Neuron
NeuroView

The Autism Sequencing Consortium: Large-Scale, High-Throughput Sequencing in Autism Spectrum Disorders

Joseph D. Buxbaum,[1,*] Mark J. Daly,[2] Bernie Devlin,[3] Thomas Lehner,[4] Kathryn Roeder,[5] Matthew W. State,[6,*] and The Autism Sequencing Consortium[7]
[1]Seaver Autism Center, Departments of Psychiatry, Neuroscience, and Genetics and Genomic Sciences, and the Friedman Brain Institute, Mount Sinai School of Medicine, New York, NY 10029, USA
[2]Broad Institute and Translational Genetics Unit, Harvard Medical School, Boston, MA 02114, USA
[3]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[4]National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA
[5]Department of Statistics and Lane Center for Computational Biology, Carnegie-Mellon University, Pittsburgh, PA 15213, USA
[6]Program on Neurogenetics, Child Study Center and Department of Psychiatry, Yale University, New Haven, CT 06520, USA
[7]A full list of The Autism Sequencing Consortium members can be found in Table S3 available online
*Correspondence: joseph.buxbaum@mssm.edu (J.D.B.), matthew.state@yale.edu (M.W.S.)
http://dx.doi.org/10.1016/j.neuron.2012.12.008

Research during the past decade has seen significant progress in the understanding of the genetic architecture of autism spectrum disorders (ASDs), with gene discovery accelerating as the characterization of genomic variation has become increasingly comprehensive. At the same time, this research has highlighted ongoing challenges. Here we address the enormous impact of high-throughput sequencing (HTS) on ASD gene discovery, outline a consensus view for leveraging this technology, and describe a large multisite collaboration developed to accomplish these goals. Similar approaches could prove effective for severe neurodevelopmental disorders more broadly.

Introduction

The past decade has seen tremendous advances in the genetics of autism spectrum disorders (ASDs). Rapidly evolving genomic technologies combined with the availability of increasingly large study cohorts has led to a series of highly reproducible findings (Betancur, 2011; Devlin et al., 2011; Devlin and Scherer, 2012), highlighting the contribution of rare variation in both DNA sequence and chromosomal structure, placing limits on the risk conferred by individual, common genetic polymorphisms, underscoring the role of de novo germline mutation, suggesting a staggering degree of genetic heterogeneity, demonstrating the highly pleiotropic effects of ASD-associated mutations, and identifying, definitively, an increasing number of specific genes and chromosomal intervals conferring risk. This progress marks a long-awaited emergence of the field from a period of tremendous uncertainty regarding viable approaches to gene discovery. At the same time, the findings underscore the scale of the challenges ahead.

Genetics of ASD before High-Throughput Sequencing

Twin studies have consistently identified a significant genetic component of ASD risk (Hallmayer et al., 2011; Ronald and Hoekstra, 2011) and gene discovery dates back over a decade (Betancur, 2011; Devlin and Scherer, 2012). Recent analyses demonstrate that common polymorphisms carry substantial risk for ASD (Anney et al., 2012; Klei et al., 2012). However, common polymorphisms have so far proven difficult to identify and replicate, probably because the relative risk conferred by these loci is small and cohort sizes have not yet reached those found necessary to identify common polymorphisms contributing to other complex psychiatric disorders (Devlin et al., 2011). In contrast, a focus on rare and de novo mutation has already been highly productive in uncovering an appreciable fraction of population risk and identifying variation conferring relatively larger biological effects.

An example of the considerable traction provided by a focus on rare inherited and de novo variation can be found in the earliest successes in ASD genetics. The protein products of risk genes for patients ascertained with nonsyndromic ASD, including *NLGN4X*, *NRXN1*, and *SHANK3*, colocalize at the postsynaptic density in excitatory glutamatergic synapses with those coded for by genes first identified in syndromic subjects, including *FMRP*, *PTEN*, *TSC1*, and *TSC2* (note, however, that as gene identification continues, "syndromic" genes are being identified in nonsyndromic cases and vice versa). These results are cause for optimism with regard to the prospects for identifying treatments that will have efficacy well beyond the boundaries suggested by mutation-defined subgroups. Moreover, the themes highlighted in these studies presaged the current era of gene discovery, not only with regard to the contribution of rare alleles and sporadic germline mutations, but also by providing the first concrete evidence of the tremendous pleiotropy and variable penetrance that are now considered characteristic of ASD risk loci.

Analyses of chromosome microarrays have provided compelling evidence that

submicroscopic variations in chromosomal structure, called copy number variation (CNV), contribute to ASD risk (Betancur, 2011; Cooper et al., 2011; Pinto et al., 2010; Sanders et al., 2011). Certain CNVs are recurrent, often due to either the presence of low-copy repeats or subtelomeric deletions, and within some of these, the attendant risk has been related to a single gene (e.g., *NRXN1* in 2p16.3, *SHANK3* in 22q13.3 deletions, and *MBD5* in 2q23.1) (Betancur, 2011). With the widespread use of microarrays in the clinical setting, accompanied by increasingly large-scale analyses of research cohorts, the field is beginning to consolidate population level data for CNV with some clear findings: (1) between 5%–10% of previously unexplained cases will carry an ASD-CNV; (2) both de novo and transmitted CNV confer risk; (3) rare CNV generally confers larger risks than are typically associated with common variants; however, many of these high-risk regions appear to contribute to ASD through a complex pattern of inheritance; and (4) the majority of confirmed ASD loci show both variable expressivity and pleiotropic effects.

A recent analysis of structural variation in ASD families from the Simons Simplex Collection, focusing on comprehensively assessed quartets of mother, father, ASD proband, and unaffected sibling (Sanders et al., 2011), serves as a useful illustration. Large, rare de novo CNV showed a 3-fold increase in probands relative to their matched siblings, yielding a highly significant difference. Moreover, the de novo events in probands were found to carry about ten more genes on average even after accounting for CNV size. Among the many results from these data, one of special salience is that no matter how inherited CNVs were parsed for analysis, no significant difference between probands and siblings emerged, even though there were many more inherited than de novo CNVs. A plausible interpretation of these results is that de novo events that alter gene function have a much higher signal-to-noise ratio than inherited CNVs that also effect gene function; put another way, gene-rich de novo CNVs are highly likely to be capturing one or more ASD genes, while inherited gene-rich CNVs are less likely on average to harbor ASD genes.

With regard to pursuing biological studies, a drawback of CNVs is their tendency to encompass multiple genes. Accordingly, if the genetic architecture of sequence variation in ASD mirrored that suggested by CNV, HTS would represent an extremely important addition to the genomic armamentarium.

The ASC, Sequencing, and Gene Discovery in ASD

Against this backdrop, the Autism Sequencing Consortium (ASC) was formed in 2010, in anticipation of both the tremendous impact HTS would have on ASD genomics as well as the many challenges the field would face as a consequence. Now including more than 20 research groups, the ASC has as its goal to collectively exploit sequencing approaches to resolve a substantial fraction of the genetic factors involved in ASD. While there are probably many hundreds of undiscovered ASD loci, emerging data provide sufficient empirical evidence upon which to develop sound and systematic approaches to identifying these loci.

From the outset, the effort to constitute the group and define the objectives for the ASC was faced with the challenge of balancing the obvious benefits of working cooperatively with the strongly held conviction that a diversity of approaches and the presence of multiple competing efforts has played, and will continue to play, an indispensable role in the field's rapid progress. The participating investigators undertook an effort to address the range of related issues, including data sharing, prospectively and prior to the widespread availability of HTS data.

In 2011, the ASC held an open meeting of investigators, funders, and other stakeholders to refine and crystallize the plans and proposals. The meeting, which included more than 100 onsite participants (see Table S1 available online) and additional web participants, was organized around three working groups: (1) sequence technology, data harmonization, and statistical inference (B. Devlin and M. Daly, Chairs); (2) samples and phenotypes (J. Buxbaum and M. Gill, Chairs); and (3) future directions (T. Lehner and M. State, Chairs). Working groups addressed a variety of issues including study designs, statistical approaches, sample availability and composition, data normalization, bioinformatics challenges, and the integration of gene discovery into broader efforts at translational neuroscience (Table S2). The meeting was video cast and can be accessed at http://videocast.nih.gov/pastevents.asp. We present a synthesis and summary of that meeting, reflecting both a current view of the field and consensus recommendations for gene discovery.

In light of the high degree of genetic heterogeneity in ASD, it was apparent that HTS would provide a powerful platform for gene discovery. Whole-genome sequencing (WGS) can detect structural variation of all types, ranging from gross chromosomal rearrangements to CNV and insertion deletions (indels), while also providing highly sensitive single-base resolution. Similarly, whole-exome sequencing (WES) can reliably detect single-nucleotide variants (SNVs) in the coding segments of the genome, many indels, and some CNV. Of course, both technologies provide the ability to identify rare alleles to a degree that is not possible on genotyping platforms.

To date, four large-scale ASD WES studies have been carried out in trios, namely a proband with ASD and the biological parents, or in quads, a trio plus an unaffected sibling (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). Among the 1,000 families assessed by the four studies, the rate of de novo loss-of-function (LoF) variation was consistently found to be significantly higher in cases compared to controls, allowing for the development of rigorous statistical approaches to identifying specific risk genes. Indeed, six ASD genes were identified, *CHD8*, *DYRK1A*, *GRIN2B*, *KATNAL2*, *POGZ*, and *SCN2A*, because they carried recurrent, highly damaging de novo events. While *SCN2A* has been previously implicated in epilepsy, none of these genes were known to carry ASD risk. Another key finding, one that will prove useful for gene discovery, was that roughly half of all de novo LoF mutations seen in ASD probands fall in ASD genes, with about 12% of ASD subjects showing a de novo LoF mutation.

These WES studies found a background rate of missense de novo variation that is more than 10-fold higher than that for

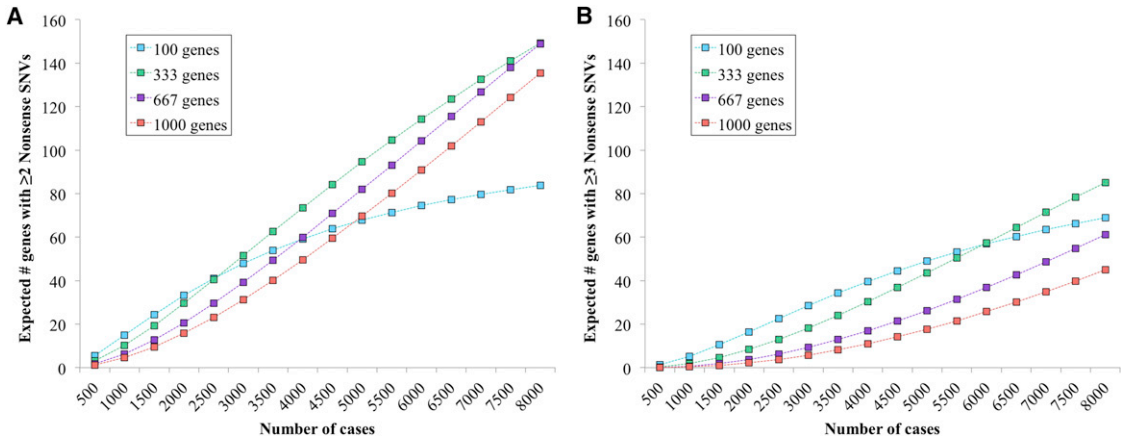

Figure 1. Expected Yield of Identified ASD Risk Genes as a Function of the Number of Trios Evaluated for De Novo LoF Variants
Analyses (data not shown) have demonstrated that for 8,000 families, a threshold of three or more de novo LoF variants (right) is sufficient to declare genome-wide significance. Using two or more de novo LoF as a criteria (left) would identify ~80 likely ASD genes with an FDR of 0.1. As we learn more about annotation of truly damaging missense de novo variants, the rate of discovery will rise dramatically—and we anticipate that this will happen soon by our work and that of others such as the 1,000 genomes project. In addition, other forms of discovery are not modeled but will lead to additional gene discovery. The analyses underlying this figure are further described in Sanders et al. (2012).

LoF alleles. These missense changes undoubtedly include risk alleles; however, only a 5%–10% excess of such mutations was found in ASD cohorts, a difference that did not reach significance collectively across studies. Accordingly, it is not yet possible to confidently assign risk to this broad category of mutation, nor to establish an agreed upon threshold for the significance of observing ''recurrent'' de novo missense mutations in a given gene. Given the relevance of LoF alleles, this difficulty surely reflects the signal-to-noise problems engendered by neutral background variation and the difficulties that attend differentiating the subset of truly functional missense variations.

The interpretation of case-control exome sequencing has also not been as straightforward as family studies evaluating de novo LoF events. For example, WES of a sample of 1,000 cases and 1,000 controls and inspection of the six novel ASD genes just described showed, in hindsight, only a slight excess of LoF mutations in *KATNAL2* and *CHD8* in cases, a difference that did not approach statistical significance (Neale et al., 2012). Indeed, across the entire genome, no genes were found to harbor a sufficiently large excess of rare alleles in cases versus controls to support a significant association after accounting for multiple comparisons (Liu et al., personal communication). These results are consistent with the multiple lines of evidence supporting a large number of ASD risk genes scattered throughout the genome. Methods to extract signal from case-control studies, alone and in combination with de novo data, are rapidly evolving. Still, it seems reasonable to conclude that large studies, involving tens of thousands of subjects, will be necessary to identify risk loci using standard analyses of mutation burden in case-control samples.

The ASC's Proposed Path Forward

The path forward is either WES or WGS in large cohorts. Because of its higher signal-to-noise ratio, discovery of de novo mutations, especially LoF mutations that cluster in the same gene among unrelated individuals, is an immediately productive approach to gene discovery and will be emphasized. As the number of trios or quads sequenced grows linearly, the rate of gene identification is predicted to accelerate (Figure 1). Based on the first results from the ASC sites, the value of expanding efforts in search of recurrent de novo events is clear. If HTS were to be performed on 8,000 families, and even ignoring other sources of key information, the experiment should yield between 40–60 novel ASD genes and a large number of additional genes falling just short of significance that could readily be confirmed via targeted sequencing in additional large patient cohorts (Figure 1).

Efforts of this scale are underway. To give some examples, the Simons Foundation has committed to sequencing more than 2,600 quartets, the ARRA Autism Sequencing Consortium has finished 400 families, Genome Canada is supporting the sequencing of 1,000 trios and families, and the UK10K project is targeting ~800 ASD cases in the 10,000 to be sequenced. Autism Speaks, in partnership with the Beijing Genomics Institute, is committed to whole-genome sequencing of 60 families and has proposed an ultimate target of 2,000 families.

A key related question is whether an even higher yield of ASD genes can be gleaned simply by making more effective use of data generated in ongoing experiments. In fact, it is a near certainty that there will be significant traction in evaluating other types of mutations beyond de novo LoF variants. Ongoing research promises to refine the interpretation of various classes of mutations, including inherited variation from family and case-control analyses, for which the chief obstacle is the high frequency of apparently neutral rare variation in the genome. In addition, there are already emerging successes focusing on recessive and X-linked LoF variation. These efforts may be aided through the study of sequence data in unusual high-risk extended pedigrees that are also available. Thus, based on refined interpretation of sequence, we

expect to identify additional ASD genes. Progress in this area will also require methods to combine data on inherited variation with data on de novo events.

The ASC recognizes that a focus on DNA sequence, by itself, is insufficient. There are additional sources of information that can be brought to bear to identify novel ASD genes (Figure 2). RNA-seq and Chip-seq studies of typical and ASD brains offer an increasingly accurate picture of gene coexpression and regulatory networks, thereby identifying processes altered in ASD, both by themselves and by overlap with genes identified as disrupted in ASD. And RNA-seq studies of peripheral samples (blood or induced neural cells) have the potential to survey thousands of individuals to identify ASD-related biological signatures. Bearing in mind that one out of every two de novo LoF events found in probands hits an ASD gene, these kinds of biological information can be exploited to separate the ASD "wheat from the chaff." Moreover, CNVs have already identified many regions of the genome as harboring one or more ASD genes, so there will be ways of combining CNV and sequence information to identify additional ASD genes. If other sources of information prove as useful as we anticipate, the yield of ASD genes could easily amplify well beyond that predicted by Figure 1, paving the way for systems biological and neurobiological follow-up. In addition, understanding gene-environment interaction and gene-environment correlation remains an important long-term goal in ASD, and such approaches will be enormously facilitated by this gene discovery.

Beyond gene discovery, integration of information as depicted in Figure 2 holds the promise for clarifying the etiology and biology of ASD. Eventually we foresee identifying ASD-related biological signatures to define subgroups enriched for disruptions in specific pathways and, ultimately, to identify subsets of patients amenable to specific treatments. For brain and blood samples, it is also now possible to interrogate epigenetic modifications, mechanisms that are likely to play a substantial role in ASD. Other potentially uncharacterized risks include rare disruption in the mitochondrial genome and alterations to the microbiome. The microbiome, thought to contribute as much as 10% of the metabolites in the bloodstream, has recently been shown to affect behavior in model systems. If it is a mediator of ASD risk, it would be particularly amenable to intervention.

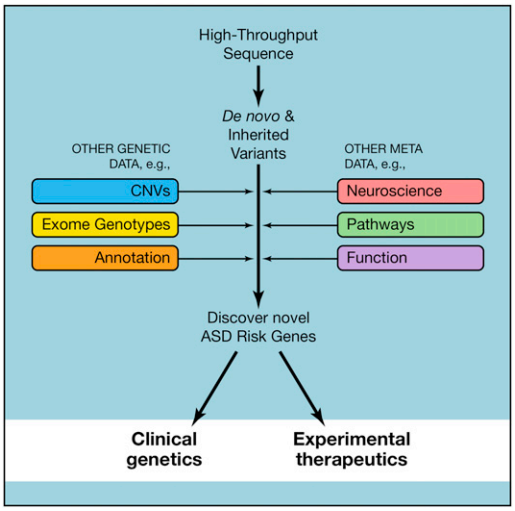

Figure 2. A Pathway for Discovering Novel Risk Genes for Autism Spectrum Disorders

Beyond the genes identified by de novo events, as illustrated in Figure 1, inherited variation will also prove useful for identifying risk genes, the simplest to interpret being variation acting recessively. Potentially imbuing far greater discovery power will be two other sources of data: genetic data of various kinds, such as genomic regions already implicated in risk for ASD by copy number variation, and metadata from other complementary fields, such as neuroscience and so called "omics" (e.g., pathways of functionally related genes). With these additional sources of information, it will be possible to identify a substantial fraction of ASD genes, providing sufficient grist for clinical geneticists to predict risk and for pharmacologists to develop therapeutics.

The ASC Challenges Ahead

Collaboration

The empirical data to develop the ASC strategy involved three ASC groups who shared their data prior to publication (Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). This is a model that we strongly favor in the ASC, as it strikes a workable balance between preserving intellectual diversity and competitiveness while still reaping the benefits of cooperative research.

Samples

Approximately 8,000–10,000 families are available and poised for discovery efforts among the groups contributing to the ASC, and these all should be sequenced with HTS approaches. However, we believe the collection of additional ASD cohorts remains a vitally important priority that would dramatically accelerate gene discovery, validation and characterization of mutation spectra in ASD-risk genes, clarify genotype-phenotype relationships, and provide a critical substrate for ongoing effort to identify shared neurobiological mechanisms and treatment targets among patients with diverse genetic etiologies.

Bioinformatics

WES is currently favored over WGS because of its lower-cost, lower-informatics overhead and ease of interpretation. However, WGS provides a more comprehensive view of both sequence and structural variation, does not require target capture, and is able to better interrogate regions of high GC content that may be particularly prone to de novo mutation. The transition from WES to WGS over the next several years is likely, given the trajectory of costs and the steady introduction of new sequencing technologies. These developments will undoubtedly contribute further to the understanding of ASD but, in our view, should not delay current WES efforts, which are already driving new studies of the biology of ASD.

Sequencing and analyzing data from tens of thousands of samples generates a volume of data that overwhelms standard approaches to data storage and backup. Movement of data is cumbersome, time consuming, or infeasible. Because fair collaboration among ASC researchers requires that all participants have equal access to all data and equal opportunity to analyze it, and because variant detection remains a work in progress, the ASC solution is to create a bioinformatics infrastructure to collate data at a single site for analysis. A strength of this approach is that it has capacity for massive data sharing and joint analyses, thereby accelerating progress while avoiding the pitfalls of beginning data harmonization post hoc once individual studies have been completed and

published. Nonetheless, the ASC recognizes the prerogative of individual groups to investigate their own data freely.

Foundational Resources for Functional and Systems Biological Analyses

As novel genes and pathways are identified, functional analyses will take these findings forward to understand mechanisms of pathophysiology. While elegant functional approaches exist, high-throughput methods will be essential. This need is even more acute when one considers that many variants of unknown significance will be identified, so that augmenting genetic findings with in vitro assays could help determine whether a particular gene plays a bona fide role in ASD.

Integration with Other Psychiatric Disorders

ASC data will be further enhanced by HTS efforts focused on disorders that are already showing overlapping risk loci, including intellectual disability, epilepsy, and schizophrenia. It is reasonable to predict that knowledge about all these disorders will be enhanced by collaboration and open sharing of data and results.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two tables and a full list of The Autism Sequencing Consortium members and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2012.12.008.

ACKNOWLEDGMENTS

The authors thank the National Institute of Mental Health (NIMH), the National Human Genome Research Institute (NHGRI), and the Seaver Foundation for supporting the ASC meetings and calls and for facilitating and encouraging broad participation. The authors also thank Jessica Brownfeld for help with organization and manuscript preparation.

REFERENCES

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., Bölte, S., Bolton, P.F., Bourgeron, T., et al. (2012). Hum. Mol. Genet. *21*, 4781–4792.

Betancur, C. (2011). Brain Res. *1380*, 42–77.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). Nat. Genet. *43*, 838–846.

Devlin, B., and Scherer, S.W. (2012). Curr. Opin. Genet. Dev. *22*, 229–237.

Devlin, B., Melhem, N., and Roeder, K. (2011). Brain Res. *1380*, 78–84.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., et al. (2011). Arch. Gen. Psychiatry *68*, 1095–1102.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). Neuron *74*, 285–299.

Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D., et al. (2012). Mol. Autism *3*, 9.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., et al. (2012). Nature *485*, 242–245.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Nature *485*, 246–250.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Nature *466*, 368–372.

Ronald, A., and Hoekstra, R.A. (2011). Am. J. Med. Genet. B. Neuropsychiatr. Genet. *156B*, 255–274.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Neuron *70*, 863–885.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). Nature *485*, 237–241.