLA and Yale University. The NIMH cohort encompassed children enrolled in an ongoing longitudinal study of autism clinical subtypes who received clinical genetics testing. A total of 106 children with autism were included in this analysis. Typically developing children were included if they did not show signs of an ASDs or developmental delay.

The lymphoblast cell lines (LCLs) of the subjects were grown in RPMI 1640 medium with 2 mM L-glutamine and 25 mM HEPES (Invitrogen, Carlsbad, CA, USA), 10% fetal bovine serum, and 1 × Antibiotic-Antimycotic solution (Invitrogen) at 37°C in a humidified 5% CO2 chamber. Cells were grown to a density of 6×10^5/ml. Special attention was given to maintain all the cell lines in the same conditions to minimize environmental variation.

Microarray experiments

A total of 9×10^6 of lymphoblasts were seeded in a T-75 flask in 30 ml of fresh medium. After 24 h, total RNA was extracted from the cells using an RNeasy Mini Kit with DNase treatment (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. RNA quantity and quality were measured by ND-100 (Nanodrop, Wilmington, DE, USA) and 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), respectively. For AGRE samples, mRNA was hybridized on the Illumina Whole Human Genome Array Human REF-8 version 2.0 according to the manufacturer’s protocol. For SSC and NIMH samples, mRNA was hybridized on the Illumina Whole Human Genome Array Human REF-8 version 3.0 according to the manufacturer’s protocol.

Sample quality control

GenomeStudio was used to convert image to numerical data as per our typical protocols [35, 42, 43]. *SampleNetwork* function in R was applied for pre-processing, which including the following steps: 1) Samples (chips) were cross-correlated using expression levels for all probe sets. These inter-array correlations (IACs) [60] were averaged for each array and compared to the resulting distribution of IACs for the dataset [43]. Samples with an average IAC < 2.0 standard deviation below the mean IAC for the dataset were removed. 2) Following sample removal, quantile normalization [60] was performed in R. 3) To eliminate batch effects, additional normalization was performed using the R package *ComBat* [61] with the default parameters. *ComBat* successfully eliminated batch effects as evidenced by hierarchical clustering and significant improvement of mean IAC. We only used probes with evidence of robust expression (detection p-value <= 0.05 in at least 50% samples).

In each cohort, there are potential cofounders as gender and age that can affect the analysis. So we applied linear regression to regress out gender and age effects in each cohort and used the residuals for follow up analysis. In NIMH cohort, additional five cofounders are reported including fasting status, medication, special diets, supplement taken status and sedation status (Table S1). To keep consistent between the NIMH and other two cohorts, we only regress out the gender and age in NIMH as well, but utilized *Limma* package to detect the differentially expressed (DEX) genes for each cofounder in NIMH. We then removed genes that are altered (p-value < 0.005) in any of the five cofounders. One special case is for sedation status since it's highly cofounded with the autistic case status: 90 out of 99 (91%) autistic cases are sedated while none of the controls are sedated. If we used *Limma* to detect the DEX genes between sedated and un-sedated in the cohort, it's likely we will also find genes that are potentially associated with autistic trait. To overcome this issue, we applied differential expression analysis for sedation in

autistic cases only (compare the 90 sedated cases versus 9 un-sedated cases). In total, 1330 out of 9096 (15%) DEX genes were identified in least one cofounder and are removed from the NIMH gene list.

Genome-wide differential expression (DE) analysis

The *Limma* [62] package in R was applied for standard differential expression analysis each cohort.

Classification analysis

We utilized differentially expressed (DEX) gene lists from each cohort to build prediction models with two popular prediction methods: Random Forest (RF) and Support Vector Machine (SVM) (Methods). We first used the DEX genes to classify case from control in each cohort. This is called classification step. At this step, 10-fold cross validation is applied to measure the classification accuracy (area under curve (AUC) value to indicate the accuracy). After that, DEX gene lists from one certain cohort were used to predict the disease status in the other two cohorts. This can help to test the prediction power of a model in independent cohorts, which is called prediction step. Classifier with all expressed genes as features was used as background test.

Weighted co-expression network analysis

To identify clusters of co-expressed genes that are potentially related to ASD, WGCNA was applied to three cohorts respectively following the standard procedure for generating a signed network[63, 64]. In short, pairwise Pearson correlation coefficients were calculated for the expressed probes across all samples, and converted to connection strengths, defined as $[(1+\text{correlation})/2]^{\beta}$ where $\beta=10$ [63]. These adjacency matrices were then used as the basis for

defining topological overlap (TO), which measures the common connections between a pair of genes. For each network, we then used average linkage hierarchical clustering to group genes on the basis of TO dissimilarity measure (1 – TO). Finally, modules were defined using a dynamic tree-cutting algorithm[64]. The module eigengene (ME), defined as the first principle component of a given module, was used to represent characteristic expression patterns of individual modules. Module preservation was done by using the *modulePreservation* function (with default settings) in R. This function applies a permutation-based method to evaluate the robustness of the connections within a module in independent datasets.

Pathway analysis

DAVID GO database and MetaCore by GeneGO (Thompson Reuters) were used for pathway analyses. For both analyses, the background was set to the total list of genes expressed in our dataset. The statistical significance threshold level for all GO enrichment analyses was $p<0.05$.

eQTL analysis

Genome-wide e-QTL analysis was conducted on AGRE cohort using efficient mixed-model association (EMMAX). All expressed probe (10773 probes) and 1040268 SNPs were used as the input. To remove the gender and age effects, the residuals of the expression values were used after regressing out gender and age. Kinship matrix was measured using the standard EMMAX parameters to control for the population and family structures. The output of eQTL had a lamda = 0.98, indicating the population and family structures were well controlled. We define SNPs spanning in the transcript boundary plus the 500kb upstream and downstream regions as *cis*-eQTL. Bonferroni corrected p-value (4.5e-12) is used as cutoff.

Figure 1-1

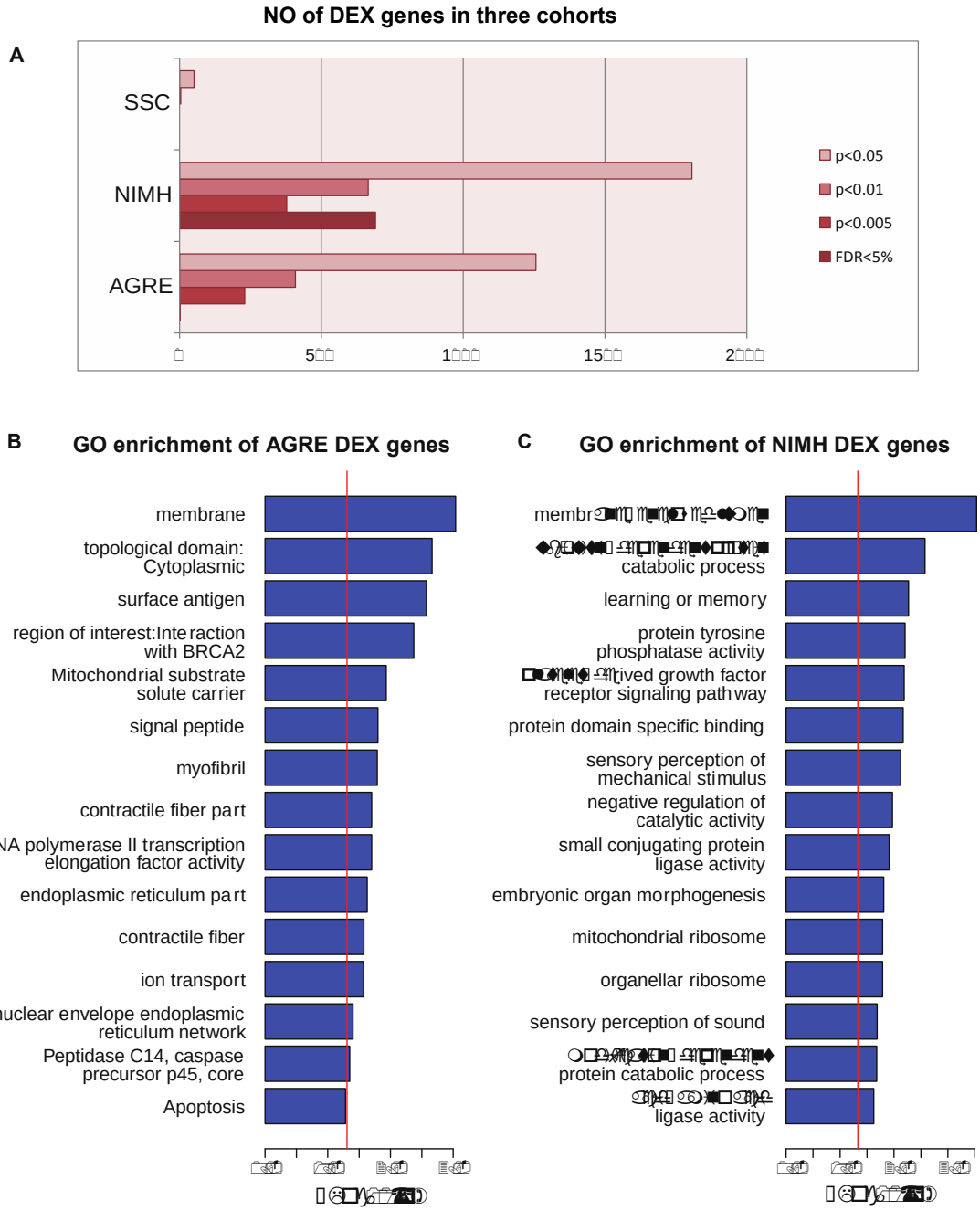

A
NO of DEX genes in three cohorts
SSC
NIMH
AGRE
p<0.05
p<0.01
p<0.005
FDR<5%
B
GO enrichment of AGRE DEX genes
membrane
topological domain: Cytoplasmic
surface antigen
region of interest:Interaction with BRCA2
Mitochondrial substrate solute carrier
signal peptide
myofibril
contractile fiber part
RNA polymerase II transcription elongation factor activity
endoplasmic reticulum part
contractile fiber
ion transport
nuclear envelope endoplasmic reticulum network
Peptidase C14, caspase precursor p45, core
Apoptosis
C
GO enrichment of NIMH DEX genes
learning or memory
protein tyrosine phosphatase activity
protein domain specific binding
sensory perception of mechanical stimulus
negative regulation of catalytic activity
small conjugating protein ligase activity
embryonic organ morphogenesis
mitochondrial ribosome
organellar ribosome
sensory perception of sound

Figure 1-1. Differential expression analysis between ASD cases and controls. A) Barplot shows the number of genes as differentially expressed in AGRE, NIMH and SSC with different cutoffs: p-value < 0.05, p-value < 0.01, p-value < 0.005 and FDR < 5%. B) Barplot shows the pathways with significant p-value (p-value < 0.05) from DAVID GO enrichment analysis based on differentially expressed genes (p-value < 0.01) in AGRE and NIMH.

Figure 1-2

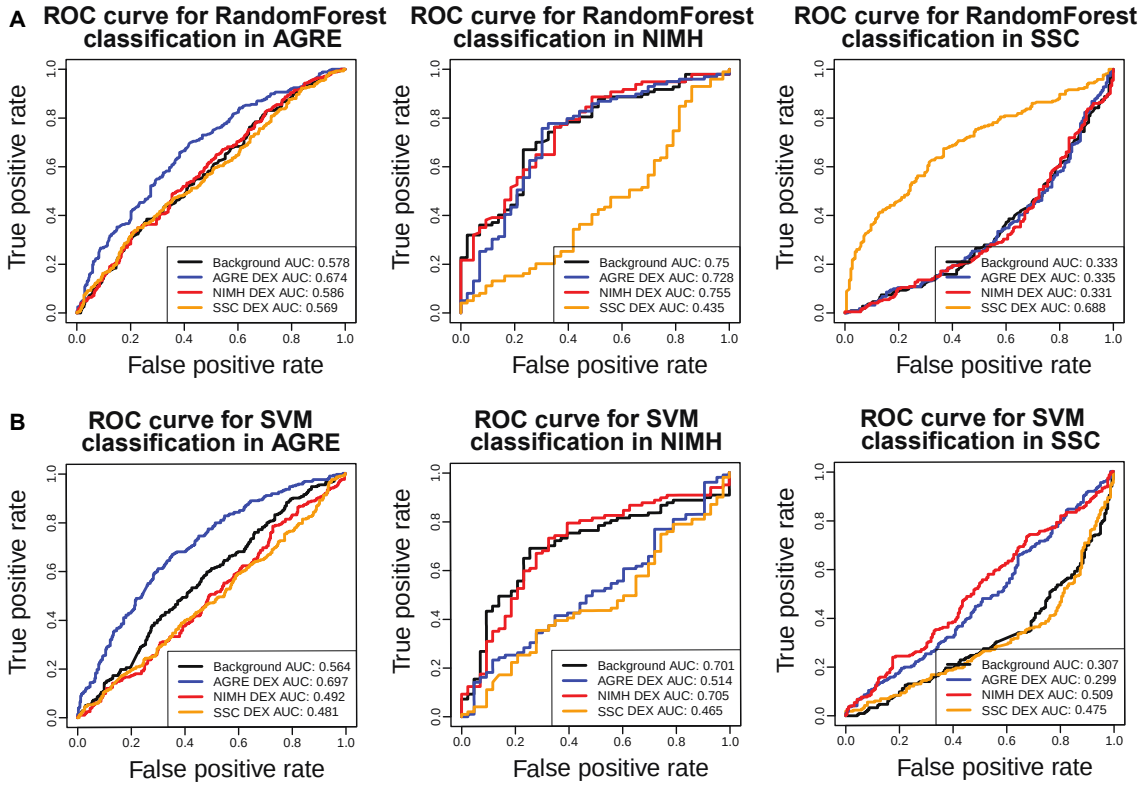

A
ROC curve for RandomForest classification in AGRE
Background AUC: 0.578
AGRE DEX AUC: 0.674
NIMH DEX AUC: 0.586
SSC DEX AUC: 0.569
ROC curve for RandomForest classification in NIMH
Background AUC: 0.75
AGRE DEX AUC: 0.728
NIMH DEX AUC: 0.755
SSC DEX AUC: 0.435
ROC curve for RandomForest classification in SSC
Background AUC: 0.333
AGRE DEX AUC: 0.335
NIMH DEX AUC: 0.331
SSC DEX AUC: 0.688
B
ROC curve for SVM classification in AGRE
Background AUC: 0.564
AGRE DEX AUC: 0.697
NIMH DEX AUC: 0.492
SSC DEX AUC: 0.481
ROC curve for SVM classification in NIMH
Background AUC: 0.701
AGRE DEX AUC: 0.514
NIMH DEX AUC: 0.705
SSC DEX AUC: 0.465
ROC curve for SVM classification in SSC
Background AUC: 0.307
AGRE DEX AUC: 0.299
NIMH DEX AUC: 0.509
SSC DEX AUC: 0.475
True positive rate
False positive rate

Figure 1-2. Classification analysis reveals weak signals for predicting ASD status based on LCLs expression. A) Each ROC curve shows the RandomForest classification signals based on differentially expressed genes and background gene list in AGRE, NIMH and SSC cohort respectively. Area under curve (AUC) value is reported for each classification model. B) Each ROC curve shows the Support Vector Machine (SVM) classification signals based on differentially expressed genes and background gene list in three cohorts. AUC value is reported for each classification model.

Figure 1-3

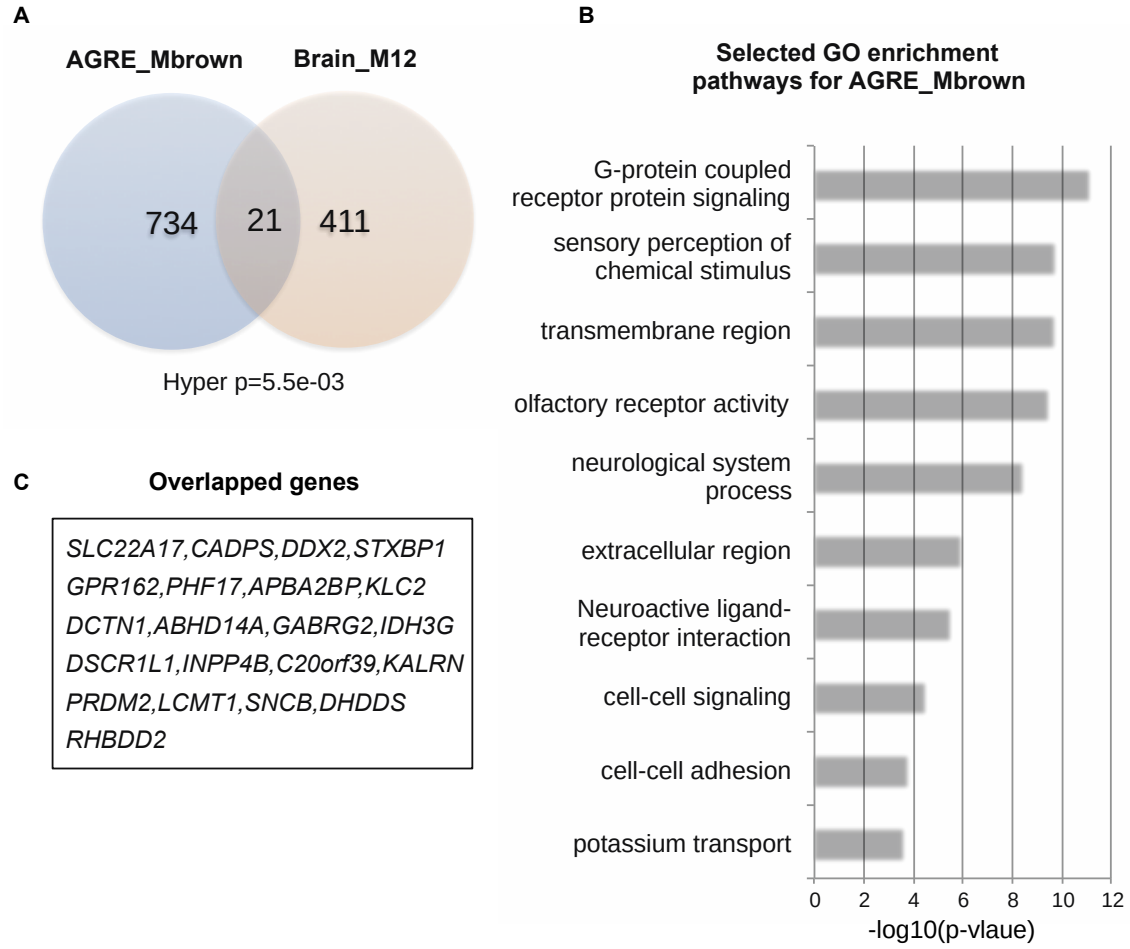

A
AGRE_Mbrown
Brain_M12
734
21
411
Hyper p=5.5e-03
B
Selected GO enrichment pathways for AGRE_Mbrown
G-protein coupled receptor protein signaling
sensory perception of chemical stimulus
transmembrane region
olfactory receptor activity
neurological system process
extracellular region
Neuroactive ligand-receptor interaction
cell-cell signaling
cell-cell adhesion
potassium transport
0
2
4
6
8
10
12
-log10(p-vlaue)
C
Overlapped genes
SLC22A17,CADPS,DDX2,STXBP1
GPR162,PHF17,APBA2BP,KLC2
DCTN1,ABHD14A,GABRG2,IDH3G
DSCR1L1,INPP4B,C20orf39,KALRN
PRDM2,LCMT1,SNCB,DHDDS
RHBDD2

Figure 1-3. Network analysis identified AGRE_Mbrown module to be ASD related. A) Venn-diagram shows the significant overlap (hypergeometric p-value = 5.5e-03) between AGRE_Mbrown module and M12, an ASD module identified in brain samples [21]. B) The list of overlapped genes between AGRE_Mbrown and M12. C) Selected top GO enrichment pathways associated with AGRE_Mbrown.

Table 1-1. Summary table of sample information.

Source	NO. of samples	NO. of cases	NO. of controls	Family structure
AGRE	627	283	344	Multiplex families
NIMH	142	99	43	Mixed multiplex+simplex families
SSC	409	221	188	Simplex families

Table 1-2. List of genes as differentially expressed in both LCLs and brain.

Gene.Name	LCLs Resource
CMKOR1	**15q dup vs. control; Nishimura et al. (2007) Human molecular genetics [22]** **FMR1 vs. control; Nishimura et al. 2007 Human Molecular Genetics [22]**
DKFZP564O0823	**AGRE cohort** **NIMH cohort**
PITPNC1	**Hu et al. (2006) BMC Genomics [25]** **15q dup vs. control; Nishimura et al. (2007) Human molecular genetics [22]** **FMR1 vs. control; Nishimura et al. (2007) Human molecular genetics [23]**
PRKCB1	**15q dup vs. control; Nishimura et al. (2007) Human molecular genetics [22]** **FMR1 vs. control; Nishimura et al. (2007) Human molecular genetics [22]**
VIM	**AGRE cohort** **FMR1 vs. control; Nishimura et al. (2007) Human molecular genetics [22]** **15q dup vs. control; Nishimura et al. (2007) Human molecular genetics [22]**
ALDH4A1	AGRE cohort
ARMC8	Hu et al. (2009) Autism Res. [23]
ATP2B2	Hu et al. (2009) PLOS ONE [24]
ATP6V0D1	NIMH cohort
CCDC50	Hu et al. (2009) Autism Res. [23]
CD44	Hu et al. (2009) Autism Res. [23]
CD74	NIMH cohort
CIRBP	NIMH cohort
CPNE3	AGRE cohort
DNAJB1	16p11.2 del vs. control; Luo et al. (2012) AJHG [32]
DNAJB1	7q11.23 dup vs. control; Luo et al. (2012) AJHG [32]
ELMOD1	Hu et al. (2006) BMC Genomics [25]
FCGBP	Hu et al. (2009) Autism Res. [23]
GNA12	NIMH cohort
GNA13	NIMH cohort
HIST2H2AC	NIMH cohort
HSPB1	AGRE cohort
IFITM3	Seno et al. (2011) Brain Res. [26]
INPPL1	16p11.2 dup vs. control; Luo et al. (2012) AJHG [32]

ITGB5	7q11.23 dup vs. control; Luo et al. (2012) AJHG [32]
ITPR1	AGRE cohort
KIF1C	NIMH cohort
LCP1	Hu et al. (2009) Autism Res. [23]
LOC400566	Seno et al. (2011) Brain Res. [26]
LY96	NIMH cohort
LYPD1	16p11.2 dup vs. control; Luo et al. (2012) AJHG [32]
MSI2	AGRE cohort
NAP1L5	AGRE cohort
NEFM	AGRE cohort
NQO1	NIMH cohort
PFTK1	Hu et al. (2009) Autism Res. [23]
PLEKHC1	7q11.23 dup vs. control; Luo et al. (2012) AJHG [32]
PLOD2	NIMH cohort
PLTP	AGRE cohort
PNKD	NIMH cohort
PREPL	16p11.2 del vs. control; Luo et al. (2012) AJHG [32]
RHBDF2	NIMH cohort
SAT	NIMH cohort
SLC16A9	AGRE cohort
SLC29A1	16p11.2 dup vs. control; Luo et al. (2012) AJHG [32]
SLC2A5	AGRE cohort
SMYD2	AGRE cohort
SOX9	Seno et al. (2011) Brain Res. [26]
TARBP1	NIMH cohort
TESC	NIMH cohort
TNFRSF1A	16p11.2 dup vs. control; Luo et al. (2012) AJHG [32]
TNPO1	NIMH cohort
TRIM37	NIMH cohort
UCHL1	Seno et al. (2011) Brain Res. [26]
VHL	NIMH cohort
ZFP36	FMR1 vs. control; Nishimura et al. (2007) Human Molecular Genetics [22]

Table 1-3. eQTLs significantly associated with AHI1 expression.

SNP	Satistics	Pvalue	Loci	Minor Allele	Major Allele	eQTL Type	In_gene	In_500kb
rs6931735	-0.119	5.11E-28	135666504	G	A	cis	YES	YES
rs2064430	-0.111	3.31E-25	135684449	T	C	cis	YES	YES
rs7772681	-0.109	1.42E-23	135690706	C	T	cis	YES	YES
rs13208164	-0.103	1.74E-15	135762978	A	G	cis	YES	YES
rs11154801	-0.149	3.46E-42	135781048	A	C	cis	YES	YES
rs717120	-0.106	1.12E-13	135806822	C	T	cis	No	YES
rs9389295	-0.116	6.77E-28	135889916	C	T	cis	No	YES
rs9402715	-0.117	6.48E-27	135901390	G	A	cis	No	YES
rs9402717	-0.107	8.35E-25	135912471	T	C	cis	No	YES
rs719885	-0.139	2.99E-34	135981048	G	A	cis	No	YES
rs3734213	-0.140	1.43E-34	135983001	C	T	cis	No	YES
rs4896157	-0.124	1.45E-27	135983812	G	A	cis	No	YES
rs9494312	-0.098	3.44E-17	136009657	T	G	cis	No	YES
rs12210389	-0.102	6.19E-20	136024808	A	G	cis	No	YES
rs9321521	-0.091	3.12E-15	136031343	A	G	cis	No	YES
rs7739635	-0.092	7.30E-16	136039471	T	C	cis	No	YES
rs10223338	-0.109	2.38E-21	136043720	T	C	cis	No	YES
rs12202212	-0.104	5.80E-20	136060840	T	C	cis	No	YES
rs1475069	-0.092	1.21E-15	136097927	C	A	cis	No	YES
rs2038551	-0.092	9.83E-16	136100870	T	G	cis	No	YES
rs9399161	-0.086	2.67E-14	136122157	G	A	cis	No	YES
rs947583	-0.094	6.57E-16	136175352	C	T	cis	No	YES

Figure 1-S1

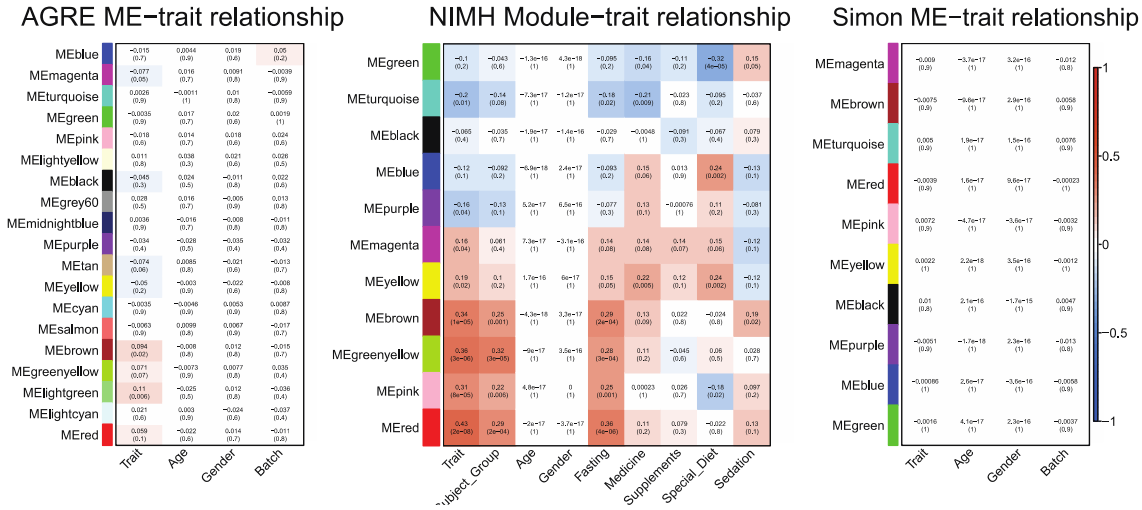

AGRE ME−trait relationship
MEblue
MEmagenta
MEturquoise
MEgreen
MEpink
MElightyellow
MEblack
MEgrey60
MEmidnightblue
MEpurple
MEtan
MEyellow
MEcyan
MEsalmon
MEbrown
MEgreenyellow
MElightgreen
MElightcyan
MEred
Trait
Age
Gender
Batch
NIMH Module−trait relationship
MEgreen
MEturquoise
MEblack
MEblue
MEpurple
MEmagenta
MEyellow
MEbrown
MEgreenyellow
MEpink
MEred
Trait
Subject_Group
Age
Gender
Fasting
Medicine
Supplements
Special_Diet
Sedation
Simon ME−trait relationship
MEmagenta
MEbrown
MEturquoise
MEred
MEpink
MEyellow
MEblack
MEpurple
MEblue
MEgreen
Trait
Age
Gender
Batch
1
0.5
0
−0.5
−1

Figure 1-S1. Module Eigengene correlation with ASD trait in AGRE, NIMH and SSC cohort respectively. X-axis is the trait and other variables including age, gender and batch. In NIMH, five confounders are also included. Y-axis is the module labels in each network. Red indicates positive correlation while blue indicates negative correlation.

Figure 1-S2

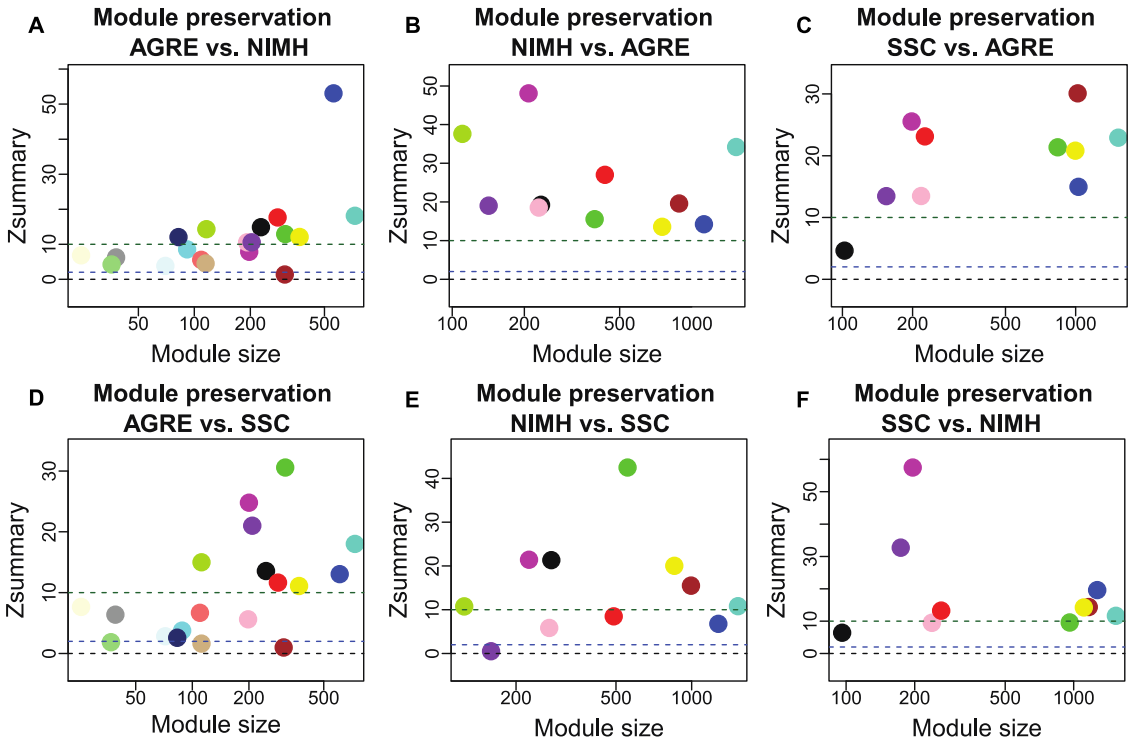

A
Module preservation
AGRE vs. NIMH
B
Module preservation
NIMH vs. AGRE
C
Module preservation
SSC vs. AGRE
D
Module preservation
AGRE vs. SSC
E
Module preservation
NIMH vs. SSC
F
Module preservation
SSC vs. NIMH
Zsummary
Module size

Figure 1-S2. Module preservation analysis across cohorts. ModulePreservation function in R is used to measure the preservation level of modules. In each panel, reference cohort is labeled firstly in the title followed by test cohort. Y-axis shows Z summary scores based on multiple network statistics and x-axis shows the module size.

1.9 References

1. Abrahams BS, Geschwind DH: Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics* 2008, 9(5):341-355.

2. Miles JH: Autism spectrum disorders--a genetics review. *Genetics in Medicine : Official Journal of the American College of Medical Genetics* 2011, 13(4):278-294.

3. Jorde LB, Hasstedt SJ, Ritvo ER, Mason-Brothers A, Freeman BJ, Pingree C, McMahon WM, Petersen B, Jenson WR, Mo A: Complex segregation analysis of autism. *American Journal of Human Genetics* 1991, 49(5):932-938.

4. Bolton PF, Pickles A, Murphy M, Rutter M: Autism, affective and other psychiatric disorders: patterns of familial aggregation. *Psychol Med* 1998, 28(2):385-395.

5. Ronald A, Hoekstra RA: Autism spectrum disorders and autistic traits: a decade of new twin studies. *American Journal of Med Genet Neuropsychiatr Genet* 2011, 156B(3):255-274.

6. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012, 485(7397):237-241.

7. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA *et al*: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011, 70(5):863-885.

8. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A *et al*: De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012, 74(2):285-299.

9. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J *et al*: Strong association of de novo copy number mutations with autism. *Science* 2007, 316(5823):445-449.

10. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y *et al*: Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics* 2008, 82(2):477-488.

11. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD *et al*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012, 485(7397):246-250.

12. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM *et al*: Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009, 459(7246):528-533.

13. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT *et al*: A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics* 2010, 19(20):4072-4082.

14. Prandini P, Pasquali A, Malerba G, Marostica A, Zusi C, Xumerle L, Muglia P, Da Ros L, Ratti E, Trabetti E *et al*: The association of rs4307059 and rs35678 markers with autism spectrum disorders is replicated in Italian families. *Psychiatric Genetics* 2012, 22(4):177-181.

15. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, Moreno-De-Luca D, Yu TW, Fombonne E, Geschwind D *et al*: Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism* 2012, 3(1):9.

16. Devlin B, Scherer SW: Genetic architecture in autism spectrum disorder. *Current Opinion in Genetics & Development* 2012, 22(3):229-237.

17. Stein JL, Parikshak NN, Geschwind DH: Rare inherited variation in autism: beginning to see the forest and a few trees. *Neuron* 2013, 77(2):209-211.

18. Geschwind DH, Levitt P: Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology* 2007, 17(1):103-111.

19. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH: Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 2013, 155(5):1008-1021.

20. Purcell AE, Jeon OH, Zimmerman AW, Blue ME, Pevsner J: Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 2001, 57(9):1618-1628.

21. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011, 474(7351):380-384.

22. Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, Sigman M, Steindler C, Pellegrini S, Schanen NC, Warren ST *et al*: Genome-wide expression profiling

of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics* 2007, 16(14):1682-1698.

23. Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME, Luu T, Lai Y, Lee NH: Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Research : Official Journal of the International Society for Autism Research* 2009, 2(2):78-97.

24. Hu VW, Nguyen A, Kim KS, Steinberg ME, Sarachana T, Scully MA, Soldin SJ, Luu T, Lee NH: Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis. *PloS One* 2009, 4(6):e5775.

25. Hu VW, Frank BC, Heine S, Lee NH, Quackenbush J: Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics* 2006, 7:118.

26. Ghahramani Seno MM, Hu P, Gwadry FG, Pinto D, Marshall CR, Casallo G, Scherer SW: Gene and miRNA expression profiles in autism spectrum disorders. *Brain Research* 2011, 1380:85-97.

27. Gregg JP, Lit L, Baron CA, Hertz-Picciotto I, Walker W, Davis RA, Croen LA, Ozonoff S, Hansen R, Pessah IN *et al*: Gene expression changes in children with autism. *Genomics* 2008, 91(1):22-29.

28. Enstrom AM, Lit L, Onore CE, Gregg JP, Hansen RL, Pessah IN, Hertz-Picciotto I, Van de Water JA, Sharp FR, Ashwood P: Altered gene expression and function of peripheral blood natural killer cells in children with autism. *Brain, Behavior, and Immunity* 2009, 23(1):124-133.

29. Kong SW, Collins CD, Shimizu-Motohashi Y, Holm IA, Campbell MG, Lee IH, Brewster SJ, Hanson E, Harris HK, Lowe KR *et al*: Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PloS One* 2012, 7(12):e49475.

30. Glatt SJ, Tsuang MT, Winn M, Chandler SD, Collins M, Lopez L, Weinfeld M, Carter C, Schork N, Pierce K *et al*: Blood-based gene expression signatures of infants and toddlers with autism. *Journal of the American Academy of Child and Adolescent Psychiatry* 2012, 51(9):934-944 e932.

31. Kong SW, Shimizu-Motohashi Y, Campbell MG, Lee IH, Collins CD, Brewster SJ, Holm IA, Rappaport L, Kohane IS, Kunkel LM: Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. *Neurogenetics* 2013, 14(2):143-152.

32. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, Cai C, Ou J, Lowe JK *et al*: Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *American Journal of Human Genetics* 2012, 91(1):38-55.

33. Voineagu I: Autism: from genetics to biomarkers. *Disease markers* 2012, 33(5):223-224.

34. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, Dilullo NM, Parikshak NN, Stein JL *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature doi:101038/nature10945* 2012.

35. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2010, 474(7351):380-384.

36. Legido A, Jethva R, Goldenthal MJ: Mitochondrial dysfunction in autism. *Seminars in Pediatric Neurology* 2013, 20(3):163-175.

37. Avdjieva-Tzavella D, Mihailova S, Lukanov C, Naumova E, Simeonov E, Tincheva R, Toncheva D: Mitochondrial DNA mutations in two bulgarian children with autistic spectrum disorders. *Balkan journal of medical genetics : BJMG* 2012, 15(2):47-54.

38. Gu F, Chauhan V, Kaur K, Brown WT, LaFauci G, Wegiel J, Chauhan A: Alterations in mitochondrial DNA copy number and the activities of electron transport chain complexes and pyruvate dehydrogenase in the frontal cortex from subjects with autism. *Translational Psychiatry* 2013, 3:e299.

39. Essa MM, Subash S, Braidy N, Al-Adawi S, Lim CK, Manivasagam T, Guillemin GJ: Role of NAD(+), Oxidative Stress, and Tryptophan Metabolism in Autism Spectrum Disorders. *International Journal of Tryptophan Research : IJTR* 2013, 6(Suppl 1):15-28.

40. Skafidas E, Testa R, Zantomio D, Chana G, Everall IP, Pantelis C: Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *Molecular Psychiatry* 2012.

41. Cai C, Langfelder P, Fuller TF, Oldham MC, Luo R, van den Berg LH, Ophoff RA, Horvath S: Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics* 2010, 11:589.

42. Miller JA, Horvath S, Geschwind DH: Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* 2010, 107(28):12698-12703.

43. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: Functional organization of the transcriptome in human brain. *Nature Neuroscience* 2008, 11(11):1271-1282.

44. Sicca F, Imbrici P, D'Adamo MC, Moro F, Bonatti F, Brovedani P, Grottesi A, Guerrini R, Masi G, Santorelli FM *et al*: Autism with seizures and intellectual disability: possible causative role of gain-of-function of the inwardly-rectifying K+ channel Kir4.1. *Neurobiology of Disease* 2011, 43(1):239-247.

45. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: Efficient control of population structure in model organism association mapping. *Genetics* 2008, 178(3):1709-1723.

46. Lotan A, Lifschytz T, Slonimsky A, Broner EC, Greenbaum L, Abedat S, Fellig Y, Cohen H, Lory O, Goelman G *et al*: Neural mechanisms underlying stress resilience in Ahi1 knockout mice: relevance to neuropsychiatric disorders. *Molecular Psychiatry* 2013.

47. Alvarez Retuerto AI, Cantor RM, Gleeson JG, Ustaszewska A, Schackwitz WS, Pennacchio LA, Geschwind DH: Association of common variants in the Joubert syndrome gene (AHI1) with autism. *Human Molecular Genetics* 2008, 17(24):3887-3896.

48. Torri F, Akelai A, Lupoli S, Sironi M, Amann-Zalcenstein D, Fumagalli M, Dal Fiume C, Ben-Asher E, Kanyas K, Cagliani R *et al*: Fine mapping of AHI1 as a schizophrenia susceptibility gene: from association to evolutionary evidence. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 2010, 24(8):3066-3082.

49. Lajonchere CM, Consortium A: Changing the landscape of autism research: the autism genetic resource exchange. *Neuron* 2010, 68(2):187-191.

50. Caliskan M, Cusanovich DA, Ober C, Gilad Y: The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics* 2011, 20(8):1643-1652.

51. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* 2009, 10(3):184-194.

52. Philippi A, Roschmann E, Tores F, Lindenbaum P, Benajou A, Germain-Leclerc L, Marcaillou C, Fontaine K, Vanpeene M, Roy S *et al*: Haplotypes in the gene encoding protein kinase c-beta (PRKCB1) on chromosome 16 are associated with autism. *Molecular Psychiatry* 2005, 10(10):950-960.

53. Robinson EB, Howrigan D, Yang J, Ripke S, Anttila V, Duncan LE, Jostins L, Barrett JC, Medland SE, Macarthur DG *et al*: Response to 'Predicting the diagnosis of autism spectrum disorder using gene pathway analysis'. *Molecular Psychiatry* 2013.

54. Hulse AM, Cai JJ: Genetic variants contribute to gene expression variability in humans. *Genetics* 2013, 193(1):95-108.

55. Bushel PR, McGovern R, Liu L, Hofmann O, Huda A, Lu J, Hide W, Lin X: Population differences in transcript-regulator expression quantitative trait loci. *PLoS One* 2012, 7(3):e34286.

56. Gaffney DJ, Veyrieras JB, Degner JF, Pique-Regi R, Pai AA, Crawford GE, Stephens M, Gilad Y, Pritchard JK: Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* 2012, 13(1):R7.

57. Becker J, Wendland JR, Haenisch B, Nothen MM, Schumacher J: A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals. *European Journal of Human Genetics : EJHG* 2012, 20(1):97-101.

58. Min JL, Taylor JM, Richards JB, Watts T, Pettersson FH, Broxholme J, Ahmadi KR, Surdulescu GL, Lowy E, Gieger C *et al*: The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS One* 2011, 6(7):e22070.

59. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008, 4(10):e1000214.

60. Gold DL, Wang J, Coombes KR: Inter-gene correlation on oligonucleotide arrays: how much does normalization matter? *American Journal of Pharmacogenomics* 2005, 5(4):271-279.

61. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8(1):118-127.

62. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004, 3:Article3.

63. Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005, 4:Article17.

64. Langfelder P, Zhang B, Horvath S: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008, 24(5):719-720.

CHAPTER 2: Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders

2.1 Abstract

Copy Number Variants (CNVs) are a major contributor to the pathophysiology of autism spectrum disorder, but the functional impact of CNVs remains largely unexplored. Since brain tissue is not available from most samples, we interrogated gene expression in lymphoblasts from 244 families with discordant siblings in the Simons Simplex Collection in order to identify potentially pathogenic variation. Our results reveal that the overall frequency of significantly mis-expressed genes (which we refer to here as outliers) identified in probands and unaffected siblings do not differ. However, in probands, but not their unaffected siblings, the group of outlier genes is significantly enriched in neural-related pathways including neuropeptide signaling, synaptogenesis and cell adhesion. We demonstrate that outlier genes cluster within the most pathogenic CNVs (rare *de novo* CNVs) and can be used to prioritize rare CNVs of potentially unknown significance. Several non-recurrent CNVs with significant gene expression alterations are identified, including deletions on chromosome 3q27, 3p13 and 3p26, and duplications at 2p15, suggesting these as potential novel ASDs loci. In addition, we identify distinct pathways disrupted in 16p11.2 microdeletions, microduplications and 7q11.23 duplications, and show that specific genes within the 16p CNV interval correlate with differences in head circumference, an Typically developing children were included if they did not show signs of an ASDs or developmental delay ASDs relevant phenotype. This study provides evidence that pathogenic structural variants have functional impact on transcriptome alterations

in ASDs at a genome-wide level, and demonstrates the utility of this approach for prioritization of genes for subsequent functional analysis.

2.2 Introduction

Autism, also known as autism spectrum disorders (ASDs [MIM 209850]), is a heterogeneous syndrome defined by impairments in three core domains: social interaction, language and range of interests [1, 2]. Autistic disorder is not viewed in isolation, but rather as one of several entities collectively referred to as the autism spectrum disorders (ASDs) [1]. Both family [3, 4] and twin studies [5] indicate ASDs are highly heritable neuropsychiatric disorders. A growing body of literature reveals that rare mutations or structural variations dramatically increase disease risk [6-11]. This evidence suggests rare genetic variation plays a larger role in ASDs than previously suspected [2, 12-14].

The discovery of rare and recurrent copy number variation (CNV) as important pathogenic mutations in ASDs was a watershed in ASDs genetics [7, 8]. Recurrent CNVs such as those at 16p11.2, 22q11.2, 1q21.1, 7q.23 and 15q11-q13 show statistically significant association with ASDs [15-17],[18, 19]. However, the functional impact of these CNVs on downstream RNA expression at both a collective and individual level remains largely unknown. Since CNVs alter copy number and presumably must act via changes in downstream gene expression, an initial study that explored the transcriptome-wide effects of CNVs in human lymphoblast cell lines reported that changes in gene copy number explained roughly 20% of detected transcriptional alterations [20]. Although widely assumed, it remains unknown whether rare CNVs identified in autistic individuals have similar effects on transcription levels and subsequent pathophysiology. Evidence certainly exists for the association of rare CNVs as a

group in ASDs, but the paucity of cases prohibits proof of genetic association for most individual rare CNVs. Alternative lines of evidence, such as gene expression data may provide converging evidence for functional alterations related to a particular CNV, and would thus be of significant utility.

No risk locus has been identified with a frequency exceeding ~1% in affected samples, consistent with heterogeneity [18, 21]. Our experimental strategy is predicated on the assumption that, instead of treating ASDs cases as a group, analyses of individuals at the resolution of the single gene would yield valuable insight. First, we analyzed gene expression variance in families with discordant siblings (one affected and one unaffected) from the Simons Simplex Collection (SSC). Since brain or neuronal tissue is not available from large numbers of individuals with ASDs, we used lymphoblasts, which although not expressing all relevant central nerve system (CNS) genes, do provide useful data for a significantly overlapping set of genes expressed in the CNS [22-24]. To assess which dysregulated genes could direct us to pathogenic mutations, we investigated expression variance in each subject and identified genes with significant deviations in expression in individuals' lymphoblasts. To explore the functional impact of CNVs in ASDs at a genome-wide scale, our interrogation utilized the overlap of structural variation data in a recently published manuscript [18] with transcriptional data in a subset of the same population. Our data support the notion that the intersection of gene expression with mutation data, such as CNV calls, or SNVs derived from exome sequence data, represents an efficacious approach for identifying new mutations and prioritizing autism susceptibility genes associated with chromosomal structural variation.

2.3 Neural-related pathways are altered in the lymphoblast cell lines (LCLs) of probands, but not siblings

Gene expression profiling was performed using LCLs from 439 individuals in 244 Simplex families, consisting of one proband and their unaffected sibling. Data collection occurred in two stages: first, we analyzed 386 individuals from 196 families, and then we prioritized 53 individuals with *de novo* CNVs from Sanders et al. 2011 (42 probands, 8 siblings and 3 mothers who carry 16p11.2 events) (Methods). Data was cleaned to control confounding factors such as batch, race and gender effects (Methods, Figure 2-S1). Four hundred and twelve microarrays, accounting for 221 probands, 188 siblings and 3 mothers, containing a total 11,150 expressed probes remained for analysis (360 from stage 1, 52 from stage 2) (Figure 2-1). Since the genetic contribution to ASDs includes rare mutations of intermediate to large effect size, differential gene expression is more likely to occur as a consequence within the CNV region in those specific cases, relative to other cases and controls. Based on this, a simple statistical framework was applied to identify "outlier genes" in individuals, defined as those whose expression is either two or three standard deviations (SDs) from the overall mean expression for that gene across the cohort (Methods) [23]. We initially took a strict, conservative approach by defining an outlier gene as having a ±3SD deviation (99.7% confidence interval) from the mean expression of that gene across all samples (Methods). Probands and siblings had a similar number of outlier genes per individual (8.1 vs. 10.2, $p = 0.60$ for down-regulated genes, and 16.6 vs. 17.6, $p = 0.76$ for up-regulated genes; un-paired t-test), similar to what is observed when all CNVs are treated as a homogeneous class of events [18]. Restricting analysis to brain-expressed genes[25] demonstrated a modest, but significant enrichment of outlier genes expressed in human fetal brain[25] in probands versus siblings (77% vs. 73%, Chi-square $p = 1.5E-03$).

However, no such enrichment was observed in human adult brain [26] (76% vs 76%, Chi-square p = 0.95; 81% vs. 81%, Chi-square p = 0.93). This agrees with most models of ASDs origin that posit a fetal or prenatal origin in most cases [1, 27-30].

We next used MetaCore by GeneGO and DAVID GO to explore whether the outlier genes had divergent biological functions, or were related to specific pathways (Methods). To control for effects related to transformation, we removed differentially expressed genes (DEX) known to be caused by EBV transformation [31]. Remarkably, in addition to several non-neural pathways, a significant enrichment of neural-related pathways in probands was observed. GeneGO (Figure 2-2) captured signal transduction, neuropeptide signaling pathways (p = 1.3E-06), development, neurogenesis, and synaptogenesis (p = 3.8E-03). DAVID GO (Table S2) also captured enrichment of similar CNS-related pathways, none of which were enriched in siblings (Table S2). This is not solely due to CNV (see below overlap analysis), as >90% of the genes in GeneGO neural pathways are outside CNV. Analyses of the stage-one samples in isolation revealed the same enrichment phenomena, a clear indication that sample selection bias had no impact on the results, confirming the robustness of the GO observations. Thus, despite profiling a peripheral non-neural tissue, we identified significant neural pathways previously related to ASDs [32], including some identified in a recent pathway analysis of SCC CNVs [33]. Our investigation also identified several previously known ASDs susceptibility genes to be classified as outliers, including *OXTR* [MIM 167055], *PCDH9* [MIM 603581], *CNTN4* [MIM 607280] and *UBE3A* [MIM 601623] (Table S3).

2.4 Copy number variation affects transcript levels in both probands and siblings

We next asked whether CNVs result in transcriptional changes and, conversely, can dysregulated genes aid in characterizing structural chromosomal variation? We compared CNVs identified in the Simons Simplex Cohort, which represents the most extensively validated cohort of CNV calls in ASDs [18]. In this study, three independent algorithms were used to identify a robust set of CNVs. Over 500 qPCR were done in random selected individuals representing 403 *de novo* and 120 transmitted events, providing a high confidence group of CNVs. These CNV data were integrated with microarray gene expression data, resulting in 330 samples characterized by both genotyping data and expression data (Figure 2-1).

To analyze the functional impact of CNVs on expression, linear regression was employed to interpret the relationship between copy number and the standard expression value (Z score) by taking a random sample, conditional on copy number status (Methods). We found a significant correlation between copy number and extreme expression ($\beta = 0.524$, p-value = 1.30E-05); that is genes in regions of duplication or deletion were far more likely to show extreme expression values compared with the genome background. We increased statistical power with a larger sample of outliers by assessing the percentage of CNVs bearing dysregulated genes in 330 samples, using a cutoff of $\pm$ 2SD (95% confidence interval) (Methods). By calculating the percentage of CNVs with dysregulated genes, 238 out of 2215 CNVs (10.7%) were found to contain at least one dysregulated gene, with a similar ratio between probands (11.5%) and siblings (9.7%) (Methods). Next, we calculated an Odds Ratio (OR) by comparing the average ratio of outlier genes among all expressed genes in the genome to the average ratio of outlier genes from expressed genes within CNVs of the 330 cases and siblings (Methods). We observed increased odds that outlier genes would be present in CNVs versus elsewhere in the genome (In probands: OR=4.3, Bonferroni p = 2.97E-102; In siblings: OR=2.6, Bonferroni p = 2.16E-21).

Moreover, in both probands and siblings, the direction of differential expression strongly correlated with the direction of copy number change. This presents further evidence that outliers are not random. The expected direction of dysregulation was observed in 92% of events (down-regulation in deletions and up-regulation in duplications) (Table S5).

Previous studies have suggested that CNVs can affect not only the transcriptional level of genes within them, but also genes in nearby regions up to 500kb on either side [20, 34]. We observed that 18.3% of CNVs have dysregulated genes within 500kb 5' or 3' in both probands and siblings, a significant enrichment compared to the rest of the genome (Bonferroni corrected p = 1.4E-07 for probands; Bonferroni corrected p = 1.5E-06 for siblings; Fisher's Exact test) (Methods). Interestingly, these changes were less likely to show the expected directionality shift compared with those inside the CNV. Only 43% changed in the direction of CNV dosage, indicating a more complex mechanism of regulation (Table S5). Furthermore, our linear regression model did not capture a significant relationship between copy number and the expression value of these nearby genes ($\beta = 0.029$, p-value = 0.234) (Methods), indicating that the relationship between *cis* gene expression and copy number is not linear.

2.5 Outlier genes are enriched in large rare de novo CNVs

Previous studies have shown that rare CNVs, especially rare *de novo* CNVs, are associated with autism [7, 18, 35]. Here, in general, the rarer the CNV, the higher the chance that it harbors an expression outlier (p = 4.9e-19; Methods). Based on the degree of CNV pathogenicity suggested by previous studies (rare *de novo*>rare transmitted>common), we next investigated whether there was an observable gradient in transcriptional change. Since rare *de novo* CNVs may be larger or contain more genes than rare transmitted CNVs and common

CNVs (Sanders et al. 2011; Levy et al. 2011), we used two methods to control for the potential confounding effect of CNV size (Methods). We calculated the proportion of dysregulated genes within a given CNV by dividing the number of dysregulated genes by the number of expressed genes within CNVs. This yielded a significantly higher proportion of dysregulated genes in rare *de novo* CNVs versus rare transmitted CNVs and common CNVs in probands ($p < 2.0E-16$, Kruskal-Wallis test) (Figure 2-3A). We then compared an arbitrary cohort of CNV matched for gene number in probands (16 rare *de novo* CNVs, 18 rare transmitted CNVs and 31 common CNVs). This comparison detected significantly more dysregulated genes in probands' rare *de novo* CNVs compared with the other two CNV classes ($p = 1.5E-05$, Kruskal-Wallis test) (Figure 2-3B). The results signify that, not only are genic segments enriched in rare *de novo* CNVs in probands, but these rare *de novo* CNVs are enriched in dysregulated genes even after correction for gene number within the CNV.

We next performed an independent assessment of predictions of CNV pathogenicity based on the gene expression data, employing a recently developed bioinformatics method for the assessment of haploinsufficiency (HI) [36]. To assess haploinsufficiency on a gene-by-gene level and correct for the potential confound of CNV size, we calculated HI probabilities (pHI), which estimate the likelihood of being haploinsufficient for each dysregulated gene involved in rare deletions in probands versus siblings [36]. We combined rare *de novo* CNVs with rare transmitted CNVs to increase statistical power, and focused our analysis on deletions because deletions, not duplications, are associated with HI. A significantly higher pHI probabllity was observed in probands than in siblings, consistent with increased pathogenicity of CNV in probands (Figure 2-3C). We also compared dysregulated to non-dysregulated genes within the same CNV. Importantly, the pHI of genes that are down-regulated in probands is significantly

greater than in those genes that do not change expression within rare deletions, showing a relationship between expression dysregulation and predicted pathogenicity (Bonferroni p = 4.4E-02, Mann Whitney U test) (Figure 2-3C). In contrast, down-regulated genes in siblings actually have a lower HI than non-differentially expressed genes within rare deletions, as would be predicted based on the presumed relative non-pathogenicity of these expression changes (Bonferroni p = 0.25, Mann Whitney U test; Figure 2-3C). As a control, we tested the gene pHI in common deletions in probands versus siblings as a control. No difference was observed as expected based on the presumed lack of pathogenicity of these events (Figure 2-3D).

2.6 Transcriptional data aids prioritization of small and non-recurrent CNV

We next reasoned that gene expression could help prioritize the potential pathogenicity of rare non-recurrent CNV, an important step, since even large *de novo* CNVs occur in 1-2 % of controls. To identify whether genes within a defined genomic region were significantly dyregulated, the percentage of dysregulated genes within each CNV was compared to random expectations on the genome background (Methods). Twenty-seven out of 40 rare *de novo* CNVs identified in probands have significantly more dysregulated genes in comparison with the genome background ($p < 0.05$, permutation test) (Table 1). Our analysis highlights a number of non-recurrent CNVs that have not previously been shown to be associated with ASDs, including deletions at 3q27, 3p13 and 3p26, and duplications at 2p15 and 13q14. To verify the altered expression detected by microarrays, we selected 12 genes in 8 corresponding non-recurrent CNVs to validate by qPCR (Methods). Nine of 12 (75%) genes were confirmed by qPCR, supporting the robustness of these analyses (Figure 2-S5A, B).

We next examined whether expression data could inform our analysis of small, potentially pathogenic CNVs. Figure 2-4 shows four examples of small rare CNVs observed in probands with a relatively high ratio of outlier genes. Both the copy number variations and expression alterations in these four examples were validated by PCR (Methods). One example involves a case with both a 16p11.2 deletion event (Figure 2-4D) and a small rare deletion at Xq28, affecting the expression level of gene *TMLHE* [MIM 300777][37]. Although not previously reported as associated with autism, *TMLHE* is an outlier gene in two affected siblings, but not in the unaffected sibling in one family from AGRE [38], suggesting *TMLHE* as a putative autism candidate gene. Since transcription levels are affected within these CNVs, the data presented in Figure 2-4 clearly warrant follow up in additional cohorts.

2.7 Transcriptional alterations in recurrent CNVs 16p11.2 duplications and deletions and 7q11.23 duplications

To determine whether gene expression analysis could help differentiate 16p11.2 deletions and duplications [MIM 611913] and identify dysregulated candidate genes, we conducted an examination of the effects of the 16p11.2 CNV on gene expression within the interval (Figure 2-5). First, we validated the dysregulation of 3 genes of interest from across the interval, *ALDOA* [MIM 103850] *MAPK3* [MIM 601795] and *CORO1A* [MIM 605000] in 5 cases of 16p11.2 deletion using qPCR to provide technical validation of a cross-section of the microarray results (Figure 2-S5C).

This examination generated several notable observations. Using a multivariate linear regression model, we observed a positive correlation between transcription level and 16p11.2 copy number, highlighting the group of genes most correlated with 16p11.2 dosage: *MAPK3* ($p <$

2E-16), *YPEL3* [MIM 609724] (p < 2E-16), *CORO1A* (p = 6E-15) and *KCTD13* [MIM 608947] (p = 1E-13) (Figure 2-5A) (Methods). Second, deletions had a larger effect on transcriptional level, and contained more genes with altered expression compared with duplications (Figure 2-5), in agreement with a recently published 16p11.2 mouse model [39]. We also studied the expression pattern in the 3 mothers carrying 16p11.2 events (2 duplications and one mosaic deletion) (Figure 2-S3). Consistent with their lack of clinical ASDs diagnosis, carriers look similar to controls, with few changes in gene expression relative to cases (p = 8.5E-05, Kruskal-Wallis test) (Figure 2-S3). This suggests that changes in expression levels may at least partially explain the molecular mechanism of incomplete penetrance of 16p11.2 events observed in parents and some offspring.

To determine *trans*-regulation of 16p11.2 events and explore whether 16p11.2 duplications and deletions affected similar or divergent biological pathways, we performed genome-wide differential expression (DEX) analysis, focusing on changes outside of the CNV region (Methods). Seventy DEX genes were observed in 16p11.2 deletion cases, and 135 genes DEX in 16p11.2 duplication cases (p<0.01). Strikingly, no overlap was evident in DEX genes between the two conditions. GO enrichment analysis revealed that in deletions, pathways containing DEX genes were enriched in neural-related ontologies, whereas no such enrichment was observed in duplications (Figure 2-7A,B)(Methods). This suggests that 16p11.2 deletions and duplications interrupt distinct molecular pathways, providing a functional basis for the different phenotypes observed in these two conditions.

Previous studies indicate that 16p11.2 deletion cases have significant macrocephaly while cases with duplications have microcephaly [40-47]. To explore if variance in gene expression in

the 16p11.2 region can be related to variance in head circumference, we applied a multivariate linear regression model (Methods). The most significantly associated genes within the CNV are *TAOK2* [MIM 613199], *CORO1A*, *KCTD13* and *QPRT* [MIM 606248] (Figure 2-S4). This is not a circular association reflecting the confounding of DE with gene dosage and HC, as several of the genes most associated with HC, including *TAOK2*, are not among the most DE genes in the region. Remarkably, the changes in these genes' expression accounted for more than 50% of the variance in HC. Given the sample size, this should be treated as a preliminary observation that warrants follow-up. But, it suggests that alterations in gene expression in peripheral blood can be related to disease-relevant central nervous system phenotypes.

Another recurrent event associated with autism is the 7q11.23 William-Beuren Syndrome [MIM 194050] region duplication [18, 19]. Similar to the 16p11.2 events, we observed multiple dysregulated genes in this region consistently changing in all three cases including *BCL7B* [MIM 605846], *EIF4H* [MIM 603431] and *LAT2* [MIM 605719] (Figure 2-6). Outside of the region, we observed 85 genes to be differentially expressed (DEX) genes in individuals with 7q11.23 duplications ($p < 0.01$). GO analysis identified several developmental pathways enriched in this gene list including forebrain development, determination of bilateral symmetry, and hippocampus development, providing another demonstration that CNS relevant pathways can be recovered from peripheral blood.

To explore whether genome-wide expression changes were sufficient to separate the different genotypes from each other and controls, we performed principle component analysis (PCA) for 16p11.2 and 7q11.23 cases, and compared them with 20 controls and 20 sporadic cases in different families (Methods). This analysis (Figure 2- 7D) suggests that 16p11.2

deletions and duplications may be distinct from each other, consistent with the analysis of gene expression within the CNV and the GO analysis of *trans* effects of each mutation on genome-wide expression. Furthermore, although the number is small, the 7q11.23 cases appear to cluster more with the 16p11.2 deletion cases, consistent with the observation that both disrupt CNS related gene ontology categories, whereas the 16p11.2 duplication cases do not. Interestingly, sporadic autism cases clustered with the controls (Figure 2-7D), consistent with the absence of significant shared genome-wide gene expression changes differentiating between randomly selected cases versus controls (Figure 2-S6). To further study the relationship between recurrent variants that are associated with autism, we compared the DEX genes from 16p11.2 deletion/duplications, 7q11.23 duplications and DEX genes identified previously in 15q11-13 duplications (15qdup) and fragile X mutation carriers with autism (FMR1-FM) [22]. Interestingly, *RIMS3* [MIM 611600] [48] is DEX in 16p11.2dup, 7q11.23dup, 15qdup and FMR1-FM, evidence for convergent dysregulation in multiple forms of ASDs.

2.8 Discussion

These results demonstrate the utility of gene expression analysis in evaluating the functional consequences of rare functional structural variations in a human neuropsychiatric disease, autism. Given the difficulty in interpreting whole genome level data in the context of rare variation, our data demonstrate a wealth of transcriptional alterations that are associated with structural variation. By integrating expression and genomic data, we show that the more pathogenic classes of CNVs are associated with increased odds of harboring transcriptional alterations either within or nearby the CNV, consistent with previous studies that demonstrate the impact of CNVs on genome-wide expression [20, 34]. We also found the CNVs only explain a

proportion of outlier genes; further studies are needed to identify potential mutations or epigenetic modifications within those outlier genes that may contribute to the expression alterations. Additionally, for recurrent CNVs known to be associated with autism, *cis* and *trans* expression analyses suggest distinct molecular mechanisms for 16p11.2 deletions and duplications.

It is well recognized that any method based on expression profiling would be optimal in the tissue most involved in the disorder, the central nervous system (CNS), preferably during early brain development when ASDs unfold. There is no doubt that this analysis has missed some disease relevant genes that are not expressed in lymphoblasts [49]. Unfortunately, post mortem brain tissue is only available from a very small number of inidividuals and tissue from early developmental stages is not available. Thus, the use of lymphoblasts has the advantage that these cells are widely available and permit a high-throughput, genome-wide analysis. Advances in induced pluripotent stem cell (IPSc) technology may eventually permit analyses of neuronal development in vitro [50]. Our successful use of expression data in lymphoblasts supports the use of such an approach in the future for determining the functional consequences of rare SNVs and CNVs. This is especially germane given the recent results of exome sequencing in ASDs [51-54]. These studies reveal an excess of rare *de novo* nonsense SNVs in ASDs, and to a lesser extent missense SNVs. Except in a few cases, the extent to which a given variant is functional is hard to predict. Thus, integration of gene expression with SNV data would likely be helpful.

Analysis of outliers was performed independently from analysis of CNV, with equivalent numbers of outliers in probands and siblings. However, GO analysis demonstrates that there is specific enrichment of CNS pathways in outliers detected in probands, supporting the hypothesis

that ASDs risk in simplex families is associated with the position and size of the CNV and not necessarily their overall burden. The GO pathways dysregulated specifically in probands also include known autism candidate genes, for example, the *oxytocin receptor (OXTR)* [55] *and ubiquitin protein ligase E3A (UBE3A)* [17]. We also observe enrichment of non-neural pathways in probands as well. Although some of these are not annotated as neural in GeneGO, they include signaling pathways that play crucial roles in neural development, such as BMP, TGF-□ or FGF signaling. Few pathways are enriched in siblings and all are non-neural, consistent with the interpretation that these likely represent noise, such as that introduced during the EBV transformation process [56] or based on the effect of variability in genetic background.

The pathogenic role of *de novo* CNVs in ASDs has been previously established [7, 17, 35, 57, 58]. Although it has been assumed that underlying changes in gene expression contribute to pathogenicity, previously this has not been demonstrated. If a CNV encompasses a region where biologically critical genes are more likely to be haploinsufficient, then it has a higher chance of having a functional impact on transcription [36]. We observed a higher pHI of genes in rare CNVs only in probands and not in sibling CNVs, providing independent validation of the outlier analysis by showing clear differences between the functional impact of these CNVs on expression. Previous studies have shown that many factors may contribute to pathogenicity of CNVs, including size, gene density, segmental duplication density, enrichment of certain functional pathways, and a higher than average expression correlation than the genome background [59, 60]. Here we show that analysis of peripheral blood gene expression can provide a useful and direct assessment of the functional consequences of chromosomal structural variation in a neuropsychiatric condition.

Assessment of the functional, potentially pathogenic impact of individual rare non-recurrent CNVs in disease remains an important challenge. Here, we use the outlier approach to identify novel candidate ASDs loci at 12p11.22, 15q23, 1p34.3, 3q27 and 3p26.2. For example, the 3p26.2 deletion in one proband contains 3 expressed genes: the *inositol 1,4,5-triphosphate receptor, type 1 (ITPR1* [MIM 147265]*), SET domain and mariner transposase fusion gene (SETMAR* [MIM 609834]*),* and *sulfatase modifying factor 1 (SUMF1* [MIM 607939]*),* all of which are down-regulated. Although none of these genes have been previously associated with autism, they are all functionally linked to the nervous system. Another example is a 100kb deletion at 3q27.2, which includes only one gene, the *SR-like splicing factor SFRS10/TRA2b (Htra2-beta1*; also known as *TRA2b* [MIM 602719]), which was down-regulated in the probands. *TRA2b* has recently been implicated in activity dependent regulation of RNA-splicing via interaction with *DARPP-32* [MIM 604399] [64]. This is particularly interesting given the involvement of another neuronal splicing factor regulated by neuronal activity [65], *Fox1/A2BP1* [MIM 605104] in ASDs [66], and previous data implicating activity dependent regulation of gene expression in ASDs [67].

This study also provides the exploration of 16p11.2 micro-deletions and micro-duplications. Previously, it was unclear which genes are dysregulated in or near the 16p11.2 region, or if there is a common expression signature shared by 16p11.2 cases. Our analysis shows a significant positive correlation of expression level and copy number as recently observed in mouse models [39] and highlights genes with the most consistent alterations across all 16p11.2 cases, including *the potassium channel tetramerisation domain containing 13 (KCTD13), Aldolase A, fructose-bisphosphate (ALDOA), and MYC-associated zinc finger protein (MAZ* [MIM 600999]). Potassium channel proteins like *KCNJ3* [MIM 601534],

KCNMA1 [MIM 600150] have been associated with neurodevelopmental abnormalities [68, 69]. *ALDOA* is involved in glycolysis and energy balance, which is important for synaptic metabolism and neurotransmitter release [70]. *MAZ* enhances the NMDA receptor subunit type 1 activity during neuronal differentiation [71]. This study provides a source for candidate gene prioritization for future functional and mutational analyses. Although our analysis of differential expression highlighted different molecular pathways disrupted in 16p duplications and deletion, one needs to also consider that we could be missing some common pathways that are only expressed in brain. In this regard, it is notable that 7q11.23 cases cluster with the 16p del cases in terms of global gene expression changes in lymphoblasts. Within the 7q11.23 duplications, we found that *STX1A* [MIM 186590], *CLIP2* [MIM 603432] and *LIMK1* [MIM 601329] are up-regulated, but, we do not see alterations in *GTF2I* [MIM 601679] and *CYLN2* [MIM 603432], which were previously shown to be dysregulated in 7q11.23 duplications by qRT-PCR [72]. This may be due to differences in techniques or the phenotypes assayed, and further studies in larger samples will permit more precise expression-phenotype correlations. We hypothesize that the observed expression differences are likely related to the phenotypic differences observed in reciprocal 7q11.23 events and provide a starting point for connecting specific genes to phenotypes in subjects with 7q11.23 CNV.

We also provide the molecular correlates of a clinical phenotype, head circumference (HC), in ASDs [40-45]. This is especially interesting, since 16p11.2 deletions are highly penetrant for ASDs (associated with macrocephaly), while 16p11.2 duplication cases (associated with microcephaly) are less penetrant for ASDs. Here in 16p11.2 events, we demonstrate a significant correlation between HC and expression of several genes within the CNV including *TAOK2*, which showed the largest such correlation. *TAOK2* interacts with the JNK mitogen-

activated protein kinase pathway [73], which has been shown to control survival, proliferation and differentiation of cells composing the central and peripheral nervous system [74], providing a biologically plausible link between this gene and a brain growth phenotype that can be tested in future studies in neural tissues and model organisms.

In summary, we present the largest genome-wide expression profiling study in ASDs and integrate this transcriptional data with genomic data. Each of these data sets, gene expression and CNV, is complementary and either alone is less powerful than the combination of the two. These data highlight the utility of this approach for prioritization of mutations and specific genes for further downstream functional or mutational analysis an approach that should have widespread utility given the proliferation of genome sequencing and analysis of structural variation. This is especially true for rare, non-recurrent variants for which standard statistical tests of association are underpowered. We show that the intersection of such events with expression permits a statistical analysis of individual events, facilitating prioritization of individual rare CNV. These results elucidate the genome-wide functional impact of CNVs, and may help to explain complex phenotypes related to brain growth, such as head circumference, all of which will help to link genotype to phenotype in complex neuropsychiatric disorders, such as autism.

2. 9 Methods

Individuals and lymphoblast cell lines (LCLs) analyzed in this study

We analyzed individuals from Simon Simplex Collection (SSC) in two stages. In the first stage, we collected 386 individuals from 196 families (190 matched sib pairs plus 5 siblings and 1 proband). In the second stage, we prioritized 53 samples with *de novo* CNVs (42 probands, 8 siblings and 3 mothers who carry 16p11.2 events) [18]. Phenotype information can be found at

Simons Foundation Autism Research Initiative (SFARI) database and inclusion information is shown in supplemental Table 2-S1. This study was approved by the Institutional Review Boards at all participating institutions, including UCLA and Yale University. The lymphoblast cell lines (LCLs) of the subjects were grown in RPMI 1640 medium with 2 mM L-glutamine and 25 mM HEPES (Invitrogen, Carlsbad, CA, USA), 10% fetal bovine serum, and 1 × Antibiotic-Antimycotic solution (Invitrogen) at 37°C in a humidified 5% CO2 chamber. Cells were grown to a density of 6×10^5/ml. Special attention was given to maintain all the cell lines in the same conditions to minimize environmental variation.

Microarray experiments

A total of 9×10^6 of lymphoblasts were seeded in a T-75 flask in 30 ml of fresh medium. After 24 h, total RNA was extracted from the cells using an RNeasy Mini Kit with DNase treatment (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. RNA quantity and quality were measured by ND-100 (Nanodrop, Wilmington, DE, USA) and 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), respectively. mRNA was hybridized on the Illumina Whole Human Genome Array Human REF-8 version 3.0 according to the manufacturer’s protocol.

Sample quality control

GenomeStudio was used to convert image to numerical data as per our typical protocols [26, 75, 76]. Four hundred and thirty-nine samples (chips) were cross-correlated using expression levels for all probe sets. These inter-array correlations (IACs) [77] were averaged for each array and compared to the resulting distribution of IACs for the dataset [75]. Samples with an average IAC < 2.0 standard deviation below the mean IAC for the dataset were removed.

Those remaining were clustered using average linkage and 1-IAC as a distance metric to identify the 27 samples with poor quality (6%). Following sample removal, quantile normalization [77] was performed in R. To eliminate batch effects, additional normalization was performed using the R package *ComBat* [78] using the default parameters. *ComBat* successfully eliminated batch effects as evidenced by hierarchical clustering and significant improvement of mean IAC (Figure 2-S1). After data pre-processing, 412 microarrays remained for follow up analysis, 333 of which had genomic array data and expression data. Three samples (of 333) are mothers of probands. We used the remaining 330 samples for all of the analyses except the 16p11.2 event analysis. Among 412 samples, we have 168 pairs of individuals (each pair is from the same family). Ninety-eight out of 168 pairs are gender-matched. To control for potential confounding factors, linear regression was used to remove gender and age effects. We checked the average CNV number per individual, and with the exception of African Americans (60 CNVs per individual), there was no effect of ancestry on CNV frequency (35 CNVs per individual). Since African American samples only comprise 3% of our cohort, we retained them to have more statistical power and a better overlap between microarray and genetic data.

Probe quality control

We only used probes with evidence of robust expression (detection p-value <= 0.05 in at least 50% samples). By filtering out non-expressed probes, 11150 probes (corresponding to 9524 genes) remained for analysis. To study the functional impact of CNVs on expression, we filtered the 9524 genes by restriction to genes that had 30 or more markers (SNPs and monomorphic probes) covering them. For any of these "high-quality" genes that had multiple gene expression

reads, we took the average expression for each gene. This resulted in a set of 8006 unique genes with gene expression values.

Outlier gene analysis

For outlier gene analysis, we calculated the Z statistic for each gene using the *scale* function in R. We calculated the mean and standard deviation for each expressed gene in cases and controls separately. A cutoff (2SD/ 3SD) was selected to define whether a gene is an outlier gene in probands or siblings. For outlier analysis performed not in conjunction with CNV data, we used a more stringent cutoff (3SD). Subsequently, for the comparison of the overlap between CNV and transcriptional alterations, we used 2SD as a cutoff. These different thresholds were used for two major reasons. When analyzing expression changes in isolation, we used the more conservative 3SD cutoff to increase stringency. When we integrate genotyping and expression data, we relaxed the statistical threshold to 2 SD so as to increase power by increasing the number of potentially dysregulated, outlier genes. We use the term "outlier genes" unless we have evidence that the gene is also affected at the genetic level by a CNV. In that case, we call the gene dysregulated to reflect the concept that it is contained within a structural variantion and shows significant alteration in gene expression.

Odds ratio analysis

An odds ratio (OR) was calculated with the *epitools* library in R using the Wald method, an unconditional likelihood estimation method. For calculating the odds ratio of dysregulated genes within CNVs (or near CNVs) versus the genome background, the genes not within CNVs in a certain individual are used as the control group. The two-by-two contingency matrix was made for calculating the odds ratio: the two columns are: 1) sum of the gene number within

CNVs for all probands or all siblings; 2) sum of the gene number in genome elsewhere but not CNVs for all probands or all siblings; the two rows are: 1) sum of the dysregulated genes; 2) sum of the normally expressed genes. Bonferroni correction [79] was used to correct for multiple testing of the OR analysis.

Integrating expression data with CNV data

The CNV list was taken from Sanders et al. (2011) Table 2-S4 and Table 2-S8. The criteria for sub grouping CNVs were as described [18] and *de novo* CNVs are determined by the CNV calling algorithm descried therein. Rare CNVs are defined as CNVs with less than 50% overlap with those in the Database of Genomic Variants (DGV) [80].

Multivariate linear regression analysis of expression and copy number

For analysis of the relationship between gene expression (genes within CNVs and genes nearby [500kb]) and copy number, we applied a GEE [81] model using the *geeglm* function in R. We regressed out the effects of age and sex from the standardized gene expression data using a linear model. We then used the residuals obtained from the linear model as the continuous, predicted variable for our expression value analysis. Next, we: 1) obtained a biased sample of 100 gene expression residuals in which the copy number variants were equally represented (50 were duplications and 50 were deletions); 2) matched each subject with CNVs to a subject with no CNVs, matching by gene; 3) fit a GEE linear model (which in all instances that follow is used to account for any unknown, within-individual correlation amongst gene expressions) between the gene expression residuals and the two predictor variables, proband status and copy number ; 4) repeated steps 1-3 for 500 runs to obtain a distribution of coefficients and p-values for each predictor. To measure the effects of rareness and size of CNVs on outlier status in gene

expression, we defined as outliers those standardized gene expression scores with absolute value ≥2, which we then encoded as a binary variable. We then used a GEE model with a binomial link for a logistic regression to accommodate the binary nature of the outlier status variable. Rareness was defined as in Sanders et al. (2011). The contrast is with genes falling in CNVs that do not meet these criteria. The estimated size of the CNV was entered as a continuous variable. To study the *cis*-regulation of CNVs, we performed a similar analysis by using genes 500kb upstream and downstream of CNVs. The predictors are the same as for genes within CNVs.

Detecting outlier genes within CNVs

For 330 individuals, a total of 12,068 CNVs were identified by Sanders et al. (2011). Two thousand, two hundred and fifteen out of 12,068 CNVs contain at least one gene expressed in LCLs. This list of 2215 CNVs was used to study the functional impact of copy number on transcription. For expressed genes within these CNVs, we identified outliers, as genes that are ±2SD from the mean expression in all samples. With this method, 10.7% (238 out of 2215) CNVs contain at least one outlier gene.

Enrichment analysis of outlier genes in rare de novo CNV

To compare the dysregulated genes residing in rare *de novo* CNVs versus rare transmitted CNVs and common CNVs, we analyzed all CNVs containing as least one gene expressed in the LCLs, which led to 38 rare *de novo* CNVs from 37 probands, 419 rare transmitted CNVs from 170 probands and 353 common CNVs from 184 probands. We used two methods to control for the gene number in each type of CNV: 1) we compared the ratio of dysregulated genes (number of dysregulated genes divided by the number of genes expressed) between these three groups. The Kruskal-Wallis test, a general form of multi-group non-parametric test was used; 2) we

matched CNVs for gene number content: 16 rare *de novo* CNVs, 18 rare transmitted and 31 common CNVs matched for gene number. The Bonferroni correction was used to correct for multiple testing.

To compare the dysregulated genes residing in rare *de novo* CNVs between probands and siblings, we compared 38 rare *de novo* CNVs found in 37 probands and 3 rare *de novo* CNVs found in 3 siblings. We calculated the ratio of dysregulated genes within each CNV and compared the rank difference of the ratio by the Mann Whitney U test.

Permutation test of outlier genes in the whole genome

To compute the empirical p-value of the significance of the number of dysregulated genes within each rare *de novo* CNV, a permutation test was applied. We randomly picked one individual, one chromosome region and selected the adjacent genes to match the number of expressed genes in each rare *de novo* CNV, and then calculated the number of dysregulated genes in this randomly picked region. A hundred thousand permutations were performed for each rare *de novo* CNV.

Multivariate linear regression analysis of expression and copy number at 16p11.2 and 7q11.23

The *geeglm* function in R was used to fit a linear regression model between the copy number and expressed genes in 16p11.2 and 7q11.23 respectively: expression value ~ copy number + age + gender. General estimating equations were used to correct for family structure. For 16p11.2 events, we fitted the model by treating the copy number both as a quantitative variable and a factor variable; both methods provided similar results. The p-value from the

quantitative variable approach is reported in Figure 2-5. For 7q11.23 duplications, we fitted the model by treating the copy number as a quantitative variable.

Genome-wide differential expression (DE) analysis

The *Limma* [82] package in R was applied for standard differential expression analysis in the cases of 16p11.2 deletions and duplications, and 7q11.23 duplications. Controls were chosen from the pool of all controls with a matched gender ratio to specific cases. In total, seven 16p11.2 deletions (6 males and 1 female), and 120 controls (100 males and 20 females) were used for DE analysis, while six 16p11.2 duplications (5 males and 1 female), and 117 controls (100 males and 17 females) were used. In total, three 7q11.23 duplications (2 females and 1 male) and 142 controls (46 males and 96 females) were used.

Multivariate linear regression analysis of phenotype

The *lm* function in R was used to fit a linear regression model between the expressed genes in 16p11.2 region and head circumference phenotype, adjusted for age and gender [83]. Age, gender and expression value were used together as predictors and the expression value of each gene was normalized by the *scale* function in R program before fitting the linear model.

Principle component analysis (PCA)

The *prcomp* function in R was used to calculate the first two principle components. Seven 16p11.2 deletions, six 16p11.2 duplications and three 7q11.23 cases were used. The 20 sporadic cases and 20 controls were selected randomly in our samples. Samples were clustered by the differentially expressed genes ($p<0.01$) identified in 16p11.2 duplications and deletions, and 7q11.23 duplications.

Pathway analysis

DAVID GO database and MetaCore by GeneGO (Thompson Reuters) were used for pathway analyses. For both analyses, the background was set to the total list of genes expressed in our dataset. The statistical significance threshold level for all GO enrichment analyses was $p<0.05$.

qRT-PCR validation for copy number variation

Quantitative polymerase chain reaction (qPCR) was used to confirm the presence or absence of predicted CNVs in lymphoblast DNA. Two control primers were designed within 'house-keeping genes': *RPP21* [MIM 612524] and *ZNF80* [MIM 194553], genes in which no CNVs were reported in the Database of Genomic Variants (DGV). 1ul of DNA with the concentration of 0.2ug/ul was used for qPCR reaction by 2X MyTaq Red Mix (Bioline). A pooled sample from 96 normal siblings of Simons Simplex Collection was used as the control sample. Quantitative RT-PCR was performed on the ABI Prism 7900 (Applied Biosystems, Foster City, CA, USA) using Platinum SYBR Green qPCR SuperMix UDG with ROX (Invitrogen). Thermal cycling consisted of an initial step at 50°C for 2 min followed by another step at 95°C for 2 min and 45 cycles of 95°C for 15 s and 60°C for 30s. The primers used for qRT-PCR are listed in Table 2-S6. The following formula was used to estimate copy number [18]:

Estimated copy number = $2(-\Delta\Delta CT)$

Where:

- $\Delta\Delta CT$ = (CT Region:Sample – CT Ref:Sample) – (CT Region:Control – CT Ref:Control)

- CT Region:Sample = mean CT values for the region of interest and sample of interest (e.g. ExpPrimer1 and ExpSample1)
- CT Ref:Sample = mean CT values for the reference region and sample of interest (e.g. RNase P Primer and ExpSample1)
- CT Region:Control = mean CT values for the region of interest and the control sample (e.g. ExpPrimer1 and CT_pooled_control)
- CT Ref:Control = mean CT values for the reference region and the control sample (e.g. RNase P Primer and CT_pooled_control)

qRT-PCR validation for expression alteration

Five hundred nanograms of total RNA was used to make cDNA by SuperScript III First-Strand Synthesis SuperMix (Invitrogen) and random hexamers (Invitrogen). The qRT-PCR was performed on an ABI Prism 7900 (Applied Biosystems, Foster City, CA, USA) using Platinum SYBR Green qPCR SuperMix UDG with ROX (Invitrogen). Thermal cycling consisted of an initial step at 50°C for 2 min followed by another step at 95°C for 2 min and 45 cycles of 95°C for 15 s and 60°C for 30s. Data were normalized by the quantity of *glyceraldehyde-3-phosphate dehydrogenase (GAPDH* [MIM 138400]*).* The gene Ct value of targeted probands was compared to the average Ct values from 5 unaffected siblings, matched for gender and age. The primers used are listed in Table 2-S6.

The ΔCt, ΔΔCt and fold change of the tested gene were calculated by following formula:

ΔCt for each sample:

ΔCt= Ct (tested gene) – Ct (*GAPDH*)

ΔΔCt for each sample:

ΔΔCt= ΔCt of tested gene in the targeted proband – average ΔCt of test gene in siblings

Fold change for up-regulated genes:

Fold change = 2 ^ (-ΔΔCt)

Fold change for down-regulated genes:

Fold change = -2 ^ ΔΔCt

Figure 2-1

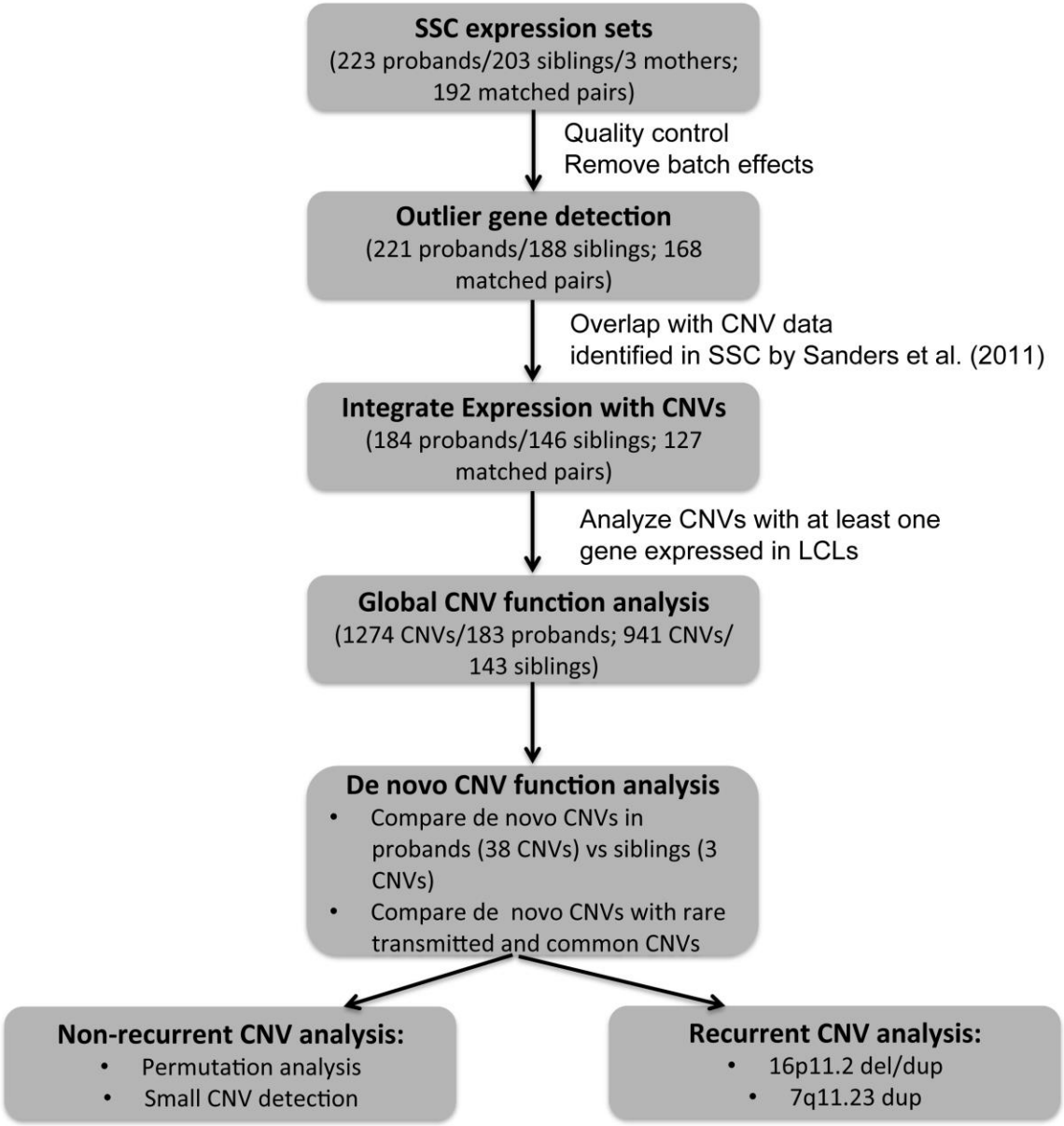

SSC expression sets
(223 probands/203 siblings/3 mothers; 192 matched pairs)
Quality control
Remove batch effects
Outlier gene detection
(221 probands/188 siblings; 168 matched pairs)
Overlap with CNV data
identified in SSC by Sanders et al. (2011)
Integrate Expression with CNVs
(184 probands/146 siblings; 127 matched pairs)
Analyze CNVs with at least one gene expressed in LCLs
Global CNV function analysis
(1274 CNVs/183 probands; 941 CNVs/ 143 siblings)
De novo CNV function analysis
• Compare de novo CNVs in probands (38 CNVs) vs siblings (3 CNVs)
• Compare de novo CNVs with rare transmitted and common CNVs
Non-recurrent CNV analysis:
• Permutation analysis
• Small CNV detection
Recurrent CNV analysis:
• 16p11.2 del/dup
• 7q11.23 dup

Figure 2-1. Flow chart of expression data analysis and integration with CNV data in the Simons Simplex Collection (SSC). Quality control was done before any data analysis (Figure S1, Methods). The numbers of individuals and CNVs used for down stream analysis is shown in the flow chart.

Figure 2-2

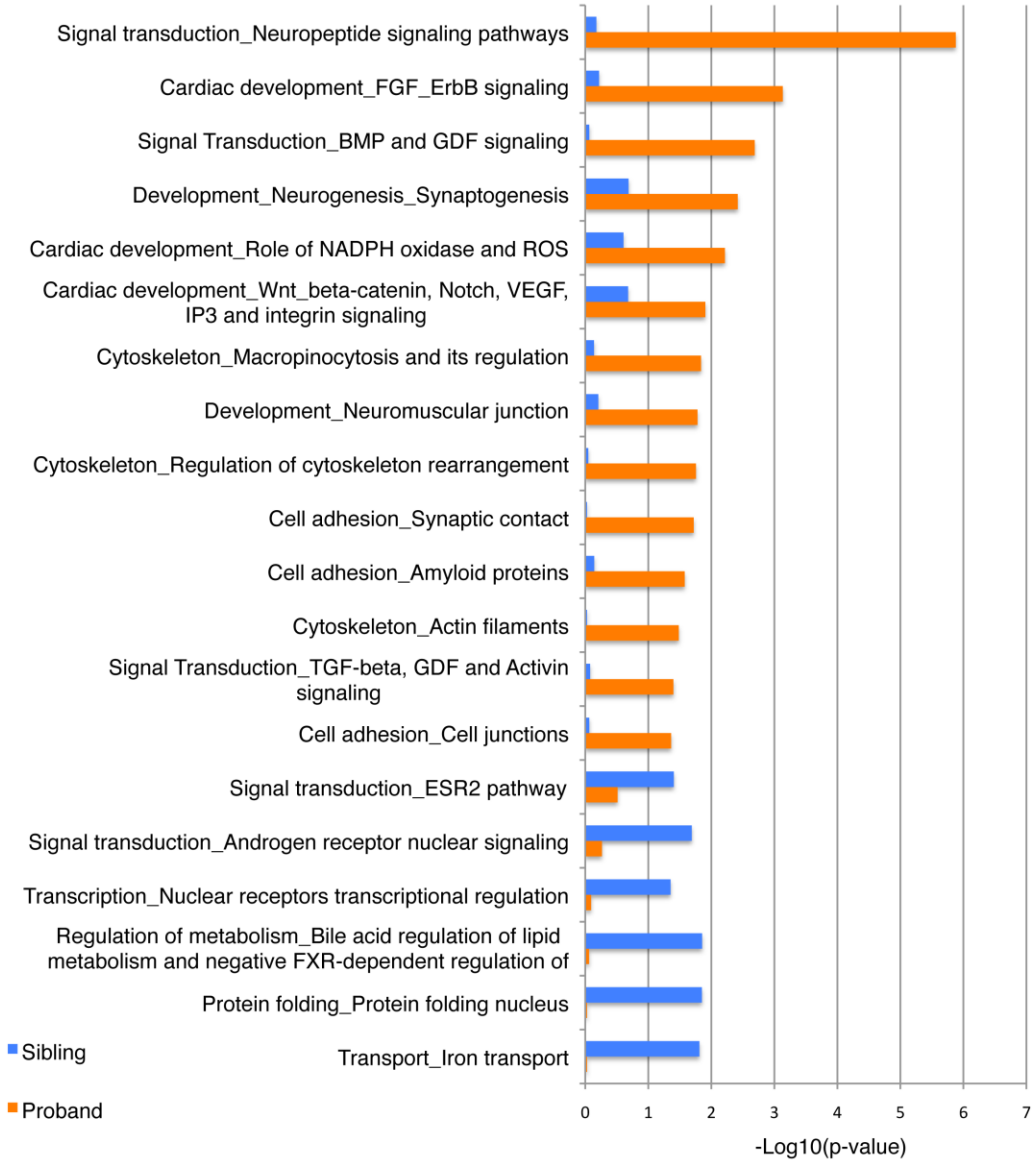

Signal transduction_Neuropeptide signaling pathways
Cardiac development_FGF_ErbB signaling
Signal Transduction_BMP and GDF signaling
Development_Neurogenesis_Synaptogenesis
Cardiac development_Role of NADPH oxidase and ROS
Cardiac development_Wnt_beta-catenin, Notch, VEGF, IP3 and integrin signaling
Cytoskeleton_Macropinocytosis and its regulation
Development_Neuromuscular junction
Cytoskeleton_Regulation of cytoskeleton rearrangement
Cell adhesion_Synaptic contact
Cell adhesion_Amyloid proteins
Cytoskeleton_Actin filaments
Signal Transduction_TGF-beta, GDF and Activin signaling
Cell adhesion_Cell junctions
Signal transduction_ESR2 pathway
Signal transduction_Androgen receptor nuclear signaling
Transcription_Nuclear receptors transcriptional regulation
Regulation of metabolism_Bile acid regulation of lipid metabolism and negative FXR-dependent regulation of
Protein folding_Protein folding nucleus
Transport_Iron transport
Sibling
Proband
0
1
2
3
4
5
6
7
-Log10(p-value)

Figure 2-2. Neural-related pathways are enriched in probands versus siblings. GeneGO was used to run the ontology analysis for outlier genes identified in probands and siblings respectively. The –log10 p-value is shown with the pathways that were significant (with uncorrected p-value <0.05) in either probands or siblings.

Figure 2-3

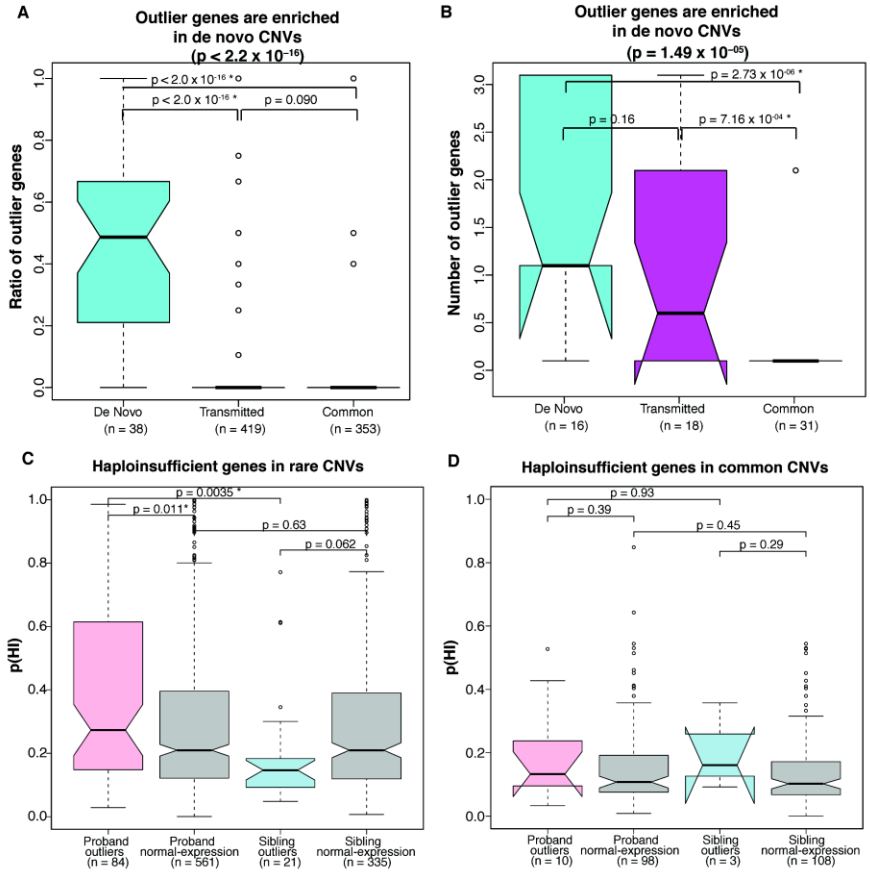

A
Outlier genes are enriched in de novo CNVs (p < 2.2 x 10⁻¹⁶)
Ratio of outlier genes
p < 2.0 x 10⁻¹⁶ *
p < 2.0 x 10⁻¹⁶ *
p = 0.090
De Novo (n = 38)
Transmitted (n = 419)
Common (n = 353)
B
Outlier genes are enriched in de novo CNVs (p = 1.49 x 10⁻⁰⁵)
Number of outlier genes
p = 2.73 x 10⁻⁰⁶ *
p = 0.16
p = 7.16 x 10⁻⁰⁴ *
De Novo (n = 16)
Transmitted (n = 18)
Common (n = 31)
C
Haploinsufficient genes in rare CNVs
p(HI)
p = 0.0035 *
p = 0.011*
p = 0.63
p = 0.062
Proband outliers (n = 84)
Proband normal-expression (n = 561)
Sibling outliers (n = 21)
Sibling normal-expression (n = 335)
D
Haploinsufficient genes in common CNVs
p(HI)
p = 0.93
p = 0.39
p = 0.45
p = 0.29
Proband outliers (n = 10)
Proband normal-expression (n = 98)
Sibling outliers (n = 3)
Sibling normal-expression (n = 108)

Figure 2-3. Outlier genes are enriched in rare *de novo* CNVs in probands. A) The boxplot depicts the ratio of dysregulated genes (number of dysregulated genes within CNV versus the total number of genes within that CNV) in the three types of CNVs (rare *de novo* CNVs, rare transmitted CNVs and common CNVs) respectively. The Krusakal-Wallis test p-value is shown. **B)** The boxplot shows the number of dysregulated genes in three types of CNVs with matched gene number. **C)** The boxplot of HI scores for down-regulated genes (2SD) in rare deletions in probands and siblings versus normal expressed genes within CNVs. The HI score of dysregulated genes in rare deletions in probands is significantly higher compared to the normal expressed genes, while the HI score of dysregulated genes in rare deletions in siblings is significantly lower compared to normal expressed genes (Mann Whitney U test). **D)** The boxplot of HI scores for down-regulated genes (2SD) in common deletions in probands and siblings versus normal expressed genes within CNVs. The Mann Whitney U test p-value is shown for each pair-wise comparison. A star indicates a statistically significant p-value after Bonferroni correction ($p < 0.017$ in A and B, $p< 0.0125$ in C and D). Error bars for these four panels are defined as the 1.5 times the interquartile range.

Figure 2-4

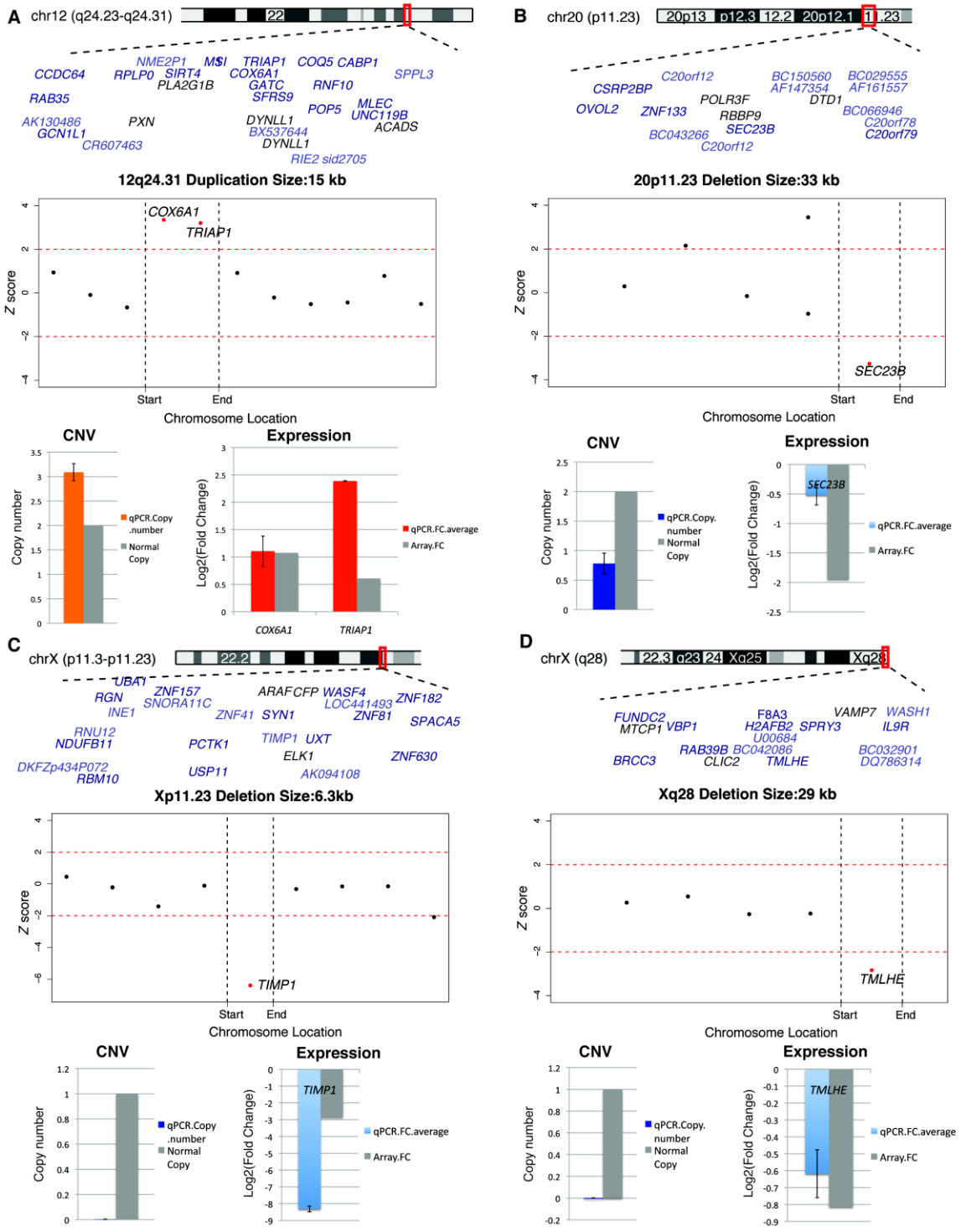

A
chr12 (q24.23-q24.31)
NME2P1 MSI TRIAP1 COQ5 CABP1
CCDC64 RPLP0 SIRT4 COX6A1 SPPL3
PLA2G1B GATC RNF10
RAB35 SFRS9
POP5 MLEC
AK130486 PXN DYNLL1 UNC119B
GCN1L1 BX537644 ACADS
CR607463 DYNLL1
RIE2 sid2705
12q24.31 Duplication Size:15 kb
COX6A1
TRIAP1
Z score
Start
End
Chromosome Location
CNV
Copy number
qPCR.Copy .number
Normal Copy
Expression
Log2(Fold Change)
qPCR.FC.average
Array.FC
COX6A1
TRIAP1
B
chr20 (p11.23)
20p13 p12.3 12.2 20p12.1 11.23
C20orf12 BC150560 BC029555
CSRP2BP AF147354 AF161557
POLR3F DTD1
OVOL2 ZNF133 RBBP9 BC066946
BC043266 SEC23B C20orf78
C20orf12 C20orf79
20p11.23 Deletion Size:33 kb
SEC23B
Z score
Start
End
Chromosome Location
CNV
Copy number
qPCR.Copy. number
Normal Copy
Expression
Log2(Fold Change)
SEC23B
qPCR.FC.average
Array.FC
C
chrX (p11.3-p11.23)
22.2
UBA1
ZNF157 ARAF CFP WASF4 ZNF182
RGN SNORA11C LOC441493
INE1 ZNF41 SYN1 ZNF81 SPACA5
RNU12
NDUFB11 PCTK1 TIMP1 UXT
ELK1 ZNF630
DKFZp434P072 USP11 AK094108
RBM10
Xp11.23 Deletion Size:6.3kb
TIMP1
Z score
Start
End
Chromosome Location
CNV
Copy number
qPCR.Copy .number
Normal Copy
Expression
Log2(Fold Change)
TIMP1
qPCR.FC.average
Array.FC
D
chrX (q28)
22.3 q23 24 Xq25 Xq28
FUNDC2 F8A3 VAMP7 WASH1
MTCP1 VBP1 H2AFB2 SPRY3 IL9R
U00684
RAB39B BC042086 BC032901
BRCC3 CLIC2 TMLHE DQ786314
Xq28 Deletion Size:29 kb
TMLHE
Z score
Start
End
Chromosome Location
CNV
Copy number
qPCR.Copy. number
Normal Copy
Expression
Log2(Fold Change)
TMLHE
qPCR.FC.average
Array.FC

Figure 2-4. Outlier genes highlight small, but likely functional CNV. A) Depicts a small duplication with the high ratio of dysregulated genes. **B, C, D)** Depicts small deletions with the high ratio of dysregulated genes. The Z scores of all expressed genes within the CNV interval and within 500kb upstream and downstream are shown. Outlier genes (2SD, red) within the CNV are shown. A barplot was used to show the qPCR validation for both copy number change and the expression alteration. Error bars show the standard deviation of three replicates of qPCR experiments.

Figure 2-5

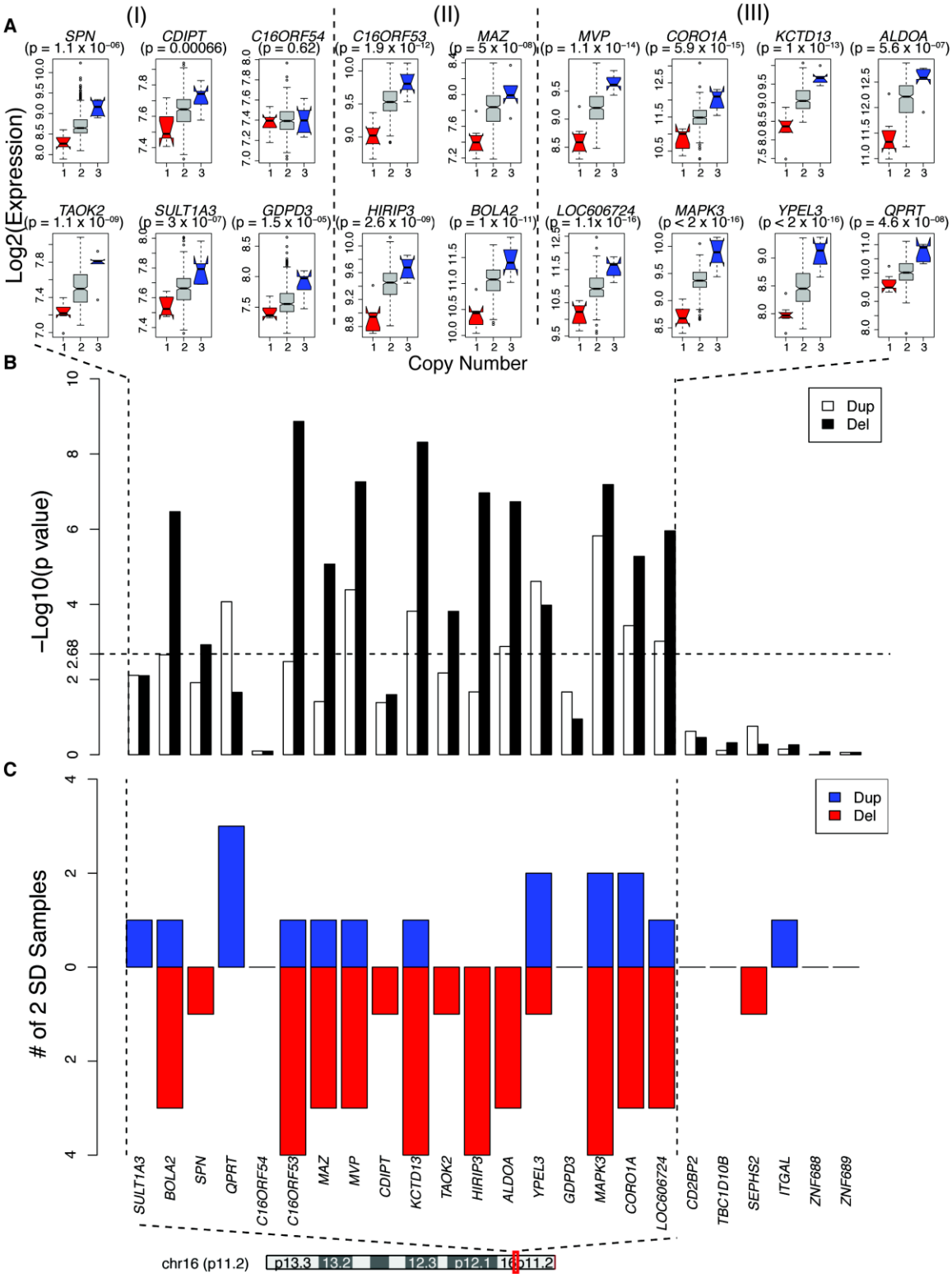

A
(I)
(II)
(III)
SPN (p = 1.1 x 10−06)
CDIPT (p = 0.00066)
C16ORF54 (p = 0.62)
C16ORF53 (p = 1.9 x 10−12)
MAZ (p = 5 x 10−08)
MVP (p = 1.1 x 10−14)
CORO1A (p = 5.9 x 10−15)
KCTD13 (p = 1 x 10−13)
ALDOA (p = 5.6 x 10−07)
TAOK2 (p = 1.1 x 10−09)
SULT1A3 (p = 3 x 10−07)
GDPD3 (p = 1.5 x 10−05)
HIRIP3 (p = 2.6 x 10−09)
BOLA2 (p = 1 x 10−11)
LOC606724 (p = 1.1x 10−16)
MAPK3 (p < 2 x 10−16)
YPEL3 (p < 2 x 10−16)
QPRT (p = 4.6 x 10−08)
Log2(Expression)
Copy Number
B
Dup
Del
−Log10(p value)
2.68
C
Dup
Del
of 2 SD Samples
SULT1A3
BOLA2
SPN
QPRT
C16ORF54
C16ORF53
MAZ
MVP
CDIPT
KCTD13
TAOK2
HIRIP3
ALDOA
YPEL3
GDPD3
MAPK3
CORO1A
LOC606724
CD2BP2
TBC1D10B
SEPHS2
ITGAL
ZNF688
ZNF689
chr16 (p11.2)
p13.3
13.2
12.3
p12.1
11.2

Figure 2-5. Gene expression in the 16p11.2 duplication and deletion interval. A) For each of the expressed genes within the 16p11.2 interval, the Log2 expression level is shown for deletions (red), duplications (blue) and controls (grey). The p-value is calculated using a multivariable linear regression model with 16p11.2 cases and 398 controls without a known 16p11.2 event (Methods). Twelve out of 19 expressed genes in deletions have at least a 1.3-fold change, while 8 out of 19 genes in duplications show a 1.3-fold or greater change. Group I represents genes that don't reach 1.3 fold change in either duplications or deletions; Group II represents genes that have larger than 1.3 fold change in deletions only; Group III represents genes that have larger than 1.3 fold change in both duplications and deletions (dash line separated). Error bars for these four panels are defined as the 1.5 times the interquartile range. **B)** The –Log of the p-value (T-test) for duplications and deletions respectively are shown on the y-axis for each gene within the 16p11.2 region and within 500kb upstream and downstream. The dashed vertical line shows the p-value threshold after Bonferroni correction (corrected for 24 genes, p-value < 2.1E-03). **C)** Genes showing expression deviating by at least two standard deviations from the mean across 13 samples (7 deletions, 6 duplications) with 16p11.2 CNVs.

Figure 2-6

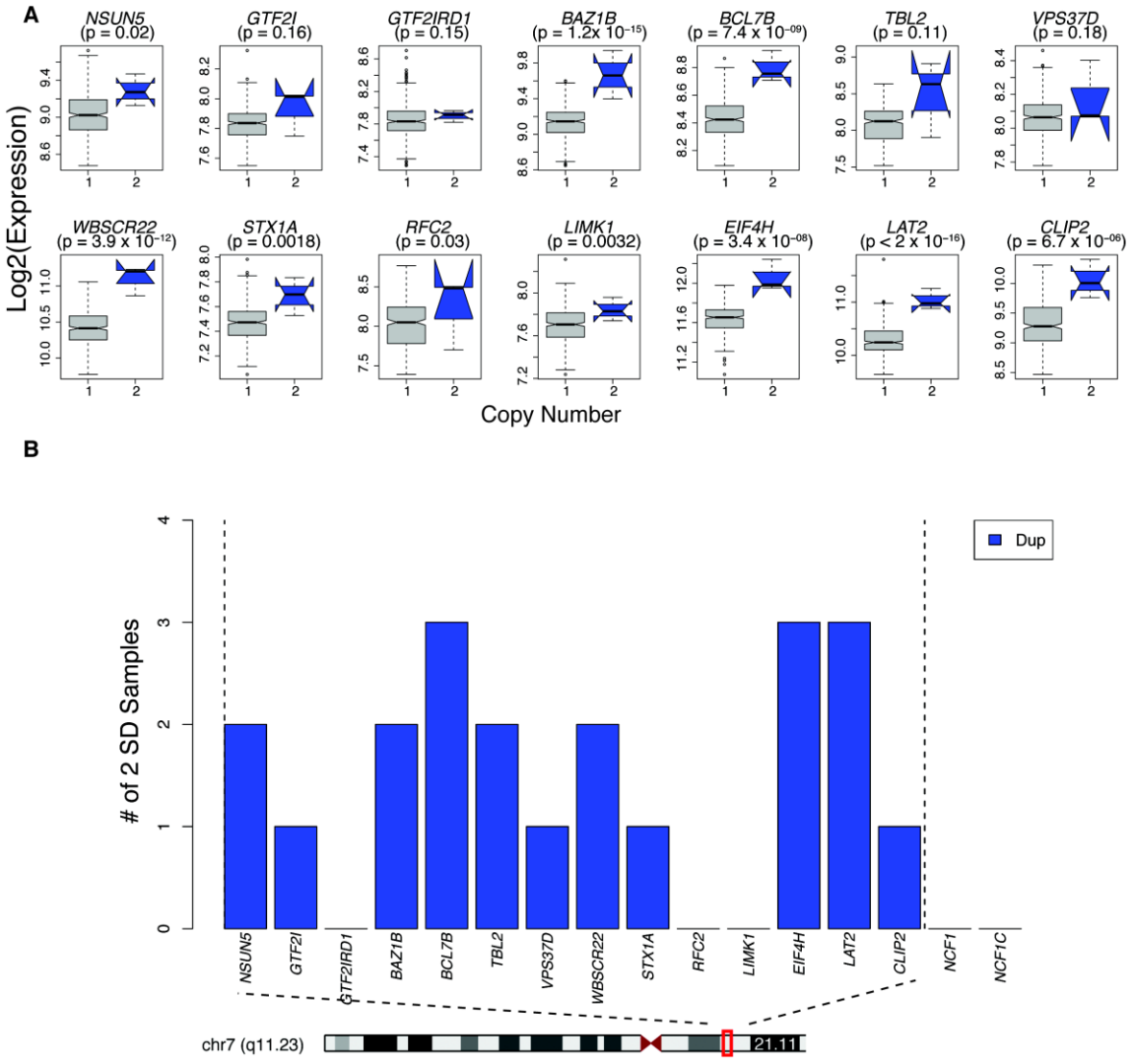

A
NSUN5 (p = 0.02)
GTF2I (p = 0.16)
GTF2IRD1 (p = 0.15)
BAZ1B (p = 1.2x 10⁻¹⁵)
BCL7B (p = 7.4 x 10⁻⁰⁹)
TBL2 (p = 0.11)
VPS37D (p = 0.18)
WBSCR22 (p = 3.9 x 10⁻¹²)
STX1A (p = 0.0018)
RFC2 (p = 0.03)
LIMK1 (p = 0.0032)
EIF4H (p = 3.4 x 10⁻⁰⁸)
LAT2 (p < 2 x 10⁻¹⁶)
CLIP2 (p = 6.7 x 10⁻⁰⁶)
Log2(Expression)
Copy Number
B
Dup
of 2 SD Samples
NSUN5
GTF2I
GTF2IRD1
BAZ1B
BCL7B
TBL2
VPS37D
WBSCR22
STX1A
RFC2
LIMK1
EIF4H
LAT2
CLIP2
NCF1
NCF1C
chr7 (q11.23)
21.11

Figure 2-6. Gene expression in the 7q11.23 interval. A) For each of the expressed genes within the 7q11.23 interval, the $\log_2$ expression level is shown for duplications (blue) and controls (grey). The p-value is calculated using multivariate linear regression with 7q11.23 duplications and 411 controls without a known 7q11.23 event (Methods). Error bars for these four panels are defined as the 1.5 times the interquartile range. **B)** Genes showing expression deviating by at least two standard deviations from the mean across 3 samples with 7q11.23 duplications.

Figure 2-7

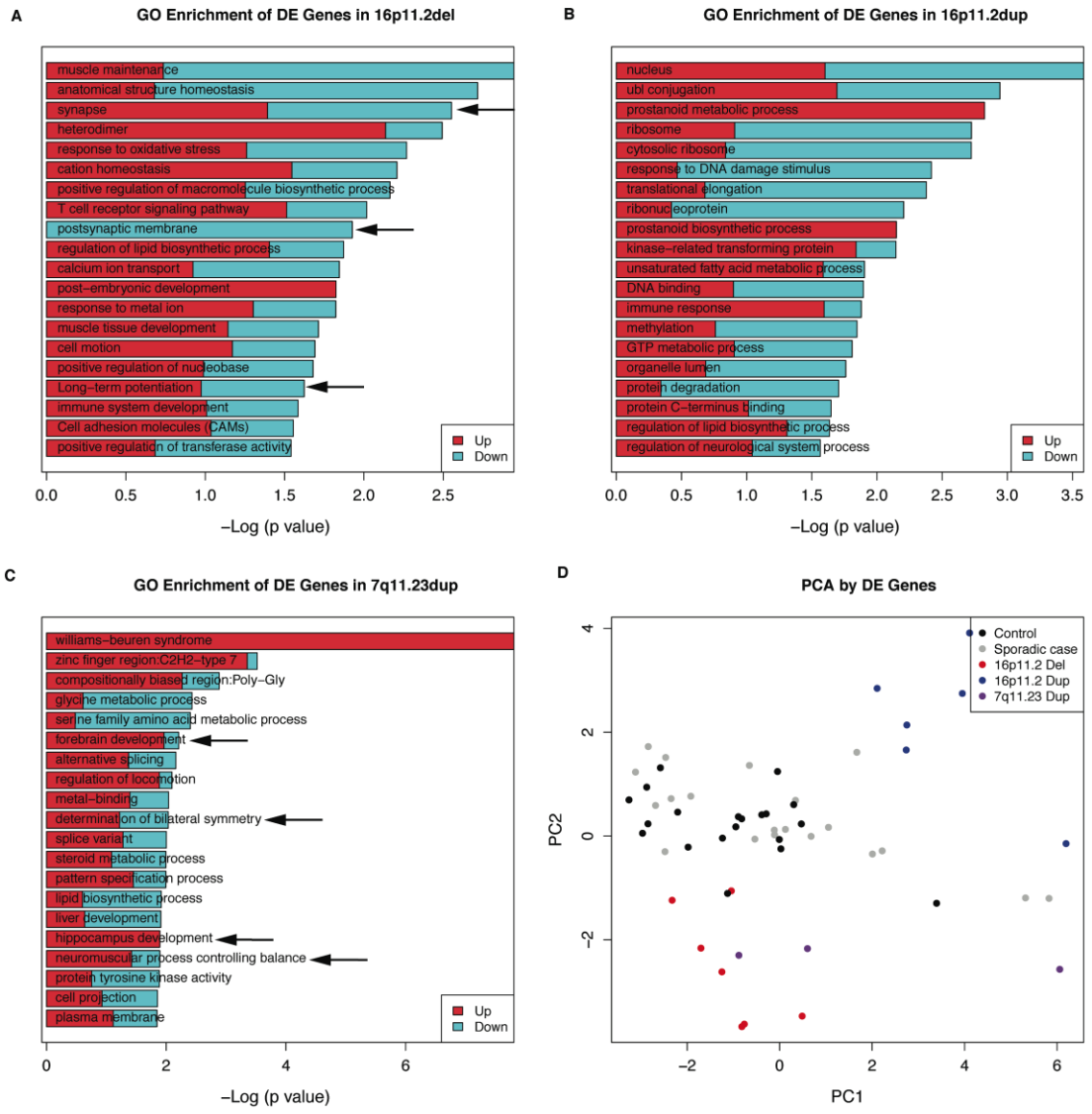

A
GO Enrichment of DE Genes in 16p11.2del
muscle maintenance
anatomical structure homeostasis
synapse
heterodimer
response to oxidative stress
cation homeostasis
positive regulation of macromolecule biosynthetic process
T cell receptor signaling pathway
postsynaptic membrane
regulation of lipid biosynthetic process
calcium ion transport
post−embryonic development
response to metal ion
muscle tissue development
cell motion
positive regulation of nucleobase
Long−term potentiation
immune system development
Cell adhesion molecules (CAMs)
positive regulation of transferase activity
Up
Down
0.0 0.5 1.0 1.5 2.0 2.5
−Log (p value)
B
GO Enrichment of DE Genes in 16p11.2dup
nucleus
ubl conjugation
prostanoid metabolic process
ribosome
cytosolic ribosome
response to DNA damage stimulus
translational elongation
ribonucleoprotein
prostanoid biosynthetic process
kinase−related transforming protein
unsaturated fatty acid metabolic process
DNA binding
immune response
methylation
GTP metabolic process
organelle lumen
protein degradation
protein C−terminus binding
regulation of lipid biosynthetic process
regulation of neurological system process
Up
Down
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
−Log (p value)
C
GO Enrichment of DE Genes in 7q11.23dup
williams−beuren syndrome
zinc finger region:C2H2−type 7
compositionally biased region:Poly−Gly
glycine metabolic process
serine family amino acid metabolic process
forebrain development
alternative splicing
regulation of locomotion
metal−binding
determination of bilateral symmetry
splice variant
steroid metabolic process
pattern specification process
lipid biosynthetic process
liver development
hippocampus development
neuromuscular process controlling balance
protein tyrosine kinase activity
cell projection
plasma membrane
Up
Down
0 2 4 6
−Log (p value)
D
PCA by DE Genes
Control
Sporadic case
16p11.2 Del
16p11.2 Dup
7q11.23 Dup
PC2
PC1
−2 0 2 4 6

Figure 2-7. GO enrichment analysis and principle component analysis highlight distinct molecular pathways in 16p11.2 duplications and deletions. A) GO enrichment analysis of the 307 genes ($p < 0.05$) showing altered expression in deletions (DAVID). The –log of the uncorrected p-value is shown in A-C. **B)** GO enrichment analysis of the 698 genes ($p < 0.05$) showing altered expression in duplications (DAVID). **C)** GO enrichment of the 439 genes ($p <0.05$) showing altered expression in 7q11.23 duplications (DAVID). **D)** Scatter plot of the first two components of 16p11.2 cases, 7q11.23 cases, sporadic autism cases and controls. Samples are clustered based on Principle Component Analysis (PCA). Seven 16p11.2 deletions probands (red), 6 16p11.2 duplications (green), 3 7q11.23 duplications (purple) probands were included. As a comparison group, 20 randomly selected sporadic probands (blue) and 20 randomly selected controls (black) were included. The first two principle components were used to form two-dimensional space. The merged list of differentially expressed (DEX) genes ($p<0.01$) in 16p11.2 duplications and deletions, as well as 7q11.23 duplications was utilized for PCA.

Table 2-1. Gene dysregulation in *de novo* CNVs.

Individual	Loci	Type	Size(kb)	%Outlier genes	Empirical p-value[a]	Outlier genes
12184.p1	12p11.22	Deletion	13,000	63%	1.00E-05	>10 genes
11233.p1	15q23	Deletion	5,000	53%	1.00E-05	*ADPGK,BBS4,KIF23,MYO9A,NPTN,PARP6,PKM2,RPLP1*
11090.p1	16p11.2	Deletion	600	47%	1.00E-05	*ALDOA,BOLA2,C16ORF53,CORO1A,HIRIP3,KCTD13, LOC606724,MAPK3,MAZ*
11540.p1	16p11.2	Deletion	600	58%	1.00E-05	>10 genes
12451.p1	16p11.2	Deletion	600	62%	1.00E-05	*ALDOA,C16ORF53,CDIPT,CORO1A,HIRIP3,KCTD13, MAPK3,MAZ,MVP,YPEL3*
11435.p1	16p13.3	Deletion	1,200	76%	1.00E-05	>10 genes
11080.p1	1p34.3	Duplication	5,000	64%	1.00E-05	> 10 genes
12239.p1	22q11.21	Deletion	1,400	93%	1.00E-05	>10 genes
11129.p1	7q11.23	Duplication	1,400	57%	1.00E-05	*BAZ1B,BCL7B,EIF4H,LAT2,NSUN5,STX1A,TBL2, WBSCR22*
12420.p1	1q21.1	Duplication	1,000	71%	3.00E-05	*ACP6,BCL9,CHD1L,GPR89A,PRKAB2*
12032.p1	3p13	Deletion	5,000	67%	5.00E-05	*ARL6IP5,C3ORF64,SUCLG2,TMF1,FOXP1,LMOD3*
11154.p1	7q11.23	Duplication	1,000	43%	0.00011	*BAZ1B,BCL7B,CLIP2,EIF4H,LAT2,WBSCR22*
11046.p1	3p26.2	Deletion	700	100%	0.00012	*ITPR1,SETMAR,SUMF1*
12343.p1	13q14.11	Duplication	500	75%	0.00039	*ELF1,MRPS31,WBP4*
11551.p1	16p13.2	Duplication	500	75%	0.00039	*CARHSP1,PMM2,USP7*
12594.p1	7q11.23	Duplication	300	75%	0.00039	*BCL7B,NSUN5,TBL2*
12647.p1	16p11.2	Duplication	500	32%	0.00046	*BOLA2,CORO1A,KCTD13,MAPK3,MVP,SULT1A3*
11353.p1	17q12	Deletion	1600	50%	0.00106	*AATF,ACACA,TADA2L*
12235.p1	9q34.11	Duplication	600	36%	0.00108	*ODF2,PTGES2,SET,SLC27A4*
12435.p1	16p11.2	Duplication	600	25%	0.00365	*CORO1A,IMAA,MAZ,SPN*
11433.p1	16p11.2	Deletion	500	21%	0.006	*ALDOA,KCTD13,MVP,SPN*
11555.p1	16p11.2	Duplication	700	21%	0.006	*C16ORF53,LOC606724,MAPK3,QPRT*
11435.p1	9p24.2	Duplication	3,000	33%	0.01022	*DOCK8,KIAA0020*
11962.p1	10q11.23	Duplication	1,700	100%	0.02	*CSTF2T*
12339.p1	3q27.2	Deletion	100	100%	0.02	*SFRS10*
12224.p1	22q13.1	Deletion	200	50%	0.035	*ADSL*
11343.p1	2p15	Duplication	1,700	50%	0.035	*XPO1*
12007.p1	15q11.2	Duplication	2,200	33%	0.05	*UBE3A*
11680.p1	16p11.2	Deletion	500	12%	0.05	*MAPK3,MVP*
12100.p1	16p11.2	Deletion	600	12%	0.05	*C16ORF53,HIRIP3*
11532.p1	17p13.1	Duplication	800	33%	0.05	*FAM64A*
12295.s1	19p13.3	Duplication	300	50%	0.00038	*C19ORF22,POLRMT,PTBP1,RNF126*
12117.s1	17q23.1	Duplication	2,000	67%	0.0026	*APPBP2,PPM1D*

[a] p-value is calculated by permutation test (Methods).

Figure 2-S1

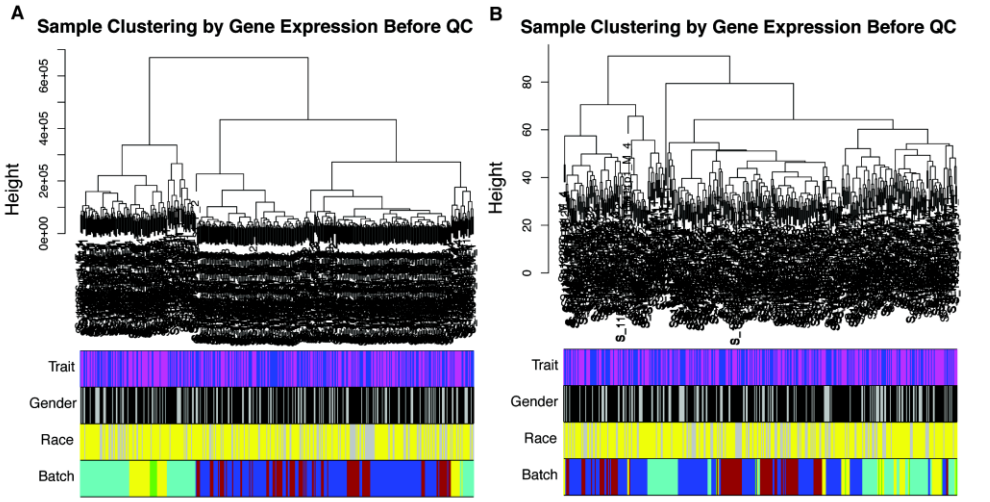

A
Sample Clustering by Gene Expression Before QC
Height
0e+00
2e+05
4e+05
6e+05
Trait
Gender
Race
Batch
B
Sample Clustering by Gene Expression Before QC
Height
0
20
40
60
80
Trait
Gender
Race
Batch

Figure 2-S1. Data pre-processing to remove outlier chips and correct for batch effects.

A) Hierarchical clustering of samples before data quality control (QC). Color bars show the trait (case: magenta; control: cyan), gender (male: black; female: grey), race (Caucasian: yellow; non-Caucasian: grey) and batch of each sample. Batch is defined based on the hybridization date. **B)** Hierarchical clustering after quality control including removing outlier chips, quantile normalization and combat for removing batch effects.

Figure 2-S2

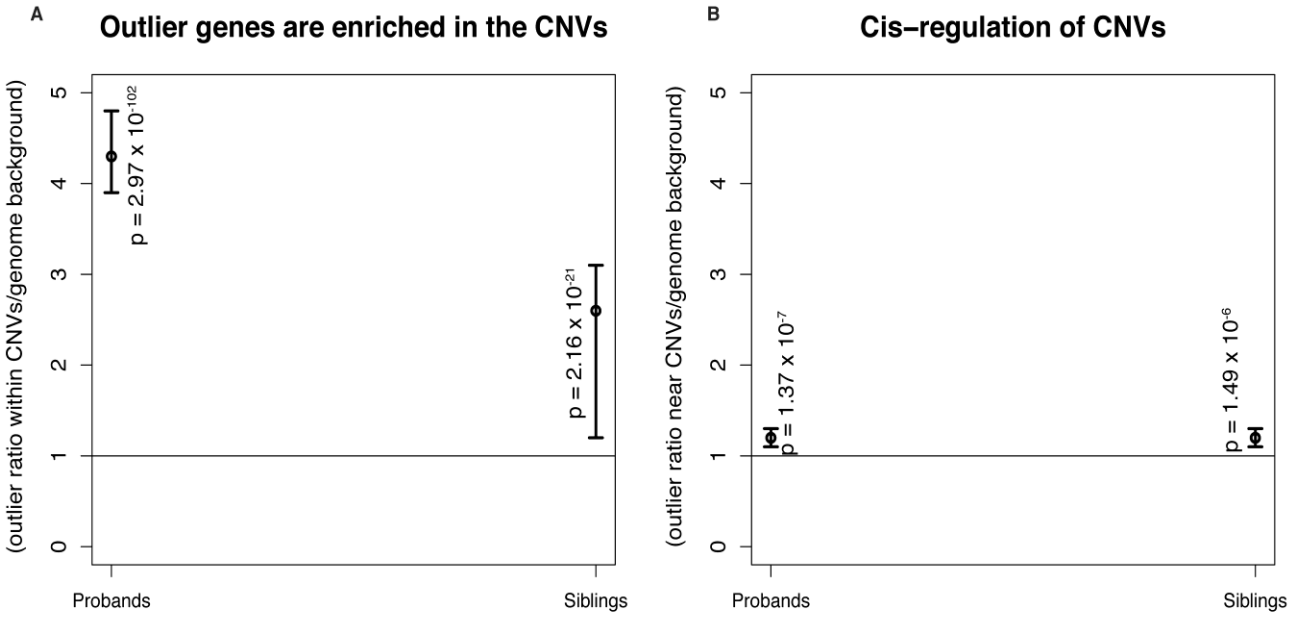

A
Outlier genes are enriched in the CNVs
(outlier ratio within CNVs/genome background)
0
1
2
3
4
5
p = 2.97 x 10-102
p = 2.16 x 10-21
Probands
Siblings
B
Cis−regulation of CNVs
(outlier ratio near CNVs/genome background)
p = 1.37 x 10-7
p = 1.49 x 10-6
Probands
Siblings

Figure 2-S2. CNVs affect the expression of genes within CNVs and up to 500kb surrounding them.

A) Odds ratio (OR) of the percentage of dysregulated genes (2SD) within CNVs compared to the percentage of dysregulated genes out of 9524 genes (11150 expressed probes) across the genome (background). Bar height shows the 95% confidence interval (CI). The CNVs comprise all CNVs each individual has, including both rare and common CNVs. **B)** Odds ratio of the percentage of dysregulated genes in the 500kb surrounding region of probands and siblings compared to the ratio of dysregulated genes in the genome background. The OR is significant for both probands and siblings for genes within CNVs, as well as genes within 500kb nearby (p value by Fisher's exact test).

Figure 2-S3

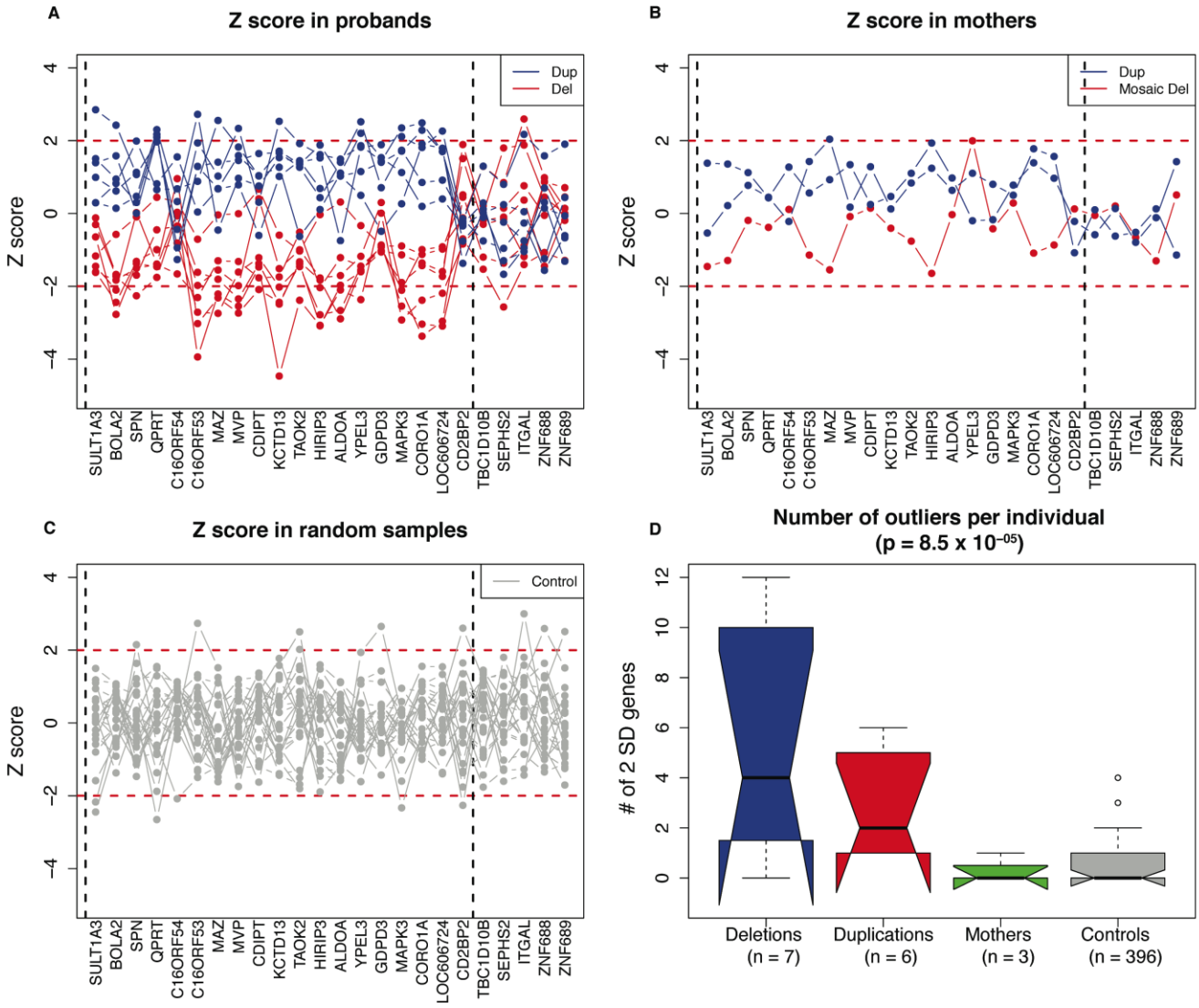

A
Z score in probands
Dup
Del
Z score
B
Z score in mothers
Dup
Mosaic Del
Z score
C
Z score in random samples
Control
Z score
D
Number of outliers per individual
(p = 8.5 x 10⁻⁰⁵)
of 2 SD genes
Deletions (n = 7)
Duplications (n = 6)
Mothers (n = 3)
Controls (n = 396)
SULT1A3
BOLA2
SPN
QPRT
C16ORF54
C16ORF53
MAZ
MVP
CDIPT
KCTD13
TAOK2
HIRIP3
ALDOA
YPEL3
GDPD3
MAPK3
CORO1A
LOC606724
CD2BP2
TBC1D10B
SEPHS2
ITGAL
ZNF688
ZNF689

Figure 2-S3. Dysregulation of genes within 16p11.2 and the closely surrounding region in probands, carriers and controls.

A) Z scores of 18 expressed genes within 16p11.2 and 6 expressed genes residing 500kb upstream or downstream in probands (7 deletions: red; and 6 duplications: blue). Genes on x-axis are aligned based on their location on chromosome. The 16p11.2 boundaries are shown with vertical dashed lines. 2SD is used as the cutoff to define outlier genes (horizontal dashed lines). **B)** Z scores of the same 24 genes in 3 mothers who carry the 16p11.2 events, but are unaffected (2 duplications: blue; and 1 mosaic deletion: red). **C)** Z scores of the same 24 genes in 20 randomly picked individuals (either probands or siblings) without known 16p11.2 events. **D)** The boxplot shows the number of outlier genes within 16p11.2 region per individual in different sample groups (p value = 8.5×10^{-05}, Kruskal-Wallis test).

Figure 2-S4

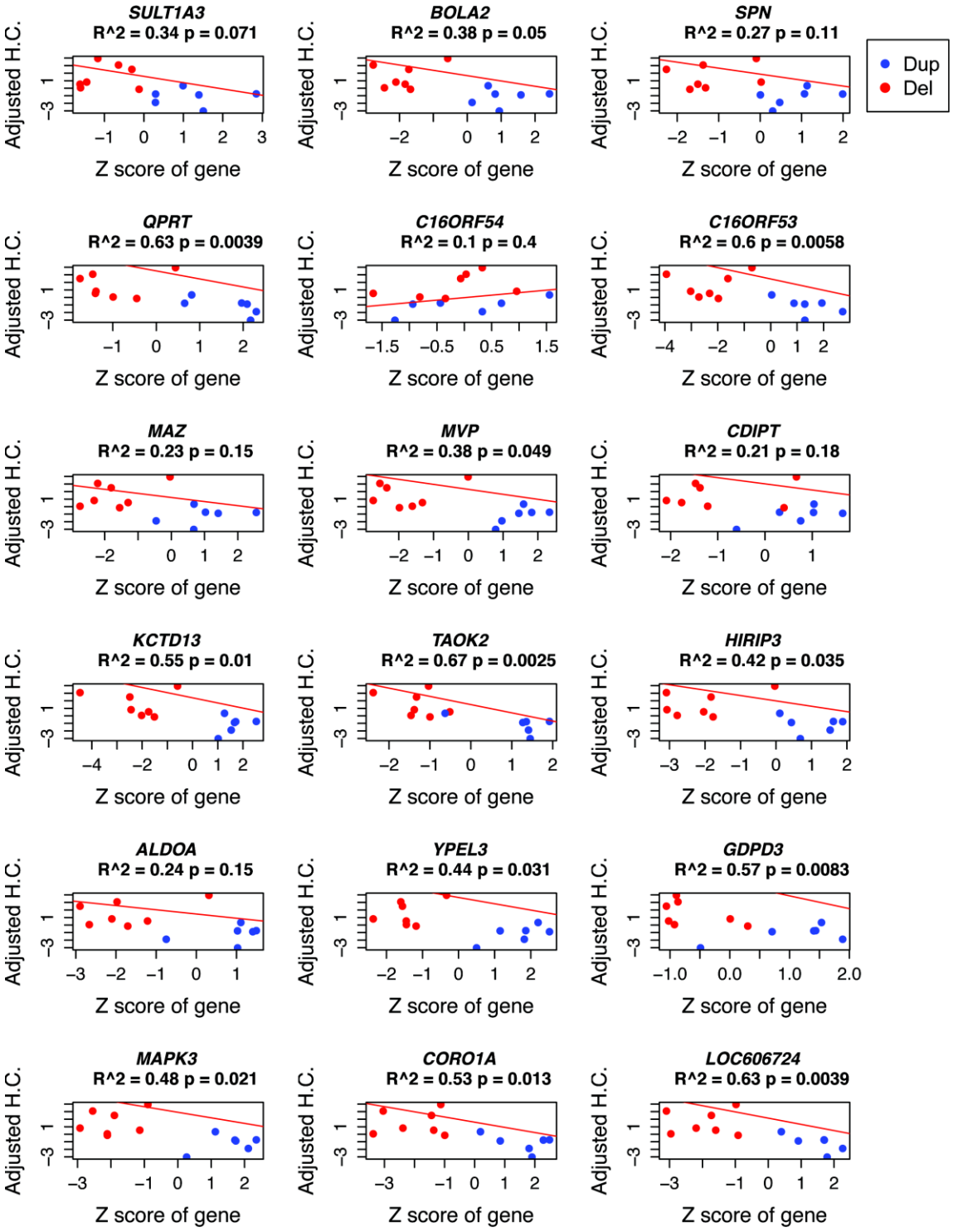

SULT1A3
R^2 = 0.34 p = 0.071
BOLA2
R^2 = 0.38 p = 0.05
SPN
R^2 = 0.27 p = 0.11
QPRT
R^2 = 0.63 p = 0.0039
C16ORF54
R^2 = 0.1 p = 0.4
C16ORF53
R^2 = 0.6 p = 0.0058
MAZ
R^2 = 0.23 p = 0.15
MVP
R^2 = 0.38 p = 0.049
CDIPT
R^2 = 0.21 p = 0.18
KCTD13
R^2 = 0.55 p = 0.01
TAOK2
R^2 = 0.67 p = 0.0025
HIRIP3
R^2 = 0.42 p = 0.035
ALDOA
R^2 = 0.24 p = 0.15
YPEL3
R^2 = 0.44 p = 0.031
GDPD3
R^2 = 0.57 p = 0.0083
MAPK3
R^2 = 0.48 p = 0.021
CORO1A
R^2 = 0.53 p = 0.013
LOC606724
R^2 = 0.63 p = 0.0039
Adjusted H.C.
Z score of gene
Dup
Del

Figure 2-S4. Correlation of head circumference and gene expression within 16p11.2.

The Z scores of 18 expressed genes within 16p11.2 region (x-axis) and adjusted head circumference (HC; y-axis) are shown. A multivariate linear regression model is fitted (variables used are standardized expression value (z score), age and gender; Methods). R-square of the linear regression model and p-value of the correlation between standardized expression value and HC is shown.

Figure 2-S5

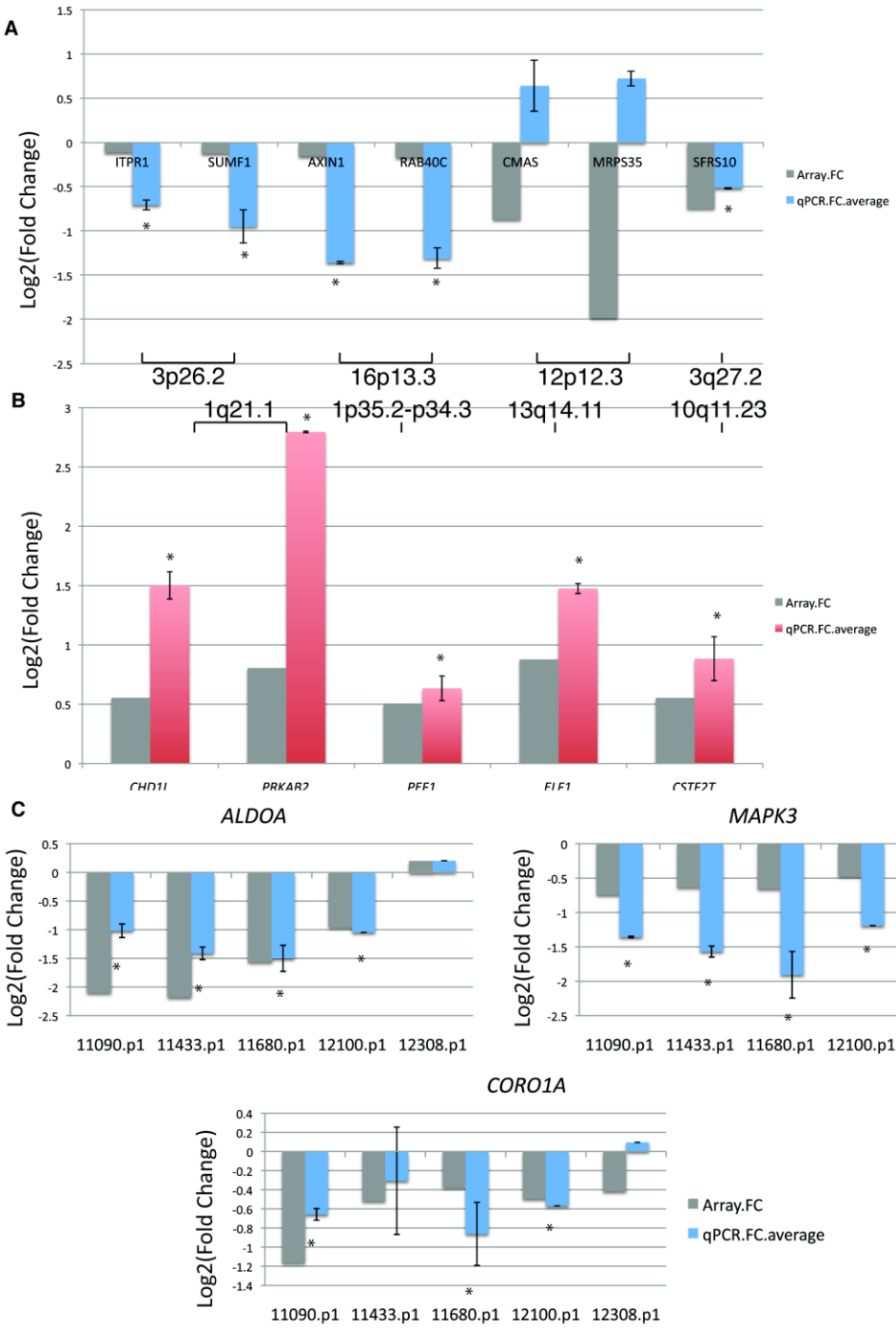
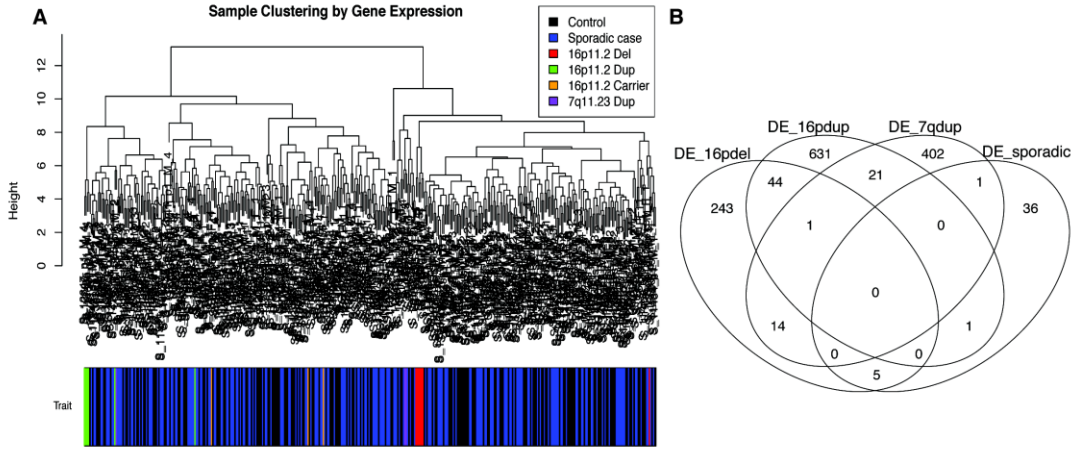

A
Log2(Fold Change)
ITPR1
SUMF1
AXIN1
RAB40C
CMAS
MRPS35
SFRS10
Array.FC
qPCR.FC.average
3p26.2
16p13.3
12p12.3
3q27.2
B
1q21.1
1p35.2-p34.3
13q14.11
10q11.23
Log2(Fold Change)
Array.FC
qPCR.FC.average
C
ALDOA
MAPK3
CORO1A
Log2(Fold Change)
11090.p1
11433.p1
11680.p1
12100.p1
12308.p1
Array.FC
qPCR.FC.average

Figure 2-S5. Confirmation of the outlier genes by qRT-PCR.

A) Eight down-regulated genes in 4 probands tested by qRT-PCR (Methods). Seventy-five percent of them are validated, showing at least 1.3-fold change (*). The CNV harboring each gene is shown. **B)** Five up-regulated genes in 4 probands are validated by qRT-PCR. One hundred percent of them are validated (* highlights genes with at least 1.3-fold change by qRT-PCR). **C)** Three genes down-regulated in 16p11.2 deletions are validated in 5 probands. Results represent the log 2 fold change of each gene on microarray and qRT-PCR (1.3-fold change: *). Error bars indicate the standard deviation of three replicates of qPCR results.

Figure 2-S6



C

	brain DEX (444)	15q11-13dup (80)	FMR1-FM (120)	Seno et al (33)	Hu et al (45)
sporadic(43)	*CEBPD*	NA	NA	NA	*HMBOX1*
16pdel(307)	*D4S234E, HERC6,ITPR1, PREPL,PRKCE, PTK2B,TNNT2*	NA	ZFP36	NA	NA
16pdup(698)	*RTPRT,SLC29A1*	*RIMS3,NR3C1*	*RIMS3,NR3C1*	NA	NA
7q11.23dup(439)	*LCMT1,PRKCB1,PRKCE*	*CYF1P1,RIMS3, CHST12,PRKCB1*	*PSEN2,GNG2, MBTPS1,RIMS3, CHST12, PRKCB1*	*LEF1,MFGE8, PLD1*	*SRD5A1*

Figure 2-S6. Differential expression analysis in the Simons Simplex Collection (SSC).

A) Sample clustering analysis for all sporadic cases (blue), controls (black), 16p11.2 deletions (red), 16p11.2 duplications (blue), 16p11.2 carriers (orange) and 7q11.23 duplications. **B)** Venn diagram of the overlap of DEX genes (p value < 0.05) identified in different groups (DEX, differentially expressed genes) **C)** DEX overlap with autism brain[26], recurrent events: 15q11-13dup, FMR1-FM [22] and LCLs [84, 85].

Table 2-S1 to Table 2-S6

On line resources: http://www.cell.com/AJHG/retrieve/pii/S0002929712002674

2.10 References

1. Geschwind DH: Advances in autism. *Annual Review Medicine* 2009, 60:367-380.

2. Miles JH: Autism spectrum disorders--a genetics review. *Genet Medicine* 2011, 13(4):278-294.

3. Jorde LB, Hasstedt SJ, Ritvo ER, Mason-Brothers A, Freeman BJ, Pingree C, McMahon WM, Petersen B, Jenson WR, Mo A: Complex segregation analysis of autism. *American Journal of Human Genetics* 1991, 49(5):932-938.

4. Bolton PF, Pickles A, Murphy M, Rutter M: Autism, affective and other psychiatric disorders: patterns of familial aggregation. *Psychol Med* 1998, 28(2):385-395.

5. Ronald A, Hoekstra RA: Autism spectrum disorders and autistic traits: a decade of new twin studies. *American Journal of Med Genet Neuropsychiatr Genet* 2011, 156B(3):255-274.

6. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS *et al*: Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010, 466(7304):368-372.

7. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J *et al*: Strong association of de novo copy number mutations with autism. *Science* 2007, 316(5823):445-449.

8. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y *et al*: Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics* 2008, 82(2):477-488.

9. Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP, Pakula AT, Eisenhauer NJ, Uhas K, Weik L *et al*: Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *American Journal of Human Genetics* 2010, 87(5):618-630.

10. Bourgeron T, Leboyer M, Delorme R: [Autism: more evidence of a genetic cause]. *Bull Acad Natl Med* 2009, 193(2):299-304; discussion 304-295.

11. Jamain S, Quach H, Betancur C, Rastam M, Colineaux C, Gillberg IC, Soderstrom H, Giros B, Leboyer M, Gillberg C *et al*: Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature Genetics* 2003, 34(1):27-29.

12. State MW: The genetics of child psychiatric disorders: focus on autism and Tourette syndrome. *Neuron* 2010, 68(2):254-269.

13. McClellan J, King MC: Genomic analysis of mental illness: a changing landscape. *JAMA* 2010, 303(24):2523-2524.

14. Geschwind DH: Autism: many genes, common pathways? *Cell* 2008, 135(3):391-395.

15. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S *et al*: 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genetics* 2009, 41(2):160-162.

16. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T *et al*: Association between microdeletion and microduplication at 16p11.2 and autism. *New England Journal of Medicine* 2008, 358(7):667-675.

17. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP *et al*: Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009, 459(7246):569-573.

18. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA *et al*: Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011, 70(5):863-885.

19. Treadwell-Deering DE, Powell MP, Potocki L: Cognitive and behavioral characterization of the Potocki-Lupski syndrome (duplication 17p11.2). *J Dev Behav Pediatr* 2010, 31(2):137-143.

20. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C *et al*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007, 315(5813):848-853.

21. Abrahams BS, Geschwind DH: Advances in autism genetics: on the threshold of a new neurobiology. *Nature Review Genetics* 2008, 9(5):341-355.

22. Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, Sigman M, Steindler C, Pellegrini S, Schanen NC, Warren ST *et al*: Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics* 2007, 16(14):1682-1698.

23. Coppola G, Karydas A, Rademakers R, Wang Q, Baker M, Hutton M, Miller BL, Geschwind DH: Gene expression study on peripheral blood identifies progranulin mutations. *Ann Neurol* 2008, 64(1):92-96.

24. Voineagu I, Huang L, Winden K, Lazaro M, Haan E, Nelson J, McGaughran J, Nguyen LS, Friend K, Hackett A *et al*: CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. *Mol Psychiatry* 2011, 17(1):4-7.

25. Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, Geschwind DH, Mane SM, State MW, Sestan N: Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 2009, 62(4):494-509.

26. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2010, 474(7351):380-384.

27. Rodier PM, Ingram JL, Tisdale B, Nelson S, Romano J: Embryological origin for autism: developmental anomalies of the cranial nerve motor nuclei. *J Comp Neurol* 1996, 370(2):247-261.

28. Ploeger A, Raijmakers ME, van der Maas HL, Galis F: The association between autism and errors in early embryogenesis: what is the causal mechanism? *Biol Psychiatry* 2010, 67(7):602-607.

29. Walsh CA, Morrow EM, Rubenstein JL: Autism and brain development. *Cell* 2008, 135(3):396-400.

30. Barnby G, Abbott A, Sykes N, Morris A, Weeks DE, Mott R, Lamb J, Bailey AJ, Monaco AP: Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. *American Journal of Human Genetics* 2005, 76(6):950-966.

31. Caliskan M, Cusanovich DA, Ober C, Gilad Y: The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics* 2011, 20(8):1643-1652.

32. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011, 474(7351):380-384.

33. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D: Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 2011, 70(5):898-907.

34. Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A: Segmental copy number variation shapes tissue transcriptomes. *Nature Genetics* 2009, 41(4):424-429.

35. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI, Alvarez Retuerto AI, Imielinski M, Hadley D, Bradfield JP *et al*: Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* 2009, 5(6):e1000536.

36. Huang N, Lee I, Marcotte EM, Hurles ME: Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 2010, 6(10):e1001154.

37. Celestino-Soper PB, Shaw CA, Sanders SJ, Li J, Murtha MT, Ercan-Sencicek AG, Davis L, Thomson S, Gambin T, Chinault AC *et al*: Use of array CGH to detect exonic copy number variants throughout the genome in autism families detects a novel deletion in TMLHE. *Human Molecular Genetics* 2011, 20(22):4360-4370.

38. Celestino-Soper PcBS, Violante S, Crawford EL, Luo R, Lionel AC, Delaby E, Cai G, Sadikovic B, Lee K, Lo C *et al*: A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism. *Proceedings of the National Academy of Sciences of the United States of America doi:10 1073/pnas1120210109* 2012.

39. Horev G, Ellegood J, Lerch JP, Son YE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M *et al*: Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proceedings of the National Academy of Sciences of the United States of America* 2011, 108(41):17076-17081.

40. Bijlsma EK, Gijsbers AC, Schuurs-Hoeijmakers JH, van Haeringen A, Fransen van de Putte DE, Anderlid BM, Lundin J, Lapunzina P, Perez Jurado LA, Delle Chiaie B *et al*: Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *Eur Journal Med Genet* 2009, 52(2-3):77-87.

41. Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, Szatmari P, Joseph-George AM, Mackay S, Whitten K, Noble B *et al*: Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J Med Genet* 2009, 47(3):195-203.

42. Hanson E, Nasir RH, Fong A, Lian A, Hundley R, Shen Y, Wu BL, Holm IA, Miller DT: Cognitive and behavioral characterization of 16p11.2 deletion syndrome. *J Dev Behav Pediatr* 2010, 31(8):649-657.

43. Kumar RA, Marshall CR, Badner JA, Babatz TD, Mukamel Z, Aldinger KA, Sudi J, Brune CW, Goh G, Karamohamed S *et al*: Association and mutation analyses of 16p11.2 autism candidate genes. *PLoS One* 2009, 4(2):e4582.

44. Shinawi M, Liu P, Kang SH, Shen J, Belmont JW, Scott DA, Probst FJ, Craigen WJ, Graham BH, Pursley A *et al*: Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet* 2010, 47(5):332-341.

45. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O *et al*: Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics* 2009, 41(11):1223-1227.

46. Miller DT, Nasir R, Sobeih MM, Shen Y, Wu BL, Hanson E: 16p11.2 Microdeletion 2009 Sep 22 [Updated 2011 Oct 27]. In: Pagon RA, Bird TD, Dolan CR, et al., editors. GeneReviews™ [Internet]. Seattle (WA): University of Washington, Seattle; 1993-.

47. Kumar RA, Sudi J, Babatz TD, Brune CW, Oswald D, Yen M, Nowak NJ, Cook EH, Christian SL, Dobyns WB: A de novo 1p34.2 microdeletion identifies the synaptic vesicle gene RIMS3 as a novel candidate for autism. *J Med Genet* 2010, 47(2):81-90.

48. Cai C, Langfelder P, Fuller TF, Oldham MC, Luo R, van den Berg LH, Ophoff RA, Horvath S: Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics* 2010, 11:589.

59. Dolmetsch R, Geschwind DH: The Human Brain in a Dish: The Promise of iPSC-Derived Neurons. *Cell* 2011, 145(6):831-834.

50. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A *et al*: De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012, 74(2):285-299.

51. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, Dilullo NM, Parikshak NN, Stein JL *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature doi:101038/nature10945* 2012.

52. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V *et al*: Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature, doi:101038/nature11011* 2012.

54. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD *et al*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature doi:101038/nature10989* 2012.

53. Pobbe RL, Pearson BL, Blanchard DC, Blanchard RJ: Oxytocin receptor and Mecp2(308/Y) knockout mice exhibit altered expression of autism-related social behaviors. *Physiol Behav doi:101016/jphysbeh201202024* 2012.

54. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C *et al*: Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* 2008, 4(11):e1000287.

55. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE: De novo rates and selection of large copy number variation. *Genome Res* 2010, 20(11):1469-1481.

56. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L, Walsh T, McCarthy SE, Baker C *et al*: A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genetics* 2010, 42(3):203-209.

57. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P *et al*: Origins and functional impact of copy number variation in the human genome. *Nature* 2009, 464(7289):704-712.

58. Hehir-Kwa JY, Wieskamp N, Webber C, Pfundt R, Brunner HG, Gilissen C, de Vries BB, Ponting CP, Veltman JA: Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS Comput Biol* 2010, 6(4):e1000752.

59. Novak MJ, Sweeney MG, Li A, Treacy C, Chandrashekar HS, Giunti P, Goold RG, Davis MB, Houlden H, Tabrizi SJ: An ITPR1 gene deletion causes spinocerebellar ataxia 15/16: a genetic, clinical and radiological description. *Mov Disord* 2010, 25(13):2176-2182.

60. Schorge S, van de Leemput J, Singleton A, Houlden H, Hardy J: Human ataxias: a genetic dissection of inositol triphosphate receptor (ITPR1)-dependent signaling. *Trends Neurosci* 2010, 33(5):211-219.

61. Settembre C, Annunziata I, Spampanato C, Zarcone D, Cobellis G, Nusco E, Zito E, Tacchetti C, Cosma MP, Ballabio A: Systemic inflammation and neurodegeneration in a mouse model of multiple sulfatase deficiency. *Proceedings of the National Academy of Sciences of the United States of America* 2007, 104(11):4506-4511.

62. Benderska N, Becker K, Girault JA, Becker CM, Andreadis A, Stamm S: DARPP-32 binds to tra2-beta1 and influences alternative splicing. *Biochim Biophys Acta* 2010, 1799(5-6):448-453.

63. Lee JA, Tang ZZ, Black DL: An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons. *Genes Dev* 2009, 23(19):2284-2293.

64. Martin CL, Duvall JA, Ilkin Y, Simon JS, Arreaza MG, Wilkes K, Alvarez-Retuerto A, Whichello A, Powell CM, Rao K *et al*: Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. *Am journal med genet neuropsychiatr Genet* 2007, 144B(7):869-876.

65. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, Hill RS, Mukaddes NM, Balkhy S, Gascon G, Hashmi A *et al*: Identifying autism loci and genes by tracing recent shared ancestry. *Science* 2008, 321(5886):218-223.

66. Newbury DF, Warburton PC, Wilson N, Bacchelli E, Carone S, Lamb JA, Maestrini E, Volpi EV, Mohammed S, Baird G *et al*: Mapping of partially overlapping de novo deletions across an autism susceptibility region (AUTS5) in two unrelated individuals affected by developmental delays with communication impairment. *American Journal Med Genet A* 2009, 149A(4):588-597.

67. Laumonnier F, Roger S, Guerin P, Molinari F, M'Rad R, Cahard D, Belhadj A, Halayem M, Persico AM, Elia M *et al*: Association of a functional deficit of the BKCa channel, a synaptic regulator of neuronal excitability, with autism and mental retardation. *Am Journal Psychiatry* 2006, 163(9):1622-1629.

68. Pellerin L: Food for thought: the importance of glucose and other energy substrates for sustaining brain function under varying levels of activity. *Diabetes Metab* 2011, 36 Suppl 3:S59-63.

69. Okamoto S, Sherman K, Bai G, Lipton SA: Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal differentiation. *Brain Res Mol Brain Res* 2002, 107(2):89-96.

70. Van der Aa N, Rooms L, Vandeweyer G, van den Ende J, Reyniers E, Fichera M, Romano C, Delle Chiaie B, Mortier G, Menten B *et al*: Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. *Eur Journal Med Genet* 2009, 52(2-3):94-100.

71. Moore TM, Garg R, Johnson C, Coptcoat MJ, Ridley AJ, Morris JD: PSK, a novel STE20-like kinase derived from prostatic carcinoma that activates the c-Jun N-terminal kinase mitogen-activated protein kinase pathway and regulates actin cytoskeletal organization. *J Biol Chem* 2000, 275(6):4311-4322.

71. Wagner B, Sibilia M: Methods to study MAP kinase signalling in the central nervous system. *Methods Mol Biol* 2010, 661:481-495.

72. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: Functional organization of the transcriptome in human brain. *Nature Neuroscience* 2008, 11(11):1271-1282.

73. Miller JA, Horvath S, Geschwind DH: Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* 2010, 107(28):12698-12703.

74. Gold DL, Wang J, Coombes KR: Inter-gene correlation on oligonucleotide arrays: how much does normalization matter? *Am Journal Pharmacogenomics* 2005, 5(4):271-279.

75. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8(1):118-127.

76. Bonferroni CE: Teoria statistica delle classi e calcolo delle probabilit `a. *Pubblicazioni del r istituto superiore di scienze economiche e commerciali di firenze 8* 1936(8): 3--62.

77. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW: Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res* 2006, 115(3-4):205-214.

78. Liang KY, Zeger SL: Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika* 1986, 73(1):13-22.

79. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004, 3:Article3.

80. Roche AF, Mukherjee D, Guo SM, Moore WM: Head circumference reference data: birth to 18 years. *Pediatrics* 1987, 79(5):706-712.

81. Ghahramani Seno MM, Hu P, Gwadry FG, Pinto D, Marshall CR, Casallo G, Scherer SW: Gene and miRNA expression profiles in autism spectrum disorders. *Brain Res* 2011, 1380:85-97.

82. Hu VW, Nguyen A, Kim KS, Steinberg ME, Sarachana T, Scully MA, Soldin SJ, Luu T, Lee NH: Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis. *PLoS One* 2009, 4(6):e5775.

CHAPTER 3: Understand autism risk genes at the transcriptomic network level

The abstract introduction, results and methods quoted below are directly from published papers. I have the permission of all authors and the letters are appended at the end of my thesis:

3.1 Abstract

" Genome-wide transcriptional profiling was used to characterize the molecular underpinnings of neocortical organization in rhesus macaque, including cortical areal specialization and laminar cell-type diversity. Microarray analysis of individual cortical layers across sensorimotor and association cortices identified robust and specific molecular signatures for individual cortical layers and areas, prominently involving genes associated with specialized neuronal function. Overall, transcriptome-based relationships were related to spatial proximity, being strongest between neighboring cortical areas and between proximal layers. We observed that laminar patterns were more similar between macaque and human compared to mouse, as was the unique V1 profile that was not observed in mouse. These data provide a unique resource detailing neocortical transcription patterns in a nonhuman primate with great similarity in gene expression to human. Meanwhile, genome-wide transcriptome profiling has been done at human fetal brain, in which the expression of each laminar was collected. Together these data provide a rich freely accessible resource for 1) understanding mechanisms of macaque and human brain, 2) study the brain-related diseases.

Genetic studies have identified dozens of autism spectrum disorder (ASD) susceptibility genes, raising two critical questions: 1) do these genetic loci converge on specific biological processes, and 2) where does the phenotypic specificity of ASD arise, given its genetic overlap

with intellectual disability (ID)? To address this, we mapped ASD and ID risk genes onto co-expression networks representing developmental trajectories and transcriptional profiles representing fetal and adult cortical laminae. ASD genes tightly coalesce in modules that implicate distinct biological functions during human cortical development, including early transcriptional regulation and synaptic development. At a circuit level, ASD genes are enriched in superficial cortical layers and glutamatergic projection neurons. Furthermore, we show that the patterns of ASD and ID risk genes are distinct, providing a novel biological framework for investigating the pathophysiology of ASD.

3.2 Introduction

The mammalian neocortex is characterized by its stereotyped laminar cytoarchitecture and regional variations in cellular architecture that differentiate cortical areas. As emphasized by Brodmann over a century ago through the creation of cytoarchitectonic cortical maps [1], cortical organization is conserved across species, particularly between humans and nonhuman primates [2]. Gene expression is increasingly used as an empirical means of differentiating and delineating cortical areas, for example through identification of area-specific gene markers [3] or boundary mapping based on differences in neurotransmitter receptor expression [4]. Whole-genome transcriptional profiling has particular potential to elucidate cortical areal specification and specialization through identification of differentially regulated genes and molecular pathways that underlie cytoarchitectural and functional areal identity [5].

The major factor that differentiates different cortical areas is their distinct laminar organization, reflecting the composition of specific cell types within each layer. Many gene markers have been identified through mining genome-wide cellular resolution gene expression data resources in the

Allen Mouse Brain Atlas [6] (http://www.brain-map. org) and by using targeted approaches [7]. In addition, transcriptional profiling using DNA microarrays or RNA sequencing has been successful in identifying molecular signatures for discrete cortical layers in mice [8-10] using punches or laser microdissection, as well as in specific excitatory and inhibitory cortical cell types using selective genetic or tracer-based cell labeling and live isolation methods [11-13].

Rhesus macaque provides a tractable nonhuman primate model system to analyze the transcriptional organization of the primate neocortex. Macaque is genetically and physiologically similar to humans, with a sequence identity of approximately 93% [14]. Many elements of cortical cytoarchitecture are similar in macaque and human, including specialized primary visual cortex and dorsal and ventral visual streams. In this study, we aimed to understand organizational principles of the primate neocortex using transcriptional profiling analysis of individually isolated cortical layers from a variety of well-defined cortical regions in the adult rhesus macaque and to compare rhesus gene expression patterns in homologous cortical areas and cell types in human and mouse. The entire microarray data set is also available through the NIH Blueprint NHP Atlas website (http://blueprintnhpatlas.org).

Autism Spectrum Disorders (ASDs) are a heterogeneous neurodevelopmental disorder, in which hundreds of genes have been implicated [15, 16]. Analysis of copy number variation (CNV) and exome sequencing [17-20] have identified rare de novo variants (RDNVs) that alter dozens of protein coding genes in ASD, none of which account for more than 1% of ASD cases [21]. This, and the fact that a significant fraction (40-60%) of ASDs is explained by common variation [22], points to a heterogeneous genetic architecture.

These findings raise several issues. Based on the background human mutation rate [23], most genes affected by only one observed RDNV to date are likely false positives that do not increase risk for ASDs [24]. It is therefore essential to develop approaches that prioritize singleton variants, especially missense mutations. Furthermore, given the heterogeneity of ASD, it would valuable to identify common pathways, cell-types, or circuits disrupted within ASD itself. Recent studies combining gene expression, protein-protein interactions (PPIs), and other systematic gene annotation resources suggest some molecular convergence in subsets of ASD risk genes [25-28]. Yet, it remains unclear how the large number of genes implicated through different methods may converge to affect human brain development, which is critical to a mechanistic understanding of ASDs [15]. Additionally, ASDs have considerable overlap with ID at the genetic level, so identifying molecular pathways and circuits that confer the phenotypic specificity of ASDs would be of considerable utility [29, 30].

Here, we took a stepwise approach to determine if genes implicated in ASDs affect convergent pathways during in vivo human neural development, and whether they are enriched in specific cells or circuits. Here we assessed shared neurobiological function among these genes, including enrichment in layer-specific patterns from micro-dissected human fetal and adult primate cortical laminae. Our integration of large transcriptomic profiling data with disease relevant gene lists permits rigorous interrogation of biological convergence and specificity in ASDs that takes its heterogeneity into consideration and enables comparison of ASDs with ID.

3.3 Transcriptome profiling of macaque and human brain

To analyze transcriptional profiles associated with major laminar and areal axes of cortical organization, laser microdissection (LMD) was used to selectively isolate individual

cortical layers in ten discrete areas of the neocortex from two male and two female adult rhesus monkeys. These areas spanned primary sensorimotor cortices (S1, M1, A1, and V1), higher-order visual areas (V2, MT, and TE), and frontal cortical areas (DLPFC, OFC, and ACG). In each cortical region, samples were isolated from layers definable on the basis of lightly stained Nissl sections used for the sample preparation, taking care to avoid layer boundaries. In most areas, 5 layers were isolated (L2, L3, L4, L5, and L6), although in M1, OFC, and ACG no discernible L4 could be isolated. Eight layers were sampled in V1 to include the functionally specialized and cytoarchitecturally distinct sublayers of L4 (4A, 4B, 4Ca, and 4Cb). For a nonneocortical comparator data set, samples were also isolated from subfields of the hippocampus (CA1, CA2, CA3, and dentate gyrus) and from the magno-, parvo-, and koniocellular layers of the dorsal lateral geniculate nucleus (LGN). Collectively, the selected regions allowed for interrogation of differences in gene expression between cortical areas and layers located distal or proximal to each other, and from regions that comprise specific functional types or streams. Representative pre- and postcut images from each structure are shown in Figure S1, available online, and stereotaxic locations of sampled cortical regions in Table S1. RNA was isolated from LMD samples, and 5 ng total RNA per sample was amplified to generate sufficient labeled probe for use on Affymetrix rhesus macaque microarrays.

Multiple analytical methods were used independently to identify the most robust patterns of gene expression. Principle component analysis (PCA) can often illustrate the major organizational features of microarray data sets, and we initially applied it to the whole sample set comprising 225 cortical, hippocampal, and thalamic samples across all 52,865 probes. A significant proportion of the variance was ac- counted for by the first three components (12.5%, 8.7%, and 6.8%, respectively; Figure S2). Samples from major structures (cortex, hippocampus,

and thalamus) cluster together, have highly distinct molecular signatures and appear well segregated. Considering the cortical samples alone, the first three components accounted for a similar proportion of variance (13.6%, 8.5%, and 6.6%, respectively), and plotting samples by areal or laminar class revealed striking organization along two orthogonal axes reflecting the areal and laminar dimensions of the neocortex. Remarkably, the spatial relationships between neocortical samples are recapitulated by the transcriptional relationships between samples. Samples align in a rostral to caudal orientation by cortical area (4,923), and individual animals (2,347; Table 3-S2). Importantly, there was a high degree of overlap between the sets of genes varying by cortical region and layer, suggesting that a substantial proportion of the genes differentiating cortical areas vary within specific cortical layers. Gene set analysis of both areal and laminar genes showed enrichment for genes associated with axonal guidance signaling and ephrin receptor signaling, synaptic long-term potentiation (LTP) and neuronal activities (Table 3-S2). Gene expression patterns associated with gender and individual animals were also identified by ANOVA (Figure S2), and individual-associated differences were enriched with genes related to metabolism, mitochondria, and antigen presentation (Table 3-S2). Gender-specific gene expression was observed both on sex and autosomal chromosomes (Figure 3-S2), and there was significant overlap ($p < 1e-09$) between the individual-related genes identified here and gender-related genes identified in human brain [31].

We next applied WGCNA to identify sets, or modules, of highly coexpressed genes by searching for genes with similar patterns of variation across samples as defined by high topological overlap [32]. Applied to the entire set of neocortical samples, WGCNA revealed a series of gene modules (named here as colors) related to different features of the data set. Gene assignment to modules and gene ontology analysis for the whole cortex network are shown in

Table 3-S3. The majority of these modules correlated with laminar and regional patterns as described below. Several modules were related to gender and individual differences, as previously observed in humans [33]. In Figure 3-1G, the lightyellow module was strongly enriched in male versus female samples (upper panel), while the grey60 module was selectively lowest in samples originating from one particular animal. The top (hub) genes in the lightyellow module were on the Y chromosome, including the putative RNA helicase DDX3Y and the 40S ribosomal protein RPS4Y1.

The most striking features were the robust molecular signatures associated with different cortical layers. A wide variety of transcriptional patterns were associated with individual cortical layers or subsets of layers. For example, ANOVA of laminar expression in all cortical regions and clustering of these genes identified large gene sets enriched in specific subsets of (generally proximal) layers. Notably, the majority of these laminar patterns are consistent across different cortical areas, reflecting conserved laminar and cellular architecture across the cortex. Gene set analysis suggests these layer-associated clusters are associated with neuronal function, including neuronal activity, LTP/LTD, calcium, glutamate and GABA signaling. Consistent with functional studies of superficial layer synaptic plasticity, genes and pathways involved in LTP and calcium signaling were most represented in L2 and L3. Pathways related to cholesterol metabolism were enriched in deeper layers, likely reflecting the greater proportion of oligodendrocytes closer to the underlying white matter. Similarly, many of the gene modules identified through WGCNA of all cortical samples were correlated with specific cortical layers. By ANOVA-based clustering and WGCNA, proximal layers showed the strongest correlations, with superficial L2 and L3 highly correlated with one another, and the deeper L4–6 highly correlated as well (dendrograms in Figures 3-1B, 1E, and 1F).

Individual layers showed highly specific gene expression signatures. Layer-enriched expression patterns were identified by searching for genes with high correlation to layer-specific artificial template patterns [34] (Table 3-S5). Figure 3-1C shows cohorts of genes with remarkably layer-specific expres- sion that was relatively constant across all cortical areas. These observations demonstrate the specificity of the laminar dissec- tions with minimal interlaminar contamination, and also the constancy of laminar gene expression across the neocortex.

WGCNA gene modules derived from the whole cortex network also showed highly layer-enriched expression, demonstrating the robustness of our findings. For example, the black module contains genes enriched in superficial L2 (hub genes plotted in Figure 3-1D, top row). While some layer-specific genes could be identified by targeted analyses, the dominant patterns were more complex, with most network modules being associated with combinations of layers, typically proximal to one another. For example, individual modules were enriched in L2–4 (salmon), L3–5 (greenyellow), L4–5 (royalblue) and in a gradient increasing from L2 to L6 (red). This tendency for coexpression between adjacent layers is also apparent in the heatmap representation of gene clusters in Figures 3-1A and 3-1E. Gene ontology (GO) analysis of these modules provides some insight into their functional relevance (Table 3-S3). The greenyellow module was enriched for genes associated with axons and neuron projec- tions, potentially related to long-range pyramidal projection neurons in L3 and L5. The red module showed enrichment in genes associated with myelination, consistent with the presence of oligodendrocytes in deep layers, and this module was highly correlated with oligodendrocyte-associated gene networks in other studies (data not shown).

Interestingly, the expanded L4 of V1 displayed a distinct sig- nature from the rest of L4 (see top of middle box in Figure 3-2A). To explore this further, we performed ANOVA and

WGCNA selectively on samples from V1 (Figures 3-2E and 2F; Tables 3-S6 and 3-S7). A comparison between V1 ANOVA-derived laminar differential expression and membership in whole cortex WGCNA modules is in Table S8. Similar to the whole cortex analysis, robust clusters and network modules were associated with individual cortical layers. As shown in the unsupervised hierarchical 2D clustering of ANOVA results in Figure 2E, individual samples from each layer cluster together, and neighboring cortical layers are most similar to one another. Interestingly, L4A clusters with more superficial layers, while L4B, L4Ca, and L4Cb display a distinct transcriptional pattern, most easily seen by the dendrograms based on ANOVA and network analysis in Figures 3-2E and 2F.

To investigate whether layer specificity of gene expression may relate to selective patterns of connectivity, we examined the relationship between thalamocortical inputs and their targets in V1. L4Ca and L4Cb receive input selectively from magnocellular (M) and parvocellular (P) divisions of the LGN, respectively. Hypothesizing that there may be substantial shared gene expression patterns selective for specific pairs of connected neurons, we searched for genes that were differentially expressed between the thalamic inputs and between the cortical targets. One thousand two probes were differentially expressed between L4Ca and L4Cb (t test, $p < .01$) and 825 probes between M and P. Surprisingly, these gene sets did not significantly overlap (13/1,827; $p = 0.08$). Although the possibility certainly exists that specific ligand-receptor pairs are associated with this selective connectivity, it would appear that the specificity of these connections is not associated with specific large-scale correlated gene expression patterns.

To validate the specificity of the microarray findings and test hypotheses about laminar enrichment based on ANOVA and WGCNA, we examined a set of genes displaying layer-

enriched patterns using in situ hybridization (ISH) in areas V1 and V2. Overall the laminar specificity of gene expression and variations between cortical areas predicted by microarrays were confirmed by cellular-level analysis and illustrate the high information content of layer-specific expression profiling and gene specificity of the microarray probesets. For example, GPR83 is selectively expressed in L2 of all cortical areas, both by microarray and ISH analysis. Laminar specificity was confirmed for RORB (L3–5), PDYN (L4–5), CUX2 (L2–4), and SV2C (L3–4 enriched). Specificity for deep cortical layers was prominent, as shown for PDE1A (L5–6), NR4A2 (L5–6), COL24A1 (L6) and RXFP1 (L5–6). Differences in laminar specificity were some- times apparent between V1 and V2 (generally V1 versus all other areas); CUX2 was expressed in L2 through L4Cb in V1 but more limited to L2 and L3 in V2, and SV2C was highest in L4B in V1, but highest in L3 in area V2.

Both ANOVA and WGCNA analysis identified gene clusters enriched in specific subsets of cortical regions. As illustrated in the dendrograms from both methods, the strongest relationships between cortical areas were based on areal proximity rather than functionally connectivity. For example, the caudal visual areas V1, V2, and MT showed highly correlated patterns of gene expression, while the functionally related but distal visual region TE had greater transcriptional similarity to its proximal neighbor A1 in temporal cortex. Strong relationships were observed for the adjacent primary motor and sensory cortices M1 and S1 and for the frontal DLPFC and OFC regions. Differentially expressed genes showed enrichment in specific subsets of (generally proximal) cortical areas, generally related to neuronal development and function (axon guidance, neuronal activities, LTP/LTD; Table 3-S9). Areal expression also had a strong laminar signature, easily visualized by grouping these ANOVA-derived genes by cortical layer (Figure 3-S3).

Parallel relationships between cortical areas were observed by WGCNA demonstrating the robustness of these observations, with individual gene modules showing enrichment in specific cortical regions. Module eigengenes revealed additional patterning, including rostrocaudal gradients and laminar components to areal patterning. For example the tan module, reflected a caudal low to rostral high patterning enriched in deep L5 and L6. Another gene module (purple, upper right) had an opposite gradient from high caudal to low rostral, in this case enriched in L3 and L4. Other modules were more area-specific: in V2, MT, DLPFC, and OFC (blue) or lowest in V1, V2, and MT, with enrichment in L2 and L3 (pink).

3.4 Laminar and cellular enrichment patterns of autism genes

Deficits in cortical patterning and layering have been observed in ASD [28], we therefore tested whether ASD-affected genes are enriched in the developing laminae of fetal cortex and the terminally differentiated laminae of adult cortex (Experimental Procedures). We compared multiple ASD gene lists with the ID gene sets for enrichment in laminae of the developing and adult cortex, and found a sharp contrast in laminar enrichment between ASD and ID genes (Figure 3-3A-B). Additionally, in adult, asdM12 exhibits strongly significant enrichment in L3 ($Z > 2.7$, FDR < 0.01), while other ASD lists follow a similar trend of superficial layer enrichment ($Z > 2$, $p < 0.05$). In contrast, the "ID all" and "ID only" gene sets follow a trend of lower layer enrichment (Figure 3-3B), an across-layer pattern that is significantly different from all of the ASD lists (Figure 3-3C-D, Extended Experimental Procedures).

We also observed a similar trend in superficial layer enrichment for the modules that are enriched in asdM12 genes (M13, M16, and M17; Figure 3-3F). M13 and M16 also exhibit weaker enrichment in L5 and L6. Module-level analysis in fetal brain also highlighted a

difference between the RDNV enriched modules, M2 and M3. Although both M2 and M3 are most highly expressed in early human fetal development (prior to PCW 17), M2 reaches its peak later and is enriched in the cortical plate (CPi/CPo), whereas M3 peaks earlier, consistent with its enrichment in the germinal zone (VZ, SZi, SZo; Figure 3-3E). In adult, this distinction is no longer present (Figure 3-3F), with both M2 and M3 showing enrichment in superficial layers (L2, L4). We also asked whether any of these gene sets or modules were enriched for cell-type specific markers paralleling the observed laminar enrichment. We observed enrichment in this set of well-curated upper layer glutamatergic neuron markers among asdM12, M2, and M3 genes (Extended Experimental Procedures, Figure 3-S4C-D), which agrees with the L2-4 enrichment of asdM12 and ASD risk gene modules.

Figure 3-4 highlights adult layer-level expression patterns of several strong ASD candidate genes with enriched expression in superficial layers (e.g. SHANK2, CNTNAP2) and shows that many genes affected by RDNVs recurrently in the 965 ASD probands (e.g. SCN2A, POGZ) also show superficial layer enrichment (Figures 3-4A-B). Although some prenatally expressed genes have low expression levels in adult cortex, we use these mature laminae for cell-marker enrichment analyses because laminar expression patterns are more clearly delineated relative to PCW 15-21 (Figure 3-4A and 4E, Figures 3-S4A-B). Furthermore, neuronal migration in humans persists into the third trimester, and upper layer neuronal identity is not finalized until after PCW 28 [37]. Out of the 6 genes with recurrent RDNVs in probands in which we can detect layer preference, 5 are predominantly expressed in superficial layers in adult. Some of the genes in Figure 4 also show expression in a lower layer (NLGN1, SCN2A, ITPR1, MLL3), though superficial layer enrichment is stronger (larger differential expression t-value in Table 3-S1A).

3.5 Discussion

Our analyses offer a genome-wide neurobiological context to begin to unify the genetics of ASD, providing robust evidence of both molecular pathway and circuit-level convergence. Integration of ASD genes with developmental co-expression networks and laminar expression data connects multiple ASD risk enriched modules to glutamatergic neurons in upper cortical layers (L2-L4), tying ASD risk genes to specific brain circuitry. The observation of convergent biology in ASD stands in striking contrast with ID, which does not show the same level of developmental or anatomical specificity. Laminar enrichment in the "ASD/ID overlap" genes show a similar pattern as the "ASD only" genes (in L2, Figure 3-4B). Therefore disruption in ID genes that also cause ASD likely affects superficial layers compared to disruption in genes causing ID only; our analyses lead to the prediction that specific disruption of cortical-cortical connectivity, for example by targeting upper layer glutamatergic neurons which predominantly comprise inter- and intra-hemispheric projections, is more likely to affect core ASD phenotypes such as social behavior, rather than general intellectual ability alone.

Our analysis further links specific molecules and pathways to the cortical-cortical intra- and inter-hemispheric disconnection that has been hypothesized as a shared circuit-level deficit unifying diverse ASD etiologies [16, 38]. An illustrative example is the disruption of ARID1B, a BAF complex member, which harbors a RDNV and is a hub of M3. Severe mutations in ARID1B cause corpus callosum abnormalities, ID, and ASD [39, 40]. Another BAF complex member, SMARCC2, implicated by RDNVs in probands, controls cortical thickness by repressing the pool of intermediate progenitors, which preferentially contribute to forming cortical layers 2-4 [41], providing another molecular link to inter- and intra-hemispheric connectivity. Other single-gene examples linking M2 and M3 RDNV-affected genes to cortical cytoarchitecture and cortical connectivity exist, but our analysis makes the first systematic

connection between genes disrupted in ASD and this circuit-level disruption. As additional genes in the early fetal co-expression modules are found to harbor recurrent RDNVs, cortical-cortical connectivity will be a valuable phenotype to assess in both animal models and human patients.

3.6 Methods

Developmental expression data

BrainSpan developmental RNA-seq data (obtained from www.brainspan.org) summarized to GENCODE10 [42] gene-level reads per kilobase million mapped reads (RPKM) values were used (Extended Experimental Procedures for data preprocessing, see Table S1D for sample details). Only neocortical regions were used in our analysis and only genes with a normalized RPKM value of 1 in at least one region at one time point for 80% of the available samples were considered expressed.

Weighted Gene Co-expression Network Analysis

We used the R package WGCNA [43] to construct co-expression networks, as previously done [28] and described in detail in Extended Experimental Methods. The modules were characterized using GO Elite to control the network-wide false discovery rate, with all enriched pathways comprising at least 10 genes at $Z > 2$ and FDR < 0.01 [44]. All network plots were constructed using the igraph package in R.

Gene sets

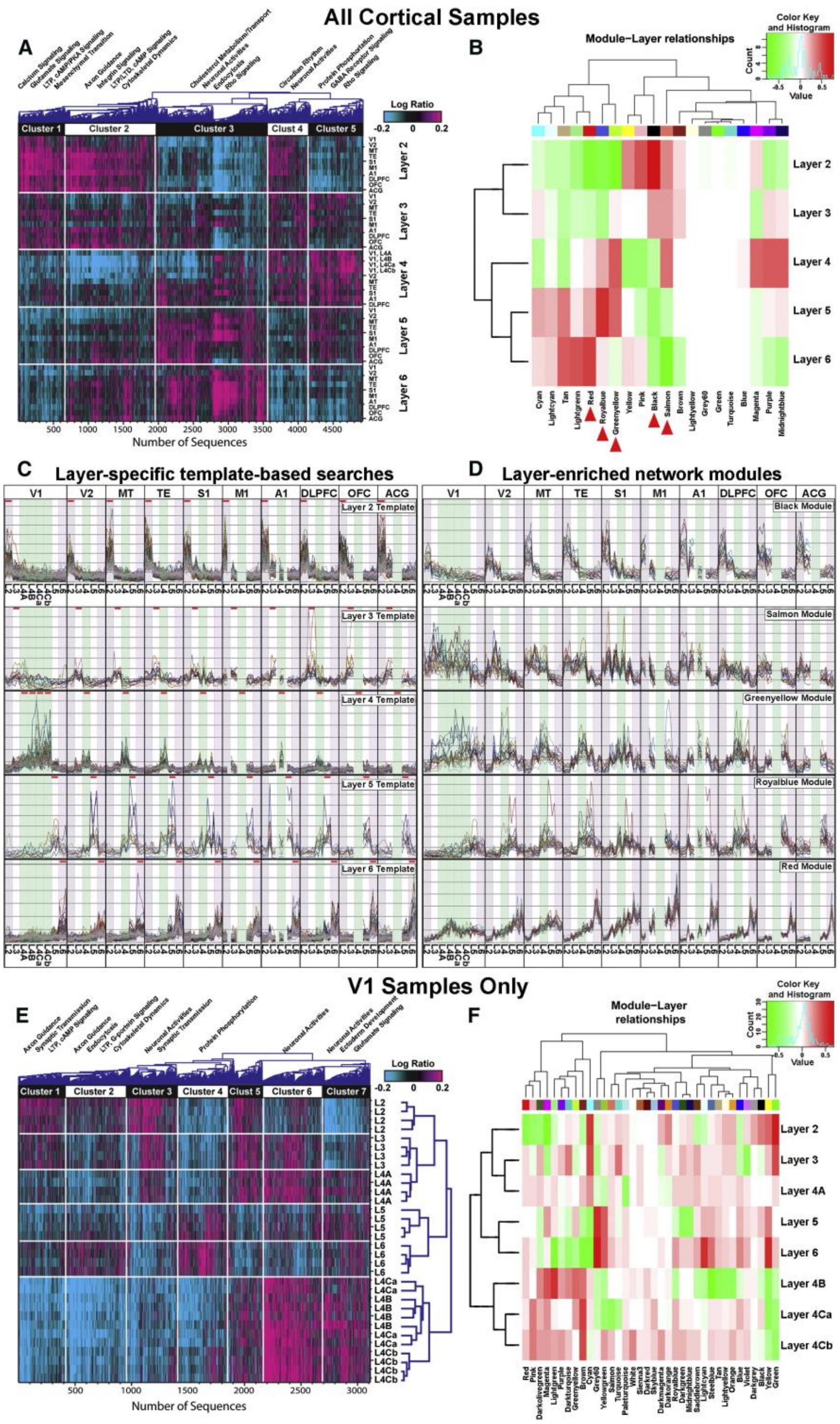
The SFARI ASD set was compiled using the online SFARI gene database, AutDB. We used the "Gene Score," which classifies evidence levels, to restrict our set to those categorized as S (Syndromic) and evidence levels 1-4 (high confidence - minimal evidence). We obtained

asdM12 and adsM16 from a prior, independent gene expression study that profiled expression changes in ASD cortex and applied WGCNA to identify modules of dysregulated genes ASD [28]. We curated ID genes from four reviews cataloging genes causing "ID all" [45-48] resulting in 401 genes. For candidate lists, we used the HUGO gene nomenclature to find updated gene symbols. We obtained RDNVs from four publications [17-20], and split them into discovery and validation sets as discussed in the results (see Extended Experimental Procedures for further details about gene sets).

Layer-specific and Cell-type Marker Enrichment

We utilized human fetal neocortical laminar gene expression datasets from BrainSpan, at PCW 15/16 and PCW 21 and primate neocortical laminar gene expression data from a published study [49]. For laminar specificity, differential expression of each gene in each layer was calculated against background, resulting in t-values for each gene in each layer (Table S1A). We quantified the skew of differential expression t-values of each gene set in each layer, applied a FDR cut-off across all enrichments in all layers (Z = 2.7, FDR = 0.01), and computed bootstrapped confidence intervals to assess enrichment of gene sets in layers. To quantify cell-marker relationships, we used an analogous method, replacing the t-value by the correlation of each gene to a set of known cell marker genes in the adult layer data (Table S1A). Statistical comparison of enrichment trends across layers between ASD and ID gene sets set was performed by comparing the distribution of scores across layers using a permutation analysis (Extended Experimental Methods).*"*

Figure 3-1



All Cortical Samples
A
B
Module−Layer relationships
C
Layer-specific template-based searches
D
Layer-enriched network modules
V1 Samples Only
E
F
Module−Layer relationships
Number of Sequences

Figure 3-1. Robust Transcriptional Signatures of Cortical Laminar Structure.

(A) 1D clustering of genes showing differential laminar expression (ANOVA $p < 10^{-12}$) across all cortical samples, with selected enriched gene sets for each cluster (BF corrected $p < .01$).

(B) Module versus layer relationships based on whole cortex WGCNA, with individual modules showing strong correlations to individual cortical layers (red indicates high correlations). Red arrows under module names indicate modules shown in (D). (C) Identification of genes selectively expressed in specific cortical layers. Displayed are genes with correlation coefficient > 0.7 (L2) or > 0.6 (L3–6) to artificial templates (red bars) in each layer across all cortical areas (see Experimental Procedures). (D) Layer-enriched network modules. Plotted are the top 30 (black, salmon) or top 20 (greenyellow, royalblue, red) hub genes from 5 modules showing different patterns of laminar enrichment. Individual gene profiles in C and D were normalized by the mean expression value for that gene for display on same scale. ANOVA (E) and WGCNA (F) of V1 samples only. (E) 2D clustering of genes showing differential laminar expression among V1 samples (ANOVA, $p < 10^{-3}$), with selected gene set enrichment results (BF corrected $p < 0.1$). (F) Module versus layer relationships based on V1 WGCNA. Dendrograms in (B), (E), (F) show strongest relationships between proximal layers, and a distinct signature associated with the specialized L4 sublayers of V1. Module assignment and gene set enrichment for WGCNA, ANOVA and template analyses are provided in Table S3. Whole-Cortex WCGNA Module Gene Assignment and GO Analysis, Table S4. Gene Set Annotation of Whole-Cortex ANOVA Laminar Gene Clusters, Table S5. Layer-Enriched Gene Sets Derived from

Correlations to Artificial Layer-Specific Gene Templates, Table S6. Gene Set Annotation of V1 ANOVA Laminar Gene Clusters, Table S7. V1 WCGNA Module Gene Assignment and GO Analysis and Table S8. Comparison of V1 ANOVA Laminar Genes with WGCNA Whole-Cortex Modules and Layer-Enriched Genes Derived from Correlation to Artificial Gene Templates.

Figure 3-2

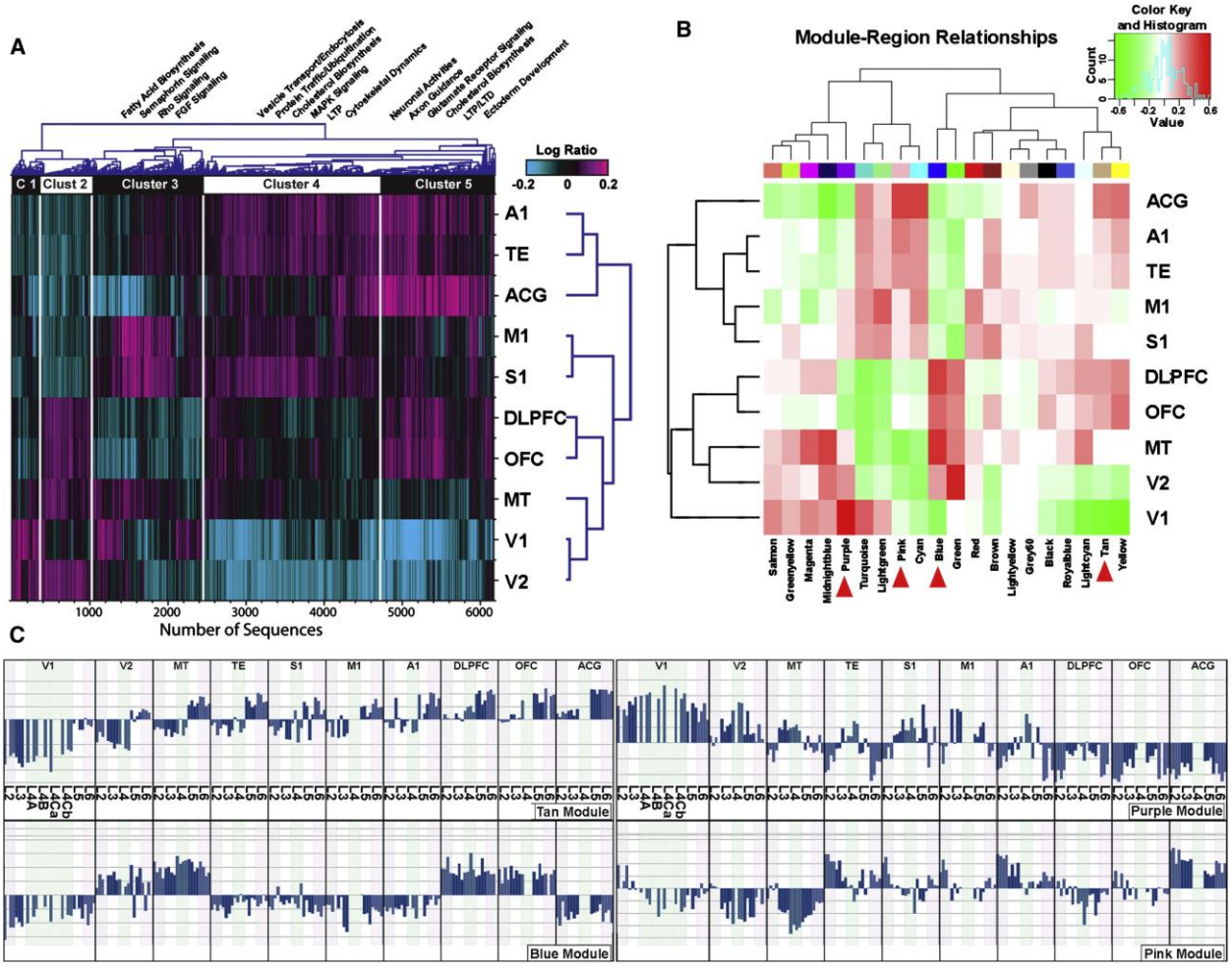

A
Fatty Acid Biosynthesis
Semaphorin Signaling
Rho Signaling
FGF Signaling
Vesicle Transport/Endocytosis
Protein Tariff/Ubiquitination
Cholesterol Biosynthesis
MAPK Signaling
LTP
Cytoskeletal Dynamics
Neuronal Activities
Axon Guidance
Glutamate Receptor Signaling
Cholesterol Biosynthesis
LTP/LTD
Ectoderm Development
Log Ratio
-0.2 0 0.2
C 1
Clust 2
Cluster 3
Cluster 4
Cluster 5
A1
TE
ACG
M1
S1
DLPFC
OFC
MT
V1
V2
1000 2000 3000 4000 5000 6000
Number of Sequences
B
Module-Region Relationships
Color Key and Histogram
Count
Value
ACG
A1
TE
M1
S1
DLPFC
OFC
MT
V2
V1
Salmon
Greenyellow
Magenta
Midnightblue
Purple
Turquoise
Lightgreen
Pink
Cyan
Blue
Green
Red
Brown
Lightyellow
Grey60
Black
Royalblue
Lightcyan
Tan
Yellow
C
V1 V2 MT TE S1 M1 A1 DLPFC OFC ACG
Tan Module
Purple Module
Blue Module
Pink Module

Figure 3-2. Molecular Signatures of Cortical Regions.

(A) 2D cluster of genes differentially expressed between cortical regions (ANOVA, $p < 10^{-12}$), with selected enriched gene sets for specific clusters (BF corrected $p < 0.01$). (B) Module versus region relationships based on whole cortex WGCNA. Individual modules show strong correlations to subsets of cortical regions, and proximal regions show the strongest similarity. (C) Module eigengene plots for tan (upper left, caudal to rostral high gradient), purple (upper right, rostral to caudal high gradient), blue (lower left, V2, MT, DLPFC, OFC high), and pink (TE, A1, ACG high) modules. Further ANOVA is provided in Table S9 and Figure S3.

Figure 3-3

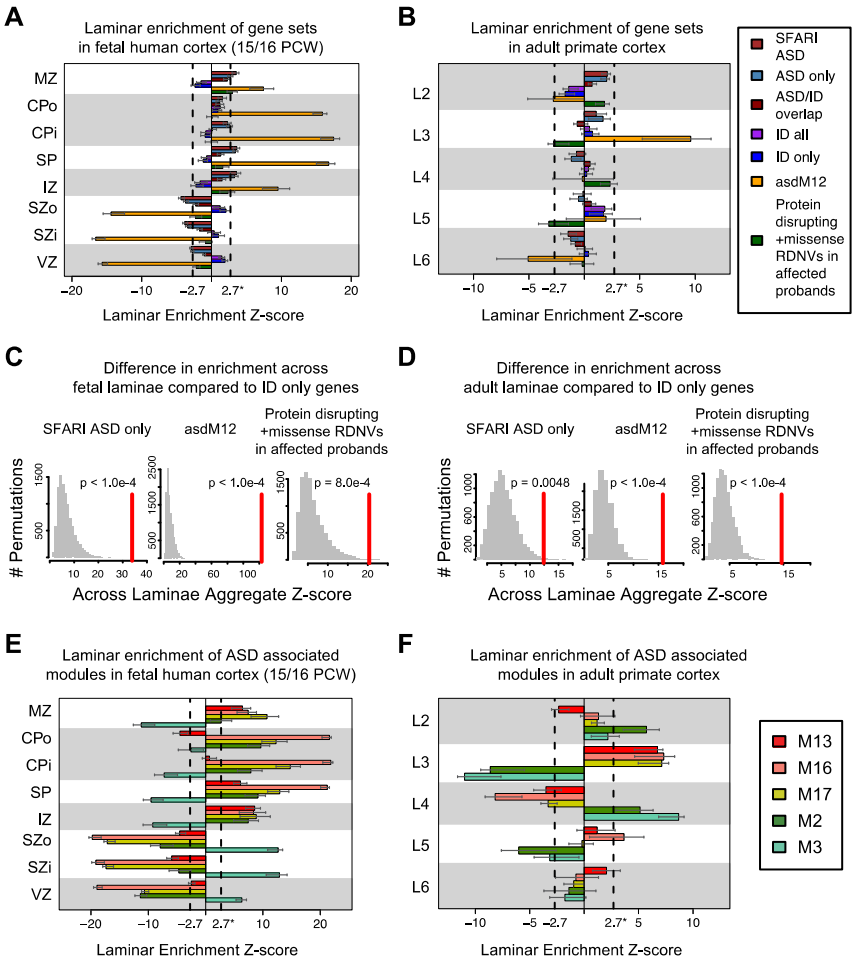

A
Laminar enrichment of gene sets in fetal human cortex (15/16 PCW)
MZ
CPo
CPi
SP
IZ
SZo
SZi
VZ
−20 −10 −2.7 2.7* 10 20
Laminar Enrichment Z-score
B
Laminar enrichment of gene sets in adult primate cortex
L2
L3
L4
L5
L6
−10 −5 −2.7 2.7* 5 10
Laminar Enrichment Z-score
SFARI ASD
ASD only
ASD/ID overlap
ID all
ID only
asdM12
Protein disrupting +missense RDNVs in affected probands
C
Difference in enrichment across fetal laminae compared to ID only genes
SFARI ASD only
asdM12
Protein disrupting +missense RDNVs in affected probands
Permutations
p < 1.0e-4
p < 1.0e-4
p = 8.0e-4
Across Laminae Aggregate Z-score
D
Difference in enrichment across adult laminae compared to ID only genes
SFARI ASD only
asdM12
Protein disrupting +missense RDNVs in affected probands
# Permutations
p = 0.0048
p < 1.0e-4
p < 1.0e-4
Across Laminae Aggregate Z-score
E
Laminar enrichment of ASD associated modules in fetal human cortex (15/16 PCW)
MZ
CPo
CPi
SP
IZ
SZo
SZi
VZ
−20 −10 −2.7 2.7* 10 20
Laminar Enrichment Z-score
F
Laminar enrichment of ASD associated modules in adult primate cortex
L2
L3
L4
L5
L6
−10 −5 −2.7 2.7* 5 10
Laminar Enrichment Z-score
M13
M16
M17
M2
M3

Figure 3-3. Enrichment for laminar differential expression of gene sets and associated developmental co-expression modules in fetal human and adult primate cortex.

(A) In fetal cortex, ASD sets (SFARI, asdM12, and RDNV-affected) are enriched for differential expression in laminae containing post mitotic neurons, whereas genes implicated in ID are weakly enriched in germinal layers. A high Z-score for a gene set in a layer corresponds to differential expression across the gene set in that layer. (B) In adult cortex, asdM12 sets show strong enrichment in layer 3, whereas ID genes are weakly enriched in layer 5. (C) and (D) Summing the Z-score across layers in A) and B) and comparing to randomly permuted sets of genes of similar size demonstrates that, in both fetal and adult cortex, the laminar distribution of multiple ASD implicated gene sets is significantly distinct from that of genes implicated only in ID. (E) SFARI/asdM12 associated developmental co-expression modules M13, M16, and M17 follow enrichment trends similar to the SFARI/asdM12 gene set in fetal brain. However, the modules strongly associated with the RDNV affected genes, M2 and M3, show distinct enrichment patterns. (F) ASD-associated modules are predominantly enriched in superficial layers 2-4 of adult cortex. Additionally, M16 shows weak enrichment in L5. In contrast to fetal cortex, M2 and M3 are in enriched in the same laminae in adult suggesting they serve distinct functions during cortical development that contribute to superficial cortical layers 2-4. Dashed lines in bar plots indicate Z = 2.7 (equivalent to FDR = 0.01), error bars indicate 95% bootstrapped CIs. Laminae: Marginal Zone (MZ), Outer/Inner Cortical Plate (CPo/CPi), Subplate (SP), Intermediate Zone (IZ), Outer/Inner Subventricular Zone (SZo/SZi), Ventricular Zone (VZ), and adult cortical layers 2-6 (L2-6). See also Figure S4.

Figure 3-4

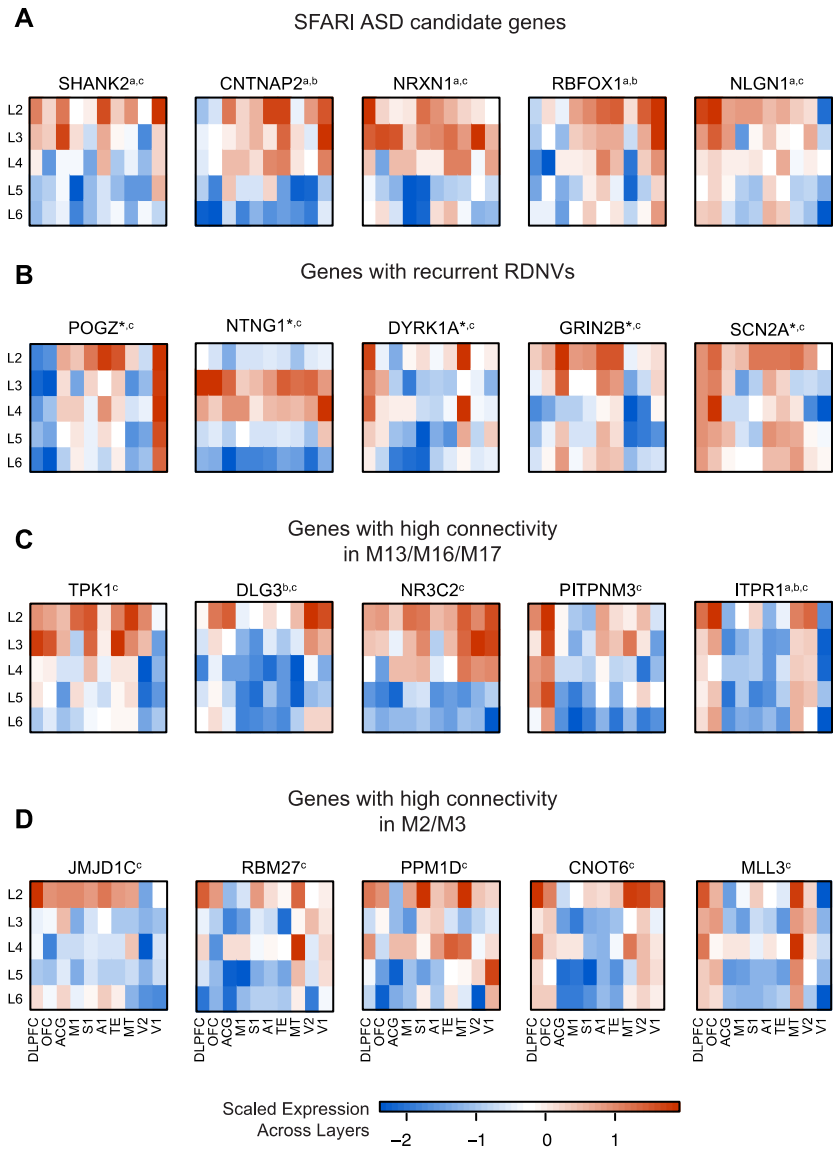

A
SFARI ASD candidate genes
SHANK2a,c
CNTNAP2a,b
NRXN1a,c
RBFOX1a,b
NLGN1a,c
B
Genes with recurrent RDNVs
POGZ*,c
NTNG1*,c
DYRK1A*,c
GRIN2B*,c
SCN2A*,c
C
Genes with high connectivity in M13/M16/M17
TPK1c
DLG3b,c
NR3C2c
PITPNM3c
ITPR1a,b,c
D
Genes with high connectivity in M2/M3
JMJD1Cc
RBM27c
PPM1Dc
CNOT6c
MLL3c
L2
L3
L4
L5
L6
DLPFC
OFC
ACG
M1
S1
A1
TE
MT
V2
V1
Scaled Expression Across Layers
−2
−1
0
1

Figure 3-4. Laminar patterns for genes implicated in ASD.

(A) SFARI candidate genes for ASD. (C) Genes with strong recurrent RDNV evidence across studies. Genes not displayed include *TBR1* (lower layer enriched), *CHD8* (no layer enrichment detected), *CUL3* (no layer enrichment detected), and *KATNAL2* (not detected in these data). (B) Genes with high connectivity in M13, M16, and M17. (C) RDNV genes with high connectivity in M2 and M3.[a] indicates membership in SFARI ASD, [b] indicates membership in asdM12, [c] indicates the gene is affected by a RDNV, * indicates recurrent RDNVs. Color bar values represent scaled expression (standard deviation of the mean-centered expression value across layers). All genes shown have $t > 2$ for enrichment in an upper layer (L2 or L3) over background, and $t < 2$ for lower layers (L5 or L6). Regions: dorsolateral prefrontal (DLPFC), orbitofrontal (OFC), anterior central gyrus (ACG), primary motor (M1), primary somatosensory (S1), primary auditory (A1), higher-order visual area TE (TE), higher-order visual area MT/5 (MT), secondary visual cortex (V2), primary visual cortex (V1).

All supplementary figures and tables as listed in the following websites:

http://www.cell.com/neuron/retrieve/pii/S0896627312002255

http://www.cell.com/abstract/S0092-8674(13)01349-4

3.6 References

1. Barbier EL, Marrett S, Danek A, Vortmeyer A, van Gelderen P, Duyn J, Bandettini P, Grafman J, Koretsky AP: Imaging cortical anatomy by high-resolution MR at 3.0T: detection of the stripe of Gennari in visual area 17. *Magnetic Resonance in Medicine : official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine* 2002, 48(4):735-738.

2. Zilles K, Amunts K: Centenary of Brodmann's map--conception and fate. *Nature Reviews Neuroscience* 2010, 11(2):139-145.

3. Takahata T, Komatsu Y, Watakabe A, Hashikawa T, Tochitani S, Yamamori T: Differential expression patterns of occ1-related genes in adult monkey visual cortex. *Cerebral Cortex* 2009, 19(8):1937-1951.

4. Zilles K, Palomero-Gallagher N, Schleicher A: Transmitter receptors and functional anatomy of the cerebral cortex. *Journal of Anatomy* 2004, 205(6):417-432.

5. Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, Geschwind DH, Mane SM, State MW, Sestan N: Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 2009, 62(4):494-509.

6. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ *et al*: Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 2007, 445(7124):168-176.

7. Molyneaux BJ, Arlotta P, Menezes JR, Macklis JD: Neuronal subtype specification in the cerebral cortex. *Nature Reviews Neuroscience* 2007, 8(6):427-437.

8. Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, Garcia-Moreno F, Molnar Z, Margulies EH, Ponting CP: A transcriptomic atlas of mouse neocortical layers. *Neuron* 2011, 71(4):605-616.

9. Hoerder-Suabedissen A, Wang WZ, Lee S, Davies KE, Goffinet AM, Rakic S, Parnavelas J, Reim K, Nicolic M, Paulsen O *et al*: Novel markers reveal subpopulations of subplate neurons in the murine cerebral cortex. *Cerebral Cortex* 2009, 19(8):1738-1750.

10. Wang WZ, Oeschger FM, Lee S, Molnar Z: High quality RNA from multiple brain regions simultaneously acquired by laser capture microdissection. *BMC Molecular Biology* 2009, 10:69.

11. Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, Macklis JD: Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* 2005, 45(2):207-221.

12. Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML *et al*: Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell* 2008, 135(4):749-762.

13. Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, Huang ZJ, Nelson SB: Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience* 2006, 9(1):99-107.

14. Rhesus Macaque Genome S, Analysis C, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL *et al*: Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 2007, 316(5822):222-234.

15. Berg JM, Geschwind DH: Autism genetics: searching for specificity and convergence. *Genome Biology* 2012, 13(7):247.

16. Geschwind DH, Levitt P: Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology* 2007, 17(1):103-111.

17. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A *et al*: De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012, 74(2):285-299.

18. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V *et al*: Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012, 485(7397):242-245.

19. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD *et al*: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012, 485(7397):246-250.

20. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL *et al*: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012, 485(7397):237-241.

21. Devlin B, Scherer SW: Genetic architecture in autism spectrum disorder. *Current Opinion in Genetics & Development* 2012, 22(3):229-237.

22. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, Moreno-De-Luca D, Yu TW, Fombonne E, Geschwind D *et al*: Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism* 2012, 3(1):9.

23. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB *et al*: A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012, 335(6070):823-828.

24. Gratten J, Visscher PM, Mowry BJ, Wray NR: Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nature Genetics* 2013, 45(3):234-238.

25. Ben-David E, Shifman S: Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Molecular Psychiatry* 2013, 18(10):1054-1056.

26. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D: Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 2011, 70(5):898-907.

27. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, Hill DE, Zoghbi HY: Protein interactome reveals converging molecular pathways among autism disorders. *Science Translational Medicine* 2011, 3(86):86ra49.

28. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011, 474(7351):380-384.

29. Geschwind DH: Genetics of autism spectrum disorders. *Trends in Cognitive Sciences* 2011, 15(9):409-416.

30. Matson JL, Shoemaker M: Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities* 2009, 30(6):1107-1114.

31. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G *et al*: Spatio-temporal transcriptome of the human brain. *Nature* 2011, 478(7370):483-489.

32. Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 2005, 4:Article17.

33. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH: Functional organization of the transcriptome in human brain. *Nature Neuroscience* 2008, 11(11):1271-1282.

34. Lein ES, Zhao X, Gage FH: Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 2004, 24(15):3879-3889.

35. Campbell DB, Sutcliffe JS, Ebert PJ, Militerni R, Bravaccio C, Trillo S, Elia M, Schneider C, Melmed R, Sacco R *et al*: A genetic variant that disrupts MET transcription is associated with autism. *Proceedings of the National Academy of Sciences of the United States of Americ* 2006, 103(45):16834-16839.

36. Mukamel Z, Konopka G, Wexler E, Osborn GE, Dong H, Bergman MY, Levitt P, Geschwind DH: Regulation of MET by FOXP2, genes implicated in higher cognitive dysfunction and autism risk. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 2011, 31(32):11437-11442.

37. Bystron I, Blakemore C, Rakic P: Development of the human cerebral cortex: Boulder Committee revisited. *Nature Reviews Neuroscience* 2008, 9(2):110-122.

38. Belmonte MK, Allen G, Beckel-Mitchener A, Boulanger LM, Carper RA, Webb SJ: Autism and abnormal development of brain connectivity. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 2004, 24(42):9228-9231.

39. Halgren C, Kjaergaard S, Bak M, Hansen C, El-Schich Z, Anderson CM, Henriksen KF, Hjalgrim H, Kirchhoff M, Bijlsma EK *et al*: Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. *Clinical Genetics* 2012, 82(3):248-255.

40. Santen GW, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, Kant SG, Snoeck IN, Peeters EA, Hilhorst-Hofstee Y *et al*: Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. *Nature Genetics* 2012, 44(4):379-380.

41. Tuoc TC, Boretius S, Sansom SN, Pitulescu ME, Frahm J, Livesey FJ, Stoykova A: Chromatin regulation by BAF170 controls cerebral cortical size and thickness. *Developmental Cell* 2013, 25(3):256-269.

42. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D *et al*: GENCODE: producing a reference annotation for ENCODE. *Genome biology* 2006, 7 Suppl 1:S4 1-9.

43. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, 9:559.

44. Zambon AC, Gaj S, Ho I, Hanspers K, Vranizan K, Evelo CT, Conklin BR, Pico AR, Salomonis N: GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 2012, 28(16):2209-2210.

45. Inlow JK, Restifo LL: Molecular and comparative genetics of mental retardation. *Genetics* 2004, 166(2):835-881.

46. Lubs HA, Stevenson RE, Schwartz CE: Fragile X and X-linked intellectual disability: four decades of discovery. *American Journal of Human Genetics* 2012, 90(4):579-590.

47. Ropers HH: Genetics of intellectual disability. *Current Opinion in Genetics & Development* 2008, 18(3):241-250.

48. van Bokhoven H: Genetic and epigenetic networks in intellectual disabilities. *Annual Review of Genetics* 2011, 45:81-104.

49. Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, Finney EM, McWhorter MM, Serikawa K, Lemon T, Morgan R *et al*: Transcriptional architecture of the primate neocortex. *Neuron* 2012, 73(6):1083-1099.