Human Molecular Genetics, 2011, Vol. 20, No. 18 **3632–3641**
doi:10.1093/hmg/ddr283
Advance Access published on June 16, 2011

Identification of a functional rare variant in autism using genome-wide screen for monoallelic expression

Eyal Ben-David[1], Einat Granot-Hershkovitz[1], Galya Monderer-Rothkoff[1], Elad Lerer[2], Shlomit Levi[3], Maya Yaari[4], Richard P. Ebstein[4], Nurit Yirmiya[4] and Sagiv Shifman[1,*]

[1]Department of Genetics, The Institute of Life Sciences, [2]Department of Human Genetics, [3]Department of Psychiatry, Hadassah Medical School and [4]Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel

Received February 28, 2011; Revised June 5, 2011; Accepted June 13, 2011

Recent work has led to the identification of several susceptibility genes for autism spectrum disorder (ASD) and an increased appreciation of the importance of rare and *de novo* mutations. Some of the mutations may be very hard to detect using current strategies, especially if they are located in regulatory regions. We present a new approach to identify functional mutations that exploit the fact that many rare mutations disrupt the expression of genes from a single parental chromosome. The method incorporates measurement of the relative expression of the two copies of a gene across the genome using single nucleotide polymorphism arrays. Allelic expression has been successfully used to study common regulatory polymorphisms; however, it has not been implemented as a screening tool for rare mutation. We tested the potential of this approach by screening for monoallelic expression in lymphoblastoid cell lines derived from a small ASD cohort. After filtering regions shared across multiple samples, we identified genes showing monoallelic expression in specific ASD samples. Validation by quantitative sequencing demonstrated that the genes (or only part of them) are monoallelic expressed. The genes included both previously suspected risk factors for ASD and novel candidates. In one gene, named autism susceptibility candidate 2 (AUTS2), we identified a rare duplication that is likely to be the cause of monoallelic expression. Our results demonstrate the ability to identify rare regulatory mutations using genome-wide allelic expression screens, capabilities that could be expanded to other diseases, especially those with suspected involvement of rare dominantly acting mutations.

INTRODUCTION

There is growing evidence that rare and *de novo* mutations may constitute a large proportion of susceptibility variants for autism spectrum disorder (ASD) (1,2). This implies that the disease can be triggered by mutations in many different genes, and may explain why no unifying structural or neuropathological features have been conclusively identified. However, the detection of rare and *de novo* causal variants presents challenges to current genetic mapping strategies (3).

One currently favored approach is direct sequencing of large samples of cases and controls. This strategy has become feasible with advances in sequencing technologies, but it is difficult to identify the disease-causing mutations and to distinguish them from the enormous number of non-functional sequence variations in the genome. This is especially true when the same phenotype may arise from many different rare variants. Therefore, current efforts focus on coding mutations with clear functional effect, leaving out mutations affecting regulatory regions as well as epigenetic mutations (4).

One method that potentially could be used to identify rare regulatory mutations is expression profiling. Using expression arrays, one could try to identify genes that are differently expressed in specific ASD sample relative to controls. However, studies comparing the expression profiles of cases and controls suffer from low statistical power due to genetic heterogeneity, difficulty accessing appropriate tissues, small

*To whom correspondence should be addressed at: Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra campus, Jerusalem 91904, Israel. Tel: +972 26585396; Fax: +972 26586975; Email: sagiv@vms.huji.ac.il

sample size and the small differences between cases and controls. In addition, gene expression data alone do not distinguish between changes that constitute primary etiology and those that reflect secondary pathology, compensatory mechanisms or confounding influences (5,6).

We suggest a different approach to identify rare mutations, which is based on measuring the relative expression of a gene from the paternal and maternal copies (7). Mutations may produce strong allelic expression imbalance (AEI) and may even lead to monoallelic expression. We propose that in diseases in which rare variation plays a considerable role, a proportion of the genes involved may be marked by monoallelic expression. Allelic expression analysis has been widely used to study the functional role of *a priori* known mutations in genes for monogenic diseases (8). We propose traveling along the opposite route, searching for genes that are expressed in a monoallelic way and then screening for genetic or epigenetic mutations that may underlie the monoallelic expression. In such an approach, allelic expression screening is used to expose abnormalities in gene expression caused by rare variants, newly arisen mutations or epigenetic alterations.

Here we investigate the feasibility of this approach by simulation and by studying the allelic expression landscape in lymphoblastoid cell lines (LCLs) derived from an autistic cohort. We show that the main limitation of using LCLs for this method is the tendency of such cell lines to be composed of a small number of clones or even to be monoclonal (9). In such oligoclonal cell lines, the occurrences of random monoallelic expression may be evident. Despite that, we were able to identify several genes that are candidates for being associated with autism. The candidate genes show monoallelic expression in specific ASD individuals, while showing biallelic expression in the rest of the cohort, as well as in LCLs derived from healthy individuals, examined in a previous study. In some cases, the monoallelic expressed region was restricted to only parts of the gene, strengthening the notion that the observed monoallelic expression in these genes may be abnormal. In addition, the list of candidates included genes that were previously suggested to be associated with autism. To search for possible genetic causes of the monoallelic expression, we identified copy number variations (CNVs). In one candidate gene, the CNV analysis revealed a duplication that is the most plausible cause of the monoallelic expression in the same subject, providing further support for the potential of the approach.

RESULTS

Methodological overview including simulation for gene coverage

To screen for monoallelic expression across the genome, we used single nucleotide polymorphism (SNP) arrays, modifying the protocol to test RNA rather than DNA, as previously described (10). Our method requires sequence variants that differentiate the two allelic transcripts. Because relatively few transcripts contain coding SNPs that can be used in this way we designed our assay to interrogate intronic RNA by enriching for nuclear pre-mRNA. An overview of the approach can be seen in Supplementary Material, Figure S1 and a more detailed description is provided in the Materials and Methods. In brief, we excluded SNPs that are homozygote in genomic DNA (gDNA), or were determined by us not to be expressed. For the remaining SNPs, a normalized measure of the deviation from heterozygosity was calculated for each SNP in the cDNA. We performed a rolling window analysis in order to identify regions of monoallelic expression. Adjacent regions showing similar effects were merged together.

A SNP is informative for allelic expression assay only if the sample is heterozygous at the gDNA level. The degree of heterozygosity varies between individuals and between different genomic regions, and so the coverage of each gene changes in different samples. To estimate the theoretical coverage level across the entire transcriptome, we performed simulations based on genotyping data from the HapMap database, for two genotyping platforms containing around one million SNPs: Affymetrix SNP 6.0 and Illumina 1M-Duo. We permuted HapMap-phased haplotypes to assemble 1000 diplotypes. For each diplotype, we counted the number of heterozygous SNPs in each transcript. We proceeded to establish, for each transcript, the number of diplotypes harboring at least one heterozygous SNP (Supplementary Material, Fig. S2). For the Affymetrix SNP 6.0 platform, 73% of all transcripts had more than a 50% probability of being informative and 30% of all transcripts had more than an 80% probability of being informative. With the Illumina 1M-Duo, 80% of all transcripts had more than a 50% probability of being informative and 38% of all transcripts had more than an 80% probability of being informative.

In general, there are more SNPs in larger genes, thus the size of the gene, including the introns, is correlated with the probability of it being informative in a particular sample. Brain-specific genes tend to be larger and thus harbor more SNPs. Accordingly, when limiting the simulation to brain exclusive transcripts (see Materials and Methods), 83% for Affymetrix SNP 6.0 and 89% for Illumina 1M-Duo had more than a 50% probability of being informative and 49 and 64%, respectively, of all transcripts had more than an 80% probability of being informative. When all genotyped SNPs from HapMap were included, 70% of all brain exclusive transcripts had more than an 80% probability of being informative. With deep sequencing, both rare and common SNPs in heterozygote states can be assayed, further increasing the coverage and the potential of this method to screen the entire transcriptome.

Genome-wide patterns of common monoallelic expression

One of the obstacles in using allelic expression to detect rare regulatory mutations is that monoallelic expression is also associated with normal gene function. In addition to imprinting and X-inactivation, monoallelic expression is widespread in monoclonal cell lines (10). In an attempt to avoid the occurrences of random monoallelic expression, we tested all LCLs used in this study for clonality (Supplementary Material, Fig. S3). We used polymerase chain reaction (PCR) amplification of the V-D-J junctions to detect clonal immunoglobulin heavy chain (IgH) gene rearrangements. Cell lines showing evidence for monoclonality were excluded. We selected 17

LCLs isolated from blood of autistic individuals that based on the IgH assay were not clonal. Our SNP array measurements allowed us to take an additional approach to assaying the clonality of our sample. In general, clonal cells are expected to show a high degree of monoallelic expression for genes on the X chromosome, as the same X chromosome in all cells is inactivated. In polyclonal cells, each cell is expected to have a different X chromosome inactivated, resulting in biallelic expression. A cell line could also be oligoclonal (i.e. made of small number of different clones), resulting in a random subset of SNPs that will show strong AEI. While random X inactivation appears a good marker for polyclonality, skewed X chromosome inactivation has been often observed in females carrying mutations involved in X-linked syndromes or with autism (11,12). This limits our ability to confidently describe a sample as monoclonal based on degree of monoallelic expression for genes on the X chromosome. Out of the 17 LCLs, 5 originated from female cases, enabling us to test the directionality of X inactivation in these samples. We tested the proportion of heterozygote genotype calls in the gDNA remaining so in the cDNA. We ran this analysis across the different chromosomes, while focusing specifically on the X chromosome (Supplementary Material, Fig. S4). This provided us with a crude measurement of the AEI trend in the genome. Despite the large data set, comprising of hundreds of thousands of heterozygote SNPs in each individual, we found this rate to be highly variable between the samples, with the values in the X chromosome being generally lower while congruous with the autosomal trend. However, one female individual (Family id #25) exhibited a drastically lower concordance rate in the X chromosome compared with the autosomes. This sample was not excluded, as this may reflect, rather than a state of monoclonality, a state of skewed X chromosome inactivation.

To identify common regions and genes showing deviation from equal allele expression, we developed an algorithm to search for common regions showing AEI across different LCLs. The algorithm has increased power to identify common AEI regions as it takes into account all the samples together when searching for significant AEI. Permutation was used to determine an appropriate significance threshold. We identified 756 regions that included 1072 refSeq genes showing a significant deviation from equal allele expression, but not necessarily monoallelic (Supplementary Material, Table S1). Seventeen regions (2.2%) were within the known or predicted imprinted regions. Supplementary Material, Table S1 shows imprinted genes (known or suspected) that exhibit common AEI in our samples. We tested two genes showing strong AEI (*KALRN, ZNF365*) within regions of common AEI by quantitative sequencing of cDNA. As predicted, both genes were found to have monoallelic expression (Supplementary Material, Fig. S5). We generated a consensus list of genes showing very strong AEI or complete monoallelic expression across studies and samples, by comparing our results with data from HapMap LCLs that were previously published (13). This list contains 106 genes (Supplementary Material, Table S1), and includes known imprinted genes and other unknown genes that may include genes with a tendency for random monoallelic expression. Four genes have been previously associated with autism (*CNTNAP2, CNTN4, A2BP1, MDGA2*): mutations affecting one copy of these genes were previously identified in autistic cases (1,14–16). Similarly to imprinted regions, genes with random monoallelic expression may be extremely vulnerable to mutations disrupting one copy of the gene because if it is randomly monoallelic expressed the gene may be totally inactive in some cells.

Identification of monoallelic expression in autistic individuals

We reasoned that monoallelic expression that is connected with ASD would be expected to be relatively rare. Therefore, it is expected to appear in only one individual in our small cohort and also not in any LCLs from normal individuals. This strategy also enabled us to filter out possible array-specific artifacts that are expected to be common across samples. Thus, we proceeded to search for regions showing monoallelic expression in individual samples, in addition to evaluating the results against data published previously (10,13). After filtering for windows showing evidence for monoallelic expression in multiple samples, in previously described control data set and windows that are not restricted to the position of one gene, seven genes showed evidence for AEI with a $P < 1 \times 10^{-4}$, which corresponds to a false discovery rate (FDR) < 0.05. Table 1 shows the seven genes identified as unique events in LCLs from one individual with autism. All seven genes were subject to quantitative sequencing of gDNA and cDNA. Six (86%) were found to be completely monoallelic or with extreme AEI; one had moderate AEI (*BCAT1*) (Table 1). We examined the status of the seven regions in data from HapMap LCLs. For each gene, there were between 21 and 52 informative individuals, with more than 5 SNPs in heterozygote state. None of the HapMap samples showed monoallelic expression, or strong imbalance in these regions (13) [the average difference between gDNA and cDNA ratios for all samples (Δhet) was smaller than 0.2] (Table 1). Out of the seven genes, mutations in two had been previously identified as possible risk for autism: *AUTS2* and *DPYD* (17,18). To study the likelihood of observing unique AEI regions in unaffected control samples, we analyzed the allelic expression data from the 53 HapMap samples, applying a similar approach to the one we used with the affected individuals. We again observed many regions showing evidence for monoallelic expression, including previously known imprinted regions. Using the same criteria as above, we identified seven genes that show evidence for significant AEI, each unique to one HapMap sample. Permutation test showed that the rate of rare monoallelic expressed genes in the unaffected LCLs (rate = 0.13) is significantly lower ($P = 0.03$) than the rate observed in LCLs from autistic cases (rate = 0.41). In addition, none of the seven genes identified in the HapMap cell lines has been previously linked to autism.

Complex patterns of monoallelic expression

Four of the seven tested genes showed a complex pattern of AEI based on the SNP arrays, with variation in the AEI across the gene (i.e. *ADARB2, DOK6, AUTS2* and *Kif16B*)

Table 1. Genes showing AEI in ASD samples

Gene	Chr	Start[a]	End[a]	No. of SNPs[b]	Family id[c]	*P*-value[d]	Mean AEI ratio[e]	Informative HapMap LCLs[f]
NCALD	8	103035324	103256431	17	5	3.70E−06	∞	52
DPYD	1	97059098	98436514	21	115	4.20E−06	76.9	50
DOK6	18	65456402	65649653	19	115	8.40E−06	2.8–10.1	46
BCAT1	12	24861855	24983741	18	25	1.50E−05	1.9	32
AUTS2	7	69694711	69932259	18	13	4.80E−05	∞	37
ADARB2 (NCRNA00168)	10	1525739	1676615	21	26	5.10E−05	∞	21
KIF16B	20	16205484	16412676	17	112	8.80E−05	∞	52

[a]Start and end positions of the regions showing AEI.
[b]Number of informative SNPs within each region.
[c]The id number of the family of the ASD proband.
[d]Nominal *P*-values for the AEI regions.
[e]The mean ratio between alleles based on quantitative sequencing, with ∞ denoting complete monoallelic expression. Due to background in sequencing reaction, in some cases, lower values were obtained even for monoallelic regions.
[f]Number of HapMap LCLs with more than five informative SNPs in the indicated region. All HapMap LCLs were biallelic (13).

(Figs 1 and 2). In *ADARB2* (also known as *ADAR3*), the 150 kb window that showed AEI was within the first intron, encompassing a non-coding RNA gene (*NCRNA00168*) that is transcribed on the opposite strand, whereas the rest of the gene showed biallelic expression. Sequencing of the cDNA validated the monoallelic expression of this RNA gene. For *DOK6*, the AEI was quantitatively more extreme towards the 5′ end of the gene. This was also evident in the quantitative sequencing: a SNP in the seventh and last intron exhibited a ratio between alleles of 2.8, while a SNP in the fifth intron exhibited a larger ratio of 10.1 between alleles. Sequencing of cDNA generated with *DOK6*-specific primers confirmed that this effect is not caused by transcription, on the opposite strand, of a nearby gene (*CD226*). Two other genes (*AUTS2, KIF16B*) harbored a region of monoallelic expression at the 3′ part of the gene but showed biallelic expression at the 5′ of the gene. We typed additional SNPs to validate these results and to fine-map the exact region where the biallelic expression was reduced to monoallelic expression. In total, we typed 10 SNPs in *KIF16B* and 4 SNPs in *AUTS2*, in both gDNA and cDNA. For *KIF16B*, consistent with the array results, a SNP in the first intron was biallelically expressed with a small allelic imbalance (allele ratio = 2.0), whereas SNPs in the rest of the gene were monoallelic (Fig. 1). Similarly, for *AUTS2*, the region predicted to be monoallelic based on the array results starts at the middle of the fifth intron towards the end of the gene (Fig. 2). Consistent with the array results, two of the SNPs in the fifth intron, close to the fifth exon of the gene, were biallelic (median allelic ratio = 1.16), whereas two SNPs near the sixth exon, one in the fifth intron and the other in the sixth intron, were both monoallelic expressed (Fig. 2B and C).

CNV and monoallelic expression

One possible cause for AEI in autistic samples can be mutations, including CNVs. As each sample in our study was subject to two SNP arrays, including one with gDNA, we were able to analyze the arrays data for the presence of CNVs. To identify rare CNVs, we filtered out CNVs according to overlap with regions reported in the Database of Genomic Variants, and the CNV project at the Children's Hospital of Philadelphia (CHOP), which includes high-resolution mapping of CNVs in 2026 healthy individuals. After filtering and quality control, we identified 11 CNVs (Table 2, Supplementary Material, Fig. S6). Eight CNVs were also tested with Multiplex Ligation-dependent Probe Amplification (MLPA) assay, with significant agreement between the calls of the array and the MLPA assay (Supplementary Material, Fig. S7). Six genes, in five of the CNVs, were previously reported to be affected by rare CNVs in individuals with autism. Three of the genes were reported to be affected by CNVs or translocations in multiple cases (*NRXN1, SLC9A9* and *AUTS2*) (19–24) and three had only one report (*SPTBN4*, *SHKBP1* and *EXOC4*) (2). Two genes (*NRXN1* and *EXOC4*) had deletions that are highly likely to result in functional impact at the protein level. In *EXOC4*, the deletion included two exons, number 6 and 7. The deletion in *NRXN1* was the largest identified CNV. This deletion, 1.8 Mb in length, is predicted to generate a fusion gene between the first two exons of *NRXN1* and the last six exons of the adjacent gene, *FSHR*. The subject carrying the *NRXN1* deletion was reported to have a balanced translocation involving the same chromosome (46, XY, t(2,5)(p25.1;q33.1). To assess the likelihood that these CNVs are pathogenic, we first compared them to CNVs that were previously published for 3181 control individuals using the same Affymetrix platforms (25). Although the same CNVs were not reported in the control sample, other types of CNVs were found in the following genes: *NRXN1, FSHR, ERC1, EXOC4* and *AUTS2* (25). To further examine the CNVs functionality, we reexamined their allelic expression status. For only five genes affected by CNVs, we had good coverage (more than five informative SNPs) to identify monoallelic expression (*WDFY1, SLC9A9, ERC1, EXOC4* and *AUTS2*). Two genes showed evidence for expression imbalance, one of them as stated above is *AUTS2*. In *AUTS2*, a duplication of 140 kb, encompassing the fifth exon, was identified in the same individual exhibiting AEI (Fig. 2). The duplication is located upstream of the region showing monoallelic expression (Fig. 2A and B). We analyzed the total expression levels of heteronuclear RNA in this subject using the intensities of polymorphic and non-polymorphic probes (used normally to study CNVs) that are present on the SNP 6.0 Affymetrix array (Fig. 2B). Consistent

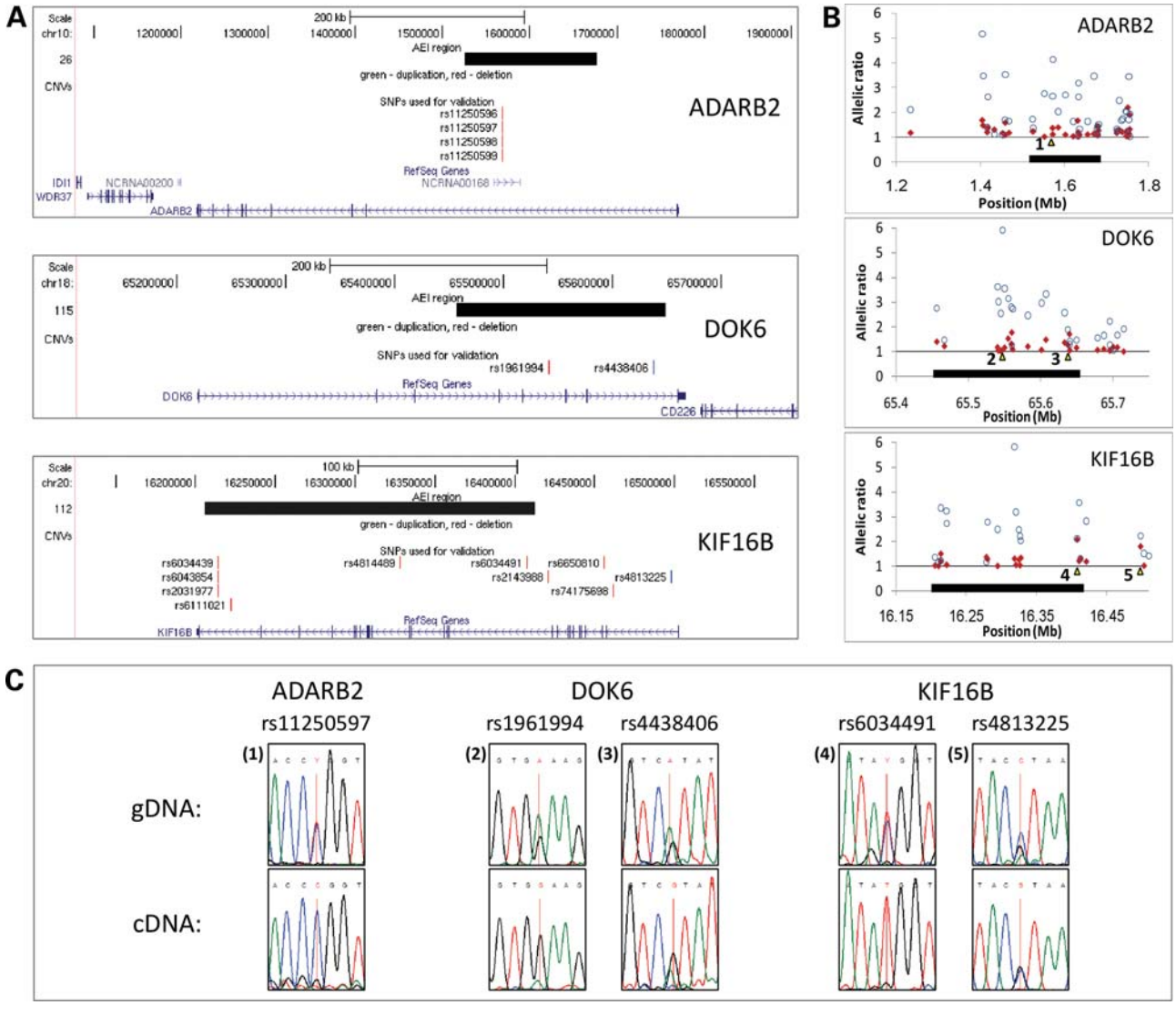

Figure 1. Complex pattern of allelic expression in *ADARB2*, *DOK6* and *KIF16B* genes. (**A**) UCSC genome browser images showing the position of the AEI region (black bar), SNPs used for validation (red marks are monoallelic SNPs and blue marks are biallelic SNPs) and RefSeq gene track. (**B**) The ratio of the raw microarray intensity data between alleles of informative SNPs. The ratio is the maximum value obtained by dividing the intensities of allele A by allele B or vice versa. The ratio for cDNA is presented in blue circles and in red diamonds for the gDNA. The marked yellow triangles show the position of the sequences presented in (**C**). The black bar represents the region showing monoallelic expression detected by the window-based analysis. (C) Sequencing results for gDNA and cDNA. The numbers on the left of the sequences correspond to the numbers of the marked triangles in (B).

with the allelic expression data, the total expression level diminished instantly downstream to the duplicated region. Comparing the alleles of the heterozygous SNPs, within the duplication and the region of monoallelic expression, with the genotypes of the parents, revealed that the duplication and the partially expressed copy of the gene both originated from the maternally derived chromosome (Fig. 2C). This suggests that the duplication has stalled *AUTS2* transcription. We analyzed both parents and the proband's sister for the presence of the duplication; both the mother and the sister carried the duplication (Supplementary Material, Fig. S7). The proband was reported to have severe intellectual disability and epilepsy, in addition to a full diagnosis of autism. The mother was reported to have mild intellectual disability and the sister was reported to have developmental delays but not autism, whereas the father was reported not to show any abnormal behavioral or cognitive phenotype. We compared the total expression of the *AUTS2* gene between the proband and his father using Affymetrix GeneChip Human Gene 1.0 ST Array. We exploited the existence of probes in each exon to study the expression of the gene across exons, which mainly represent the expression at the level of the mRNA. While the total expression of the gene was not considerably different, the first four exons showed higher expression in the proband and the last 15 exons showed higher expression in the father (Fig. 2D). This result is consistent with a truncated expression of the maternal copy of *AUTS2*. Out of the other genes that were affected by CNVs and were informative for AEI analysis, only one other gene (*ERC1*) was showing possible evidence for AEI (nominal

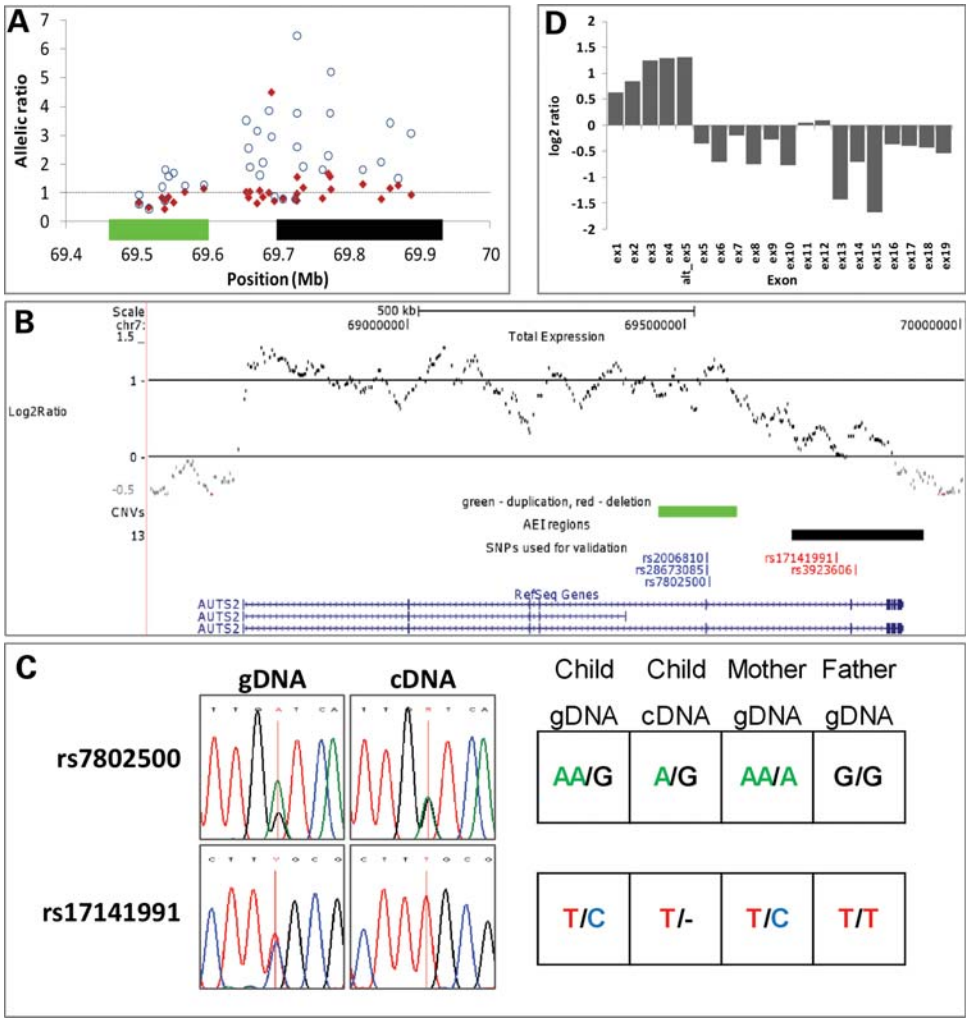

Figure 2. Monoallelic expression apparently derived by duplication in the *AUTS2* gene. (**A**) The ratio of the raw microarray intensity data between the paternal and maternal alleles of informative SNPs. The ratio for cDNA is presented in blue circles and in red diamonds for the gDNA. The green bar represents the duplicated region and the black bar the region showing monoallelic expression detected by the window-based analysis. (**B**) UCSC genome browser images showing the total expression of heteronuclear RNA based on the normalized signals of polymorphic and non-polymorphic probes. The intensities were normalized using Affymetrix genotyping console and are presented relative to a reference panel baseline intensity obtained with gDNA. Each dot is a moving average of the log2 ratio across 16 successive probes. Below are the position of the AEI region (black bar), duplicated region (in green), SNP used for validation (red marks are monoallelic SNPs and blue marks are biallelic SNPs) and RefSeq gene track. (**C**) On the left is an example of the sequencing validation results of two SNPs, rs7802500, located within the duplication and rs17141991 located within the region showing monoallelic expression. The red vertical line indicates the position of the SNP. The table on the right shows the genotyping of the proband (gDNA and cDNA) and his parents (gDNA only). The table shows that both the duplicated allele (allele A) and the non-expressing allele (allele C) are inherited from the mother. (**D**) The expression ratio (log2) between the proband and his father for different exons of *AUTS2*. Consistent with a truncated transcription of one copy of the gene in the proband, exons 1–4 and the alternative exon 5 show higher expression in the proband (average ratio = 2.1) and exons 5–19 show higher expression in the father (average ratio = 1.5).

$P = 0.00013$) in the same subject carrying the CNV (a 9 kb intronic deletion in either the 10th, 13th or 14th intron, depending on the splicing isoform). Validation attempts using quantitative sequencing revealed a monoallelic SNP only in the first intron of the *ERC1* gene, which was not consistent with the results of the array and was far from the deletion region (data not shown).

DISCUSSION

AEI can result from different types of epigenetic or genetic variation between the two copies of the gene. So far, allelic expression has been mainly studied in the context of regulatory polymorphisms, imprinted genes or random monoallelic expression (26). *Cis*-acting common regulatory variants are usually associated with a small degree of AEI, which is observed across different samples (27). In contrast, classical imprinted genes show monoallelic expression across samples in a parent-of-origin manner (26). Random monoallelic expression is apparent in single or clonal cells, in different directions across different cells (10). We explored another possible mechanism for monoallelic expression, which is the effect of rare functional mutations. Monoallelic expression caused by rare mutations is expected to be unique to the individual that carry the mutation, and depending on the type and location of the mutation may be restricted to only part of the gene.

Table 2. CNVs in ASD samples

Family id no[a]	CNV state	Copy number	Chr[b]	CNV Size (kb)	Start[c]	End[c]	Genes in CNV region	Inheritance[d]
15	Loss	1	2	1822	49131265	50953285	*NRXN1, FSHR*	NA
119	Gain	4	2	23	224440369	224462934	*WDFY1*	P[e]
115	Loss	1	3	12	144782914	144794876	*SLC9A9*	P
112	Gain	3	20	109	754677	863863	*FAM110A, ANGPT4*	NA
52	Gain	3	19	15	45762682	45777854	*SPTBN4, SHKBP1*	NA
26	Gain	3	9	129	5606095	5734973	*KIAA1432*	P
26	Loss	1	12	9	1324438	1333073	*ERC1*	NA
10	Loss	1	7	75	132668829	132744296	*EXOC4*	P
10	Gain	3	X	171	109805	281199	*PPP2R3B, PLCXD1, GTPBP6*	M
13	Gain	3	3	130	180619034	180748723	*GNB4*	M
13	Gain	3	7	140	69455261	69595511	*AUTS2*	M

[a]The id number of the family of the ASD proband.
[b]Chromosome number.
[c]Start and end positions of CNV.
[d]Estimated mode of inheritance: M, maternal; P, paternal.
[e]The parent has duplication and the child triplication.

In this study, we developed a method to identify rare genetic variation by analyzing monoallelic expression across the genome. One of the major purposes of this work was to test the feasibility of this approach to identify rare mutations. Despite several limitations of the current study, including the small sample size and the use of LCLs, we were able to demonstrate the potential of this approach. There are two main disadvantages for using LCLs for identifying autism-associated mutations. First, many brain-specific genes cannot be analyzed. Secondly, the low number of clones composed in each cell line may produce a bias toward one allele. Since random monoallelic expression is expected to be observed in multiple cell lines, including cell lines from non-autistic individuals, we concentrated on cases where the monoallelic expression was only observed in one specific LCL from an affected individual. We propose that future studies using this method should attempt to use tissues or primary culture.

Among the genes that showed monoallelic expression in ASD samples, the most convincing case for monoallelic expression that reflects a rare genetic variant was for *AUTS2*. *AUTS2* was first identified by a study of a monozygotic twin pair concordant for autism with an identical balanced translocation in a novel gene (23). Since the first study, *de novo* translocations and an inversion truncating the *AUTS2* gene were reported in five additional unrelated subjects with one or more of the following symptoms: autism, mental retardation and epilepsy (18,22,28). Using our method, we were able to show that a duplication of the fifth exon can result in a similar truncated transcript, exemplifying the power of the method in moving from a biological phenotype to its probable genetic cause. The function of this gene is still unknown, but recent studies show that the gene is expressed in developing mouse brain, and may have a critical role in the development of cortical regions (29,30).

In summary, the method described in this study, which is based on genome-wide allelic expression analysis, can be used to identify different types of rare genetic or epigenetic mutations affecting gene expression. It can be also used to test the functional effect of mutations identified by other means and to provide new insight into genetic or epigenetic mechanisms. This method could be expanded to additional complex or monogenic diseases, especially ones with suspected involvement of rare variation. Furthermore, the increasing feasibility of whole genome sequencing suggests that large screens could be performed with deep sequencing in the future. Not only would deep sequencing increase the coverage of the screen, more importantly allelic expression can provide a biological phenotype to help identify non-coding risk variants in sequencing studies.

MATERIALS AND METHODS

Participants

LCLs derived from 17 subjects with ASD (12 males, 5 females) were included. A detailed description of the entire cohort was previously described (31). Subjects were diagnosed with DSM IV autistic disorder ($n = 14$) or pervasive developmental disorder-not otherwise specified (PDD-NOS; $n = 3$), using the Autism Diagnostic Interview—Revised (ADI-R), and the Autism Diagnostic Observation Scale—Generic (ADOS-G). The subjects' level of functioning was assessed using standard intelligence measures selected according to the subjects' age and abilities and a standard daily living skills interview. Average IQ scores of subjects included in this study were 51.2 (SD = 28.3).

DNA and RNA extraction

gDNA was extracted from the cell lines with QIAamp DNA Blood Mini Kit (Qiagen). Nuclei were isolated using Nuclei EZ Prep kit NUC101 kit (Sigma) according to the manufacturer's protocol. The nuclei were isolated from 3 to 20 million cells. Isolated nuclei were stored in NucleiPure storage solution (Sigma) at -80°C. Total RNA was extracted from the nuclei using TRI reagent (Sigma), and then diluted appropriately and treated with Turbo DNAfree (Ambion) according to the manufacturer's protocol for 'strong DNA contamination'. Absence of DNA in the RNA samples was

verified by PCR amplification of at least 100 ng of RNA with primers for an intergenic region. The DNA and RNA samples were quantified using the NanoDrop ND-1000 UV-Vis spectrophotometer (NanoDrop Technologies), and the integrity of the RNA was examined by capillary electrophoresis with a Bioanalyzer using RNA 6000 Nano Labchips (Agilent). The DNA-free total RNA was converted to double-stranded cDNA as previously described (32). Briefly, reverse transcription of 5 μg of RNA (37°C, Superscript II; Invitrogen) was followed by second-strand synthesis using a mix of DNA polymerase I, DNA ligase and RNAse H (BioLabs). The resulting DNA was cleaned up by phenol-chloroform extraction and ethanol precipitation and resuspended in DEPC dH_2O. cDNA samples were adjusted to a final concentration of 100 ng/μl. In all other respects, the processing of cDNA samples for the SNP arrays was identical to processing of gDNA samples and performed according to Affymetrix guidelines for the SNP 6.0 arrays. Global gene expression was performed on total RNA using Affymetrix Gene ST1.0 microarray according to the manufacture protocol, analyzed using Affymetrix Expression Console.

Clonality analysis

To assess clonality, IgH gene rearrangement was examined. PCR amplification of V-D-J junctions was performed using published primers targeting framework 3 VH region (FRIII) (33). NALM-6 lymphoma cell line DNA served as the positive control for monoclonal IgH gene rearrangement, while DNA from whole blood served as the polyclonal control. gDNA from the cell lines was amplified by PCR using a semi-nested protocol as previously described (34). Briefly, first- and second-round reactions contained 0.4 μM of each primer, 2.5 mmol/l $MgCl_2$, 200 μmol/l dNTP, 0.0002 U of Taq DNA polymerase with 1× buffer (Qiagen HotStarTaq) in a 10 μl reaction. PCR cycling was performed at 95°C for 15 min for one cycle, followed by 35 cycles at 95°C for 30 s, 58°C for 30 s and 72°C for 30 s. The final cycle was followed by a 10 min extension at 72°C. Agarose gel (4%) electrophoresis was used to analyze the PCR products. Samples containing polyclonal B cells exhibited a smear of bands between 80 and 120 bp, whereas samples containing monoclonal B cells exhibited one or two crisp bands in the size of 80–120 bp (representing mono- or biallelic rearrangement). Samples exhibiting three or more crisp bands in the size between 80 and 120 bp were considered to be oligoclonal.

Coverage estimation

Our method depends on the existence of heterozygous SNPs inside transcripts, and hence we tested its transcriptome-wide coverage. Genotyping data for 60 CEU trios from the Hapmap project (retrieved from the Hapmap website, releases 2.2 and 3) were phased using Beagle 3.0.4 software (35). We sampled 1000 diplotypes randomly from the resulting haplotype pool, and estimated the probability of heterozygosity in each transcript by calculating the number of diplotypes which held a heterozygous SNP in each transcript. Measures were calculated for all transcripts, as well as for the subset of transcripts which are brain exclusive. Brain exclusive transcripts were determined by two different methods. First, the Absent/Present calls from U133A chip from the Brain Atlas 2 project (Geo Accession: GSE1133) were used to locate genes in which an absent call was determined for all tissues apart from the brain, and a present call was determined for at least one of the brain tissues. The second approach, used by Shi *et al*. (36), was sifting the same data, this time the quantitative expression values, for genes in which the highest expression was in one of the brain tissues, and it was twice as high as any non-brain tissue. Both methods yielded the same list of genes.

Determining regions of AEI in individual samples

SNP genotyping was performed with the Birdseed v2 algorithm (37) implemented in the Affymetrix Genotyping Console. The SNP genotype calling was done separately on gDNA and cDNA. We filtered out SNPs which were homozygote in the gDNA, or were determined by us not to be expressed (combined signal for both alleles <3000). All SNPs passing intensity filter were included even if not in annotated genes. For informative SNPs, a normalized distance from heterozygosity was calculated for each SNP in the cDNA, using the confidence score from Affymetrix Birdseed V2 genotyping algorithm (37). This measure is distributed between 0 and 1, with lower values corresponding with higher confidence, so we subtracted it from 1 in our filtered SNPs, resulting in a measure for distance from heterozygosity. For the data from HapMap LCLs, we used Δhet (13) as the distance from heterozygosity. Distances were transformed to a uniform distribution of probabilities by ranking the confidence score among SNPs, and dividing it by the total number of SNPs (so for instance, the score corresponding with the largest distance from heterozygote state would be 1 divided by the number of SNPs). On the resulting uniform distribution of probabilities, a statistical analysis could be performed. As AEI might not be limited to within the boundaries of a single gene, we performed a rolling window analysis to identify regions of monoallelic expression. For each window of five consecutive and informative SNPs, we calculated a combined *P*-value over the probabilities for all SNPs within the window, based on Fisher's method and a Chi-square distribution (given, for a window of k SNPs, by $\chi^2 = -2\sum_{i=1}^{k} \log_e(p_i)$). We filtered out windows which included SNPs with over 1 Mb between them. Consecutive windows showing AEI ($P < 0.05$) were combined if the distance between them was less than 1 Mb. For the resultant merged window, Fisher's statistic and corresponding combined *P*-value was recalculated on all SNPs within the merged window. To calculate an empirical FDR under the null hypothesis of no AEI, we counted the number of *P*-values below different thresholds across all samples in 100 simulations that permuted randomly the rank of all informative SNPs for each sample, followed by window-based analysis as shown above. To test for enrichment of the rate of unique AEI regions in LCLs from autistic cases relative to HapMap LCLs, we used the global test of CNV burden in cases versus controls implemented in PLINK. Significance was assessed with permutations (100 000 permutations, one-sided test). All other statistical analyses were performed using the R project for statistical

computing (http://www.r-project.org). Window analysis was performed using the R 'zoo' package.

Widespread AEI loci detection

To establish widespread AEI, we integrated the SNP data for all individuals in our autism sample into a combined measure. First, expressed SNPs were filtered, and their confidence score normalized, according to the criteria supplied above. For each SNP, a combined probability measure was calculated by applying Fisher's method (see above) across all heterozygote individuals. This resulted in a single estimate per SNP of its distance from heterozygosity across all the different individuals. A rolling window analysis was then performed as described above. A significant threshold was arrived at by permutation testing. To generate a consensus list of monoallelic expressing genes, we compared the list of genes showing a significant AEI in our study with a list of genes that show a very strong AEI also in HapMap LCLs (13). First, for each transcript, we identified the informative individuals in this data set. Informative individuals were defined as those with more than three informative SNPs, SNPs which were expressed and heterozygote, as determined by the criteria described in Ge *et al.* (13). For each informative individual, the mean Δhet of the transcript was calculated. Transcripts in which at least one individual showed strong AEI (Δhet >0.3) and in which at least half of the informative individuals showed at least a modest AEI (Δhet >0.1) were determined to harbor evidence for monoallelic expression.

Monoallelic expression validation

Heterozygous SNPs showing monoallelic expression were validated by sequencing of PCR products amplified from gDNA and cDNA samples, with primers designed using Primer 3.0 (http://frodo.wi.mit.edu/primer3/). Touchdown PCR was performed using the following steps: denaturation at 95°C for 15 min followed by 20 cycles with denaturation at 94°C for 1 min, annealing for 1 min with a decreasing temperature profile (decreased by 0.5°C every cycle from 65 to 55°C) and elongation at 72°C for 30 s. The last 20 cycles had a denaturation temperature of 94°C for 1 min, an annealing temperature of 55°C for 1 min and an elongation temperature of 72°C for 30 s. The PCR program was finalized with a 5 min step of elongation at 72°C. Amplicon size was verified by agarose gel electrophoresis. The PCR reactions were treated with exonuclease I and shrimp alkaline phosphatase, incubated at 37°C for 30 min followed by 80°C for 10 min and then sequenced using the ABI PRISM 3730xl DNA Analyzer. The allelic expression ratio between alleles was estimated based on the signal ratio in cDNA versus gDNA using the PeakPicker software (38) (http://genomequebec.mcgill.ca/publications/pastinen/).

Detection of CNV events

To detect CNV events, we used two different algorithms, the Canary algorithm implemented in the Affymetrix Genotyping Console software and the PennCNV algorithm (39). For Canary, we defined a minimum window size of five SNPs, with no restrictions on segment size or overlap with known segments. For pennCNV, we followed the pipeline for analyzing Affymetrix genome-wide 6.0 array data denoted in the developers' website. Only events reported by both methods were taken for downstream analysis. As we were interested in rare variation, we needed to filter out events which are abundant in the general population, and appear in healthy individuals. We tested our results for overlap with CNVs in the Database of Genomic Variants (40) and the CNV project at the Children's Hospital of Philadelphia (CHOP) which includes CNVs from 2026 control individuals (41), and with CNVs that were detected in 3181 control individuals using the same Affymetrix platforms (25). CNV validation was carried out using MLPA technology, with the EK-1 kit (MRC-Holland, Amsterdam, The Netherlands), following MRC-Holland recommendations. Primers and probes were supplied by IDT (Integrated DNA Technologies, Coralville, IA, USA). Capillary electrophoresis analysis was performed using an ABI PRISM 3730xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) and analyzed using the Peak Scanner™ Software v1.0.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank Prof. Benjamin Yakir for fruitful discussions and statistical advice and Prof. Jonathan Flint for critical reading of the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the Legacy Heritage Fund program of the Israel Science Foundation (grant no. 1998/08).

REFERENCES

1. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–449.
2. 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) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.
3. Carvajal-Carmona, L.G. (2010) Challenges in the identification and use of rare disease-associated predisposition variants. *Curr. Opin. Genet. Dev.*, **20**, 277–281.
4. Teer, J.K. and Mullikin, J.C. (2010) Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, **19**, R145–R151.
5. Bray, N.J. (2008) Gene expression in the etiology of schizophrenia. *Schizophr. Bull.*, **34**, 412–418.
6. Iwamoto, K. and Kato, T. (2006) Gene expression profiling in schizophrenia and related mental disorders. *Neuroscientist*, **12**, 349–361.
7. Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
8. Morgan, N.V., Morris, M.R., Cangul, H., Gleeson, D., Straatman-Iwanowska, A., Davies, N., Keenan, S., Pasha, S., Rahman, F., Gentle, D. *et al.* (2010) Mutations in SLC29A3, encoding an equilibrative

nucleoside transporter ENT3, cause a familial histiocytosis syndrome (Faisalabad histiocytosis) and familial Rosai-Dorfman disease. *PLoS Genet.*, **6**, e1000833.
9. Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T. and Todd, J.A. (2008) Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE*, **3**, e2966.
10. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. and Chess, A. (2007) Widespread monoallelic expression on human autosomes. *Science*, **318**, 1136–1140.
11. Gong, X., Bacchelli, E., Blasi, F., Toma, C., Betancur, C., Chaste, P., Delorme, R., Durand, C.M., Fauchereau, F., Botros, H.G. *et al.* (2008) Analysis of X chromosome inactivation in autism spectrum disorders. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **147B**, 830–835.
12. Plenge, R.M., Stevenson, R.A., Lubs, H.A., Schwartz, C.E. and Willard, H.F. (2002) Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *Am. J. Hum. Genet.*, **71**, 168–173.
13. Ge, B., Pokholok, D.K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D.J., Le, J., Koka, V., Lam, K.C., Gagne, V. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216–1222.
14. Arking, D.E., Cutler, D.J., Brune, C.W., Teslovich, T.M., West, K., Ikeda, M., Rea, A., Guy, M., Lin, S., Cook, E.H. *et al.* (2008) A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.*, **82**, 160–164.
15. Roohi, J., Montagna, C., Tegay, D.H., Palmer, L.E., DeVincent, C., Pomeroy, J.C., Christian, S.L., Nowak, N. and Hatchwell, E. (2009) Disruption of contactin 4 in three subjects with autism spectrum disorder. *J. Med. Genet.*, **46**, 176–182.
16. Bucan, M., Abrahams, B.S., Wang, K., Glessner, J.T., Herman, E.I., Sonnenblick, L.I., Alvarez Retuerto, A.I., Imielinski, M., Hadley, D., Bradfield, J.P. *et al.* (2009) Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.*, **5**, e1000536.
17. Carter, M., Nikkel, S., Fernandez, B., Marshall, C., Noor, A., Lionel, A., Prasad, A., Pinto, D., Joseph-George, A., Noakes, C. *et al.* (2010) Hemizygous deletions on chromosome 1p21.3 involving the DPYD gene in individuals with autism spectrum disorder. *Clin. Genet*. doi: 10.1111/j.1399-0004.2010.01578.x.
18. Huang, X.L., Zou, Y.S., Maher, T.A., Newton, S. and Milunsky, J.M. (2010) A de novo balanced translocation breakpoint truncating the autism susceptibility candidate 2 (AUTS2) gene in a patient with autism. *Am. J. Med. Genet. A*, **152A**, 2115–2119.
19. Feng, J., Schroer, R., Yan, J., Song, W., Yang, C., Bockholt, A., Cook, E.H. Jr, Skinner, C., Schwartz, C.E. and Sommer, S.S. (2006) High frequency of neurexin 1beta signal peptide structural variants in patients with autism. *Neurosci. Lett.*, **409**, 10–13.
20. Kim, H.G., Kishikawa, S., Higgins, A.W., Seong, I.S., Donovan, D.J., Shen, Y., Lally, E., Weiss, L.A., Najm, J., Kutsche, K. *et al.* (2008) Disruption of neurexin 1 associated with autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 199–207.
21. Morrow, E.M., Yoo, S.Y., Flavell, S.W., Kim, T.K., Lin, Y., Hill, R.S., Mukaddes, N.M., Balkhy, S., Gascon, G., Hashmi, A. *et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science*, **321**, 218–223.
22. Bakkaloglu, B., O'Roak, B.J., Louvi, A., Gupta, A.R., Abelson, J.F., Morgan, T.M., Chawarska, K., Klin, A., Ercan-Sencicek, A.G., Stillman, A.A. *et al.* (2008) Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am. J. Hum. Genet.*, **82**, 165–173.
23. Sultana, R., Yu, C.E., Yu, J., Munson, J., Chen, D., Hua, W., Estes, A., Cortes, F., de la Barra, F., Yu, D. *et al.* (2002) Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. *Genomics*, **80**, 129–134.
24. Huang, X.L., Zou, Y.S., Maher, T.A., Newton, S. and Milunsky, J.M. (2010) A de novo balanced translocation breakpoint truncating the autism susceptibility candidate 2 (AUTS2) gene in a patient with autism. *Am. J. Med. Genet. A*, **152A**, 2112–2114.
25. International Schizophrenia Consortium. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature.*, **455**, 237–241.
26. Khatib, H. (2007) Is it genomic imprinting or preferential expression?. *Bioessays*, **29**, 1022–1028.
27. Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, **11**, 533–538.
28. Kalscheuer, V.M., FitzPatrick, D., Tommerup, N., Bugge, M., Niebuhr, E., Neumann, L.M., Tzschach, A., Shoichet, S.A., Menzel, C., Erdogan, F. *et al.* (2007) Mutations in autism susceptibility candidate 2 (AUTS2) in patients with mental retardation. *Hum. Genet.*, **121**, 501–509.
29. Bedogni, F., Hodge, R.D., Elsen, G.E., Nelson, B.R., Daza, R.A., Beyer, R.P., Bammler, T.K., Rubenstein, J.L. and Hevner, R.F. (2010) Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. *Proc. Natl Acad. Sci. USA*, **107**, 13129–13134.
30. Bedogni, F., Hodge, R.D., Nelson, B.R., Frederick, E.A., Shiba, N., Daza, R.A. and Hevner, R.F. (2010) Autism susceptibility candidate 2 (Auts2) encodes a nuclear protein expressed in developing brain regions implicated in autism neuropathology. *Gene Expr. Patterns*, **10**, 9–15.
31. Yirmiya, N., Rosenberg, C., Levi, S., Salomon, S., Shulman, C., Nemanov, L., Dina, C. and Ebstein, R.P. (2006) Association between the arginine vasopressin 1a receptor (AVPR1a) gene and autism in a family-based study: mediation by socialization skills. *Mol. Psychiatry*, **11**, 488–494.
32. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
33. Diss, T.C., Pan, L., Peng, H., Wotherspoon, A.C. and Isaacson, P.G. (1994) Sources of DNA for detecting B cell monoclonality using PCR. *J. Clin. Pathol.*, **47**, 493–496.
34. Lobo, A., Okhravi, N., Adamson, P., Clark, B.J. and Lightman, S. (2007) Protocol for the use of polymerase chain reaction in the detection of intraocular large B-cell lymphoma in ocular samples. *J. Mol. Diagn.*, **9**, 113–121.
35. Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.
36. Shi, P., Bakewell, M.A. and Zhang, J. (2006) Did brain-specific genes evolve faster in humans than in chimpanzees?. *Trends Genet.*, **22**, 608–613.
37. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
38. Ge, B., Gurd, S., Gaudin, T., Dore, C., Lepage, P., Harmsen, E., Hudson, T.J. and Pastinen, T. (2005) Survey of allelic expression using EST mining. *Genome Res.*, **15**, 1584–1591.
39. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. and Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.
40. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
41. Shaikh, T.H., Gai, X., Perin, J.C., Glessner, J.T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L.K., D'Arcy, M. *et al.* (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.*, **19**, 1682–1690.