BRIEF COMMUNICATIONS ARISING

Bias towards large genes in autism

ARISING FROM I. F. King *et al. Nature* **501,** 58–62 (2013); doi:10.1038/nature12504

In an important recent paper, King *et al.*[1] reported that inhibition of TOP1 and other topoisomerases reduces the expression of extremely long genes. They also showed that the list of large genes affected by TOP1 inhibition is enriched with candidate genes for autism spectrum disorders (ASD); however, the list of candidate genes that was used contains many genes with limited evidence for association with ASD[2]. Here we demonstrate that the size of the genes among ASD candidate genes is biased towards extremely large genes only for genes identified to be disrupted by copy number variations (CNVs). Thus, our analysis suggests that the association between large genes and ASD is mainly driven by the method that implicated the genes in ASD. There is a Reply to this Brief Communication Arising by Zylka, M. J. *et al. Nature* **512,** http://dx.doi.org/10.1038/nature13584 (2014).

The literature on ASD mentions many candidate genes, yet convincing evidence was yielded only for few of them. This is reflected in the SFARI Gene database used by King *et al.*[1]. This database currently contains 588 genes, for which the evidence for association with ASD varies considerably. To address this concern, a gene scoring module was developed in SFARI Gene 2.0 to estimate the evidence level for individual genes[2]. For example, Table 1 of King *et al.*[1] lists 49 ASD candidate genes that were affected by Top1 inhibitors, but only three of these genes are considered strong candidates, four genes have suggestive evidence, and an additional two genes are involved in syndromes associated with ASD. To retest the association between gene length and ASD, we first selected genes in the SFARI database that had a score of at least suggestive evidence. The gene with the strongest evidence, *CHD8*, is not particularly large (~50 kilobases), but the list also contains some of the largest genes in the genome like *AUTS2* and *NRXN1* (~1,000 kb). The SFARI Gene database is based on studies that focused on specific genes and on genome-wide studies, which are considered to be unbiased. However, as we show below, even in genome-wide studies there are biases influenced by the type of study.

The genome-wide search for ASD risk genes has been performed mainly by searching for rare and *de novo* variants, using two main methods: microarrays to identify CNVs, and exome sequencing to identify single nucleotide variations (SNVs) predicted to alter the protein sequence. Under the assumption of uniform mutation rate, the probability for a coding SNV is not strongly related to the total size of the gene, which is mainly determined by the length of the introns and untranslated regions. However, this is not true for CNVs. Most CNVs identified in ASD are large and contain multiple genes[3,4]. Therefore, it is hard to associate any particular gene with ASD. In contrast, when a gene is extremely large there is a higher probability for it to become an ASD risk gene because it was the only gene affected by the CNV. Accordingly, we hypothesized that large genes will be associated with ASD mainly if implicated by CNVs.

Following the above hypothesis, we divided the ASD genes from the SFARI Gene database into two groups on the basis of the mutation type that implicated the gene in ASD: genes affected by CNVs or by SNVs. We plotted the distribution of gene sizes separately for each group. As can be seen in Fig. 1a, the distributions were notably different. "Genes with CNVs" showed bimodal distribution of short and large genes, whereas "genes with SNVs" were relatively short.

To further study the association between gene size and mutation types, we focused on six studies that searched for *de novo* CNVs and SNVs, mostly in the same cohort (the Simons Simplex Collection)[3–8]. We compared the distribution of gene lengths between all coding genes in the genome, brain-specific genes, and genes with *de novo* SNVs or CNVs, in both ASD probands and unaffected siblings (Fig. 1b). Consistent with King *et al.*[1], and as was previously reported[9], we found that brain-specific genes are significantly larger than the average gene in the genome ($P = 6 \times 10^{-22}$). Whereas genes identified to be disrupted by *de novo* SNVs in ASD had a similar distribution of sizes as brain-specific genes ($P = 1$), the size of genes with *de novo* CNVs were significantly larger than either group ($P < 3 \times 10^{-3}$, $P < 5 \times 10^{-5}$, respectively). Furthermore, there was no difference in gene size between affected versus unaffected children for both *de novo* CNVs and SNVs ($P = 1$).

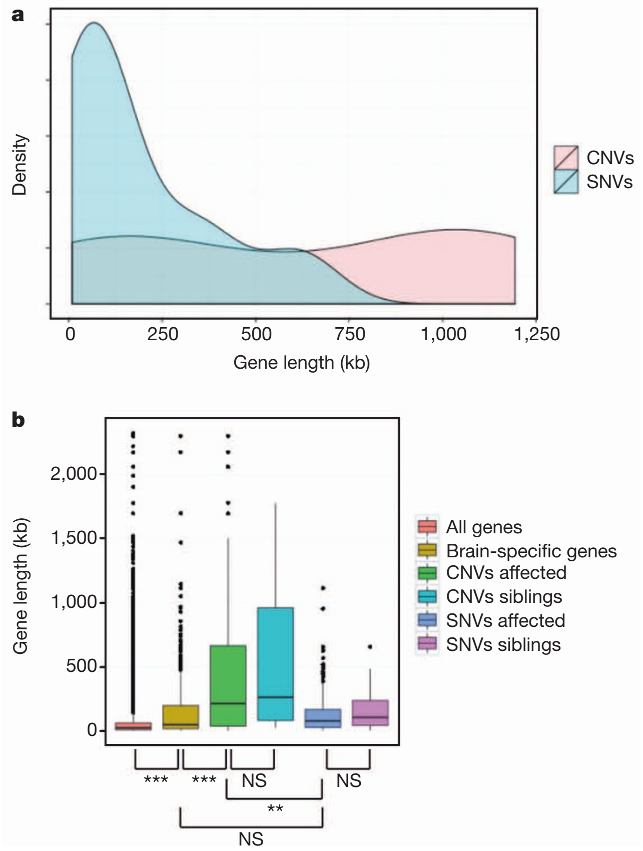

Figure 1 | Association between gene size and mutation types. a, Density plots of gene length for ASD genes with evidence for association according to the SFARI Gene database. Genes were divided to two groups on the basis of the type of mutations that implicated the gene in ASD, CNVs or SNVs. **b**, The distribution of gene length is presented by box plots for all genes in the genome, brain-specific genes and genes with *de novo* SNVs or CNVs identified by recent genome-wide studies. NS, non-significant; $^{**}P < 0.01$; $^{***}P < 0.001$.

In summary, our analysis suggests that the association between large genes and ASD that was observed by King *et al.*[1] stems mainly from the method of implicating genes based on CNVs, and is not an inherent property of ASD risk genes.

Methods

The list of ASD genes was constructed based on the SFARI Gene database (accessed 11 December 2013). We discarded genes with no or minimal support for association (score >3). Because of the difficulties to replicate genetic association and linkage results of ASD, candidate genes were considered only if the evidence was based on rare variants. The length of each gene was determined based on the largest transcript in the refSeq table of the UCSC Genome Browser (hg19 assembly). Genes were divided into CNVs or SNVs groups on the basis of the majority of studies that associated

©2014 Macmillan Publishers Limited. All rights reserved

the gene with ASD. In addition, we studied the length of genes found to be disrupted by *de novo* SNVs (nonsense, frameshift or splice site mutations)[5–8], and single genes affected by *de novo* CNVs[3,4]. To test for differences in the distribution of gene sizes in different groups we performed a Kruskal–Wallis rank sum test on all groups (using the kruskal.test function in The R Project for Statistical Computing), followed by a pair-wise Mann–Whitney-Wilcoxon Test (using the wilcox.test function in R). *P* values were adjusted for multiple tests using the Bonferroni correction.

Shahar Shohat[1] & Sagiv Shifman[1]
[1]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel.
email: sagiv@vms.huji.ac.il

Received 17 January; accepted 23 May 2014.

1. King, I. F. *et al.* Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501,** 58–62 (2013).
2. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4,** 36 (2013).
3. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (2011).
4. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
5. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
6. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
7. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
8. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
9. Ben-David, E. *et al.* Identification of a functional rare variant in autism using genome-wide screen for monoallelic expression. *Hum. Mol. Genet.* **20,** 3632–3641 (2011).

Author Contributions Both authors conceived the project, analysed the data and wrote the paper.

Competing Financial Interests Declared none.

doi:10.1038/nature13583

Zylka *et al.* reply

REPLYING TO S. Shohat & S. Shifman *Nature* **512,** http://dx.doi.org/10.1038/nature13583 (2014)

Shohat and Shifman's analysis[1] indicates that long autism spectrum disorder (ASD) genes are overrepresented in the SFARI Gene/AutDB database (as of 11 December 2013) owing to the discovery method. We agree with their analysis and with the need to consider the strength of evidence behind each candidate gene. When our study was underway[2], SFARI Gene provided the only comprehensive list of autism candidate genes with confidence values. Subsequent to our publication, more genes have been added to this database and scored, highlighting the rapid pace of advances in the ASD field and the changing confidence behind each ASD gene.

We agree with Shohat and Shifman that the proportion of ASD genes that are long may drop as more ASD genes are identified. We did not account for how the discovery method used to identify a given ASD candidate, be it based on copy number variant (CNV) or single nucleotide variant (SNV), might affect average gene length in our study. However, it is also undeniable that many long genes are considered candidates in ASD pathology, such as *NRXN1* and *CNTNAP2*. Moreover, our mechanistic findings are not in dispute. Indeed, three other groups came to the same conclusion as us—that topoisomerases preferentially facilitate expression of long genes[3–5]. Our study demonstrates an essential role for topoisomerases in transcriptional elongation of long neuronal genes and suggests a critical role for these enzymes in neurodevelopmental disorders like autism.

Shohat and Shifman[1] also suggest that the SFARI Gene database contains many genes with weak links to ASD pathology. In our study, we did not rank genes as stronger or weaker ASD candidates, and treated all equally. However, when the degree of evidence behind each candidate is taken into account, using the gene scoring module in SFARI Gene 2.0[6] (as of 1 April 2014), it remains clear that numerous strong ASD candidate genes are very long (>200 kilobases). Thus, we feel our conclusion linking topoisomerases and gene length with autism is still warranted, but this remains to be tested more rigorously pending *in vivo* studies with animal models and additional human genetic studies.

Future studies are likely to validate additional long genes as strong ASD candidates. For example, *NRXN3* and *CNTN5* (1.5 and 1.3 megabase, respectively) are not yet scored in SFARI Gene, yet these genes are deleted in patients with ASD[7,8] and both are significantly reduced in topotecan-treated neurons[2].

Ultimately, we agree that making conclusions about the nature of ASD genes is complicated by factors like the ones Shohat and Shifman describe[1] as well as by the evolving knowledge of autism genetics. Regardless, we identified a transcriptional mechanism that affects the expression of long genes, a number of which are currently classified as strong ASD candidates.

Mark J. Zylka[1,2,3], Ben D. Philpot[1,2,3] & Ian F. King[1]
[1]Department of Cell Biology and Physiology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
[2]Carolina Institute for Developmental Disabilities, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
[3]UNC Neuroscience Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
email: zylka@med.unc.edu

1. Shohat, S. & Shifman, S. Bias towards large genes in autism. *Nature* **512,** http://dx.doi.org/10.1038/nature13583 (2014).
2. King, I. F. *et al.* Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501,** 58–62 (2013).
3. Solier, S. *et al.* Transcription poisoning by topoisomerase I is controlled by gene length, splice sites, and miR-142–3p. *Cancer Res.* **73,** 4830–4839 (2013).
4. Teves, S. S. & Henikoff, S. Transcription-generated torsional stress destabilizes nucleosomes. *Nature Struct. Mol. Biol.* **21,** 88–94 (2014).
5. Veloso, A. *et al.* Genome-wide transcriptional effects of the anti-cancer agent camptothecin. *PLoS ONE* **8,** e78190 (2013).
6. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4,** 36 (2013).
7. Vaags, A. K. *et al.* Rare deletions at the neurexin 3 locus in autism spectrum disorder. *Am. J. Hum. Genet.* **90,** 133–141 (2012).
8. van Daalen, E. *et al.* Social responsiveness scale-aided analysis of the clinical impact of copy number variations in autism. *Neurogenetics* **12,** 315–323 (2011).

doi:10.1038/nature13584

©2014 Macmillan Publishers Limited. All rights reserved