The American Journal of Human Genetics

Supplemental Data

Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV

Michael H. Duyzend, Xander Nuttle, Bradley P. Coe, Carl Baker, Deborah A. Nickerson, Raphael Bernier, and Evan E. Eichler

Supplemental Figures

Supplemental Appendix

Supplemental Tables

Supplemental Web Resources

Supplemental References

Supplemental Figures

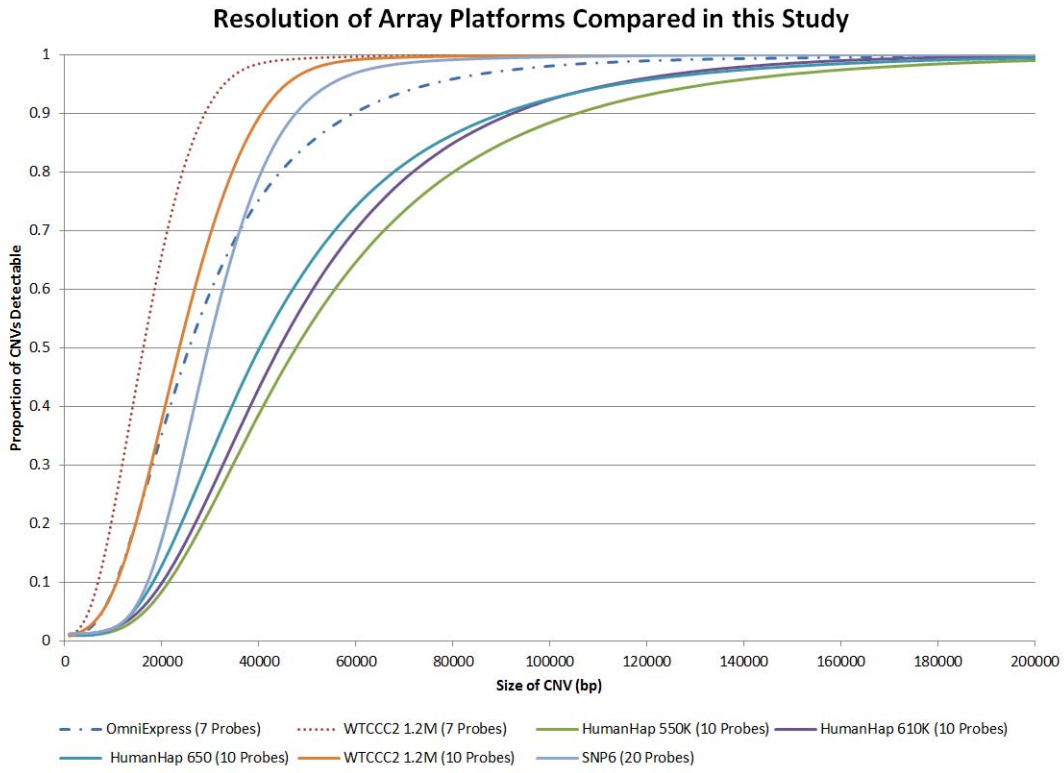


Figure S1: The sensitivity of the SNP microarray platforms used in this study. The number of probes required in each simulated CNV is listed. We used the method for calculating sensitivity by Coe *et al*.[1]

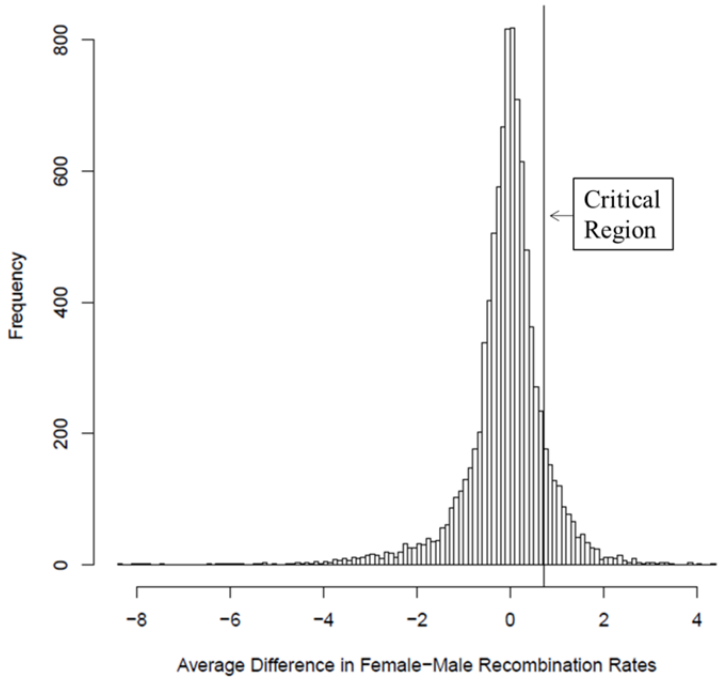


Figure S2: **Average difference in female-male recombination rates for 550 kbp windows genome-wide.** The average difference in female and male recombination rates were calculated for ten thousand 550 kbp windows (the same size as the 16p11.2 critical region), excluding sex chromosomes, regions with segmental duplications, and gaps. The 16p11.2 critical region is in the 87th percentile for the average difference amongst these regions.

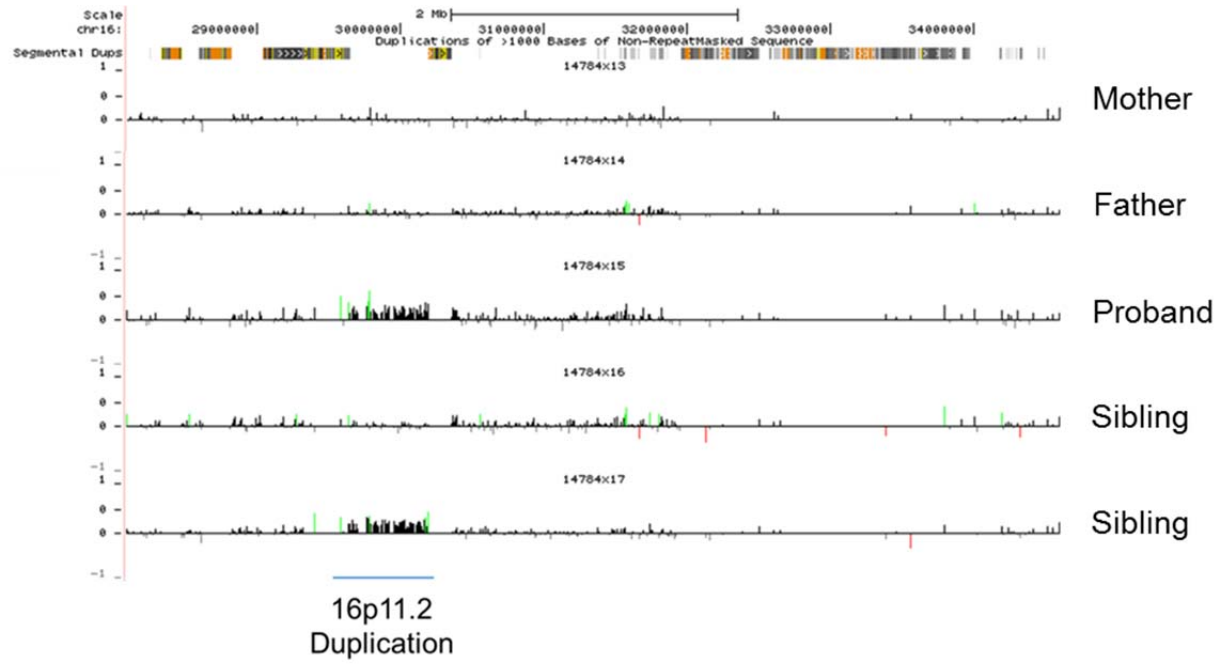


Figure S3: Incorrect inheritance pattern for family 14784. This proband (14784.x15) was labelled as having a de novo duplication, but the sibling (14784.x17), listed as a non-carrier, also carries a duplication. Since the event is not present in the parents (14784.x13, mother; 14784.x14, father), this is likely the result of germline mosaicism.

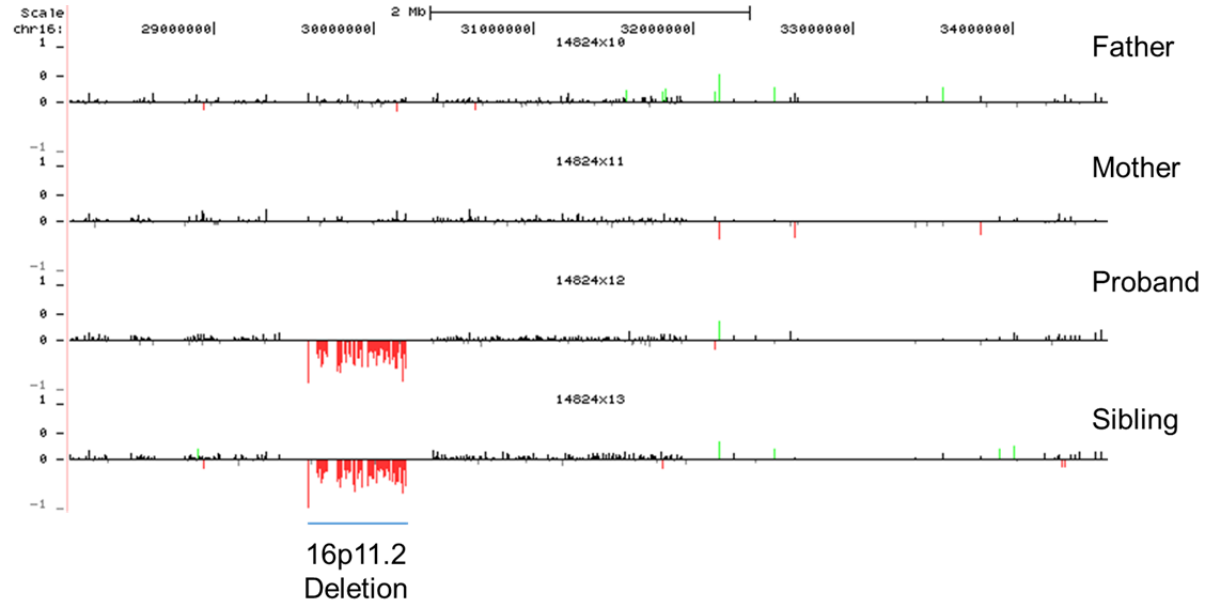


Figure S4: Monozygotic twins in family 14824 with a de novo 16p11.2 deletion.

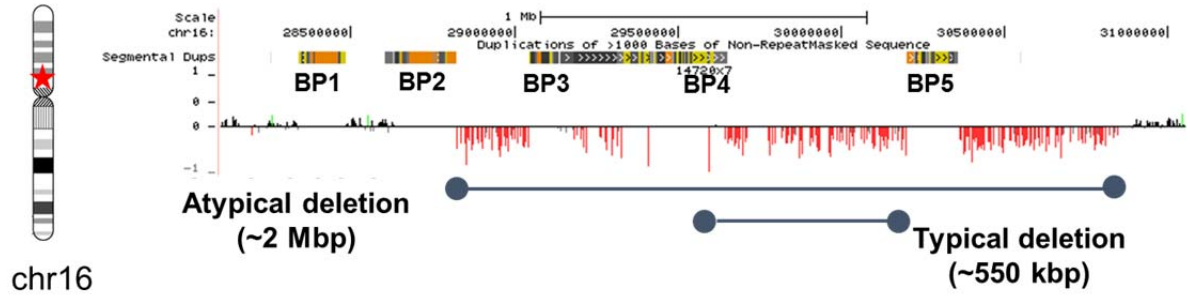


Figure S5: Atypical 16p11.2 deletion. One individual in the cohort (14720.x10) had an atypical de novo 16p11.2 deletion that extends from breakpoint 2 beyond breakpoint 5 with nearly 2 Mbp of genomic DNA deleted.

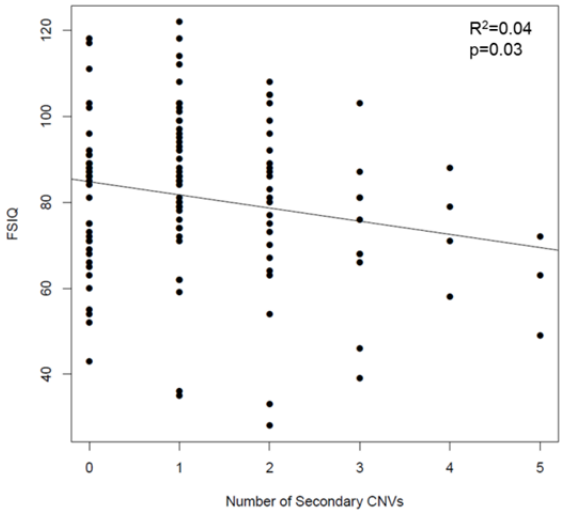
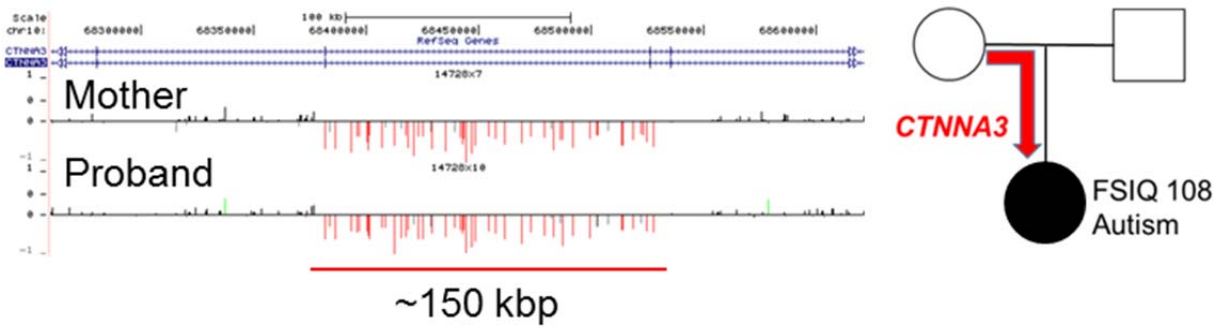


Figure S6: Correlation of FSIQ and the number of secondary CNVs in screened probands. There is a modest, statistically significant correlation between the FSIQ and number of secondary CNVs present in the screened probands (R^2=0.04, p=0.03).

a.



b.

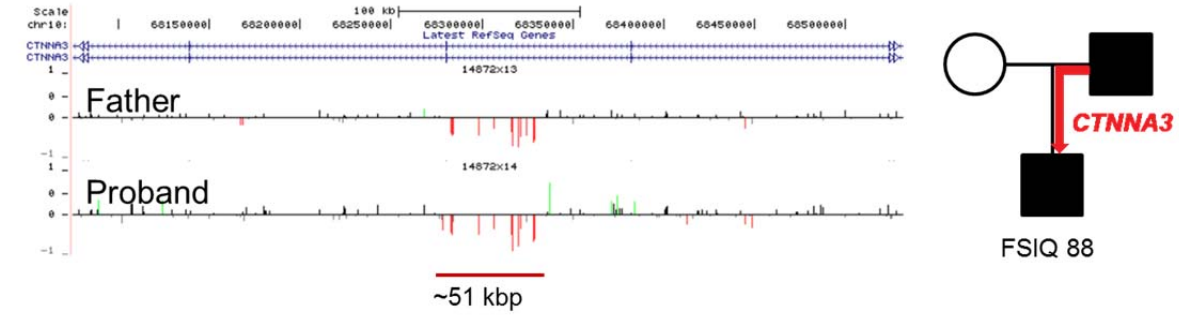


Figure S7: Secondary deletions of *CTNNA3* in 16p11.2 duplication families. We found a recurrent secondary deletion in two 16p11.2 duplication families affecting a gene associated with autism risk variants, αT-catenin, *CTNNA3*. (a) A ~150 kbp deletion involving *CTNNA3* is transmitted from mother to daughter who also carries a 16p11.2 de novo duplication. (b) A ~50 kbp deletion transmitted from father to son. Both father and son carry the 16p11.2 duplication. While both of these CNVs are individually rare, there are a similar number of cases and controls with events across *CTNNA*.[2] However, two large studies have found genetic association between *CTNNA3* and autism[3,4] and rare deletions have been identified in individuals with ASD.[5]

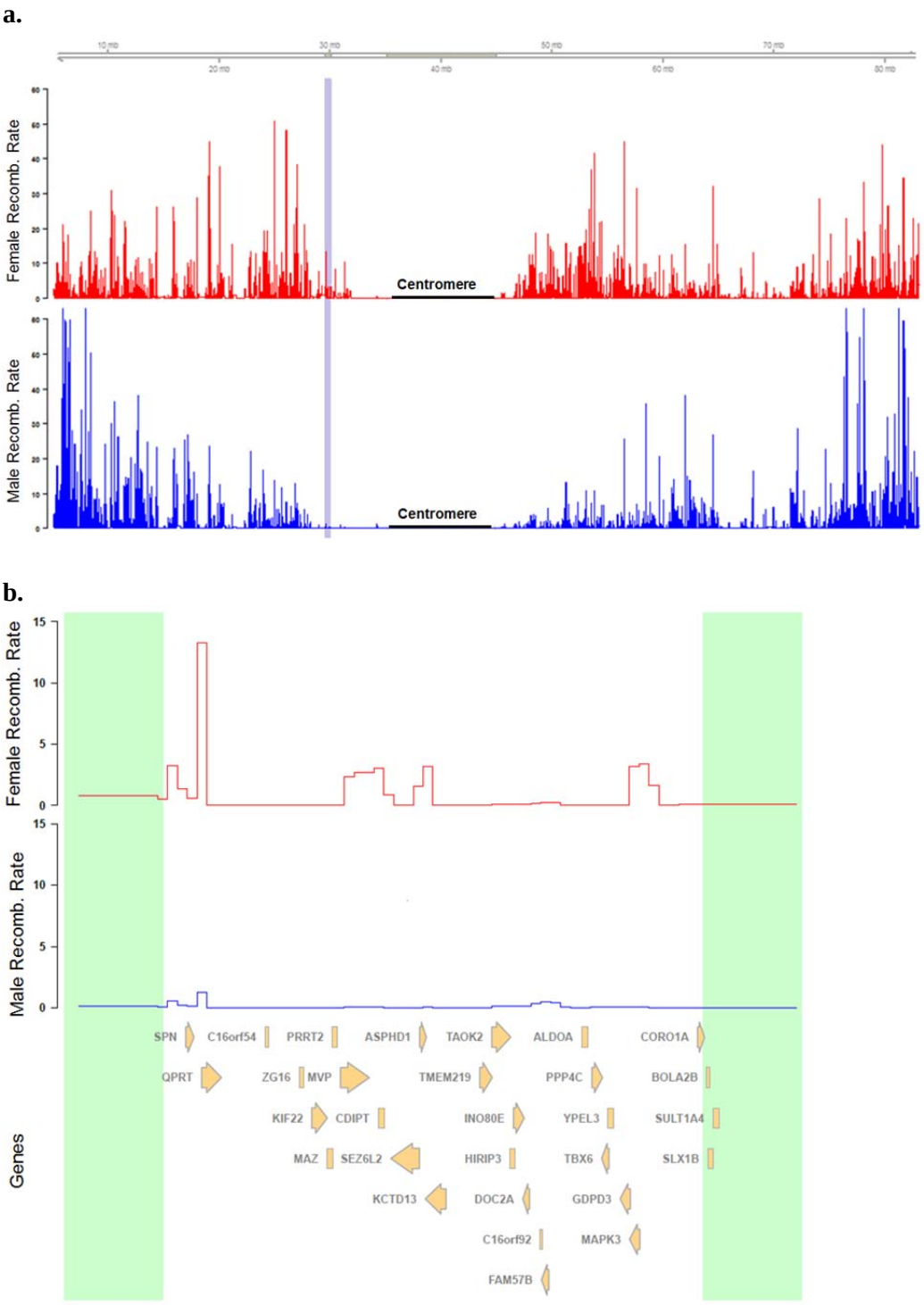


Figure S8: Male and female recombination rates across chromosome 16. a) Not including the telomeres, males have a decreased recombination rate across chromosome 16 compared to females, using data from Kong *et al.*[6]. The lavender bar is the critical region. b) The standardized female recombination rate in females (red) and males (blue) across the 16p11.2 critical region. Note the maternal recombination hotspot in the left-hand most side of the region. (Plotted using the GViz R package[7].)

Supplemental Appendix

This Supplemental Appendix includes the methods for determining the parent-of-origin and mechanism of unequal crossing over of de novo 16p11.2 CNVs.

Method for determining parent-of-origin of de novo 16p11.2 CNVs

De novo 16p11.2 deletions

We used the B-allele frequency data from the SNP microarrays to infer if de novo events originated on the maternal or paternal haplotype. For this analysis, we used data from probes (112, for OmniExpress arrays) falling in the 16p11.2 critical region. In deletion individuals, only one copy of the critical region remains. Therefore, possible SNP genotypes for probes are A or B with corresponding B-allele frequency of 0 or 1. This is tabulated below:

Deletion:		
Affected Haplotype:	*Unaffected Haplotype:*	*B-allele Frequency:*
NA	A	0
NA	B	1

In the case where we had genotype information on both parents (trios), we used the parental genotypes to infer the inheritance of the unaffected haplotype in the proband.

We define the probabilities as follows:

$P(mother|probes)$: Probability of inheritance of the unaffected haplotype from mother given the probes

$P(father|probes)$: Probability of inheritance of the unaffected haplotype from father given the probes

$P(mother)$: Prior probability of inheritance of the unaffected haplotype from mother

$P(probes)$: Posterior probability of the observed probes

By Bayes' theorem,

$$P(mother|probes) = \frac{P(probes \mid mother) * P(mother)}{P(probes)}$$

And similarly for the father,

$$P(father|probes) = \frac{P(probes \mid father) * P(mother)}{P(probes)}$$

Since the prior probability of inheritance of the unaffected haplotype from either parent is 0.5, we have that:

$$P(mother) = \frac{1}{2}$$

and,

$$P(father) = \frac{1}{2}$$

Since the unaffected haplotype must come from either the mother or the father, we also have that,

$$P(mother \mid probes) + P(father \mid probes) = 1$$

Using this relationship, we see that,

$$P(probes) = \frac{P(probes \mid father) + P(probes \mid mother)}{2}$$

To determine the quantity $P(probes \mid mother)$, the probability of observing the set of probes given that the unaffected haplotype comes from the mother, we make the assumption that probe signals are independent and recognize that,

$$P(probes \mid mother) = \prod_{i=1}^{n} P(probe_i \mid mother)$$

where $probe_i$ is the i^{th} probe out of n in the critical region.

To compute $P(probe_i \mid mother)$, we recognize that each site in the unaffected haplotype will have a genotype of either A or B. Since the mother is diploid over the critical region, at each site she has possible genotypes of AA, AB, or BB. Assuming the probability of a genotyping error is 0.001, as suggested by Illumina (see **Supplemental Web Resources**), we build the following probability table:

Mother Genotype at probe *i*	Proband Genotype at probe *i*	P($probe_i$|*mother*)
AA	A	0.999
AB	A	0.5
BB	A	0.001
AA	B	0.001
AB	B	0.5
BB	B	0.999

The approach is identical to compute $P(probes|father)$.

In the cases where we had only one parent available, we estimated the probability of the unobserved parent using the known allele frequencies for that particular probe from the 1000 Genomes Project, when it existed. In these cases, even though we did not have SNP microarray data from the missing parent, in all cases the 16p11.2 CNV was determined to be de novo by clinical microarray or another method. We model the genotype of the missing parent at each probe using the allele frequencies calculated from the 1000 Genomes Project and we only use probes that are present in dbSNP 140 (95 probes). For the missing parent, the probability $P(probe_i|MissingParent)$ now depends on the genotype frequencies. Let a_i be the allele frequency of the A allele at a probe i and b_i be the allele frequency of the B allele at a probe i. Then, assuming the probability of a genotyping error is 0.001, we construct the following probability table:

Missing Parent Genotype at probe *i*	P(Missing Parent Genotype at probe *i*)	Proband Genotype at probe *i*	P($probe_i$|MissingParent)
AA	a_i^2	A	0.999
AB	$2*a_i*b_i$	A	0.5
BB	b_i^2	A	0.001
AA	a_i^2	B	0.001
AB	$2*a_i*b_i$	B	0.5
BB	b_i^2	B	0.999

Then, using Hardy-Weinberg we have:

$$P(probe_i = A|MissingParent) = 0.999 * a_i^2 + 0.5 * 2 * a_i * b_i + 0.001 * b_i^2$$

and

$$P(probe_i = B|MissingParent) = 0.999 * b_i^2 + 0.5 * 2 * a_i * b_i + 0.001 * a_i^2$$

In this way, we calculate the exact probability that the unaffected haplotype was inherited from the mother or father.

In cases where a different version of the same array was run (i.e., Omni1 vs. Omni2), we selected only those probes present on each array in the family.

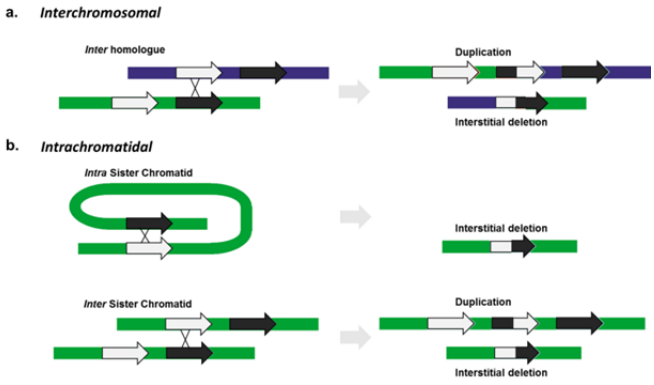
De novo 16p11.2 duplications

Duplication individuals have three copies of the critical region. The possible SNP genotypes for each probe over the critical region in duplication individuals are AAA, AAB, ABB, and BBB with corresponding B-allele frequencies of 0, 1/3, 2/3, and 1, respectively. Due to the possibilities of either inter or intrachromosomal mechanisms of crossing over, we evaluate the genotypes of both parents and the proband to determine the parent-of-origin. In particular, we compute all possible outcomes of rearrangement given a particular parent-of-origin and mechanism of crossing over (**Table S14**). At each probe *i*, this provides the probabilities of parent-of-origin and, in some cases, mechanism of unequal crossing over. We count the number of partially informative markers (probability >0.5 and <1) and fully informative markers (probability=1) to determine the parent-of-origin of the observed event (**Table S5**).

Genotype			Parent-of-Origin		Mechanism	
Mother	**Father**	**Proband**	**Prob. Mother**	**Prob. Father**	**Prob. Inter**	**Prob. Intra**
AA	AA	AAA	0.50	0.50	0.50	0.50
AA	AA	AAB	NA	NA	NA	NA
AA	AA	ABB	NA	NA	NA	NA
AA	AA	BBB	NA	NA	NA	NA
AA	AB	AAA	0.66	0.33	0.33	0.66
AA	AB	AAB	0.50	0.50	0.50	0.50
AA	AB	ABB	0.00	1.00	0.00	1.00
AA	AB	BBB	NA	NA	NA	NA
AA	BB	AAA	NA	NA	NA	NA
AA	BB	AAB	1.00	0.00	0.50	0.50
AA	BB	ABB	0.00	1.00	0.50	0.50
AA	BB	BBB	NA	NA	NA	NA
AB	AA	AAA	0.33	0.66	0.33	0.66
AB	AA	AAB	0.50	0.50	0.50	0.50
AB	AA	ABB	1.00	0.00	0.00	1.00
AB	AA	BBB	NA	NA	NA	NA
AB	AB	AAA	0.50	0.50	0.00	1.00
AB	AB	AAB	0.50	0.50	0.66	0.33
AB	AB	ABB	0.50	0.50	0.66	0.33
AB	AB	BBB	0.50	0.50	0.00	1.00
AB	BB	AAA	NA	NA	NA	NA
AB	BB	AAB	1.00	0.00	0.00	1.00
AB	BB	ABB	0.50	0.50	0.75	0.25
AB	BB	BBB	0.33	0.66	0.33	0.66
BB	AA	AAA	NA	NA	NA	NA
BB	AA	AAB	0.00	1.00	0.50	0.50
BB	AA	ABB	1.00	0.00	0.50	0.50
BB	AA	BBB	NA	NA	NA	NA
BB	AB	AAA	NA	NA	NA	NA
BB	AB	AAB	0.00	1.00	0.00	1.00
BB	AB	ABB	0.50	0.50	0.75	0.25
BB	AB	BBB	0.66	0.33	0.33	0.66
BB	BB	AAA	NA	NA	NA	NA
BB	BB	AAB	NA	NA	NA	NA
BB	BB	ABB	NA	NA	NA	NA
BB	BB	BBB	0.50	0.50	0.50	0.50

Method for determining mechanism of unequal crossing over of de novo 16p11.2 CNVs

Nonallelic homologous recombination (NAHR) can occur between homologous chromosomes or between or within sister chromatids:



Mechanisms of unequal crossing over. (a) NAHR can occur between homologous chromosomes or (b) between or within sister chromatids.

We used the below method to determine the mechanism of unequal crossover.

Quads:
When possible, we used full quads in order to perfectly phase the parents using the sibling.

(1) *Phase the proband and sibling*

We phased the proband and sibling into maternal and paternal alleles by comparison with the parental genotypes. To determine the mechanism of unequal crossing over, we consider only the haplotype on which the de novo 16p event occurred (maternal or paternal). The possibilities are shown below:

Mother	Father	Child	Maternal Allele	Paternal Allele
AA	AA	AA	A	A
AA	AB	AA	A	A
AA	AB	AB	A	B
AA	BB	AB	A	B
AB	AA	AA	A	A
AB	AA	AB	B	A
AB	AB	AA	A	A
AB	AB	AB	NA	NA
AB	AB	BB	B	B
AB	BB	AB	A	B
AB	BB	BB	B	B
BB	AA	AB	B	A
BB	AB	AB	B	A
BB	AB	BB	B	B
BB	BB	BB	B	B

Only those sites that can distinguish between the two chromosomes in the parent on which the 16p11.2 de novo event originated are informative. For example, if the event occurred on the maternal haplotype, then only heterozygous genotypes in the mother are informative. In total, we considered a total of 314 markers telomeric and 314 markers centromeric of the critical region as possibly informative. For the rest of this discussion, we assume for simplicity that the 16p11.2 de novo CNV occurs on the maternal haplotype.

(2) *Compare the proband and sibling maternal haplotypes*

We assume that no crossing over event has occurred in the 16p11.2 region in the sibling on the haplotype of interest. Therefore, the maternal haplotype present in the sibling should represent perfectly one of the two maternal chromosomes. We determine if the maternal alleles flanking the critical region match the sibling alleles (314 markers on the left, 314 markers on the right). There are four possibilities, shown below:

Left Flank	**Right Flank**	**Conclusion**
Match	Match	Intrachromosomal
Don't Match	Match	Interchromosomal
Match	Don't Match	Interchromosomal
Don't Match	Don't Match	Intrachromosomal

(3) *Calculate a probability that the event occurred by an inter or intrachromosomal mechanism*

We developed a probability model to calculate the probability of an interchromosomal versus an intrachromosomal event given the number of alleles in the left and right flanks (called flank 1 and flank 2 for simplicity) that match the sibling markers. Each marker gets the number 0 if the proband does not match the sibling and 1 if the proband matches the sibling. We notice that the probability of an intrachromosomal event is the probability of the left flank and right flanks of the proband either matching the sibling or both not matching the sibling. We also assume independence of the flanks, so we have that:

$$P(Intrachromosomal|Probes)$$
$$= P\left(((flank1 = 0 \cap flank2 = 0) \cup (flank1 = 1 \ \cap flank2 = 1))|probes\right)$$
$$= P(flank1 = 0|probes)P(flank2 = 0|probes) + P(flank1 = 0|probes)P(flank2 = 0|probes)$$

Similarly, if the flanks do not match each other, than we have an interchromosomal event. That is:

$$P(Interchromosomal|Probes)$$
$$= P\left(((flank\ 1 = 0 \cap flank\ 2 = 1) \cup (flank\ 1 = 0 \ \cap flank\ 2 = 1))|probes\right)$$
$$= P(flank1 = 0|probes)P(flank2 = 1|probes) + P(flank1 = 0|probes)P(flank2 = 1|probes)$$

From this we note that:

$$P(Intrachromosomal|Probes) + P(Interchromosomal|Probes) = 1$$

Using this relationship, we only have to calculate $P(Intrachromosomal|Probes)$. To do so, we must calculate each of the four components that make up this probability. We will show how to do this for the first two components corresponding to flank 1, as the second two follow. Using Bayes' theorem, we have that:

$$P(flank1 = 0|probes) = \frac{P(flank1 = 0) * P(probes|flank1 = 0)}{P(probes)}$$

and also that:

$$P(flank1 = 1|probes) = \frac{P(flank1 = 1) * P(probes|flank1 = 1)}{P(probes)}$$

We assume the probability that proband and sibling match in flank 1 is equal to the probability that they do not match. Therefore,

$P(flank1 = 0) = 0.5$
and,
$P(flank1 = 1) = 0.5$

To calculate *P(probes)* we note that:

$$P(flank1 = 0|probes) + P(flank1 = 1|probes) = 1$$

Expanding this, we have that:

$$P(probes) = \frac{P(probes|flank1 = 0) + P(probes|flank1 = 1)}{2}$$

Finally, we assume independence of individual markers and note that:

$$P(probes|flank1 = 0) = \prod_{probes} P(probe_i|flank1 = 0)$$

and,

$$P(probes|flank1 = 1) = \prod_{probes} P(probe_i|flank1 = 1)$$

To compute the probabilities for the individual probes, we take that the probability of a genotyping error is 0.001. Therefore, we construct the following probability table:

Sibling/Proband Markers	$P(probe_i\|flank1=0)$	$P(probe_i\|flank1=1)$
Don't Match (0)	0.999*0.001*2	$0.999^2*0.001^2$
Match (1)	$0.999^2*0.001^2$	0.999*0.001*2

To determine the accuracy of the assumption that no crossing over events occurred in the sibling, we asked how many crossover events were predicted between the leftmost and rightmost marker in our analysis. Based on the data from Kong *et al.*[1], the genetic distance between the leftmost and rightmost markers in our analysis is 6.20 centimorgans for the female and 0.45 centimorgans for the male, which corresponds to a probability of crossover of 6.2% for the female and 0.45% for the male. Therefore, the number of false positives (an intrachromosomal event being interpreted as an interchromosomal event or vice versa) is 1 in 16 for de novo events originating on the maternal haplotype and 1 in 222 for events originating on the paternal haplotype.

Trios:

In the trio case, we do not have a sibling for phasing the haplotypes of the parent-of-origin. Therefore, we performed statistical phasing of the mothers. The approach was similar to that used for the quads:

(1) *Phase the proband*

We phased the proband into maternal and paternal alleles by comparison with the parental genotypes. As before, we are only concerned with the informative markers that allow us to distinguish between parental haplotypes, i.e., those cases when the parent-of-origin is heterozygous for a particular allele.

(2) *Phase the parent-of-origin haplotypes*

Since a sibling is not present to phase the haplotypes of the parent-of-origin, we instead need to statistically phase the parent-of-origin haplotypes. To do this we used the program Beagle v4.0 (http://faculty.washington.edu/browning/beagle/beagle.html) and used the phased 1000 Genomes Project phase 3 chromosome 16 reference panel for phasing (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5/).
The exact command used is below:

java -Xmx10G -jar beagle.r1399.jar gt=Input.vcf.gz ref=chr16.1kg.phase3.v5.vcf.gz out= phased_data.out impute=false

(3) *Compare the proband to the parent-of-origin*

Next, we compare the proband's haplotype from the parent-of-origin of the 16p11.2 event to one of the haplotypes from the parent-of-origin. As before, the probes will either match or not match in each flank and we use the same probability model as before to compute the probability of an inter or intrachromosomal event.

Duplications

As mentioned above, for duplications certain combinations of maternal, paternal, and proband genotypes are partially or perfectly informative of a mechanism of unequal crossing over. When the flanking marker-based approach did not yield a consistent result, we compared this approach to the approach using probes in the critical region (**Table S5**).

Supplemental Web Resources
Beagle: http://faculty.washington.edu/browning/beagle/beagle.html
cnvPartition Algorithm: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_cnv_plug_ins.pdf
Genotype Rare Variants Tech Note: http://support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_genotyping_rare_variants.pdf

Supplemental References

1. Coe, B.P., Ylstra, B., Carvalho, B., Meijer, G.A., Macaulay, C., and Lam, W.L. (2007). Resolving the resolution of array CGH. Genomics *89*, 647–653.

2. Coe, B.P., Witherspoon, K., Rosenfeld, J.A., van Bon, B.W.M., Vulto-van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E.L.M., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat. Genet. *46*, 1063–1071.

3. Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.A., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528–533.

4. Miyashita, A., Arai, H., Asada, T., Imagawa, M., Matsubara, E., Shoji, M., Higuchi, S., Urakami, K., Kakita, A., Takahashi, H., et al. (2007). Genetic association of CTNNA3 with late-onset Alzheimer's disease in females. Hum. Mol. Genet. *16*, 2854–2869.

5. 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). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

6. Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

7. Hahne,F, Durinck,S., Ivanek,R., Mueller,A., Lianoglou,S., Tan,G., and Parsons, L. Gviz: Plotting data and annotation information along genomic coordinates. R package version 1.12.1.