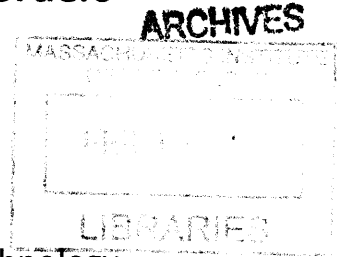


# Small RNA and A-to-I Editing in Autism Spectrum Disorders

by

Alal Eran

B.Sc. Computer Science, Ben Gurion University (2004)



Submitted to the Harvard-MIT Division of Health Sciences and Technology  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN  
BIOINFORMATICS AND INTEGRATIVE GENOMICS  
at the  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
February 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author .....  
Harvard-MIT Division of Health Sciences and Technology  
January 24, 2013

Certified by .....  
Isaac S. Kohane, MD, PhD  
Director, Countway Library of Medicine  
Professor of Pediatrics and Health Sciences Technology, Harvard Medical School  
Thesis Supervisor

Certified by .....  
Louis M. Kunkel, PhD  
Director, Program in Genomics, Children's Hospital Boston  
Professor of Pediatrics and Genetics, Harvard Medical School  
Thesis Supervisor

Accepted by .....  
Emery N. Brown, MD, PhD  
Director, Harvard-MIT Division of Health Sciences and Technology  
Professor of Computational Neuroscience and Health Sciences and Technology



# Small RNA and A-to-I Editing in Autism Spectrum Disorders

by  
Alal Eran

Submitted to the Harvard-MIT Division of Health Sciences and Technology  
on February 4, 2013, in partial fulfillment of the requirements for the degree of  
DOCTOR OF PHILOSOPHY IN BIOINFORMATICS AND INTEGRATIVE GENOMICS

## Abstract

One in every 88 children is diagnosed with Autism Spectrum Disorders (ASDs), a set of neurodevelopmental conditions characterized by social impairments, communication deficits, and repetitive behavior. ASDs have a substantial genetic component, but the specific cause of most cases remains unknown. Understanding gene-environment interactions underlying ASD is essential for improving early diagnosis and identifying critical targets for intervention and prevention.

Towards this goal, we surveyed adenosine-to-inosine (A-to-I) RNA editing in autistic brains. A-to-I editing is an epigenetic mechanism that fine-tunes synaptic function in response to environmental stimuli, shown to modulate complex behavior in animals. We used ultradeep sequencing to quantify A-to-I recoding of candidate synaptic genes in postmortem cerebella from individuals with ASD and neurotypical controls. We found unexpectedly wide distributions of human A-to-I editing levels, whose extremes were consistently populated by individuals with ASD. We correlated A-to-I editing with isoform usage, identified clusters of correlated sites, and examined differential editing patterns. Importantly, we found that individuals with ASD commonly use a dysfunctional form of the editing enzyme ADAR1.

We next profiled small RNAs thought to regulate A-to-I editing, which originate from one of the most commonly altered loci in ASD, 15q11. Deep targeted sequencing of *SNORD115* and *SNORD116* transcripts enabled their high-resolution detection in human brains, and revealed a strong gender bias underlying their expression. The consistent 2-fold upregulation of 15q11 small RNAs in male vs. female cerebella could be important in delineating the role of this locus in ASD, a male dominant disorder.

Overall, these studies provide an accurate population-level view of small RNA and A-to-I editing in human cerebella, and suggest that A-to-I editing of synaptic genes may be informative for assessing the epigenetic risk for autism.

Thesis Supervisor: Isaac S. Kohane, MD, PhD

Title: Director, Countway Library of Medicine, and Professor of Pediatrics and Health Sciences Technology, Harvard Medical School

Thesis Supervisor: Louis M. Kunkel, PhD

Title: Director, Program in Genomics, Children's Hospital Boston, and Professor of Pediatrics and Genetics, Harvard Medical School

## **Acknowledgments**

The work presented here and my ability to conduct it would not have been possible without the endless inspiration, invaluable mentorship, and exceptional training provided by Isaac Kohane and Louis Kunkel. Thanks dads for letting me learn from the best. I would also like to thank my heroes and role models David Bartel and Emery Brown for sharing their brainpower as part of my thesis committee. I thank the families of children with autism that enabled this thesis by donating the brains of their loved ones. I am also grateful for generous financial support from the Nancy Lurie Marks Family Foundation, Jan and Ruby Krouwer, the Simons Foundation, and Roche Diagnostics. Many thanks to my wonderful colleagues, mentors, lab mates, and friends that made work a joyful means to spend 16 hours a day, especially Alvin Kho, Ben Reis, Beth Gilbert, Billy Li, Christin Collins, Daniele Skopek, David Margulies, Dennis Wall, Dick Bennett, Genri Kawahara, Jill McCarthy, Kayla Vatalaro, Laurie Ward, Lily Rodriguez, Luke Hutchison, Marie Boyle, Marielle Thorne, Natassia Vieira, Nathan Palmer, Patty Canningham, Raj Manrai, Shachar Reichman, Stephanie Brewster, Tram Tran, and Yuko Motohashi.

I wouldn't have survived the grad school roller coaster without the best friends in the world Adaya Cohen, Alison Hill, Ami and Hanna Levy-Moonshine, Ashley Bischof, Christine Savage, Clare Poynton, Dana Maor, Efrat Reuven, Elana Erez, Elena Helman, Jill McCarthy, Kay Everett, Kayla Vatalaro, Oded Shaham, and Sarah Calvo. Special thanks to the Ernst brigade - Shoshka, Chaim, Chamoodandoon, Talooosh, Doritkes, and Jakey - my home away from home. Todah rabah to Abba, Ema, Oded, and Savta for always believing in me and giving me the power to never stop exploring. To the love of my life, Amit, we made it! We will never be apart again.

# Contents

<b>Chapter 1</b> Introduction	<b>6</b>
<b>Chapter 2</b> Comparative A-to-I RNA Editing in Autistic and Neurotypical Cerebella	<b>41</b>
<b>Chapter 3</b> Relationships Between A-to-I Editing and Autism-Implicated Small RNA	<b>88</b>
<b>Chapter 4</b> Implications and Future Directions	<b>118</b>
<b>Appendix A</b> <i>Comment on “Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients”</i>	<b>121</b>
<b>Appendix B</b> <i>Haplotype structure enables prioritization of disease markers and candidate genes in autism spectrum disorder</i>	<b>125</b>
<b>Appendix C</b> <i>Whole genome sequencing of six unrelated patients with autism reveals novel candidate genes and pathways affected by rare and nonsynonymous variants</i>	<b>162</b>
<b>Appendix D</b> <i>Distinctive patterns of miRNA expression in primary muscular disorders</i>	<b>191</b>
<b>Appendix E</b> Supplementary material for chapter 2	<b>205</b>
<b>Appendix F</b> Supplementary material for chapter 3	<b>231</b>

# Chapter 1:

## Introduction

Autism spectrum disorders (ASDs) are common neurodevelopmental disorders of complex genetic etiology, characterized by deficits in reciprocal social interaction and repetitive behaviors<sup>1</sup>. Significant recent progress has been made in deciphering the molecular basis of ASD but the cause of the majority of cases remains unknown<sup>2</sup>. Understanding gene-environment interactions that lead to ASD is needed to improve early diagnosis and identify critical targets for intervention and prevention. This thesis explores the potential role of adenosine-to-inosine (A-to-I) RNA editing in autistic brains and its regulation by small RNA.

A-to-I editing is a neurodevelopmentally-regulated transcriptional modification shown to modulate complex behavior in animals<sup>3-9</sup>. Mice and flies with altered editing levels recapitulate several aspects of the human ASD<sup>3, 4, 6, 9, 10</sup>. Recent animal studies demonstrate the causal relationship between editing levels and specific maladaptive behaviors<sup>3-8, 10</sup>, and between specific environmental exposures and editing levels<sup>11, 12</sup>. Thus, investigating the involvement of A-to-I editing in ASD would shed light on the role

of an epigenetic mechanism that fine-tunes neuro-physiological properties in response to environmental stimuli.

Small noncoding RNA have important functions both upstream and downstream of A-to-I editing. The brain-specific small nucleolar RNAs (snoRNA) *mbii-52* were shown to inhibit editing of the serotonin receptor *HTR2C* in vivo<sup>13</sup> and in vitro<sup>14</sup>, via base-pairing to and around the editing sites. microRNAs (miRNA), key regulators of gene expression<sup>15</sup>, are thought to be prime A-to-I editing targets thanks to their double stranded nature<sup>16</sup>, a prerequisite for A-to-I editing<sup>17</sup>.

The genomic technology revolution offers unprecedented opportunities to explore human RNA editing in a cost-effective, large-scale manner. With the availability of well-phenotyped, high quality postmortem brain tissue samples, we were able to begin characterizing the impact, scope, and distribution of A-to-I RNA editing in ASD. In the work presented here, I first examined A-to-I editing which directly alters synaptic function in postmortem cerebella from individuals with ASD and neurotypical controls (Chapter 2), and then investigated its potential regulation by snoRNA transcribed from the genomic region most commonly altered in ASD (Chapter 3). In the near future I will examine the editing and expression levels of miRNA, master regulators of brain development and favorable editing targets (Chapter 4).

In this chapter, I introduce the ASD phenotype and its underlying genomic architecture, discuss the emerging role of synaptic abnormalities in ASD, describe how A-to-I editing may regulate synaptic function and synaptogenesis in response to environmental stimuli, and review our current understanding of how such editing may be regulated by small RNA. Finally, I summarize the goals and organization of this dissertation.

## **Autism is clinically and genetically heterogeneous, with an emerging common theme of synaptic abnormalities**

Autism is not a distinct, categorical disorder, rather a spectrum of social deficits, communication impairments, and repetitive behavior<sup>2</sup>. It is currently estimated that 1 in every 88 children in the United States has autism spectrum disorder (ASD), with boys nearly five times more likely to be affected than girls<sup>18</sup>. Although twin and family studies provide substantial evidence that ASD is one of the most heritable complex disorders<sup>19-21</sup>, the specific variants causing or increasing the risk for ASD remain largely elusive. Recent advances in autism genetics, brought about with the genomics technology revolution, have highlighted its extreme locus heterogeneity, revealing a role for de novo mutations<sup>22-25</sup>, copy-number variants<sup>26-29</sup>, common variants<sup>30, 31</sup>, and rare single nucleotide variants<sup>32, 33</sup>. This has accelerated a growing realization that ASD is comprised of a multitude of etiologies with incompletely overlapping symptomatology and clinical course, which likely converge at the mechanistic or pathway levels<sup>1, 34-37</sup>. Furthermore, recent large scale exome sequencing studies suggest that not only do different individuals with ASD carry different likely-deleterious variants, but a single individual may have multiple different variants in likely candidate genes<sup>22-25</sup>. Therefore, there might exist a spectrum of genetic variants underlying the spectrum of clinical manifestations, making ASD extremely heterogeneous on both the molecular and clinical levels.

### **The ASD phenotype**

Autism, first described by Leo Kanner in 1943<sup>38</sup>, is a complex behavioral disorder defined by developmental deficits in three domains: social interaction, communication, and repetitive stereotypical behavior. ASD includes autism, Asperger syndrome, Rett syndrome and pervasive developmental disorder not otherwise specified (PDD-NOS)<sup>39</sup>.



Behind this definition lies a broad spectrum of disease manifestation, ranging from debilitating impairments to mild personality disorders<sup>40</sup>. Social deficits may include impairment in the use of nonverbal behaviors such as eye gaze, facial expression, body gestures, failure to develop peer relationships, and lack of social reciprocity<sup>39</sup>. Communication impairments range from complete absence of spoken language to a delay in its acquisition. Those that do develop adequate speech exhibit a significant impairment in the ability to initiate or sustain a conversation, as well as stereotyped use of language. Affected individuals also exhibit restricted and stereotyped patterns of behavior, interests, and activities, including abnormal preoccupation with certain activities and adherence to routines or rituals<sup>39</sup>.

ASD occurs in all racial, ethnic and social groups<sup>18</sup>. The dramatic rise in ASD prevalence over the past few years may be attributed to changes in the diagnostic criteria<sup>41</sup>, younger age of diagnosis<sup>42</sup>, and increased awareness of the education systems and medical community<sup>43</sup>. Most in the autism research community have adopted the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview–Revised (ADI-R) to behaviorally phenotype children evaluated for ASD<sup>44</sup>. The ADOS is a standardized protocol for the *observation* of social and communicative behavior of children. The ADI-R is a semi-structured, investigator-administered *interview* for caregivers of affected children and adults. By definition, these instruments are sensitive to the assessor's perception of the child's specific behavior on the day of the interview/observation. Moreover, it is only feasible to examine delay in verbal communication around the age of two, wasting precious time of potential intervention within the sensitive developmental time window. Biomarker identification is therefore a top priority, as it would enable both early and objective diagnoses.

Medications have not been proven to correct the core deficits of ASDs and are not the primary treatment. Early educational intervention is the basis for treating and managing

ASD<sup>45</sup>. Applied behavior analysis, the most common type of intervention, is a learning paradigm characterized by discrete presentation of stimuli with behavioral responses followed by immediate feedback, with intense reinforcement through repeated trials of instruction<sup>46</sup>. Effective behavioral intervention programs throughout the country emphasize the importance of entering the program as early as possible<sup>47</sup>. Therefore, the development of molecular diagnostic approaches has an immediate potential to improve outcomes by enabling early behavioral intervention.

Most children with ASD remain within the spectrum as adults and chronic management is required. There are currently two drugs approved by the Food and Drug Administration (FDA) for treating maladaptive behaviors in ASD. Risperidone, an atypical antipsychotic, was the first medication approved for the symptomatic treatment of aggressive behavior, deliberate self-injury, and temper tantrums in children and adolescents with ASDs<sup>48</sup>. Another atypical antipsychotic, aripiprazole, was approved in 2009 for the treatment of irritability in children and adolescents with autism<sup>49</sup>. Both risperidone<sup>50</sup> and aripiprazole<sup>51</sup> target the serotonin receptor HTR2C, a functional<sup>52, 53</sup> and positional<sup>28, 54</sup> candidate for ASD, whose A-to-I editing will be examined in Chapter 2 and its regulation by small RNA in Chapter 3.

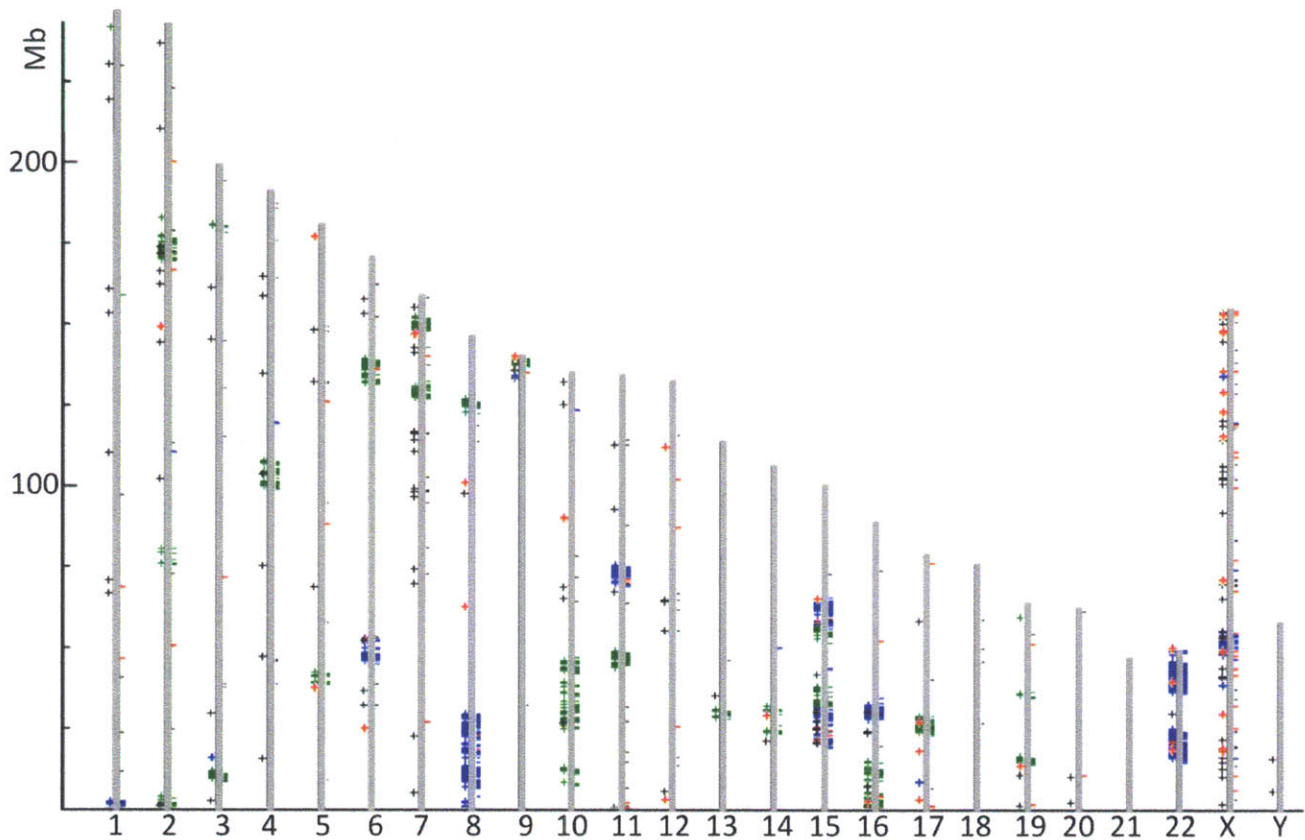
### **Two decades of autism genetics**

Although the underlying cause of ASD remains largely unknown, it is clear that an interaction between genetic predisposition and environmental factors at a sensitive period in neuronal development plays a key role. Infectious, metabolic, and environmental etiologies have been suggested, but evidence from twin and family studies has implicated genetic factors as playing a major role in determining the risk of ASD<sup>55-57</sup>. Twin studies show a concordance of 60%–92% for monozygotic twins and 0–30% for dizygotic pairs, depending on phenotypic definitions<sup>19, 58</sup>. The rate of ASD in

siblings is 9.5-11.5%<sup>59</sup>, much higher than the general population prevalence. Together, these findings have led to heritability estimates of up to 90%, making ASD one of the most heritable complex disorders<sup>60</sup>.

Despite its high heritability, efforts to identify the genetic causes of ASD have enjoyed only limited success. Numerous susceptibility loci have been identified, yet few have been replicated, supporting the notion that the genetic complexity of ASD outmatches the proportion of the autistic population sampled to date. Here I will review central trends in the past two decades of autism genetic research, including family-based linkage studies, genome-wide association studies, copy number variation studies, and the recent convergence on synaptic abnormalities in ASD.

Family-based linkage studies were the predominant approach in autism genetics during the late 1990s and early 2000s. More than 300 linkage studies have been performed, implicating nearly the entire human genome in ASD. Not surprisingly, very few of them have been replicated (**Figure 1**). Linkage studies are powerful tools to identify the cause of disease in a studied family. However, neurodevelopment is a complex mechanism, and an autism-causing mutation in one family may be absent from another, despite the shared phenotype. Genome-wide-significant linkage peaks (LOD > 3.63, MLS > 3.93, or  $Z_{lr} > 4.1$  according to the Lander-Kruglyak standards<sup>61</sup>) were mapped to 1p, 2q, 3q, 7q, 12q, 15q, 17q, and 21q<sup>62-67</sup>. The best-replicated region is 7q<sup>64, 68-71</sup> and as such is the main location of studied candidate genes, including *RELN*, *FOXP2*, *EN2*, *CADPS2*, and *MET*<sup>72-75</sup>. 17q is also among the few replicated regions<sup>63</sup>, and is home to the much-studied *SLC6A4* gene<sup>76</sup> (**Figure 1**).



**Figure 1** Genomic regions related to the etiology of ASD, as of November 2012. Blue marks represent recurrent copy number variation, green marks depict replicated linkage findings, red marks are genes causing syndromes in which ASD is a common symptom, and black marks are genes implicated in ASD with high confidence, using multiple lines of evidence, including segregation patterns, expression profiles, and functional studies. Forward strand genes are marked on the left of each chromosome with a plus sign, reverse genes are noted on the right by a minus sign. Data was obtained from AutismKB<sup>77</sup>.

Then came genome-wide association studies (GWAS). GWAS have dominated human genetics between 2005-2011, and were first applied to ASD in 2008<sup>78</sup>. Their main importance is in the rejection of the *common disease common variant* hypothesis, and the realization that *complex diseases* are indeed complex, and are likely caused by the

interaction of multiple genes with environmental factors. The most successful autism GWAS associated both deletions and duplications in 16p11.2 with autism<sup>78</sup> (**Figure 1**). 16p11 is spanned by segmental duplications and therefore harbors common structural variation in humans. Although many large deletion and duplication syndromes include autism as a comorbidity<sup>79</sup>, individuals with either 16p11 deletion syndrome or 16p11 duplication syndrome tend to be behaviorally typical (MIM IDs #613604, #614671). Genotyping more than 5000 individuals with ASD<sup>30, 31, 80, 81</sup> has associated only one more common variant with the disorder<sup>31</sup>, with an odds ratio of 1.19, supporting the genomic heterogeneity of ASD.

One of the most influential studies in modern autism genetic research was Michael Wigler and Jonathan Sebat's 2007 report of large copy number variants (CNVs) associated with "sporadic"<sup>a</sup> but not "familial" autism<sup>82</sup>. Their study initiated two important developments that have since dominated the field: (i) the notion of two independently transmitted forms of ASD, and (ii) an increased focus on the contribution of CNVs to the ASD phenotype.

**Sporadic vs. familial ASD.** Despite a rich history of twin and family studies supporting a substantial heritable component underlying ASD, the hypothesis that there could exist an independent, mostly environmentally-driven form of autism has led to the establishment of the Simons Simplex Collection (SSC), a heavily funded consortium that recruited<sup>83</sup>, phenotyped<sup>84</sup>, genotyped<sup>26, 27, 29, 85</sup>, and is currently sequencing<sup>22, 24, 25, 86</sup> ~10,000 individuals from families with only one affected child around the United States. According to the new model<sup>87</sup>, a fraction of ASD could be caused by de novo mutations in the parental germline with significantly reduced penetrance in female offspring compared to males. This model is supported by a paternal age effect in autism<sup>88</sup>, and fits

---

<sup>a</sup> Sporadic/simplex autism is defined as having only one affected proband in the family, whereas familial/multiplex autism refers to having multiple affected offspring.

the discrepancy in concordance rates between monozygotic and dizygotic twins<sup>19</sup>. Furthermore, it is a conveniently testable model using modern genomic technologies.

Consequently, several large-scale genome-wide studies have recently characterized the patterns of de novo CNVs<sup>26, 27, 29, 85</sup> (see below) and single nucleotide variants<sup>22-25, 86</sup> (SNVs) in “sporadic autism” SSC families. Collectively, these studies support only a modest role for de-novo mutations in “sporadic” ASD. Even if each likely deleterious de novo mutation were to cause ASD, the overall contribution of de novo mutations in protein-coding regions of the genome is currently limited to ~20% of SSC cases<sup>22-27, 29, 85</sup>. In fact, the rate of de novo mutations in large cohorts of simplex families was shown to be similar to that of multiplex families<sup>36</sup> and to the expected background distribution<sup>23</sup>. It seems that de novo mutations might be just one of several types of genomic variation that increase the risk for one of the most heritable complex traits. Aggressive sequencing efforts currently under way should provide sufficient power to fully confirm or reject the Wigler-Sebat conjecture in the next couple of years.

**CNVs in ASD.** The combination of the Wigler-Sebat report<sup>82</sup> with the ability to measure CNVs using SNP chips<sup>89</sup> has dusted off past cytogenetic findings in ASD and pushed CNV studies to center stage. Historically, since the development of karyotyping techniques in the 1960s it was clear that large chromosomal abnormalities typically result in developmental delay and intellectual disabilities<sup>90</sup>. With gradual improvements in the detection resolution, first via fluorescent in situ hybridization (FISH)<sup>91</sup> and then with array-based comparative genomic hybridization (aCGH)<sup>92</sup>, it became evident that even smaller deletions and duplications are significantly associated with ASD<sup>79</sup>. Recent large scale CNV studies in SSC families have revealed a significantly higher incidence of rare de novo but not inherited CNVs in probands as compared to their unaffected sibs<sup>26, 27, 29, 85</sup>. In case-control studies, rare genic CNVs but not all CNVs were associated with ASD<sup>36, 80</sup>. These and other findings have contributed to the realization that each CNV per se is neither sufficient nor necessary to cause ASD, and that ASD is a lot more complex than

previously thought<sup>35</sup>. Because ASD is a common disorder with a shared phenotype, individually rare etiologies must converge at some level. One common theme among the numerous molecular findings in ASD, including CNVs<sup>26, 36</sup>, is that many different mutations may result in shared synaptic alterations.

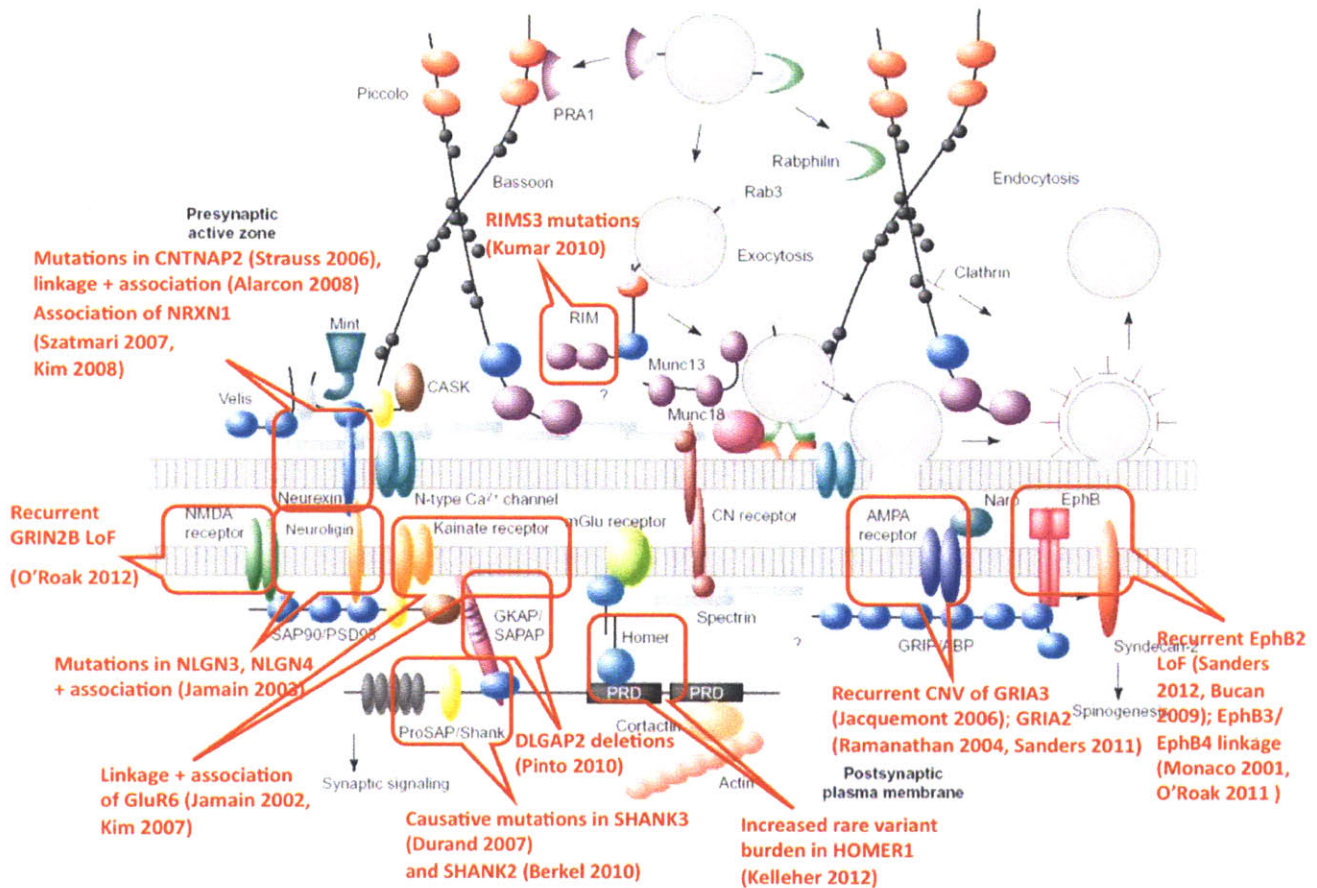
### **Synaptic alterations in ASD**

The divergent concordance rates strongly suggest that ASD is not a single-gene Mendelian disease, and several multi-locus models have been proposed<sup>65, 93, 94</sup>. As shown in **Figure 1**, numerous loci have been implicated in ASD, each altered in a small proportion of families, and causing disease in combination with other genomic and environmental factors. The current paradigm is that these most likely converge at the pathway or mechanism level across individuals, giving rise to the high prevalence of ASD. An open question is what is the functional relationship between bona fide autism genes?

One approach initiated by Thomas Bourgeron's group is that many autism genes are involved in synapse formation and synaptic homeostasis<sup>1, 26, 37</sup>. Their 2003 report of causative mutations in the X-linked neuroligins *NLGN3* and *NLGN4* suggested that an overall defect in synaptogenesis may predispose to ASD<sup>95</sup> (**Figure 2**). The previous year they reported suggestive linkage and association of *GRIK2*<sup>96</sup>, a glutamate receptor expressed during brain development and modified by A-to-I editing<sup>97</sup>. In 2007 they identified causative mutations in *SHANK3*, a protein that regulates the structural organization of dendritic spines by binding to neuroligins<sup>98</sup> (**Figure 2**). *SHANK3* mutations were found in <1% of individuals with ASD, but this functional convergence was sufficient to influence the Autism Genome Project (AGP) to search hard for association and linkage to *SHANK3* and other neuroligin binding partners<sup>99</sup>. Following a report of causative mutations in *CNTNAP2*, a neurexin gene<sup>100</sup>, the AGP reported biased

transmission of neurexin 1 (*NRXN1*) SNPs in a genome wide scan of 1400 families<sup>99</sup>. Together, these findings started an avalanche of supportive interpretation of genome-wide scans<sup>22-26, 36, 54, 80, 101, 102</sup> and the development of mouse models of synaptic protein loss of function which demonstrated “autistic like features”, including *SHANK1*<sup>103</sup>, *SHANK2*<sup>104</sup>, *SHANK3*<sup>105</sup>, *NLGN1*<sup>106</sup>, *NLGN3*<sup>107</sup> and *NLGN4*<sup>108</sup>. **Figure 2** visualizes the functional relationships between these synaptic genes, highlighting related interacting proteins that have since been implicated in the disorder. It is still unknown to what extent mutations in these genes predispose to ASD and in which context. Hence, identifying mechanisms of gene-environment interactions and their contribution to the observed synaptic alterations could be highly informative for risk assessment<sup>109</sup>. A-to-I RNA editing is potentially one such mechanism, linking environmental stimuli with synaptic transmission.





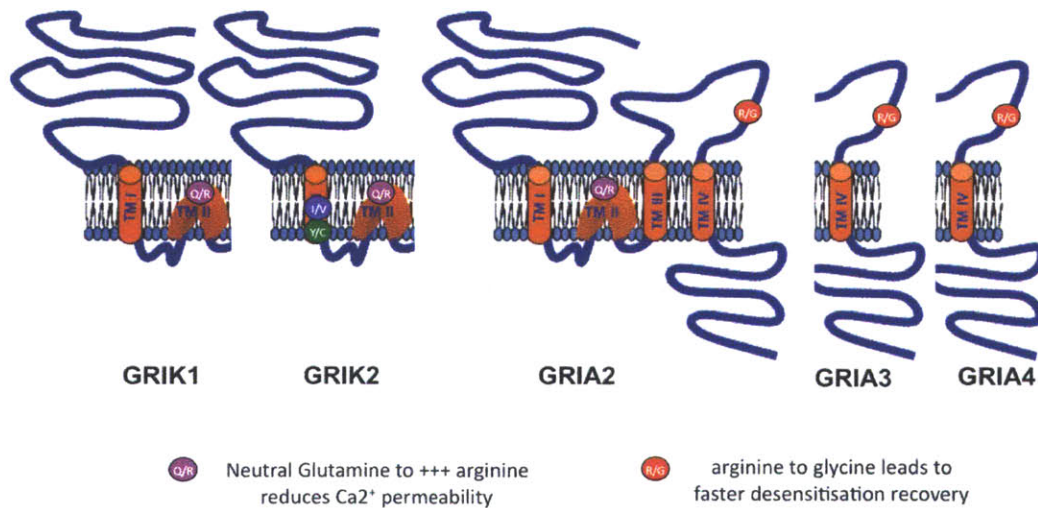
**Figure 2** Synaptic genes confidently implicated in ASD using multiple lines of evidence. Cell adhesion molecules such as cadherins (CDH), protocadherins (PCDH), neuroligins (NLGN1-4) and neurexins (NRXN1-3, CNTNAP2) are involved in synaptic recognition and assembly, and were implicated in ASD via genome-wide linkage and association studies, direct resequencing, and functional studies in mouse models. Within the postsynaptic density, scaffold proteins such as SHANK1-3 and DLGAP2 assemble the postsynaptic signaling complex and connect between membrane proteins and the actin cytoskeleton. Glutamate receptors, including NMDA, AMPA, and kainate receptors, play a central role in producing excitatory currents and maintaining synaptic plasticity. See figure for the specific genomic approaches taken to implicate these in ASD. Abbreviations: LoF, loss of function.

## **A-to-I RNA Editing fine-tunes synaptic function in response to environmental stimuli**

Surprisingly little is known about A-to-I RNA editing in humans, despite its essential roles. In animals, site-specific adenosine-to-inosine RNA base conversions, carried out by adenosine deaminase acting on RNA (ADAR) enzymes, modulate complex behavior and exhibit precise regional specificity in the brain<sup>5, 8, 110</sup>. A-to-I RNA editing is an efficient means to increase RNA complexity, thereby fine-tuning both gene function and dosage<sup>111-113</sup>. The cellular machinery recognizes inosine as guanosine, so A-to-I editing of codons and splicing signals directly modifies protein-coding gene function<sup>113-118</sup>, while editing of microRNAs<sup>119-122</sup> and their binding sites<sup>123</sup> alters gene expression. This should be particularly important in the human brain, the single most complex organ in cellular diversity, connectivity, morphogenesis, and responses to environmental stimuli<sup>124</sup>.

Synaptic function is a major target of A-to-I editing<sup>125</sup>, which can fine-tune neurophysiological properties in response to environmental stimuli<sup>126</sup>. Canonical signaling pathways acting on the editing enzymes link A-to-I RNA editing to environmental cues: ADARB1 function requires inositol hexakisphosphate<sup>127</sup>, and the expression of ADAR is interferon-inducible<sup>128</sup>. Several recoding events that directly alter synaptic strength or duration in response to environmental signals have been characterized in rodents<sup>114-118</sup>. For example, the serotonin receptor *HTR2C* undergoes editing in five sites, which dramatically alters its G-protein coupling activity, and hence the relationship between serotonin levels and postsynaptic signal transduction<sup>5, 118</sup>. This editing is regulated by exposure to acute stress and chronic treatment with antidepressants<sup>12</sup>. Another example is the neurodevelopmentally-regulated editing of transcripts encoding the AMPA receptors GRIA2, GRIA3 and GRIA4<sup>129</sup>, where arginine to glycine (R/G) recoding of the ligand binding domains leads to faster desensitization recovery<sup>114</sup>. Moreover, glutamine to arginine (Q/R) editing of the transmembrane

domains in mRNAs encoding the kainate receptors GRIK1 and GRIK2 reduces the receptors' calcium permeability<sup>115</sup>, with varying degrees of editing throughout mouse neurodevelopment<sup>129</sup>. Since 0 to 100% of mRNA molecules can be edited at any given point<sup>130</sup>, Q/R and R/G editing of ionotropic glutamate receptor transcripts provides an efficient means for fine-tuning the glutamatergic synapse in response to the changing environment (**Figure 3**).



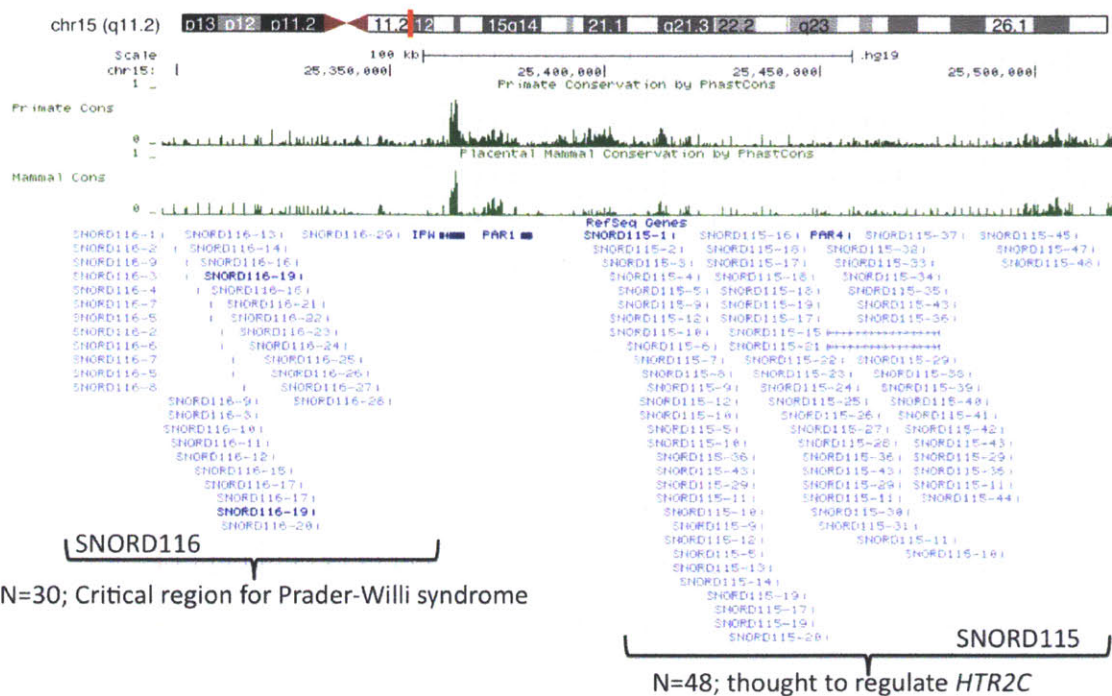
**Figure 3** Ionotropic glutamate receptors regulated by A-to-I RNA recoding. Three AMPA receptors and two kainate glutamate receptors are subject to A-to-I RNA editing. Q/R editing of the transmembrane domains of GRIK1, GRIK2, and GRIA2 dramatically alters their calcium permeability, while R/G recoding of the intracellular domain of GRIA2, GRIA3, and GRIA4 tunes their kinetics. A-to-I editing of dendritic RNA granules can therefore provide an incredibly efficient means for rapid fine-tuning of synaptic strength and duration in response to environmental cues, contributing to synaptic plasticity.

Although few A-to-I RNA editing studies have been conducted in humans, it has been postulated to be one of the molecular mechanisms connecting environmental inputs and behavioral outputs<sup>126, 131</sup>. The increased editing in humans<sup>132</sup> as compared to other animals<sup>133</sup>, including nonhuman primates<sup>134</sup>, has been proposed to generate molecular complexity that might constitute the basis of higher-order cognition<sup>126</sup>. Hence, characterizing A-to-I editing in typically and atypically developed individuals, such as those with ASD, may shed light on environment-dependent epigenetic mechanisms central to human neurodevelopment.

Several lines of evidence support an examination of the link between A-to-I editing and ASD. First, model organisms with altered A-to-I editing exhibit maladaptive behaviors characteristic of ASD<sup>5, 110, 135</sup>, sometimes with seizures<sup>110</sup> or Prader-Willi-like symptoms<sup>6</sup>, both of which are typically detected in 25% of children with ASD<sup>136, 137</sup>. Second, altered editing of mRNAs encoding the serotonin receptor HTR2C has been detected in a mouse model of autism<sup>4</sup> and in disorders that aggregate in families with ASD<sup>138</sup>, including schizophrenia<sup>139</sup> and major depression<sup>140</sup>. Third, a fly model of Fragile X syndrome, the most common single gene cause of ASD<sup>141</sup>, was recently shown to exhibit significant editing alterations, via a direct interaction between the Fragile X gene dFMR1 and the editing enzyme dADAR<sup>9</sup>. Finally, independent genomic studies have implicated variants in synaptic genes, the most edited type of genes<sup>125</sup>, as a recurring theme in ASD<sup>1, 26, 34, 37</sup>.

## **The imprinted brain-specific 15q11 snoRNAs may guide A-to-I mRNA editing**

The most common genetic lesion in ASD is copy number variation of the imprinted 15q11–13 locus<sup>4</sup>. Two small nucleolar RNA (snoRNA) clusters are transcribed from its paternal allele, *SNORD115* (*HBII-52*) and *SNORD116* (*HBII-85*), both highly expressed in the brain<sup>142</sup> (**Figure 4**). While the precise function of these snoRNAs remains to be elucidated, the mouse homolog of *SNORD115* was shown to regulate A-to-I RNA editing of the serotonin receptor Htr2c<sup>13, 14</sup>, a functional<sup>143</sup> and positional<sup>144</sup> candidate for ASD. Mice lacking *Snord115* show serotonin-dependent behavioral changes and Htr2c editing alterations<sup>13</sup>, while mice heterozygous for a paternal duplication of 15q11-13 display social abnormalities and altered Htr2c editing<sup>4</sup>. As to *SNORD116*, its absence was shown to be the cause of the key characteristics of the Prader-Willi phenotype<sup>145</sup>, which includes ASD in 23% of cases<sup>146</sup>. Members of the *SNORD116* cluster are predicted to regulate the alternative splicing of several neuronal genes<sup>147</sup>.



**Figure 4** Small RNA clusters expressed from the paternal allele of 15q11. The *SNORD115* cluster contains 48 paralogous genes thought to regulate A-to-I editing of the serotonin receptor *HTR2C*. Immediately upstream is the *SNORD116* cluster, which contains 30 genes thought to regulate the alternative splicing of several neuronal genes. Deletion of *SNORD116* is sufficient to cause the core symptoms of Prader-Willi syndrome, a neurodevelopmental condition in which a quarter of the patients also fall within the autism spectrum.

### Small RNA and A-to-I editing in Mendelian traits comorbid with ASD

Mendelian (i.e. single gene) causes of ASD are rare and collectively account for up to 10% of ASD<sup>2</sup>, but may be highly informative of the underlying molecular mechanisms.

**Fragile X Syndrome (FXS)** is the most common known monogenic form of autism, with about a third of FXS patients exhibiting comorbid ASD<sup>148</sup>, and roughly 5% of individuals

with ASD having FXS<sup>149</sup>. Absence of the Fragile X Mental Retardation Protein (FMRP) causes FXS, a severe form of intellectual disability<sup>9</sup>. FMRP is an RNA binding protein (RBP) that regulates dendritic protein synthesis, with essential roles at the synapse, as evidenced by its severe loss of function phenotype in humans, and the altered neuronal development and circuit formation shown in knockout mice and flies<sup>150</sup>. FMRP associates with many RNA granules<sup>151</sup>, several members of the RNAi machinery<sup>152, 153</sup>, including the RNA induced silencing complex (RISC)<sup>154</sup>, synaptic miRNA<sup>155</sup>, and locally-translating polyribosomes in dendrites<sup>153, 156</sup>. Furthermore, it was recently shown that the fly FMRP, dFMRP, modulates A-to-I editing levels via direct interaction with dADAR<sup>9</sup>. The exact relationship between loss of FMRP function and the FXS and ASD phenotypes remains to be elucidated. The involvement of FMRP in the RNAi and RNA editing pathways suggests that examining small RNA and editing in ASD could advance our understanding of its underlying molecular basis.

**Prader Willi Syndrome (PWS)** is a neurodevelopmental disorder due to the loss of paternally expressed genes from the imprinted 15q11-13 region, with 23% of patients exhibiting comorbid ASD<sup>146</sup>. The PWS locus contains the SNORD115 and SNORD116 snoRNAs. It is currently thought that the key features of PWS are caused by the absence of the SNORD116 cluster, with other genes in the region making more subtle phenotypic contributions<sup>145</sup>. A mouse model of PWS shows altered A-to-I editing of HTR2C, linked to mbii-52 loss<sup>13</sup> (the mouse homolog of SNORD116).

## **Dissertation goals and organization**

My dissertation seeks to examine A-to-I RNA editing in autistic brains and its potential regulation by small RNA. Chapter 2 describes several years of work in which I set up a system to accurately survey A-to-I RNA editing levels in postmortem brain tissue from individuals with ASD. Using ultradeep cDNA pyrosequencing coupled to gDNA genotyping we carried out the first quantitation of A-to-I RNA editing levels in ASD and found that when compared to matched neurotypical individuals, some individuals with ASD show consistently exaggerated levels of synaptic A-to-I RNA editing. As part of this study, we examined the editing patterns of the serotonin receptor *HTR2C*, a central pharmacological target in ASD, whose potential regulation by small RNA is the subject of Chapter 3. This chapter describes a targeted small RNA sequencing approach used to quantify the relationships between the expression of imprinted brain-specific small nucleolar RNA genes from an autism-implicated locus and A-to-I editing of *HTR2C*. Chapter 4 describes the implications of these studies and others that I plan to complete in the near future, with an overarching goal of understanding RNA-mediated gene-environment interactions underlying ASD, and more generally complex human behavior in health and disease.

The appendices include related research that set the foundations for the experimental and computational techniques I have utilized in my main projects. Appendix A contains a publication in which I had examined the alternative splicing of an ASD candidate gene, *CADPS2*, in brain and blood of individuals with ASD. I later used a similar approach to profile alternative splicing of the editing enzyme ADARB1 in Chapter 2. Appendix B contains a publication in which I had reviewed the rich autism literature which helped me understand the historical and social context of my work, as well as write Chapter 1. Appendix C contains a manuscript in which I had designed and carried out large scale genotyping to validate whole genome sequencing of individuals with ASD. This work utilized an experimental approach designed in Chapter 2, and taught me the great



advantages of approaching the problem at the RNA level. Appendix D contains a publication in which I developed a meta-analysis approach to infer microRNA function. I will use this approach in a follow up study described in Chapter 4. Finally, Appendix E contains supplementary material for Chapter 2, and Appendix F contains supplementary material for Chapter 3.

## References

1. Toro R, Konyukh M, Delorme R, Leblond C, Chaste P, Fauchereau F *et al.* Key role for gene dosage and synaptic homeostasis in autism spectrum disorders. *Trends Genet* 2010; **26**(8): 363-372.
2. Geschwind DH. Genetics of autism spectrum disorders. *Trends Cogn Sci* 2011; **15**(9): 409-416.
3. Jepson JE, Savva YA, Yokose C, Sugden AU, Sahin A, Reenan RA. Engineered alterations in RNA editing modulate complex behavior in *Drosophila*: regulatory diversity of adenosine deaminase acting on RNA (ADAR) targets. *J Biol Chem* 2011; **286**(10): 8325-8337.
4. Nakatani J, Tamada K, Hatanaka F, Ise S, Ohta H, Inoue K *et al.* Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism. *Cell* 2009; **137**(7): 1235-1246.
5. Mombereau C, Kawahara Y, Gundersen BB, Nishikura K, Blendy JA. Functional relevance of serotonin 2C receptor mRNA editing in antidepressant- and anxiety-like behaviors. *Neuropharmacology* 2010; **59**(6): 468-473.
6. Morabito MV, Abbas AI, Hood JL, Kesterson RA, Jacobs MM, Kump DS *et al.* Mice with altered serotonin 2C receptor RNA editing display characteristics of Prader-Willi syndrome. *Neurobiol Dis* 2010; **39**(2): 169-180.
7. Dracheva S, Lyddon R, Barley K, Marcus SM, Hurd YL, Byne WM. Editing of serotonin 2C receptor mRNA in the prefrontal cortex characterizes high-novelty

- locomotor response behavioral trait. *Neuropsychopharmacology* 2009; **34**(10): 2237-2251.
8. Tonkin LA, Saccomanno L, Morse DP, Brodigan T, Krause M, Bass BL. RNA editing by ADARs is important for normal behavior in *Caenorhabditis elegans*. *EMBO J* 2002; **21**(22): 6025-6035.
  9. Bhogal B, Jepson JE, Savva YA, Pepper AS, Reenan RA, Jongens TA. Modulation of dADAR-dependent RNA editing by the *Drosophila* fragile X mental retardation protein. *Nat Neurosci* 2011; **14**(12): 1517-1524.
  10. Horsch M, Seeburg PH, Adler T, Aguilar-Pimentel JA, Becker L, Calzada-Wack J *et al*. Requirement of the RNA-editing enzyme ADAR2 for normal physiology in mice. *J Biol Chem* 2011; **286**(21): 18614-18622.
  11. Du Y, Stasko M, Costa AC, Davisson MT, Gardiner KJ. Editing of the serotonin 2C receptor pre-mRNA: Effects of the Morris Water Maze. *Gene* 2007; **391**(1-2): 186-197.
  12. Englander MT, Dulawa SC, Bhansali P, Schmauss C. How stress and fluoxetine modulate serotonin 2C receptor pre-mRNA editing. *J Neurosci* 2005; **25**(3): 648-651.
  13. Doe CM, Relkovic D, Garfield AS, Dalley JW, Theobald DE, Humby T *et al*. Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. *Hum Mol Genet* 2009; **18**(12): 2140-2148.
  14. Vitali P, Basyuk E, Le Meur E, Bertrand E, Muscatelli F, Cavaille J *et al*. ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *J Cell Biol* 2005; **169**(5): 745-753.
  15. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009; **136**(2): 215-233.
  16. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol* 2006; **7**(12): 919-931.
  17. Wu D, Lamm AT, Fire AZ. Competition between ADAR and RNAi pathways for an extensive class of RNA targets. *Nat Struct Mol Biol* 2011; **18**(10): 1094-1101.

18. Autism, Developmental Disabilities Monitoring Network Surveillance Year Principal I, Centers for Disease C, Prevention. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR Surveill Summ* 2012; **61**(3): 1-19.
19. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological medicine* 1995; **25**(1): 63-77.
20. Folstein S, Rutter M. Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry* 1977; **18**(4): 297-321.
21. Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg IC, Jakobsson G *et al.* A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *J Child Psychol Psychiatry* 1989; **30**(3): 405-416.
22. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012; **485**(7397): 237-241.
23. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012; **485**(7397): 242-245.
24. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; **485**(7397): 246-250.
25. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012; **74**(2): 285-299.
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. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 2011; **70**(5): 886-897.

28. 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): 863-885.
29. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, Hill DE *et al.* Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med* 2011; **3**(86): 86ra49.
30. 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.
31. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009; **459**(7246): 528-533.
32. Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, Hill RS *et al.* Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet* 2012; **8**(4): e1002635.
33. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG *et al.* Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* 2012.
34. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011; **474**(7351): 380-384.
35. State MW, Levitt P. The conundrums of understanding genetic risks for autism spectrum disorders. *Nat Neurosci* 2011.
36. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010; **466**(7304): 368-372.
37. Bourgeron T. A synaptic trek to autism. *Curr Opin Neurobiol* 2009; **19**(2): 231-234.
38. Kanner L. Autistic disturbances of affective contact. *Nervous child* 1943; **2**: 217-250.

39. Association AP. *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*, 2000.
40. Bailey A, Phillips W, Rutter M. Autism: towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *Journal of child psychology and psychiatry, and allied disciplines* 1996; **37**(1): 89-126.
41. Shattuck PT. The contribution of diagnostic substitution to the growing administrative prevalence of autism in US special education. *Pediatrics* 2006; **117**(4): 1028-1037.
42. Wazana A, Bresnahan M, Kline J. The autism epidemic: fact or artifact? *J Am Acad Child Adolesc Psychiatry* 2007; **46**(6): 721-730.
43. Coo H, Ouellette-Kuntz H, Lloyd JE, Kasmara L, Holden JJ, Lewis ME. Trends in Autism Prevalence: Diagnostic Substitution Revisited. *J Autism Dev Disord* 2007.
44. Mazefsky CA, Oswald DP. The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting. *Autism* 2006; **10**(6): 533-549.
45. Hilton JC, Seal BC. Brief report: comparative ABA and DIR trials in twin brothers with autism. *J Autism Dev Disord* 2007; **37**(6): 1197-1201.
46. Granpeesheh D, Tarbox J, Dixon DR. Applied behavior analytic interventions for children with autism: a description and review of treatment research. *Ann Clin Psychiatry* 2009; **21**(3): 162-173.
47. Myers SM, Johnson CP. Management of children with autism spectrum disorders. *Pediatrics* 2007; **120**(5): 1162-1182.
48. Scahill L, Koenig K, Carroll DH, Pachler M. Risperidone approved for the treatment of serious behavioral problems in children with autism. *J Child Adolesc Psychiatr Nurs* 2007; **20**(3): 188-190.
49. Owen R, Sikich L, Marcus RN, Corey-Lisle P, Manos G, McQuade RD *et al.* Aripiprazole in the treatment of irritability in children and adolescents with autistic disorder. *Pediatrics* 2009; **124**(6): 1533-1540.

50. Rauser L, Savage JE, Meltzer HY, Roth BL. Inverse agonist actions of typical and atypical antipsychotic drugs at the human 5-hydroxytryptamine(2C) receptor. *J Pharmacol Exp Ther* 2001; **299**(1): 83-89.
51. Zhang JY, Kowal DM, Nawoschik SP, Lou Z, Dunlop J. Distinct functional profiles of aripiprazole and olanzapine at RNA edited human 5-HT<sub>2C</sub> receptor isoforms. *Biochem Pharmacol* 2006; **71**(4): 521-529.
52. Veenstra-VanderWeele J, Blakely RD. Networking in autism: leveraging genetic, biomarker and model system findings in the search for new treatments. *Neuropsychopharmacology* 2012; **37**(1): 196-212.
53. Cook EH, Leventhal BL. The serotonin system in autism. *Curr Opin Pediatr* 1996; **8**(4): 348-354.
54. Gai X, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Mol Psychiatry* 2011.
55. Rutter M. Genetic studies of autism: from the 1970s into the millennium. *Journal of abnormal child psychology* 2000; **28**(1): 3-14.
56. Folstein SE, Rosen-Sheidley B. Genetics of autism: complex aetiology for a heterogeneous disorder. *Nature reviews* 2001; **2**(12): 943-955.
57. Lamb JA, Parr JR, Bailey AJ, Monaco AP. Autism: in search of susceptibility genes. *Neuromolecular medicine* 2002; **2**(1): 11-28.
58. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: A decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* 2011.
59. Constantino JN, Todorov A, Hilton C, Law P, Zhang Y, Molloy E *et al.* Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. *Mol Psychiatry* 2012.
60. Yang MS, Gill M. A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence. *Int J Dev Neurosci* 2007; **25**(2): 69-85.
61. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics* 1995; **11**(3): 241-247.

62. Auranen M, Vanhala R, Varilo T, Ayers K, Kempas E, Ylisaukko-Oja T *et al.* A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. *American Journal of Human Genetics* 2002; **71**(4): 777-790.
63. Cantor RM, Kono N, Duvall JA, Alvarez-Retuerto A, Stone JL, Alarcon M *et al.* Replication of Autism Linkage: Fine-Mapping Peak at 17q21. *Am J Hum Genet* 2005; **76**(6).
64. IMGSAC. A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *American journal of human genetics* 2001; **69**(3): 570-581.
65. Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* 1999; **65**(2): 493-507.
66. Shao Y, Cuccaro ML, Hauser ER, Raiford KL, Menold MM, Wolpert CM *et al.* Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *Am J Hum Genet* 2003; **72**(3): 539-548.
67. Molloy CA, Keddache M, Martin LJ. Evidence for linkage on 21q and 7q in a subset of autism characterized by developmental regression. *Mol Psychiatry* 2005; **10**(8): 741-746.
68. IMGSAC. A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Human molecular genetics* 1998; **7**(3): 571-578.
69. Barrett S, Beck JC, Bernier R, Bisson E, Braun TA, Casavant TL *et al.* An autosomal genomic screen for autism. Collaborative linkage study of autism. *American journal of medical genetics* 1999; **88**(6): 609-615.
70. Philippe A, Martinez M, Guilloud-Bataille M, Gillberg C, Rastam M, Sponheim E *et al.* Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. *Human molecular genetics* 1999; **8**(5): 805-812.
71. Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J *et al.* A genomic screen of autism: evidence for a multilocus etiology. *American journal of human genetics* 1999; **65**(2): 493-507.

72. Sadakata T, Washida M, Iwayama Y, Shoji S, Sato Y, Ohkura T *et al.* Autistic-like phenotypes in *Cadps2*-knockout mice and aberrant CADPS2 splicing in autistic patients. *The Journal of clinical investigation* 2007; **117**(4): 931-943.
73. Skaar DA, Shao Y, Haines JL, Stenger JE, Jaworski J, Martin ER *et al.* Analysis of the RELN gene as a genetic risk factor for autism. *Molecular psychiatry* 2005; **10**(6): 563-571.
74. Gauthier J, Joober R, Mottron L, Laurent S, Fuchs M, De Kimpe V *et al.* Mutation screening of FOXP2 in individuals diagnosed with autistic disorder. *American journal of medical genetics* 2003; **118**(2): 172-175.
75. Benayed R, Gharani N, Rossman I, Mancuso V, Lazar G, Kamdar S *et al.* Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. *Am J Hum Genet* 2005; **77**(5): 851-868.
76. Cho IH, Yoo HJ, Park M, Lee YS, Kim SA. Family-based association study of 5-HTTLPR and the 5-HT2A receptor gene polymorphisms with autism spectrum disorder in Korean trios. *Brain research* 2007; **1139**: 34-41.
77. Xu LM, Li JR, Huang Y, Zhao M, Tang X, Wei L. AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res* 2012; **40**(Database issue): D1016-1022.
78. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008; **358**(7): 667-675.
79. Martin CL, Ledbetter DH. Autism and cytogenetic abnormalities: solving autism one chromosome at a time. *Curr Psychiatry Rep* 2007; **9**(2): 141-147.
80. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI *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.
81. Anney R, Klei L, Pinto D, Almeida J, Bacchelli E, Baird G *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet* 2012; **21**(21): 4781-4792.



82. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T *et al.* Strong association of de novo copy number mutations with autism. *Science* 2007; **316**(5823): 445-449.
83. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 2010; **68**(2): 192-195.
84. Davidson J, Goin-Kochel RP, Green-Snyder LA, Hundley RJ, Warren Z, Peters SU. Expression of the Broad Autism Phenotype in Simplex Autism Families from the Simons Simplex Collection. *J Autism Dev Disord* 2012.
85. 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): 863-885.
86. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 2012; **44**(4): 471.
87. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K *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.
88. Reichenberg A, Gross R, Weiser M, Bresnahan M, Silverman J, Harlap S *et al.* Advancing paternal age and autism. *Archives of general psychiatry* 2006; **63**(9): 1026-1032.
89. Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res* 2004; **64**(9): 3060-3071.
90. Judd LL, Mandell AJ. Chromosome studies in early infantile autism. *Arch Gen Psychiatry* 1968; **18**(4): 450-457.
91. Sukumar S, Wang S, Hoang K, Vanchiere CM, England K, Fick R *et al.* Subtle overlapping deletions in the terminal region of chromosome 6q24.2-q26: three cases studied using FISH. *Am J Med Genet* 1999; **87**(1): 17-22.

92. Fan YS, Jayakar P, Zhu H, Barbouth D, Sacharow S, Morales A *et al.* Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. *Hum Mutat* 2007; **28**(11): 1124-1132.
93. Ashley-Koch AE, Mei H, Jaworski J, Ma DQ, Ritchie MD, Menold MM *et al.* An analysis paradigm for investigating multi-locus effects in complex disease: examination of three GABA receptor subunit genes on 15q11-q13 as risk factors for autistic disorder. *Annals of human genetics* 2006; **70**(Pt 3): 281-292.
94. Pickles A, Bolton P, Macdonald H, Bailey A, Le Couteur A, Sim CH *et al.* Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *American Journal of Human Genetics* 1995; **57**(3): 717-726.
95. Jamain S, Quach H, Betancur C, Rastam M, Colineaux C, Gillberg IC *et al.* Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature genetics* 2003; **34**(1): 27-29.
96. Jamain S, Betancur C, Quach H, Philippe A, Fellous M, Giros B *et al.* Linkage and association of the glutamate receptor 6 gene with autism. *Mol Psychiatry* 2002; **7**(3): 302-310.
97. Greger IH, Khatri L, Kong X, Ziff EB. AMPA receptor tetramerization is mediated by Q/R editing. *Neuron* 2003; **40**(4): 763-774.
98. Durand CM, Betancur C, Boeckers TM, Bockmann J, Chaste P, Fauchereau F *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature genetics* 2007; **39**(1): 25-27.
99. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature genetics* 2007; **39**(3): 319-328.
100. Strauss KA, Puffenberger EG, Huentelman MJ, Gottlieb S, Dobrin SE, Parod JM *et al.* Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. *The New England journal of medicine* 2006; **354**(13): 1370-1377.
101. Feyder M, Karlsson RM, Mathur P, Lyman M, Bock R, Momenan R *et al.* Association of mouse Dlg4 (PSD-95) gene deletion and human DLG4 gene

- variation with phenotypes relevant to autism spectrum disorders and Williams' syndrome. *Am J Psychiatry* 2010; **167**(12): 1508-1517.
102. Hussman JP, Chung RH, Griswold AJ, Jaworski JM, Salyakina D, Ma D *et al.* A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. *Mol Autism* 2011; **2**(1): 1.
  103. Silverman JL, Turner SM, Barkan CL, Tolu SS, Saxena R, Hung AY *et al.* Sociability and motor functions in Shank1 mutant mice. *Brain Res* 2011; **1380**: 120-137.
  104. Schmeisser MJ, Ey E, Wegener S, Bockmann J, Stempel AV, Kuebler A *et al.* Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. *Nature* 2012; **486**(7402): 256-260.
  105. Bangash MA, Park JM, Melnikova T, Wang D, Jeon SK, Lee D *et al.* Enhanced Polyubiquitination of Shank3 and NMDA Receptor in a Mouse Model of Autism. *Cell* 2011; **145**(5): 758-772.
  106. Blundell J, Blaiss CA, Etherton MR, Espinosa F, Tabuchi K, Walz C *et al.* Neuroligin-1 deletion results in impaired spatial memory and increased repetitive behavior. *J Neurosci* 2010; **30**(6): 2115-2129.
  107. Tabuchi K, Blundell J, Etherton MR, Hammer RE, Liu X, Powell CM *et al.* A neuroligin-3 mutation implicated in autism increases inhibitory synaptic transmission in mice. *Science* 2007; **318**(5847): 71-76.
  108. Jamain S, Radyushkin K, Hammerschmidt K, Granon S, Boretius S, Varoqueaux F *et al.* Reduced social interaction and ultrasonic communication in a mouse model of monogenic heritable autism. *Proc Natl Acad Sci U S A* 2008; **105**(5): 1710-1715.
  109. Grafodatskaya D, Chung B, Szatmari P, Weksberg R. Autism spectrum disorders and epigenetics. *J Am Acad Child Adolesc Psychiatry* 2010; **49**(8): 794-809.
  110. Jepson JE, Savva YA, Yokose C, Sugden AU, Sahin A, Reenan RA. Engineered alterations in RNA editing modulate complex behavior in *Drosophila*: regulatory diversity of adenosine deaminase acting on RNA (ADAR) targets. *J Biol Chem* 2010.
  111. Maas S. Gene regulation through RNA editing. *Discov Med* 2010; **10**(54): 379-386.

112. Mattick JS. RNA as the substrate for epigenome-environment interactions: rRNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays* 2010; **32**(7): 548-552.
113. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* 2010; **79**: 321-349.
114. Lomeli H, Mosbacher J, Melcher T, Hoyer T, Geiger JR, Kuner T *et al.* Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* 1994; **266**(5191): 1709-1713.
115. Kohler M, Burnashev N, Sakmann B, Seeburg PH. Determinants of Ca<sup>2+</sup> permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. *Neuron* 1993; **10**(3): 491-500.
116. Rula EY, Lagrange AH, Jacobs MM, Hu N, Macdonald RL, Emeson RB. Developmental modulation of GABA(A) receptor function by RNA editing. *J Neurosci* 2008; **28**(24): 6196-6201.
117. Feldmeyer D, Kask K, Brusa R, Kornau HC, Kolhekar R, Rozov A *et al.* Neurological dysfunctions in mice expressing different levels of the Q/R site-unedited AMPAR subunit GluR-B. *Nat Neurosci* 1999; **2**(1): 57-64.
118. Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E *et al.* Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 1997; **387**(6630): 303-308.
119. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* 2007; **315**(5815): 1137-1140.
120. Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG *et al.* Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res* 2008; **36**(16): 5270-5280.
121. Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K. RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep* 2007; **8**(8): 763-769.

122. Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhataar R *et al.* Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol* 2006; **13**(1): 13-21.
123. Borchert GM, Gilmore BL, Spengler RM, Xing Y, Lanier W, Bhattacharya D *et al.* Adenosine deamination in human transcripts generates novel microRNA binding sites. *Hum Mol Genet* 2009; **18**(24): 4801-4807.
124. Mehler MF, Mattick JS. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. *Physiol Rev* 2007; **87**(3): 799-823.
125. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 2009; **324**(5931): 1210-1213.
126. Mattick JS, Mehler MF. RNA editing, DNA recoding and the evolution of human cognition. *Trends Neurosci* 2008; **31**(5): 227-233.
127. Macbeth MR, Schubert HL, Vandemark AP, Lingam AT, Hill CP, Bass BL. Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* 2005; **309**(5740): 1534-1539.
128. Patterson JB, Samuel CE. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol Cell Biol* 1995; **15**(10): 5376-5388.
129. Wahlstedt H, Daniel C, Enstero M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res* 2009; **19**(6): 978-986.
130. Wulff BE, Nishikura K. Substitutional A-to-I RNA editing. *WIREs RNA* 2010; **1**(1): 90-101.
131. Jepson JE, Reenan RA. RNA editing in regulating gene expression in the brain. *Biochim Biophys Acta* 2008; **1779**(8): 459-470.
132. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM *et al.* Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 2011; **333**(6038): 53-58.

133. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* 2004; **2**(12): e391.
134. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J *et al.* Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proc Natl Acad Sci U S A* 2010; **107**(27): 12174-12179.
135. Singh M, Zimmerman MB, Beltz TG, Johnson AK. Affect-related behaviors in mice misexpressing the RNA editing enzyme ADAR2. *Physiol Behav* 2009; **97**(3-4): 446-454.
136. Canitano R. Epilepsy in autism spectrum disorders. *Eur Child Adolesc Psychiatry* 2007; **16**(1): 61-66.
137. Veltman MW, Craig EE, Bolton PF. Autism spectrum disorders in Prader-Willi and Angelman syndromes: a systematic review. *Psychiatr Genet* 2005; **15**(4): 243-254.
138. Daniels JL, Forssen U, Hultman CM, Cnattingius S, Savitz DA, Feychting M *et al.* Parental psychiatric disorders associated with autism spectrum disorders in the offspring. *Pediatrics* 2008; **121**(5): e1357-1362.
139. Sodhi MS, Burnet PW, Makoff AJ, Kerwin RW, Harrison PJ. RNA editing of the 5-HT(2C) receptor is reduced in schizophrenia. *Mol Psychiatry* 2001; **6**(4): 373-379.
140. Gurevich I, Tamir H, Arango V, Dwork AJ, Mann JJ, Schmauss C. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* 2002; **34**(3): 349-356.
141. Hagerman R, Hoem G, Hagerman P. Fragile X and autism: Intertwined at the molecular level leading to targeted treatments. *Mol Autism* 2010; **1**(1): 12.
142. Cavaille J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B *et al.* Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A* 2000; **97**(26): 14311-14316.
143. Zafeiriou DI, Ververi A, Vargiami E. The serotonergic system: its role in pathogenesis and early developmental treatment of autism. *Curr Neuropharmacol* 2009; **7**(2): 150-157.

144. Marco EJ, Skuse DH. Autism-lessons from the X chromosome. *Soc Cogn Affect Neurosci* 2006; **1**(3): 183-193.
145. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE *et al*. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet* 2008; **40**(6): 719-721.
146. Dykens EM, Lee E, Roof E. Prader-Willi syndrome and autism spectrum disorders: an evolving story. *J Neurodev Disord* 2011; **3**(3): 225-237.
147. Bazeley PS, Shepelev V, Talebizadeh Z, Butler MG, Fedorova L, Filatov V *et al*. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene* 2008; **408**(1-2): 172-179.
148. Belmonte MK, Bourgeron T. Fragile X syndrome and autism at the intersection of genetic and neural networks. *Nat Neurosci* 2006; **9**(10): 1221-1225.
149. Schaefer GB, Mendelsohn NJ. Genetics evaluation for the etiologic diagnosis of autism spectrum disorders. *Genet Med* 2008; **10**(1): 4-12.
150. Bassell GJ, Warren ST. Fragile X syndrome: loss of local mRNA regulation alters synaptic development and function. *Neuron* 2008; **60**(2): 201-214.
151. Aschrafi A, Cunningham BA, Edelman GM, Vanderklish PW. The fragile X mental retardation protein and group I metabotropic glutamate receptors regulate levels of mRNA granules in brain. *Proc Natl Acad Sci U S A* 2005; **102**(6): 2180-2185.
152. Jin P, Zarnescu DC, Ceman S, Nakamoto M, Mowrey J, Jongens TA *et al*. Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. *Nat Neurosci* 2004; **7**(2): 113-117.
153. Ishizuka A, Siomi MC, Siomi H. A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes Dev* 2002; **16**(19): 2497-2508.
154. Caudy AA, Myers M, Hannon GJ, Hammond SM. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev* 2002; **16**(19): 2491-2496.

155. Edbauer D, Neilson JR, Foster KA, Wang CF, Seeburg DP, Batterton MN *et al.* Regulation of synaptic structure and function by FMRP-associated microRNAs miR-125b and miR-132. *Neuron* 2010; **65**(3): 373-384.
156. Stefani G, Fraser CE, Darnell JC, Darnell RB. Fragile X mental retardation protein is associated with translating polyribosomes in neuronal cells. *J Neurosci* 2004; **24**(33): 7272-7276.



## **Chapter 2:**

# **Comparative A-to-I RNA Editing in Autistic and Neurotypical Cerebella**

**Alal Eran, Jin Billy Li, Kayla Vatalaro, Jillian McCarthy, Fedik Rahimov,  
Christin Collins, Kyriacos Markianos, David M. Margulies, Emery N. Brown,  
Sarah E. Calvo, Isaac S. Kohane, Louis M. Kunkel**

This chapter is presented in the context of its contemporary science and originally appeared in *Molecular Psychiatry* 2012 Aug 7. doi: 10.1038/mp.2012.118.

Corresponding Supplementary Material can be found in Appendix E.

Adenosine-to-inosine (A-to-I) RNA editing is a neurodevelopmentally-regulated epigenetic modification shown to modulate complex behavior in animals. Little is known about human A-to-I editing, but it is thought to constitute one of many molecular mechanisms connecting environmental stimuli and behavioral outputs. Thus, comprehensive exploration of A-to-I RNA editing in human brains may shed light on gene-environment interactions underlying complex behavior in health and disease. Synaptic function is a main target of A-to-I editing, which can selectively recode key amino acids in synaptic genes, directly altering synaptic strength and duration in response to environmental signals. Here we performed a high-resolution survey of synaptic A-to-I RNA editing in a human population, and examined how it varies in autism, a neurodevelopmental disorder in which synaptic abnormalities are a common finding. Using ultra-deep (>1000x) sequencing, we quantified the levels of A-to-I editing of 10 synaptic genes in postmortem cerebella from 14 neurotypical and 11 autistic individuals. A high dynamic range of editing levels was detected across individuals and editing sites, from 99.6% to below detection limits. In most sites, the extreme ends of the population editing distributions were individuals with autism. Editing was correlated with isoform usage, clusters of correlated sites were identified, and differential editing patterns examined. Finally, a dysfunctional form of the editing enzyme ADARB1 was found more commonly in postmortem cerebella from individuals with autism. These results provide a population-level, high-resolution view of A-to-I RNA editing in human cerebella, and suggest that A-to-I editing of synaptic genes may be informative for assessing the epigenetic risk for autism.

## **Introduction**

Site-specific adenosine-to-inosine (A-to-I) RNA base conversions, carried out by adenosine deaminase acting on RNA (ADAR) enzymes, exhibit precise regional specificity

in the brain and modulate complex behavior in model organisms<sup>1-3</sup>. A-to-I RNA editing is an efficient means to increase RNA complexity, thereby fine-tuning both gene function and dosage<sup>4-6</sup>. The cellular machinery recognizes inosine as guanosine, so A-to-I editing of codons and splicing signals directly modifies protein-coding gene function<sup>6-11</sup>, while editing of microRNAs<sup>12-14</sup> and their binding sites<sup>15</sup> alters gene expression. This is particularly important in the human brain, the single most complex organ in cellular diversity, connectivity, morphogenesis, and responses to environmental stimuli<sup>16</sup>.

Synaptic function is a major target of A-to-I editing<sup>17</sup>, which can fine-tune neurophysiological properties in response to environmental stimuli<sup>18, 19</sup>. Canonical signaling pathways acting on the editing enzymes link A-to-I RNA editing to environmental cues: *ADARB1* function requires inositol hexakisphosphate<sup>20</sup>, and the expression of *ADAR* is interferon-inducible<sup>21</sup>. Several recoding events that directly alter synaptic strength or duration in response to environmental signals have been characterized in rodents<sup>7-11</sup>. For example, mRNAs of the serotonin receptor HTR2C undergo editing in five sites, which dramatically alters its G-protein coupling activity, and hence the relationship between serotonin levels and postsynaptic signal transduction<sup>2, 11</sup>. This editing is regulated by exposure to acute stress and chronic treatment with antidepressants<sup>22</sup>. Another example is the neurodevelopmentally-regulated editing of transcripts encoding the AMPA receptors GRIA2, GRIA3 and GRIA4<sup>23</sup>, where arginine to glycine (R/G) recoding of the ligand binding domains leads to faster desensitization recovery<sup>7</sup>. Moreover, glutamine to arginine (Q/R) editing of the transmembrane domains in the kainate receptors GRIK1 and GRIK2 reduces their calcium permeability<sup>8</sup>, with varying degrees of editing throughout mouse neurodevelopment<sup>23</sup>. Since 0 to 100% of mRNA molecules can be edited at any given point<sup>24</sup>, Q/R and R/G editing of ionotropic glutamate receptors provides an efficient means for fine-tuning the glutamatergic synapse in response to the changing environment.

Little is known about A-to-I RNA editing in humans, but it has been postulated to be one

of the molecular mechanisms connecting environmental inputs and behavioral outputs<sup>18, 25</sup>. The increased editing in humans<sup>26</sup> as compared to other animals<sup>27</sup>, including nonhuman primates<sup>28</sup>, has been proposed to generate molecular complexity that might constitute the basis of higher-order cognition<sup>18</sup>. Therefore, characterizing A-to-I editing in typically and atypically developed individuals may shed light on environment-dependent epigenetic mechanisms central to human neurodevelopment. Autism Spectrum Disorders (ASD; [MIM 209850]) are highly heritable common neurodevelopmental disorders of complex genetic etiology, characterized by deficits in reciprocal social interaction and repetitive behaviors<sup>29</sup>. Several studies characterize synaptic abnormalities in ASD<sup>29-32</sup>. However, the mechanisms for gene-environment interactions and their contribution to the observed synaptic alterations remain unknown. The number of candidate genes is rapidly increasing<sup>33-35</sup> and a main challenge is to identify the context in which they confer risk to ASD. Recent twin studies suggest that the contribution of environmental factors to ASD is larger than previously thought, with lower monozygotic concordance estimates (77 to 88%), and a surprisingly high dizygotic concordance of 31% (as compared to a sibling recurrence rate of 19%)<sup>36, 37</sup>. Hence, the identification of mechanisms governing gene-environment interactions relevant to ASD could be informative for risk assessment<sup>38</sup>. A-to-I RNA editing is potentially one such mechanism, linking environmental stimuli with synaptic transmission.

Several lines of evidence support an examination of the link between A-to-I editing and ASD. First, model organisms with altered A-to-I editing exhibit maladaptive behaviors characteristic of ASD<sup>2, 39</sup>, sometimes with seizures<sup>1, 40</sup> or Prader-Willi-like symptoms<sup>41</sup>, both of which are typically detected in 25% of children with ASD<sup>42, 43</sup>. Second, altered editing of the serotonin receptor *HTR2C* has been detected in a mouse model of autism<sup>44</sup> and in disorders that aggregate in families with ASD<sup>45</sup>, including schizophrenia<sup>46</sup> and major depression<sup>47</sup>. Third, a fly model of Fragile X syndrome, the most common single gene cause of ASD<sup>48</sup>, was recently shown to exhibit significant editing alterations,

via a direct interaction between the Fragile X gene *FMR1* and the editing enzyme *dADAR*<sup>49</sup>. Finally, independent genomic studies have implicated variants in synaptic genes, the most edited type of genes<sup>17</sup>, as a recurring theme in ASD<sup>29-32</sup>.

Here we focus on neurodevelopmentally-regulated A-to-I editing that directly alters synaptic function. We precisely quantify and compare the levels of editing across individuals, and characterize the distinct editing landscapes of ten synaptic genes acting in the human cerebellum, contrasting postmortem brains from typically developed individuals with brains isolated from individuals with ASD. We then (i) specifically examine editing patterns in the glutamatergic and serotonergic systems, (ii) correlate editing with isoform usage, and (iii) identify clusters of correlated sites. Importantly, we find that the relative usage of a dysfunctional form of the editing enzyme *ADARB1* is significantly higher in ASD.

## Materials and Methods

### Subjects

Thirty fresh-frozen cerebellar samples of deidentified individuals with nonsyndromic autism and carefully-matched neurotypical individuals were obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank (NICHD BTB) and the Harvard Brain Tissue Resource Center (HBTRC), through the Autism Tissue Program (ATP). To minimize confounding factors, matching was based on age, gender, race, and post-mortem-interval (PMI). **Supplementary Table 1** details the 25 samples that passed quality control measures.

### Selection of Target Genes

With the accumulation of multidisciplinary studies highlighting synaptic alterations in ASD<sup>29, 31, 32, 34, 50-54</sup>, we chose to focus on neurodevelopmentally-regulated A-to-I

recoding that alters synaptic function. All synaptic genes shown to undergo developmentally dependent A-to-I recoding by Wahlstedt et al.<sup>23</sup> were included in this study (**Supplementary Table 2**).

## **Molecular Methods**

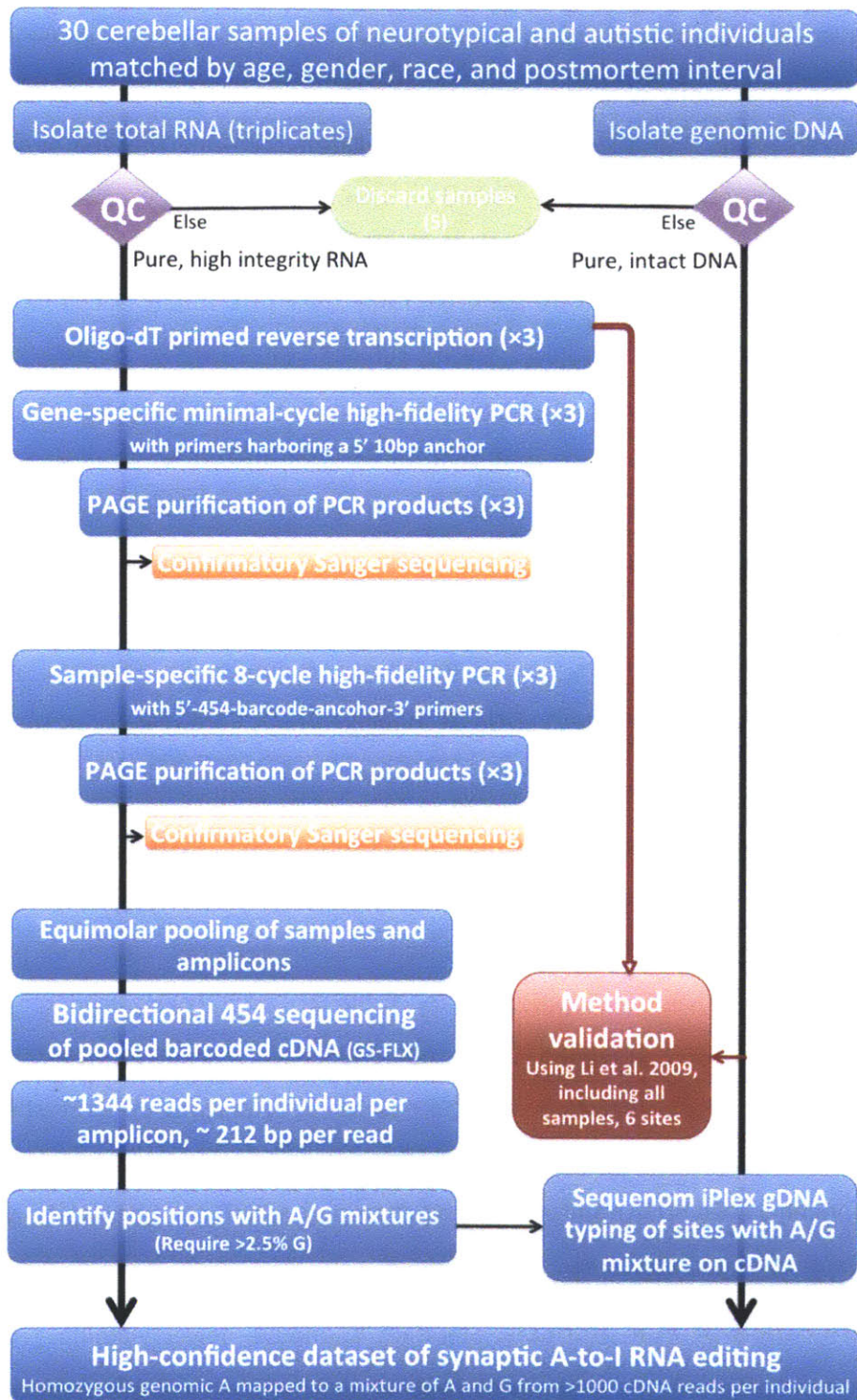
### **RNA isolation and quality assurance**

Tissue samples were disrupted and RNA isolated in triplicate (mirVana™ miRNA Isolation Kit, Life Technologies, Grand Island, NY, USA), followed by DNase I treatment (DNA-free™, Life Technologies). Samples that seemed intact on 1% agarose gels were quantified on the ND-1000 Spectrophotometer (NanoDrop, Wilmington, DE, USA) and their integrity analyzed using the Agilent 2100 Bioanalyzer Eukaryote Total RNA Nano Series II (Agilent Technologies, Santa Clara, CA, USA). Only samples with RNA Integrity Number (RIN) >7 were included in the study, minimizing post mortem degradation artifacts<sup>55</sup>. The mean RIN of the samples used was 8.0 (**Supplementary Table 1**).

### **454 cDNA library preparation**

A stringent PCR-based library preparation protocol was designed to ensure unbiased templates for multiplexed 454 sequencing, as shown in **Figure 1**. Target selection was based on two rounds of PCR, intended to minimize barcode-based amplification biases: the first was gene specific and the second incorporated a sample-specific barcode to each amplicon.

**cDNA synthesis.** cDNA was synthesized from 250ng of each triplicate total RNA prep, using oligo dT priming. Invitrogen's Superscript III First-Strand Synthesis System for RT-PCR was utilized, following the manufacturer's recommendations (Life Technologies, Grand Island, NY, USA).



**Figure 1** Overview of the experimental approach.

**Universal tag optimization.** To avoid barcode-based amplification biases, we employed a two-round PCR approach, where the first PCR is gene specific and includes a universal tag at the 5' end of each primer, and the second PCR is sample specific, primes off the universal tag, and adds a barcode immediately following the 454 sequencing primer. To minimize the read length wasted on a universal tag, we designed an optimal 10bp tag, which enables efficient priming off such a short sequence. For optimal tag selection, all possible 10-mers were generated and their GC content calculated. Those 258,047 10-mers with 50% GC were subject to fastPCR quality analysis (<http://primerdigital.com/fastpcr.html>), yielding a quality score in the range of 5 to 104. Then 10-mers with a quality score  $\geq 90$  that do not form dimers with the 454 primers were selected by attaching each 10-mer to all possible 5'-454 primer-barcode-3' forward and reverse combinations and testing for homo- and hetero-dimerization using fastPCR. The best 10-mers were *tcgatcagca* (added at the 5' of all forward primers) and *tacgatgcgt* (reverse). All primers were synthesized and purified by polyacrylamide gel electrophoresis (PAGE) by Integrated DNA Technologies (IDT, Coralville, IA, USA).

**Gene-specific amplification.** Triplicate PCRs were performed as detailed in **Supplementary Table 3** using Invitrogen's Accuprime PFX system, following the manufacturer's protocol, to ensure high fidelity, specificity and yield (Life Technologies). Primers were designed to amplify all Refseq isoforms, targeting the regions flanking the known editing sites listed in **Supplementary Table 2**. FastPCR was used for primer design, and all reactions amplified across splice junctions, except for the intronless KCNA1. To minimize PCR artifacts, a minimal number of cycles was used. Starting with 20 cycles, PCR products were analyzed on the QIAxcel Gel Electrophoresis System (QIAGEN, Valencia, CA, USA) and were subject to two additional amplification cycles until the expected band appeared.

**PAGE purification of PCR products.** Standard 8% native polyacrylamide gels were prepared in-house and triplicate gene-specific PCR products were separated by



electrophoresis next to a 10bp ladder (Life Technologies) at 140V. The gels were stained with SYBR® Gold Nucleic Acid Gel Stain (Life Technologies) and the correct product sizes (**Supplementary Table 3**) excised on Life Technologies' Safe Imager™ blue light transilluminator. To prevent cross-contamination, separate gels, staining boxes, and scalpels were used for each sample. The excised bands were transferred to 0.5mL tubes that were punctured at the bottom, placed inside a 1.5mL tube, and spun at 16,300 g for 10 minutes. The resulting shredded gel pieces were eluted in 300uL of 10mM Tris-HCl and 1mM EDTA overnight at 37°C. The next day the gel slurry was spun for 5 minutes at 16,300 g and the supernatant transferred to a new 1.5mL tube. 300uL of 10mM Tris-HCl and 1mM EDTA was added to the leftover shredded gel pieces and incubated for 2 hours at 37°C. The gel pieces were removed using Spin-X® Centrifuge Tube Filters (Corning, Amsterdam, The Netherlands). The eluted cDNA was isopropanol-precipitated with 0.7 volume of isopropanol and 1/300 volume GlycoBlue (Life Technologies) as a carrier. The cDNA pellet was then washed with 70% ethanol four times, and resuspended in 20uL nuclease free water. The concentration and purity of the products was determined using the ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA).

**Confirmatory Sanger sequencing.** 20% of the gene-specific amplification products were sequenced bidirectionally on the Applied Biosystems 3730 DNA Analyzer (Life Technologies), to confirm that they had the correct sequence.

**Sample-specific amplification.** 100ng of the purified gene-specific triplicate PCR products were amplified to add a sample specific barcode fused to the 454 sequencing primer. The standard 10bp 454 MID barcodes were used and every individual with ASD and its matched neurotypical sample shared a barcode (and later sequenced on opposite sides of the PicoTiterPlate). *454 Primer A* was fused to the barcode in the forward primer and *454 Primer B* in the reverse. Matched samples were handled together and amplified on the same plate. Life Technologies' Accuprime PFX system was

used, following the manufacturer's protocol, in eight cycles of PCR. 20pmol of primers were used in each reaction.

**PAGE purification and confirmatory Sanger sequencing of sample-specific PCR products** were performed as described above for the gene-specific PCR.

**Quantitation.** The purified sample-specific PCR products were quantitated using Quant-iT™ PicoGreen® dsDNA Reagent (Life Technologies), to enable equimolar pooling for sequencing. Batches were made on the gene level; that is, all barcoded products of the same gene-specific PCR were analyzed together. Samples were diluted 1:20 and two volumes of each dilution were quantitated: 5uL and 8uL, using the Victor3 Multiplate Reader (PerkinElmer, Waltham, MA, USA), in black round-bottom plates (Corning), following the manufacturer protocol. Lambda DNA samples were included on each plate, used to create a standard curve from 0 to 50ng, and also served as controls with a specific quantity of 12.5ng and 3.125ng, in duplicate. For each amplicon, a sample's quantitation was considered successful if the coefficient of variation (CV) between the quantities determined for the 8uL product and 1.6\* 5uL product was  $\leq 10\%$ . The protocol was repeated with larger volumes until the CV between the two dilutions was  $\leq 10\%$ . The final concentration was the average between the two dilutions.

**Equimolar pooling.** Samples were first pooled to amplicons based on their Picogreen readout and then amplicons were pooled based on their product size. Separate pools were made for ASD and neurotypical samples in a final concentration of 5ng/uL. An aliquot was taken for Bioanalyzer and NanoDrop quality assessment, which showed that all amplicon bands existed in each pool and had a relatively similar intensity.

### **454 sequencing**

Bidirectional GS FLX sequencing was performed as described<sup>56</sup> by the 454 Life Sciences Sequencing Center (Branford, CT, USA). ASD and neurotypical pools were sequenced on

opposite sides of a 2-region PicoTiterPlate. 457,104 reads were obtained, containing 85,709,299 high quality bases (**Supplementary Figure 1**).

## **DNA isolation**

About 2g of the same cerebellar tissue samples were disrupted with a mortar and pestle on dry ice, and genomic DNA (gDNA) was extracted using the QIAamp DNA Mini Kit (QIAGEN, Valencia, CA, USA).

## **Genomic DNA genotyping**

Sequenom iPLEX genotyping (Sequenom, Inc, San Diego, CA, USA) was done at Boston Children's Hospital's SNP Genotyping Facility. Twenty-five SNPs were genotyped in four populations. Every sample was genotyped in triplicate, and all triplicates had to agree on the genotype to be considered successful. All passing SNPs had a genotype success rate of >96.7%.

## **Method Validation**

To validate our findings, cDNA and gDNA from all samples were independently analyzed by the parallel capture and sequencing method described by Li et al.<sup>17</sup> Padlock probes were designed to hybridize to the editing site at position 4 of the *Extension Arm*, and ten base-pairs of the captured sequence separated the *Ligation* and *Extension Arms* (**Supplementary Table 4**). To capture editing sites on cDNA and gDNA, 135ng of padlock probes were mixed with 1µg of gDNA or 200ng of cDNA, and incubated with Taq polymerase and ligase as previously described<sup>17</sup>. Then the linear DNA part of the padlock was digested by exonuclease I and III, and the circles created by the extended and ligated probes were amplified as described<sup>17</sup>. Following purification, all samples were pooled and sequenced in two lanes of an Illumina Genome Analyzer, as described<sup>17</sup>. The resulting FASTQ files were analyzed with perl scripts that deconvoluted

the samples based on their barcode sequences, aligned reads to reference sequences using BLAST with a four-letter word size, and counted the number of fully-mapped reads that contained each allele and had no more than 3 mismatches or indels with the reference sequence.

About a quarter of the sites examined by ultradeep 454 sequencing were also assayed by padlock-capture and Illumina sequencing. On average, 3420 unidirectional Illumina cDNA reads and 627 unidirectional Illumina gDNA reads were obtained for each sample at each of the six sites assayed by both methods: *CYFIP2* K/E, *FLNA* Q/R, *GRIA2* Q/R+4, *GRIA2* R/G, *GRIK1* Q/R, and *KCNA1* I/V. More than 98% of the genomic reads mapping to these editing sites contained adenosines. Therefore, adenosine/guanosine mixtures on cDNA with >2.5% G were considered to be representative of A-to-I editing. The editing levels were modeled by a beta-binomial distribution as detailed below, and compared to those determined by ultradeep 454 cDNA sequencing coupled to gDNA genotyping.

The results of the two editing detection methods were tightly correlated, with an average Pearson's *r* of 0.923 (mean *p* = 0.002) across six sites. See **Supplementary Table 5** for details.

### ***ADARB1* Isoform Usage Analysis**

**Relative *ADARB1* isoform usage analysis with semiquantitative PCR.** ThermoStart Taq DNA Polymerase (ABgene, Epsom, UK) was used to amplify cDNA triplicates in 30 cycles, 65°C annealing, and primers which detect all known *ADARB1* isoforms: aaacagtctccgccagtcaa (forward, exon 4) and caggtcaccaaactaccagg (reverse, exon 6). The normal isoforms yielded a 1057bp product and the dysfunctional NR\_027672 isoform a 122bp product. The relative frequency of the latter was quantitated using QIAGEN's QIAxcel Gel Electrophoresis System. Both bands were isolated using Invitrogen's E-Gel SizeSelect System (Life Technologies), and sequenced bidirectionally on ABI 3730 DNA Analyzers (Life Technologies). The skipping of exon 5, which contains

the enzyme's double stranded RNA binding domains, was confirmed in all 122bp bands, and its inclusion confirmed in all 1057bp products. The statistical significance of the difference between inactive *ADARB1* isoform usage in individuals with ASD and neurotypical individuals was assessed using the Mann-Whitney U test ( $p=0.003$ ).

**ADARB1 isoform quantitation using real-time RT-qPCR.** Real-time qPCR was performed on triplicate cDNA preparations from all 25 samples using TaqMan® Gene Expression Assays (Applied Biosystems), with primers and probe sequences shown in **Supplementary Table 6**. The samples were first analyzed on the BioMark™ 96.96 Dynamic Array (Fluidigm) in nanoliter reaction volumes, and then in larger volumes (10 $\mu$ L) on the ABI 7900HT System (Applied Biosystems). A panel of eleven commonly used endogenous control genes was quantitated alongside the *ADARB1* isoforms on the BioMark™ system: ACTB, GAPDH, B2M, 18S, PPIA, HPRT1, GUSB, TBP, RPLP0, TFRC and PGK1 (TaqMan® assays Hs99999903\_m1, Hs99999905\_m1, Hs99999907\_m1, Hs99999901\_s1, Hs99999904\_m1, Hs99999909\_m1, Hs99999908\_m1, Hs99999910\_m1, Hs99999902\_m1, Hs99999911\_m1, and Hs99999906\_m1, respectively). The two most stable genes across all samples were GAPDH and RPLP0, as determined by the geNorm algorithm (<http://medgen.ugent.be/~jvdesomp/genorm/>). To increase transcript levels to detectable concentrations in nanoliter reaction chambers of the BioMark™ array without altering their ratios in the transcriptome, cDNA was pre-amplified with a diluted pool of TaqMan® assays for 14 cycles and the final reactions were diluted 1:5 with TE. Expression levels were normalized to GAPDH using the delta Ct method<sup>76</sup>, and the Pearson correlation between Fluidigm BioMark™ and ABI 7900HT quantities was 0.67 ( $p= 4.65e-11$ ). This demonstrates platform-specific biases and is consistent with previous findings<sup>77</sup>. The Mann-Whitney U test was used to assess the significance of the difference in quantities of the dysfunctional *ADARB1* isoform relative to GAPDH in individuals with ASD and neurotypical individuals ( $p=0.007$ ).

**Relative ADARB1 isoform usage analysis in an independent set of postmortem cerebella.** In addition to the 25 samples whose synaptic editing levels have been surveyed, an independent set of postmortem cerebella from five neurotypical individuals and four individuals with ASD was examined (**Supplementary Table 7**) using semi-quantitative RT-PCR, as detailed above.

**Relative ADARB1 isoform usage analysis using RNAseq. RNA-seq.** Ten of the surveyed samples were subject to high-throughput RNA sequencing on Applied Biosystem's SOLiD platform (Life Technologies). Poly(A)+ RNA was captured (Oligotex mRNA Mini Kit, QIAGEN), heat-fragmented (95°C for 20 minutes), and prepared for sequencing (Small RNA Expression Kit, Life Technologies) that yielded 114M reads. Life Technologies' WT Analysis software package was used to align the reads to the reference genome and expressed sequence tags, identifying 33.4M uniquely mapped reads. Of those, an average of 877 reads per sample mapped to *ADARB1* exons. The relative frequency of the dysfunctional NR\_027672 isoform was measured as 1- the relative frequency of isoforms that include the double-stranded RNA binding domains (dsRBDs). The latter was measured as the RPKM (Reads Per Kilobase of exon model per Million mapped reads) ratio between the skipped, dsRBDs containing exon, and its immediate 3' constitutive exon:

Relative frequency of the dysfunctional *ADARB1* isoform NR\_027672 =

$$1 - \frac{RPKM (ADARB1 \text{ dsRBDs exon})}{RPKM (ADARB1 \text{ constitutive exon})}$$

**Correlation between editing and splicing of *FLNA*.** When we amplified across the *FLNA* Q/R site using one primer in exon 43 and one in exon 44, an unexpected 720bp band of varying brightness appeared in all samples. Bidirectional Sanger sequencing revealed that it is a result of intron 43 retention. Its relative abundance was quantitated

using QIAGEN's QIAxcel electrophoresis system, and correlated with Q/R editing levels detected by 454 sequencing.

**Gene expression profiling.** Affymetrix Exon 1.0 ST arrays (Santa Clara, CA, USA) were used to measure global gene expression of each RNA sample. Following quantile normalization, gene expression levels were calculated using the Probe Log Iterative ER (PLIER) algorithm and differences in Gene Core expression were determined using the Mann-Whitney U test.

### **Computational Methods**

**454 variant calling and quantification.** The 454 GS Amplicon Variant Analyzer software was used to deconvolute samples, align reads to reference sequences, and call variants and haplotypes based on bidirectional sequence changes from the reference.

**Beta-binomial modeling of editing levels.** For each sample in each site, the posterior editing density,  $f(\text{editing})$ , was a beta distribution with parameters  $\alpha=1+\text{number of reads with G mapped to the site}$ , and  $\beta=1+\text{number of reads with A mapped to the site}$  (**Supplementary Figure 2**).

**Differential editing analyses.** Kolmogorov-Smirnov (KS) tests were used to assess the significance of differences in continuous measurements, such as those of editing levels between neurotypical individuals and individuals with ASD. To compare discrete distributions, such as the count of *GRIK2* or *HTR2C* isoforms resulting from combinatorial RNA editing, Pearson's chi square test was used. The total number of reads belonging to each of the possible editing isoforms were summed across each group and only isoforms supported by  $> 5$  reads in both groups were included in the analysis (following this test's assumptions).

**Correlation between dysfunctional *ADARB1* Isoform frequency and overall editing levels.** For each sample  $i$ ,  $RF_i$  is the relative frequency of the dysfunctional

*ADARB1* isoform. A sample's overall editing level,  $E_i$ , is the sum of its standardized editing levels across all sites,  $\sum Z_{ij}$ , where  $Z_{ij}$  is the standardized editing level of sample  $i$  at site  $j$ . Pearson's correlation was calculated between these two metrics,  $RF$  and  $E$ , across all samples, to measure their linear dependence.

**Association between editing and splicing of AMPA Receptors.** Two-by-two contingency tables were used to summarize the relationships between isoform selection and editing (**Supplementary Table 8**), and Fisher's exact test was used to determine the significance of the association between them.

**Correlations among editing sites.** The Pearson correlation coefficient was calculated to quantify the linear relationships between editing at different sites, among all individuals. Biclustering was then used to identify modules of tightly correlated sites, with the EXPANDER software<sup>57</sup>.

**Independence of editing at neighboring sites.** To summarize the relationships between editing at *GRIA2* Q/R and Q/R+4 among all individuals, a two-by-two contingency table is shown in **Supplementary Table 8**. Fisher's exact test was used for power and significance calculations.

**Multiple testing correction.** All p-values were Benjamini-Hochberg corrected for multiple testing<sup>58</sup> to ensure that the false discovery rate of this entire study is below 0.05.

## Results

### Precise Multiplex Quantitation of A-to-I RNA Editing in Human Cerebella

For a high-resolution view of A-to-I RNA editing in a human brain population, ultra-deep (>1000x) 454 cDNA sequencing and genomic DNA (gDNA) genotyping were used to



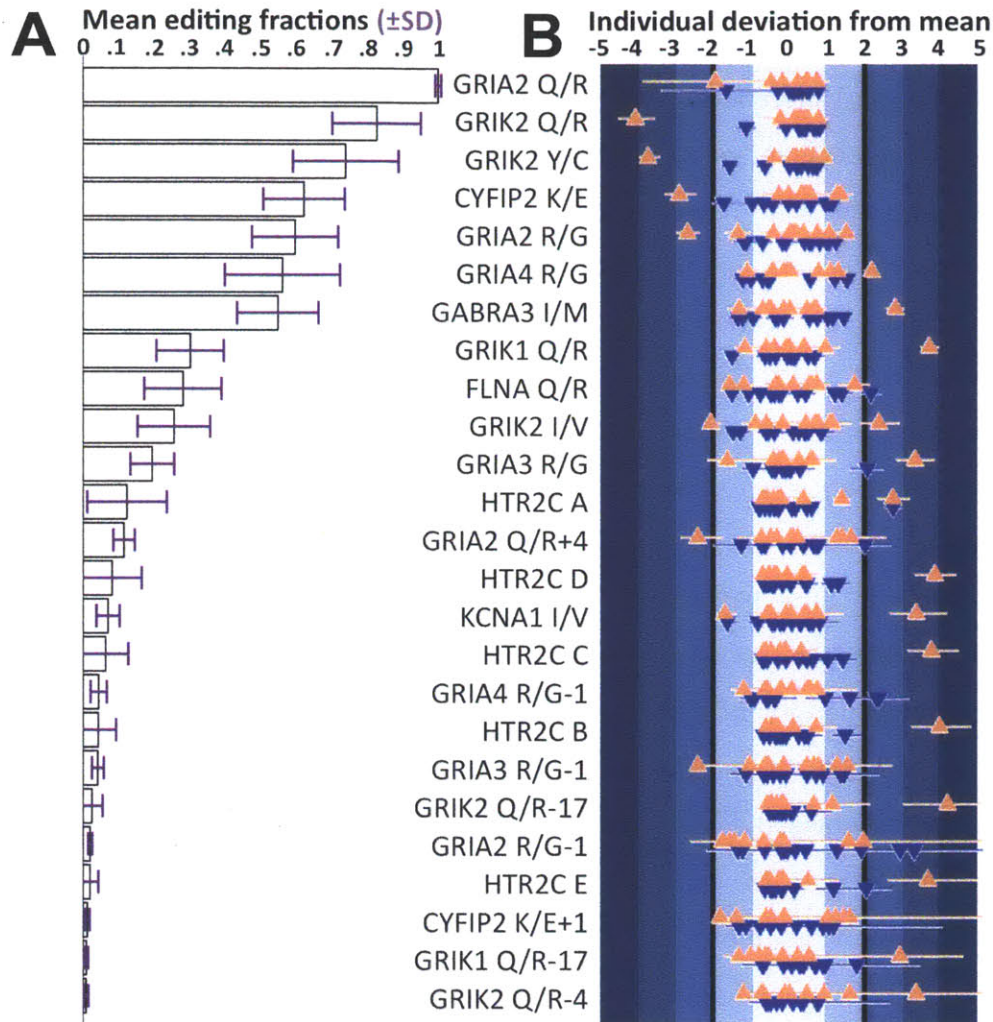
detect A/G mixtures on cDNA mapping to homozygous genomic A/A (**Figure 1**). This study focused on all ten synaptic genes shown in mouse to undergo neurodevelopmentally-regulated A-to-I recoding<sup>23</sup> that results in well-characterized neurophysiological alterations<sup>7-9, 59-61</sup> (**Supplementary Table 2**). Editing was measured in postmortem cerebellum, one of several brain regions implicated in ASD<sup>62</sup> by both imaging<sup>63</sup> and autopsy<sup>64</sup> studies. Cerebellar tissue samples were obtained from neurotypical individuals and individuals with non-syndromic ASD, matched by gender, age, race, and postmortem interval (**Supplementary Table 1**). Pooled, barcoded, bidirectional 454 cDNA sequencing yielded on average 1344 reads of 212bp per individual per amplicon (**Supplementary Figure 1**). gDNA was genotyped at 18 well-characterized editing sites with known functional consequences and 7 positions aligned to a mixture of guanosines and adenosines on cDNA. A-to-I edited sites were identified by the presence of a bidirectional cDNA A/G mixture at homozygous A/A gDNA positions. Editing levels were modeled by a beta-binomial distribution, resulting in a posterior editing density for each individual at each site. This model considers both the fraction of edited reads and the total number of reads covering a site, to produce a distribution describing the level of editing and our confidence in that measurement (**Supplementary Figure 2**).

This approach provided a high confidence dataset of synaptic A-to-I RNA editing from individuals with ASD and matched neurotypical individuals, with an average 95% confidence interval length of 0.038. For independent validation, the same cDNA and gDNA were analyzed by parallel capture and >3000x Illumina sequencing<sup>17</sup> in six sites among all individuals. A tight correlation between the two editing detection methods is shown in **Supplementary Figure 3** and **Supplementary Table 5**, with an average correlation coefficient of 0.923.

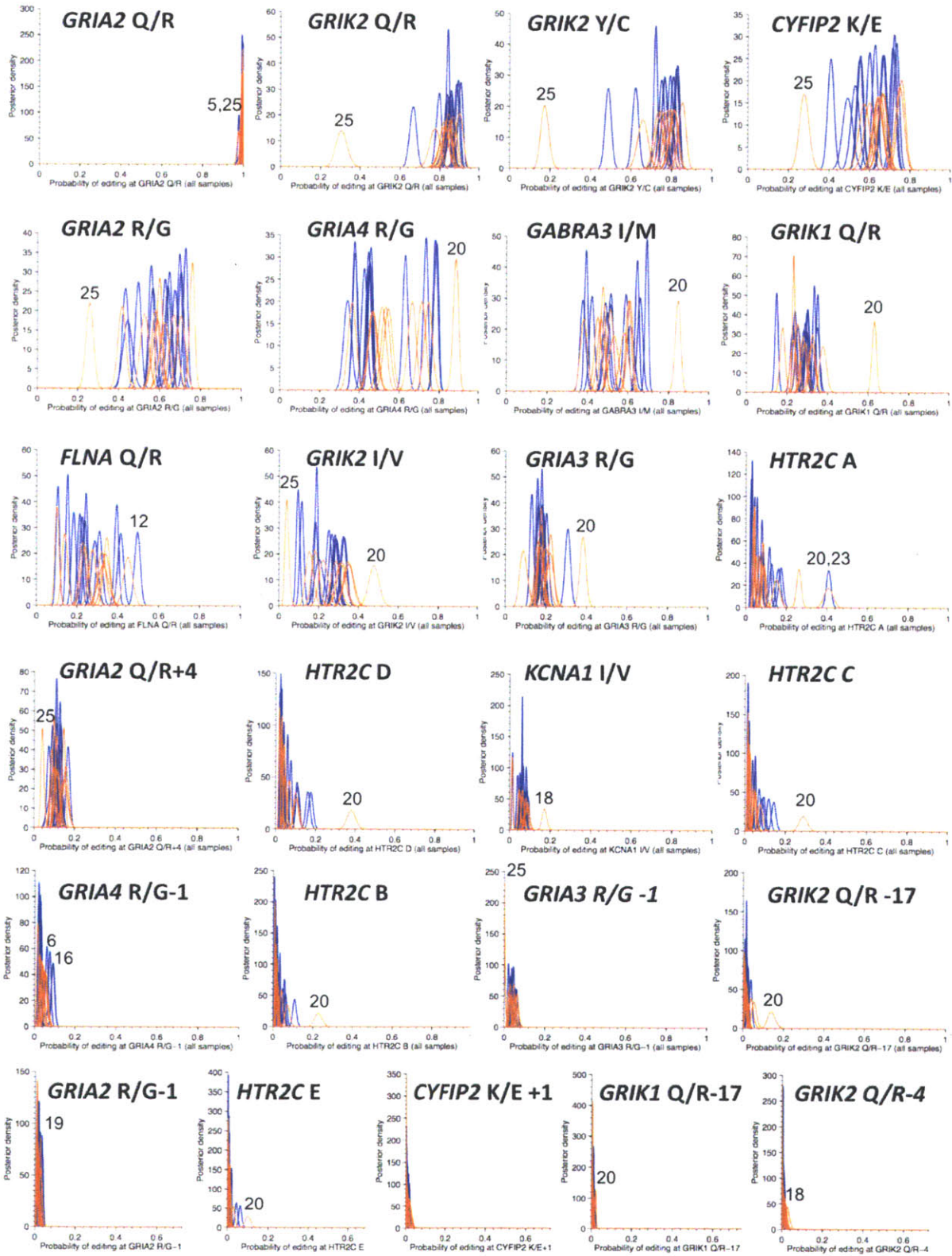
### **High Dynamic Range of Editing across Sites and Individuals**

The editing levels of 25 sites were robustly measured and found to range from 2.5% to

99.6% (Figure 2A and Figure 3).



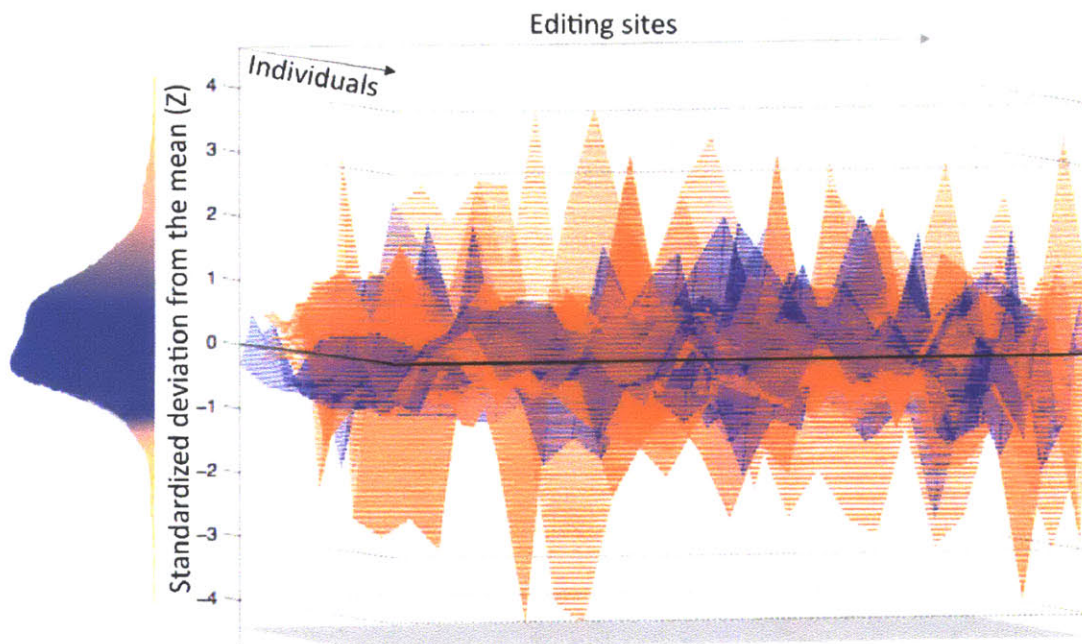
**Figure 2** Individuals with ASD at the extremes of the population editing distributions in 20 of 25 sites. **(A)** High dynamic range of mean editing across sites, from 99.6% to below detection limit. Standard deviation bars, shown in purple, indicate the large variability in neurodevelopmentally-regulated editing among carefully matched individuals. **(B)** The individual standardized deviations from the mean editing level across all sites show that the extreme of the population editing distributions tend to be individuals with ASD (orange triangles), and they are the major contributor to the large variability shown in **(A)**. Horizontal lines denote the standardized 95% confidence interval of an individual's posterior editing distribution. Synaptic editing levels at least two standard deviations away from the mean are highly informative of ASD, with a positive predictive value of 78%.



**Figure 3** Probabilities of editing among all individuals and all sites. This panel summarizes the results of our study, showing the posterior editing densities among all individuals at each site. Sites are sorted according to the mean population editing level. Individuals with ASD are shown in orange and neurotypical individuals in blue. A high dynamic range of editing and large variability between carefully-matched individuals can be observed. Individuals with ASD consistently lie at the extremes of the population editing distributions. In 20 of 25 sites, outliers are individuals with ASD. Samples located two or more standard deviations away from the mean are labeled according to the serial numbers noted in **Supplementary Table 1**.

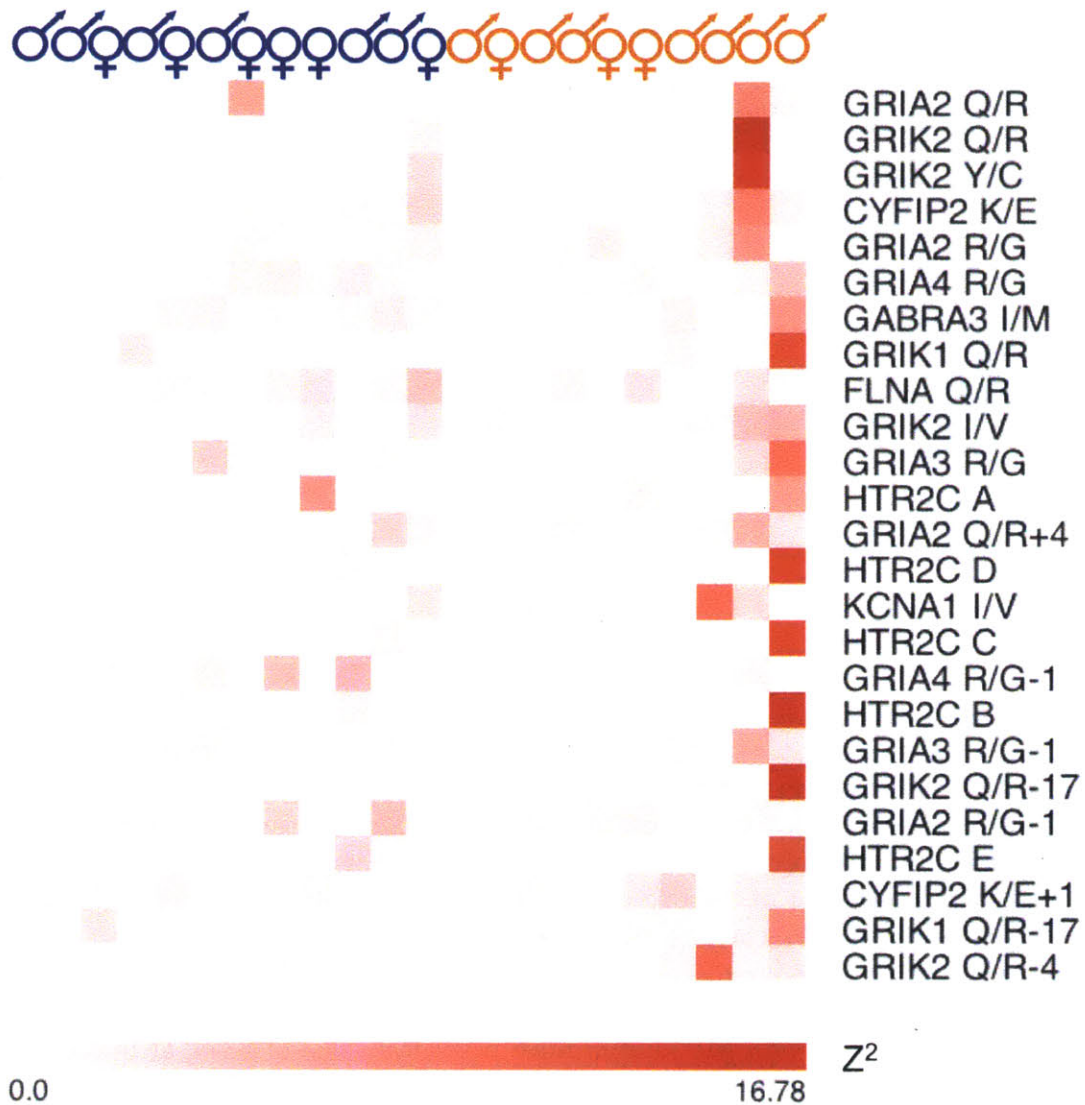
Unexpectedly broad distributions of editing levels were found across neurotypical individuals and carefully-matched individuals with ASD in those sites previously shown to be neurodevelopmentally regulated in mice<sup>23</sup> (average coefficient of variation 54.8%). Only one site, *GRIA2* Q/R, showed similar levels of editing among all individuals (CV=0.5%), consistent with reports that editing of this site is essential and unchanged throughout development<sup>23, 65</sup>. No relationships between editing levels and age, gender, race, postmortem interval, or RNA integrity were detected (**Supplementary Figure 4**).

By inspection of the individual posterior editing densities across all sites (**Figure 3**), we noticed that the extremes of most editing distributions are individuals with ASD. To quantify this observation, the point estimates of the editing levels of each individual at each site were transformed to Z scores and those  $\geq 2$  or  $\leq -2$  were considered extreme, representing editing levels that are at least two standard deviations away from the mean. In 20 of 25 sites, individuals with ASD were at the extreme of the spectrum of editing seen for that site (**Figure 2B**). Having extreme synaptic editing levels is highly informative of ASD, with a positive predictive value  $p(ASD | |Z_{editing}| > 2)$  of 0.78. Outliers at different sites are different individuals with ASD (**Figures 3 and 4**), and editing of more than three standard deviations away from the mean was specific to individuals with ASD (**Figure 4**).



**Figure 4** Differences in the landscapes of synaptic editing in individuals with ASD (orange) and neurotypical individuals (blue). The contours of the standardized deviation from the mean editing level (Z-scores, y-axis) are shown across sites (x-axis) and individuals (coming out of the page), with matched samples overlaid on one another. Individuals with ASD are consistently located at the extremes of the examined cohort's editing distributions, and outliers at different sites are different individuals with ASD. On the left is the summary distribution of all Z-scores, showing that extreme editing is specific to individuals with ASD. The positive predictive value of being at the extreme  $p(\text{ASD} | |Z| > 2)$  is 0.78.

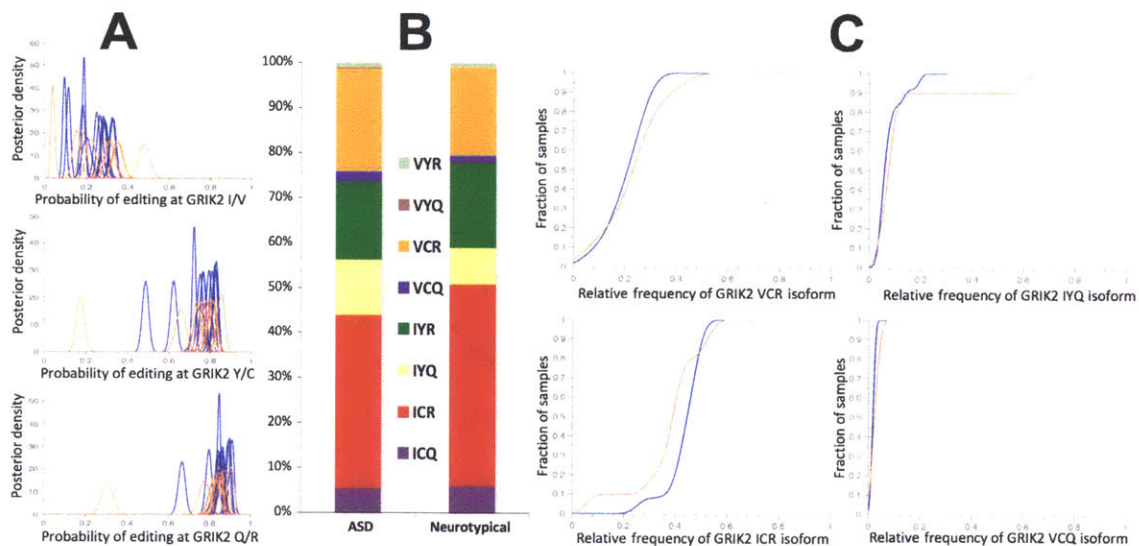
The overall editing variance in ASD was more than two-fold that of neurotypical individuals (ASD variance=0.58, median neurotypical variance of equally sized subsamples=0.24, Brown-Forsythe  $p=5.6e-3$ ). The major contributors to this increased variance are 30% of the individuals with ASD (**Figure 5**).



**Figure 5** Individual contributions to the variance. With the sample variance defined as  $S^2 = (N-1) \cdot \sum_i (X_i - \mu)^2$ , this heatmap portrays  $(X_{ij} - \mu_j)^2$  for each individual  $i$  at site  $j$ . Specifically,  $X_{ij}$  is the point estimate of the degree of A-to-I editing of individual  $i$  at site  $j$ ;  $\mu_j$  is the mean editing level of the entire population at site  $j$ ; and to enable a convenient visual comparison of all sites, each measurement was normalized by the site variance. Thus, it is the equivalent of the square of the Z score for each individual at each site. Neurotypical individuals are shown in blue, individuals with ASD in orange, and sites are ordered according to the mean population editing level.

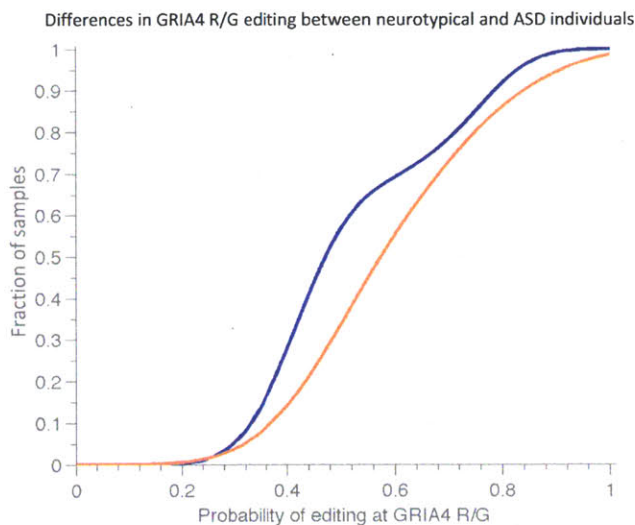
## Patterns of Editing in the Glutamatergic and Serotonergic Systems

Next, editing patterns were specifically examined in the glutamatergic and serotonergic systems, which have been shown to be implicated in ASD<sup>29, 66</sup>. *GRIK2*, whose genomic locus has been repeatedly linked and associated with ASD<sup>66</sup>, is edited in three sites, leading to the formation of eight protein isoforms that differ in their calcium permeability<sup>23</sup>. Differential *GRIK2* editing between individuals with ASD and neurotypical individuals is depicted in **Figure 6**, both on a single site and isoform-wide levels.



**Figure 6** Differences in editing-mediated *GRIK2* isoforms between neurotypical individuals (blue) and individuals with ASD (orange). *GRIK2* is edited at three sites: I/V, Y/C (both part of TM1) and Q/R (TM2), altering the receptor's calcium permeability. Editing at these sites was found to be tightly correlated (**Figure 11B**). (A) Large variability in *GRIK2* editing at each site among all individuals. The extremes of the observed editing spectra are individuals with ASD. Consult Supplementary Figure 5 for more details. (B) Combinatorial editing of the three sites results in eight isoforms, with distinct molecular properties. Significant group differences in *GRIK2* isoform distributions between individuals with ASD and neurotypical individuals were identified ( $p < 1e-4$ , chi-square test). (C) The strongest differences are in IYQ and VCQ isoforms, which are ~1.5-fold over-represented in ASD. Shown are differences in the cumulative density functions (CDFs) of the VCR, ICR, IYQ, and VCQ isoforms.

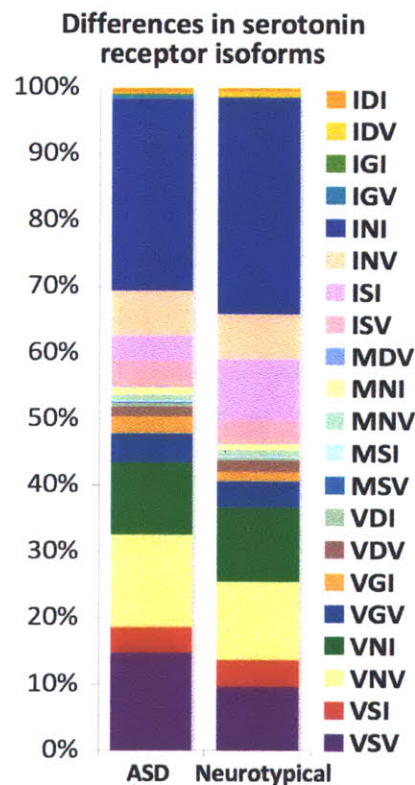
Kolmogorov-Smirnov (KS) tests were used to assess the significance of differences in continuous measurements, such as those of editing levels between neurotypical individuals and individuals with ASD. To compare discrete distributions, such as the count of *GRIK2* or *HTR2C* isoforms resulting from combinatorial RNA editing, Pearson's chi square test was used. The relative frequencies of *GRIK2* isoforms were significantly different between individuals with ASD and neurotypical individuals ( $p < 1e-4$ , chi-square test). Differences were also detected in the editing of *GRIA4* R/G ( $p < 3.6e-2$ , KS test, **Figure 7**). Other comparisons of glutamate receptors did not reach statistical significance, as this study is not powered to detect small effect sizes (**Supplementary Figure 5**).



**Figure 7** Differences in *GRIA4* R/G editing between individuals with ASD (orange) and neurotypical individuals (blue). Shown are kernel-smoothed cumulative distribution functions (ksCDFs) of *GRIA4* R/G editing. On a group level, editing at this site is increased in ASD ( $p = 3.58e-2$ , KS test).



Editing was also examined in the serotonin receptor *HTR2C*, which is targeted by the only FDA-approved drugs to treat autistic symptoms, risperidone and aripiprazole. This receptor undergoes RNA editing at five sites, which dramatically alters its G-protein coupling activity. *HTR2C* editing may create up to 24 different protein isoforms, characterized by distinct molecular and behavioral phenotypes<sup>2, 6</sup>. Comparing the relative isoform frequencies of individuals with ASD and neurotypical individuals revealed significant differences ( $p < 1e-4$ , chi square test), the strongest in IDV, MNV and ISI isoforms - all under-represented in ASD (**Figure 8**). However, the variance within ASD is much larger than that between neurotypical and ASD individuals, suggesting that this finding should be considered with caution and revisited with a larger sample size.



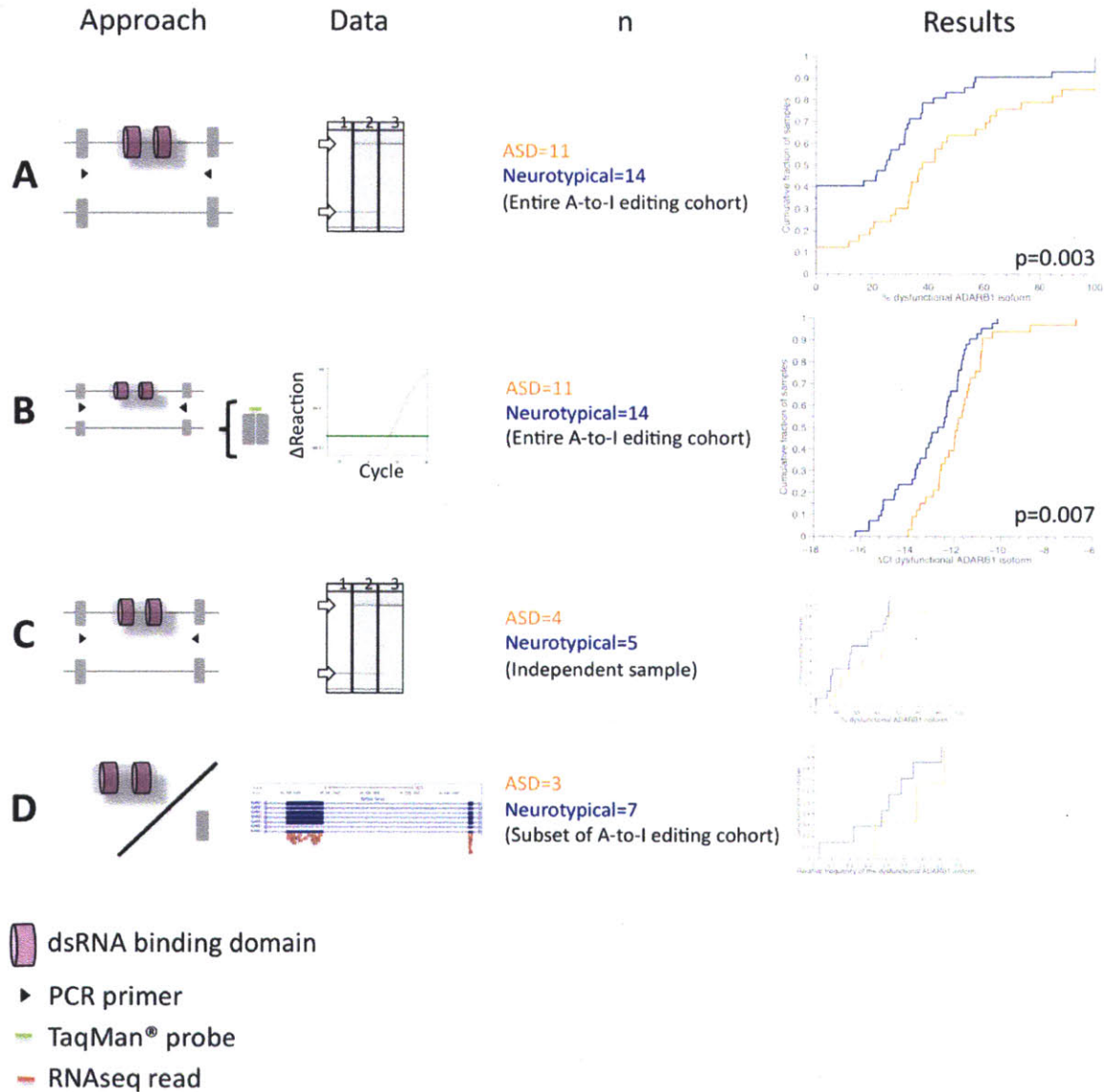
**Figure 8** Differences in the relative abundance of *HTR2C* isoforms between individuals with ASD and neurotypical individuals ( $p < 1e-4$ , chi-square test). 21 of the possible 24 isoforms created by combinatorial editing of five *HTR2C* sites were detected. The

greatest differences were in the IDV, MNV and ISI isoforms, all under-represented in ASD by >2-fold.

### **Differential Relative *ADARB1* Isoform Usage in Individuals with ASD**

The sites examined here are predominantly edited by *ADARB1* (**Supplementary Table 2**). To explore the regulatory basis of the observed editing differences, *ADARB1* expression and relative isoform usage were examined. In rodents, the function of *ADARB1* has been shown to be developmentally regulated via alternative splicing that creates an inactive protein, while its overall expression remains fairly constant throughout development<sup>23, 67</sup>. In humans, a dysfunctional *ADARB1* isoform (*ADAR2g*, NR\_027672), resulting from alternative skipping of the exon harboring two double stranded RNA binding domains (dsRBDs), has been found to constitute about 20% of adult cerebellar *ADARB1* mRNA<sup>68</sup>. As such, the presence of this dsRBDs-encoding exon was assayed by semi-quantitative RT-PCR in all samples. While individuals with ASD and neurotypical individuals showed similar levels of overall *ADARB1* expression (**Supplementary Figure 6**), the relative usage of the inactive *ADARB1* isoform was significantly more common in individuals with ASD (**Figure 9A**,  $p=3.0e-3$ , Mann-Whitney U test). Exon skipping was confirmed by bidirectional sequencing in all samples. The abundance of the inactive *ADARB1* isoform was also quantitated in all samples using TaqMan® Gene Expression Assays (**Figure 9B**). The dysfunctional form was over-expressed in individuals with ASD ( $p=7.3e-3$ , Mann-Whitney U test), with a significant correlation between semi-quantitative and quantitative PCR results (Pearson's  $p=2.0e-5$ ). Additionally, an increased relative usage of the dysfunctional *ADARB1* isoform in ASD was observed in an independent cohort of nine cerebellar tissue samples (**Figure 9C**). The mean relative frequency of the exon-skipped isoform was higher in individuals with ASD in this set as well ( $61.3\% \pm 6.0\%$  vs.  $50.3\% \pm 3.3\%$  in neurotypical individuals), and in all 34 samples there was a significant increase in the relative usage of the dysfunctional *ADARB1* isoform in ASD ( $p=0.003$ ,

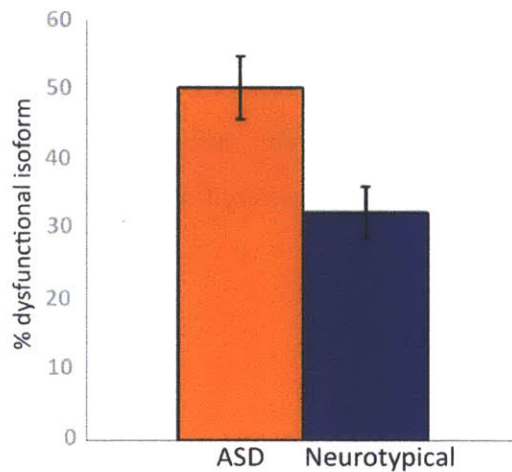
Mann Whitney U test, **Figure 10**). Ten of the surveyed samples were also subject to RNA-seq (**Figure 9D**), and the relative *ADARB1* exon usage determined by RNA-seq correlated with the semi-quantitative RT-PCR results (**Supplementary Figure 7**).



**Figure 9** Overview of *ADARB1* isoform usage analyses. **(A)** The relative frequency of the inactive *ADARB1* isoform (NR\_027672) was assayed in all 25 samples using semi-

quantitative PCR. Individuals with ASD were found to use the dysfunctional form more frequently than neurotypical individuals ( $p=0.003$ , Mann-Whitney U test). This isoform lacks the enzyme's dsRBDs and has been shown to be untranslated<sup>68</sup>. ADARB1 is the predominant editing enzyme of the sites analyzed in this study. The mean relative frequency of the dysfunctional *ADARB1* isoform was  $46.3\% \pm 5.6\%$  in ASD and  $26\% \pm 4.5\%$  in neurotypical individuals, similar to that previously reported in neurotypical human cerebellum<sup>68</sup>. All 25 samples included in the RNA editing survey were also part of this analysis. **(B)** The expression of the dysfunctional isoform was quantitated in all samples using TaqMan® Gene Expression Assays, relative to *GAPDH*. The dysfunctional form was found to be upregulated in individuals with ASD as compared to neurotypical individuals ( $p=0.007$ , Mann-Whitney U test, relative to *GAPDH*), with a significant correlation between quantitative and semi-quantitative PCR results (Pearson's  $p= 2.0e-5$ ). **(C)** A small independent cohort of postmortem cerebella was collected (**Supplementary Table 7**) and assayed as in (A). Individuals with ASD demonstrated a higher relative usage of the inactive form in this set as well (shown here), and overall the dysfunctional form was more frequently used by individuals with ASD than controls ( $p=0.003$  in all 34 samples, Mann-Whitney U test, **Figure 10**). **(D)** Ten samples have undergone polyA+ selected RNA-seq, and the relative frequency of the dysfunctional isoform was measured using the RPKM (Reads Per Kilobase of exon model per Million mapped reads) ratio between the skipped, dsRBDs containing exon, and its immediate 3' constitutive exon. This measurement correlates with the results in (A) and was on average higher among the examined individuals with ASD (**Supplementary Figure 7**).

The relative frequency of the dysfunctional isoform is correlated with overall editing levels (Pearson's  $r= 0.48$ ,  $p=3.2e-2$ ), as measured by the sum of the standardized editing scores across all sites.

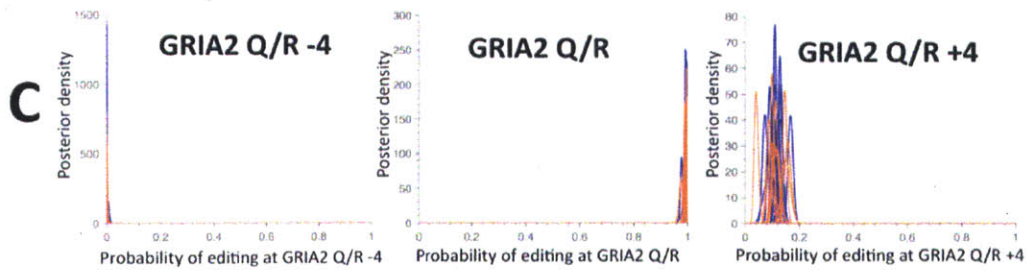
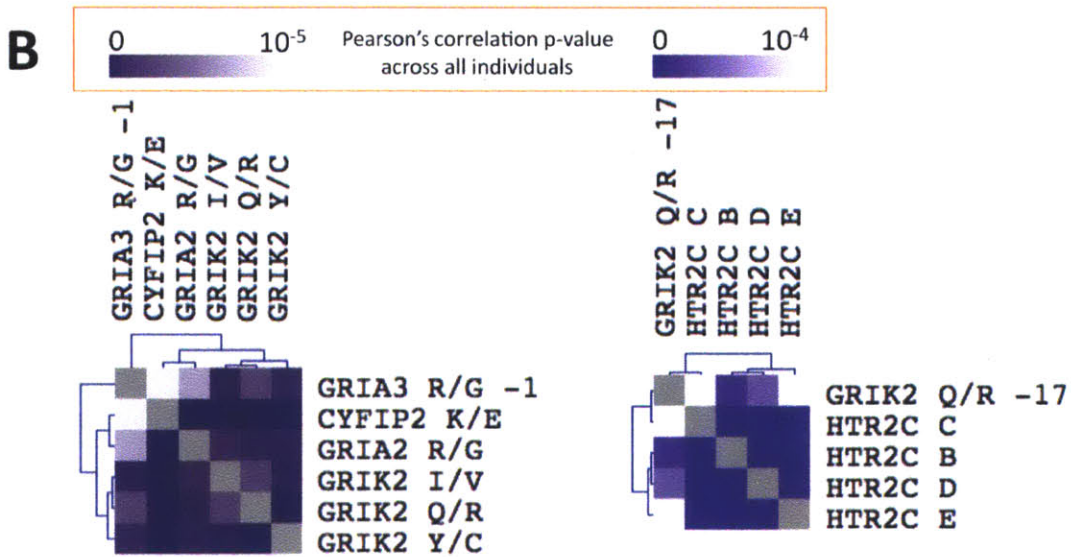
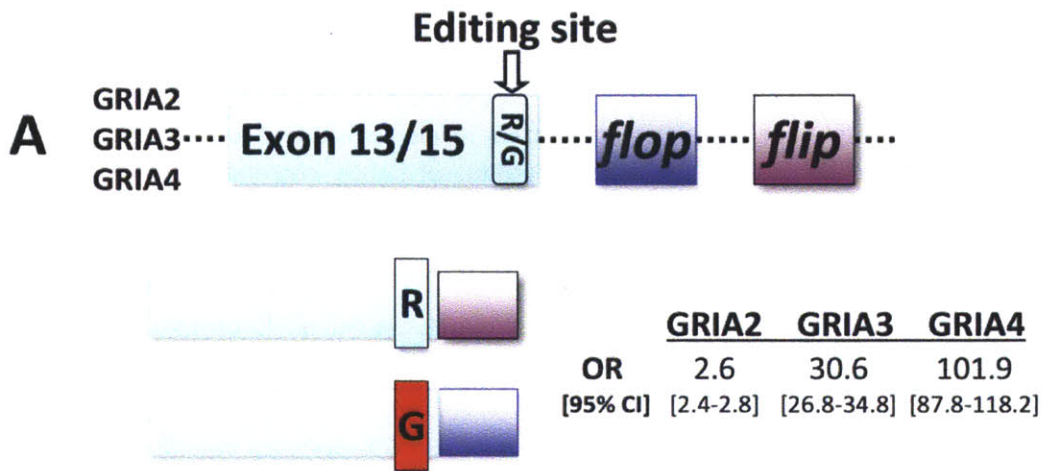


**Figure 10** Relative *ADARB1* isoform usage in 34 postmortem cerebella. The presence of the *ADARB1* exon containing the double stranded RNA binding domain was assayed in an independent sample of 5 neurotypical individuals and 4 individuals with ASD using semi-quantitative PCR. Individuals with ASD demonstrated higher relative usage of the dysfunctional NR\_027672 isoform in this set as well, and in all 34 samples the significance of the difference between neurotypical individuals and individuals with ASD was  $p=0.003$  (Mann-Whitney U test).

### **Tight Relationships between Recoding and Splicing in Three AMPA Receptors and *Filamin-A***

The long 454 reads (**Supplementary Figure 1**) were used to study regulatory characteristics of human RNA editing, including the previously hypothesized<sup>6; 69</sup> relationships between recoding and alternative splicing selection in three AMPA receptors (*GRIA2*, *GRIA3*, and *GRIA4*) and filamin-A (*FLNA*). *GRIA2*, *GRIA3*, and *GRIA4* each contain two mutually exclusive exons that modulate desensitization kinetics, termed *flip* and *flop*<sup>7</sup>. Editing at the 3' end of the exon immediately preceding the *flip/flop* module was strongly associated with the *flop* isoform in *GRIA4* (OR=101.9 [95%

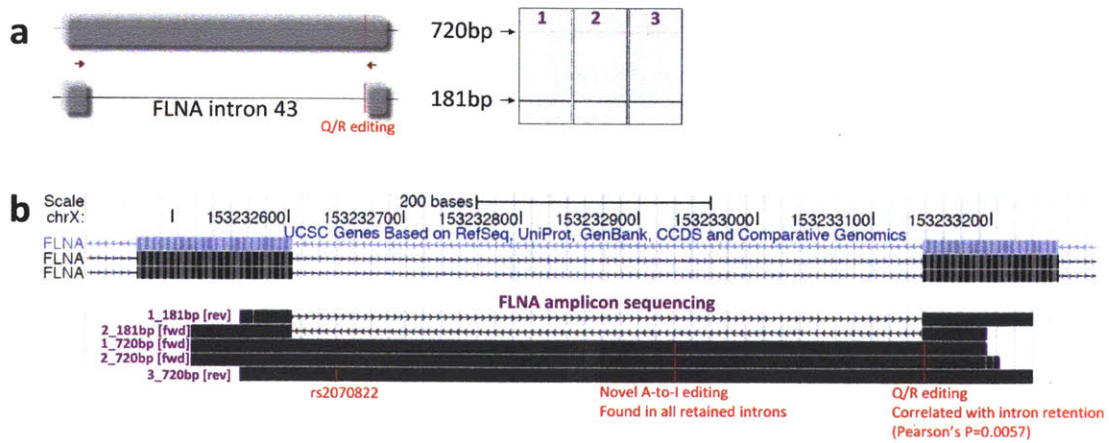
CI 87.8-118.2]), *GRIA3* (OR= 30.6 [26.8-34.8]), and to a lesser extent *GRIA2* (OR= 2.6 [2.4-2.8]) (All p-values <1e-300, Fisher's exact test, **Figure 11A** and **Supplementary Table 8**). Consistent with the detection of differential editing at *GRIA4* R/G between individuals with ASD and neurotypical individuals, differential *GRIA4* isoform usage was also identified ( $p < 3.6 \times 10^{-2}$ , KS test, **Supplementary Figure 8**).



**Figure 11** Relationships between editing and splicing, and among sites. **(A)** Editing at the 3' end of the exon immediately preceding the *flip/flop* module is strongly associated with the *flop* isoform in *GRIA4* (OR=101.9 [95% CI 87.8-118.2]), *GRIA3* (OR= 30.6 [26.8-34.8]), and to a lesser extent *GRIA2* (OR= 2.6 [2.4-2.8]) (All p-values <1e-300, Fisher's exact test). **(B)** The linear correlation of editing levels across sites was measured among all individuals and biclustered to reveal modules of tightly correlated sites, suggesting coregulation by the same factors. The tightest module contains 6 sites in 4 genes, demonstrating highly correlated editing across these sites among all individuals. Consult Supplementary Figure 5 for broader perspective. **(C)** Three neighboring sites at a 4 bp distance from one another undergo independent editing ( $p=1$ , power>0.97, Fisher's exact test), with *GRIA2* Q/R -4 fully unedited, *GRIA2* Q/R fully edited, and *GRIA2* Q/R+4 variably edited across individuals, between 4-17%.

Another relationship between recoding and splicing was examined in *FLNA*, where editing at the 3' end of exon 43 was found to be significantly correlated with intron 43 retention (Pearson's  $p<6e-3$ , **Figure 12**). This intron retention introduces a frameshift and is not included in any human Refseq *FLNA* isoform. Alignment of the retained intron sequence to expressed sequence tags and short transcriptome reads revealed its expression in several tissues from multiple organisms. Thus it likely represents a conserved splicing event that possibly regulates *FLNA* dosage. All retained introns were found to contain an unknown A-to-I editing site at their center, verified by gDNA genotyping. Whether this editing is a cause or effect of the intron retention remains to be elucidated.





**Figure 12** Correlation between editing and splicing of Filamin-A (*FLNA*) (a) *FLNA* is edited at the 3' end of exon 43, resulting in a glutamine to arginine change. When we amplified across the Q/R site an unexpected 720bp band of varying brightness appeared in all samples, in addition to the expected 181bp product. This is visualized by the gel at the top, which includes 3 representative samples (b) Sequencing of the additional 720bp band in all samples revealed that it is a result of intron 43 retention. Shown is the UCSC genome browser illustration of exon 43, intron 43, and exon 44 of human *FLNA* with their locations on the X chromosome. On the top are all the Refseq *FLNA* isoforms, all lacking intron 43. On the bottom is the alignment of bidirectional Sanger sequence of the PCR products from the gel shown in (a). The red marks designate differences from the reference human genome. For all individuals, the relative abundance of the retained intron isoform was quantitated using the QIAxcel electrophoresis system and correlated with Q/R editing levels detected on 454 sequencing. A tight correlation between Q/R editing and intron 43 retention was observed ( $p=0.0057$ , Pearson correlation). All retained introns were found to contain a novel A-to-I editing site, verified by gDNA genotyping. Fifteen of the 25 individuals showed an additional A to G change at the 3' end of the retained intron, but gDNA genotyping showed that this is a genomic SNP (rs2070822). There was no significant difference in the minor allele frequency of this SNP between individuals with ASD and neurotypical individuals.

### *cis* and *trans* Relationships across Editing Sites

To learn more about the regulation of human RNA editing, clusters of tightly correlated editing sites across all individuals were identified. First, Pearson's correlation was calculated to quantify the linear dependencies between editing at different sites, among

all individuals. Biclustering was then used to identify modules of tightly correlated sites with an average correlation coefficient  $> 0.7$  between all pairs of sites in the cluster, across all samples (**Supplementary Table 9**). The cluster with the strongest correlation contained 6 sites in 4 genes (*GRIA2* R/G, *CYFIP2* K/E, *GRIA3* R/G-1, and *GRIK2* I/V, Q/R and Y/C) (Pearson's  $r=0.8$ ,  $p<1e-4$ ), suggesting that these sites may be co-regulated (**Figure 11B**). Some neighboring sites, including *GRIA2* Q/R and Q/R+4, showed independent editing (Fisher's Exact  $p=1$ , Power $>0.97$ , **Figure 11C** and **Supplementary Table 8**), suggesting that their editing is either spatiotemporally distinct and/or carried out by different complexes.

## Discussion

This study represents an initial examination in ASD of A-to-I RNA editing, a form of gene-environment interaction that fine-tunes synaptic function in response to environmental stimuli. Human genes that undergo editing are more often involved in processes that are affected in neurodevelopmental disorders<sup>17, 18</sup>, and A-to-I editing is overall increased in the human lineage<sup>26-28</sup>. While little is known about the impact of RNA editing in humans, there is compelling evidence supporting a key role for A-to-I editing in modulating complex behavior in animals<sup>1-3, 41, 44, 70</sup>. Mice and flies with altered editing levels recapitulate several behavioral homologues to human ASD<sup>1, 41, 44</sup>. Gene-specific studies in animals demonstrate the causal relationship between editing levels and specific maladaptive behaviors<sup>1, 10</sup>, and between specific environmental exposures and editing levels<sup>71</sup>. Therefore, the results presented here shed light on the role of an epigenetic mechanism that connects environmental signals and downstream behavioral outputs, integrating genetic and environmental information.

DNA sequence variation in a number of different loci is strongly associated with ASD, but no individual locus is altered in more than 2% of cases<sup>66</sup>. In contrast, 30% of the

individuals with ASD examined here showed extreme levels of synaptic RNA editing, suggesting that A-to-I editing may be both a marker for and mechanism of ASD. It has been recently shown that typical synaptic protein synthesis occurs within an optimal range and deviations in either direction can lead to cognitive impairments<sup>72</sup>. This study suggests that a similar inverse-U relationship might exist between synaptic A-to-I editing levels and the ASD phenotype. In ASD, it now appears that increased variance can be found in different levels of function, ranging from synaptic A-I editing and protein synthesis<sup>72</sup>, through mitochondrial<sup>73</sup> and immune<sup>74</sup> function, to cellular and systems neuroanatomy<sup>75</sup>. A plausible generalization of these findings is that many biological processes exhibit increased variance in ASD as a result of many etiologies disrupting ASD function and homeostasis. We speculate that altered A-to-I editing could act as a common compensatory mechanism for the wide range of synaptic abnormalities in ASD, as affected neurons try to maintain synaptic homeostasis. Alternatively, altered editing could be a direct consequence of ASD-causing mutations. In either case, larger sets of genes, brain regions, and individuals should be examined to understand the potential contribution of A-to-I editing to ASD.

## **Data Availability**

Access to the raw sequence reads can be found at the National Database for Autism Research under accession number NDARCOL0001951.

## Acknowledgments

We thank Oliver St. Lawrence, Jamie Jett, Benjamin Boese and Tim Harkins at 454, our wonderful lab mates, Prof. David Bartel, Thutrang Nguyen, Eran Mick, and Elena Helman for their tremendous help. Human tissue was obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, contract HHSN275200900011C, Re. No. N01-HD-9-0011, and from The Harvard Brain Tissue Resource Center (HBTRC), through the Autism Tissue Program. HBTRC is supported by Grant MH068855. The Molecular Genetics Core Facility at Children's Hospital Boston Intellectual and Developmental Disabilities Research Center (IDDRC) is supported by Grant NIH-P30-HD18655. This study was generously supported by the Nancy Lurie Marks Family Foundation, The Roche Applied Science Sequencing Grant Program, Autism Speaks, Simons Foundation, and NIH Grant 1R01MH085143-01.

## References

1. Jepson JE, Savva YA, Yokose C, Sugden AU, Sahin A, Reenan RA. Engineered alterations in RNA editing modulate complex behavior in *Drosophila*: regulatory diversity of adenosine deaminase acting on RNA (ADAR) targets. *J Biol Chem* 2011; **286**(10): 8325-8337.
2. Mombereau C, Kawahara Y, Gundersen BB, Nishikura K, Blendy JA. Functional relevance of serotonin 2C receptor mRNA editing in antidepressant- and anxiety-like behaviors. *Neuropharmacology* 2010; **59**(6): 468-473.

3. Tonkin LA, Saccomanno L, Morse DP, Brodigan T, Krause M, Bass BL. RNA editing by ADARs is important for normal behavior in *Caenorhabditis elegans*. *EMBO J* 2002; **21**(22): 6025-6035.
4. Maas S. Gene regulation through RNA editing. *Discov Med* 2010; **10**(54): 379-386.
5. Mattick JS. RNA as the substrate for epigenome-environment interactions: rRNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. *Bioessays* 2010; **32**(7): 548-552.
6. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* 2010; **79**: 321-349.
7. Lomeli H, Mosbacher J, Melcher T, Hoyer T, Geiger JR, Kuner T *et al.* Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* 1994; **266**(5191): 1709-1713.
8. Kohler M, Burnashev N, Sakmann B, Seeburg PH. Determinants of Ca<sup>2+</sup> permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. *Neuron* 1993; **10**(3): 491-500.
9. Rula EY, Lagrange AH, Jacobs MM, Hu N, Macdonald RL, Emeson RB. Developmental modulation of GABA(A) receptor function by RNA editing. *J Neurosci* 2008; **28**(24): 6196-6201.

10. Feldmeyer D, Kask K, Brusa R, Kornau HC, Kolhekar R, Rozov A *et al.* Neurological dysfunctions in mice expressing different levels of the Q/R site-unedited AMPAR subunit GluR-B. *Nat Neurosci* 1999; **2**(1): 57-64.
11. Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E *et al.* Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 1997; **387**(6630): 303-308.
12. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* 2007; **315**(5815): 1137-1140.
13. Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K. RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep* 2007; **8**(8): 763-769.
14. Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R *et al.* Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol* 2006; **13**(1): 13-21.
15. Borchert GM, Gilmore BL, Spengler RM, Xing Y, Lanier W, Bhattacharya D *et al.* Adenosine deamination in human transcripts generates novel microRNA binding sites. *Hum Mol Genet* 2009; **18**(24): 4801-4807.
16. Mehler MF, Mattick JS. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. *Physiol Rev* 2007; **87**(3): 799-823.

17. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 2009; **324**(5931): 1210-1213.
18. Mattick JS, Mehler MF. RNA editing, DNA recoding and the evolution of human cognition. *Trends Neurosci* 2008; **31**(5): 227-233.
19. Garrett S, Rosenthal JJ. RNA Editing Underlies Temperature Adaptation in K<sup>+</sup> Channels from Polar Octopuses. *Science* 2012.
20. Macbeth MR, Schubert HL, Vandemark AP, Lingam AT, Hill CP, Bass BL. Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* 2005; **309**(5740): 1534-1539.
21. Patterson JB, Samuel CE. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol Cell Biol* 1995; **15**(10): 5376-5388.
22. Englander MT, Dulawa SC, Bhansali P, Schmauss C. How stress and fluoxetine modulate serotonin 2C receptor pre-mRNA editing. *J Neurosci* 2005; **25**(3): 648-651.
23. Wahlstedt H, Daniel C, Enstero M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res* 2009; **19**(6): 978-986.
24. Wulff BE, Nishikura K. Substitutional A-to-I RNA editing. *WIREs RNA* 2010; **1**(1): 90-101.

25. Jepson JE, Reenan RA. RNA editing in regulating gene expression in the brain. *Biochim Biophys Acta* 2008; **1779**(8): 459-470.
26. Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* 2012.
27. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* 2004; **2**(12): e391.
28. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J *et al.* Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proc Natl Acad Sci U S A* 2010; **107**(27): 12174-12179.
29. Toro R, Konyukh M, Delorme R, Leblond C, Chaste P, Fauchereau F *et al.* Key role for gene dosage and synaptic homeostasis in autism spectrum disorders. *Trends Genet* 2010; **26**(8): 363-372.
30. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011; **474**(7351): 380-384.
31. Bourgeron T. A synaptic trek to autism. *Curr Opin Neurobiol* 2009; **19**(2): 231-234.



32. 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.
33. Anney RJ, Kenny EM, O'Dushlaine C, Yaspán BL, Parkhomenka E, Buxbaum JD *et al.* Gene-ontology enrichment analysis in two independent family-based samples highlights biologically plausible processes for autism spectrum disorders. *Eur J Hum Genet* 2011.
34. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010; **466**(7304): 368-372.
35. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, Hill DE *et al.* Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med* 2011; **3**(86): 86ra49.
36. Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T *et al.* Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Arch Gen Psychiatry* 2011.
37. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: A decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* 2011.
38. Grafodatskaya D, Chung B, Szatmari P, Weksberg R. Autism spectrum disorders and epigenetics. *J Am Acad Child Adolesc Psychiatry* 2010; **49**(8): 794-809.

39. Singh M, Zimmerman MB, Beltz TG, Johnson AK. Affect-related behaviors in mice misexpressing the RNA editing enzyme ADAR2. *Physiol Behav* 2009; **97**(3-4): 446-454.
40. Jepson JE, Savva YA, Yokose C, Sugden AU, Sahin A, Reenan RA. Engineered alterations in RNA editing modulate complex behavior in Drosophila: regulatory diversity of adenosine deaminase acting on RNA (ADAR) targets. *J Biol Chem* 2010.
41. Morabito MV, Abbas AI, Hood JL, Kesterson RA, Jacobs MM, Kump DS *et al.* Mice with altered serotonin 2C receptor RNA editing display characteristics of Prader-Willi syndrome. *Neurobiol Dis* 2010; **39**(2): 169-180.
42. Canitano R. Epilepsy in autism spectrum disorders. *Eur Child Adolesc Psychiatry* 2007; **16**(1): 61-66.
43. Veltman MW, Craig EE, Bolton PF. Autism spectrum disorders in Prader-Willi and Angelman syndromes: a systematic review. *Psychiatr Genet* 2005; **15**(4): 243-254.
44. Nakatani J, Tamada K, Hatanaka F, Ise S, Ohta H, Inoue K *et al.* Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism. *Cell* 2009; **137**(7): 1235-1246.
45. Daniels JL, Forssen U, Hultman CM, Cnattingius S, Savitz DA, Feychting M *et al.* Parental psychiatric disorders associated with autism spectrum disorders in the offspring. *Pediatrics* 2008; **121**(5): e1357-1362.

46. Sodhi MS, Burnet PW, Makoff AJ, Kerwin RW, Harrison PJ. RNA editing of the 5-HT(2C) receptor is reduced in schizophrenia. *Mol Psychiatry* 2001; **6**(4): 373-379.
47. Gurevich I, Tamir H, Arango V, Dwork AJ, Mann JJ, Schmauss C. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* 2002; **34**(3): 349-356.
48. Hagerman R, Hoem G, Hagerman P. Fragile X and autism: Intertwined at the molecular level leading to targeted treatments. *Mol Autism* 2010; **1**(1): 12.
49. Bhogal B, Jepson JE, Savva YA, Pepper AS, Reenan RA, Jongens TA. Modulation of dADAR-dependent RNA editing by the Drosophila fragile X mental retardation protein. *Nat Neurosci* 2011; **14**(12): 1517-1524.
50. Peca J, Ting J, Feng G. SnapShot: Autism and the synapse. *Cell* 2011; **147**(3): 706, 706 e701.
51. Stephenson DT, O'Neill SM, Narayan S, Tiwari A, Arnold E, Samaroo HD *et al.* Histopathologic characterization of the BTBR mouse model of autistic-like behavior reveals selective changes in neurodevelopmental proteins and adult hippocampal neurogenesis. *Mol Autism* 2011; **2**(1): 7.
52. Wang X, McCoy PA, Rodriguiz RM, Pan Y, Je HS, Roberts AC *et al.* Synaptic dysfunction and abnormal behaviors in mice lacking major isoforms of Shank3. *Hum Mol Genet* 2011; **20**(15): 3093-3108.

53. Gai X, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Mol Psychiatry* 2011.
54. Farra N, Zhang WB, Pasceri P, Eubanks JH, Salter MW, Ellis J. Rett syndrome induced pluripotent stem cell-derived neurons reveal novel neurophysiological alterations. *Mol Psychiatry* 2012.
55. Durrenberger PF, Fernando S, Kashfi SN, Ferrer I, Hauw JJ, Seilhean D *et al.* Effects of antemortem and postmortem variables on human brain mRNA quality: a BrainNet Europe study. *J Neuropathol Exp Neurol* 2010; **69**(1): 70-81.
56. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**(7057): 376-380.
57. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R *et al.* EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics* 2005; **6**: 232.
58. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 1995; **57**(1): 125–133.
59. Belcher SM, Howe JR. Characterization of RNA editing of the glutamate-receptor subunits GluR5 and GluR6 in granule cells during cerebellar development. *Brain Res Mol Brain Res* 1997; **52**(1): 130-138.

60. Bhalla T, Rosenthal JJ, Holmgren M, Reenan R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat Struct Mol Biol* 2004; **11**(10): 950-956.
61. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol* 2006; **7**(12): 919-931.
62. Abrahams BS, Geschwind DH. Connecting genes to brain in the autism spectrum disorders. *Arch Neurol* 2010; **67**(4): 395-399.
63. Mostofsky SH, Powell SK, Simmonds DJ, Goldberg MC, Caffo B, Pekar JJ. Decreased connectivity and cerebellar activity in autism during motor task performance. *Brain* 2009; **132**(Pt 9): 2413-2425.
64. Palmen SJ, van Engeland H, Hof PR, Schmitz C. Neuropathological findings in autism. *Brain* 2004; **127**(Pt 12): 2572-2583.
65. Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* 2000; **406**(6791): 78-81.
66. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 2008; **9**(5): 341-355.
67. Rueter SM, Dawson TR, Emeson RB. Regulation of alternative splicing by RNA editing. *Nature* 1999; **399**(6731): 75-80.

68. Kawahara Y, Ito K, Ito M, Tsuji S, Kwak S. Novel splice variants of human ADAR2 mRNA: skipping of the exon encoding the dsRNA-binding domains, and multiple C-terminal splice sites. *Gene* 2005; **363**: 193-201.
69. Levanon EY, Hallegger M, Kinar Y, Shemesh R, Djinovic-Carugo K, Rechavi G *et al.* Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res* 2005; **33**(4): 1162-1168.
70. Dracheva S, Lyddon R, Barley K, Marcus SM, Hurd YL, Byne WM. Editing of serotonin 2C receptor mRNA in the prefrontal cortex characterizes high-novelty locomotor response behavioral trait. *Neuropsychopharmacology* 2009; **34**(10): 2237-2251.
71. Du Y, Stasko M, Costa AC, Davisson MT, Gardiner KJ. Editing of the serotonin 2C receptor pre-mRNA: Effects of the Morris Water Maze. *Gene* 2007; **391**(1-2): 186-197.
72. Auerbach BD, Osterweil EK, Bear MF. Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature* 2011; **480**(7375): 63-68.
73. Rossignol DA, Frye RE. Mitochondrial dysfunction in autism spectrum disorders: a systematic review and meta-analysis. *Mol Psychiatry* 2012; **17**(3): 290-314.
74. Garbett K, Ebert PJ, Mitchell A, Lintas C, Manzi B, Mirnics K *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiol Dis* 2008; **30**(3): 303-311.

75. Lainhart JE, Lange N. Increased neuron number and head size in autism. *JAMA* 2011; **306**(18): 2031-2032.
76. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001; **25**(4): 402-408.
77. Devonshire AS, Elaswarapu R, Foy CA. Applicability of RNA standards for evaluating RT-qPCR assays and platforms. *BMC Genomics* 2011; **12**: 118.
78. Anitei M, Stange C, Parshina I, Baust T, Schenck A, Raposo G *et al.* Protein complexes containing CYFIP/Sra/PIR121 coordinate Arf1 and Rac1 signalling during clathrin-AP-1-coated carrier biogenesis at the TGN. *Nat Cell Biol* 2010; **12**(4): 330-340.
79. Riedmann EM, Schopoff S, Hartner JC, Jantsch MF. Specificity of ADAR-mediated RNA editing in newly identified targets. *RNA* 2008; **14**(6): 1110-1118.
80. Nishimoto Y, Yamashita T, Hideyama T, Tsuji S, Suzuki N, Kwak S. Determination of editors at the novel A-to-I editing positions. *Neurosci Res* 2008; **61**(2): 201-206.
81. Dillon C, Goda Y. The actin cytoskeleton: integrating form and function at the synapse. *Annu Rev Neurosci* 2005; **28**: 25-55.
82. Wang Q, Miyakoda M, Yang W, Khillan J, Stachura DL, Weiss MJ *et al.* Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *J Biol Chem* 2004; **279**(6): 4952-4961.

## **Chapter 3:**

# **Relationships between A-to-I Editing and Autism-Implicated Small RNA**

**Alal Eran, Kayla Vatalaro, Jillian McCarthy, Emery N. Brown, Isaac S.  
Kohane, Louis M. Kunkel**



Autism Spectrum Disorders (ASDs) are highly heritable neurodevelopmental disorders of mostly idiopathic etiology. Despite extreme genetic heterogeneity, one recurrent finding is copy number alterations of the imprinted 15q11-13 locus. Two small nucleolar RNA (snoRNA) clusters are transcribed from its paternal allele, *SNORD115* and *SNORD116*, both highly expressed in the brain. While the precise function of these snoRNAs remains to be elucidated, *SNORD115* transcripts are thought to regulate adenosine-to-inosine (A-to-I) RNA editing of the serotonin receptor *HTR2C*, a functional and positional candidate gene shown to be differentially edited in ASD in Chapter 2.

For an initial examination of 15q11 snoRNAs in ASD and their impact on *HTR2C* editing, we used deep targeted 454 sequencing to interrogate *SNORD115* and *SNORD116* transcripts from postmortem cerebella of 25 individuals with ASD and neurotypical controls. We first characterized commonalities and differences in 15q11 snoRNA expression among individuals. Despite non-overlapping expression patterns, cluster-wide *SNORD115* and *SNORD116* expression was found to be tightly correlated, suggestive of coregulation. Sub-cluster quantification detected a gamma-like expression distribution with a few dominant centrally located genes within each cluster. Direct inter-individual comparisons revealed a strong gender bias with a consistent two-fold upregulation in male vs. female cerebella. This finding may contribute to the observed 8-fold risk for ASD in males vs. females lacking 15q11 snoRNA expression. In addition, previously uncharacterized C/D box RNAs were detected, many of which are predicted to function as natural *cis* antisense transcripts to brain expressed genes. Finally, testing the proposed relationships between *SNORD115* expression and *HTR2C* editing revealed no strong linear dependencies, suggesting that A-to-I editing regulation is a lot more complex than previously thought. Together, these findings further implicate 15q11 in ASD and suggest that snoRNA-mediated regulation may be involved in the complex epigenetic basis of ASD in ways other than through A-to-I editing.

## Introduction

Since its discovery 50 years ago<sup>1</sup>, platelet hyperserotonemia has been the most consistent biomarker in ASD<sup>2-5</sup>. Although peripheral serotonin does not cross the blood-brain barrier, the same gene network regulates serotonin homeostasis in the blood and the brain<sup>6</sup>. Several members of this network have been reproducibly implicated in ASD, including *SLC6A4*<sup>7-10</sup>, *ASMT*<sup>11, 12</sup>, *HTR4*<sup>13, 14</sup>, *HTR1B*<sup>15, 16</sup>, *HTR2A*<sup>17, 18</sup>, *HTR2C*<sup>19, 20</sup>, *HTR3A*<sup>21</sup>, *HTR3C*<sup>22, 23</sup>, *HTR3E*<sup>22, 23</sup>, and *HTR5A*<sup>24</sup>, supporting a convergent etiology of serotonin dysregulation in ASD. According to the developmental hyperserotonemia model of ASD, high levels of serotonin acting as a developmental signal during early brain development inhibit the outgrowth of serotonergic neurons in a much studied negative feedback loop<sup>25-27</sup>. Several murine models generated to test this model recapitulate many of the social and behavioral characteristics of the disorder<sup>25, 28</sup>. Further support for this model is provided by recent epidemiological data that associated ASD with *in utero* exposure to drugs that raise blood serotonin levels, such as serotonin reuptake inhibitors<sup>29</sup>. Moreover, several studies suggest that maternal infection, a well-established rare etiology of ASD<sup>30, 31</sup>, often leads to prenatal serotonin dysregulation<sup>32, 33</sup>. Finally, imaging studies have revealed a lower activity of serotonergic systems in the brains of individuals with ASD<sup>34, 35</sup>, with worsening symptoms upon serotonin depletion<sup>36</sup>. Taken together, multiple lines of genomic, biochemical, epidemiological, and functional evidence suggest that serotonergic dysregulation may underlie ASD in many cases.

The serotonin receptor HTR2C, a G-protein coupled receptor, undergoes A-to-I RNA editing in five sites, which dramatically alters its G-protein binding activity<sup>37</sup> and uncouples the relationships between serotonin levels and downstream postsynaptic signaling<sup>38</sup>. In Chapter 2, we compared cerebellar *HTR2C* editing levels in individuals with ASD and matched neurotypical individuals. Although our findings warrant further replication in a larger cohort, we detected significant differences in editing-mediated protein isoforms in individuals with ASD, suggesting that altered A-to-I editing of HTR2C

transcripts may contribute to the serotonergic alterations in ASD. Here we wish to examine the regulation of this editing by small RNA genes transcribed from 15q11, one of the most commonly mutated loci in ASD<sup>39</sup>.

It has been suggested that the small nucleolar RNA cluster *SNORD115* may regulate HTR2C editing, via sequence complementarity to and around its editing sites<sup>40, 41</sup>. This neuron-specific<sup>42, 43</sup> C/D box RNA cluster originates from one of the genomic regions most commonly altered in ASD, 15q11<sup>39</sup>. While its precise function remains to be elucidated, mice lacking *Snord115* show serotonin-dependent behavioral changes and *Htr2c* editing alterations<sup>44</sup>. Mice heterozygous for a paternal duplication of 15q11-13 display social abnormalities and changes in *htr2c* editing<sup>40</sup>.

To characterize *SNORD115* expression patterns in a human brain population and examine their relations to HTR2C A-to-I editing, we used deep targeted sequencing to accurately quantify and directly compare sub-cluster expression of *SNORD115* families in postmortem cerebella from individuals with ASD and matched neurotypical individuals. We also profiled the neighboring *SNORD116* cluster, which is thought to be the critical region for the core symptoms of Prader-Willi syndrome, a neurodevelopmental disorder often comorbid to ASD<sup>45</sup>. We report differences and commonalities in 15q11 snoRNA expression across individuals, including a strong gender bias, which could help explain the 8-fold higher risk for ASD in males lacking 15q11 snoRNA expression, as opposed to females carrying the same mutation<sup>45</sup>. Using existing data on HTR2C editing from the same cohort<sup>46</sup>, we tested the proposed relationships between *SNORD115* expression and HTR2C editing but found no strong linear relationships, suggesting that A-to-I editing regulation is a lot more complex than previously thought.

## **Materials and Methods**

### **Subjects**

Fresh-frozen cerebellar samples of individuals with nonsyndromic autism and neurotypical individuals were obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank and the Harvard Brain Tissue Resource Center, through the Autism Tissue Program (**Supplementary Table 1**). Genotyping data from the National Database of Autism Research collections NDARCOL0001855 and NDARCOL0001870 were used to confirm that the studied samples lack large 15q copy number alterations.

### **Molecular Methods**

#### **RNA isolation and quality assurance**

RNA was isolated and assessed as previously described<sup>46</sup>. Only samples with RNA Integrity Number (RIN) >7 were included in the study (**Supplementary Table 1**).

#### **SNORD-targeted library preparation**

**Generation of a normalization spike in.** A synthetic SNORD was designed to have similar GC content to the human SNORD115 and 116 genes (42%), be amplified using the primers described below, be of equivalent size to the longer SNORD116 genes (94 nt), and not match the human genome. The spike in sequence, 5'-TGGGTCGATGATGAGAAGGTTGAGCTTAGTCCTCTTCAGCTAGTTGTGATGACTTATTAATATCA TTTGCAATACCTTTAACGCTGAGGCCCA-3', was in-vitro transcribed using Ambion's MEGashortscript™ T7 Kit (Life Technologies, Grand Island, NY, USA), DNase treated (Life Technologies) and purified by 8% urea-polyacrylamide gel electrophoresis (PAGE), following the manufacturer's protocol. The template oligo (Integrated DNA Technologies, Coralville, Iowa, USA) was 5'-

TAATACGACTCACTATAGGGTGGGTCGATGATGAGAAGGTTGAGCTTAGTCCTCTTCAGCTAGT  
TGTGATGACTTATTAATATCATTGCAATACCTTTAACGCTGAGGCCCA-3'.

**cDNA synthesis.** Random hexamers primed cDNA was synthesized from 250ng of each triplicate total RNA prep spiked with 2 fmol of the synthetic SNORD. Invitrogen's Superscript III First-Strand Synthesis System for RT-PCR was utilized, following the manufacturer's recommendations (Life Technologies).

**SNORD115 and SNORD116 selection.** The Refseq sequences of all human *SNORD115* genes (n=48) and *SNORD116* genes (n=30) were extracted from NCBI ([ftp://ftp.ncbi.nlm.nih.gov/refseq/H\\_sapiens/](ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/)) and aligned using ClustalW2<sup>47</sup>. Conserved stem sequences were identified at the 5' and 3' ends of all sequences (**Supplementary Figure 1**), and used to design degenerate primers that amplify all 78 genes. PCR with 5'-TGGRTCRCATGATGAS -3' and 5'-TGGRCCTCAGY-3' yielded two major products of ~82bp and ~95bp, corresponding to *SNORD115* and *SNORD116* genes, respectively (**Supplementary Figure 2**).

**Gene-specific amplification.** Each triplicate cDNA was amplified using the degenerate SNORD primers tagged by a 5' 10bp sequence that served as the basis for a second sample-specific PCR. Forward primer 5'-TCGATCAGCATGGRTCRCATGATGAS-3' and reverse 5'-TACGATGCGTTGGRCCTCAGY-3' were synthesized by IDT and used in 35 cycles of PCR at 46°C annealing with Invitrogen's Accuprime PFX system, following the manufacturer's protocol (Life Technologies).

**PAGE purification of PCR products** was performed as previously described<sup>46</sup>.

**Sample-specific amplification.** 100ng of the purified gene-specific triplicate PCR products were amplified to add a sample specific MID barcode fused to the 454 sequencing primer<sup>46</sup>. Matched samples were handled together in eight cycles of PCR at 50°C annealing using Life Technologies' Accuprime PFX system, following the manufacturer's recommendations. 20pmol of primers were used in each reaction.

**Quantitation and equimolar pooling.** PAGE-purified sample-specific PCR products were quantitated using Quant-iT™ PicoGreen® dsDNA Reagent (Life Technologies), to enable equimolar pooling for multiplexed sequencing. Samples were diluted 1:20 and two volumes of each dilution were quantitated using the Victor3 Multiplate Reader (PerkinElmer, Waltham, MA, USA), in black round-bottom plates (Corning), following the manufacturer protocol. Lambda DNA samples were used to create a standard curve. A sample's quantitation was considered successful if the coefficient of variation (CV) between the two dilutions was <10%. Equimolar amounts of neurotypical and ASD samples were then joined into two 5ng/uL pools.

### **454 sequencing**

Bidirectional GS FLX sequencing was performed by the 454 Life Sciences Sequencing Center (Branford, CT, USA). ASD and neurotypical pools were sequenced on opposite sides of a 2-region PicoTiterPlate, yielding 82,964 SNORD reads (**Supplementary Figure 3**).

## **Computational Methods**

### **Read and quality score clipping**

As the 454 reads are longer than the targeted SNORDs, 454 primers, sample specific barcodes, universal tags, and degenerate SNORD-specific primers were searched within the reads using perl regexps, and trimmed such that only the primer-to-primer insert of each read was kept for further analyses. Each 454 qual string was also trimmed at the positions corresponding to its read's primer-to-primer insert.

## Read alignment

Clipped reads were aligned to the human genome using Bowtie 2<sup>48</sup>, based on their quality scores. The human genome assembly hg19 was downloaded from UCSC Genome Browser<sup>49</sup> and indexed along with the synthetic SNORD spike in sequence. Bowtie 2 was used in end-to-end mode, using *very-sensitive* parameter settings, and reporting all secondary alignments. 65,946 reads were successfully mapped to 15q11, 14,865 to the synthetic SNORD, 220 to other regions in the human genome, and 1,930 reads were unmapped. Unmapped reads were very short, with an average post-clipping length of 20.4bp. Samtools was used for BAM and SAM file handling. Custom perl scripts were used for all downstream analyses.

## Simulations-based SNORD family definition, expression quantitation, and performance evaluation

The *SNORD115* and *SNORD116* gene clusters are characterized by a high degree of redundancy, making their accurate sub-cluster quantification a challenge. Several types of indistinguishable paralogs exist, including perfectly duplicated genes (e.g. *SNORD115-17*, *SNORD115-18*, and *SNORD115-19*), and highly homologous genes whose single nucleotide difference was covered by a primer and clipped from the reads (e.g. *SNORD115-1*, *SNORD115-13*, and *SNORD115-16*). These were identified by self-BLAST<sup>50</sup> and grouped into SNORD families. Expression quantitation was then performed at the SNORD family level, 85% of which contained a single gene.

To design the optimal quantitation strategy and test its performance, synthetic 454 reads corresponding to the 78 Refseq *SNORD115* and *SNORD116* genes were generated according to empirically derived error rates. A homopolymer error, the most common type of pyrosequencing error, was modeled by a normal distribution with a mean of the homopolymer length and variance of  $(0.17)^2 * \text{homopolymer length}^{51}$ . The second most common error, a nucleotide insertion, was applied to uniformly sampled reads with a

base pair probability of 0.29%<sup>52</sup>. Deletion probability was 0.27%<sup>52</sup> and mismatches were applied with 0.12% to each base<sup>52</sup>.

Several expression quantitation strategies were tested on a training set of 100,000 synthetic reads, aiming to optimize the mapping and quantification accuracy, which was assessed in comparison to the true identity and frequency of the reads. Examined considerations included utilizing the single best alignment vs. multiple top alignments with  $\leq 1$  edit distance, using a simple sum of reads vs. a weighted sum, and weighing by alignment score as a measure of mapping quality vs. weighing by read entropy as a measure of information content, defined as

$$H(read) = - \sum_{i \in \text{read alignments}} \frac{AS_i}{\sum_{j \in \text{alignments}} AS_j} * \log\left(\frac{AS_i}{\sum_{j \in \text{alignments}} AS_j}\right).$$

Different combinations of these considerations were examined and the most accurate expression quantitation approach was found to be

$$Expression(SNORD \text{ family}_i) = \sum_{r \text{ with } \max(AS_r) \in SNORD \text{ family}_i} |\max(AS_r)|^{-1}$$

That is, per SNORD family  $i$ , count all reads  $r$  whose single best alignment maps to at least one locus of SNORD family  $i$ , normalized by the total number of SNORD families best mapped to  $r$ .

A few paralogs could not be distinguished with sufficient certainty and were repeatedly cross-mapped during training. They were subsequently joined to SNORD families as detailed in **Supplementary Table 2**. The final detection resolution was at the level of 29 SNORD115 families and 26 SNORD116 families. At this resolution, the noted expression quantitation approach was tested on two additional independent simulated datasets and found to be robustly accurate. Its mapping accuracy was  $99.37 \pm 0.02\%$  and  $R^2$  goodness of fit to the true expression  $0.999 \pm 3.33e-05$ . (**Supplementary Figure 4**).



### **Secondary structure prediction**

Vienna RNA version 2.0<sup>53</sup> was used to compute minimal fold energies and plot the predicted secondary structures.

### **Linearity analyses**

Confidence intervals around correlation coefficients were calculated using Fisher's rho to Z transformation. Linear regressions were assessed using F tests.

### **Differential expression analyses**

Bayesian comparisons were used to infer expression differences between groups, including between individuals with ASD and neurotypical individuals, and males vs. females. In each comparison, the mean expression level of each group was bootstrapped and the difference between the resampled group means was calculated 10,000 times. Differential expression was considered significant if the 95% credibility interval (CI) of the group differences excluded 0.

### **Unsupervised learning**

K-means clustering and principal component analysis were used to examine the internal structure of the data.

### **Multiple testing correction**

P-values were corrected for multiple testing to ensure that the Benjamini-Hochberg false discovery rate<sup>54</sup> of this entire study is below 0.05.

## Results

### Accurate quantitation of *SNORD115* and *SNORD116* transcripts in human cerebellum

For a high-resolution view of the imprinted 15q11 snoRNAs in a human brain population, targeted ultradeep 454 sequencing was used to quantify *SNORD115* and *SNORD116* transcripts in postmortem cerebella from 25 neurotypical individuals and individuals with ASD (**Figure 1**). *SNORD115* and *SNORD116* cluster members were selected based on their conserved stem sequences (**Supplementary Figure 1**) and sequenced to an average of 3500x (**Supplementary Figure 3**). In all, 29 *SNORD115* families and 26 *SNORD116* families were detected with 99.9% accuracy (**Supplementary Figure 4**). SNORD families group highly paralogous genes that cannot be distinguished with sufficient certainty, 85% of which contain a single gene (**Supplementary Table 2**). Such family-level resolution facilitated confident sub-cluster analysis, while sequencing along a normalization spike-in enabled direct inter-individual comparisons.

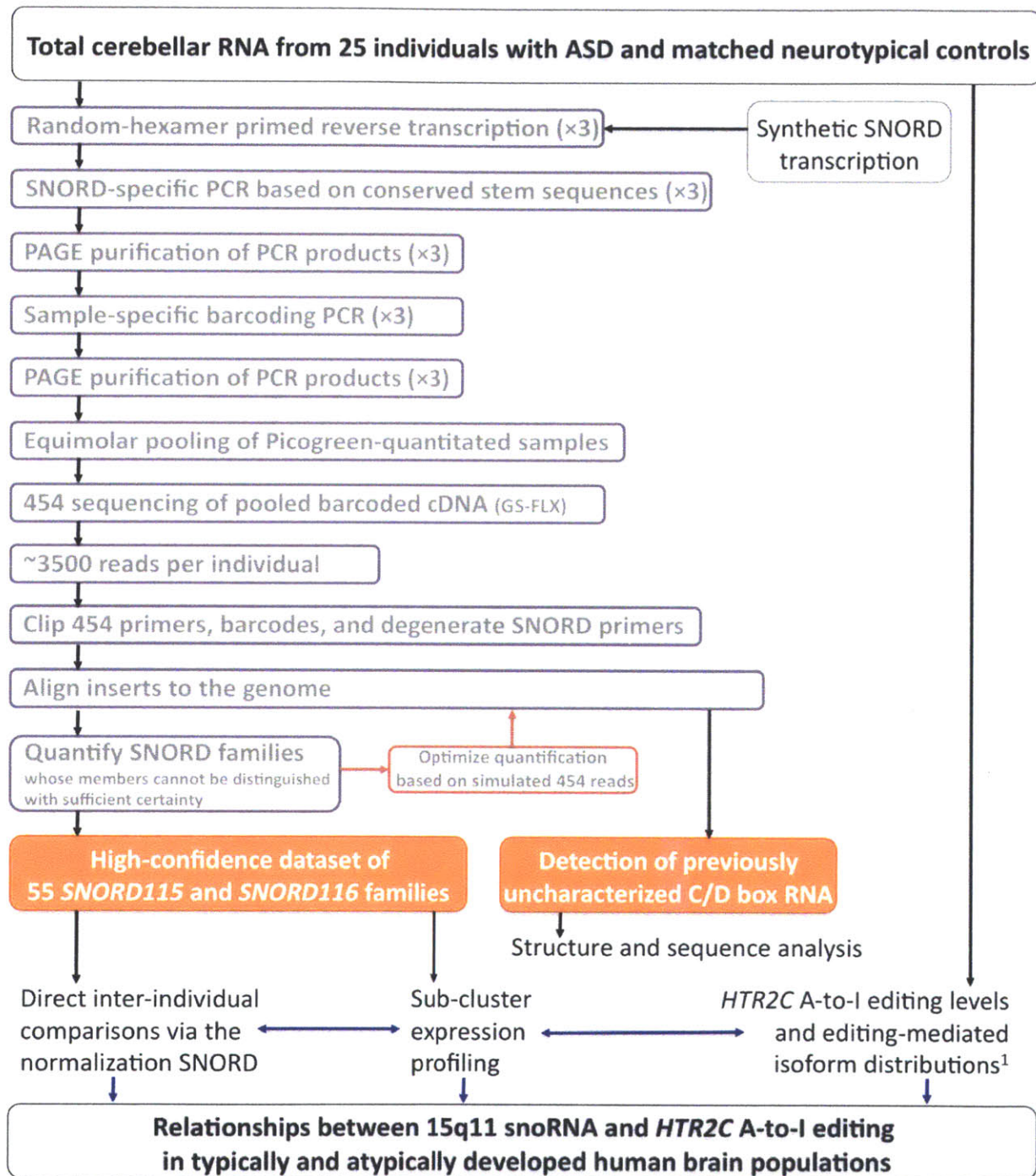
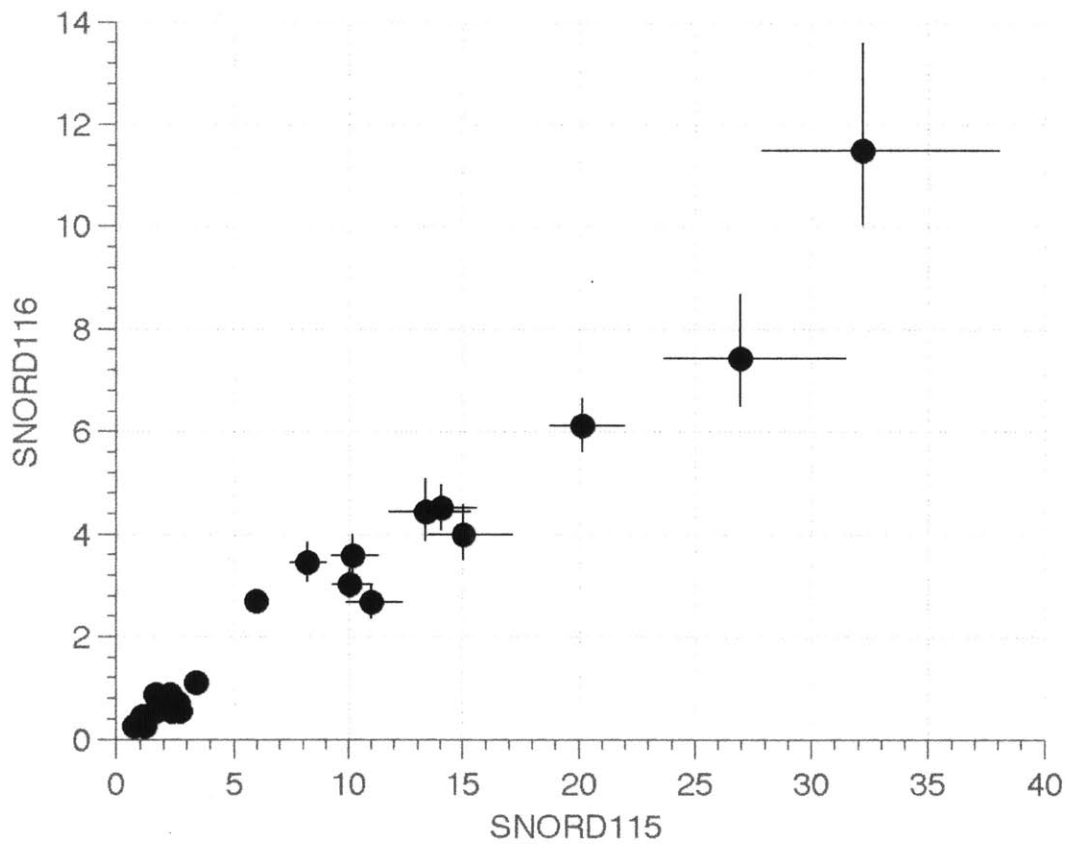


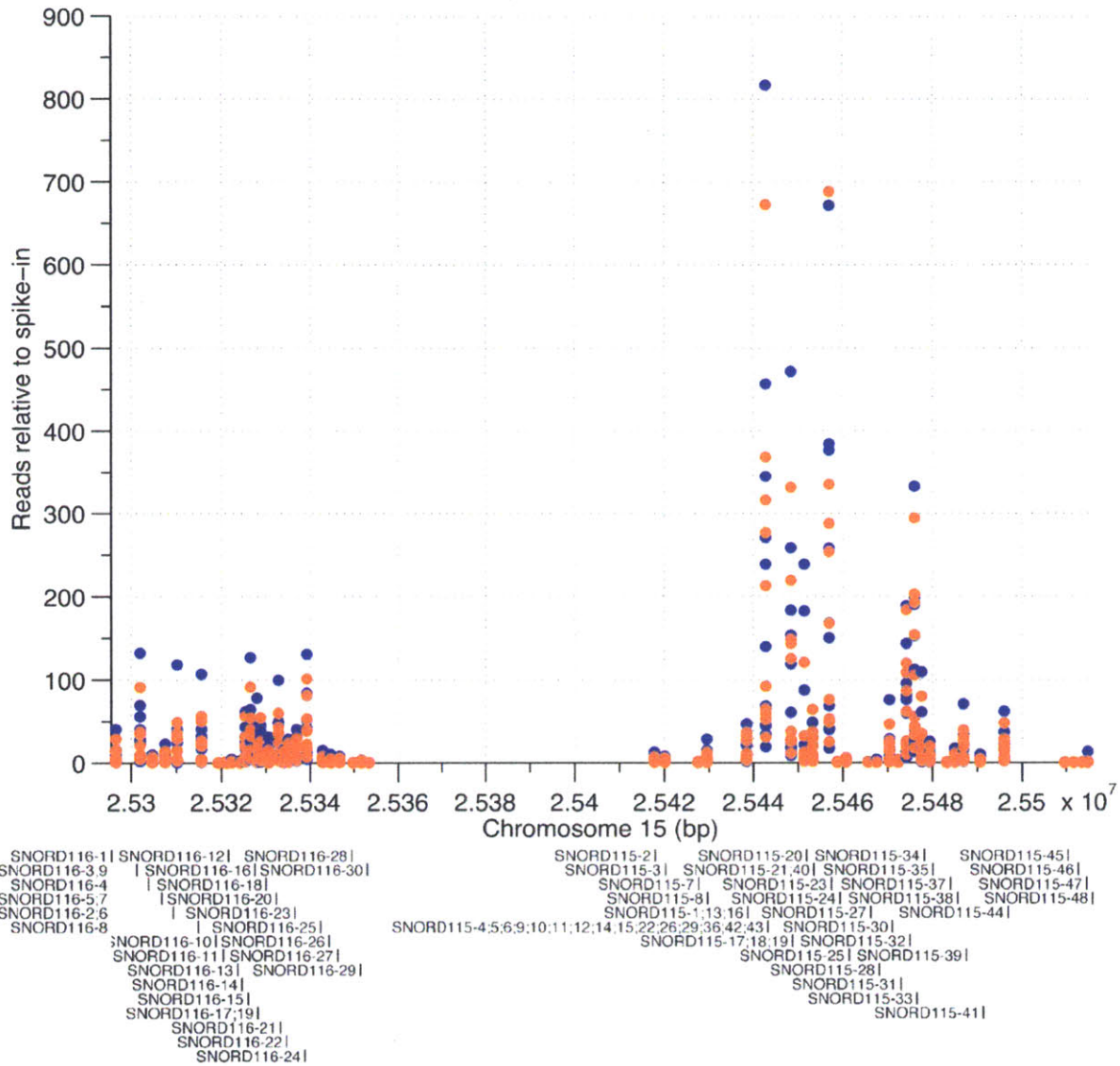
Figure 1 Overview of the experimental approach

We first compared the overall levels of *SNORD115* and *SNORD116* transcripts, and found that they are tightly correlated across individuals (**Figure 2**,  $\rho=0.982$ , 95% CI=[0.977, 0.985],  $p=7.3e-16$ ). Although these clusters have non-overlapping expression patterns<sup>43</sup>, they may be coregulated in the cerebellum.



**Figure 2** Tight correlation between overall *SNORD115* and *SNORD116* expression in human cerebellum. Reads mapping to any *SNORD115* (X axis) or *SNORD116* family (Y axis) were summed per sample and normalized by the number of reads mapping to the spike-in. Resampling-based 95% confidence intervals are shown for each measurement. A strong linear relationship exists between overall gene expression of the two major 15q11 snoRNA clusters across individuals ( $\rho=0.982$ , 95% CI=[0.977,0.985],  $p=7.3e-16$ ).

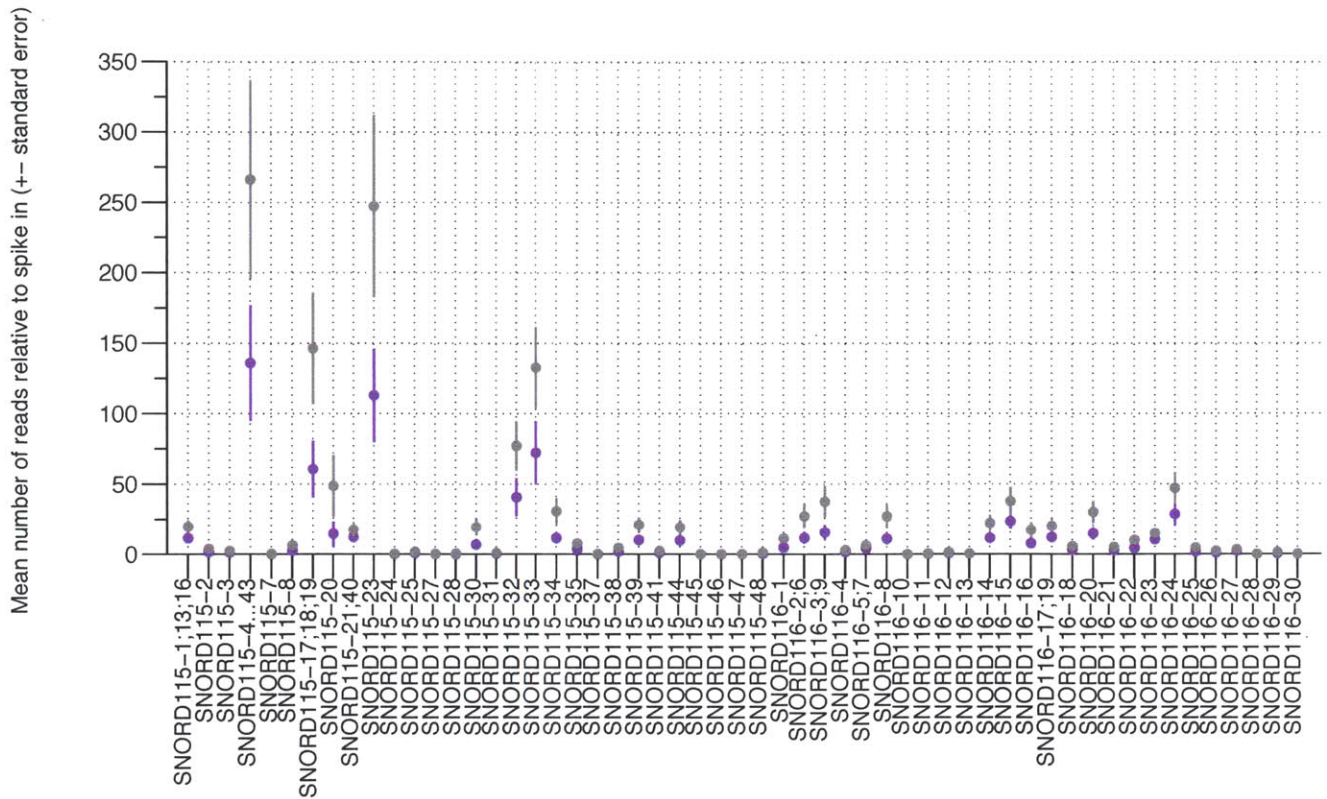
Next, intra-cluster expression distributions were inspected. In both clusters, about half of the reads belonged to 3-4 predominant single genes. *SNORD115-23*, *SNORD115-32* and *SNORD115-33* comprised  $45\pm 1.48\%$  of the overall *SNORD115* transcripts across individuals, while *SNORD116-15*, *SNORD116-20*, *SNORD116-23*, and *SNORD116-24* constituted  $45\pm 2.2\%$  of the *SNORD116* cluster expression. Within each cluster, these dominant genes are centrally located (**Figure 3**). No obvious sequence characteristics would make them preferred PCR templates over other genes (**Supplementary Figure 1**). The overall 15q11 snoRNA expression patterns could be described by a gamma distribution (**Supplementary Figure 5**).



**Figure 3.** Cerebellar *SNORD115* and *SNORD116* expression in neurotypical individuals (blue) and individuals with ASD (orange), according to their genomic organization. Sub-cluster expression was calculated at the SNORD family level (Supplementary Table 2) relative to a normalization spike-in, and plotted according to the median genomic location of all family members, noted at the bottom. The *SNORD116* cluster is the smaller upstream one (chr15:25296623-25353499) and the *SNORD115* cluster is the downstream larger one (chr15:25415870-25515005). Predominant genes in each cluster tend to be centrally located. Transcripts originating from the 3' ends tend to be underrepresented.

## Sex specific expression of 15q11 snoRNA

*SNORD115* and *SNORD116* expression levels were next compared between individuals with ASD and matched neurotypical individuals, demonstrating largely similar expression patterns (**Supplementary Figure 6**). However, an unsupervised analysis detected a strong internal structure of four clusters, which were mostly separated by sex and affected status (**Supplementary Figure 7**). A subsequent comparison of male vs. female expression identified a consistent 2-fold upregulation of every *SNORD115* and *SNORD116* family in males (**Figure 4** and **Supplementary Figure 8**). Group means were resampled 10,000 times and gender differences in SNORD family expression were considered significant if the 95% credibility interval (CI) of the difference excluded 0. The most differentially expressed SNORD families were *SNORD-115-2* (95% CI of the difference 0.08-4.79), *SNORD115-17;18;19* (95% CI 6.2-171.46), *SNORD115-30* (1.85-25.87), *SNORD115-31* (0.15-1.24), *SNORD115-34* (1.74-39.19), *SNORD116-3;9* (0.85-45.57), *SNORD116-8* (0.06-35.16), *SNORD116-22* (0.39-11.63), *SNORD116-25* (1.01-5.59), and *SNORD116-29* (0.22-1.58). This previously uncharacterized sex-specific expression pattern could be important in delineating the role of 15q11 in ASD, the most commonly implicated locus in a male dominant disorder. With the current sample size, male-specific comparisons are underpowered (**Supplementary Figure 9**), and further work should examine 15q11 SNORD expression in larger cohorts of male brains.

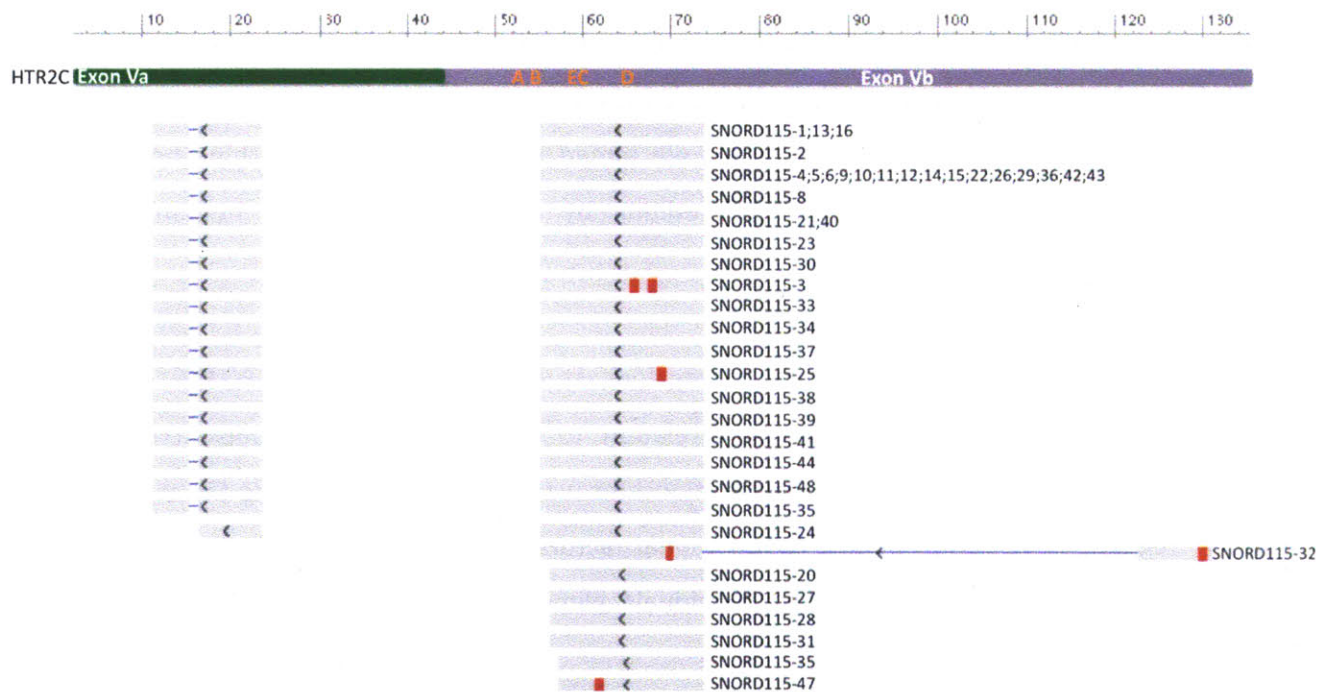


**Figure 4** Consistent 2-fold overexpression of imprinted 15q11 SNORDs in male (gray) vs. female (purple) cerebellum. Every *SNORD115* and *SNORD116* family tested was upregulated in male individuals, by an average of  $2.19 \pm 0.17$  fold. Shown is the mean number of reads relative to the spike in among same-sex individuals, per SNORD family  $\pm$  the standard error. Consult Supplementary Figure 8 for individual expression levels.



## Relationships between *SNORD115* expression and *HTR2C* editing in autism

In Chapter 2 we quantified A-to-I editing and editing-mediated *HTR2C* isoform distributions in a human brain population, comparing neurotypical individuals to individuals with ASD. Here we examine the proposed relationships between *SNORD115* expression and *HTR2C* A-to-I editing in this population. As depicted in **Figure 5**, most *SNORD115* families contain a complementary sequence to the edited region of *HTR2C*. It has therefore been suggested that their expression might affect *HTR2C* editing. However, the precise relationships between *SNORD115* expression and *HTR2C* editing remain unclear, and several contradictory findings from small-scale studies have been reported.



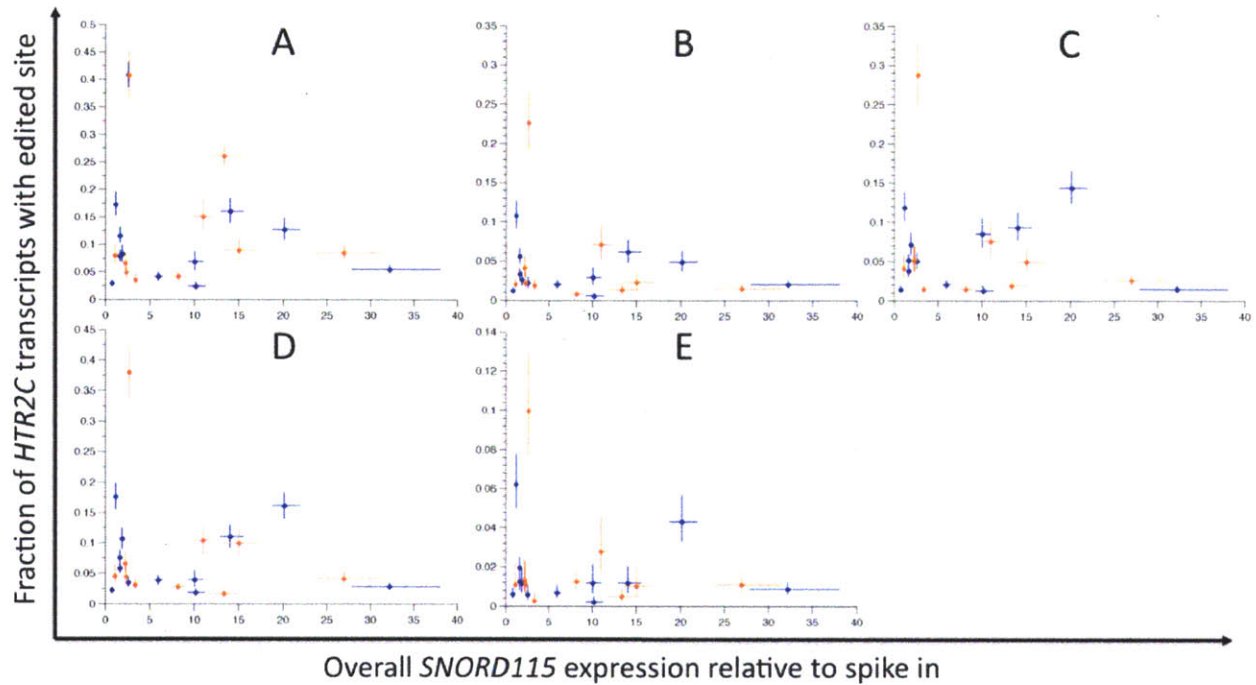
**Figure 5** Alignment of *SNORD115* families to *HTR2C*. Most *SNORD115* families share sequence complementarity to the edited region of *HTR2C*, with the exception of *SNORD115-7*, *SNORD115-17;18;19*, and *SNORD115-46*. Specifically, they are predicted to base-pair to the unedited states of E, C, and D, shown in orange, corresponding to *HTR2C* protein isoforms VNI, MNI, and INI. Additionally, 19 *SNORD115* families align to

both exon Va (green) and Vb (purple), and therefore have a potential to affect alternative splicing by binding to both exons. Arrows signify 5' to 3' orientation, red squares denote mismatches.

Vitali et al. (2005) were the first to examine the relationship between Htr2c editing and Snord115 expression<sup>41</sup>. They reported that overexpression of a generic mouse Snord115 leads to reduced editing at the C and D sites in neuronal culture. However, in 2009 Nakatani et al. showed that constitutive transgenic overexpression of Snord115 leads to increased editing in vivo<sup>40</sup>. Shortly thereafter, Snord115 knockout in another strain resulted in a suspiciously summed “overall increased editing across all five sites”<sup>44</sup>.

In humans, Kishore and Stamm first showed that four individuals with PWS harboring a deletion of the SNORD115 cluster show increased editing at sites A-C but not D, as compared to two control individuals<sup>55</sup>. However Glatt-Deeley et al. later showed that there are no consistent HTR2C editing differences between two other individuals with PWS and four other controls. Here we aimed to clarify the relationships between SNORD115 expression and HTR2C editing by surveying the largest human brain population to date.

First, for a direct comparison to previous studies, the overall *SNORD115* cluster expression level was examined in relation to single site editing levels across all individuals (**Figure 6**). Nonlinear relationships were identified, supporting the complex regulation of HTR2C editing by additional factors. Of note, overall HTR2C expression levels were rather similar across individuals, eliminating potential stoichiometric biases (**Supplementary Figure 10**).

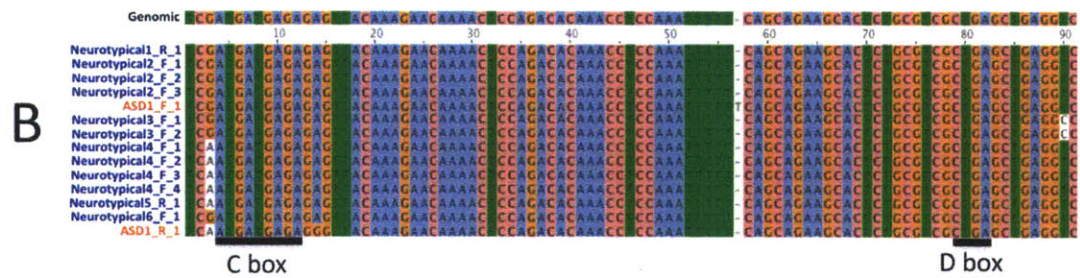
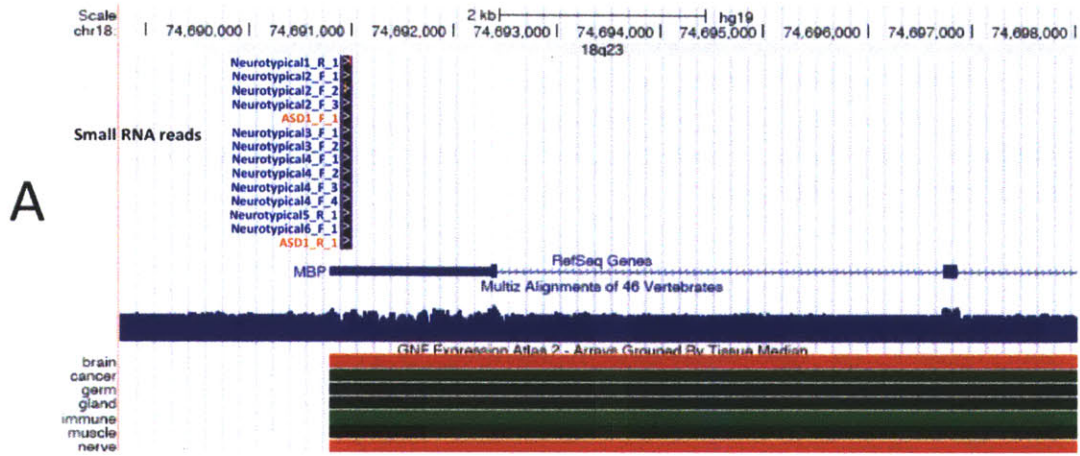


**Figure 6** Nonlinear relationships between overall SNORD115 expression levels and *HTR2C* editing in postmortem cerebella from 25 individuals with ASD (orange) and neurotypical controls (blue). Shown is the point estimate for each measurement and 95% confidence interval around it.

Next, the potential regulatory role of those *SNORD115* families aligned to the edited *HTR2C* region was examined by multivariate linear regression. One significant relationship emerged, between the expression of *SNORD115-27* and *SNORD115-8* and the usage of *HTR2C* ISV isoform, which corresponds to editing exclusively the C and D sites ( $p= 2.67e-04$ , F-test). However, the functional impact of this relationship is likely rather minor because of the low relative frequencies of the linked variables within the cellular milieu of *SNORD115* genes and *HTR2C* isoforms.

### **Detection of previously uncharacterized C/D box RNAs**

The primers used to select *SNORD115* and *SNORD116* transcripts also amplified other previously uncharacterized C/D box RNAs that share similar stem sequences. As exemplified in **Figure 7**, to be considered a putative SNORD, a newly detected transcript had to (1) be supported by more than 10 bidirectional reads from multiple samples; (2) have all mapped reads contain perfectly aligned C and D boxes (RUGAUGA and CUGA, respectively); and (3) have a stable secondary structure with a predicted minimal fold energy (MFE)  $\leq -10$  Kcal/mol. Of 21 confidently detected, previously uncharacterized C/D box RNAs, 13 (62%) originate from the opposite strand of a conserved brain expressed 3' untranslated regions (UTRs), including that of myelin basic protein (MBP, **Figure 7A**), and glial high affinity glutamate transporter member 2 (SLC1A2). These may commonly function as *cis*-antisense transcripts, fine-tuning the expression of brain mRNAs<sup>56,57</sup>.



**Figure 7** A previously uncharacterized conserved C/D box RNA is *cis*-antisense to myelin basic protein (*MBP*). (A) Fourteen reads from six neurotypical individuals and two individuals with ASD are complementary to the brain-expressed *MBP* 3' UTR in *cis*. Shown is the UCSC Genome Browser<sup>49</sup> view of the *MBP* locus on 18q23, a highly conserved region indicated by whole genome alignments of 46 vertebrates. While the SNORD reads are aligned to the forward strand, arrows on *MBP* represent its reverse orientation, and its last noncoding region, its 3' UTR, is complementary to the small RNA reads. At the bottom, alignment of probe sequences used by the Human Gene

Expression Atlas of the Genomics Institute of the Novartis Research Foundation (GNF) shows that MBP is highly expressed in the human nervous system<sup>58</sup>. (B) Multiple sequence alignment of the reads to the genome shows that all harbor the C and a D box motifs (RUGAUGA and CUGA, respectively). (C) Minimal fold energy (MFE) prediction identifies a highly stable putative structure resembling that of the known SNORDs, with free energy of 18.79 Kcal/mol.

## Discussion

In this study we examined the proposed regulation of *HTR2C* editing by *SNORD115*, and the overall contribution of 15q11 snoRNAs to ASD. We found a strong gender bias underlying *SNORD115* and *SNORD116* expression in human cerebella, showing a consistent 2-fold upregulation in males vs. females. ASD occurs in males nearly five times as much as in females, and 15q11 is one of its most commonly implicated loci. Males with Prader Willi syndrome, in which the *SNORD115* and *SNORD116* clusters are not expressed, have a 40% chance of comorbid ASD, while females with the same genomic abnormality have a 5% chance of comorbid ASD<sup>45</sup>. This 8:1 male-to-female risk ratio is nearly double that of the general population. Therefore, the finding of gender-specific *SNORD* expression could be important in delineating the role of 15q11 in ASD and must be replicated in an independent cohort.

Despite the sequence complementarity between *SNORD115* genes and *HTR2C*, our study does not support a strong dependency between the two. Before further claims are made, it would be good to test the potential interaction of ADARB1, *HTR2C* and *SNORD115*, for example by CLIP-seq. In light of these results and other small RNA findings, we speculate that the main role of *SNORD115* is in maintaining the imprinted state by regulating chromatin modification.

## **Acknowledgments**

We thank Prof. David Bartel, our wonderful lab mates, Ami Levy-Moonshine, Ofra Amir, and Elena Helman for their tremendous help. Human tissue was obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, contract HHSN275200900011C, Re. No. N01-HD-9-0011, and from The Harvard Brain Tissue Resource Center (HBTRC), through the Autism Tissue Program. HBTRC is supported by Grant MH068855. This study was generously supported by the Nancy Lurie Marks Family Foundation, The Roche Applied Science Sequencing Grant Program, Autism Speaks, Simons Foundation, and NIH Grant 1R01MH085143-01.

## References

1. Schain RJ, Freedman DX. Studies on 5-hydroxyindole metabolism in autistic and other mentally retarded children. *J Pediatr* 1961; **58**: 315-320.
2. Mulder EJ, Anderson GM, Kema IP, de Bildt A, van Lang ND, den Boer JA *et al.* Platelet serotonin levels in pervasive developmental disorders and mental retardation: diagnostic group differences, within-group distribution, and behavioral correlates. *J Am Acad Child Adolesc Psychiatry* 2004; **43**(4): 491-499.
3. Cook EH, Leventhal BL. The serotonin system in autism. *Curr Opin Pediatr* 1996; **8**(4): 348-354.
4. Hranilovic D, Bujas-Petkovic Z, Vragovic R, Vuk T, Hock K, Jernej B. Hyperserotonemia in adults with autistic disorder. *J Autism Dev Disord* 2007; **37**(10): 1934-1940.
5. Lam KS, Aman MG, Arnold LE. Neurochemical correlates of autistic disorder: a review of the literature. *Res Dev Disabil* 2006; **27**(3): 254-289.
6. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M *et al.* Spatio-temporal transcriptome of the human brain. *Nature* 2011; **478**(7370): 483-489.
7. Sutcliffe JS, Delahanty RJ, Prasad HC, McCauley JL, Han Q, Jiang L *et al.* Allelic heterogeneity at the serotonin transporter locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors. *Am J Hum Genet* 2005; **77**(2): 265-279.
8. International Molecular Genetic Study of Autism C. A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *Am J Hum Genet* 2001; **69**(3): 570-581.
9. Yonan AL, Alarcon M, Cheng R, Magnusson PK, Spence SJ, Palmer AA *et al.* A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet* 2003; **73**(4): 886-897.
10. Spence SJ, Cantor RM, Chung L, Kim S, Geschwind DH, Alarcon M. Stratification based on language-related endophenotypes in autism: attempt to replicate reported linkage. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**(6): 591-598.



11. 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): 863-885.
12. Melke J, Goubran Botros H, Chaste P, Betancur C, Nygren G, Anckarsater H *et al.* Abnormal melatonin synthesis in autism spectrum disorders. *Mol Psychiatry* 2008; **13**(1): 90-98.
13. Hu VW, Addington A, Hyman A. Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published GWAS data. *PLoS One* 2011; **6**(4): e19067.
14. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011; **474**(7351): 380-384.
15. Orabona GM, Griesi-Oliveira K, Vadasz E, Bulcao VL, Takahashi VN, Moreira ES *et al.* HTR1B and HTR2C in autism spectrum disorders in Brazilian families. *Brain Res* 2009; **1250**: 14-19.
16. Autism Genome Project C, Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 2007; **39**(3): 319-328.
17. Kazek B, Huzarska M, Grzybowska-Chlebowczyk U, Kajor M, Ciupinska-Kajor M, Wos H *et al.* Platelet and intestinal 5-HT<sub>2A</sub> receptor mRNA in autistic spectrum disorders - results of a pilot study. *Acta Neurobiol Exp (Wars)* 2010; **70**(2): 232-238.
18. Steele MM, Al-Adeimi M, Siu VM, Fan YS. Brief report: A case of autism with interstitial deletion of chromosome 13. *J Autism Dev Disord* 2001; **31**(2): 231-234.
19. Piton A, Gauthier J, Hamdan FF, Lafreniere RG, Yang Y, Henrion E *et al.* Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia. *Mol Psychiatry* 2011; **16**(8): 867-880.
20. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010; **466**(7304): 368-372.

21. Anderson BM, Schnetz-Boutaud NC, Bartlett J, Wotawa AM, Wright HH, Abramson RK *et al.* Examination of association of genes in the serotonin system to autism. *Neurogenetics* 2009; **10**(3): 209-216.
22. Rehnstrom K, Ylisaukko-oja T, Nummela I, Ellonen P, Kempas E, Vanhala R *et al.* Allelic variants in HTR3C show association with autism. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**(5): 741-746.
23. Allen-Brady K, Miller J, Matsunami N, Stevens J, Block H, Farley M *et al.* A high-density SNP genome-wide linkage scan in a large autism extended pedigree. *Mol Psychiatry* 2009; **14**(6): 590-600.
24. Coutinho AM, Sousa I, Martins M, Correia C, Morgadinho T, Bento C *et al.* Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. *Hum Genet* 2007; **121**(2): 243-256.
25. McNamara IM, Borella AW, Bialowas LA, Whitaker-Azmitia PM. Further studies in the developmental hyperserotonemia model (DHS) of autism: social, behavioral and peptide changes. *Brain Res* 2008; **1189**: 203-214.
26. Shemer AV, Azmitia EC, Whitaker-Azmitia PM. Dose-related effects of prenatal 5-methoxytryptamine (5-MT) on development of serotonin terminal density and behavior. *Brain Res Dev Brain Res* 1991; **59**(1): 59-63.
27. Whitaker-Azmitia PM, Druse M, Walker P, Lauder JM. Serotonin as a developmental signal. *Behav Brain Res* 1996; **73**(1-2): 19-29.
28. Veenstra-VanderWeele J, Muller CL, Iwamoto H, Sauer JE, Owens WA, Shah CR *et al.* Autism gene variant causes hyperserotonemia, serotonin receptor hypersensitivity, social impairment and repetitive behavior. *Proc Natl Acad Sci U S A* 2012; **109**(14): 5469-5474.
29. Croen LA, Grether JK, Yoshida CK, Odouli R, Hendrick V. Antidepressant use during pregnancy and childhood autism spectrum disorders. *Arch Gen Psychiatry* 2011; **68**(11): 1104-1112.
30. Atladottir HO, Henriksen TB, Schendel DE, Parner ET. Autism after infection, febrile episodes, and antibiotic use during pregnancy: an exploratory study. *Pediatrics* 2012; **130**(6): e1447-1454.

31. Chess S. Autism in children with congenital rubella. *J Autism Child Schizophr* 1971; **1**(1): 33-47.
32. Fatemi SH, Reutiman TJ, Folsom TD, Huang H, Oishi K, Mori S *et al.* Maternal infection leads to abnormal gene regulation and brain atrophy in mouse offspring: implications for genesis of neurodevelopmental disorders. *Schizophr Res* 2008; **99**(1-3): 56-70.
33. Winter C, Reutiman TJ, Folsom TD, Sohr R, Wolf RJ, Juckel G *et al.* Dopamine and serotonin levels following prenatal viral infection in mouse--implications for psychiatric disorders such as schizophrenia and autism. *Eur Neuropsychopharmacol* 2008; **18**(10): 712-716.
34. Makkonen I, Riikonen R, Kokki H, Airaksinen MM, Kuikka JT. Serotonin and dopamine transporter binding in children with autism determined by SPECT. *Dev Med Child Neurol* 2008; **50**(8): 593-597.
35. Chugani DC, Muzik O, Rothermel R, Behen M, Chakraborty P, Mangner T *et al.* Altered serotonin synthesis in the dentatohalamocortical pathway in autistic boys. *Ann Neurol* 1997; **42**(4): 666-669.
36. McDougle CJ, Naylor ST, Cohen DJ, Aghajanian GK, Heninger GR, Price LH. Effects of tryptophan depletion in drug-free adults with autistic disorder. *Arch Gen Psychiatry* 1996; **53**(11): 993-1000.
37. Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E *et al.* Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 1997; **387**(6630): 303-308.
38. Du Y, Stasko M, Costa AC, Davisson MT, Gardiner KJ. Editing of the serotonin 2C receptor pre-mRNA: Effects of the Morris Water Maze. *Gene* 2007; **391**(1-2): 186-197.
39. Moreno-De-Luca D, Sanders SJ, Willsey AJ, Mulle JG, Lowe JK, Geschwind DH *et al.* Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol Psychiatry* 2012.
40. Nakatani J, Tamada K, Hatanaka F, Ise S, Ohta H, Inoue K *et al.* Abnormal behavior in a chromosome-engineered mouse model for human 15q11-13 duplication seen in autism. *Cell* 2009; **137**(7): 1235-1246.

41. Vitali P, Basyuk E, Le Meur E, Bertrand E, Muscatelli F, Cavaille J *et al.* ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *J Cell Biol* 2005; **169**(5): 745-753.
42. Cavaille J, Buiting K, Kiefmann M, Lalonde M, Brannan CI, Horsthemke B *et al.* Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A* 2000; **97**(26): 14311-14316.
43. Vitali P, Royo H, Marty V, Bortolin-Cavaille ML, Cavaille J. Long nuclear-retained non-coding RNAs and allele-specific higher-order chromatin organization at imprinted snoRNA gene arrays. *J Cell Sci* 2010; **123**(Pt 1): 70-83.
44. Doe CM, Relkovic D, Garfield AS, Dalley JW, Theobald DE, Humby T *et al.* Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. *Hum Mol Genet* 2009; **18**(12): 2140-2148.
45. Veltman MW, Craig EE, Bolton PF. Autism spectrum disorders in Prader-Willi and Angelman syndromes: a systematic review. *Psychiatr Genet* 2005; **15**(4): 243-254.
46. Eran A, Li JB, Vatalaro K, McCarthy J, Rahimov F, Collins C *et al.* Comparative RNA editing in autistic and neurotypical cerebella. *Mol Psychiatry* 2012.
47. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; **23**(21): 2947-2948.
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**(4): 357-359.
49. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* 2013; **41**(D1): D64-69.
50. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000; **7**(1-2): 203-214.
51. Korbelt JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009; **10**(2): R23.

52. McElroy KE, Luciani F, Thomas T. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* 2012; **13**: 74.
53. Lorenz R, Bernhart SH, Honer Zu Siederdisen C, Tafer H, Flamm C, Stadler PF *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011; **6**: 26.
54. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 1995; **57**(1): 125–133.
55. Kishore S, Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 2006; **311**(5758): 230-232.
56. Sun M, Hurst LD, Carmichael GG, Chen J. Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Res* 2005; **33**(17): 5533-5543.
57. Faghihi MA, Wahlestedt C. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* 2009; **10**(9): 637-643.
58. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004; **101**(16): 6062-6067.

# Chapter 4:

## Implications and Future Directions

Using deep targeted sequencing of postmortem cerebella from individuals with ASD as compared to neurotypical controls, we have begun characterizing the dynamics, range, variability and regulation of a key epigenetic mechanism linking environmental stimuli and synaptic transmission, A-to-I RNA editing. This work has important implications for understanding gene-environment interactions underlying ASD, ultimately leading to earlier diagnosis and the identification critical targets for therapeutic intervention.

### Implications

#### Extreme synaptic editing in autism

DNA sequence variation in a number of different loci is strongly associated with ASD, but no individual locus is altered in more than 2% of cases<sup>66</sup>. In contrast, 30% of the individuals with ASD examined in Chapter 2 showed extreme levels of synaptic RNA editing, suggesting that A-to-I editing may be both a marker for and mechanism of ASD. Our study suggested that an inverse-U relationship might exist between synaptic A-to-I

editing levels and the ASD phenotype. We speculated that altered A-to-I editing could act as a common compensatory mechanism for the wide range of synaptic abnormalities in ASD, as affected neurons try to maintain synaptic homeostasis.

### **Gender biased 15q11 snoRNA expression and complex relationships with HTR2C editing**

The finding of a strong gender bias underlying *SNORD115* and *SNORD116* expression in human cerebella is important in delineating the role of 15q11 in ASD and must be replicated in an independent cohort. ASD occurs in males nearly five times as much as in females, and 15q11 is one of its most commonly implicated loci. Males with Prader Willi syndrome, in which the *SNORD115* and *SNORD116* clusters are not expressed, have a 40% chance of comorbid ASD, while females with the same genomic abnormality have a 5% chance of comorbid ASD. The detected two-fold upregulation of 15q11 snoRNA may play a role in this 8:1 male-to-female risk ratio, which is nearly double that of the general population. Moreover, despite the sequence complementarity between *SNORD115* genes and HTR2C, our study did not support a strong dependency between the two. Because this has been the largest survey of co-*SNORD115* expression and HTR2C editing to date, its results call for a closer evaluation of the putative relationships.

### **Future Directions**

#### **Compare genome-wide synaptic A-to-I editing in multiple brain regions from large cohorts of individuals with ASD and matched controls.**

Our findings in Chapter 2 were confined to cerebellar candidate genes from 25 individuals. To further understand the role of synaptic A-to-I editing in ASD, larger sets of genes, brain regions, and individuals must be examined. The increasing availability of

well-phenotyped brain tissue allows us to undertake a comprehensive investigation of RNA editing in ASD. We plan to use parallel capture and ultradeep sequencing of cDNA and gDNA to quantify editing in postmortem prefrontal cortex, superior temporal gyrus, and cerebellar tissue samples of individuals with ASD and carefully-matched controls.

### **Examine synaptic A-to-I editing in mouse models of ASD and Mendelian traits comorbid to ASD, and inspect the effect of pharmacological intervention on editing**

To understand the causes, impact, and potential correction of the editing alterations detected in Chapter 2, we wish to identify animal models that might recapitulate the editing changes observed in human subjects. We plan to compare editing levels in the brains of mouse models of Mendelian traits comorbid to ASD and their wild-type counterparts, including tuberous sclerosis complex (TSC), fragile X syndrome (FXS), and Duchenne muscular dystrophy (DMD). Additionally, we will examine how editing changes in response to pharmacological intervention. Drugs previously shown to correct the behavioral phenotype by specifically targeting the perturbed pathways will be used, namely rapamycin in TSC, and mGluR5 antagonists in FXS.

### **Test differentially-edited sites in peripheral blood from large cohorts of cases, unaffected sibs, and unrelated controls**

Although A-to-I editing is a brain-centric phenomenon, it might leave marks on nervous system genes also expressed in peripheral blood. Our group has been studying RNA expression in peripheral blood of individuals with ASD, revealing that blood expression profiles can accurately predict the clinical diagnosis, and highlighting commonly altered molecular pathways in ASD. We plan to test the feasibility of detecting synaptic editing these existing samples, focusing on sites found to be differentially-edited in Chapter 2.



# **Appendix A:**

## **Comment on “Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients”**

**Alal Eran, Kaitlin R. Graham, Kayla Vatalaro, JillianMcCarthy, Christin Collins, Heather Peters, Stephanie J. Brewster, Ellen Hanson, Rachel Hundley, Leonard Rappaport, Ingrid A. Holm, Isaac S. Kohane, and Louis M. Kunkel**

This work appeared in *J Clin Invest*. 2009 Apr;119(4):679-80. doi: 10.1172/JCI38620.

Author contributions: A.E. performed the experiments, analyzed the data, and wrote the manuscript.



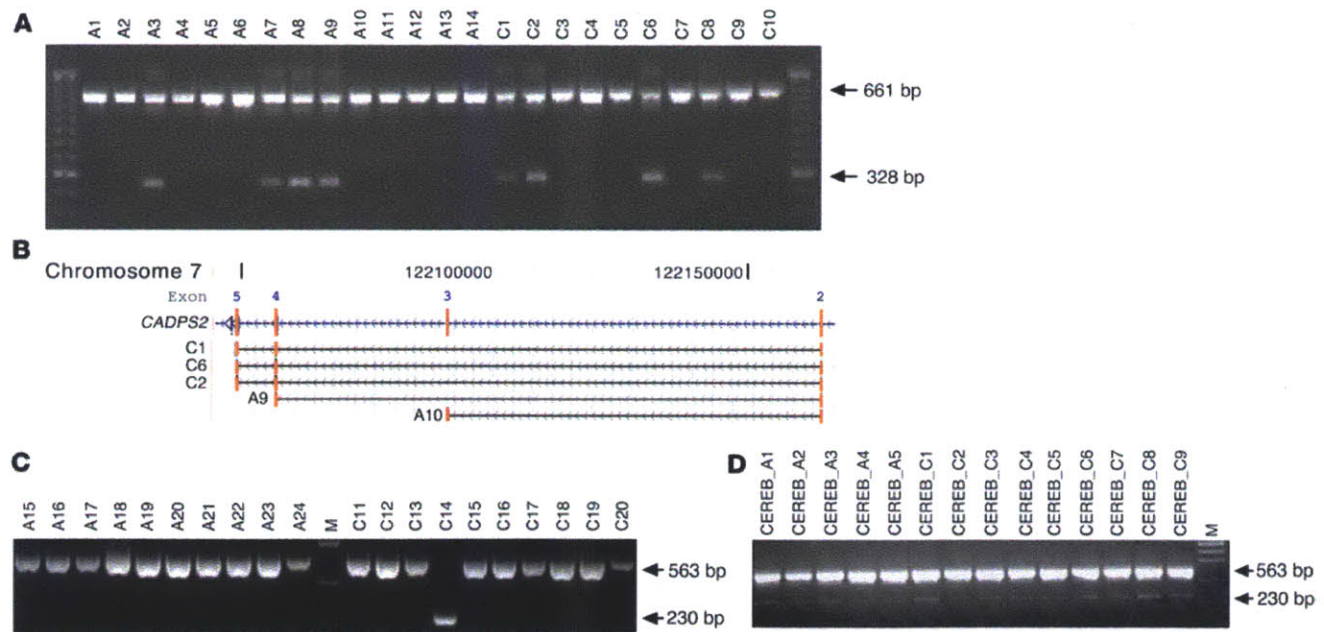
## Comment on “Autistic-like phenotypes in *Cadps2*-knockout mice and aberrant *CADPS2* splicing in autistic patients”

Sadakata et al. (1) reported that a *CADPS2* isoform lacking exon 3 is aberrantly spliced in the peripheral blood of autistic patients. However, we found this splice isoform in the blood of normal subjects at a similar frequency to that of individuals with autism spectrum disorder (ASD) (95% CI of the difference, -0.06 to 0.1). Moreover, this splice variant exists as a minor isoform in cerebellar RNA of both normal individuals and individuals with ASD. Thus, exon 3 skipping likely represents a minor isoform rather than aberrant splicing and is probably not an underlying mechanism of autism. Defects

of *CADPS2* function might contribute to autism susceptibility, but likely not through aberrant splicing.

Sadakata et al. (1) reported that 4 of 16 patients with autism expressed an exon 3-skipped variant of *CADPS2* mRNA in the blood, while the *CADPS2* mRNA of all 24 normal subjects included exon 3. They thus concluded that *CADPS2* is aberrantly spliced in autism, and they performed further experiments showing that the subcellular localization of exogenously expressed exon 3-skipped *CADPS2* is disturbed in primary cultured neocortical and cerebellar neurons.

We aimed to replicate the *CADPS2* findings in an independent set of peripheral blood samples from 41 children with ASD and 39 control children, following the Sadakata et al. protocols (Figure 1A). Furthermore, we performed sequencing (Figure 1B) and nested priming (Figure 1C) to validate the presence or absence of exon 3. Our results showed that, of 39 control samples, 1 was apparently homozygous for the exon 3-skipped allele in peripheral blood, 5 were heterozygous, and 33 were wild type. Of the 41 ASD samples, 5 were heterozygous for the exon 3-skipped isoform, while the rest were wild type.



**Figure 1**

Exon 3 skipping in *CADPS2* mRNA from 41 children with ASD and 39 control children. (A) RT-PCR of *CADPS2* mRNA in blood from subsets of patients with ASD (A1–A14) and control patients (C1–C10) following the Sadakata et al. protocols (1). The 661-bp band represents the full-length exon 1–5 fragment of *CADPS2* mRNA, while the 328-bp band is a result of exon 3 skipping. Four control samples (C1, C2, C6, and C8) and 4 ASD samples (A3, A7, A8, and A9) were heterozygous for the exon 3-skipped isoform. The flanking marker is a 50-bp ladder. The remaining samples showed only the 661-bp band (data not shown). (B) Alignment of sequences obtained from the 328-bp bands of samples C1, C2, C6, and A9 to human chromosome 7 showed that all sequences lacked exon 3. Sequencing the 661-bp band of A10 (which was representative of other samples not showing the 328-bp band) demonstrated that this fragment does include exon 3, as expected. (C) RT-PCR of blood *CADPS2* mRNA using a nested amplification. A single major band (563 bp), indicating the presence of exons 2–5, is shown in all autistic samples. Control sample C14 was apparently homozygous for a 230-bp band that resulted from skipping of exon 3. (D) RT-PCR of cerebellar *CADPS2* mRNA from individuals with ASD and control individuals showed that all cerebella contained the exon 3-skipped splice variant as a minor isoform (230-bp fragment). M, low-DNA-mass ladder (Invitrogen).



Analysis of these results showed no significant difference in the frequency of the exon 3-skipped allele in ASD versus control samples ( $P = 0.6$ , two-proportion  $z$  test). Although the samples tested here might differ from those tested by Sadakata et al. in their ethnicity, gender, or age distributions (Supplemental Figure 1 and Supplemental Tables 1 and 2; supplemental materials available online with this article; doi:10.1172/JCI38620DS1), the finding of exon 3 skipping in healthy controls at a high frequency suggests that this isoform does not represent aberrant splicing and likely is not a mechanism underlying autism.

Since Sadakata et al. extrapolate function of the exon 3-skipped isoform within the cerebellum, we additionally tested the presence of exon 3 in mRNA extracted from the cerebella of 9 control children and 5 children with ASD. All ASD and control samples were found to contain the exon 3-skipped splice variant as a minor isoform (Figure 1D).

Thus, our experiments suggest that exon 3 skipping represents a normal, minor isoform of CADPS2 in the cerebellum. As we observed no difference in prevalence of this allele between ASD and control samples, we conclude that exon 3 skipping is

likely not a mechanism underlying autism susceptibility or pathogenesis.

#### Acknowledgments

We are grateful to all the patients and their family members who participated in the Children's Hospital Boston autism study. We thank Sarah Calvo, Emery Brown, and Steve Boyden for their help. Human tissue was obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, Maryland, USA. This work was supported by the Nancy Lurie Marks Family Foundation and Autism Speaks. L.M. Kunkel is an investigator of the Howard Hughes Medical Institute.

**Alal Eran,<sup>1,2</sup> Kaitlin R. Graham,<sup>1,3</sup> Kayla Vatalaro,<sup>1</sup> Jillian McCarthy,<sup>1</sup> Christin Collins,<sup>1</sup> Heather Peters,<sup>1</sup> Stephanie J. Brewster,<sup>1</sup> Ellen Hanson,<sup>4,5</sup> Rachel Hundley,<sup>4,5</sup> Leonard Rappaport,<sup>4,6</sup> Ingrid A. Holm,<sup>1,6</sup> Isaac S. Kohane,<sup>2,6,7</sup> and Louis M. Kunkel<sup>1,3,6</sup>**

<sup>1</sup>Program in Genomics, Children's Hospital Boston, Boston, Massachusetts, USA.

<sup>2</sup>Harvard-MIT Health Sciences and Technology, Cambridge, Massachusetts, USA.

<sup>3</sup>Howard Hughes Medical Institute, Boston, Massachusetts, USA. <sup>4</sup>Developmental Medicine Center, Children's Hospital Boston, Boston, Massachusetts, USA.

<sup>5</sup>Department of Psychiatry, <sup>6</sup>Department of Pediatrics, and <sup>7</sup>Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA.

Authorship note: Alal Eran, Kaitlin R. Graham, and Kayla Vatalaro contributed equally to this work.

Conflict of interest: The authors have declared that no conflict of interest exists.

Address correspondence to: Louis M. Kunkel, Department of Genetics, Children's Hospital Boston, Boston, Massachusetts 02115, USA. Phone: (617) 355-7576; Fax: (617) 355-7588; E-mail: kunkel@enders.tch.harvard.edu.

*J. Clin. Invest.* 119:679–680 (2009). doi:10.1172/JCI38620.

1. Sadakata, T., et al. 2007. Autistic-like phenotypes in *Cadps2*-knockout mice and aberrant CADPS2 splicing in autistic patients. *J. Clin. Invest.* 117:931–943.

## Response to the letter by Eran et al.

**W**e read with great interest the comment by Eran et al. (1) regarding our recently published CADPS2 article in the *JCI* (2). We appreciate their comment that “exon 3 skipping likely represents a minor isoform rather than aberrant splicing” in the blood and postmortem cerebella of both healthy and autistic individuals. However, we are concerned about the sensitivity of detection of exon 3 skipping in their experiments and have a few replies to their letter.


First, the signal intensity of the exon 3-skipped CADPS2 band was considerably weaker than that of the normal band in some ASD and control blood samples (Figure 1A in ref. 1), in contrast to our results using samples from autism, but not pervasive development disorder — not otherwise specified (PDD-NOS), samples. Moreover, only a trace amount of skipped band was detected in all postmortem cerebella they analyzed (Figure 1D in ref. 1).

Second, they utilized RT-PCR with a nested amplification (at 70 cycles) to detect a control sample C14 with a skipped, but no normal, band (which was called “homozygous” in their comment; Figure 1C in ref. 1) and claimed that only one sample (control sample C14) was homozygous in their study. Little quantitative gain is generally noticed when increasing the number of cycles to such an extraordinary number. We are left wondering if only the skipped band would also be detected in case C14 using an ordinary RT-PCR method (similar to our method with 48 cycles), such as that used to generate the data shown in Figure 1A (1).

Third, their argument regarding one sample “homozygous for the exon 3-skipped allele” and “heterozygous” samples may not be appropriate, since the terms are usually used for genomic DNA, not mRNA. Also, it is yet unknown whether exon 3 skipping is of a *cis*- or *trans*-acting genetic origin or some other origin such as epigenetic.

Fourth, considering the results of Eran et al., we assume that a balance between exon 3-skipped and normal CADPS2 is important for the local secretion property (somato-dendritic, axonal, and synaptic secretion) of CADPS2. Our *JCI* article indicated that exon 3-skipped CADPS2 is not transported into the axons of cultured neurons and suggested that disturbance of this balance may cause a defect in local secretion. Impaired synaptic secretion should be more serious in neurons that dominantly express exon 3-skipped CADPS2 than in those that weakly express it. Thus, excessive exon 3 skipping, together with a combination of other genetic mutations, might contribute to susceptibility to autism.

Finally, we have recently succeeded in generating a mouse line expressing exon 3-skipped *Cadps2* and have confirmed that exon 3 is critical for the subcellular localization of *Cadps2* in neurons (our unpublished observations). Further studies will



shed light on the association of exon 3 skipping with disturbed brain development and behavioral traits.

**Teiichi Furuichi and  
Tetsushi Sadakata**

Laboratory for Molecular Neurogenesis,  
RIKEN Brain Science Institute, Wako, Japan.

Conflict of interest: The authors have declared that no conflict of interest exists.

Address correspondence to: Teiichi Furuichi, Laboratory for Molecular Neurogenesis, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako 351-0198, Japan. Phone: 81-48-467-5906; Fax: 81-48-467-6079; E-mail: tfuruichi@brain.riken.jp.

*J. Clin. Invest.* **119**:680–681 (2009).doi:10.1172/JCI38981.

1. Eran, A., et al. 2009. Comment on “Autistic-like phenotypes in *Cadps2*-knockout mice and aberrant CADPS2 splicing in autistic patients”. *J. Clin. Invest.* **119**:679–680.
2. Sadakata, T., et al. 2007. Autistic-like phenotypes in *Cadps2*-knockout mice and aberrant CADPS2 splicing in autistic patients. *J. Clin. Invest.* **117**:931–943.

# **Appendix B:**

## **Haplotype structure enables prioritization of disease markers and candidate genes in autism spectrum disorder**

**Badri N. Vardarajan, Alal Eran, Jae-Yoon Jung, Louis M. Kunkel, and Dennis P. Wall**

This work is currently under consideration for publication.

Author contributions: A.E. compiled the data and wrote the manuscript.

**Haplotype structure enables prioritization of common markers and candidate genes in autism spectrum disorder**

Badri N. Vardarajan<sup>1,2\*</sup>, Alal Eran<sup>3,4</sup>, Jae-Yoon Jung<sup>1</sup>, Louis M. Kunkel<sup>4</sup>, Dennis P. Wall<sup>1¶\*</sup>

<sup>1</sup>Center for Biomedical Informatics

Harvard Medical School

Boston, MA, 02115

<sup>2</sup>Gertrude H. Sergievsky Center,

Columbia University

New York, NY, 10032

<sup>3</sup>Harvard-MIT Health Sciences and Technology

77 Massachusetts Ave.

Cambridge, MA, 02139

<sup>4</sup>Division of Genetics, Program in Genomics, Boston Children's Hospital

3 Blackfan Circle, CLS 15028

Boston, MA, 02115

\* authors contributed equally

¶correspondence to: [dpwall@hms.harvard.edu](mailto:dpwall@hms.harvard.edu)

## **Abstract and Keywords**

Autism spectrum disorder (ASD) is a neurodevelopmental condition that results in behavioral, social and communication impairments. ASD has a substantial genetic component, with 60-90% trait concordance among monozygotic twins. Efforts to elucidate the causes of ASD have uncovered hundreds of susceptibility loci and candidate genes. However, due to its polygenic nature and clinical heterogeneity, only a few of these markers represent clear targets for further analyses. In the present study, we used the linkage structure associated with published genetic markers of ASD to simultaneously improve candidate gene detection while providing a means of prioritizing markers of common genetic variation in ASD. We first mined the literature for linkage and association studies of SNPs, CNVs and multi-allelic markers in Autism Genetic Resource Exchange (AGRE) families. From markers that reached genome-wide significance we calculated male-specific genetic distances, in light of the observed strong male bias in ASD. Four of 67 autism-implicated regions, 3p26.1, 3p26.3, 3q25-27 and 5p15, were enriched with differentially expressed genes in blood and brain from individuals with ASD. Of 30 genes differentially expressed across multiple expression datasets, 21 were within 10 cMs of an autism-implicated locus. Amongst them, CNTN4, CADPS2, SUMF1, SLC9A9, NTRK3 have been previously implicated in autism, while others have been implicated in neurological disorders comorbid with ASD. This work leverages the rich multimodal genomic information collected on AGRE families to present an efficient integrative strategy for prioritizing autism candidates and improving our understanding of the relationships among the vast collection of past genetic studies.

**Keywords:** autism genetics, autism spectrum disorders, AGRE, bibliome mining.

## **Introduction**

Autism spectrum disorder (ASD) is a neurodevelopmental condition that results in behavioral, social and communication impairments. It is currently estimated that 1 in every 88 children in the United States is affected with ASD, with boys five times more likely to be affected than girls<sup>1</sup>. ASD has a substantial genetic component<sup>2-4</sup>, with 88-95% monozygotic twin concordance, and an estimated heritability of 60-90%<sup>5</sup>. Studies conclude that there are multiple genetic factors that play a role in the etiology of autism. Recent findings have provided evidence in support of roles for de novo mutations<sup>6-9</sup>, common genetic variants<sup>10</sup>, rare variants<sup>11</sup>, and copy-number variation<sup>12-14</sup>. Nevertheless, the genetic basis of the majority of ASD remains largely unclear.

Contributing to the complexity, ASD linkage studies have uncovered over seventy susceptibility loci across the genome and a large number of gene candidates<sup>15, 16</sup>, but most of these findings have not been successfully replicated. The only exceptions to this trend have been linkage peaks on 17q11-17q21<sup>17-20</sup> and 7q<sup>21-25</sup>. Yet, linkage and association studies have dominated the approaches to disentangle the genetic etiology of autism for more than two decades, leaving behind a rich legacy of research findings in the biomedical literature. Reports of significant linkage peaks represent an important clue to the genetic cause of autism that should not be ignored, even in the absence of sufficient replication. However, the mechanistic relevance of the marker should still be determined. For example, a marker may designate collections of genes involved in biological processes or individual genes with mutations of high importance to the susceptibility to autism. Furthermore, these markers and their importance to the etiology of autism once they have achieved the minimum significance threshold of LOD of 3.0 or an



association p-value of  $<0.05$  (corrected for multiple testing), are usually treated as equal. Therefore, despite the fact that markers provide maps, the granularity of those maps is insufficient to direct prioritized experimental follow-up, as every marker, and every gene proximal to that marker, is equally likely to be as important. Given that markers have been identified on nearly every chromosome, the utility of linkage studies for providing specific gene leads and directing further experimental research is limited.

In the present study, we have focused on maximizing the value of previously published linkage and association findings using families from the Autism Genetic Resource Exchange project for directing further genetic analysis of autism. Specifically, our aim was to provide finer resolution to published linkage and association studies through a novel analytical strategy focused on marker-to-gene male-specific genetic distance. Our study was loosely predicated on the assumption that genes in tight linkage with a susceptibility locus are more likely to be linked with the phenotype of interest, i.e. autism, and was leveraged by the collective understanding that the disorder has a substantial male bias. As such, our work focused on reconstructing the male-specific structure of linkage disequilibrium (LD) surrounding significant autism markers, to assemble the biological concepts of genes in tight, medium, and distant LD with those markers. We examined the biological signal inherent to each concept and measured its expression in peripheral blood and postmortem brain tissue from individuals with autism as compared to controls. This strategy improves the resolution of marker-based findings by pointing to the specific genes contributing to the linkage and/or association signals, more likely to play a role in ASD. A large percentage of these genes had not been previously linked to autism but had been implicated in numerous other neurological diseases, including those with overlapping symptoms. Given the ability of this strategy to identify important and novel signal among the

rich collection of research findings from various linkage and association studies in autism, we anticipate that it will have broader applications in the study of other complex genetic disorders in which a large collection of samples had been previously typed and not immediately available for modern sequencing.

## **Materials and Methods**

### ***Autism Marker Selection***

We first mined the autism literature to identify genetic studies focusing on AGRE families. We identified 67 reports of significant linkage and association signals spanning 18 chromosomes (**Table 1**). Significance thresholds were a logarithm of the odds (LOD) score greater than 3 or corrected association p-value <0.01 (depending on the number of markers tested in the study). The search was restricted to studies performed on AGRE families because the same subjects were used to calculate the genetic map around autism markers. This strategy allowed us to capture the true rates of recombination in the studied population and avoid any potential recombination bias. Because the linkage and association studies were based on various experimental designs, we developed the strategy described below to enable their meta analysis.

Each marker was first mapped to the NCBI human genome build 36.3. Then, a 20 MB slice flanking that genomic coordinate was retrieved and the SNPs within that region were used for calculating a genetic map using the same subjects' genotypes<sup>10</sup>. The nearest SNP to the autism marker was used as the reference for calculating recombination rates with other SNPs. The recombination rates were determined with respect to the reference. We assumed that the

recombination rates between the marker and the nearest SNP was negligible enabling us to designate that SNP as a proxy for the marker.

### ***Calculation of LD structure of autism markers***

In order to establish the male-specific linkage disequilibrium (LD) structure between genes and autism markers, we created genetic maps from a 20 MB slice of the chromosome flanking each linkage locus. Specifically, we collected and assembled single nucleotide polymorphisms (SNPs) 10MB upstream and 10MB downstream of each autism marker using the SNP data for AGRE probands<sup>10</sup>. Since autism is almost five times more prevalent in males, we filtered out the females from the dataset before calculating the genetic map. These filtration procedures followed the logic that an AGRE data specific and male-only genetic map would be the most likely to provide an accurate reflection of the samples contributing to the linkage and association signals reported in the pooled studies.

To create the genetic maps for each autism marker, we estimated fine-scale recombination rates using the LDHat software package<sup>26</sup>. This program estimates recombination rates between adjacent SNPs by fitting a Bayesian model based on coalescent theory to analyze patterns of linkage disequilibrium in the data. We conducted this analysis for all 67 markers, identifying the male-specific genetic distances between the marker and genes surrounding that marker, measured in cM. For further filtering, we pruned the genetic map to 15 cMs around the marker. A process flow for the creation of the linkage disequilibrium structure sets is depicted in **Figure 1**.

### ***mRNA expression data processing***

Gene Expression Omnibus (GEO) datasets GSE6575<sup>27</sup> and GSE28521<sup>28</sup> were used to examine the expression of genes surrounding significant autism markers in individuals with ASD. The GSE6575 dataset consists of 17 samples of individuals with ASD without regression, 18 individuals with ASD with regression, 9 patients with mental retardation or developmental delay, and 12 typically developing children from the general population. In this previous study, total RNA was extracted from whole blood samples using the PaxGene Blood RNA System and run on Affymetrix U133plus2.0. For the purposes of our study, we elected to use only the 35 autistic patient samples and 12 control samples from the general population. Preprocessing and expression analyses were done with the Bioinformatics Toolbox Version 2.6 (For Matlab R2007a+). GCRMA was used for background adjustment and control probe intensities were used to estimate non-specific binding<sup>29</sup>. Housekeeping genes, gene expression data with empty gene symbols, genes with very low absolute expression values, and genes with low variance were removed from the preprocessed dataset. When compared to the 12 control samples, the t-test p-value distribution for individuals with ASD with regression was flat and non-informative. We therefore focused only the 17 samples from autistic individuals without regression for differential expression analyses.

The GSE28521 dataset consisted of post-mortem brain tissue samples from 19 autism cases and 17 controls from the Autism Tissue Project, using the IlluminaHumanRef-8 v3.0 expression beadchip panel. Three regions of the brain previously implicated in autism were profiled in each individual: superior temporal gyrus (STG, also known as Brodmann's area (BA) 41/42), prefrontal cortex (BA9) and cerebellar vermis. Raw data were formatted with  $\log_2$  transformation and normalized by quantile normalization. We considered probes with

detection p-value < 0.05 for at least half of the samples for further analysis, as described here<sup>28</sup>.

Raw p-values were generated using limma/bioconductor package in R software, and Benjamini & Hochberg multiple testing correction was applied to obtain adjusted p-values.

### ***Gene expression profiles around common autism markers***

To examine the importance of genes at varying cM distances, and to examine the level of signal relevant to autism surrounding each autism marker individually, we treated each marker region as an independent hypothesis. We then examined the differential regulation of genes within linkage disequilibrium structure (LDS) sets using the mRNA expression profiles described above. Our tests for significant differential expression deviated from standard analyses of microarray data for the primary reason that the recombination history concepts each reflected independent, prior biological knowledge. As such, we treated each concept as a separate collection of hypotheses, with the number of hypotheses being tested simultaneously equivalent to the number of genes in the set. To appropriately account for this multiple testing, we adjusted the nominal p values using the q-value calculation<sup>30</sup>, a measurement framed in terms of the false discovery rate<sup>31</sup>. All 67 recombination history concepts were investigated in this way to determine the frequencies of significant, adjusted p-values ( $q < 0.05$ ) surrounding each autism marker.

### ***Disease cross-referencing***

We mined eight existing gene-disease annotation resources for genes associated with neurological disorders considered to be closely related to autism<sup>32</sup>. Diseases included tuberous sclerosis, epilepsy, seizure disorder and many others with established behavioral similarities to

ASD. The databases examined included the Genetic Association Database<sup>33</sup>, Database of Genomic Variants (<http://projects.tcaq.ca/variation/>), dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), HuGE Navigator Navigator<sup>34</sup>, Human Gene Mutation Database (<http://www.hgmd.cf.ac.uk/ac/index.php>), Online Mendelian Inheritance in Man (<http://www.ncbi.nlm.nih.gov/omim/>), GeneCards (<http://www.genecards.org/>), and SNPedia (<http://snpedia.com/index.php/SNPedia>). Results from these resources were integrated to create a list of genes and associated gene characteristics, which was used for comparisons with the autism LDS genes.

## **Results**

More than 200 genetic studies were conducted on AGRE families between 2001-2012. These were mined to identify 67 genome-wide significant linkage and association signals for ASD (**Table 1**). Common markers for autism span 18 chromosomes, all with a logarithm of the odds (LOD) score greater than 3 or a corrected association p-value<0.01. These studies were based on various experimental designs, mostly using multiplex families with affected sib-pairs. We calibrated the positions of significant markers using NCBI human genome build 36.3<sup>35</sup>, and then aggregated all single nucleotide polymorphisms (SNPs) within a 10 MB window on either side of the marker to calculate the male-specific structure of linkage disequilibrium (LD) around each marker. Examining the recombination rates in the same subjects allows us to build a population specific genetic map, eliminating any genetic bias that might arise from considering ethnicity-matched controls.

Our calculations of recombination rates and LD between SNPs and common autism markers identified a total of 1,426 genes within 25cM of the markers. Of those, 697 protein-coding genes were within 5cMs, 450 between 5 and 10 cMs, and 212 between 10 and 15 cMs from the nearest autism locus (**Figure 2**). Both recombination rates and gene densities varied extensively among autism markers ( $28.1 \pm 7.3$  cMs in the 20 Mb region around markers, spanning  $35.4 \pm 10.4$  genes). There was a strong correlation ( $\rho=0.7$ ) between the size of the genetic map and the proportion of genes at distances  $>10$ cMs. The highest density of genes was around RFWD2 and PAPP2 on chromosome 1, in a CNV-associated region encoding 60 genes within 24 cMs. 48% and 90% of the genes fell within 5 cMs and 10 cMs, respectively, indicating that linkage disequilibrium was well preserved with increasing distance from the autism locus. In contrast, the region around a common CNV near UNQ3037 on chromosome 3 contained 73% genes at a distance greater than  $>10$  cMs.

Previous results indicate that the information content varies by marker and genetic distance, but do not directly demonstrate whether this information is of relevance to our understanding of the genetic etiology of autism. To test directly whether specific markers and/or regions surrounding those markers are more likely to contain promising new gene leads, we examined the regulatory patterns of each set independently in two expression datasets obtained from the Gene Expression Omnibus: a blood-based mRNA expression data from individuals with autism and controls (GSE6575)<sup>27</sup> and a transcriptomic analysis of post-mortem brain RNA (GSE28521). In the blood-based expression dataset although the large majority showed no change in expression, 27 marker regions (40%) contained at least one gene with significant, multiple-test corrected differential expression (**Table 2**). More than 50% of the genes around markers on 3p26 (del CNTN4, del UNQ3037), 3q (D3S3045-D3S1763), 2q (rs17420138), and 5p

(rs10513025) were differentially expressed in whole blood from individuals with ASD. In all, 79 genes were significantly enriched at  $q < 0.05$  across all the marker sets out of which 31 (39%) and 60 (76%) genes lie within 5cMs and 10 cMs of the nearest autism marker respectively, further supporting the notion that the genes proximal to the markers represent more viable autism gene leads than genes further away.

In post-mortem brain tissue data there was an abundance of signal in 64 of the 67 marker regions, which contained at least one gene at  $q\text{-value} < 0.05$ . Regions around 41 markers contained gene sets with significant differential expression, defined as more than 50% of gene differentially expressed in at least one brain region between individuals with ASD and matched controls at a  $q\text{-value}$  threshold of 0.05. Of 383 genes showing evidence of differential expression at  $q < 0.05$ , 205 (53%) and 323 (84%) lie within 5cMs and 10 cMs of the nearest autism marker, respectively.

Four markers were found to reside within a neighborhood of differentially expressed genes in both brain and blood of individuals with ASD. At least 50% of protein-coding genes around rs10513025, D3S3045-D3S1763, del CNTN4, and del UNQ3037 are differentially expressed in both tissues (**Table 2**). Three of these regions (20Mb around del CNTN4, del UNQ3037 and rs10513025) show heavy recombination and contain 73%, 68% and 47% of genes at  $>10\text{cMs}$ . Despite significant recombination within the region, genes significantly enriched for differential expression in both datasets are those closer to the autism marker. Of 30 genes found to be significantly differentially expressed in both blood and brain of individuals with ASD, 11 and 20 lie within 5cMs and 10cMs of the nearest autism marker, respectively.

Integrating a decade of genome-wide linkage and association studies, the male bias of ASD, and differential expression in both brain and blood of individuals with ASD has identified a set of



thirty prime candidates for future targeted experimentation, such as efficient targeted resequencing in very large cohorts<sup>36</sup>. Of these, CADPS2, CNTN4, NTRK3, SLC9A9, and SUMF1 have been previously implicated in ASD. Other differentially expressed genes within 20 male-specific cM of common autism markers have been implicated in disorders with shared symptoms and morbidity patterns, but not ASD per se (**Table 3**).

## **Discussion**

Despite the high heritability of autism, efforts to identify its genetic causes have enjoyed only limited success. Numerous susceptibility loci have been identified, yet few have been replicated, supporting the notion that the genetic complexity of this disorder outmatches the proportion of the autistic population that has been sampled to date. Until the sampling adequately covers the diversity of genetic systems underlying ASD, we must develop analytical approaches to make optimal use of existing results. To this end, we focused here on the development of a simple strategy aimed at targeting previously published autism markers, as well as genes genetically proximal to those markers, most likely to be causally related to autism spectrum disorder. By coupling the structure of linkage disequilibrium with knowledge of biological process and patterns of gene expression data from individuals with ASD, we were able to identify a set of markers and genes proximal to those markers likely to be most informative to the genetic basis of autism. Specific loci on a few chromosomes including three signals on chromosome 3 and one on chromosome 5 yielded the greatest signal, with a sizable percentage of adjacent genes showing highly significant differential expression in autistic blood and brain data. In support of their relevance to the genetics of autism, many of the

differentially expressed genes closely linked to the markers have already been identified as promising autism gene candidates, such as CNTN4, CADPS2, SUMF1, NTRK3 and SLC9A9. In addition, an even greater percentage of these genes have been linked to neurological diseases with high co-morbidity and behavioral similarities to autism spectrum disorder.

Overall, our strategy provides a means for meta-analysis of previous linkage and association studies to prioritize both markers and adjacent genes for further experimental analysis. While our results corroborate the general rule of thumb that genes close to loci identified via linkage and association studies are likely to be informative to the disease under study, they stress that this rule only applies to specific markers. Given the success of application to the autism research field, we expect that our analytical strategy could be of general use in the study of other similarly complex genetic diseases, such as Alzheimer's disease and type 1 diabetes.

### **Acknowledgements**

We thank our lab mates and Professors Isaac Kohane, Marco Ramoni, and Peter Tonellato for engaging discussions related to the project. This work was supported by the National Institute of Health grants 1R01MH085143-01 and 1R01MH090611-01A1.

### **Conflict of interest**

The authors declare no conflict of interest.

### **References**

1. Autism, Developmental Disabilities Monitoring Network Surveillance Year Principal I, Centers for Disease C, Prevention. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR Surveill Summ* 2012; **61**(3): 1-19.
2. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* 1995; **25**(1): 63-77.
3. Lauritsen M, Ewald H. The genetics of autism. *Acta Psychiatr Scand* 2001; **103**(6): 411-427.
4. Rutter M. Genetic studies of autism: from the 1970s into the millennium. *J Abnorm Child Psychol* 2000; **28**(1): 3-14.
5. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: A decade of new twin studies. *Am J Med Genet B Neuropsychiatr Genet* 2011.
6. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012; **485**(7397): 237-241.

7. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012; **485**(7397): 242-245.
8. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; **485**(7397): 246-250.
9. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012; **74**(2): 285-299.
10. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009; **459**(7246): 528-533.
11. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K *et al.* A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci U S A* 2007; **104**(31): 12831-12836.

12. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T *et al.* Strong association of de novo copy number mutations with autism. *Science* 2007; **316**(5823): 445-449.
13. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, Hill RS *et al.* Identifying autism loci and genes by tracing recent shared ancestry. *Science* 2008; **321**(5886): 218-223.
14. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 2007; **39**(3): 319-328.
15. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 2008; **9**(5): 341-355.
16. Freitag CM. The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol Psychiatry* 2007; **12**(1): 2-22.
17. Alarcon M, Yonan AL, Gilliam TC, Cantor RM, Geschwind DH. Quantitative genome scan and Ordered-Subsets Analysis of autism endophenotypes support language QTLs. *Mol Psychiatry* 2005; **10**(8): 747-757.

18. Cantor RM, Kono N, Duvall JA, Alvarez-Retuerto A, Stone JL, Alarcon M *et al.* Replication of autism linkage: fine-mapping peak at 17q21. *Am J Hum Genet* 2005; **76**(6): 1050-1056.
19. McCauley JL, Li C, Jiang L, Olson LM, Crockett G, Gainer K *et al.* Genome-wide and Ordered-Subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates. *BMC Med Genet* 2005; **6**: 1.
20. Yonan AL, Alarcon M, Cheng R, Magnusson PK, Spence SJ, Palmer AA *et al.* A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet* 2003; **73**(4): 886-897.
21. A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Hum Mol Genet* 1998; **7**(3): 571-578.
22. A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *Am J Hum Genet* 2001; **69**(3): 570-581.
23. Further characterization of the autism susceptibility locus AUTS1 on chromosome 7q. *Hum Mol Genet* 2001; **10**(9): 973-982.

24. Alarcon M, Cantor RM, Liu J, Gilliam TC, Geschwind DH. Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *Am J Hum Genet* 2002; **70**(1): 60-71.
25. Barrett S, Beck JC, Bernier R, Bisson E, Braun TA, Casavant TL *et al.* An autosomal genomic screen for autism. Collaborative linkage study of autism. *Am J Med Genet* 1999; **88**(6): 609-615.
26. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304**(5670): 581-584.
27. Gregg JP, Lit L, Baron CA, Hertz-Picciotto I, Walker W, Davis RA *et al.* Gene expression changes in children with autism. *Genomics* 2008; **91**(1): 22-29.
28. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*; **474**(7351): 380-384.

29. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *J Amer Stat Assoc* 2004; **99**(468): 909-917.
30. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003; **100**(16): 9440-9445.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995; **57**: 289--300.
32. Wall DP, Esteban FJ, Deluca TF, Huyck M, Monaghan T, Velez de Mendizabal N *et al.* Comparative analysis of neurological disorders focuses genome-wide search for autism genes. *Genomics* 2009; **93**(2): 120-129.
33. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nat Genet* 2004; **36**(5): 431-432.
34. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet* 2008; **40**(2): 124-125.



35. NCBI.  
<http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=3>.
36. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG *et al.* Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science* 2012.
37. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009; **459**(7246): 569-573.
38. Ma D, Salyakina D, Jaworski JM, Konidari I, Whitehead PL, Andersen AN *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet* 2009; **73**(Pt 3): 263-273.
39. Buxbaum JD, Silverman J, Keddache M, Smith CJ, Hollander E, Ramoz N *et al.* Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. *Mol Psychiatry* 2004; **9**(2): 144-150.

40. Lu AT, Cantor RM. Allowing for sex differences increases power in a GWAS of multiplex Autism families. *Mol Psychiatry* 2012; **17**(2): 215-222.
41. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI *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.
42. Ramoz N, Cai G, Reichert JG, Silverman JM, Buxbaum JD. An analysis of candidate autism loci on chromosome 2q24-q33: evidence for association to the STK39 gene. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B**(7): 1152-1158.
43. Buxbaum JD, Silverman JM, Smith CJ, Kilifarski M, Reichert J, Hollander E *et al.* Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *Am J Hum Genet* 2001; **68**(6): 1514-1520.
44. Liu X, Novosedlik N, Wang A, Hudson ML, Cohen IL, Chudley AE *et al.* The DLX1 and DLX2 genes and susceptibility to autism spectrum disorders. *Eur J Hum Genet* 2009; **17**(2): 228-235.

45. Hussman JP, Chung RH, Griswold AJ, Jaworski JM, Salyakina D, Ma D *et al.* A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. *Mol Autism* 2011; **2**(1): 1.
46. Collins AL, Ma D, Whitehead PL, Martin ER, Wright HH, Abramson RK *et al.* Investigation of autism and GABA receptor subunit genes in multiple ethnic groups. *Neurogenetics* 2006; **7**(3): 167-174.
47. Fradin D, Cheslack-Postava K, Ladd-Acosta C, Newschaffer C, Chakravarti A, Arking DE *et al.* Parent-of-origin effects in autism identified through genome-wide linkage analysis of 16,000 SNPs. *PLoS One* 2010; **5**(9).
48. 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.
49. Philippi A, Tores F, Carayol J, Rousseau F, Letexier M, Roschmann E *et al.* Association of autism with polymorphisms in the paired-like homeodomain transcription factor 1 (PITX1) on chromosome 5q31: a candidate gene analysis. *BMC Med Genet* 2007; **8**: 74.

50. Bureau A, Croteau J, Tayeb A, Merette C, Labbe A. Latent class model with familial dependence to address heterogeneity in complex diseases: adapting the approach to family-based association studies. *Genet Epidemiol* 2011; **35**(3): 182-189.
51. Serajee FJ, Zhong H, Mahbubul Huq AH. Association of Reelin gene polymorphisms with autism. *Genomics* 2006; **87**(1): 75-83.
52. Campbell DB, Sutcliffe JS, Ebert PJ, Militerni R, Bravaccio C, Trillo S *et al.* A genetic variant that disrupts MET transcription is associated with autism. *Proc Natl Acad Sci U S A* 2006; **103**(45): 16834-16839.
53. Thanseem I, Nakamura K, Miyachi T, Toyota T, Yamada S, Tsujii M *et al.* Further evidence for the role of MET in autism susceptibility. *Neurosci Res* 2010; **68**(2): 137-141.
54. Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M *et al.* A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am J Hum Genet* 2008; **82**(1): 160-164.
55. Alarcon M, Abrahams BS, Stone JL, Duvall JA, Perederiy JV, Bomar JM *et al.* Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet* 2008; **82**(1): 150-159.

56. Molloy CA, Keddache M, Martin LJ. Evidence for linkage on 21q and 7q in a subset of autism characterized by developmental regression. *Mol Psychiatry* 2005; **10**(8): 741-746.
57. Benayed R, Gharani N, Rossman I, Mancuso V, Lazar G, Kamdar S *et al.* Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus. *Am J Hum Genet* 2005; **77**(5): 851-868.
58. Autism Genome Project C, Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 2007; **39**(3): 319-328.
59. Anitha A, Nakamura K, Yamada K, Suda S, Thanseem I, Tsujii M *et al.* Genetic analyses of roundabout (ROBO) axon guidance receptors in autism. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B**(7): 1019-1027.
60. Ma DQ, Cuccaro ML, Jaworski JM, Haynes CS, Stephan DA, Parod J *et al.* Dissecting the locus heterogeneity of autism: significant linkage to chromosome 12q14. *Mol Psychiatry* 2007; **12**(4): 376-384.

61. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008; **358**(7): 667-675.
62. Buxbaum JD, Silverman JM, Smith CJ, Greenberg DA, Kilifarski M, Reichert J *et al.* Association between a GABRB3 polymorphism and autism. *Mol Psychiatry* 2002; **7**(3): 311-316.
63. Delahanty RJ, Kang JQ, Brune CW, Kistner EO, Courchesne E, Cox NJ *et al.* Maternal transmission of a rare GABRB3 signal peptide variant is associated with autism. *Mol Psychiatry* 2011; **16**(1): 86-96.
64. Miller DT, Shen Y, Weiss LA, Korn J, Anselm I, Bridgemohan C *et al.* Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J Med Genet* 2009; **46**(4): 242-248.
65. Philippi A, Roschmann E, Tores F, Lindenbaum P, Benajou A, Germain-Leclerc L *et al.* Haplotypes in the gene encoding protein kinase c-beta (PRKCB1) on chromosome 16 are associated with autism. *Mol Psychiatry* 2005; **10**(10): 950-960.

66. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 2008; **17**(4): 628-638.
67. Sutcliffe JS, Delahanty RJ, Prasad HC, McCauley JL, Han Q, Jiang L *et al.* Allelic heterogeneity at the serotonin transporter locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors. *Am J Hum Genet* 2005; **77**(2): 265-279.
68. Stone JL, Merriman B, Cantor RM, Yonan AL, Gilliam TC, Geschwind DH *et al.* Evidence for sex-specific risk alleles in autism spectrum disorder. *Am J Hum Genet* 2004; **75**(6): 1117-1123.
69. Strom SP, Stone JL, Ten Bosch JR, Merriman B, Cantor RM, Geschwind DH *et al.* High-density SNP association study of the 17q21 chromosomal region linked to autism identifies CACNA1G as a novel candidate gene. *Mol Psychiatry* 2010; **15**(10): 996-1005.
70. Campbell DB, Li C, Sutcliffe JS, Persico AM, Levitt P. Genetic evidence implicating multiple genes in the MET receptor tyrosine kinase pathway in autism spectrum disorder. *Autism Res* 2008; **1**(3): 159-168.
71. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* 2010; **19**(20): 4072-4082.

**Table 1: Autism markers identified in AGRE families between 2001-2012.** Linkage and association studies performed in AGRE families were compiled and genome-wide significant markers identified. The logarithm-of-the-odds (LOD) scores and/or association p-values are listed for each marker. Human genome Build 36.3 was used to calibrate marker position. Male specific genetic distances were calculated using dense SNP genotypes from the same individuals.

Chromosome	Marker	Median Marker Position (bp)	Male-Specific Genetic Map Units (cMs)	p-value / LOD (association/linkage)	Ref
1	dup RFWD2-PAPPA2	174522115	23.3	p=1.0e-02	37
1	rs12740310–rs3737296–rs12410279	218873645	26.7	p=5.0e-04	38
1	D1S1656	228971975	40.9	NPL = 3.21	39
1	rs6683048	235855409	37.6	p=2.3e-09	40
2	dup AK123120	13142782	41.5	p=3.57e-06	37
2	del NRXN1	50557085	20.9	p=3.30e-04	41
2	del NRXN1	51134122	21.7	p=4.7e-04	37
2	rs17420138	158585159	19.6	p=5.63e-08	40
2	rs1807984	168787136	22.9	p=7.0e-03	42
2	D2S335	172274852	22.8	HLOD=2.99 NPL=3.32	43
2	rs4519482	172671605	24.2	p=7.0e-05	44
2	204,444,539-204,446,116 LD block	204445327	26.6	p=1.8e-06	45
3	del CNTN4	1915556	27.0	p=4.7e-04	37
3	del UNQ3037	4218017	30.4	p=2.0e-03	37
3	D3S3045-D3S1763	138597603	23.2	Z=3.10 p<1.0e-03	17
3	dup NLGN1	174763176	21.5	p=1.0e-02	37
4	rs17599165	46634972	18.1	p=1.5e-03	46
4	rs1912960	46648638	18.8	p=7.3e-03	46
4	rs17599416	46668195	19.1	p=4.0e-03	46
4	rs6826933–rs17088473	61133187	17.8	HLOD = 3.79 LOD = 2.96	47
4	dup GUSBP5	144850990	22.8	p=1e-02	37



5	rs10513025	9676622	38.4	p=1.7e-06	48
5	rs1896731- rs10038113	25936438	29.3	p=3.4e-06	38
5	rs4307059	26003460	31.6	p=3.0E4e-08	10
5	rs11959298- rs6596189	134395753	23.2	p=4.0e-04	49
6	rs13193457	15453984	35.0	P=3.0e-05	50
6	del PARK2	162585788	31.6	p=4.7e-03	37
7	rs736707	102917639	21.8	p=1.40e-5	51
7	rs1858830	116099675	17.0	p=5.0e-06	52
7	rs38841	116107162	16.0	p=6.0e-04	53
7	rs7794745	146120539	36.7	LOD=3.4 p<2.14e-05	54
7	rs2710102	147205323	35.0	p=2.0e-03	55
7	D7S483	151829212	33.4	NPL=3.7,p=7.9e-05	56
7	rs1861972- rs1861973	154946830	28.1	p=3.5e-06	57
9	rs1340513	6967633	35.7	Zlr = 3.21 p = 7.0e-04	58
9	rs722628	7136888	35.7	Zlr = 3.59 p = 6.0e-03	58
9	rs536861	127353265	31.6	Zlr = 3.30 p=5.0e-04	58
10	del GRID1	87945347	30.8	p=3.1e-04	37
11	rs2421826	35187181	24.5	Zlr = 3.57	58
11	rs1358054	36163248	25.6	Zlr = 3.77 p = 8.0e-03	58
11	rs6590109	124264258	37.7	p=9.0e-03	59
12	rs1445442	63577561	25.0	HLOD=4.51	60
14	del MDGA2	46796374	22.1	p=1.3e-04	41
15	del OR4M2,OR4N	19844860	26.1	p=9.48e-12	41
15	del LOC650137	19915407	24.8	p=9.48e-12	41
15	dup UBE2A	23184355	38.4	p=9.27e-06	41
15	dup 15q11-13	23704547	37.5	p=1.0e-05	37
15	maternal dup 15q11-13	23750000	36.1	p approaching 0	61
15	GABRB3 155CA-2	24559869	38.2	MTDT p=2.0e-03	62
15	rs25409	24569934	39.2	p=8.0e-03	63
15	dup 15q13 BP4-BP5	29508500	42.5	p approaching 0	64
15	rs11855650- rs10520676	77364734	23.0	HLOD = 3.09 LOD = 3.62	47
16	FE0DBACA18ZG03v	19408579	32.6	p=1.6e-04	65

16	FE0DBACA7ZD06v	24133057	26.8	p=1.4e-05	65
16	del/dup 16p11.2	30300000	45.5	p=1.1e-04	61, 66
17	D17S1294- D17S1800	26183756	22.2	HLODREC=5.8 p=1.59e-07	67
17	D17S1294	26860299	20.9	MLS=3.2 Male Only MLS=4.3	68
17	D17S1299	36247989	25.7	MLS=3.6	18
17	D17S2180	44028199	24.7	MLS=4.1	18
17	rs757415 and rs12603112	46020488	22.8	p=1.9e-05	69
17	del BZRAP1	53747037	29.0	p=8.0e-04	41
19	del MADCAM1	451915	17.8	p=6.0e-04	41
19	rs344781	48866628	23.9	p=6.0e-03	70
20	rs723477	237362	22.8	NPL LOD=3.81	48
20	rs16999397- rs200888	958294	24.5	HLOD = 3.36 LOD = 3.38	47
20	rs4141463	14695471	34.8	p=3.7e-08	71
21	D21S1437	20568713	28.4	NPL=3.4 p=3.5e-04	56

**Table 2: Differential expression of genes around common autism markers.**

For each marker region, the table lists the percentage of genes found to be differentially expressed in blood and brain of individuals with ASD at a significance level of  $q < 0.05$ .

Marker	Blood (GSE6575)		Brain (GSE28521)	
	Number of surrounding genes with expression data	% genes significantly differentially expressed in individuals with ASD	Number of surrounding genes with expression data	% genes significantly differentially expressed in individuals with ASD
<i>del CNTN4</i>	16	100.0%	12	100.0%
<i>rs10513025</i>	12	75.0%	12	100.0%
<i>D3S3045-D3S1763</i>	22	50.0%	26	100.0%
<i>[CNV: UNQ3037]</i>	19	89.5%	21	66.7%
<i>rs11855650-rs10520676</i>	21	28.6%	18	100.0%
<i>del NRXN1</i> (Glessner 2009)	35	5.7%	27	100.0%
<i>del NRXN1</i> (Bucan 2009)	32	6.3%	24	100.0%
<i>rs6683048</i>	43	4.7%	30	100.0%
<i>dup AK123120</i>	13	0.0%	17	100.0%
<i>rs4307059</i>	16	0.0%	7	100.0%
<i>rs1896731-rs10038113</i>	15	0.0%	7	100.0%
<i>rs7794745</i>	37	0.0%	23	100.0%
<i>rs2710102</i>	38	0.0%	25	100.0%
<i>rs736707</i>	30	0.0%	23	100.0%
<i>rs344781</i>	28	0.0%	16	100.0%
<i>D7S483</i>	29	0.0%	21	95.2%
<i>rs1861972-rs1861973</i>	27	0.0%	19	94.7%
<i>del GRID1</i>	28	14.3%	17	94.1%
<i>del BZRAP1</i>	19	0.0%	14	92.9%
<i>rs17599165</i>	24	0.0%	13	84.6%
<i>rs1912960</i>	24	0.0%	13	84.6%
<i>rs17599416</i>	24	0.0%	13	84.6%
<i>rs757415-rs12603112</i>	25	4.0%	22	81.8%
<i>dup NLGN1</i>	22	0.0%	16	81.3%
<i>D17S2180</i>	27	0.0%	21	76.2%
<i>Chr2:204444539-204446116 LD block</i>	25	36.0%	10	70.0%

rs1807984	25	0.0%	20	70.0%
D1S1656	42	0.0%	36	66.7%
del MDGA2	17	0.0%	11	63.6%
FE0DBACA18ZG03v	30	0.0%	22	63.6%
rs12740310- rs3737296- rs12410279	28	0.0%	21	61.9%
D2S335	28	0.0%	18	61.1%
rs4519482	28	0.0%	18	61.1%
FE0DBACA7ZD06v	24	0.0%	18	61.1%
rs38841	22	9.1%	23	60.9%
rs1858830	22	9.1%	23	60.9%
dup GUSBP5	17	0.0%	12	58.3%
rs723477	9	11.1%	16	56.3%
del PARK2	35	11.4%	20	50.0%
dup RFWD2- PAPPA2	32	0.0%	24	50.0%
rs6826933- rs17088473	12	0.0%	10	50.0%
rs16999397-rs200888	13	0.0%	18	50.0%
dup UBE2A	14	7.1%	13	46.2%
maternal dup 15q11- 13	15	6.7%	13	46.2%
GABRB3 155CA-2	15	6.7%	13	46.2%
rs25409	15	6.7%	13	46.2%
dup 15q11-13	15	6.7%	13	46.2%
rs4141463	10	20.0%	18	44.4%
D21S1437	10	0.0%	7	42.9%
del/dup 16p11.2	15	0.0%	12	41.7%
dup 15q13 BP4-BP5	23	8.7%	19	31.6%
rs11959298 - rs6596189	23	0.0%	16	31.3%
rs1358054	29	31.0%	26	30.8%
rs1340513	24	0.0%	13	30.8%
rs722628	24	0.0%	13	30.8%
D17S1299	20	0.0%	20	30.0%
rs536861	33	0.0%	25	28.0%
del OR4M2-OR4N	10	10.0%	11	27.3%
del LOC650137	10	10.0%	11	27.3%
rs2421826	26	46.2%	23	26.1%
D17S1294-D17S1800	21	0.0%	17	17.6%
D17S1294	21	0.0%	17	17.6%
rs6590109	24	0.0%	18	11.1%
rs13193457	27	0.0%	21	9.5%

rs17420138	17	100.0%	17	0.0%
rs1445442	24	0.0%	0	0.0%
del MADCAM1	8	0.0%	5	0.0%

**Table 3: Top candidate genes based on integrating a decade of genome-wide linkage and association studies, the autism male bias, and differential expression in brain and blood of individuals with ASD.** Listed are genes located within 20 male-specific cM of genome-wide significant autism markers, which are also differentially expressed in both brain and blood of individuals with ASD. Of these, 19 genes (63%) were previously implicated in neurological disorders with high degrees of overlap in symptomatology and morbidity to ASD.

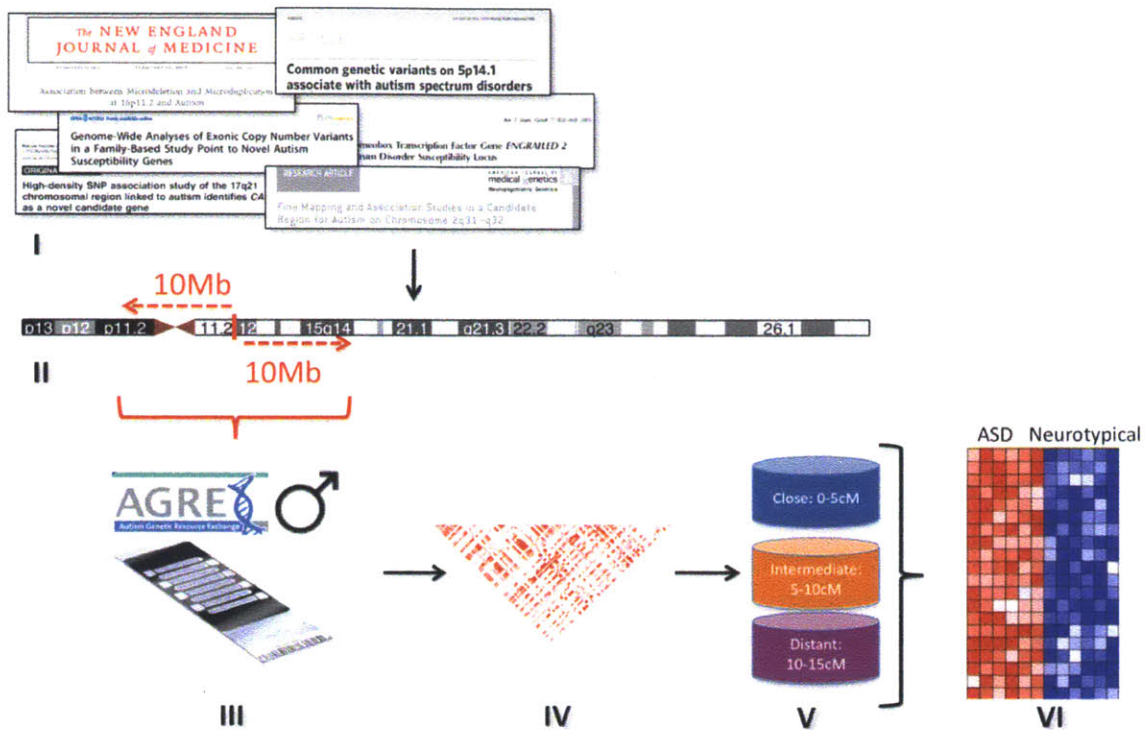
Gene	Differential expression in blood		Differential expression in brain		Male specific genetic distance from marker (cM)	Association with disorders co-morbid to ASD*
	t-test p	FDR	t-test p	FDR		
TRIM44	9.5e-02	4.8e-02	5.1e-03	1.4e-02	0.84	
ITPR1	7.7e-01	9.3e-04	1.9e-03	1.5e-03	1.09	4, 7, 15, 16,
IREB2	2.3e-02	4.0e-02	1.2e-02	2.8e-03	1.39	4, 10
CNTN4	1.4e-01	2.4e-02	3.2e-01	5.1e-03	1.88	4, 6, 8, 16, 18, 19
NMNAT3	2.0e-01	4.5e-02	4.6e-01	8.3e-03	2.39	10
RAB6B	2.0e-01	4.5e-02	1.0e-03	1.9e-04	3.02	
CADPS2	4.1e-04	3.7e-03	1.0e-05	4.9e-05	3.34	5, 6
SPTBN1	1.5e-03	9.1e-03	1.6e-01	3.1e-06	3.56	1,16
TMEM108	2.1e-01	4.5e-02	1.0e-01	2.8e-03	4.09	
ACPL2	8.3e-02	2.8e-02	8.2e-02	2.4e-03	4.43	
ADCY2	1.4e-01	3.2e-02	3.9e-02	3.9e-03	4.77	16
NSUN2	1.1e-02	1.1e-02	7.7e-01	3.8e-02	6.62	16
PANK1	1.4e-02	3.8e-02	7.7e-03	2.8e-03	7.14	
SUMF1	1.4e-01	2.4e-02	4.9e-01	6.7e-03	7.31	9, 10, 13, 21, 22
TANC1	1.2e-01	1.7e-03	6.0e-02	3.8e-02	7.31	4, 6, 17, 19, 20
SLC23A2	1.6e-02	2.7e-02	2.4e-02	1.8e-02	8.35	
EPB41L5	2.5e-03	1.3e-02	3.5e-02	1.8e-02	8.65	
ALKBH3	1.6e-01	3.7e-02	3.0e-05	2.4e-04	9.00	
SLC9A9	7.5e-02	2.8e-02	4.3e-04	1.8e-04	9.04	5, 6, 8, 15, 18, 20
NTRK3	3.0e-02	4.0e-02	7.9e-02	9.1e-03	9.67	2, 3, 5, 6, 7, 11, 15, 16, 17, 23
PLSCR4	5.4e-02	2.8e-02	8.4e-03	5.0e-04	12.30	7, 16
MYO10	1.4e-01	3.2e-02	1.4e-01	9.2e-03	12.64	10
KCNMA1	1.3e-02	3.8e-02	2.4e-01	2.2e-02	13.86	2, 8, 10, 15, 16, 18

SMYD3	4.8e-04	9.5e-03	2.6e-02	1.8e-03	14.32	
ATP2B2	5.3e-02	2.4e-02	1.1e-03	9.4e-04	14.77	14, 16, 20
ALDH18A1	9.1e-03	3.8e-02	6.0e-01	4.1e-02	15.93	8, 10, 12, 18, 19, 20
LMCD1	1.3e-01	2.4e-02	5.4e-01	6.7e-03	16.72	
ATG7	2.7e-01	4.5e-04	3.4e-04	7.1e-04	16.78	10
SYN2	2.6e-01	2.6e-02	4.6e-03	2.9e-03	18.41	7, 8, 15, 16
MKRN2	2.3e-01	2.6e-02	2.0e-02	9.5e-03	18.70	

\* List of disorders

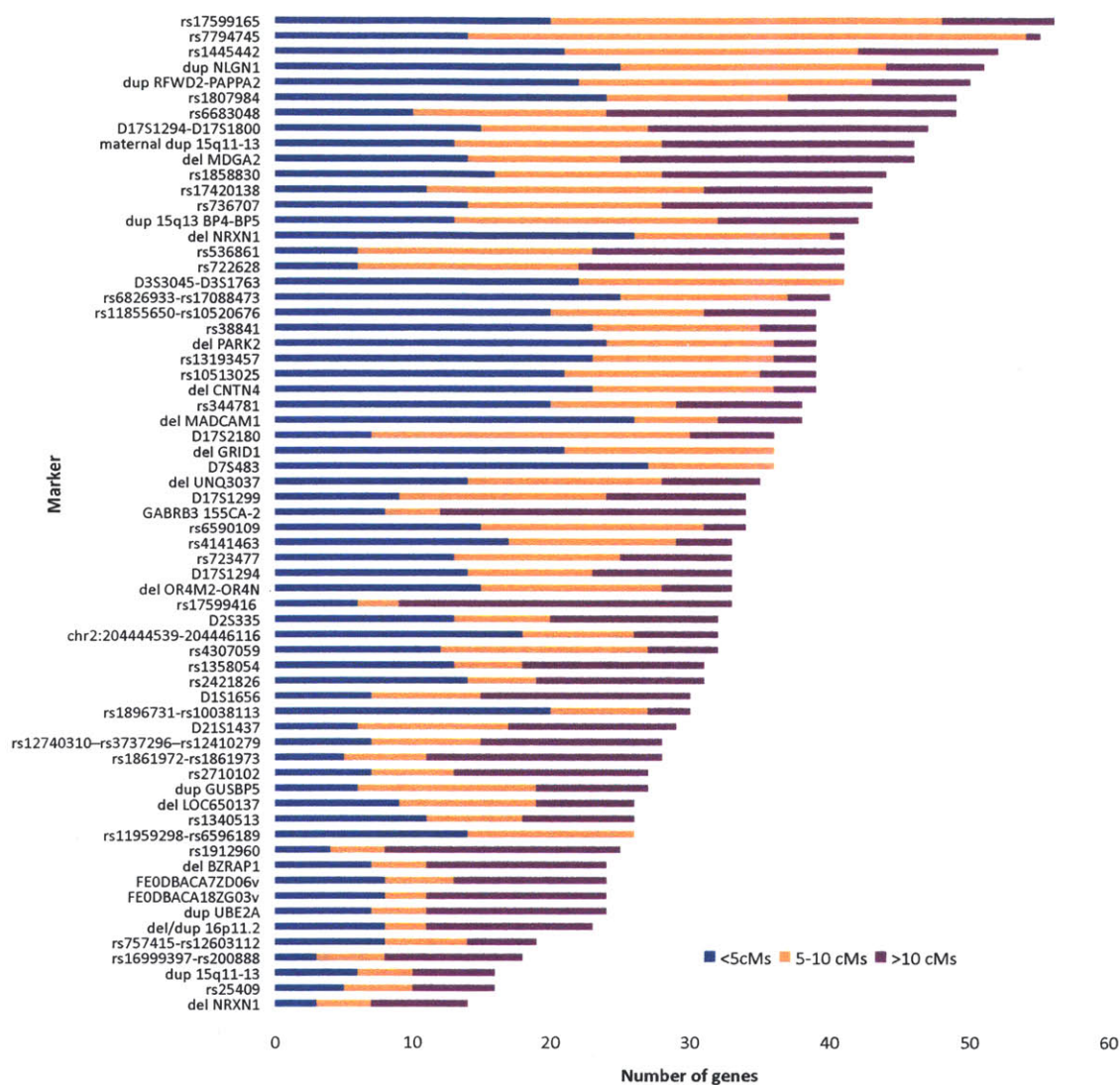
1 Neurofibromatosis, 2 Tuberos sclerosis, 3 Anxiety disorders, 4 Ataxia, 5 Attention deficit disorder, 6 Autistic disorder, 7 Bipolar disorder, 8 Seizures, 9 Cerebral palsy, 10 Dementia, 11 Depressive disorder, 12 Down syndrome, 13 Dystonia, 14 Encephalomyelitis, 15 Epilepsy, 16 Schizophrenia, 17 Hydrocephalus, 18 Mental Retardation, 19 Microcephaly, 20 Multiple sclerosis, 21 Neuroacanthocytosis, 22 Neuroaxonal dystrophies, 23 Obsessive-compulsive disorder

## Figures



**Figure 1** Integrative genomics workflow for prioritizing candidate genes for further experimentation. (I) The rich collection of genetic studies performed on AGRE families between 2001-2012 was mined to identify genome-wide significant linkage and association signals. (II) Markers were remapped to the current genome build and flanking regions extracted. (III) SNP genotypes of AGRE male probands were compiled to enable male-specific genetic distance calculations in the same subjects. (IV) Regional recombination rates between markers and SNPs were calculated, and (V) protein coding genes within 20 male-specific cM from the markers identified. (VI) The expression profiles of these genes were examined in brain and blood of individuals with ASD relative to neurotypical individuals. Genes found to be differentially expressed in both tissues and located within the male-specific vicinity of a significant autism marker are considered prime candidates for further studies. Of 30 genes that satisfy these criteria, 19 were previously implicated in disorders that share symptoms and morbidity patterns with ASD.





**Figure 2** Number of genes within 20cM of significant autism markers. Genetic distances were calculated using male-only AGRE proband SNPs<sup>10</sup>. Genes were grouped into three distance bins indicating the extent of recombination with the autism marker. The figure displays the number of genes in tight linkage with the marker and therefore the extent of recombination around each locus.

## **Appendix C:**

# **Whole genome sequencing of six unrelated patients with autism reveals novel candidate genes and pathways affected by rare and nonsynonymous variants**

**Sek-Won Kong, Michael Hsing, Alal Eran, Malcolm G. Campbell, Louis M. Kunkel, and Isaac S. Kohane**

Author contribution: A.E. performed the experiments

**Title:** Whole genome sequencing of six unrelated patients with autism reveals novel candidate genes and pathways affected by rare and nonsynonymous variants

SW Kong,<sup>1,2</sup> M Hsing,<sup>1</sup> A Eran,<sup>2</sup> AR Papa,<sup>3</sup> MG Campbell,<sup>2</sup> LM Kunkel,<sup>4</sup> IS Kohane<sup>1,2,\*</sup>

<sup>1</sup>Informatics Program, Children's Hospital Boston, 300 Longwood Avenue, Boston, MA 02115, USA, <sup>2</sup>Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA, <sup>3</sup>Boston University, 24 Cummington Street, Boston, MA 02215, USA

<sup>4</sup>Department of Genetics, Children's Hospital Boston, 300 Longwood Avenue, Boston, MA 02115, USA

\* Correspondence: [Isaac\\_Kohane@harvard.edu](mailto:Isaac_Kohane@harvard.edu) (I.S.K.)

## **Abstract**

Autism Spectrum Disorders (ASD) is a common developmental brain disorder with high heritability; however, the cause of ASD in a majority of cases remains unknown. To identify genomic variants implicated in ASD, whole genomes from six unrelated patients were sequenced using a sequencing-by-ligation method, and genome-wide copy number variations (CNV) in probands and their parents were profiled using genotyping microarrays. Rare or novel deleterious variants at conserved loci were found more frequently in the ASD genomes than in 9 ethnicity-matched HapMap genomes after controlling for the total number of variants (analysis of covariance p-value 0.0030). These variants were present in 412 genes in each ASD genome on average; and a total of 1,245 genes were found containing rare or novel deleterious variants at conserved loci in at least one of the 6 ASD genomes, but none of the 9 HapMap genomes. Among these, 32 genes including *AVPR1A*, *CADPS2*, *DISC1*, *ITGB3*, and *MET* were known ASD candidate genes from the literature. Notably, the ABC transporters pathway was enriched with rare or novel deleterious variants in 3 of 6 ASD genomes, suggesting the presence of common biological pathways that may be implicated in a subgroup of patients. Several small copy number variations that were not identified using genotyping microarrays were found using the read-depth analysis of WGS data. Together these findings highlight the potential of whole genome sequencing to uncover the complex genetic architecture of variants, genes and pathways associated with ASD.

**Keywords:** autism spectrum disorders; whole genome sequencing; synaptic genes; immune response

## Introduction

Autism spectrum disorders (ASD) constitute a broad category of highly heritable brain disorders characterized by developmental delays in communication, social interaction, and abnormal behavior. The prevalence of ASD has been continuously increasing to the present estimates of 1 in 110.<sup>1</sup> Despite this high heritability, the cause of ASD in the majority of patients remains unknown. Over 200 candidate genes have been previously reported from linkage analysis and genome-wide association studies (GWAS); however, each of these genes alone is associated with only a small proportion of cases possibly due to the genetic heterogeneity of the disease.<sup>2</sup>

Recent developments in next-generation sequencing technology have brought about a new era in personal genome sequencing.<sup>3,4</sup> A few proof-of-concept studies using whole-genome sequencing (WGS) or whole-exome sequencing (WES) have already proven the technology useful in identifying rare disease-causing mutations. These approaches have been successful in identifying novel variants in known disease genes<sup>5</sup> and in discovering novel disease genes.<sup>6-10</sup> To this end, O’Roak and colleagues sequenced 20 trios with a sporadic ASD to discover deleterious *de novo* mutations in probands.<sup>11</sup> They identified 21 *de novo* mutations in protein coding regions, and found 4 patients with severe clinical manifestations who also presented protein-altering mutations that might be causatively associated with their diseases. Interestingly, the authors suggested a multi-hit model in one particular sporadic case involving a *de novo* mutation of *FOXP1* and a rare inherited missense variant in *CNTNAP2*. The implications of these results are two-fold. First, they proved that the sequencing of individuals with similar clinical features such as severe autistic symptoms would be informative in identifying disease-associated variants. Second, they show that the examination of genetic background for the presence of rare disease-linked variants is crucial to understanding the heterogeneous genetic architecture of ASD.

Here we present a comprehensive analysis of whole genomes from six unrelated patients with ASD using a sequencing-by-ligation technology. We utilized previously studied blood gene

expression profiles as a surrogate for endophenotype. Three patients each were selected based on their enrichment for immune and synaptic pathways respectively. Each of the 6 ASD genomes was found to contain a total of 3.7 ~ 4.0 million genomic variants including 9,002 ~ 10,454 nonsynonymous variants in protein coding regions. Among these nonsynonymous variants, a total of 437 ~ 503 variants were possible loss of function (LOF) changes including frame-shift, nonsense, misstart, nonstop and splice-site disruption. We filtered these for the variants that were rare (defined as < 1% allele frequency (AF) in an ethnicity-matched population) or novel, as well as nonsynonymous, and located at evolutionarily conserved loci, as such variants are more likely to have a functional impact than common variants and play important roles in diseases.<sup>3,12</sup> Our filtering approach was purposefully stringent in its selection of possible disease-linked genes to reduce false positives. Despite the genetic heterogeneity of ASD, different variants among the 6 ASD patients converged to genes previously implicated in the disease; moreover, we found a difference in genetic burden due to rare variants between two groups characterized by gene expression endophenotype.

## **Materials and Methods**

### ***Patients and whole genome sequences***

The patients with ASD were recruited from the Developmental Medicine Center and the Division of Genetics at the Children's Hospital Boston with the Institutional Review Board approval. These patients underwent diagnostic assessment using the Autism Diagnostic Observation Schedule and Autism Diagnostic Interview Revised, as well as comprehensive clinical testing including cognitive testing, language measures, medical history, height, weight, and head circumference measurements, and behavioral questionnaires. Six male patients with autistic disorder have European ancestry (as defined by the HapMap consortium designation EUR), and their clinical and demographic characteristics are listed in **Supplementary Table S1**. These

patients were recruited for WGS analysis according to their blood gene expression profiles in a separate study (*manuscript in preparation*). In that study of 97 patients with ASD and 73 age and gender matched controls, we used pathway enrichment analysis to identify the subset of cases that were specifically enriched for the two most significant gene sets in ASD compared to controls overall, i.e., synaptic or immune response pathways (**Supplementary Methods**). For the current WGS study, we chose three patients—A-0030-P1, A-0042-P1, and A-0076-P1—from those enriched for immune response pathways, and the remaining three—A-0050-P1, A-0069-P1, and A-0091-P1—from those enriched for synaptic pathways such as long term potentiation and gap junction. These patients were highlighted in **Supplementary Figure S1** in which 90 ASD and 55 controls were projected to synaptic and immune response pathways. Hereafter, we refer to the two subgroups as the immune group and the synaptic group. We used Complete Genomics sequencing and assembly technology.<sup>7, 13</sup> For patient A-0050-P1, a genomic DNA (gDNA) sample was extracted from saliva, while all other gDNA samples were extracted from whole blood. There was no difference between the yield and concentration of gDNA from saliva and from blood. The integrity and size of gDNA was checked by pulse-field gel electrophoresis using an agarose gel.

#### ***Annotation and filtering of genomic variants to prioritize the ASD-associated genes***

We developed a WGS analysis pipeline to annotate, filter, and analyze all genomic variants as previously described.<sup>14</sup> The data sources and detailed processing steps for annotation are listed in **Supplementary Methods**. Briefly, the pipeline focuses on two major annotation modules: 1) AF estimated from general European populations of three resources (dbSNP build 132 (dbSNP132),<sup>15</sup> the 1000 Genomes Project (1000GP),<sup>16</sup> and 200 exomes (200Exomes)<sup>17</sup>), and 2) functional impact estimation based on gene model and sequence conservation using the Genomic Evolutionary Rate Profiling (GERP) score for each base position.<sup>18</sup> The combination of these two

analysis modules enables variant filtering and ranking genes and gene sets in the subsequent steps. We performed the following variant filtering process on the 6 ASD genomes and 9 EUR HapMap genomes. First, variants from the 15 genomes (6 ASD + 9 HapMap) were filtered based on the same criteria of 1) *Rare* or *Novel*, 2) *Nonsynonymous*, and 3) highly *Conserved* (GERP score > 2). Hereafter, RNNC will denote the variants meeting these criteria. Second, the number of variants that met the above criteria in each of 29,091 RefSeq transcripts was counted for each of the 15 genomes. Third, the genes that tend to have one or more variants across individuals or to have frequent rare or novel variants (i.e., ‘hypervariable’) were excluded to minimize false positive findings as described in Kohane et al.<sup>14</sup> A total of 118 hypervariable genes met the above criteria, and were excluded for further analysis (**Supplementary Table S2**). Finally, the genes that had one or more RNNC variants in any of 9 EUR HapMap genomes were excluded.

#### ***Gene set analysis for the genes with RNNC variants***

For the genes with RNNC variants in each genome or all those genes together, we searched for enriched pathways using a hypergeometric test. Similar to the hypervariable genes described above, several pathways were frequently highly ranked across all case and non-case genomes. Thus, the gene sets also significantly enriched in non-case genomes were excluded for further evaluation (see **Supplementary Methods**). The gene sets examined in the analysis included (1) a set of 27 known ASD candidate genes listed in Figure 1, Table 2, and Table 3 of Abrahams and Geschwind (referred to here as the ‘Geschwind’ gene set),<sup>2</sup> (2) a set of 2,868 synaptic genes from the SynDB [2006\_Sep.human],<sup>19</sup> and (3) biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) from DAVID functional annotation system.<sup>20</sup>

#### ***Copy number variation analysis***



The gDNA samples from 6 probands and 11 parents were hybridized to the Affymetrix Genome-Wide Human SNP 6.0 arrays in accordance with the manufacturer's protocol. For patient A-0042-P1, only the paternal profile was obtained. After identifying possible CNV loci as described in **Supplementary Methods**, we searched NCBI dbVar and the other structural genomic variants databases<sup>21</sup> to exclude CNV regions that 100% overlapped with common CNV regions. Additionally, the CNV calls from 45 non-cases provided by the Complete Genomics were used to exclude the loci that overlapped more than 50% with hypervariable CNV regions across 45 non-case genomes sequenced using the same platform.

#### ***Genotyping of discovered variants in a larger population***

Fourteen variants present in at least one ASD genome but absent in 9 EUR HapMap genomes were genotyped for 102 patients with ASD including 6 cases from this study and 67 control samples using the Sequenom MassARRAY platform and the iPLEX genotyping protocol (Sequenom, Inc., San Diego, CA, USA)(see **Supplementary Methods**).

## **Results**

#### ***Overview of genomic variants found in the six ASD genomes***

A total of 8,157,540 unique variants that differ from the human reference genome sequence (GRCh36.1, hg18) were found in the 6 ASD genome sequences. Among them, 6,540,491 were SNPs and the other variants included short insertion, deletion and substitution of less than 200 bps. The total number of SNPs per individual ranged from 3,216,075 to 3,409,570 with heterozygous to homozygous ratios of 1.55 to 1.60; among them, averages of 2,025,269 heterozygous (range 1,960,015 - 2,075,080) and 1,292,706 homozygous (range 1,246,596 - 1,334,490) variants were found per individual. The average transition to transversion ratio for

SNPs was 2.13. The numbers of different variant types are enumerated in **Supplementary Table S3**.

An average of 6,242 homozygous and 12,898 heterozygous variants were found in protein-coding sequences including splice sites (defined by the RefSeq, August 2009 release for hg18) across 6 ASD genomes. As illustrated in **Figure 1**, of the common variants - AF  $\geq$  5% in ethnicity-matched populations - similar proportions of them were synonymous (51%) and nonsynonymous (49%). However, significantly higher proportions of the rare and novel variants were nonsynonymous (69 and 68 % respectively, Chi-square test with Yates' correction  $p = 0.0063$  and  $0.0001$  respectively).

#### ***A variant filtering procedure identified new ASD candidate genes***

All variants were filtered based on AF, resulting in ~0.6 million rare or novel variants (AF < 1%). Among the rare or novel variants, ~0.2 million were located within protein-coding genes including exons, introns, UTRs and splice sites. Of these, ~1,286 variants were nonsynonymous including ~462 variants—corresponding to ~412 genes—at highly conserved loci. As a union, a total of 1,788 genes had more than one RNNC variant in at least one ASD genome.

Among RNNC variants, we found ~ 462 (range 388-545) per genome in the 6 ASD cases and ~ 400 (range 345-451) per genome in the 9 controls were identified. Total number of variants, rare or novel, and rare or novel nonsynonymous variants were not significantly different between the two groups (Wilcoxon rank sum test  $p$ -value 0.568, analysis of covariance (ANCOVA)  $p$ -values 0.134 and 0.246 respectively, **Figure 2A-C**). After controlling for the total number of variants identified, the 6 ASD genomes had significantly more RNNC variants than 9 EUR HapMap genomes (ANCOVA  $p$ -value 0.0030, **Figure 2D**). This result must be regarded as tentative due to possible unidentified confounders. For instance, the ASD samples were derived

from fresh blood, whereas the HapMap samples came from lymphoblastoid cell lines although immortalization has been documented to result in more not fewer variants<sup>22</sup>.

To reduce false positives, we excluded 534 genes in which one or more RNNC variants were found among 9 EUR HapMap genomes. We identified ~227 genes with RNNC variants (range 192 - 292) per genome that were mostly private to each case. The 1,254 genes were further prioritized based on the number of ASD genomes with RNNC variants (**Supplementary Table S4**). As listed in **Table 1**, 32 known ASD candidate genes including *ITGB3*, *AVPR1A*, *CADPS2*, *DISC1*, and *MET* had RNNC variants. Among 11 genes that had RNNC variants in 3 of the 6 ASD genomes, *FAT1* was implicated in bipolar affective disorder,<sup>23</sup> and *PLEKHH1* is expressed in synapse according to the SynDB. *FSTL5*,<sup>24</sup> and *RPS6KB2*,<sup>25</sup> and *WDR85*<sup>26</sup> were previously implicated in neuropsychiatric disorders. The ATP-binding cassette, sub-family A, member 4 (*ABCA4*) had RNNC variants in the 3 patients of immune group. The same nonsynonymous variant – p.Leu1970Phe – was found in two patients, and 3 different RNNC variants were found in the patient A-0030-P1. The 3 bps deletion of the Cyclin-D-binding Myb-like transcription factor 1 (*DMTF1*) (7q21, p.Asp66del) was found in three probands although its functional role in neural tissue has not been reported yet. Interestingly, this 3 bps deletion is on a known protein domain that interacts with Cyclin D2, which has been implicated in cortical development,<sup>27</sup> adult neurogenesis,<sup>28</sup> and GABA cell dysfunction in schizophrenia and bipolar disorders.<sup>29</sup>

### ***Enriched pathways for the genes with RNNC variants***

We tested whether the genes with RNNC variants were enriched for biological pathways or known ASD candidate genes. A gene set of known candidate genes from Geschwind (N=27) was significantly enriched in the ASD genomes, suggesting that the frequencies of RNNC variants among known ASD candidate genes were higher compared to random chance (hypergeometric p-value 0.014). Among the biological pathways, the extracellular matrix (ECM) receptor

interaction pathway and ATP-binding cassettes (ABC) transporters were significantly enriched for the 1,254 genes (hypergeometric p-values  $9.41 \times 10^{-6}$  and  $1.23 \times 10^{-4}$  with false discovery rates 0.01 and 0.15% respectively). When each genome was analyzed separately, the ABC transporter pathway was enriched in 3 cases of the immune group but none in the synaptic group (**Table 2**). The 11 genes of this pathway, *ABCA8*, *ABCG8*, *TAP1*, *ABCC3*, *ABCC10*, *ABCC1*, *ABCC2*, *ABCA4*, *ABCB6*, *ABCA13*, and *ABCA5* had RNNC variants in the immune group. *ABCA7* - another member of ABC transporter family - had a RNNC variant in A-0050-P1 of the synaptic group. In the synaptic group, we identified RNNC variants in the genes of Long-term depression pathway such as *ITPR3*, *PLA2G4B*, *PLCB1*, and *PRKCB*, and the glutamatergic synaptic genes such as the neuronal and epithelial glutamate transporter, *SLC1A1* and retinal glutamate transporter, *SLC1A7*. However, no single pathway was enriched across all 3 cases of synaptic group. Significant pathways are listed in **Supplementary Table S6**.

#### ***de novo and inherited copy number variations***

We identified small CNVs (10-100kbps) with averages of 17 losses (range 8-39) and 6 gains (range 4-8) per individual. Among these 140 CNV regions, 32 CNVs (6 gains and 26 losses) were not found in DGV or dbVar databases (**Table 3**). Twenty-six genes were found in the CNV regions, and four genes: *DLC1* (8p22), *SGCZ* (8p22), *CCDC9* (19q13.32), and *PLXNB3* (Xq28), were affected in two or more individuals. A majority of CNV regions smaller than 100 kbps were not discovered with genotyping microarrays. For larger CNVs of > 100kbps, a total of 9 CNVs (4 gains and 5 losses) were identified by read-depth analysis using WGS. These loci were also found with genotyping arrays, and 6 CNVs were inherited CNVs (**Table 3**). Interestingly, the gain at the 8p22 region that encompasses *DLC1* and *SGCZ* was a *de novo* change in A-0050-P1, but the same locus gain was maternally inherited in A-0091-P1. Furthermore, the CNVs of 8p22 have been reported in two independent studies of ASD.<sup>30,31</sup>

### ***Genotyping of 14 RNNC variants in large cohorts***

We selected a set of 14 RNNC variants that were present in at least one ASD genome but not found in any of 9 EUR HapMap genomes. These variants were found in the protein coding regions of genes including *ALPK2*, *CROCC*, *FLG*, *FMN1*, *MLL3*, and a potential hypervariable gene, *HYDIN*, as well as a CTT deletion that introduces a frameshift in *TSC2* (rs34638836).<sup>32</sup> To further examine the association of the 14 RNNC variants with ASD, those variants were genotyped in 184 additional individuals, including 102 individuals with ASD, 15 unaffected family members of 6 patients with ASD recruited for original study, and 67 unrelated neurotypical controls. The frameshift deletion in *TSC2*, rs34638836, had a greater frequency in cases (4 in 204 chromosomes genotyped but none in 134 chromosomes from controls, OR= 3.22, 95% Confidence Interval 0.36-28.49). Furthermore, we found that CTT deletion of *TSC2* was transmitted from the unaffected mother to the case (A-0091-P1). The remaining variants did not show any significant differences in frequencies between ASD and neurotypical controls (**Supplementary Table S7**). The RNNC variants in *HYDIN* were private except for the variant c.6050C>T, supporting the idea that this gene tends to have private RNNC variants across many individuals.

### **Discussion**

The phenotypic heterogeneity behind complex diseases such as ASD presents a great challenge to identify common genetic burdens in a patient population.<sup>33</sup> This study demonstrated the use of a filtering and analysis approach to reduce this genetic complexity from millions of sequence variants per genome to a set of candidate genes and enriched pathways shared among the patients with ASD. By filtering for RNNC (rare/novel, non-synonymous and conserved) variants, we identified highly affected genes and excluded potential false positives (i.e., hypervariable genes)

based on the differential RNNC variant frequency between case and non-case genomes. The enrichment of previously known ASD candidate genes validated this method, while the common pathways pointed to new mechanistic insights to be gleaned from WGS analysis.

Several genes directly implicated in synaptic transmission such as *SLC6A3*, *CHRNA3*, and *KCNJ5* had RNNC variants. *SLC6A3*, a member of dopamine transporter family, is known for the association with neuropsychiatric disorders such as attention deficit hyperactivity disorders, bipolar disorders, and Parkinson's disease. *Slc6a3* null mice show impaired spatial learning and memory, and delayed habituation to the novel environment compared to the wild type mice.<sup>34</sup> The mutations in gamma subunit of acetylcholine receptor encoded by the *CHRNA3* have been discovered in patients with multiple pterygium syndrome, which is a congenital disease characterized by webbing of the neck, elbows, and/or knees and joint contractures.<sup>35</sup> *KCNJ5* is a voltage-gated potassium channel, and physically interacts with the dopamine receptor D4 (DRD4)<sup>36</sup> which has been implicated in ASD and attention-deficit hyperactivity disorder.<sup>37,38</sup>

Interestingly, four genes of the GABAergic synapse pathway (*CCNA1B*, *GABRP*, *GABRR2*, and *GNB3*) were enriched with RNNC variants in the immune group, but not in the synaptic group. The GABA pathway appears to play an important role in immune responses in the brain.<sup>39</sup> The functionally deleterious variants across GABAergic pathway genes may result in heightened response to immune modulators such as chemokines, which in turn, can cause excitotoxic damages to neuronal cells. This finding suggests that blood gene expression profiles may serve as an endophenotype that can help to illuminate the genetic heterogeneity of ASD. Despite the fact that the GABAergic pathway as a whole is not enriched in the immune group, the presence of these variants points to immune response as a potential signature for a subgroup of ASD.

Also notable was the enrichment of ABC transporters in the immune group. The ABC transporters are ubiquitously found in all human tissues, and play important roles in transporting molecules across membranes, including blood-brain barrier (BBB).<sup>40</sup> Mutations in ABC

transporters have been associated with a wide range of Mendelian diseases such as bare lymphocyte syndrome (MIM ID# 604571), cystic fibrosis (MIM ID# 219700), Stargardt disease (MIM ID# 248200), and Tangier disease (MIM ID# 205400).<sup>41</sup> In the BBB, the functional role of ABC transporters in neuroinflammatory diseases such as multiple sclerosis and Alzheimer's disease is thought to be associated with transporting inflammatory mediators.<sup>42-45</sup> These findings, together with previous reports on post-mortem brain studies<sup>46-48</sup> and on epidemiological overlap with other autoimmune disorders<sup>49</sup>, suggest the implication of immune response pathways in the pathophysiology of ASD.

Our filtering approach is purposefully stringent in selecting candidate genes to reduce false positive incidental findings, but as a result several genes implicated in ASD might have been excluded (false negatives).<sup>14</sup> For instance, a RNNC variant in *DIAPH3* was found in A-0050-P1 (NP\_001035982.1, p.Pro614Thr), and two different RNNC variants (c.1808C>T and c.2974C>T) were found in one of the 9 EUR HapMap genomes. Thus *DIAPH3* was excluded by our gene-level filtering criteria. Interestingly, Vorstman et al. reported an ASD case with the same variant (p.Pro614Thr) that was paternally inherited was discovered with maternally inherited deletion of 13q21.2 (i.e., loss of one copy of *DIAPH3*).<sup>50</sup> Another example of possible false negative is *TSC2*, which has been previously implicated in ASD. Two ASD genomes (A-0091-P1 and A-0050-P1) had RNNC variants including the CTT deletion; however, one of the HapMap genomes (NA12004) had a RNNC variant in *TSC2*.

While our findings are intriguing, our study is limited by a small number of samples, and its lack of genotypes in parents and unaffected siblings. We compared 6 ASD genomes with the known variants in dbSNP132, 1000GP, 200Exomes, and 9 unrelated HapMap genomes with EUR ancestry. However, clinical and detailed phenotypic information from these large-scale databases is not readily available. Without parental genomes, we could not determine whether RNNC variants identified in this study were *de novo* or inherited except for 14 variants selected for

genotyping assay. We attempted, unsuccessfully, to check for inheritance among small CNVs using genotyping microarrays. Some loci identified with genotyping arrays were not evident with CNV estimation based on read-depth using WGS. The CNV probes are not evenly distributed along the genome due to microarray design limitation, thus some loci tend to have sparse probes compared to the other loci. This uneven distribution of genes along the chromosome and the probes has been associated with false positive CNV calls depending on the array platform used.<sup>51</sup> There are also genomic regions which high-throughput sequencing technologies cannot sequence, but except for those regions, WGS presented more accurate and higher resolution CNV calls in general. Another limitation of current sequencing techniques is false negatives due to uncalled variants from sequencing an individual once and from highly repetitive or unsequenceable regions. We sequenced one genome (patient ID: A-0050-P1) twice to estimate the overall variants call accuracy in the sequenceable portion of genome. When the called variants between two genomic sequences were compared, 82.72% of calls were identical to each other, and 16.45% were called only in one run, but not called in the other. Among the all variants that were called in two genomic sequences, 0.83% of variants were discordant.

At the conclusion of this initial study, we see two obvious avenues for continued work. First, as more WGS from healthy individuals become available, our approach will become more fruitful as we will be able to relax the stringency of our candidate gene selection procedure. Second, we hope that other groups will incorporate our approach of combining endophenotype based case-selection with WGS, a method that seems to be beneficial in identifying subclasses of diseases with complex phenotypes. Our study showed that the ABC transporter genes were enriched with RNNC variants in the 3 cases with transcriptomic immune signature. The use of endophenotypes to select cases for WGS holds promise for the future study of ASD.



**Conflict of Interest**

The authors declare no conflict of interest.

**Acknowledgements**

We gratefully acknowledge all the participating families and their contributions. This work was supported by grants from Simons Foundation (95117, to L.M.K. and I.S.K.), Nancy Lurie Marks Family Foundation (to L.M.K. and I.S.K.), Molecular Genetics Core laboratory supported by NICHD (P30HD018655, to L.M.K.), and Charles H. Hood Foundation (S.W.K.).

## References

1. Kogan MD, Blumberg SJ, Schieve LA, Boyle CA, Perrin JM, Ghandour RM *et al.* Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. *Pediatrics* 2009; **124**(5): 1395-1403.
2. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 2008; **9**(5): 341-355.
3. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**(6): 415-425.
4. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010; **11**(10): 685-696.
5. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England journal of medicine* 2010; **362**(13): 1181-1191.
6. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* 2010; **42**(1): 30-35.
7. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010; **328**(5978): 636-639.
8. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature genetics* 2010; **42**(6): 483-485.
9. Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C *et al.* Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat* 2010; **31**(8): 918-923.

10. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* 2010; **42**(9): 790-793.
11. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* 2011; **43**(6): 585-589.
12. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010; **8**(1): e1000294.
13. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010; **327**(5961): 78-81.
14. Kohane IS, Hsing M, Kong SW. Taxonomizing, sizing, and overcoming the incidentalome. *Genet Med* 2012.
15. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 2001; **29**(1): 308-311.
16. Consortium GP. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**(7319): 1061-1073.
17. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics* 2010; **42**(11): 969-972.
18. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* 2005; **15**(7): 901-913.
19. Zhang W, Zhang Y, Zheng H, Zhang C, Xiong W, Olyarchuk JG *et al.* SynDB: a Synapse protein DataBase based on synapse ontology. *Nucleic acids research* 2007; **35**(Database issue): D737-741.

20. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; **4(5)**: P3.
21. Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J *et al.* Public data archives for genomic structural variation. *Nature genetics* 2010; **42(10)**: 813-814.
22. Londin ER, Keller MA, D'Andrea MR, Delgrosso K, Ertel A, Surrey S *et al.* Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* 2011; **12**: 464.
23. Abou Jamra R, Becker T, Georgi A, Feulner T, Schumacher J, Stromaier J *et al.* Genetic variation of the FAT gene at 4q35 is associated with bipolar affective disorder. *Molecular psychiatry* 2008; **13(3)**: 277-284.
24. Magri C, Sacchetti E, Traversa M, Valsecchi P, Gardella R, Bonvicini C *et al.* New copy number variations in schizophrenia. *PLoS One* 2010; **5(10)**: e13422.
25. Vazquez-Higuera JL, Mateo I, Sanchez-Juan P, Rodriguez-Rodriguez E, Pozueta A, Calero M *et al.* Genetic variation in the tau kinases pathway may modify the risk and age at onset of Alzheimer's disease. *J Alzheimers Dis* 2011; **27(2)**: 291-297.
26. Verhoeven WM, Kleefstra T, Egger JI. Behavioral phenotype in the 9q subtelomeric deletion syndrome: a report about two adult patients. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2010; **153B(2)**: 536-541.
27. Glickstein SB, Monaghan JA, Koeller HB, Jones TK, Ross ME. Cyclin D2 is critical for intermediate progenitor cell proliferation in the embryonic cortex. *J Neurosci* 2009; **29(30)**: 9614-9624.
28. Jedynak P, Jaholkowski P, Wozniak G, Sandi C, Kaczmarek L, Filipkowski RK. Lack of cyclin D2 impairing adult brain neurogenesis alters hippocampal-dependent behavioral tasks without reducing learning ability. *Behav Brain Res* 2012; **227(1)**: 159-166.

29. Benes FM. Regulation of cell cycle and DNA repair in post-mitotic GABA neurons in psychotic disorders. *Neuropharmacology* 2011; **60**(7-8): 1232-1242.
30. Gai X, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Molecular psychiatry* 2011.
31. Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U *et al.* Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nature genetics* 2010; **42**(6): 489-491.
32. Wilson PJ, Ramesh V, Kristiansen A, Bove C, Jozwiak S, Kwiatkowski DJ *et al.* Novel mutations detected in the TSC2 gene from both sporadic and familial TSC patients. *Hum Mol Genet* 1996; **5**(2): 249-256.
33. State MW, Levitt P. The conundrums of understanding genetic risks for autism spectrum disorders. *Nat Neurosci* 2011; **14**(12): 1499-1506.
34. Gainetdinov RR, Wetsel WC, Jones SR, Levin ED, Jaber M, Caron MG. Role of serotonin in the paradoxical calming effect of psychostimulants on hyperactivity. *Science* 1999; **283**(5400): 397-401.
35. Morgan NV, Brueton LA, Cox P, Grealley MT, Tolmie J, Pasha S *et al.* Mutations in the embryonal subunit of the acetylcholine receptor (CHRNA7) cause lethal and Escobar variants of multiple pterygium syndrome. *American journal of human genetics* 2006; **79**(2): 390-395.
36. Lavine N, Ethier N, Oak JN, Pei L, Liu F, Trieu P *et al.* G protein-coupled receptors form stable complexes with inwardly rectifying potassium channels and adenylyl cyclase. *J Biol Chem* 2002; **277**(48): 46010-46019.
37. Gadow KD, Devincent CJ, Olvet DM, Pisarevskaya V, Hatchwell E. Association of DRD4 polymorphism with severity of oppositional defiant disorder, separation anxiety disorder and repetitive behaviors in children with autism spectrum disorder. *Eur J Neurosci* 2010; **32**(6): 1058-1065.

38. Sharp SI, McQuillin A, Gurling HM. Genetics of attention-deficit hyperactivity disorder (ADHD). *Neuropharmacology* 2009; **57**(7-8): 590-600.
39. Bhat R, Axtell R, Mitra A, Miranda M, Lock C, Tsien RW *et al.* Inhibitory role for GABA in autoimmune inflammation. *Proc Natl Acad Sci U S A* 2010; **107**(6): 2580-2585.
40. Gottesman MM, Ambudkar SV. Overview: ABC transporters and human disease. *J Bioenerg Biomembr* 2001; **33**(6): 453-458.
41. Dean M, Rzhetsky A, Allikmets R. The human ATP-binding cassette (ABC) transporter superfamily. *Genome research* 2001; **11**(7): 1156-1166.
42. Cotte S, von Ahsen N, Kruse N, Huber B, Winkelmann A, Zettl UK *et al.* ABC-transporter gene-polymorphisms are potential pharmacogenetic markers for mitoxantrone response in multiple sclerosis. *Brain* 2009; **132**(Pt 9): 2517-2530.
43. Kooij G, Mizze MR, van Horssen J, Reijkerkerk A, Witte ME, Drexhage JA *et al.* Adenosine triphosphate-binding cassette transporters mediate chemokine (C-C motif) ligand 2 secretion from reactive astrocytes: relevance to multiple sclerosis pathogenesis. *Brain* 2011; **134**(Pt 2): 555-570.
44. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature genetics* 2011; **43**(5): 429-435.
45. Krohn M, Lange C, Hofrichter J, Scheffler K, Stenzel J, Steffen J *et al.* Cerebral amyloid-beta proteostasis is regulated by the membrane transport protein ABCC1 in mice. *J Clin Invest* 2011; **121**(10): 3924-3931.
46. Vargas DL, Nascimbene C, Krishnan C, Zimmerman AW, Pardo CA. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann Neurol* 2005; **57**(1): 67-81.

47. Garbett K, Ebert PJ, Mitchell A, Lintas C, Manzi B, Mirnics K *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiol Dis* 2008; **30**(3): 303-311.
48. Fatemi SH, Folsom TD, Reutiman TJ, Lee S. Expression of astrocytic markers aquaporin 4 and connexin 43 is altered in brains of subjects with autism. *Synapse* 2008; **62**(7): 501-507.
49. Atladottir HO, Pedersen MG, Thorsen P, Mortensen PB, Deleuran B, Eaton WW *et al.* Association of family history of autoimmune diseases and autism spectrum disorders. *Pediatrics* 2009; **124**(2): 687-694.
50. Vorstman JA, van Daalen E, Jalali GR, Schmidt ER, Pasterkamp RJ, de Jonge M *et al.* A double hit implicates DIAPH3 as an autism risk gene. *Molecular psychiatry* 2011; **16**(4): 442-451.
51. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics* 2007; **39**(7 Suppl): S16-21.

## Tables

**Table 1.** Top candidate genes with novel/rare and nonsynonymous variants at evolutionary conserved loci only in the 6 ASD genomes but none of the 9 control genomes. The genes shown in the table were selected from the 1,254 candidate gene list based on their presence in 3 or more ASD genomes and previously reported ASD candidate genes. The complete catalog of variants is listed in **Supplementary Table S4** and supporting references for 37 genes are listed in **Supplementary Table S5**. The variants in bold represent deleterious SNP as predicted by CONDEL, and those in red were not found in 56 HapMap genomes from Complete Genomics (see **Supplementary Methods**).



Gene symbol	Transcript ID	Number of ASD genomes	A0030	A0042	A0050	A0069	A0076	A0091	SynDB	Reference
Number of ASD genomes >= 3										
<i>ABCA4</i>	NM_000350.2	3	p.Glu471Lys, p.Val1589Met, p.Pro1948Leu	p.Leu1970Phe			p.Leu1970Phe			
<i>DMTF1</i>	NM_021145.2	3		p.Asp66del		p.Asp66del	p.Asp66del			
<i>FAM71A</i>	NM_153606.2	3		p.Lys555*	p.Ala112Glu	p.Asn536Asp				
<i>FAT1</i>	NM_005245.3	3	p.Gln2286Asp		p.Val3147Gly			p.Thr746Ala	Yes	
<i>FSTL5</i>	NM_020116.2	3	p.Asn285Ser		p.Asn291Tyr	p.Leu391Pro				
<i>LOC10013203</i>	XM_00172653	3	p.Ser23fs	p.Ser23fs		p.Ser23fs				
<i>LOC10013227</i>	XM_00172212	3			p.Arg76Asn	p.Arg76Asn	p.Arg76Asn			
<i>PLEKHH1</i>	NM_020715.2	3	p.Met438Val	p.Met438Val	p.Arg1348*				Yes	
<i>RPS6KB2</i>	NM_003952.2	3			p.Met236Thr		p.Ala104Thr	p.Asn109Ser		
<i>VWA3A</i>	NM_173615.2	3		p.Arg401Trp		p.Arg797Gln				
<i>WDR85</i>	NM_138778.2	3	p.Ser166fs		p.Arg65Cys		p.Ser166fs			
Previously reported ASD candidate genes										
<i>ITGA4</i>	NM_000885.4	2		p.Val824Ala				p.Ala5Thr, p.Val811Ile	Yes	Correia et al. 2009
<i>ITGB3</i>	NM_000212.2	2					p.Pro711Ser	p.Ile252Val	Yes	Ma et al. 2010
<i>MIB1</i>	NM_020774.2	2	Disrupt (splice site)					p.Arg613Cys	Yes	Gregg et al. 2008
<i>PLXNA4</i>	NM_020911.1	2			p.Val424Ile			p.Ala318Val		Suda et al. 2011
<i>ABCA13</i>	NM_152701.2	1					p.Thr4550Ala			de Krom et al. 2005
<i>ADRB2</i>	NM_000024.4	1						p.Val297Met	Yes	Cheslak-Postava et al. 2005
<i>AVPR1A</i>	NM_000706.3	1	p.Phe308Leu						Yes	Wassink et al. 2004
<i>BZRAP1</i>	NM_004758.1	1						p.Glu1338_Glu1339ins3	Yes	Bucan et al. 2009
<i>CADPS2</i>	NM_017954.9	1		p.Lys925Arg						Sadakata et al. 2007
<i>DCX</i>	NM_000555.2	1			p.Glu54Lys				Yes	Brooks-Kayal 2010
<i>DISC1</i>	NM_018662.2	1						p.Arg575Lys	Yes	Kilpinen et al. 2008
<i>ESR1</i>	NM_000125.2	1				p.Arg269Cys			Yes	Chakrabarti et al. 2002
<i>FOXP2</i>	NM_148898.2	1		p.Pro447Ser						Gauthier et al. 2002
<i>GFAP</i>	NM_002055.2	1		p.Pro47Leu					Yes	Fatemi 2011
<i>HIRIP3</i>	NM_003609.2	1				p.Lys80Arg				Kumar et al. 2009
<i>JMJD1C</i>	NM_032776.1	1			p.Thr973Arg					Castermans 2007
<i>MCHR1</i>	NM_005297.3	1					p.Arg317Gln			de Krom et al. 2005
<i>MCM7</i>	NM_005916.3	1						p.Ile394Thr		Maestrini et al. 2011
<i>MET</i>	NM_000245.2	1	p.Thr992Ile							Mukamel 2011
<i>MOG</i>	NM_002433.3	1						p.Leu22del	Yes	Guerini et al. 2009
<i>MTNRI1A</i>	NM_005958.3	1					p.Gly166Glu		Yes	Jonsson 2010
<i>NFIL3</i>	NM_005384.2	1						p.Ser359Tyr		de Krom et al. 2005
<i>NPY1R</i>	NM_000909.4	1				p.Lys374Thr			Yes	Ramanathan 2004
<i>NTRK2</i>	NM_006180.3	1		p.Pro204His					Yes	Correia et al. 2010
<i>OPRL1</i>	NM_000913.3	1		p.Ala7Val					Yes	de Krom et al. 2005
<i>PER2</i>	NM_022817.2	1					p.Phe876Leu			Nicholas et al. 2007
<i>PIK3CG</i>	NM_002649.2	1						p.Ala197Thr		Serajee et al. 2003
<i>PRKCB1</i>	NM_002738.5	1			p.Tyr123fs				Yes	Lintas et al. 2009
<i>RAPGEF4</i>	NM_007023.3	1					p.Tyr284Cys		Yes	Woolfrey et al. 2007
<i>ROBO4</i>	NM_019055.4	1						p.Ser592Asn		Anitha et al. 2008
<i>RP1L1</i>	NM_178857.5	1		p.Thr112Ser						Glancy et al. 2009
<i>SLC1A1</i>	NM_004170.4	1				p.Asp144Val			Yes	Gadow et al. 2010
<i>SLC6A3</i>	NM_001044.3	1		p.Ser360Phe					Yes	Gadow et al. 2008
<i>SNTG2</i>	NM_018968.2	1	p.Ser204Leu						Yes	Yamakawa et al. 2007
<i>TAOK2</i>	NM_016151.2	1	p.Arg1036Gln							Kumar et al. 2009
<i>TYR</i>	NM_000372.4	1			p.Thr373Lys					Anderson et al. 2007
<i>VIPR1</i>	NM_004624.2	1		p.Gly98Cys						Asano et al. 2001

**Table 2.** Enriched pathways for the genes with rare/novel and deleterious variants across the 6 ASD genomes. We chose three patients—A-0030-P1, A-0042-P1, and A-0076-P1—from those enriched for immune response pathways, and the remaining three—A-0050-P1, A-0069-P1, and A-0091-P1—from those enriched for synaptic pathways such as long term potentiation and gap junction (see Methods). The number represents  $-\log_{10}(p\text{-value of hypergeometric test})$  for each gene set, and red shade corresponds to the color bar below from scale 0 to 10.

KEGG pathways	A-0030-P1	A-0042-P1	A-0076-P1	A-0050-P1	A-0069-P1	A-0091-P1
hsa02010:ABC transporters	1.89	2.60	2.05			
hsa00310:Lysine degradation				1.98		
hsa00020:Citrate cycle (TCA cycle)					1.35	
hsa04662:B cell receptor signaling pathway					1.31	
hsa04512:ECM-receptor interaction						5.19
hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)						2.89
hsa04510:Focal adhesion						2.88
hsa05410:Hypertrophic cardiomyopathy (HCM)						2.63
hsa05414:Dilated cardiomyopathy						1.80



**Table 3.** Copy number variations in the 6 ASD genomes. Most large (> 100 kbps) copy number variations (CNV) were identified by both genotyping microarray and whole-genome sequencing. However, smaller CNV (< 100kb) were not readily found with microarray.

Sample	CNV	Locus	Size(kb)	Inheritance	Genes	Previous reports	WGS	Affymetrix SNP 6.0
<i>Size &gt; 100 kbps</i>								
A-0042-P1	gain	4q12	245	Maternal			Yes	Yes
	gain	14q21.2	111	<i>de novo</i>			Yes	Yes
A-0050-P1	loss	6q21.31	112	Maternal			Yes	Yes
	gain	8p22	1223	<i>de novo</i>	<i>DLC1, SGCZ</i>		Yes	Yes
	loss	9p23	137	Paternal			Yes	Yes
	loss	17p12	1468	<i>de novo</i>	<i>DNAH9, MIR744, MAP2K4, FJ34690, MYOCD, RICH2, ARHGAP44, ELAC2</i>		Yes	Yes
A-0091-P1	loss	5q23.3	252	Maternal	<i>FBN2</i>		Yes	
	gain	8p22	1079	Maternal	<i>DLC1, SGCZ</i>		Yes	Yes
	loss	22q11.22	197	Maternal	<i>TOP3B</i>		Yes	Yes
<i>Size 10 - 100 kbps</i>								
A-0030-P1	loss	7p13	11				Yes	
	loss	10p12.1	18				Yes	
	loss	19q13.32	12		<i>CCDC9</i>		Yes	
	loss	20q13.13	15				Yes	
	gain	Xq28	11		<i>PLXNB3</i>		Yes	
A-0042-P1	gain	Xq23	23				Yes	
	gain	Xq28	12		<i>PLXNB3</i>		Yes	
A-0050-P1	loss	1q43	11				Yes	
	loss	9q34.11	12				Yes	
A-0069-P1	loss	12q24.21	15				Yes	
	gain	Xq28	12		<i>PLXNB3</i>		Yes	
A-0076-P1	loss	1q21.2	46		<i>MGC29891</i>		Yes	
	loss	1p31.3	11				Yes	
	loss	6p21.2	21				Yes	
	loss	7p22.2	35				Yes	
	loss	7p13	11				Yes	
	loss	7p13	12		<i>PPIA</i>		Yes	

	loss	9q34.13	11	<i>SETX</i>	Yes
	loss	9q34.11	13		Yes
	loss	9p13.3	11	<i>NOL6</i>	Yes
	loss	10q21.3	28	<i>SLC25A16</i>	Yes
	loss	10p12.1	18		Yes
	loss	14q12	11	<i>LOC161247</i>	Yes
	loss	19p13.2	10	<i>ZNF791</i>	Yes
	loss	20q13.13	15		Yes
	loss	20q11.22	11		Yes
	loss	20q11.22	15	<i>PHF20</i>	Yes
	loss	21q22.3	12	<i>NDUFV3</i>	Yes
	loss	Xq12	12		Yes
A-0091-P1	gain	11q12.3- 13.1	47	<i>RARRES3,</i> <i>HRASLS3,</i> <i>HRASLS2</i>	Yes
	loss	19q13.32	12	<i>CCDC9</i>	Yes
	gain	Xq28	11	<i>PLXNB3</i>	Yes

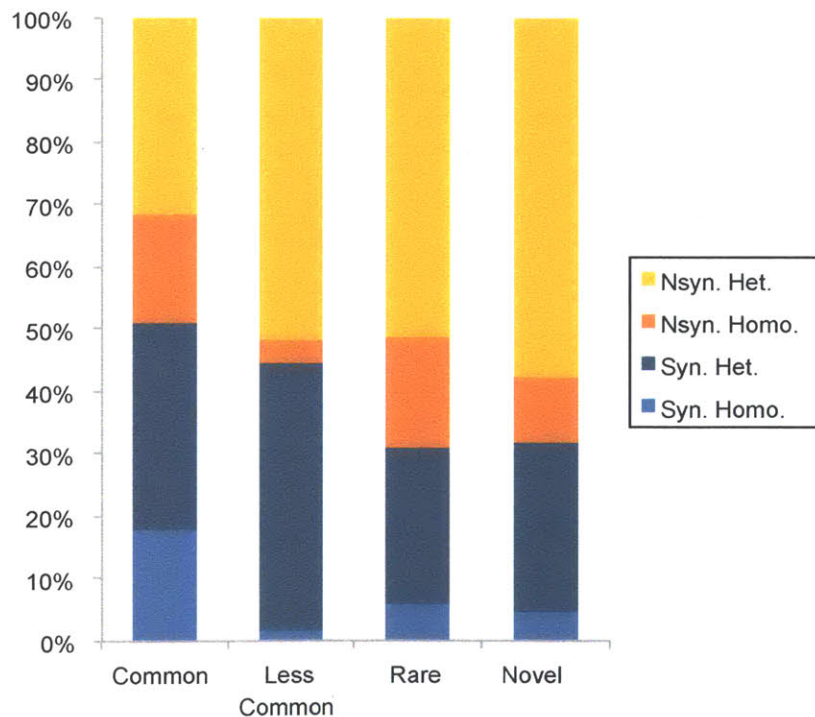
**Figure Titles and Legends**

**Figure 1. Overview of genomic variants in protein coding regions.** Based on three different categories: allele frequency, synonymous/nonsynonymous change and zygosity. A. The average number of variants from each category among the 6 ASD genomes, B. The proportion of nonsynonymous heterozygous (Nsyn. Het.), nonsynonymous homozygous (Nsyn. Homo.), synonymous heterozygous (Syn. Het.) and synonymous homozygous (Syn. Homo.) variants in the common, less common, rare or novel allele frequency category.

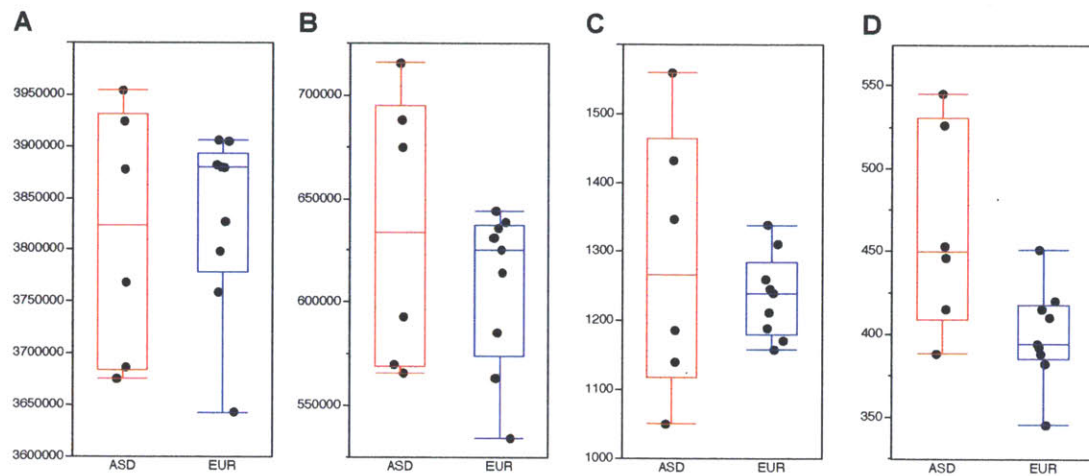
**A**

		Common (AF≥5%)	Less common (5%>AF≥1%)	Rare (AF<1%)	Novel
Nonsynonymous	Heterozygous	5312	412	540	475
	Homozygous	2900	33	189	89
Synonymous	Heterozygous	5449	345	268	228
	Homozygous	2940	13	63	39

**B**



**Figure 2.** Distribution of variants rare or novel nonsynonymous variants at conserved loci between the 6 ASD genomes and 9 control genomes. There were no significant differences in total number of variants (p-value 0.568, **Figure 2A**), rare or novel variants (p-value 0.134, **Figure 2B**), and rare or novel nonsynonymous variants (p-value 0.234, **Figure 2C**) between two groups. After controlling for total variant counts per genome, analysis of covariance showed a significant difference between the ASD and control genomes (p-value of 0.0030, **Figure 2D**).



# **Appendix D:**

## **Distinctive patterns of microRNA expression in primary muscular disorders**

**Iris Eisenberg, Alal Eran, Ichizo Nishino, Maurizio Moggio, Costanza Lamperti, Anthony A. Amato, Hart G. Lidov, Peter B. Kang, Kathryn N. North, Stella Mitrani-Rosenbaum, Kevin M. Flanigan, Lori A. Neely, Duncan Whitney, Alan H. Beggs, Isaac S. Kohane, and Louis M. Kunkel**

This work appeared in Proc Natl Acad Sci U S A. 2007 Oct 23;104(43):17016-21

Author contributions: A.E. analyzed the data and wrote parts of the manuscript

# Distinctive patterns of microRNA expression in primary muscular disorders

Iris Eisenberg<sup>a,b</sup>, Alal Eran<sup>b,c</sup>, Ichizo Nishino<sup>d</sup>, Maurizio Moggio<sup>e</sup>, Costanza Lamperti<sup>e</sup>, Anthony A. Amato<sup>f</sup>, Hart G. Lidov<sup>b,g</sup>, Peter B. Kang<sup>b,h</sup>, Kathryn N. North<sup>i</sup>, Stella Mitrani-Rosenbaum<sup>j</sup>, Kevin M. Flanigan<sup>k</sup>, Lori A. Neely<sup>l</sup>, Duncan Whitney<sup>l</sup>, Alan H. Beggs<sup>b</sup>, Isaac S. Kohane<sup>c</sup>, and Louis M. Kunkel<sup>a,b,m</sup>

<sup>a</sup>Howard Hughes Medical Institute, <sup>b</sup>Program in Genomics, Division of Genetics, <sup>c</sup>Informatics Program, and Departments of <sup>d</sup>Pathology and <sup>e</sup>Neurology, Children's Hospital, Harvard Medical School, Boston, MA 02115; <sup>d</sup>Department of Neuromuscular Research, National Institute of Neuroscience, Tokyo 187-8502, Japan; <sup>e</sup>Department of Neurology, University of Milan, 20122 Milan, Italy; <sup>f</sup>Department of Neurology, Brigham and Women's Hospital, Boston, MA 02115; <sup>g</sup>Institute for Neuromuscular Research, The Children's Hospital at Westmead, New South Wales 2145, Australia; <sup>h</sup>Goldyne Savad Institute of Gene Therapy, Hadassah-Hebrew University Medical Center, Jerusalem 91240, Israel; <sup>i</sup>Department of Human Genetics, University of Utah, Salt Lake City, UT 84132; and <sup>l</sup>US Genomics, Woburn, MA 01801

Contributed by Louis M. Kunkel, August 30, 2007 (sent for review July 25, 2007)

The primary muscle disorders are a diverse group of diseases caused by various defective structural proteins, abnormal signaling molecules, enzymes and proteins involved in posttranslational modifications, and other mechanisms. Although there is increasing clarification of the primary aberrant cellular processes responsible for these conditions, the decisive factors involved in the secondary pathogenic cascades are still mainly obscure. Given the emerging roles of microRNAs (miRNAs) in modulation of cellular phenotypes, we searched for miRNAs regulated during the degenerative process of muscle to gain insight into the specific regulation of genes that are disrupted in pathological muscle conditions. We describe 185 miRNAs that are up- or down-regulated in 10 major muscular disorders in humans [Duchenne muscular dystrophy (DMD), Becker muscular dystrophy, facioscapulohumeral muscular dystrophy, limb-girdle muscular dystrophies types 2A and 2B, Miyoshi myopathy, nemaline myopathy, polymyositis, dermatomyositis, and inclusion body myositis]. Although five miRNAs were found to be consistently regulated in almost all samples analyzed, pointing to possible involvement of a common regulatory mechanism, others were dysregulated only in one disease and not at all in the other disorders. Functional correlation between the predicted targets of these miRNAs and mRNA expression demonstrated tight posttranscriptional regulation at the mRNA level in DMD and Miyoshi myopathy. Together with direct mRNA-miRNA predicted interactions demonstrated in DMD, some of which are involved in known secondary response functions and others that are involved in muscle regeneration, these findings suggest an important role of miRNAs in specific physiological pathways underlying the disease pathology.

skeletal muscle | muscular dystrophies | inflammatory myopathies

Primary muscle disorders involve different groups of diseases, including the muscular dystrophies, inflammatory myopathies, and congenital myopathies. The diseases are defined and classified in accordance with their clinical and pathological manifestations and the distribution of predominant muscle weakness.

The muscular dystrophies are the largest heterogeneous group of >30 different inherited disorders characterized by muscle wasting and weakness of variable distribution and severity, manifesting at any age from birth to middle years, and resulting in significant morbidity and disability (1). Whereas the most characterized forms involve mutations within genes encoding structural members of the dystrophin-associated glycoprotein complex of the muscle membrane cytoskeleton, other mutations interfere with mRNA processing, alter protein posttranslational modifications, or modify enzymatic activities.

Abnormalities of dystrophin are known as the most common cause of muscular dystrophy, accounting for both Duchenne muscular dystrophy (DMD), one of the most severe types with rapidly progressive skeletal muscle weakness, and the milder Becker muscular dystrophy (BMD) phenotype (2). The highly heterogeneous

limb girdle muscular dystrophies (LGMDs) (3) is another major group of muscular dystrophies. Notably, mutated calpain-3 in patients with LGMD type 2A (LGMD2A) was the first enzyme, rather than structural protein, to be associated with muscular dystrophy (4). Mutations in dysferlin, a muscle membrane protein that plays a role in membrane repair, cause the LGMD type 2B (LGMD2B) and Miyoshi myopathy (MM) (5). Facioscapulohumeral muscular dystrophy (FSHD), a progressive muscle disease affecting mainly the muscles of the face and upper arms caused by deletions of a 3.3-kb repeat region located on 4q35.2 (6), is an additional common type of muscular dystrophy.

Among the group of congenital myopathies, nemaline myopathy (NM) is the most common nondystrophic congenital myopathy and is characterized by relatively nonprogressive proximal weakness of often, but not always, congenital onset and the presence of nemaline rod structures in the affected myofibers (7). Mutations in six different genes encoding the thin filament proteins and other skeletal muscle proteins account for the majority of disease cases.

Clinical and histopathologic overlap between the inherited muscular disorders, and the distinct idiopathic inflammatory myopathies is also being increasingly recognized (8). Polymyositis (PM), the most common of the inflammatory myopathies, is a T cell-mediated pathology in which a cellular immune response is a key feature in promoting muscle damage. Inclusion body myositis (IBM) is suspected to be a primary inflammatory myopathy, like dermatomyositis (DM) and PM, or a primary degenerative myopathic disorder, such as a dystrophy with secondary inflammation (9). The general distinction between immune-mediated and non-immune-mediated muscle diseases becomes less defined as more is learned of the complex, underlying pathogenic mechanisms in both inflammatory myopathies and muscular dystrophies.

Currently, although the number of genes identified increases every year, adding to our understanding and revealing the overall complexity of the pathogenesis of the various muscular disorders, and despite the well documented histological pathology of

Author contributions: I.E. and L.M.K. designed research; I.E. and L.A.N. performed research; I.N., M.M., C.L., A.A.A., H.G.L., P.B.K., K.N.N., S.M.-R., K.M.F., and A.H.B. contributed new reagents/analytic tools; I.E., A.E., L.A.N., D.W., and I.S.K. analyzed data; and I.E. and L.M.K. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: miRNA, microRNA; DMD, Duchenne muscular dystrophy; BMD, Becker muscular dystrophy; FSHD, facioscapulohumeral muscular dystrophy; LGMD, limb-girdle muscular dystrophies; LGMD2A, LGMD type 2A; LGMD2B, LGMD type 2B; NM, nemaline myopathy; MM, Miyoshi myopathy; IBM, inclusion body myositis; DM, dermatomyositis; PM, polymyositis; PCA, principal component analysis; ECM, extracellular matrix.

<sup>m</sup>To whom correspondence should be addressed. E-mail: kunkel@enders.tch.harvard.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0708115104/DC1](http://www.pnas.org/cgi/content/full/0708115104/DC1).

© 2007 by The National Academy of Sciences of the USA



dystrophic tissue, the underlying molecular pathways remain poorly understood, and the decisive secondary factors responsible for the variability in the clinical phenotypes are still mainly unknown. Gene expression profiling of human and mouse normal and diseased skeletal muscle has generated more detailed insight in the molecular process underlying the different conditions (10–14). However, although each of these studies has identified a number of genes in various functional categories that are differentially expressed in the disease states, the substantial underlying disease mechanisms remain to be elucidated.

MicroRNAs (miRNAs) are a class of small, endogenous non-coding RNA molecules that posttranscriptionally regulate gene expression. Several hundred mammalian miRNAs have been identified, many of which are tissue-specific and/or temporally regulated in their expression (15). The function of only a small fraction of these has been described in detail and point to their involvement in a variety of developmental and physiological processes (16, 17).

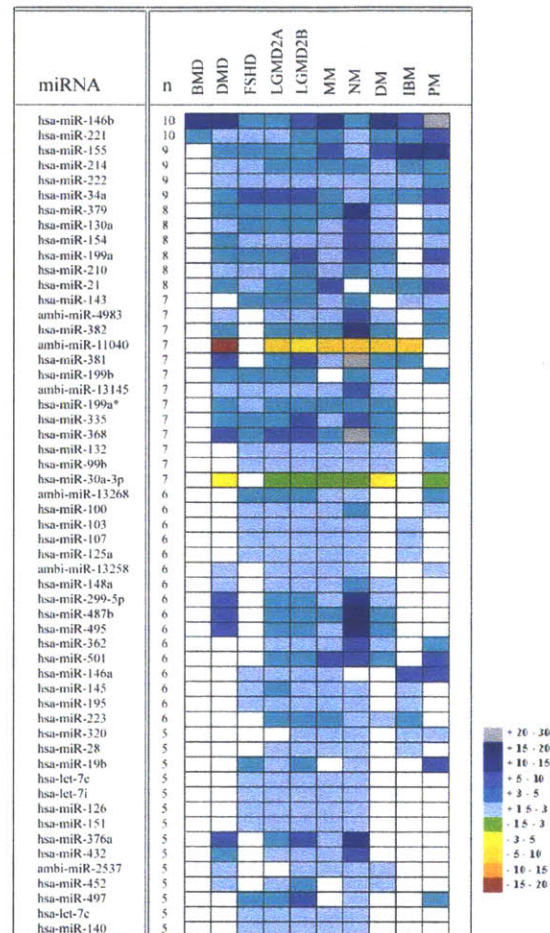
Not surprisingly, miRNAs have been shown to play an important role in the regulation of muscle development. miRNA-1 and miRNA-133 are expressed in cardiac and skeletal muscle and are transcriptionally regulated by the myogenic differentiation factors MyoD, Mef2, and SRF (18–21). In *Drosophila*, deletion of the single miRNA-1 gene results in a defect in muscle differentiation or maintenance (20, 21). In contrast, overexpression of miRNA-1 in mouse cardiac progenitors has a negative effect on proliferation, where it targets the transcription factor Hand2, involved in myocyte expansion (21). Similar to the heart, miRNA-1 overexpression in cultured skeletal myoblasts promotes skeletal muscle differentiation, as does the related but skeletal muscle-specific miR-206 (18, 22) that has also been shown to mediate MyoD-dependent inhibition of follistatin-like-1 and Utrophin genes in myoblasts (23).

In light of their involvement in modulating cellular phenotypes, we hypothesized that miRNAs might be involved in the regulation of the pathological pathways leading to muscle dysfunction, and they might be different among the different pathways that lead to myofiber degeneration. In the present study, we describe a comprehensive miRNA expression profile in muscle tissues from a broad spectrum of primary muscle disorders, aiming to identify new or modifying elements involved in the regulatory networks of muscle and interpret these results within the framework of previous mRNA expression analysis that allowed the examination of the molecular pathophysiological pathways of dystrophic muscle. Further analyses of the overall differentially expressed miRNAs were applied to select potential target genes and unravel biological signaling pathways, potentially targeted by these miRNAs.

## Results

**Overview.** To identify miRNAs that might be involved in the secondary pathological pathways in various muscle diseases and discriminate between the miRNAs possibly involved in the underlying pathways, either specific to a given disease or shared among disease types, we have carried out a comparative miRNA expression profiling across a panel of 10 different groups of muscle disorders (DMD, BMD, FSHD, LGMD2A, LGMD2B, MM, NM, IBM, DM, and PM) with different clinical and pathological characteristics [detailed in supporting information (SI) Fig. 4 and SI Table 2] and unaffected human skeletal muscle.

miRNA expression microarrays containing 428 human miRNAs from the miRBase database and novel human miRNA (Ambi-miRs) were used in this study. In total, a subset of 185 human miRNAs, corresponding to 43% of human miRNA probes present on the array, were found to be differentially expressed at a significant level ( $P < 0.05$ , false discovery rate  $< 0.05$ ) in at least one of the 10 muscle conditions compared with the control panel (SI Table 3). Interestingly, of the differentially expressed miRNAs, most were up-regulated in the different diseases (39 in DMD, 62 in FSHD, 88 in LGMD2A, 87 in LGMD2B, 69 in MM, 140 in NM, 20 in IBM, 37 in PM, and 35 in DM) as compared with normal



**Fig. 1.** miRNAs common to various muscular disorders. The list includes 55 commonly dysregulated miRNAs in five or more types of muscular disorders. A color scale represents the relative intensity of the expression signal by means of fold change compared with the control group, with gray indicating high expression and red low expression. For a complete list see SI Table 4.

muscle tissue. Whereas a total of 151 different miRNA genes were found to be consistently up-regulated relative to the control, only 28 miRNAs were down-regulated among the various conditions. Overlaying this broad commonality is the up-regulation of specific miRNAs and the specific down-regulation of others that allows us to assign a distinctive signature to each of the 10 conditions (Fig. 1 and SI Table 4).

In addition, a set of six miRNAs (30b, 92, 361, 423, 29a, and 29b), was found to be expressed in an inconsistent pattern in few of the conditions (DMD, NM, FSHD, and LGMD2B) such that in one disease the miRNA is down-regulated, whereas in others it is up-regulated, and vice versa (SI Table 4). This finding might point out the differences in the pathology and genes involved in the different regulated networks.

Among the 185 differentially expressed miRNAs, the expression profile in human tissues has been previously established for 145 (see *Materials and Methods*). Of these, 60% (87/145) are known to be expressed in adult muscle (and in other tissues), whereas the expression of the other 58 miRNAs, mostly up-regulated, was not previously detected in adult muscle to our knowledge. Moreover, almost a fifth of these nonmuscle miRNAs (11/58 miRNAs) were detected in cells of the immune system, including lymphocytes and macrophages (SI Table 4). These findings are consistent with the persistent inflammatory

response observed in many dystrophic skeletal muscles that leads to an altered extracellular environment, including an increased presence of inflammatory cells and elevated levels of various inflammatory cytokines.

**Direct Quantification of miRNA Gene Expression for Validation of Microarray Results.** The Trilogy technology (24) for quantification of miRNA expression was selected for validating miRNA microarray data. Ten different miRNAs showing distinct expression patterns in the 10 different diseases (miR-21, miR-22, miR-29c, miR-30a-3p, miR-146b, miR-221, miR-368, miR-379, Ambi-miR-693, and Ambi-miR-11040) and two miRNAs with no significant variation (miR-10b and miR-100), were quantified in two independent replicates. Thirty-nine of the RNA samples previously profiled on the arrays were analyzed on the Trilogy platform with an average of three different samples analyzed for each miRNA in any given disease. The expression of miR-146b, miR-379, miR-221, and miR-368 was below the limits of detection of this assay. The relative variations of miRNA expression levels for miR-21, miR-22, miR-29c, miR-30a-3p (except for LGMD2A), and Ambi-miR 11040 were in concordance with the normalized array data, thus validating our array results (SI Fig. 5). Ambi-miR-693, however, which was found on the arrays as down-regulated compared with muscle biopsies from unaffected individuals in all of the examined diseases (LGMD2B, MM, and NM), was found here as being up-regulated. Although we have not determined the exact cause for this discrepancy, it should be noted that there is no up-front enrichment for small RNAs in the Direct assay (unlike our microarray assays), thus we cannot exclude the possibility that precursors are also being quantified.

**Distinctive Patterns of miRNA Expression Are Associated with Different Types of Primary Muscle Disorders.** Five miRNAs (miR-146b, miR-221, miR-155, miR-214, and miR-222) (Fig. 1 and SI Table 4) were found to be consistently dysregulated in almost all samples analyzed in the study (with an exception for BMD in which the last three miRNAs were also dysregulated but with a fold change  $<1.5$  and therefore are not included in SI Table 4), across the various diseases. This finding might suggest that these miRNAs are involved in a common underlying regulatory pathway among all diseases. By contrast, other miRNAs were dysregulated only in one given disease and not in any of the others: miR-486, miR-485-5p, miR-331, miR-30e-5p, miR-30d, miR-30a-5p, miR-26a, miR-22, miR-193b, miR-101, miR-95, Ambi-miR-7075, and Ambi-miR-13156, all in muscle biopsies taken from Duchenne patients; miR-517\* in FSHD; Ambi-miR-10617 in LGMD2A; miR-301 in LGMD2B; and miR-302c\* in MM. The finding of two different miRNAs uniquely dysregulated each in one of the dysferlinopathies and not in the other might point to the involvement of a different secondary regulatory mechanism in the two different phenotypes despite their being allelic diseases. In NM a much larger set of 36 different miRNAs was uniquely dysregulated.

Among the set of miRNAs dysregulated in the various dystrophies (DMD, BMD, and FSHD), 49 in DMD and 38 in FSHD are also dysregulated in various other nondystrophic muscle diseases. Nonetheless, narrow subsets of diseases with shared miRNA profiles were identified. These include: miR-29a in DMD and FSHD; miR-30c in DMD and MM; miR-30b, miR-92, miR-29c, miR-423, miR-361, miR-299-3p, and miR-181d in DMD and NM (Fig. 1 and SI Table 4). We also noted miRNAs with a shared profile across FSHD and the following diseases: miR-16 in FSHD and LGMD2A; miR-279 in FSHD and LGMD2B; and miR-99a, miR-93, miR-455, miR-20b, miR-18a, miR-17-5p, miR-152, miR-106a, and miR-106b, all in FSHD, LGMD2A, LGMD2B, and NM.

The LGMDs analyzed in this study (LGMD2A, LGMD2B, and MM) present a tighter intragroup correlation of miRNAs compared with the other dystrophies. Expression changes for 52 miR-

NAs were shared by all three diseases, and other dysregulated miRNAs were expressed in only one of the diseases or in the two dysferlin phenotypes. In addition, a set of miRNAs was detected as specifically differentially expressed in each of the LGMDs in concordance with NM (Fig. 1 and SI Table 4).

The most extensive dysregulation of miRNAs was observed in NM with  $>150$  different miRNAs being dysregulated, and of these, 36 being dysregulated solely in this phenotype (SI Table 4). These results could be explained by the high genetic heterogeneity of the underlying disease cause, in which six different genes probably reflecting six different disease mechanisms are known to be involved in the disorder (7).

More than 60% (116) of the overall differentially expressed miRNAs in this study were implicated in at least one of the inflammatory myopathies studied, but not exclusively dysregulated in any of them. Excluding those miRNAs that were common among the three inflammatory myopathies but were also shared by most of the other diseases in the study (miR-146b, miR-221, miR-155, miR-214, and miR-222), not much overlap in the different dysregulated miRNAs was observed although  $>20$  different miRNAs were found to be dysregulated in each of the conditions (Fig. 1 and SI Table 4). These results provide an opportunity to distinguish between those inflammatory-related miRNAs and classify the other 69 dysregulated miRNAs as being involved in other pathologic processes taking place in the different affected myofibers and could give rise to the heterogeneous phenotypes.

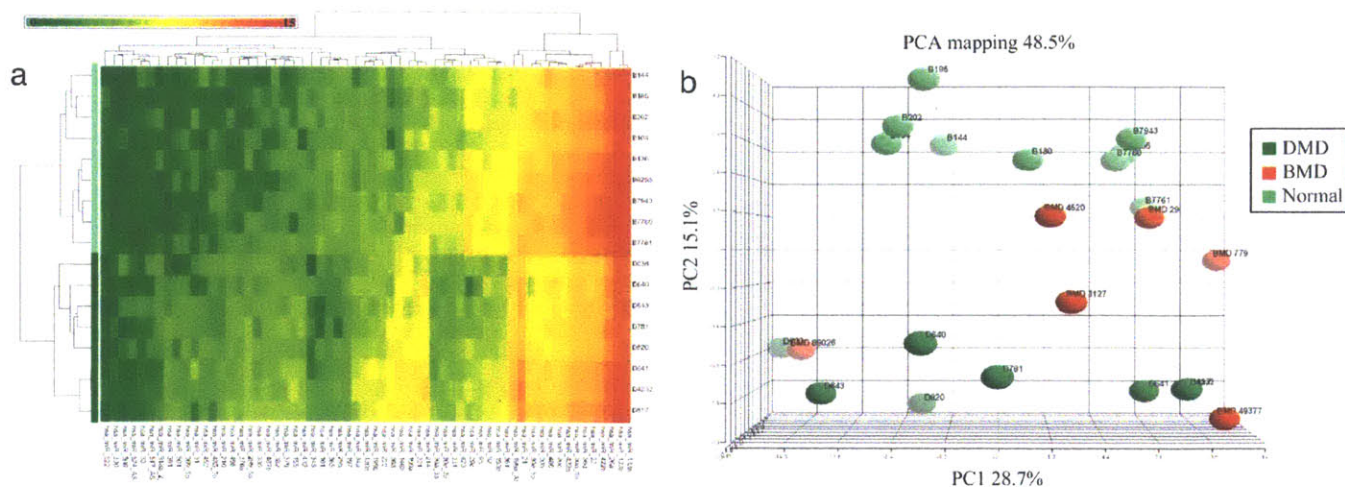
Hierarchical clustering (Fig. 2 and SI Fig. 6) and principal component analysis (PCA) of miRNAs selected by ANOVA clearly segregate and separate muscle biopsies taken from the different muscular disorders from the normal control muscle groups based on their miRNA expression profile.

**Functional Correlation Between mRNA and Predicted miRNA Targets in Muscular Disorders.** To gain insight into the function of miRNAs during the disease process, we analyzed the functional correlation between miRNAs and mRNA expression to identify differentially expressed miRNA-mRNA modules and assess the extent of miRNA effects on mRNA expression. First, we inspected the functional correlation between dysregulated mRNAs and targets of dysregulated miRNAs, which allows capturing of indirect target effects, where miRNAs bind regulatory proteins, such as transcription factors, and the latter exert the main effect.

A total of 10 mRNA and 10 miRNA datasets were analyzed (SI Table 5). A meta miRNA predictor (MAMI) enabling maximal accuracy and tunable sensitivity and specificity in predictions was applied to predict targets of differentially expressed miRNAs (A.E., C. Freifield, A. T. Kho, I.E., M. Galdzicki, K. Naxerova, M. F. Ramoni, L.M.K., and I.S.K., unpublished work).

A strong functional correlation was detected only in DMD and MM, suggesting that these two diseases have a tight posttranscriptional regulation at the mRNA level. In DMD, the correlation between the functions of down-regulated mRNAs and those of targets of up-regulated miRNAs reached significance with  $P < 0.0157$ . The functional correlation between up-regulated mRNA and targets of down-regulated miRNAs had a  $P < 0.023$ . In MM, the correlations' significances were  $P < 2.01E-06$  and  $P < 0.03413$  for functions of down-regulated mRNAs with targets of up-regulated miRNAs and up-regulated mRNAs with targets of down-regulated miRNAs, respectively.

miRNA-mRNA modules shared by DMD and MM include extracellular matrix (ECM) processes and cytoskeletal organization. Ion channel activity module was down-regulated in MM, and a strong down-regulation of a transcriptional activity module was observed in DMD. SI Fig. 7 shows an example of the miRNA-mRNA ECM module in DMD, where the overexpressed ECM protein-coding genes are regulated by both direct interaction with down-regulated miRNAs and through their mRNA targets. The overall network structure reveals tight posttranscriptional regula-



**Fig. 2.** Unsupervised hierarchical clustering and PCA of miRNA expression differentiate DMD from normal muscle. Sixty miRNAs with significantly different expression between DMD ( $n = 8$ , D) and normal individuals ( $n = 9$ , B) were identified by ANOVA. (a) Hierarchical clustering of 17 samples and 60 genes. Each row represents an individual, and each column represents a miRNA gene. A color code represents the relative intensity of the expression signal, with red indicating high expression and green indicating low expression. (b) PCA of ANOVA-selected miRNAs. In this plot, the first principle components (PC1) axis accounted for 28.7% of the variance in the data set and is a result of noise, possibly introduced by different muscle types and genders. The second principle component (PC2) accounts for 15.1% of the variance and segregates DMD from normal individuals. BMD samples ( $n = 6$ ) are found as intermediate between DMD and normal muscle, with a distribution consistent with their phenotypic characteristics. The profiles from more severely affected patients (BMD49377 and BMD89026) are found with those of DMD patients, whereas the mildly affected BMD patients (BMD29 and BMD4620) are close to normal muscle.

tion whose alteration might contribute to secondary pathological processes in the dystrophic muscle in DMD, by either direct miRNA targeting or through secondary proteins.

**Inference of miRNA Functions in Dystrophic Muscle Pathology.** Further insights into the biological pathways potentially regulated by miRNAs in the dystrophic process were obtained by direct comparison between the genes previously found as dysregulated in DMD (11) and the predicted target genes for the 62 differentially expressed miRNAs found in DMD. Fifty-seven mRNA–miRNA interactions were identified, representing 28 genes as targeted by at least one miRNA and dysregulated in DMD. About 42% of these genes were predicted to be targeted by multiple miRNAs (Table 1).

Earlier expression studies have demonstrated that significantly more mRNAs are overexpressed in dystrophic muscle than under-expressed compared with unaffected muscle (11), most likely because of an increase in protein turnover caused by the degenerative and regenerative nature of the disease. In the present analysis, muscle structure and regeneration and ECM genes were among these predicted miRNA targets. Interactions like proenkephalin–miR-29c; collagen, type I, alpha2–miR-29c; trophinin–miR-29c; RUNX1–miR-30a-5p, and PDE4D–miR-199a demonstrated high reciprocal fold change of the relevant miRNA and mRNA. At the mRNA level, dystrophin and several other structurally related proteins, which are substantially underexpressed in DMD muscle, were not predicted to correlate with any of these miRNAs.

**Dysregulated miRNAs in Muscular Disorders Are Significantly Associated with Diverse Signaling Pathways.** With the understanding that identification of mRNA targets is predictive, we analyzed the functional enrichment of predicted targets of differentially expressed miRNAs in each of the muscle phenotypes in an attempt to uncover the functional meaning among these dysregulated miRNAs.

In muscle biopsies from Duchenne patients, the 39 up-regulated miRNAs were identified to potentially target the 3' UTRs of  $\approx 5,000$  genes (of which 807 are muscle-expressed genes) and  $>4,400$  genes were identified to be targeted by the 23 miRNAs down-regulated in the disease (SI Table 6). Notably, the 11 miR-

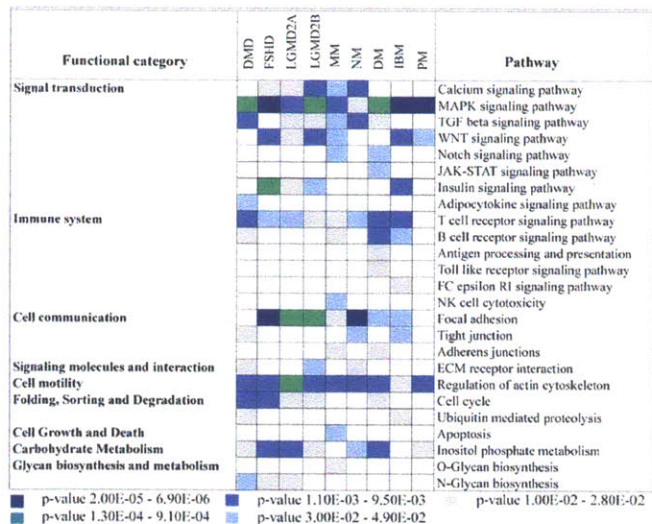
NAs dysregulated exclusively in DMD were predicted to target  $>400$  genes of diverse functions, by more than one prediction tool (SI Table 6). The detailed information for the other diseases is summarized in SI Table 6.

To analyze the role that these differentially expressed

**Table 1. Direct miRNA–mRNA predicted targeting in DMD**

miRNA	Target	miRNA fold change	Target fold change
hsa-miR-30c	VIM	-2	2
hsa-miR-148a	UCP3	2	-3
hsa-miR-130a	UBE2D1	2	-3
hsa-miR-101	TUBB2A	-2	3
hsa-miR-29c	TRO	-6	4
hsa-miR-26a	SRPX	-2	3
hsa-miR-101	SPARC	-2	2
hsa-miR-30c	RUNX1	-2	6
hsa-miR-197	PRMT2	-1	3
hsa-miR-29c	PENK	-6	8
hsa-miR-199a	PDE4D	3	-4
hsa-miR-29a	PXDN	-2	2
hsa-miR-214	LMOD1	2	-3
hsa-miR-29c	HOM-TE5-103	-6	2
hsa-miR-22	HSPG2	-3	2
hsa-miR-22	GPNMB	-3	2
hsa-miR-210	GPD1L	2	-2
hsa-miR-21	FAM50B	3	-1
hsa-miR-197	EEF1A1	-1	2
hsa-miR-26a	EPB41L3	-2	4
hsa-miR-101	CFH	-2	4
hsa-miR-29c	COL3A1	-6	6
hsa-miR-26a	COL1A2	-2	5
hsa-miR-29c	COL1A2	-6	5
hsa-miR-22	CLIC4	-3	2
hsa-miR-193b	CLIC1	-2	2
hsa-miR-30c	CD99	-2	2
hsa-miR-30a-3p	ANXA1	-4	2
hsa-miR-30c	ACTN1	-2	2

For targets predicted to interact with several miRNAs, the predicted interaction with the highest MAMI score is presented.



**Fig. 3.** Overrepresented miRNA regulatory pathways in primary muscular disorders. Fisher's exact test was used to identify significant enrichment for pathway annotations among predicted targets of the dysregulated miRNAs in the different diseases. Each column corresponds to a single disease, and each row corresponds to a KEGG pathway with an overrepresentation of miRNA targets. Pathways have been grouped in larger functional categories according to the KEGG annotation. Only pathways with at least one significant association are shown, and the confidence for enrichment of targets in a given pathway is shown by color-coding the *P* value ranges.

miRNAs play in the regulatory networks in muscular disorders, we have used the KEGG database and the DAVID bioinformatics resources (26) to identify significantly overrepresented biological pathways. Fig. 3 shows the overrepresented pathways identified in at least one muscle phenotype, using the overall predicted targets in any given disease.

Genes that were commonly targeted by the dysregulated miRNAs in muscle specimens from DMD patients, for instance, were significantly clustered in 12 biochemical pathways with some, like TGF- $\beta$  ( $P = 3.50E-03$ ), being targeted by both up-regulated and down-regulated miRNAs, suggesting an extensive miRNA regulation of this pathway in DMD. The overall analysis highlighted 25 different pathways as being significantly targeted by the different miRNAs involved in the primary muscle diseases studied. Of those, six major signal transduction signaling pathways previously described (9, 27) as involved in various aspects of muscular disorders, such as TGF- $\beta$ , calcium signaling pathway, Wnt, Notch, and MAPK signaling pathways, were found to be significantly targeted by the dysregulated miRNAs described in this study (Fig. 3).

Pathways related to the immune response were also significantly enriched in this data set in all diseases, with T cell receptor signaling pathway ( $P = 4.50E-03$ ) being most abundant. Consistent with previous studies and more recent mRNA expression analysis, the immune-related pathways were highly enriched in two of the inflammatory myopathies, DM and IBM (and surprisingly not in PM patients, maybe because of previous steroids treatment and/or the amount of inflammatory cell infiltrates). Cellular pathways, including cell motility ( $P = 6.50E-04$ ), cell communication ( $P = 2.40E-05$ ), degradation ( $P = 6.60E-03$ ), and others were also found to be extensively regulated by these miRNAs (Fig. 3).

## Discussion

Significant progress has been made in the understanding of muscle dysfunction and the causative mechanism behind the major muscular dystrophies has been explored, but knowledge of the underlying regulatory network(s) is still incomplete. Compelling evidence has demonstrated the substantial regulatory role of miRNAs in

muscle development and more recently in the etiology of cardiac failure (28). In light of these findings, we have examined miRNAs involved in major myopathological diseases in humans to gain insight into the specific regulation of genes that are disrupted in pathological muscle conditions.

A total of 185 miRNAs with statistically significant differential expression were identified in the 10 distinct forms of muscular dystrophies analyzed in the present work. Of those, a subgroup of 18 miRNAs was identified that correctly predict and distinguish the various diseases from the normal muscle tissue, with >90% accuracy in most groups (SI Table 7). In contrast to many previous studies, mostly in cancer, showing a global reduction of mature miRNA levels compared with normal tissues, an increase in abundance for many miRNAs was observed in the different muscular disorders. In this report, we provide evidence that miRNAs have a potential role in the pathophysiology of primary muscle diseases and present the complete suite of known miRNAs with altered expression in these diseases. These miRNA signatures provide the basis for a list of common target genes whose misregulation may contribute to the pathology of these disorders.

Secondary to the genetic defects, necrosis and inflammation play a crucial role in the pathogenesis of the different muscular dystrophies and myopathies, and expression profiling of various diseased muscles revealed distinct patterns of immune or immune modulatory pathways rather than nonspecific processes (11, 30). Although the immunopathology of these disorders is not fully understood, several miRNAs previously described as immune-related were found as commonly dysregulated among the various dystrophies. Together with the different patterns of dysregulated miRNAs unique to each of the different diseases, this pattern offers insights into the complexities of the inflammatory process taking place in the different affected muscle fibers.

In contrast to studies associating the overexpression of miR-155 with malignancy in humans (31), the present report describes the ubiquitous up-regulation of this immune-related miRNA (32, 33) in a completely different and unrelated context, raising intriguing questions about its functional role in the pathological process in muscle.

Despite the elucidation of several clinically relevant signal transduction pathways that can lead to disease progression, the means by which these pathways are coordinated with respect to the development and progression of muscle disease process remain obscure. Induction of miR-146 expression by activated NF- $\kappa$ B has been recently demonstrated by Taganov *et al.* (34), and its role in the immune system was also described. Evidence of perturbation of NF- $\kappa$ B signaling has been described in the process of modulating the immune response in several different dystrophies (35–37) and the inflammatory myopathies (38). It will be important to identify and analyze miR-146 downstream target genes and gain insights into the signaling pathways altered by the aberrant up-regulated expression of this miRNA in the different primary muscle diseases.

Currently, the major difficulty for functional studies of miRNAs is in determining their specific target genes at the transcriptional or translational level. Available prediction algorithms frequently predict hundreds of target genes for any single miRNA, and it is likely that this high number of genes contains a significant fraction of false positives. To restrict this high number and enrich for more reliable predicted targets, we have applied a meta predictor tool recently developed that integrates the leading prediction methods into an improved predictor. Beyond the predicted lists of targets the significant associations inferred between the sets of functional targets predicted for the overall miRNAs in each of the diseases and specific cellular pathways can be used to shape some initial hypotheses on how alteration of miRNA expression may be directly involved in different types of diseased muscles.

Furthermore, the functional correlation between the differentially expressed mRNAs and miRNAs as a module in DMD revealed a tight posttranscriptional regulation network at the

mRNA level whose alteration might contribute to increased immune response, by either direct miRNA targeting or through secondary proteins. Together with the specific mRNA-miRNA predicted interactions, some of which are directly involved in compensatory secondary response functions like connective tissue infiltration, and others that are involved in muscle regeneration, these findings raise the opportunity for therapeutic intervention at the miRNA level in preventing specific physiological pathways underlying the disease. However, because it remains difficult to estimate the true false-positive rate of the overall target prediction, a better understanding of the biological significance of these miRNAs and the alterations found in the different muscle diseases would be ultimately achieved by the development of experimental models.

**Conclusion.** Considerable advances have been made in understanding the mechanisms, both transcriptional and translational, that lead to altered gene expression under dystrophic conditions. Our results point to an additional dimension of regulation of muscle function mediated by miRNAs. An important aim for the future will be to experimentally assess the predicted targets of the miRNAs responsible for adverse skeletal muscle remodeling in the different diseases. The overall discovery of dysregulated miRNAs in the different diseases is expected not only to broaden our biological understanding of these diseases, but more importantly, to identify candidate miRNAs as potential targets for future clinical applications.

## Materials and Methods

**Patient Samples and RNA Isolation.** A total of 88 muscle specimens, representing 11 different human muscle conditions, were available for this study, all in compliance with the involved institutions approved protocols. RNAs were isolated with the mirVana miRNA Isolation Kit (Ambion, Austin, TX) according to the manufactur-

er's instruction for total RNA isolation. More detailed information is provided in *SI Text*, *SI Table 2*, and *SI Fig. 4*.

**miRNA Array Analyses.** RNA samples were processed by Asuragen Services (Austin, TX) according to the standard operating procedures of the company as described (39). For a detailed description see *SI Text*. Microarray data processing and analyses are described in detail in *SI Text*.

**miRNA Quantification Assay.** Direct miRNA detection and quantification was performed by using the Direct miRNA assay (US Genomics) as described (24).

**Assessment of miRNA and mRNA Expression in Normal Human Tissues.** Four independent data sets of miRNA expression across normal human tissues were used to assess miRNAs expression in adult skeletal muscle (25, 29, 39, 40). The assessment of mRNA expression in normal adult skeletal muscle is described in *SI Text*.

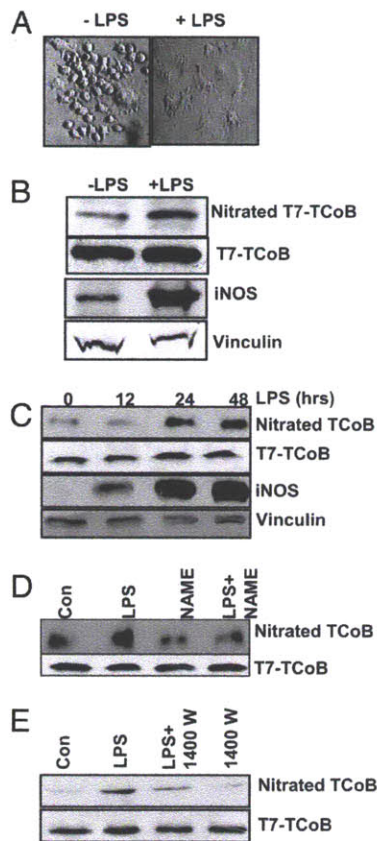
**Functional Inference of miRNA and miRNA-mRNA Correlation Analysis.** Functional inference of miRNA and miRNA-mRNA correlation analysis are detailed in *SI Text*.

We thank Drs. Tim Davison and Charles Johnson (Asuragen) for their expertise in the microarray application and excellent assistance; Dr. Marco Ramoni for data analysis; and Elicia Estrella, Kamila Naxerova, Michal Galdzicki, Joon Lee, and members of L.M.K.'s laboratory for helpful comments and suggestions. K.M.F. is supported by National Center for Research Resources Grant M01-RR00064 (to the University of Utah, Dr. Lorris Betz). M.M. and C.L. are supported by the Associazione Amici del Centro Dino Ferrari, Telethon Project GTF02008, and Eurobiobank Project QLTR-2001-02769. A.H.B. is supported by the Muscular Dystrophy Association, National Institutes of Health Grant R01-AR044345, and generous gifts from the Lee and Penny Anderson Family Foundation and the Joshua Frase Foundation. L.M.K. is an Investigator with the Howard Hughes Medical Institute.

- Davies KE, Nowak KJ (2006) *Nat Rev* 7:762-773.
- Monaco AP, Bertelson CJ, Liechti-Gallati S, Moser H, Kunkel LM (1988) *Genomics* 2:90-95.
- Laval SH, Bushby KM (2004) *Neuropathol Appl Neurobiol* 30:91-105.
- Richard I, Broux O, Allamand V, Fougereuse F, Chiannilkulchai N, Bourg N, Brenguier L, Devaud C, Pasturaud P, Roudaut C, et al. (1995) *Cell* 81:27-40.
- Bansal D, Campbell KP (2004) *Trends Cell Biol* 14:206-213.
- Tawil R, Van Der Maarel SM (2006) *Muscle Nerve* 34:1-15.
- Agrawal PB, Greenleaf RS, Tomczak KK, Lehtokari VL, Wallgren-Pettersson C, Wallefeld W, Laing NG, Darras BT, Maciver SK, Dormitzer PR, et al. (2007) *Am J Hum Genet* 80:162-167.
- Hoffman EP, Rao D, Pachman LM (2002) *Rheum Dis Clin North Am* 28:743-757.
- Dalakas MC (2006) *Nat Clin Pract Rheumatol* 2:219-227.
- Lennon NJ, Kho A, Baeskaai BJ, Perlmutter SL, Hyman BT, Brown RH, Jr (2003) *J Biol Chem* 278:50466-50473.
- Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM (2002) *Proc Natl Acad Sci USA* 99:15000-15005.
- Winokur ST, Chen YW, Masny PS, Martin JH, Ehmsen JT, Tapscott SJ, van der Maarel SM, Hayashi Y, Flanigan KM (2003) *Hum Mol Genet* 12:2895-2907.
- Sanoudou D, Haslett JN, Kho AT, Guo S, Gazda HT, Greenberg SA, Lidov HG, Kohane IS, Kunkel LM, Beggs AH (2003) *Proc Natl Acad Sci USA* 100:4666-4671.
- Greenberg SA, Sanoudou D, Haslett JN, Kohane IS, Kunkel LM, Beggs AH, Amato AA (2002) *Neurology* 59:1170-1182.
- Bartel DP (2004) *Cell* 116:281-297.
- Alvarez-Garcia I, Miska EA (2005) *Development (Cambridge, UK)* 132:4653-4662.
- Kloosterman WP, Plasterk RH (2006) *Dev Cell* 11:441-450.
- Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, Conlon FL, Wang DZ (2006) *Nat Genet* 38:228-233.
- Kwon C, Han Z, Olson EN, Srivastava D (2005) *Proc Natl Acad Sci USA* 102:18986-18991.
- Sokol NS, Ambros V (2005) *Genes Dev* 19:2343-2354.
- Zhao Y, Samal E, Srivastava D (2005) *Nature* 436:214-220.
- Kim HK, Lee YS, Sivaprasad U, Malhotra A, Dutta A (2006) *J Cell Biol* 174:677-687.
- Rosenberg MI, Georges SA, Asawachaicharn A, Analau E, Tapscott SJ (2006) *J Cell Biol* 175:77-85.
- Neely LA, Patel S, Garver J, Gallo M, Hackett M, McLaughlin S, Nadel M, Harris J, Gullans S, Rooke J (2006) *Nat Methods* 3:41-46.
- Liang Y, Ridzon D, Wong L, Chen C (2007) *BMC Genomics* 8:166.
- Dennis G, Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) *Genome Biol* 4:P3.
- Bassel-Duby R, Olson EN (2006) *Annu Rev Biochem* 75:19-37.
- van Rooij E, Sutherland LB, Qi X, Richardson JA, Hill J, Olson EN (2007) *Science* 316:575-579.
- Calin GA, Liu CG, Sevignani C, Ferracin M, Felli N, Dumitru CD, Shimizu M, Cimmino A, Zupo S, Dono M, et al. (2004) *Proc Natl Acad Sci USA* 101:11755-11760.
- Chen YW, Zhao P, Borup R, Hoffman EP (2000) *J Cell Biol* 151:1321-1336.
- Eis PS, Tam W, Sun L, Chadburn A, Li Z, Gomez MF, Lund E, Dahlberg JE (2005) *Proc Natl Acad Sci USA* 102:3627-3632.
- Rodriguez A, Vigorito E, Clare S, Warren MV, Couttet P, Soond DR, van Dongen S, Grocock RJ, Das PP, Miska EA, et al. (2007) *Science* 316:608-611.
- O'Connell RM, Taganov KD, Boldin MP, Cheng G, Baltimore D (2007) *Proc Natl Acad Sci USA* 104:1604-1609.
- Taganov KD, Boldin MP, Chang KJ, Baltimore D (2006) *Proc Natl Acad Sci USA* 103:12481-12486.
- Baghdiguian S, Martin M, Richard I, Pons F, Astier C, Bourg N, Hay RT, Chemaly R, Halaby G, Loiselet J, et al. (1999) *Nat Med* 5:503-511.
- Macaione V, Aguenouz M, Rodolico C, Mazzeo A, Patti A, Cannistraci E, Colantone L, Di Giorgio RM, De Luca G, Vita G (2007) *Acta Neurol Scand* 115:115-121.
- Acharyya S, Villalta SA, Bakkar N, Bupha-Intr T, Janssen PM, Carathers M, Li ZW, Beg AA, Ghosh S, Sahenk Z, et al. (2007) *J Clin Invest* 117:889-901.
- Monici MC, Aguenouz M, Mazzeo A, Messina C, Vita G (2003) *Neurology* 60:993-997.
- Shingara J, Keiger K, Shelton J, Laosinchai-Wolf W, Powers P, Conrad R, Brown D, Labourier E (2005) *RNA* 11:1461-1470.
- Baskerville S, Bartel DP (2005) *RNA* 11:241-247.

## Corrections

**MEDICAL SCIENCES.** For the article “Dynamic interplay between nitration and phosphorylation of tubulin cofactor B in the control of microtubule dynamics,” by Suresh K. Rayala, Emil Martin, Iraida G. Sharina, Poonam R. Molli, Xiaoping Wang, Raymond Jacobson, Ferid Murad, and Rakesh Kumar, which appeared in issue 49, December 4, 2007, of *Proc Natl Acad Sci USA* (104:19470–19475; first published November 28, 2007; 10.1073/pnas.0705149104), the authors note that, due to a printer’s error, Fig. 2 appeared incorrectly. This error does not affect the conclusions of the article. The corrected figure and its legend appear below.



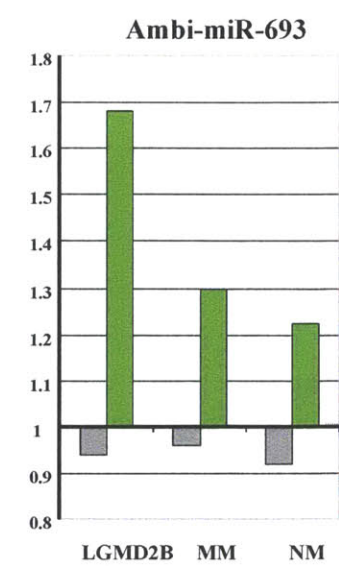
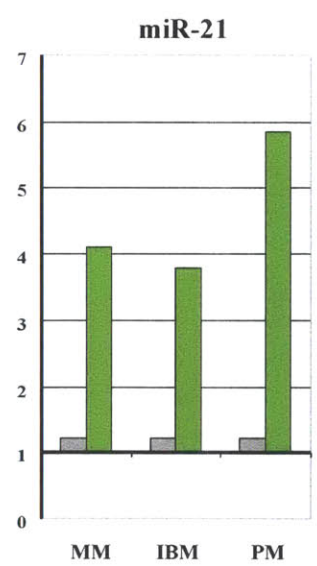
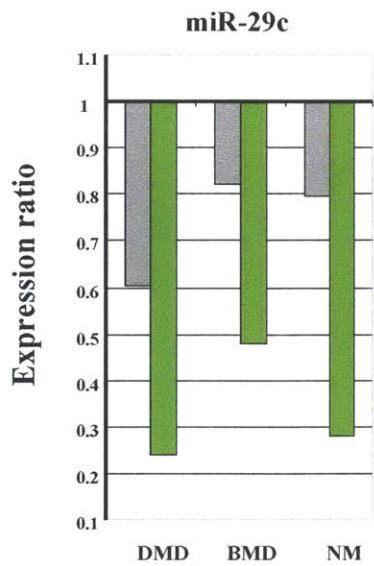
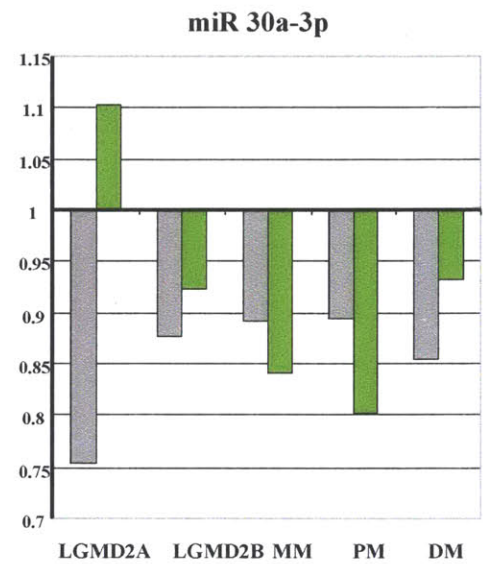
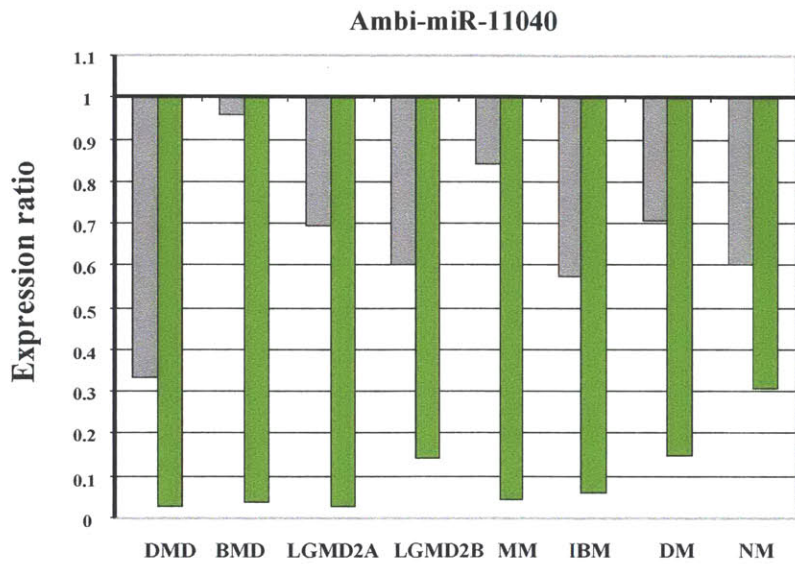
**Fig. 2.** Nitration of TCoB is iNOS-dependent. (A) Representative transmission images showing change in morphology of RAW 264.7 cells after LPS/ $\gamma$ -IFN treatment. (B) T7-TCoB was immunoprecipitated from transiently transfected RAW 264.7 cells that were treated with LPS (1  $\mu$ g/ml)/ $\gamma$ -IFN (50 units/ml) for 16 h, separated by SDS/PAGE, and immunoblotted with the indicated antibodies. (C) T7-TCoB was immunoprecipitated from transiently transfected RAW 264.7 cells that were treated with LPS/ $\gamma$ -IFN at various time points, separated by SDS/PAGE, and immunoblotted with the indicated antibodies. (D and E) T7-TCoB was immunoprecipitated from transiently transfected RAW 264.7 cells that were pretreated with either 1400W (100  $\mu$ M) or L-NAME (10 mM) for 4 h and/or LPS/ $\gamma$ -IFN, separated by SDS/PAGE, and immunoblotted with anti-nitrotyrosine and anti-T7 antibodies.

[www.pnas.org/cgi/doi/10.1073/pnas.0711492105](http://www.pnas.org/cgi/doi/10.1073/pnas.0711492105)

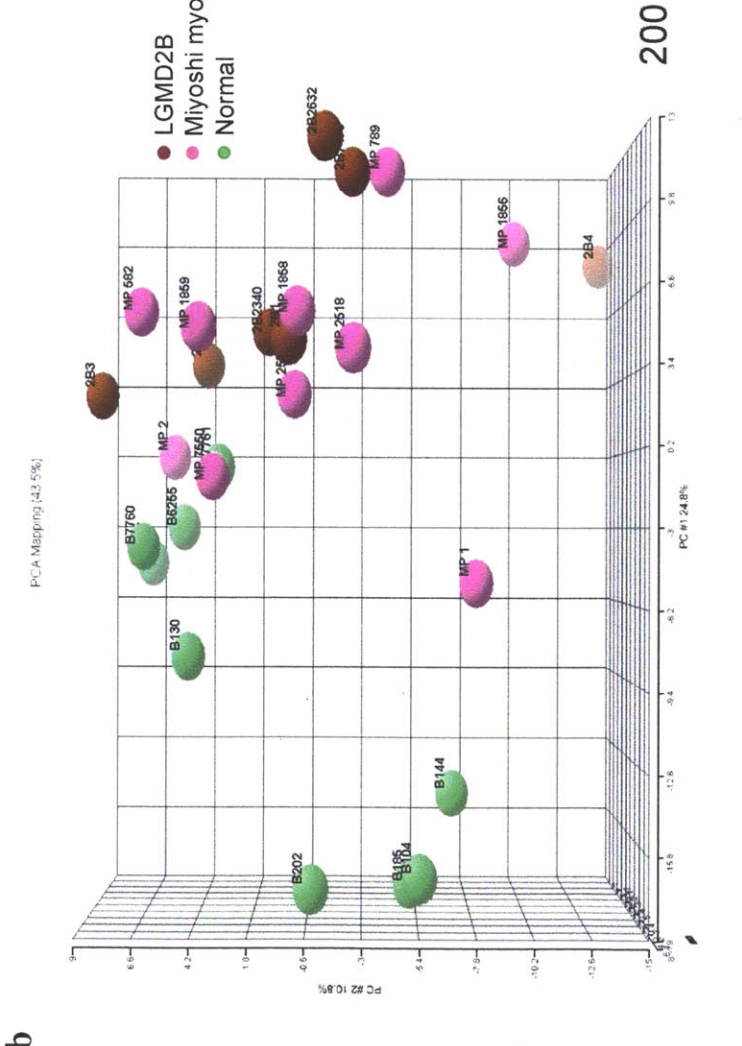
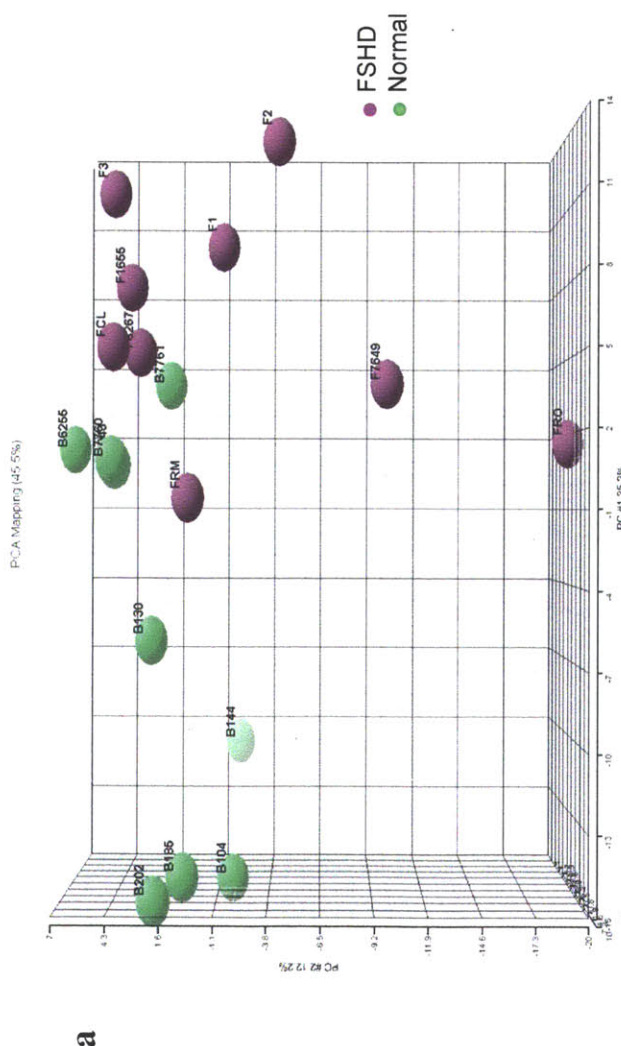
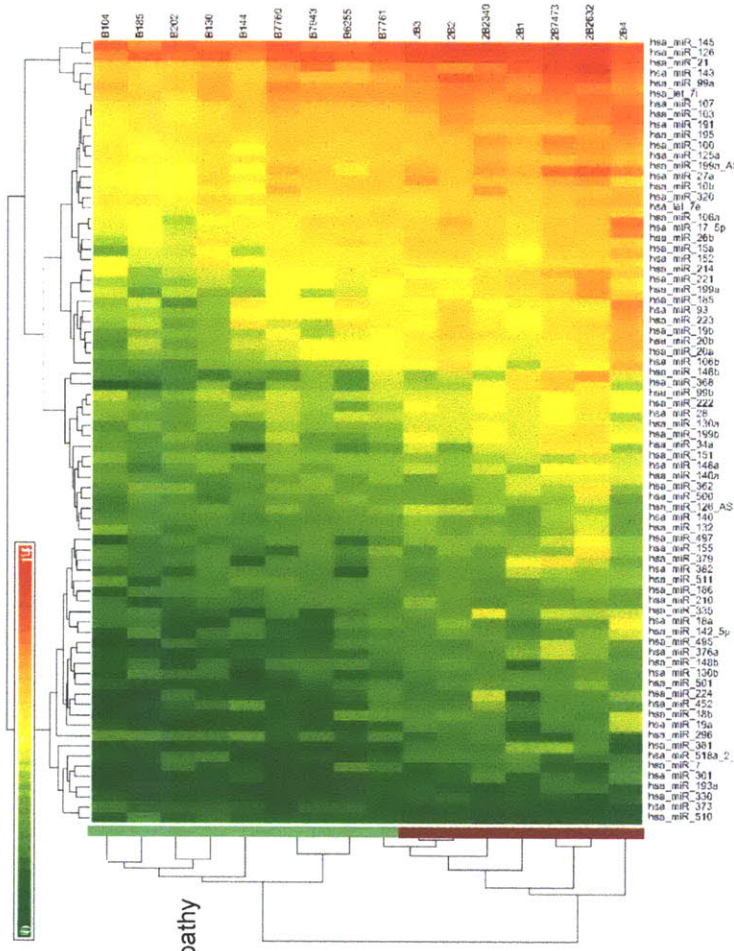
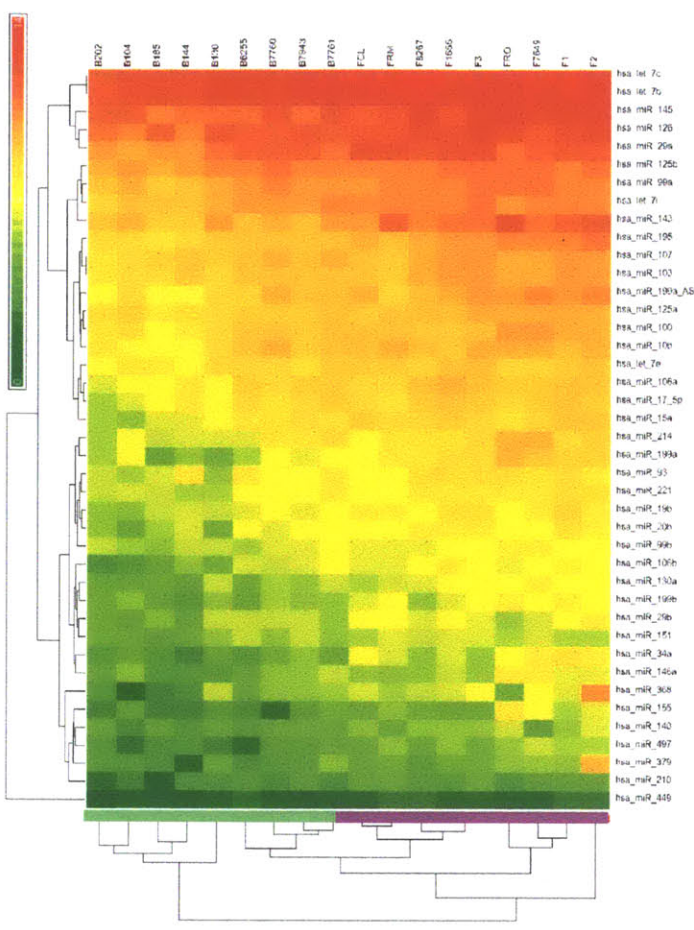
**GENETICS.** For the article “Distinctive patterns of microRNA expression in primary muscular disorders,” by Iris Eisenberg, Alal Eran, Ichizo Nishino, Maurizio Moggio, Costanza Lamperti, Anthony A. Amato, Hart G. Lidov, Peter B. Kang, Kathryn N. North, Stella Mitrani-Rosenbaum, Kevin M. Flanigan, Lori A. Neely, Duncan Whitney, Alan H. Beggs, Isaac S. Kohane, and Louis M. Kunkel, which appeared in issue 43, October 23, 2007, of *Proc Natl Acad Sci USA* (104:17016–17021; first published October 17, 2007; 10.1073/pnas.0708115104), the authors note that the affiliation information for authors Maurizio Moggio and Costanza Lamperti was incorrect in part. Their correct affiliation is “Unità Operativa di Neurologia, Centro Dino Ferrari, Università degli Studi di Milano, Istituto di Ricovero e Cura a Carattere Scientifico Fondazione Ospedale Maggiore, 20122 Milano, Italy.” The corrected affiliation line appears below.

<sup>a</sup>Howard Hughes Medical Institute, <sup>b</sup>Program in Genomics, Division of Genetics, <sup>c</sup>Informatics Program, and Departments of <sup>d</sup>Pathology and <sup>e</sup>Neurology, Children’s Hospital, Harvard Medical School, Boston, MA 02115; <sup>f</sup>Department of Neuromuscular Research, National Institute of Neuroscience, Tokyo 187-8502, Japan; <sup>g</sup>Unità Operativa di Neurologia, Centro Dino Ferrari, Università degli Studi di Milano, Istituto di Ricovero e Cura a Carattere Scientifico Fondazione Ospedale Maggiore, 20122 Milano, Italy; <sup>h</sup>Department of Neurology, Brigham and Women’s Hospital, Boston, MA 02115; <sup>i</sup>Institute for Neuromuscular Research, The Children’s Hospital at Westmead, New South Wales 2145, Australia; <sup>j</sup>Goldyne Savad Institute of Gene Therapy, Hadassah-Hebrew University Medical Center, Jerusalem 91240, Israel; <sup>k</sup>Department of Human Genetics, University of Utah, Salt Lake City, UT 84132; and <sup>l</sup>U.S. Genomics, Woburn, MA 01801

[www.pnas.org/cgi/doi/10.1073/pnas.0711290105](http://www.pnas.org/cgi/doi/10.1073/pnas.0711290105)

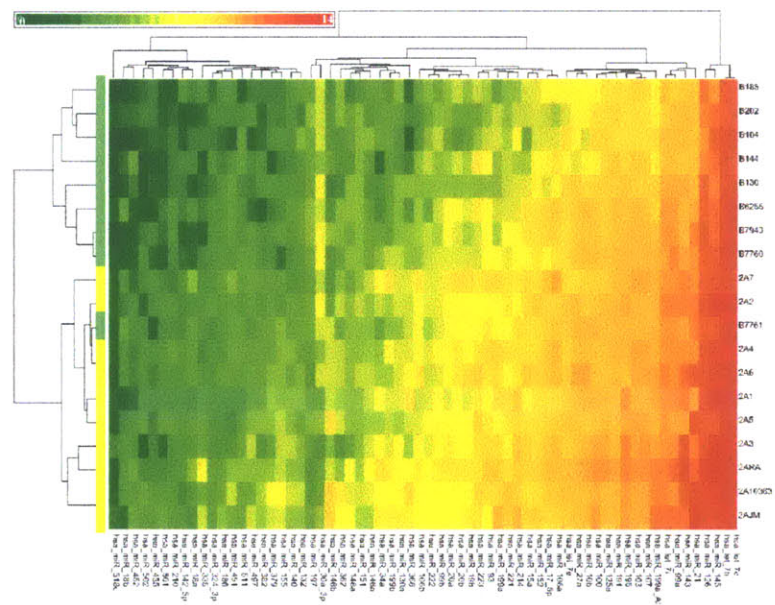
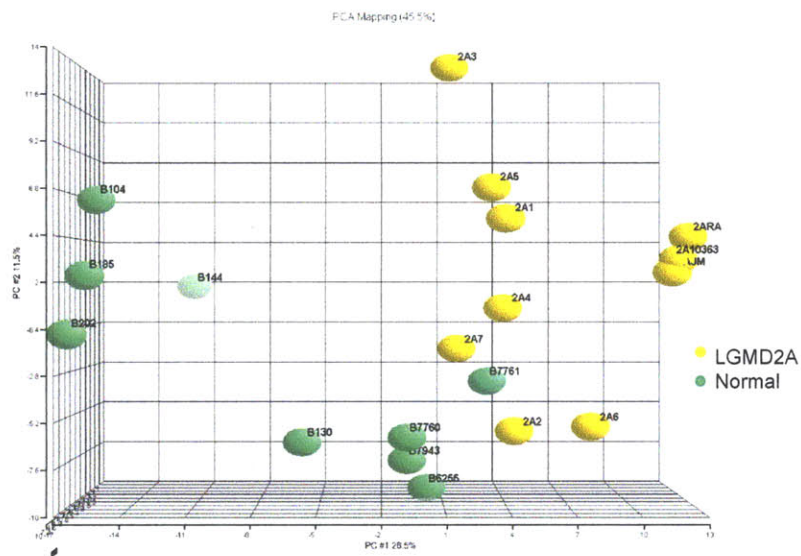


miRNA Microarray
  Direct Single molecule quantification assay

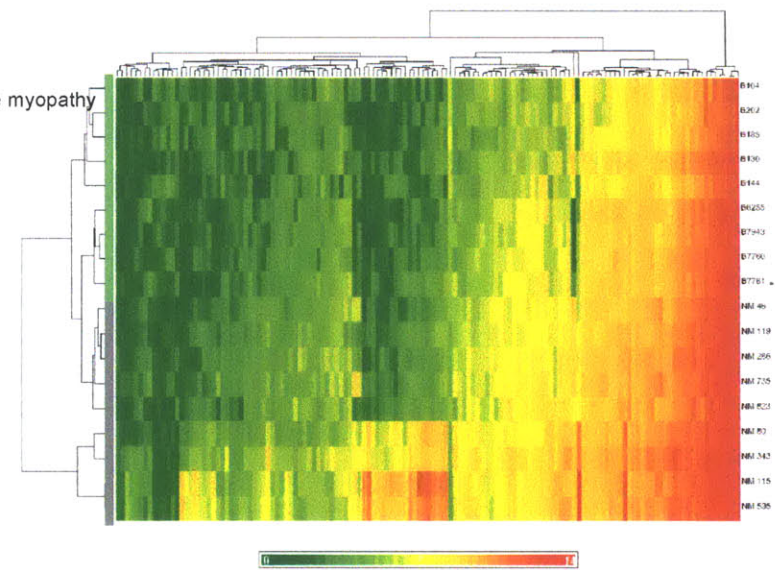
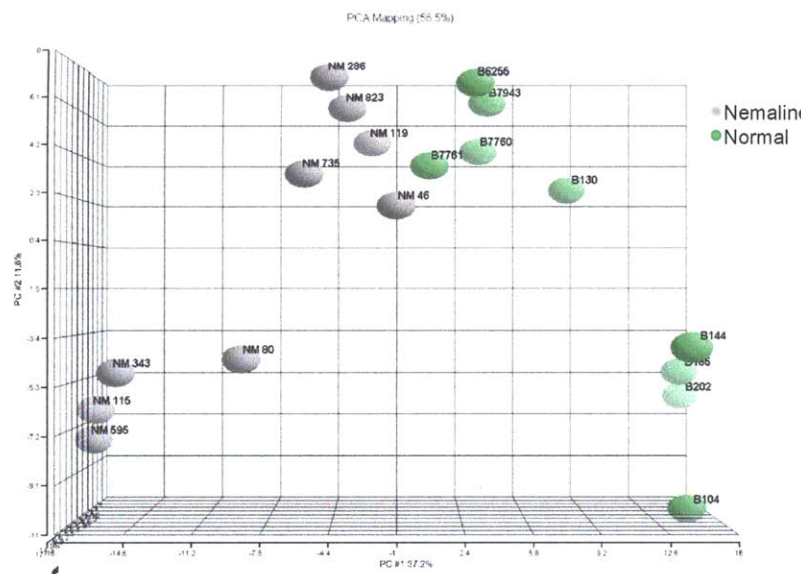




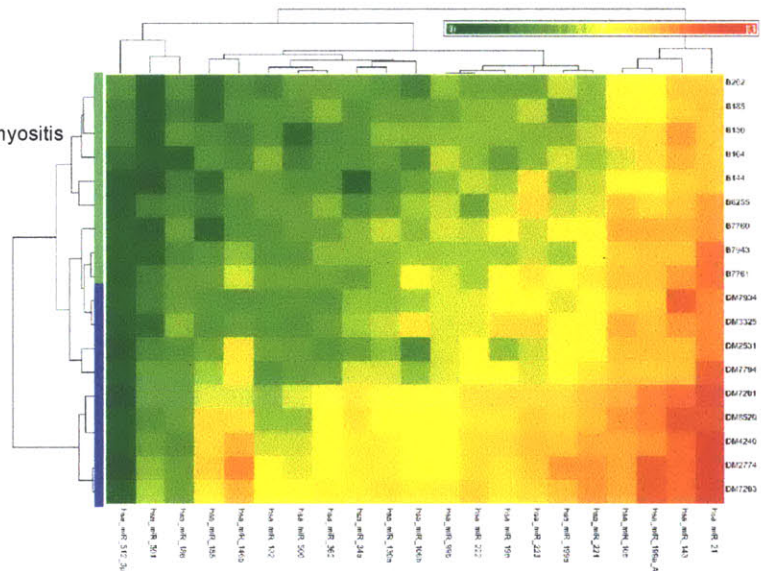
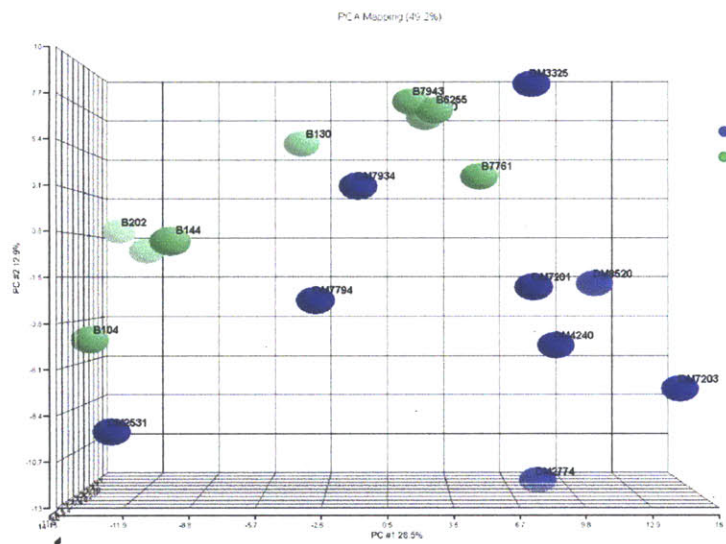
c

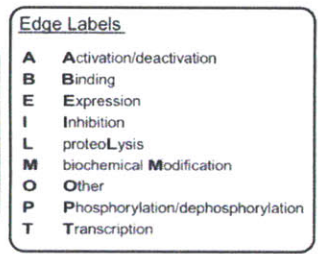
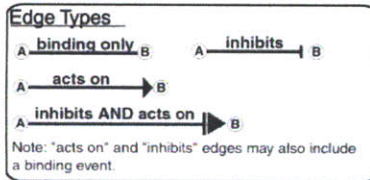
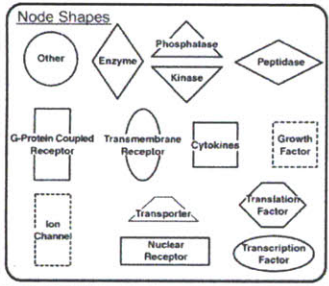
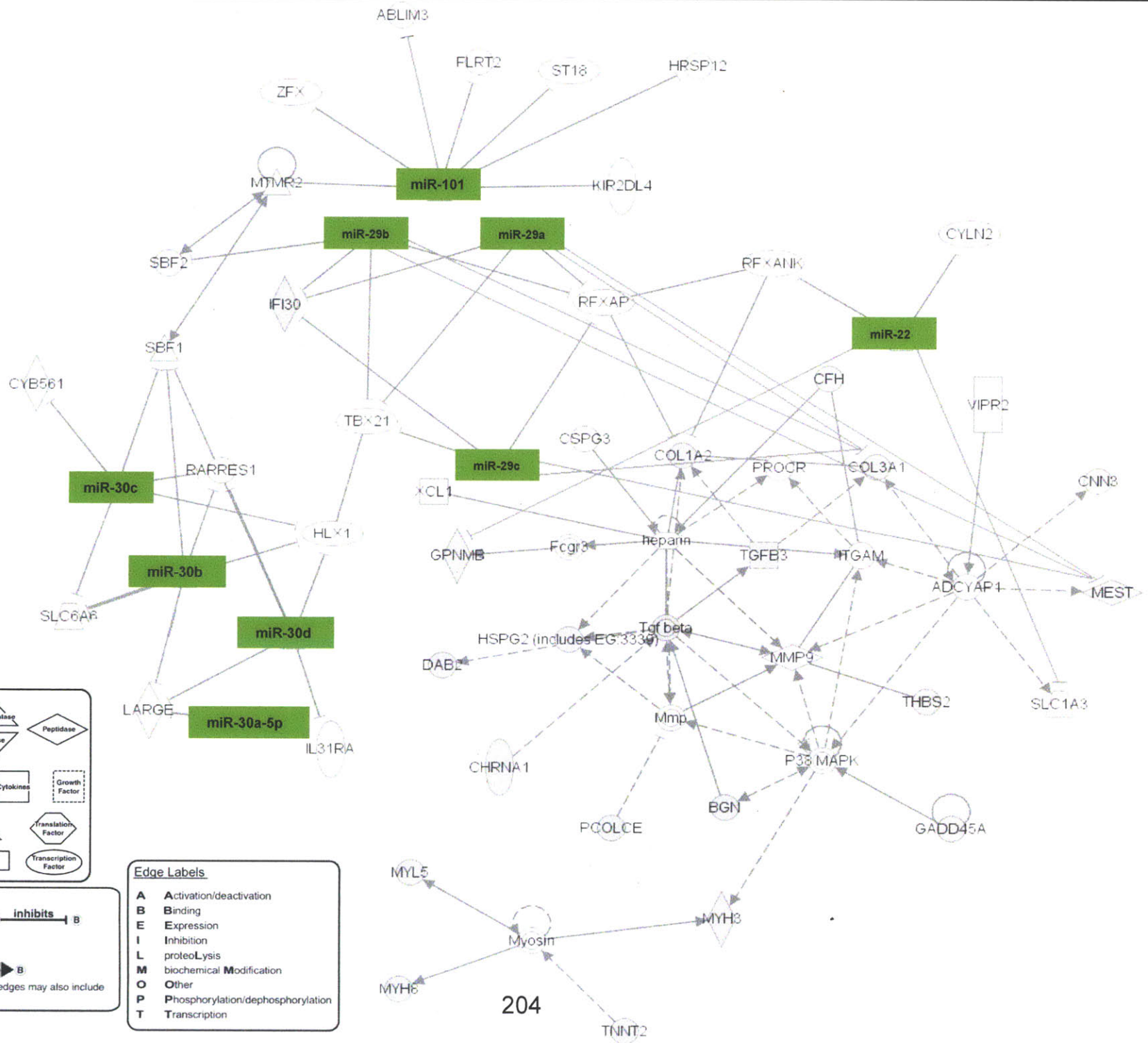


d







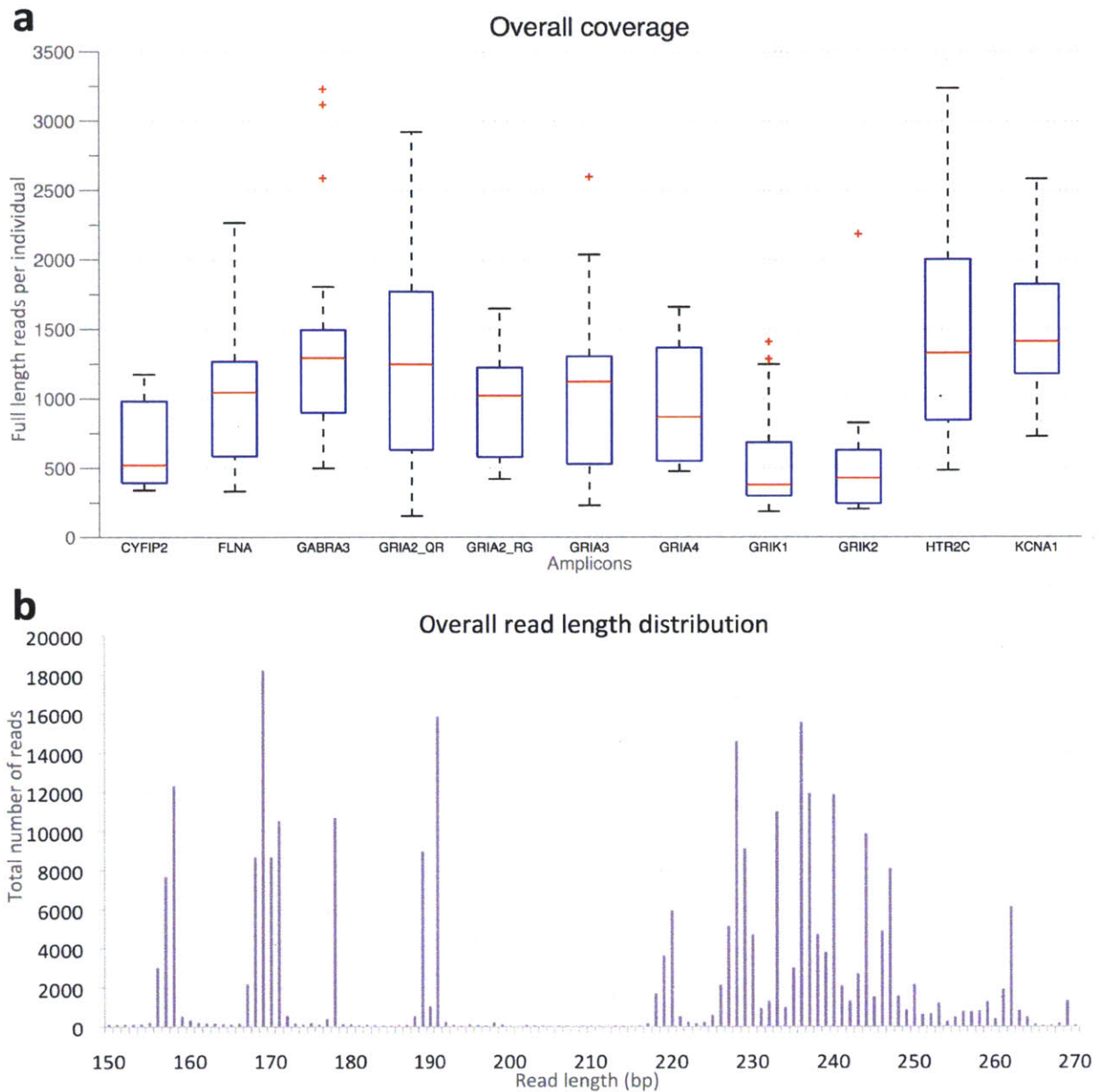


# **Appendix E:**

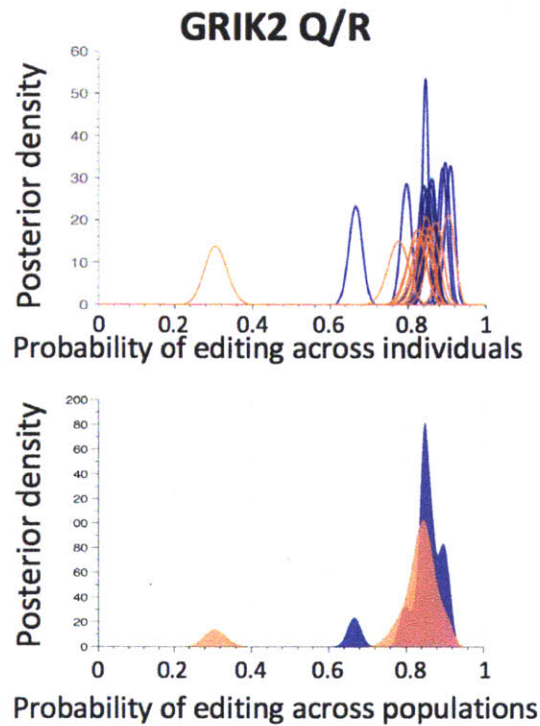
## **Supplementary Material for Chapter 2**

**Alal Eran, Jin Billy Li, Kayla Vatalaro, Jillian McCarthy, Fedik Rahimov, Christin Collins, Kyriacos Markianos, David M. Margulies, Emery N. Brown, Sarah E. Calvo, Isaac S. Kohane, Louis M. Kunkel**

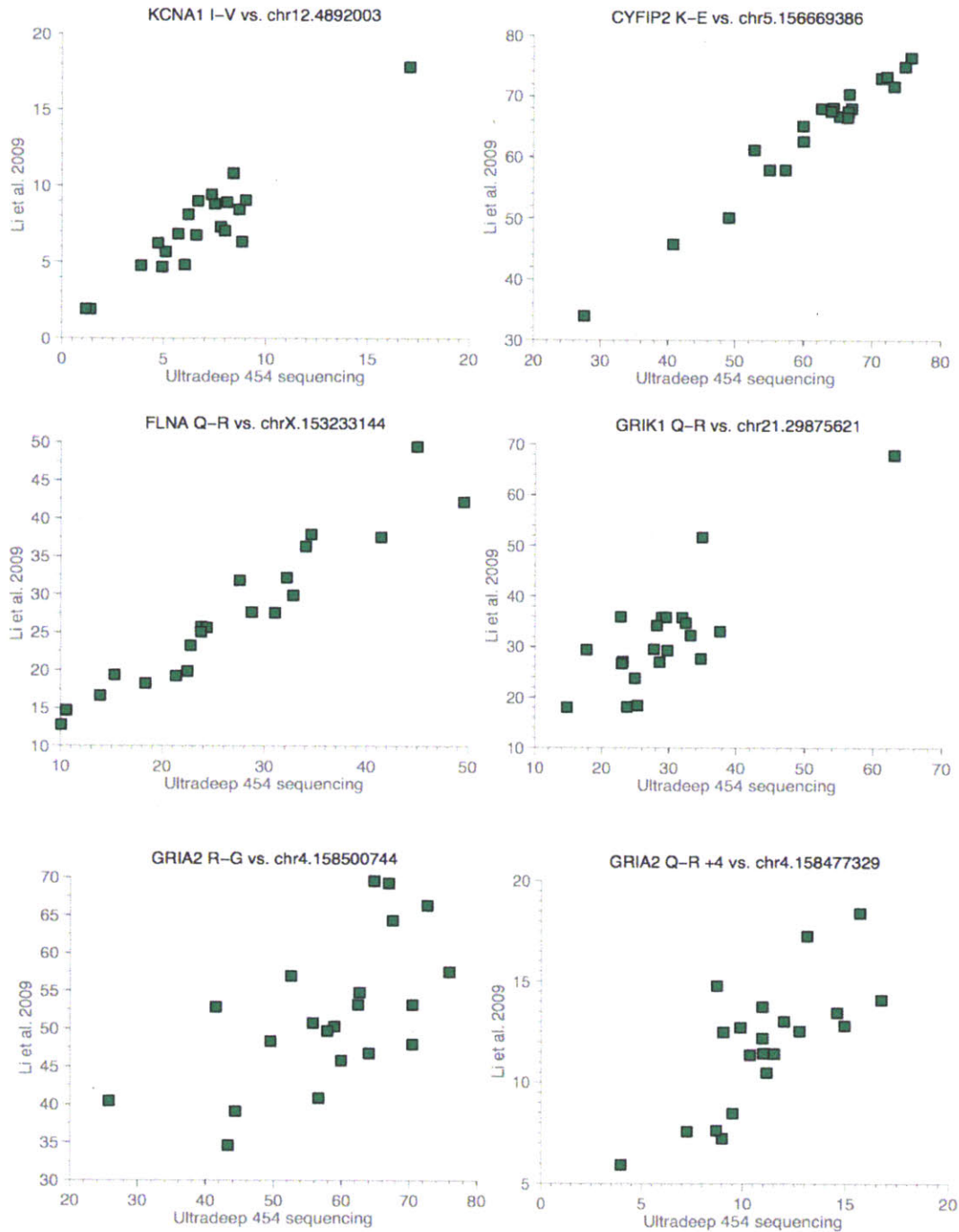
This appendix includes eight supplementary figures and nine supplementary tables



**Supplementary Figure 1** Data summary. **(a)** Overall coverage. Shown are boxplots of read counts per individual per amplicon. The mean coverage was 1344 reads per individual per amplicon, allowing for highly confident detection of RNA editing. **(b)** Overall read length distribution. Bidirectional full length reads were obtained for each amplicon, enabling the investigation of relationships across neighboring sites and between editing and splicing. Shown is the distribution of all reads in all amplicons. See **Supplementary Table 3** for amplicon lengths.

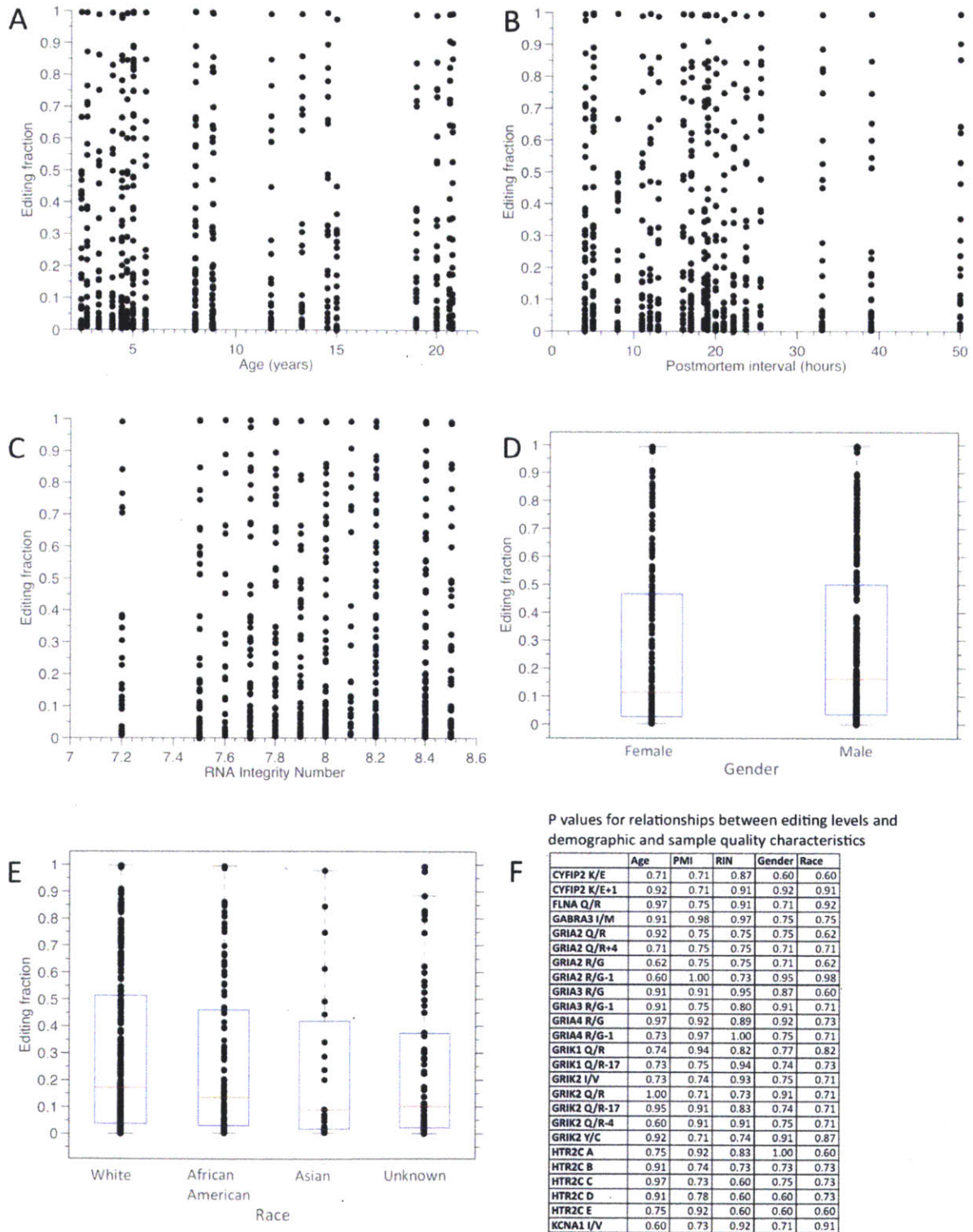


**Supplementary Figure 2** The data analysis approach depicted for one site, *GRIK2* Q/R. Editing fractions were modeled by a beta-binomial distribution, resulting in a posterior editing fraction density for each individual at each site, based on the data. On the top, a posterior editing density is shown for each human subject at *GRIK2* Q/R. The editing densities of individuals with ASD are shown in orange and neurotypical individuals in blue. The peak of each density represents a point estimate for the level of editing and its width depicts the confidence around it. A large range of editing levels is seen among individuals, from 30 to 90%. On the bottom is the corresponding population-level perspective, where individuals are grouped by their affected status and comparisons are made between groups.



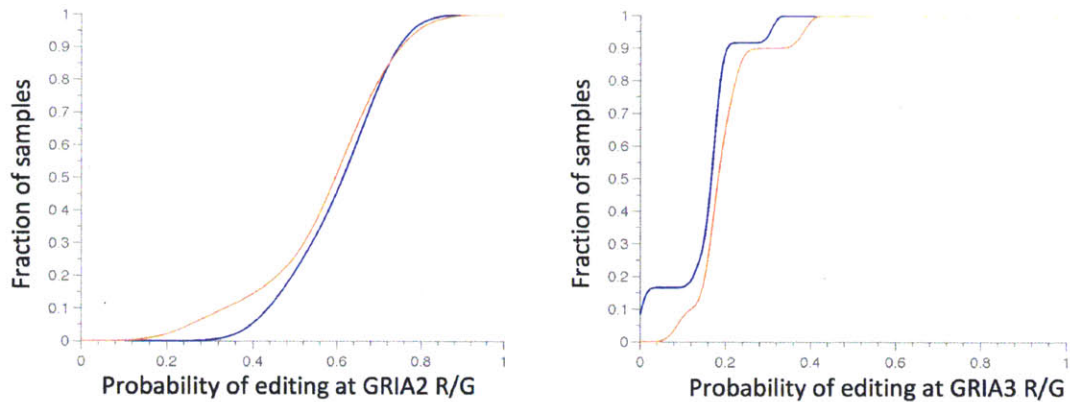
**Supplementary Figure 3** Correlation between two A-to-I editing detection methods: ultradeep 454 sequencing of PCR-selected regions, and Illumina sequencing of Padlock-captured fragments. The editing levels of six sites were independently assayed in all individuals using both detection methods. For each site, a tight correlation was observed between the results of the two methods, as detailed in **Supplementary Table 5**. Each panel depicts the correlation of editing at a single site. Overall, the mean correlation coefficient between the two methods was 0.923 (mean  $p < 2e-3$ , Pearson's correlation).



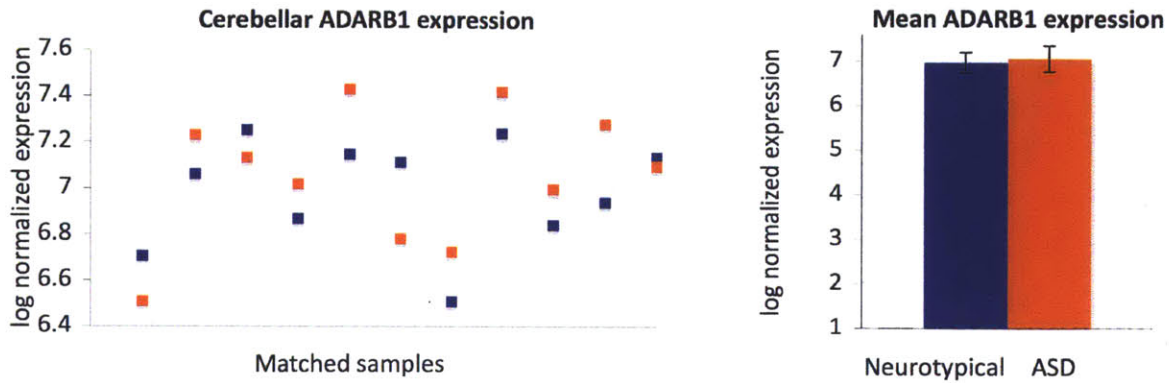


**Supplementary Figure 4** Relationships between editing levels and (A) age, (B) postmortem interval (PMI), (C) RNA integrity number (RIN), (D) gender, and (E) race. Each marker represents

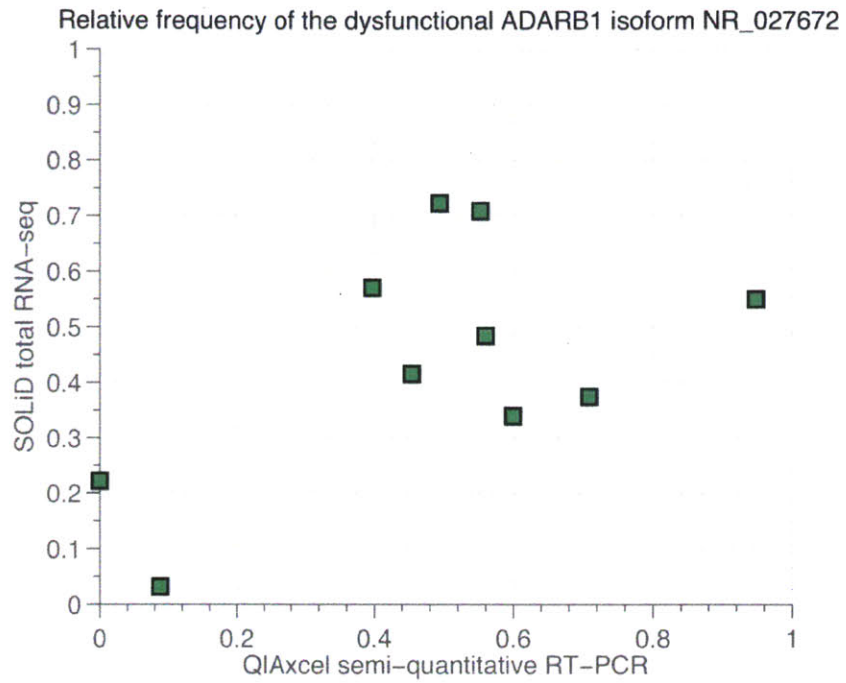
a point estimate of the degree of editing of one individual at one site, plotted against continuous characteristics in A-C, with boxplots in D and E summarizing discrete characteristics. The editing levels of all individuals at all sites were linearly regressed against continuous variables (age, PMI, and RIN), and compared between discrete groups (gender and race). No linear relationships were identified between editing fractions and age, PMI, or RIN using F tests, and no significant differences were detected across genders or races using Kruskal-Wallis tests. Benjamini-Hochberg corrected<sup>58</sup> p-values are shown in (F) for each relationship at each site.



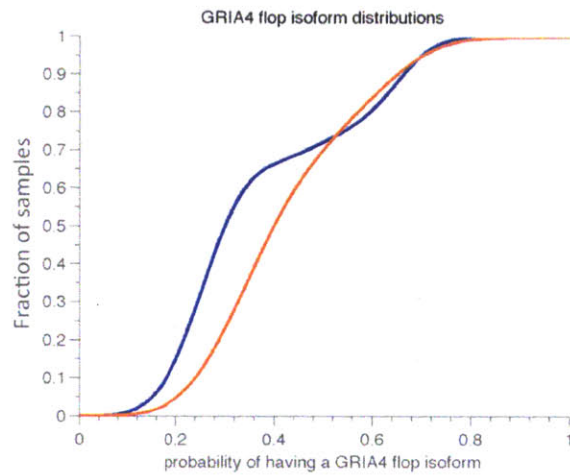
**Supplementary Figure 5** Kernel-smoothed cumulative distribution functions (ksCDFs) of *GRIA2* R/G (left) and *GRIA3* R/G editing (right) in individuals with ASD (orange) and neurotypical individuals (blue). Although the CDFs of editing in individuals with ASD and editing in neurotypical individuals are distinct, the small sample size of this study is underpowered to detect such effect sizes.



**Supplementary Figure 6** *ADARB1* expression in ASD and neurotypical individuals. No differences in *ADARB1* expression were found between individuals with ASD and best-matched neurotypical individuals. Gene core expression was measured on the Affymetrix Exon 1.0 arrays in 22 of 25 samples. On the left is the log normalized expression signal between matched samples, where individuals with ASD are shown in orange and neurotypical individuals in blue. On the right is the mean *ADARB1* expression per group, depicting a similar mean expression level between individuals with ASD and neurotypical individuals.



**Supplementary Figure 7** Correlation between relative *ADARB1* isoform detection by RNA-seq and semi-quantitative RT-PCR. The fraction of *ADARB1* transcripts that include the double-stranded RNA binding domain was measured by both RNA-seq (y-axis) and semi-quantitative RT-PCR (x-axis) in ten samples.



**Supplementary Figure 8** Differences in *GRIA4* isoform usage between individuals with ASD and neurotypical individuals. Shown are kernel-smoothed cumulative distribution functions (ksCDFs) of *GRIA4 flop* isoform usage in ASD (orange) and matched neurotypical individuals (blue). The *flop* isoform usage is higher in ASD, (and *flip* usage is respectively lower) ( $p=3.58e-2$ , KS test). *GRIA4* R/G editing is tightly associated with the *flop* isoform (OR=101.9 [95% CI 87.8-118.2]), compare this figure to Figure 7.

**Supplementary Table 1** Sample info. Matched neurotypical individuals and individuals with ASD share a barcode number (e.g. 1185 was matched to 1349).

No.	Sample ID	Source <sup>a</sup>	Barcode	Diagnosis	Age		Sex	Race	PMI <sup>b</sup>	Storage			RNA integrity RIN <sup>c</sup>	Other medical issues of note (besides autism)
					years	days				hours	years	days		
1	<b>1185</b>	NICHD BTB	1CTRL	Neurotypical	4	258	Male	White	17	6	39	8.5	None specified	
2	<b>1284</b>	NICHD BTB	9CTRL	Neurotypical	3	123	Female	African American	11	5	185	8.4	None specified	
3	<b>1349</b>	NICHD BTB	1ASD	Autism	5	220	Male	White	39	5	31	7.5	ADD with hyperactivity	
4	<b>1407</b>	NICHD BTB	12CTRL	Neurotypical	9	46	Female	African American	20	5	33	8.6	None specified	
5	<b>1499</b>	NICHD BTB	12ASD <sup>d</sup>	Neurotypical	4	170	Female	Asian	21	5	204	8.2	None specified	
6	<b>1541</b>	NICHD BTB	10CTRL	Neurotypical	20	228	Female	White	19	4	142	8.1	None specified	
7	<b>1638</b>	NICHD BTB	10ASD	Autism	20	277	Female	White	50	3	272	8.4	ADD, hyperactivity, epilepsy, irregular response to pain, schizophrenia, hearing problems, sleeping difficulty, abnormal gait, microencephaly	
8	<b>1670</b>	NICHD BTB	6CTRL	Neurotypical	13	99	Male	White	5	3	256	8	None specified	
9	<b>1706</b>	NICHD BTB	12CTRL	Neurotypical	8	214	Female	African American	20	3	231	8.2	None specified	
10	<b>1793</b>	NICHD BTB	2CTRL	Neurotypical	11	270	Male	African American	19	3	40	8	None specified	
11	<b>1860</b>	NICHD BTB	5CTRL	Neurotypical	8	2	Male	White	5	2	268	7.6	None specified	

No.	Sample ID	Source <sup>a</sup>	Barcode	Diagnosis	Age		Sex	Race	PMI <sup>b</sup>	Storage			RNA integrity RIN <sup>c</sup>	Other medical issues of note (besides autism)
					years	days				hours	years	days		
12	1864	NICHD BTB	8CTRL	Neurotypical	2	178	Female	White	8	2	250	7.9	None specified	
13	4231	NICHD BTB	2ASD	Autism	8	300	Male	African American	12	1	299	7.9	None specified	
14	4671	NICHD BTB	9ASD	Autism	4	165	Female	African American	13	0	356	8.5	None specified	
15	4721	NICHD BTB	3ASD	Autism	8	304	Male	African American	16	0	324	8.1	None specified	
16	4722	NICHD BTB	7CTRL	Neurotypical	14	198	Male	White	16	1	59	7.8	None specified	
17	4787	NICHD BTB	3CTRL	Neurotypical	12	318	Male	African American	15	0	120	8.6	None specified	
18	5144	HBTRC	4ASD	Autism	20		Male	White	23.7	6	193	7.8	Vitiligo, sleeping difficulties per parent	
19	5251	HBTRC	4CTRL	Neurotypical	19		Male	White	18.6	4	95	7.2	None specified	
20	5569	HBTRC	5ASD	Autism	5		Male	White	25.5	4	344	7.7	Allergies, hyperlexia, no/delayed reaction to heat, never felt pain, sleeping difficulty	
21	5666	HBTRC	6ASD	Autism	8		Male	White	22.2	4	230	7.5	Epilepsy, syndactyly, negative for Timothy syndrome testing	
22	6399	HBTRC	8ASD	Autism	2	278	Male	White	4	2	130	8.2	None specified	
23	6736	HBTRC	11CTRL	Neurotypical	4		Female	Unknown	17.02	1	49	8	None specified	
24	7002	HBTRC	11ASD	Autism	5		Female	Unknown	33	0	77	8.2	High tolerance to pain, vision problems	

No.	Sample ID	Source <sup>a</sup>	Barcode	Diagnosis	Age		Sex	Race	PMI <sup>b</sup>	Storage		RNA integrity RIN <sup>c</sup>	Other medical issues of note (besides autism)
					years	days				hours	years		
25	<b>7079</b>	HBTRC	7ASD	Autism	15		Male	Unknown	4			7.7	"lazy eye" Asperger Syndrome / high functioning autism

<sup>a</sup> NICHD BTB, National Institute of Child Health and Human Development Brain and Tissue Bank; HBTRC, Harvard Brain Tissue Resource Center.

<sup>b</sup> PMI, post-mortem interval.

<sup>c</sup> RIN, RNA integrity number as determined by the Agilent 2100 Bioanalyzer.

<sup>d</sup> Neurotypical sample sequenced in the ASD pool.



**Supplementary Table 2** Synaptic genes examined in this initial study. This study included all synaptic genes shown to undergo neurodevelopmentally-regulated, amino-acid altering A-to-I editing, with known functional consequences.

	Gene	Product	Genomic location (hg19)	Gene function at the synapse	Known editing sites	Functional impact of editing	Predominant editing enzyme	Refs
1	<i>CYFIP2</i>	Cytoplasmic FMR1-interacting protein 2	chr5: 156693091 - 156822604	A p53-inducible protein enriched in synaptosomes, implicated in synaptic maintenance	K/E	Potentially altering clathrin heavy chain binding	ADARB1	1-3
2	<i>FLNA</i>	Filamin A, alpha	chrX: 153576900 - 153603006	A widely expressed cytoskeletal protein that interacts with integrins, transmembrane receptor complexes, and second messengers to regulate the dynamic actin cytoskeleton, which is the major structural component of synapses.	Q/R	Might alter binding to glutamate receptor type 7, calcitonin receptor, androgen receptor, SEK-1, BRCA-2, Smad, caveolin-1, and integrin.	ADARB1	3, 4
3	<i>GABRA3</i>	Gamma-aminobutyric acid (GABA) A receptor, alpha 3	chrX: 151335634 - 151619831	A member of the GABA-A receptor gene family of heteromeric pentameric ligand-gated ion channels through which GABA, the major inhibitory neurotransmitter in the	I/M	Slows the activation of alpha3 subunit-containing GABAA receptors by altering the voltage-dependent conductance of the	ADAR+ADARB1	5

				mammalian brain, acts.		alpha3 subunit.		
4	<i>GRIA2</i>	Glutamate receptor, ionotropic, AMPA 2	chr4: 158141736 - 158287224	An inotropic glutamate receptor subunit that makes it impermeable to calcium (via RNA editing), sensitive to alpha-amino-3-hydroxy-5-methyl-4-isoxazolpropionate (AMPA). Mediates fast excitatory postsynaptic currents in neurons of the central nervous system.	Q/R <sup>a</sup> , R/G	Q/R editing results in calcium impermeability, and regulates the receptor's tetramerization and intracellular trafficking; R/G editing leads to faster desensitization recovery.	Q/R: ADARB1 R/G:ADARB1 +ADAR	6-8
5	<i>GRIA3</i>	Glutamate receptor, ionotropic, AMPA 3	chrX: 122318096 - 122624760	An AMPA receptor subunit, mediating most of the excitatory neurotransmission in the mammalian brain, and participating in synaptic plasticity and efficacy in learning and memory.	R/G	Leads to faster desensitization recovery of GRIA3-containing ionotropic glutamate receptors	ADAR	7
6	<i>GRIA4</i>	Glutamate receptor, ionotropic, AMPA 4	chr11: 105480800 - 105852819	An AMPA receptor subunit, mediating most of the excitatory neurotransmission in the mammalian brain, and participating in synaptic	R/G	Leads to faster desensitization recovery of GRIA4-containing ionotropic glutamate receptors	ADARB1	7

6	<i>GRIA4</i>	Glutamate receptor, ionotropic, AMPA 4	chr11: 105480800 - 105852819	An AMPA receptor subunit, mediating most of the excitatory neurotransmission in the mammalian brain, and participating in synaptic plasticity and efficacy in learning and memory.	R/G	Leads to faster desensitization recovery of GRIA4-containing ionotropic glutamate receptors	ADARB1	7
7	<i>GRIK1</i>	Glutamate receptor, ionotropic, kainate 1	chr21: 30925866 - 31312282	An inotropic glutamate receptor subunit sensitive to kainate that mediates fast excitatory neurotransmission.	Q/R	Reduces the calcium permeability of GRIK1-containing glutamate receptors	ADARB1+ ADAR	8, 9
8	<i>GRIK2</i>	Glutamate receptor, ionotropic, kainate 2	chr6: 101846905 - 102517957	An inotropic glutamate receptor subunit sensitive to kainate that mediates fast excitatory neurotransmission. Variants in <i>GRIK2</i> have been repeatedly linked and associated with autism.	Q/R, I/V, Y/C	Reduce the calcium permeability of GRIK2-containing glutamate receptors	ADARB1	9, 10
9	<i>HTR2C</i>	5-hydroxytryptamine (serotonin) receptor 2C	chrX: 113818551 - 114144624	A G protein-coupled receptor that stimulates phospholipase C mediated hydrolysis of phosphatidylinositol bisphosphate, leading to the mobilization of intracellular calcium and the activation of protein kinase C. Altered editing of its mRNA has been observed in a mouse model of autism and in disorders that	A,B,C, D,E	Reduce G-protein coupling functions by as much as 20-fold (depending on combinatorial editing) and decrease constitutive activity	A,B: ADAR C,E: ADARB1+ ADAR D: ADARB1	6, 8

10	<i>KCNA1</i>	Potassium voltage-gated channel, shaker-related subfamily, member 1 (episodic ataxia with myokymia)	chr12: 5019073 - 5027420	Mediates the voltage-dependent potassium ion permeability of excitable membranes.	I/V	Leads to rapid recovery from fast inactivation at negative potentials	ADARB1	11
----	--------------	-----------------------------------------------------------------------------------------------------	-----------------------------------	-----------------------------------------------------------------------------------	-----	-----------------------------------------------------------------------	--------	----

Editing at all sites except *GRIA2* Q/R was shown to be neurodevelopmentally regulated by Wahlstedt et al.<sup>12</sup>

<sup>a</sup> *GRIA2* Q/R editing is essential and unchanged throughout neurodevelopment<sup>12</sup>.

**Supplementary Table 3** PCR primers and conditions used for 454 library preparation.

<b>Amplicon</b>	<b>Forward primer<sup>a</sup></b>	<b>Reverse primer<sup>a</sup></b>	<b>Amplicon description</b>	<b>Annealing (°C)</b>	<b>PCR1 size (bp)</b>	<b>PCR2 size<sup>b</sup> (bp)</b>
<b>CYFIP2</b>	<i>tcgatcagcaggtgatgggctttggcctctac</i>	<i>tacgatgcgtctgatgctgctctgggtgca</i>	For the detection of the K/E editing site (exon 11). The forward primer (F) spans the 9-10 splice junction, and the reverse primer (R) is in exon 12	60	245	303
<b>FLNA</b>	<i>tcgatcagcagttcaacgaggaacacattcccagac</i>	<i>tacgatgcgtgctgtgcaccttggcatcga</i>	For the detection of the Q/R editing site (exon 43). F in exon 43, R in 44 of both Refseq isoforms	63	201	259
<b>GABRA3</b>	<i>tcgatcagcagctcaacagagagtctgttctctgc</i>	<i>tacgatgcgtagcccaactccgcttgggtg</i>	For the detection of the I/M editing site (exon 9). F in exon 8, R in 9	Touch-down: 68 - 0.5/cycle	231	289
<b>GRIA2 Q/R</b>	<i>tcgatcagcatggtcagcagatttagccctacga</i>	<i>tacgatgcgtacacacctccaacaatgcgcc</i>	For the detection of the Q/R and its +4 sites. F is in exon 11, R in exon 12	63	209	267
<b>GRIA2 R/G + flip/flop</b>	<i>tcgatcagcatccaaaggctatggcatcgaac</i>	<i>tacgatgcgttcaaagccaccagcattgcca</i>	For detection of the R/G site and the entire flip/flop module. F is in exon 13, R in exon 16. Flip variant is exon 15, flop is exon 14	68	270	328
<b>GRIA3</b>	<i>tcgatcagcattccaaaggctatgggtggcaacc</i>	<i>tacgatgcgtcctgccacattgctcaggctcaga</i>	Intended to detect R/G	65	218	276

<b>Amplicon</b>	<b>Forward primer<sup>a</sup></b>	<b>Reverse primer<sup>a</sup></b>	<b>Amplicon description</b>	<b>Annealing (°C)</b>	<b>PCR1 size (bp)</b>	<b>PCR2 size<sup>b</sup> (bp)</b>
<b>R/G + flip/flop</b>			editing and flip/flop isoforms that follow downstream. F is in exon 13, product spans exons 14/15 - flop/flip - and R is in exon 16.			
<b>GRIA4 R/G + flip/flop</b>	<i>tcgatcagcaggctatggagtagcaacgcc</i>	<i>tacgatgcgtgctacattgctcaggctcaaggca</i>	For detection of the R/G site and flip/flop alternative splicing. Forward is in exon 15, reverse in exon 18a/b, depending on the isoform (1/2). Flop is exon 16, flip exon 17	63	208	266
<b>GRIK1</b>	<i>tcgatcagcacatgcaaccctgactcagacgtg</i>	<i>tacgatgcgtggtcgatagagctttgggcatcag</i>	For the detection of the Q/R editing site (exon 13). F in exon 13, R in exon 14 of both Refseq isoforms	68	139	197
<b>GRIK2</b>	<i>tcgatcagcaagcccaatggtacaaaccagg</i>	<i>tacgatgcgttgctttgggcatgagctcagaac</i>	Detects the 3 editing sites in <i>GRIK2</i> (gluR6): I/V, Y/C (both in TM1 - exon 11), and Q/R (TM2 - exon 12). Forward primer is in exon 11, reverse in exon 13 of both Refseq isoforms.	63	280	338

Amplicon	Forward primer <sup>a</sup>	Reverse primer <sup>a</sup>	Amplicon description	Annealing (°C)	PCR1 size (bp)	PCR2 size <sup>b</sup> (bp)
<i>HTR2C</i>	<i>tcgatcagcaaacagcgtccatcatgcacctctg</i>	<i>tacgatgcgttttgggtcgttgagcacgcacg</i>	For the detection of A,B,C,D,E editing. F is in exon 5a, R in 6. If there is no editing / reduced editing then 5b will likely be excluded yielding a 150bp product. Intended to always have a PCR product regardless of the splice isoform.	Touch-down: 75- 0.5/cycle	150 + 245	208 + 303
<i>KCNA1</i>	<i>tcgatcagcaagatcgtgggctccttgtgtgc</i>	<i>tacgatgcgtcgtggagcaactgagcctgct</i>	For the detection of the I/V editing site (exon 2). F and R both in exon 2 as this is an intronless gene.	60	152	210
PCR2	<b>GCCTCCCTCGCGCCATCAG</b> <10bpBarcode> <i>tcgatcagca</i>	<b>GCCTTGCCAGCCCGCTCAG</b> <10bpBarcode> <i>tacgatgcgt</i>	Sample-tagging PCR that adds the barcode and 454 primers to each PCR1 product by priming off the universal tag	50		

<sup>a</sup> In italics is the universal tag added to enable sample-specific priming in PCR2.

<sup>b</sup> PCR2 adds 29\*2= 58bp to each PCR1 product.

**Supplementary Table 4** Padlock probe sequences used for validation.

Site	cDNA probe	gDNA probe
chr12.4892003	ggacacaatgacaggtacgggcagggcaattgtagcacaccagcgatggc	ggacacaatgacaggtacgggcagggcaattgtagcacaccagcgatggc
chrX.153233144	gctggtgaccttagccctgactcctgaaggctagaaacagtgaggc	gggcggtttctctcggtgcctcacctgaaggctagaaacagtgaggc
	gatagagctttgggcatcagctctgatccttgctgcatgagagctccaactcaa	
chr21.29875621	accag	tgtgacaaagataggcaaccggtgtaccttgctgcatgagagctccaactcaaacca
chr4.158477329	aaccttggcgaatatcgcatccttgctgcataaaggcacccaaggaaaac	aaccttggcgaatatcgcatccttgctgcataaaggcacccaaggaaaac
	tcactgagtttcaatactgcaagatttactggggttcttaatgaggatccttagg	atactataacaacatttagcatattgttatactattccaccaccttaatgaggatccttta
chr4.158500744	tgttgcat	ggtgttgc
chr5.156669386	actgttctctcatagtgagcactggcttaatgtatctggccagctctat	actgttctctcatagtgagcactggcttaatgtatctggccagctctat



**Supplementary Table 5** Correlation between two A-to-I editing detection methods.

Ultradeep 454 sequencing of PCR-selected ~212bp flanking the editing sites		Illumina sequencing of padlock-captured ~50bp flanking the editing sites	
Site ID	Site ID	Pearson's r	Pearson's p
<i>KCNA1</i> I/V	chr12.4892003	0.978323	3.63E-08
<i>FLNA</i> Q/R	chrX.153233144	0.968744	2.22E-07
<i>GRIK1</i> Q/R	chr21.29875621	0.965429	3.67E-07
<i>GRIA2</i> Q/R+4	chr4.158477329	0.957655	9.98E-07
<i>GRIA2</i> R/G	chr4.158500744	0.958856	8.66E-07
<i>CYFIP2</i> K/E	chr5.156669386	0.711918	9.40E-03
		<b>Mean</b>	<b>0.9234875</b>
			<b>0.001567082</b>

The linear correlation coefficients between the editing levels determined by two independent A-to-I editing detection methods were calculated, along with their p-values. Six sites were assayed independently by ultradeep 454 sequencing of PCR-selected fragments with an average length of 212bp, and by Illumina sequencing of padlock-captured fragments with an average length of 50bp. A tight correlation was observed between the two methods across all sites. See Supplementary Figure 4 for visualization. Each method has unique advantages and disadvantages. For example, only the 454-based method can detect medium-range relationships in *cis*, including those between editing sites, and between splicing and editing. The long 454 reads also allow confident alignments of closely spaced editing sites to the reference sequence, such as those in *HTR2C*. By contrast, the padlock-based capture is much more scalable and less labor intensive than the PCR-based selection, and Illumina sequencing currently yields higher coverage for a given price than 454 sequencing.

**Supplementary Table 6** Primers and probe sequences used for TaqMan® quantitation of the dysfunctional *ADARB1* isoform (NR\_027672)

Forward primer	Reverse primer	Reporter probe
TCGGTCAGGTCACCAAACCTTAC	CCAGTCAAGAAACCCTCAAAGTATTTT	TCAGCTAAAACTCATGTTTTTC

**Supplementary Table 7** Secondary cohort for additional *ADARB1* isoform analyses. All samples were obtained from the National Institute of Child Health and Human Development Brain and Tissue Bank.

Sample ID	Diagnosis	Age		Sex	Race	PMI <sup>a</sup>
		Years	Days			Hours
629	Neurotypical	7	306	Male	African American	18
1430	Neurotypical	38	187	Male	African American	26
1545	Neurotypical	45	95	Male	White	20
1908	Neurotypical	13	360	Male	White	13
4594	Neurotypical	20	131	Male	African American	10
4899	Autism	14	126	Male	White	9
5027	Autism	37	353	Male	African American	26
5115	Autism	46	135	Male	White	29
5176	Autism	22	199	Male	African American	18

<sup>a</sup> PMI, post-mortem interval.

**Supplementary Table 8** Relationships between editing of neighboring sites, and between editing and splicing of the AMPA receptors *GRIA2*, *GRIA3* and *GRIA4*.

**Strong associations between editing and splicing of three AMPA receptors**

<b><i>GRIA4</i></b>		<i>flip</i>	<i>flop</i>		
R	9147	193		Odds ratio (OR)	101.892
G	3362	7228		95% confidence interval (CI)	[87.85-118.18]
				Fisher's exact p-value	<1e-300
				Predominant isoform	<i>flip</i>

<b><i>GRIA3</i></b>		<i>flip</i>	<i>flop</i>		
R	2733	421		OR	30.554
G	709	3337		95% CI	[26.81-34.82]
				Fisher's exact p-value	<1e-300
				Predominant isoform	Roughly equal

<b><i>GRIA2</i></b>		<i>flip</i>	<i>flop</i>		
R	1388	5819		OR	2.608
G	1482	16205		95% CI	[2.41-2.82]
				Fisher's exact p-value	<1e-300
				Predominant isoform	<i>flop</i>

**Independent editing at *GRIA2***

<b><i>GRIA2</i> Q/R</b>		+4A	+4G		
Q	138	18		OR	1.015
R	25072	3319		95% CI	0.62-1.66
				Fisher's exact p-value	1.000
				Power	0.972

Reads were summarized across all samples according to the co-occurrence of the features noted in the contingency tables.

**Supplementary Table 9** Pairwise correlation of editing levels across all individuals between biclustered sites in Figure 11B.

Site 1	Site 2	Pearson's rho	Pearson's p	Benjamini and Hochberg corrected p
GRIK2 Q/R	GRIK2 Y/C	0.9642	5.383E-13	1.615E-10
HTR2C D	HTR2C E	0.9460	3.058E-11	9.022E-10
HTR2C B	HTR2C D	0.9250	7.475E-10	7.475E-08
HTR2C B	HTR2C E	0.9085	5.067E-09	3.800E-07
HTR2C C	HTR2C D	0.8497	5.623E-07	2.812E-05
HTR2C B	HTR2C C	0.8227	2.599E-06	9.748E-05
CYFIP2 K/E	GRIK2 I/V	0.8205	2.914E-06	9.713E-05
HTR2C C	HTR2C E	0.8163	3.606E-06	1.082E-04
CYFIP2 K/E	GRIK2 Y/C	0.8108	4.728E-06	1.289E-04
CYFIP2 K/E	GRIA2 R/G	0.8037	6.629E-06	1.657E-04
CYFIP2 K/E	GRIK2 Q/R	0.7935	1.051E-05	2.252E-04
GRIK2 Q/R -17	HTR2C B	0.7466	6.572E-05	1.038E-03
GRIK2 I/V	GRIK2 Y/C	0.7423	7.623E-05	1.144E-03
GRIK2 I/V	GRIA3 R/G-1	0.7454	1.621E-04	2.537E-03
GRIK2 Q/R -17	HTR2C D	0.7156	1.809E-04	2.360E-03
GRIA2 R/G	GRIK2 Q/R	0.7110	2.078E-04	2.493E-03
GRIA2 R/G	GRIK2 Y/C	0.7047	2.506E-04	2.785E-03
GRIK2 Y/C	GRIA3 R/G-1	0.7137	4.103E-04	3.535E-03
GRIA2 R/G	GRIK2 I/V	0.6866	4.173E-04	4.173E-03
GRIK2 Q/R -17	HTR2C E	0.6687	6.680E-04	6.072E-03
GRIK2 I/V	GRIK2 Q/R	0.6466	1.148E-03	9.564E-03
GRIK2 Q/R	GRIA3 R/G-1	0.6664	1.300E-03	1.300E-03
CYFIP2 K/E	GRIA3 R/G-1	0.6623	1.500E-03	3.404E-03
GRIA2 R/G	GRIA3 R/G-1	0.5932	5.800E-03	1.267E-02
GRIK2 Q/R -17	HTR2C C	0.5235	1.240E-02	5.815E-02

Pearson's correlation was calculated to quantify the linear dependencies between editing at different sites, among all individuals. Biclustering was then used to identify modules of tightly correlated sites. Shown are pairwise correlations between sites contained in the two tightest clusters.

## Supplementary References

1. Anitei M, Stange C, Parshina I, Baust T, Schenck A, Raposo G *et al.* Protein complexes containing CYFIP/Sra/PIR121 coordinate Arf1 and Rac1 signalling during clathrin-AP-1-coated carrier biogenesis at the TGN. *Nat Cell Biol* 2010; **12**(4): 330-340.
2. Riedmann EM, Schopoff S, Hartner JC, Jantsch MF. Specificity of ADAR-mediated RNA editing in newly identified targets. *RNA* 2008; **14**(6): 1110-1118.
3. Nishimoto Y, Yamashita T, Hideyama T, Tsuji S, Suzuki N, Kwak S. Determination of editors at the novel A-to-I editing positions. *Neurosci Res* 2008; **61**(2): 201-206.
4. Dillon C, Goda Y. The actin cytoskeleton: integrating form and function at the synapse. *Annu Rev Neurosci* 2005; **28**: 25-55.
5. Rula EY, Lagrange AH, Jacobs MM, Hu N, Macdonald RL, Emeson RB. Developmental modulation of GABA(A) receptor function by RNA editing. *J Neurosci* 2008; **28**(24): 6196-6201.
6. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol* 2006; **7**(12): 919-931.
7. Lomeli H, Mosbacher J, Melcher T, Hoyer T, Geiger JR, Kuner T *et al.* Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* 1994; **266**(5191): 1709-1713.
8. Wang Q, Miyakoda M, Yang W, Khillan J, Stachura DL, Weiss MJ *et al.* Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *J Biol Chem* 2004; **279**(6): 4952-4961.
9. Belcher SM, Howe JR. Characterization of RNA editing of the glutamate-receptor subunits GluR5 and GluR6 in granule cells during cerebellar development. *Brain Res Mol Brain Res* 1997; **52**(1): 130-138.
10. Kohler M, Burnashev N, Sakmann B, Seeburg PH. Determinants of Ca<sup>2+</sup> permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. *Neuron* 1993; **10**(3): 491-500.
11. Bhalla T, Rosenthal JJ, Holmgren M, Reenan R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat Struct Mol Biol* 2004; **11**(10): 950-956.

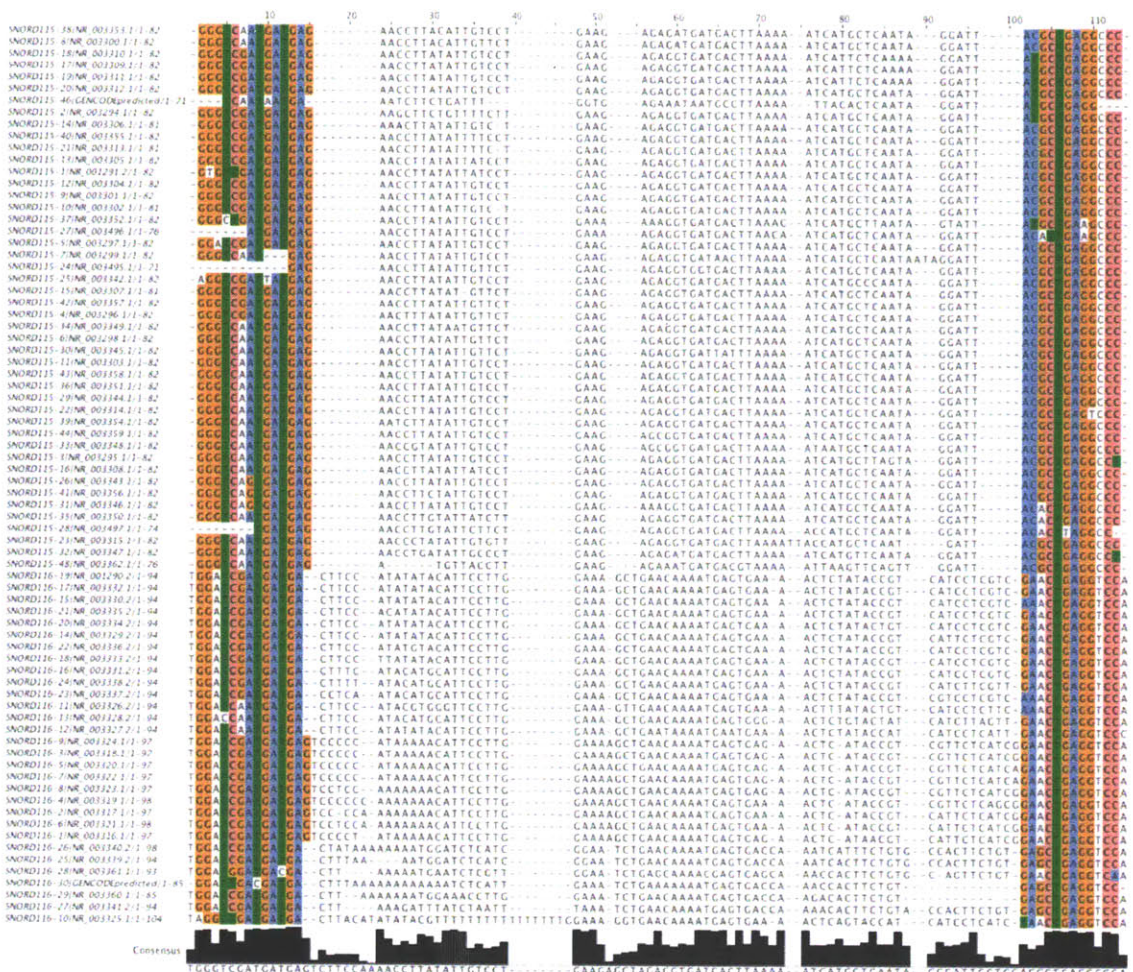
12. Wahlstedt H, Daniel C, Enstero M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res* 2009; **19**(6): 978-986.

# **Appendix F:**

## **Supplementary Material for Chapter 3**

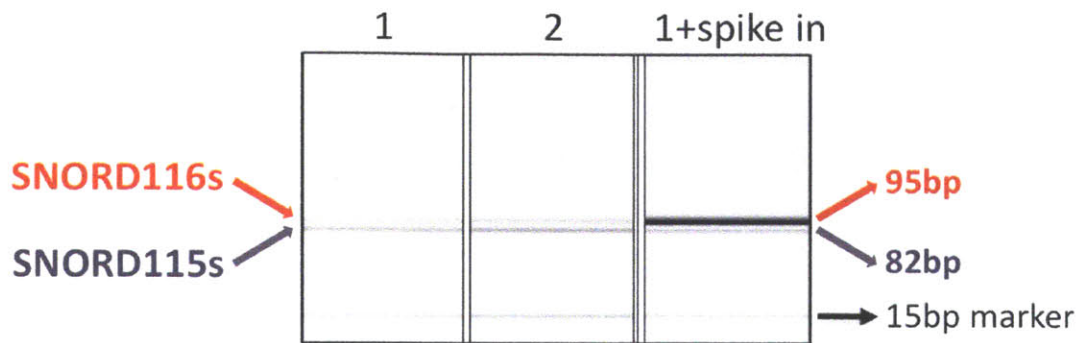
**Alal Eran, Kayla Vatalaro, Jillian McCarthy, Emery N. Brown, Isaac S.  
Kohane, Louis M. Kunkel**

This appendix includes ten supplementary figures and two supplementary tables

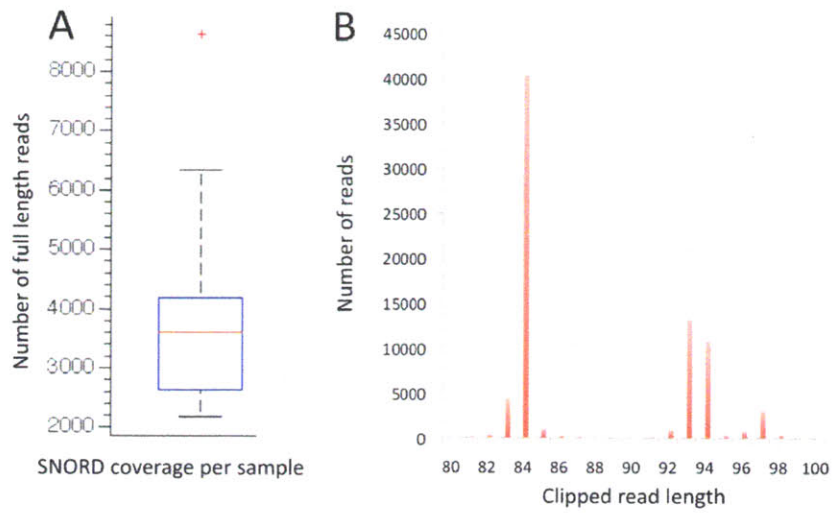


**Supplementary Figure 1** *SNORD115* and *SNORD116* sequence conservation. Multiple sequence alignment of the 48 *SNORD115* genes and 30 *SNORD116* genes highlights the conservation of their stem sequences (colored 5' and 3' ends). These regions served as the basis for SNORD-specific PCR based selection.

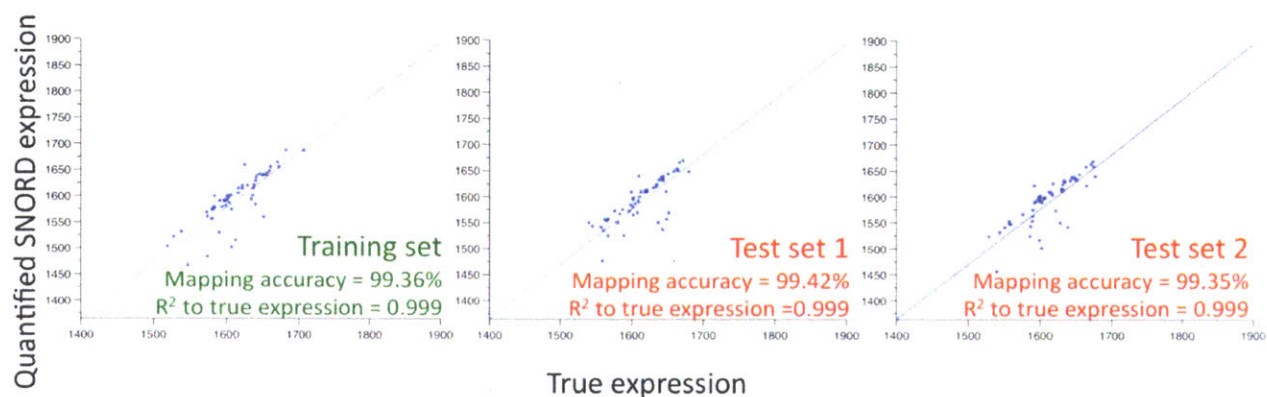




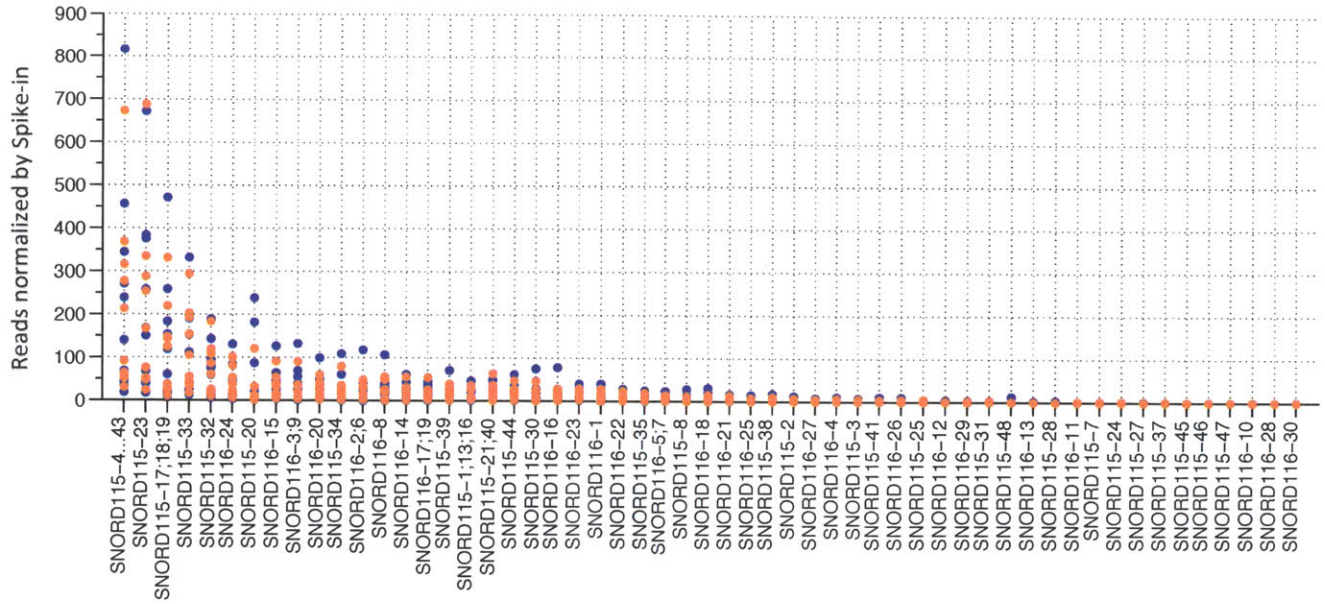
**Supplementary Figure 2** PCR-based selection of 48 *SNORD115* and 30 *SNORD116* genes yields two products of ~82bp and ~95bp respectively, shown for two independent cerebellar samples in lanes 1 and 2. To enable direct inter-individual comparisons, a synthetic SNORD was spiked into the reverse transcription reactions, shown in the right lane. Note that this is an exaggerated amount of spike in presented for visualization purposes.



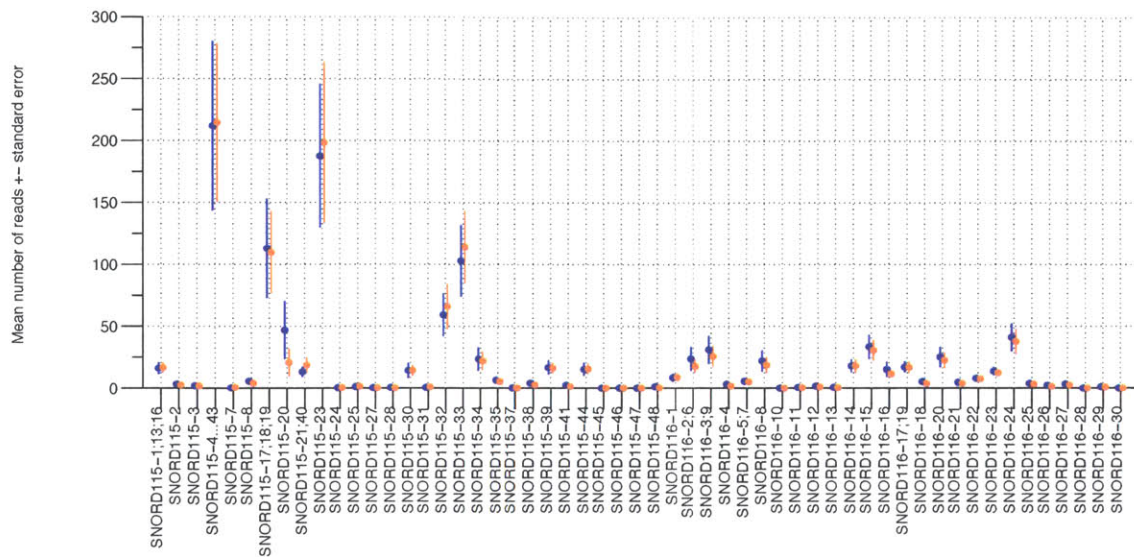
**Supplementary Figure 3** Data summary. **(A)** Coverage. Shown is a boxplot of the number of reads obtained per individual. **(B)** Clipped read length distribution. Detected SNORDs are supported by bidirectional full length reads.



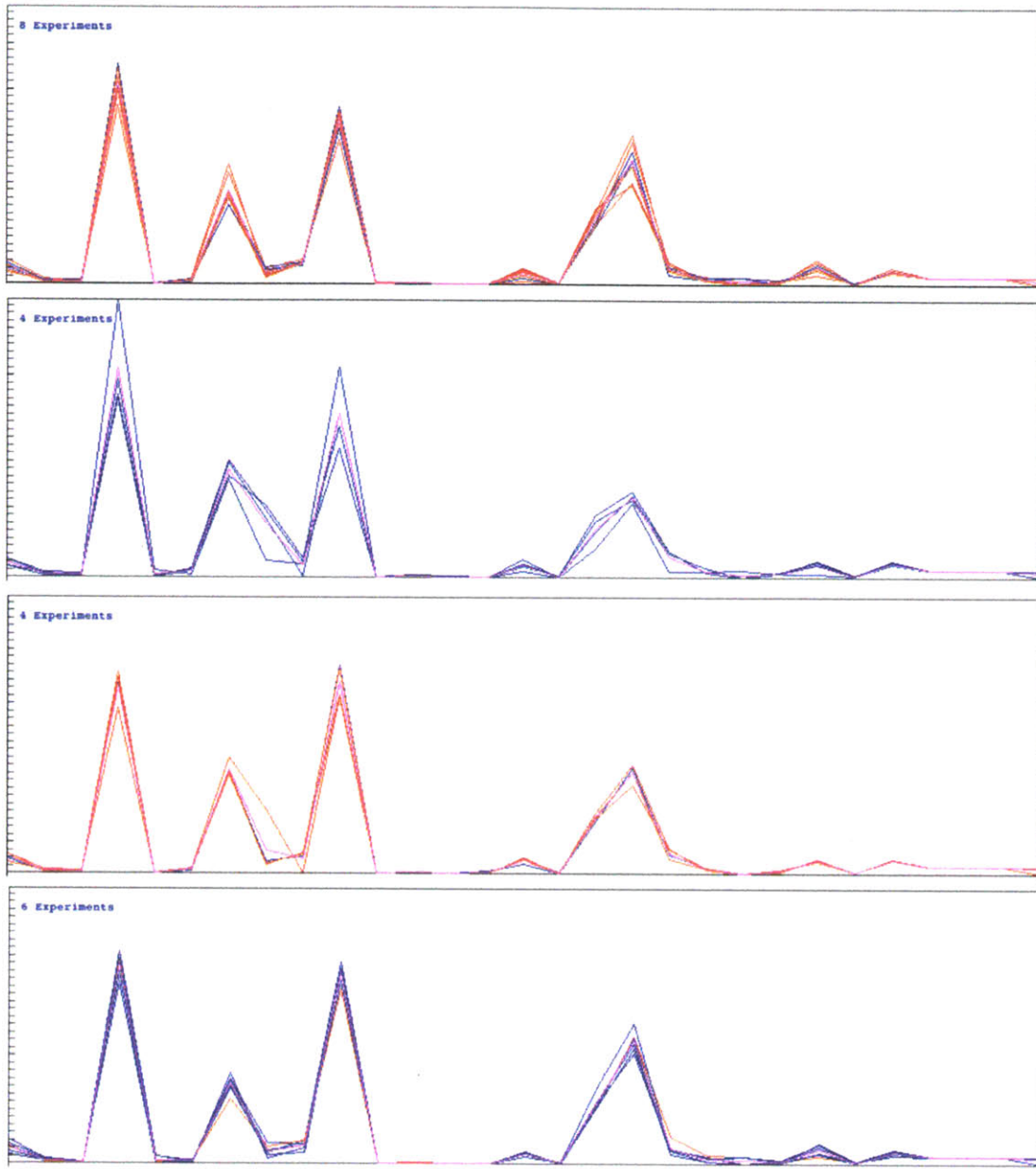
**Supplementary Figure 4** Method evaluation. Synthetic 454 reads were generated based on empirically derived error rates for each of 100,000 randomly sampled *SNORD115* and *SNORD116* Refseq sequences. One set of simulated 454 reads was used for method development, whose main challenge was the high sequence similarity among paralog members of the *SNORD115* and *SNORD116* clusters. Different considerations were tested and tuned until a highly accurate quantification approach was identified, based on the single best alignment of primer-to-primer inserts to 55 SNORD families defined by their members' cross-mapping. This approach was tested on two additional independent simulated datasets, and found to be robustly accurate. Its mapping accuracy is  $99.37 \pm 0.02\%$  and  $R^2$  goodness of fit to the true expression is  $0.999 \pm 3.33e-05$ .



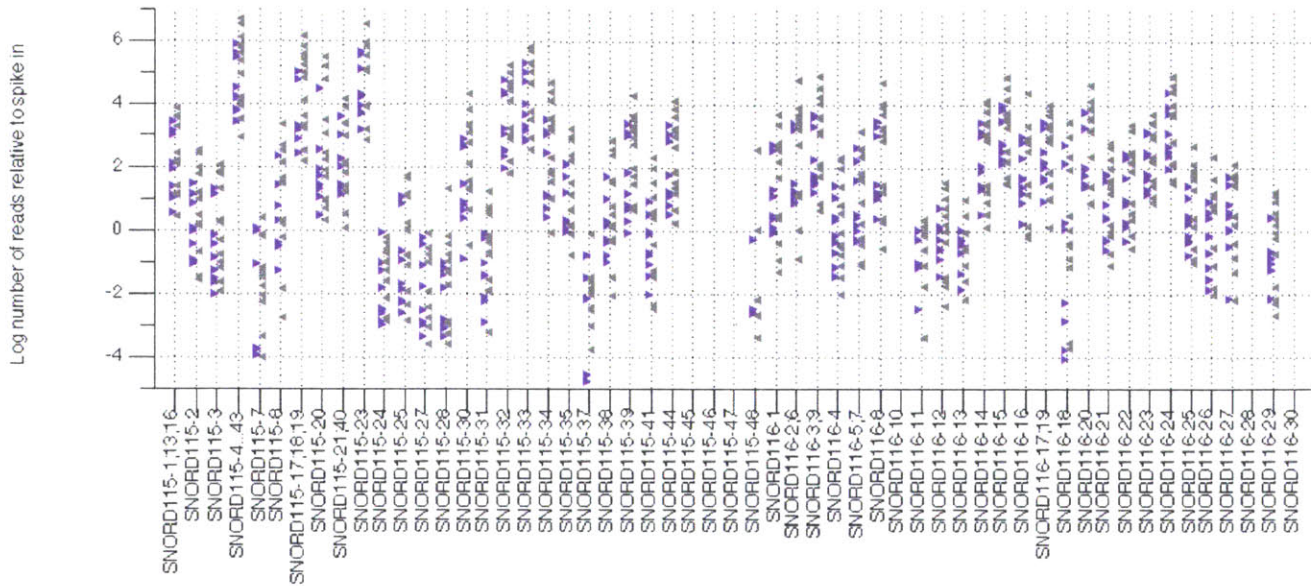
**Supplementary Figure 5** 15q11 snoRNA expression can be described by a gamma distribution with parameters  $\hat{\alpha} = 0.17$  (95% CI=[0.16,0.18]), and  $\hat{\beta} = 117.5$  (95% CI=[101.3,136.2]). Expression was quantified at the SNORD family level, grouping highly paralogous genes that cannot be confidently distinguished from one another (Supplementary Table 2).



**Supplementary Figure 6** Largely similar 15q11 snoRNA expression patterns between all individuals with ASD (orange) and all neurotypical individuals (blue). The mean number of reads relative to the spike in is shown per population  $\pm$  the standard error. Subsequent unsupervised analysis showed a 2-fold upregulation of all SNORDs in male cerebella, confounding this comparison. Consult Figure 4.

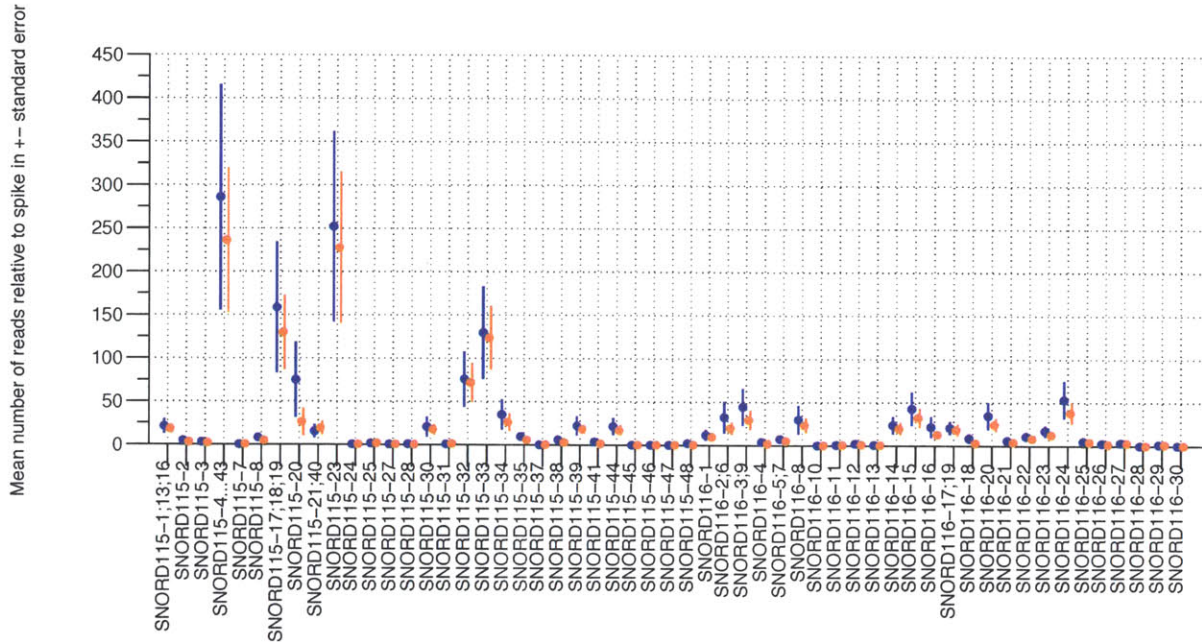


**Supplementary Figure 7** K-means clustering identifies four clusters that are mostly separated by sex and affected status, suggesting that gender may be a strong factor affecting SNORD expression. Shown is *SNORD115* expression across individuals, in which four groups were identified: the first cluster contains six individuals with ASD and two controls, 75% of which are male; the second contains four neurotypical controls; the third contains three males with ASD and one neurotypical male; the fourth contains five neurotypical controls and one individual with ASD, 85% of which are female.



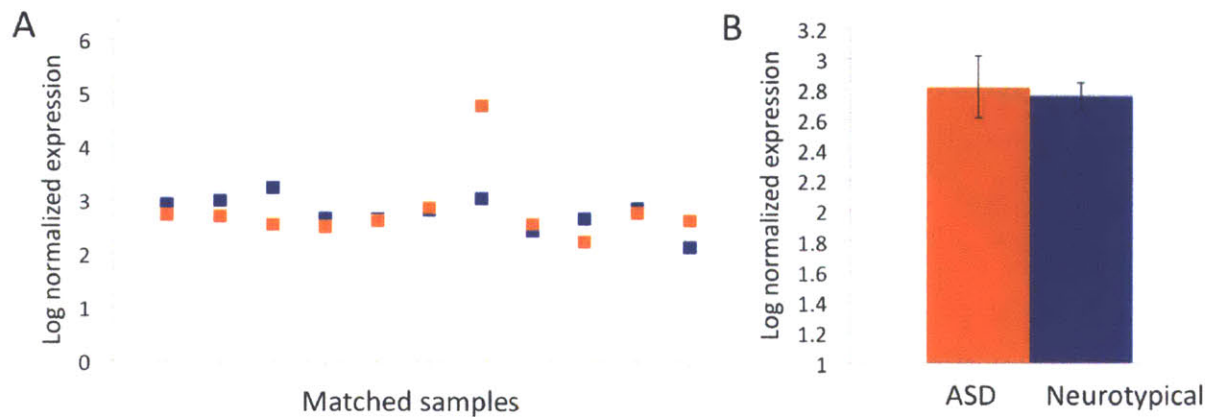
**Supplementary Figure 8** Individual log-transformed SNORD family expression in females (purple) and males (gray). The log number of reads relative to the spike-in is shown for each individual in each SNORD family. *SNORD115-2*, *SNORD115-17;18;19*, *SNORD115-30*, *SNORD115-31*, *SNORD115-34*, *SNORD116-3;9*, *SNORD116-8*, *SNORD116-22*, *SNORD116-25*, and *SNORD116-29* were found to be significantly differentially expressed at the group level.

### Male only *SNORD115* and *SNORD116* expression



**Supplementary Figure 9** Male-only SNORD expression in individuals with ASD (orange) and neurotypical individuals (blue). A comparison of seven males with ASD to six neurotypical males is underpowered. The trend of region-wide downregulation in ASD should be tested in a larger cohort of male brains. Compare to Figure 4 and supplementary figure 6.





**Supplementary Figure 10** Similar *HTR2C* expression in individuals with ASD (orange) and neurotypical individuals (blue). Gene core expression was measured on the Affymetrix Exon 1.0 arrays in 22 of 25 samples. No differences in *HTR2C* expression were found. **(A)** Log normalized expression signal between matched samples. **(B)** Mean *HTR2C* expression per group, depicting a similar level of overall *HTR2C* expression between individuals with ASD and neurotypical individuals.

**Supplementary Table 1** Sample info

No.	Sample ID	Source <sup>a</sup>	Barcode	Diagnosis	Age		Sex	Race	PMI <sup>b</sup>	Storage		RNA integrity RIN <sup>c</sup>	Other medical issues of note (besides autism)
					years	days				hours	years		
1	<b>1185</b>	NICHD BTB	1CTRL	Neurotypical	4	258	Male	White	17	6	39	8.5	None specified
2	<b>1284</b>	NICHD BTB	9CTRL	Neurotypical	3	123	Female	African American	11	5	185	8.4	None specified
3	<b>1349</b>	NICHD BTB	1ASD	Autism	5	220	Male	White	39	5	31	7.5	ADD with hyperactivity
4	<b>1407</b>	NICHD BTB	12CTRL	Neurotypical	9	46	Female	African American	20	5	33	8.6	None specified
5	<b>1499</b>	NICHD BTB	12ASD <sup>d</sup>	Neurotypical	4	170	Female	Asian	21	5	204	8.2	None specified
6	<b>1541</b>	NICHD BTB	10CTRL	Neurotypical	20	228	Female	White	19	4	142	8.1	None specified
7	<b>1638</b>	NICHD BTB	10ASD	Autism	20	277	Female	White	50	3	272	8.4	ADD, hyperactivity, epilepsy, irregular response to pain, schizophrenia, hearing problems, sleeping difficulty, abnormal gait, microencephaly
8	<b>1670</b>	NICHD BTB	6CTRL	Neurotypical	13	99	Male	White	5	3	256	8	None specified
9	<b>1706</b>	NICHD BTB	12CTRL	Neurotypical	8	214	Female	African American	20	3	231	8.2	None specified
10	<b>1793</b>	NICHD BTB	2CTRL	Neurotypical	11	270	Male	African American	19	3	40	8	None specified
11	<b>1860</b>	NICHD BTB	5CTRL	Neurotypical	8	2	Male	White	5	2	268	7.6	None specified
12	<b>1864</b>	NICHD BTB	8CTRL	Neurotypical	2	178	Female	White	8	2	250	7.9	None specified
13	<b>4231</b>	NICHD BTB	2ASD	Autism	8	300	Male	African American	12	1	299	7.9	None specified
14	<b>4671</b>	NICHD BTB	9ASD	Autism	4	165	Female	African American	13	0	356	8.5	None specified
15	<b>4721</b>	NICHD BTB	3ASD	Autism	8	304	Male	African American	16	0	324	8.1	None specified

No.	Sample ID	Source <sup>a</sup>	Barcode	Diagnosis	Age		Sex	Race	PMI <sup>b</sup>	Storage		RNA integrity RIN <sup>c</sup>	Other medical issues of note (besides autism)
					years	days				hours	years		
16	<b>4722</b>	NICHD BTB	7CTRL	Neurotypical	14	198	Male	White	16	1	59	7.8	None specified
17	<b>4787</b>	NICHD BTB	3CTRL	Neurotypical	12	318	Male	African American	15	0	120	8.6	None specified
18	<b>5144</b>	HBTRC	4ASD	Autism	20		Male	White	23.7	6	193	7.8	Vitiligo, sleeping difficulties per parent
19	<b>5251</b>	HBTRC	4CTRL	Neurotypical	19		Male	White	18.6	4	95	7.2	None specified
20	<b>5569</b>	HBTRC	5ASD	Autism	5		Male	White	25.5	4	344	7.7	Allergies, hyperlexia, no/delayed reaction to heat, never felt pain, sleeping difficulty
21	<b>5666</b>	HBTRC	6ASD	Autism	8		Male	White	22.2	4	230	7.5	Epilepsy, syndactyly, negative for Timothy syndrome testing
22	<b>6399</b>	HBTRC	8ASD	Autism	2	278	Male	White	4	2	130	8.2	None specified
23	<b>6736</b>	HBTRC	11CTRL	Neurotypical	4		Female	Unknown	17.02	1	49	8	None specified
24	<b>7002</b>	HBTRC	11ASD	Autism	5		Female	Unknown	33	0	77	8.2	High tolerance to pain, vision problems "lazy eye"
25	<b>7079</b>	HBTRC	7ASD	Autism	15		Male	Unknown	4			7.7	Asperger Syndrome / high functioning autism

<sup>a</sup> NICHD BTB, National Institute of Child Health and Human Development Brain and Tissue Bank; HBTRC, Harvard Brain Tissue Resource Center.

<sup>b</sup> PMI, post-mortem interval.

<sup>c</sup> RIN, RNA integrity number as determined by the Agilent 2100 Bioanalyzer.

<sup>d</sup> Neurotypical sample sequenced in the ASD pool.

**Supplementary Table 2** SNORD family definitions

Three layers of redundancy are color-coded in red, green, and grey. Red are perfect paralogs (exact duplicate genes). In green are genes whose primer-to-primer sequence is identical. In gray are genes with 99% similarity that could not be confidently distinguished from one another in the simulations. The remaining 85% of the SNORD families contain a single gene.

SNORD Family	Members
SNORD116-1	SNORD116-1
SNORD116-2;6	SNORD116-2, SNORD116-6
SNORD116-3;9	SNORD116-3, SNORD116-9
SNORD116-4	SNORD116-4
SNORD116-5;7	SNORD116-5, SNORD116-7
SNORD116-8	SNORD116-8
SNORD116-10	SNORD116-10
SNORD116-11	SNORD116-11
SNORD116-12	SNORD116-12
SNORD116-13	SNORD116-13
SNORD116-14	SNORD116-14
SNORD116-15	SNORD116-15
SNORD116-16	SNORD116-16
SNORD116-17;19	SNORD116-17, SNORD116-19
SNORD116-18	SNORD116-18
SNORD116-20	SNORD116-20
SNORD116-21	SNORD116-21
SNORD116-22	SNORD116-22
SNORD116-23	SNORD116-23
SNORD116-24	SNORD116-24
SNORD116-25	SNORD116-25
SNORD116-26	SNORD116-26
SNORD116-27	SNORD116-27
SNORD116-28	SNORD116-28
SNORD116-29	SNORD116-29
SNORD116-30	SNORD116-30
SNORD115-1;13;16	SNORD115-1, SNORD115-13, SNORD115-16
SNORD115-2	SNORD115-2
SNORD115-3	SNORD115-3
SNORD115-4;5;6;9;10;11;12;14;15;22;26;29;36;42;43	SNORD115-4, SNORD115-5, SNORD115-6 SNORD115-9, SNORD115-10, SNORD115-11 SNORD115-12, SNORD115-14, SNORD115-15 SNORD115-22, SNORD115-26, SNORD115-29 SNORD115-36, SNORD115-42, SNORD115-43

SNORD115-7	SNORD115-7
SNORD115-8	SNORD115-8
SNORD115-17;18;19	SNORD115-17, SNORD115-18, SNORD115-19
SNORD115-20	SNORD115-20
SNORD115-21;40	SNORD115-21, SNORD115-40
SNORD115-23	SNORD115-23
SNORD115-24	SNORD115-24
SNORD115-25	SNORD115-25
SNORD115-27	SNORD115-27
SNORD115-28	SNORD115-28
SNORD115-30	SNORD115-30
SNORD115-31	SNORD115-31
SNORD115-32	SNORD115-32
SNORD115-33	SNORD115-33
SNORD115-34	SNORD115-34
SNORD115-35	SNORD115-35
SNORD115-37	SNORD115-37
SNORD115-38	SNORD115-38
SNORD115-39	SNORD115-39
SNORD115-41	SNORD115-41
SNORD115-44	SNORD115-44
SNORD115-45	SNORD115-45
SNORD115-46	SNORD115-46
SNORD115-47	SNORD115-47
SNORD115-48	SNORD115-48