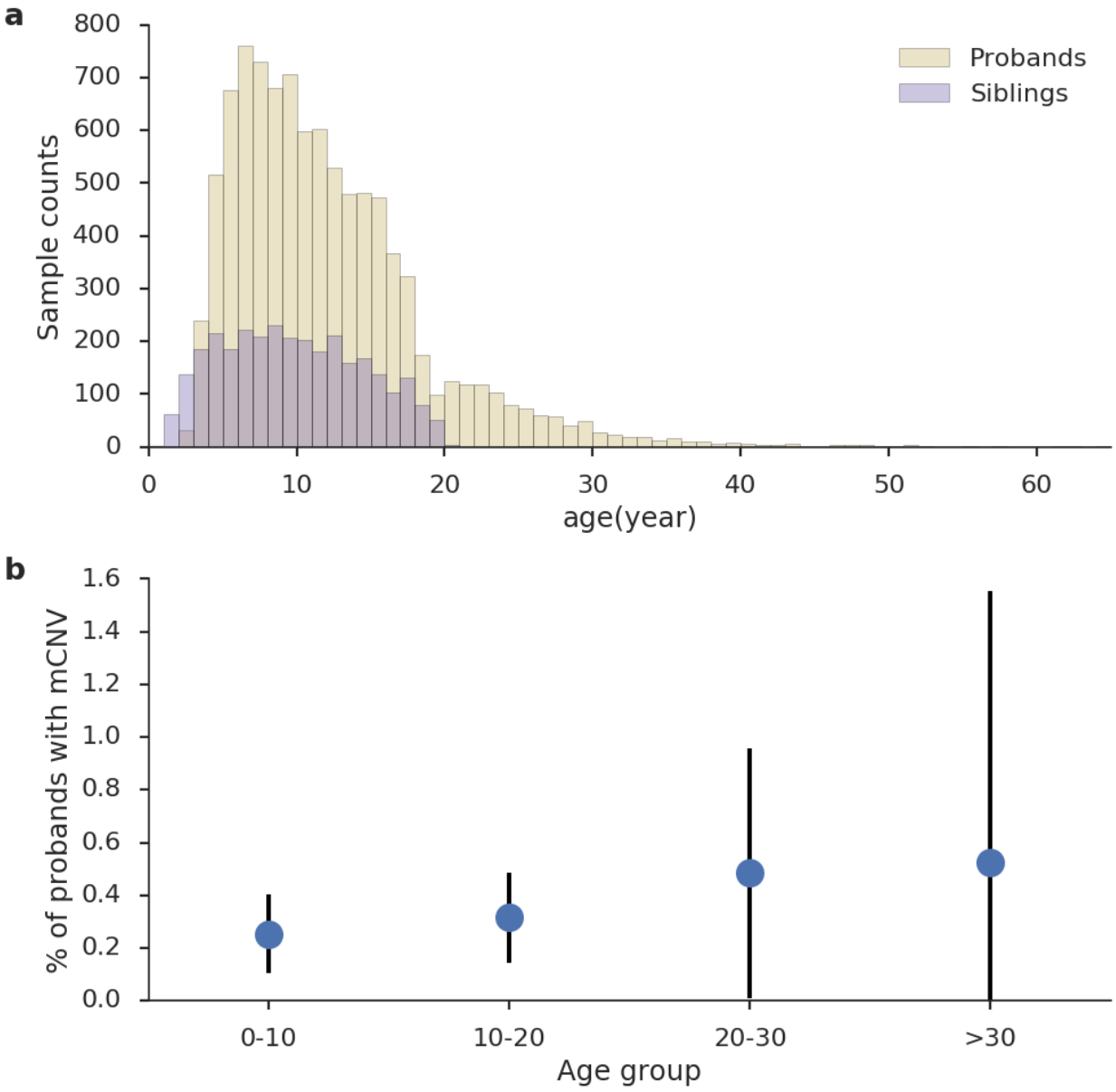
Supplementary Information:

Large mosaic copy number variations confer autism risk

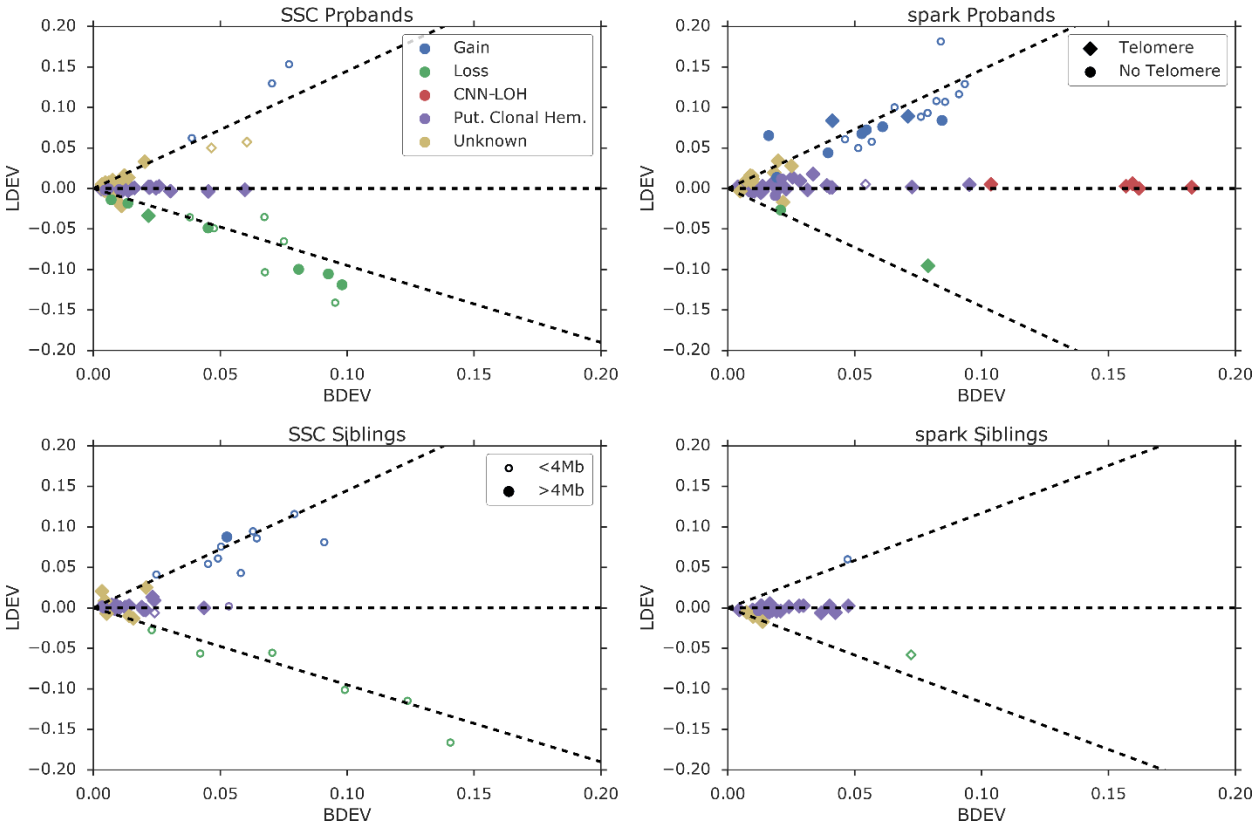
Maxwell A. Sherman[1,2,3,*], Rachel E. Rodin[3,4], Giulio Genovese[3,5,6], Caroline Dias[4,7], Alison R. Barton[2,3], Ronen E. Mukamel[2,3], Bonnie Berger[1,8], Peter J. Park[9,*,§], Christopher A. Walsh[3,4,*,§], Po-Ru Loh[2,3,*,§]

Contents

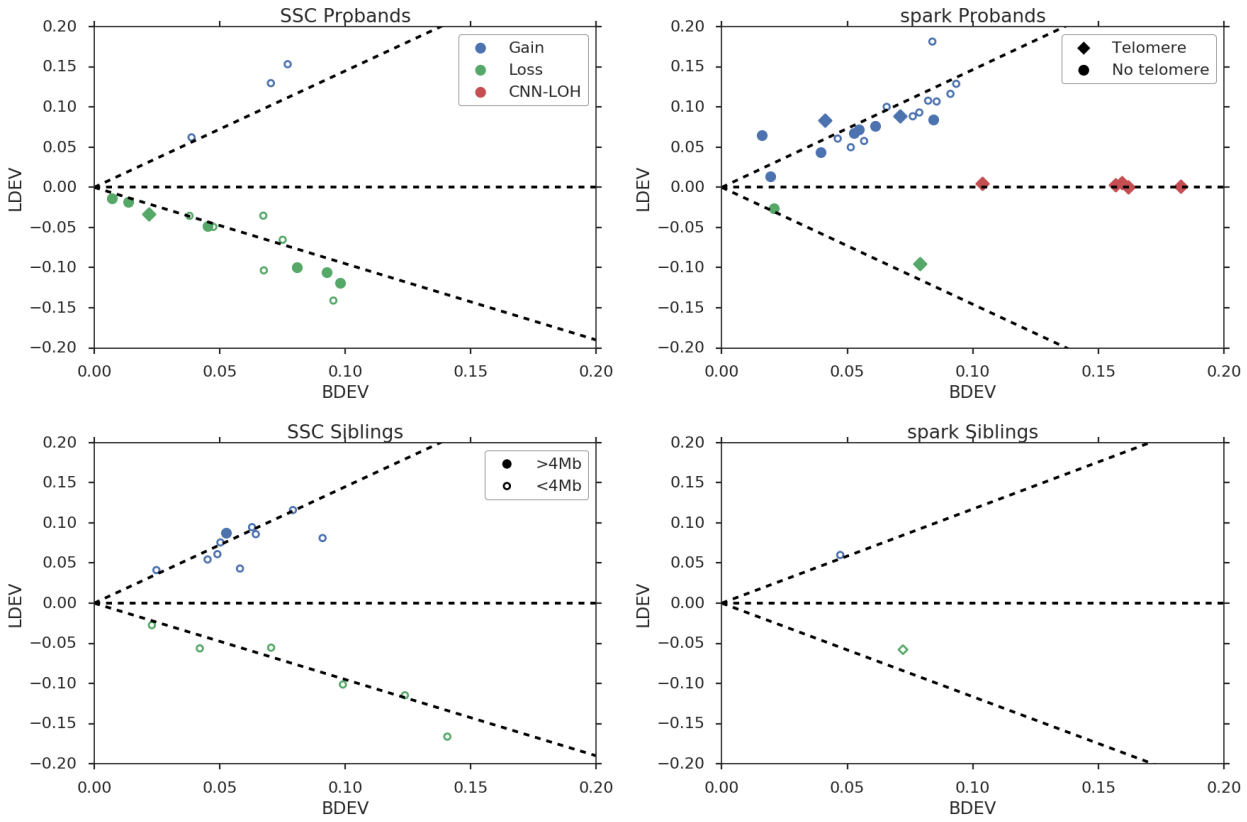
1. Supplementary Figures



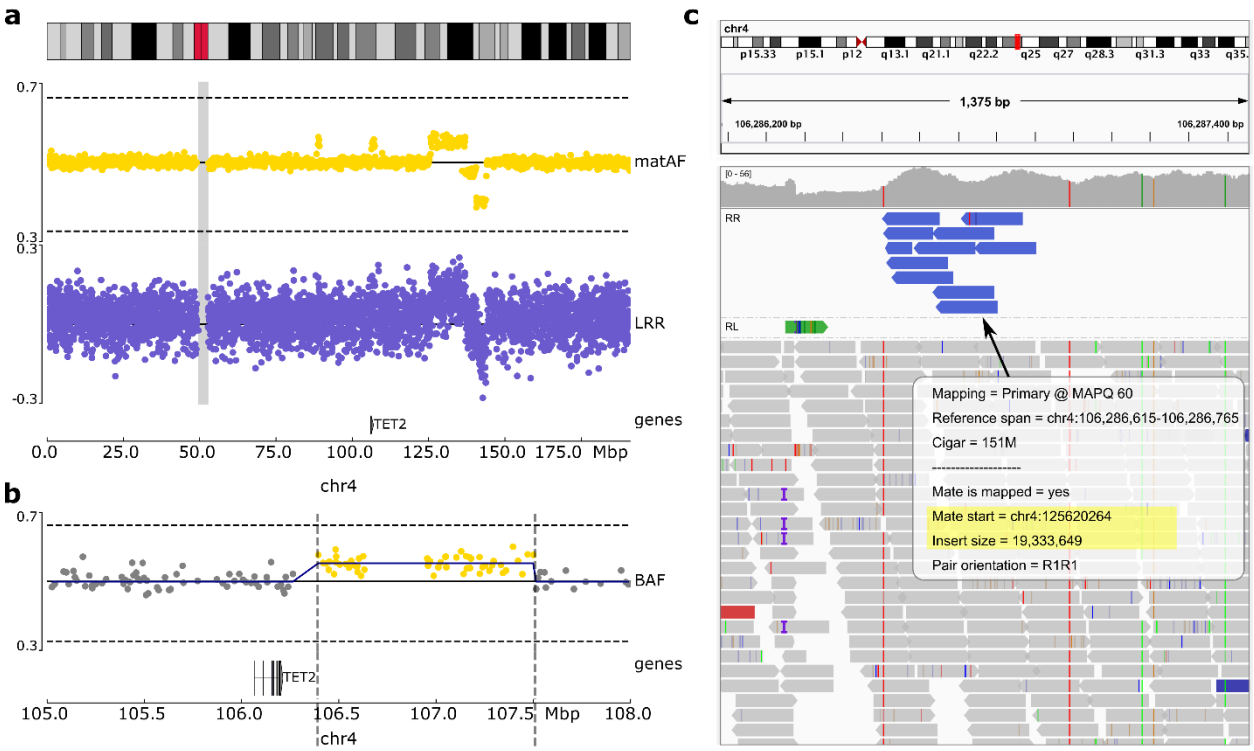
Supplementary Figure 1: **a**, Age distribution of SPARK probands and siblings and **b,** rate of mCNVs in SPARK probands by age. The increase in rate with age is not significant (P = 0.095 logistic regression; P = 0.40 comparison of rate in 0-10 age group to rate in >30 age group by Fisher's exact test). Probands and siblings in SSC were between age 3-18 at enrollment, and individual-level age information in SSC was not available to us.



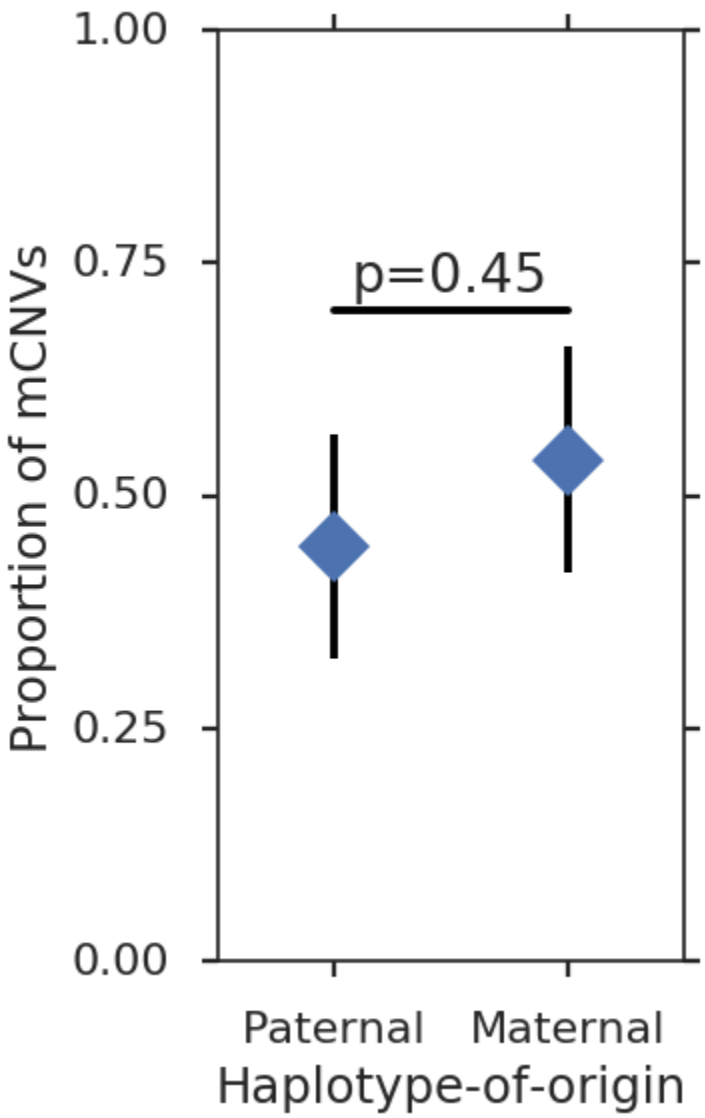
Supplementary Figure 2: Plots of LRR deviation from 0 (LDEV) versus B allele frequency deviation from 0.5 (BDEV) in all mosaic events including putative clonal hematopoietic events and events of unknown copy-number state. Gains fall along an upwards diagonal line; losses fall along a downwards diagonal line; and CNN-LOH fall along the horizontal axis. Dashed lines are the expected duplication, deletion, and CNN-LOH trends as inferred by the EM algorithm fit on parental data (Methods). Marker color indicates inferred mosaic copy state. Small, unfilled markers are events <4 Mb and large filled markers are events >4 Mb; circles indicate events which do not extend to telomeres and diamonds indicate those which do extend to telomeres. Event type is indicated by color as listed in the legend of SSC Probands.



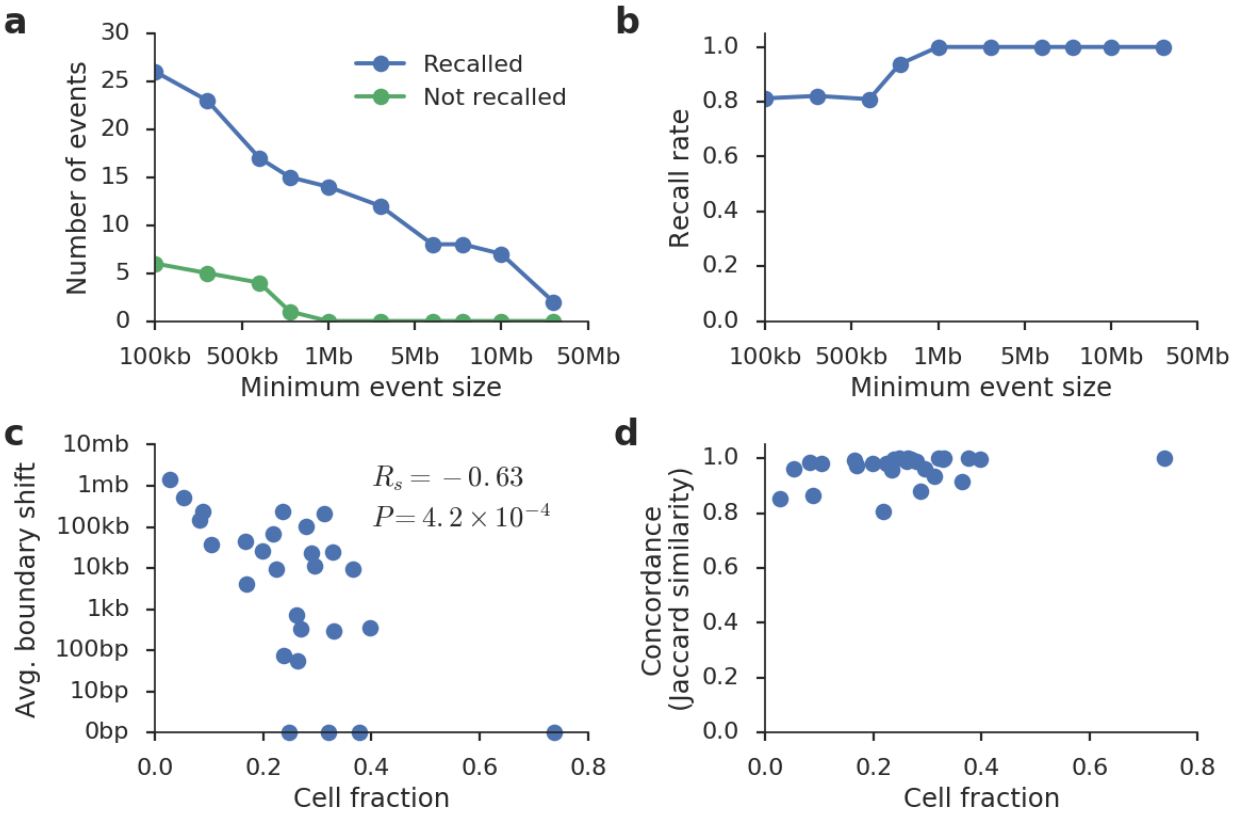
Supplementary Figure 3: Plots of LRR deviation from 0 (LDEV) versus B allele frequency deviation from 0.5 (BDEV) for putative early-developmental mosaic events that are included in burden analyses. Dashed lines are the expected duplication, deletion, and CNN-LOH trends as inferred by the EM algorithm fit on parental data (Methods). Marker color indicates inferred mosaic copy-number state. Small, unfilled markers are events <4 Mb and large filled markers are events >4 Mb; circles indicate events which do not extend to telomeres and diamonds indicate those which do extend to telomeres. Event type is indicated by color as listed in the legend of SSC Probands.



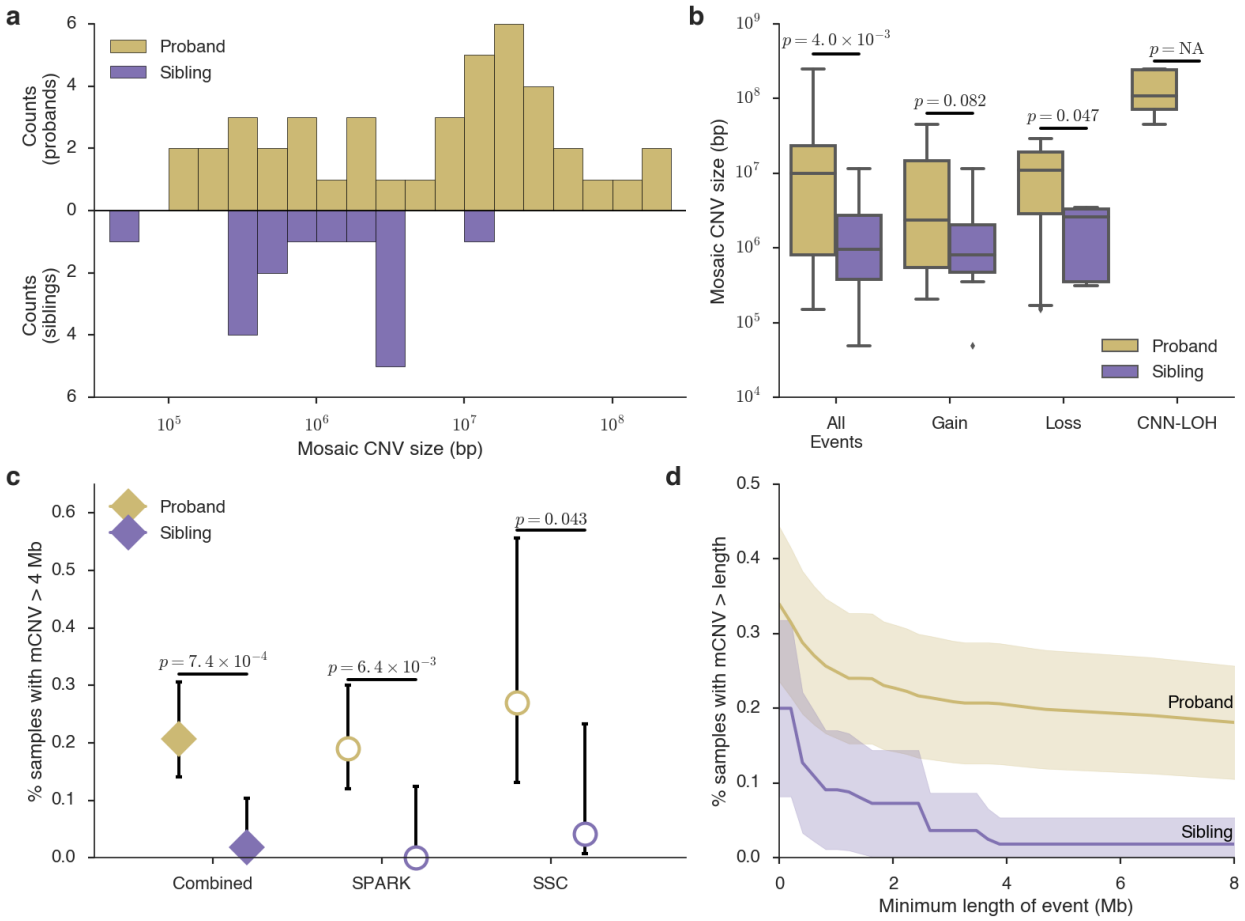
Supplementary Figure 4: SSC sibling 13391.s1 carries a complex rearrangement reminiscent of chromothripsis on chromosome 4. a) Maternal allele frequency (matAF, gold) of heterozygous SNPs and LRR signal (purple) of all array-typed SNPs across chr4 of 13391.s1. Multiple mosaic duplications and deletions are apparent. The location of the gene *TET2*, which has been implicated in myeloproliferative disorders, is as indicated. b) Zoom-in of matAF of heterozygous SNPs in the region around *TET2*. The approximate boundaries of the duplicated segment as determined by MoChA are marked with dashed grey lines. SNPs falling within the duplication are colored gold; SNPs outside the duplication are colored grey. c) IGV plot of left breakpoint of duplicated segment shown in (b). Reads with discordantly mapped mates are colored blue. Mapping information is shown for one representative blue read, demonstrating that it maps to chr4:106,286,615 and its mate maps to the start of the large duplicated segment spanning approximately 4:125610859-137063398 as determined by MoChA. The breakpoint occurs ~85 kb downstream of *TET2*. Reads were aligned to GRCh38 and positions were converted to GRCh37 coordinates via the UCSC liftOver tool.



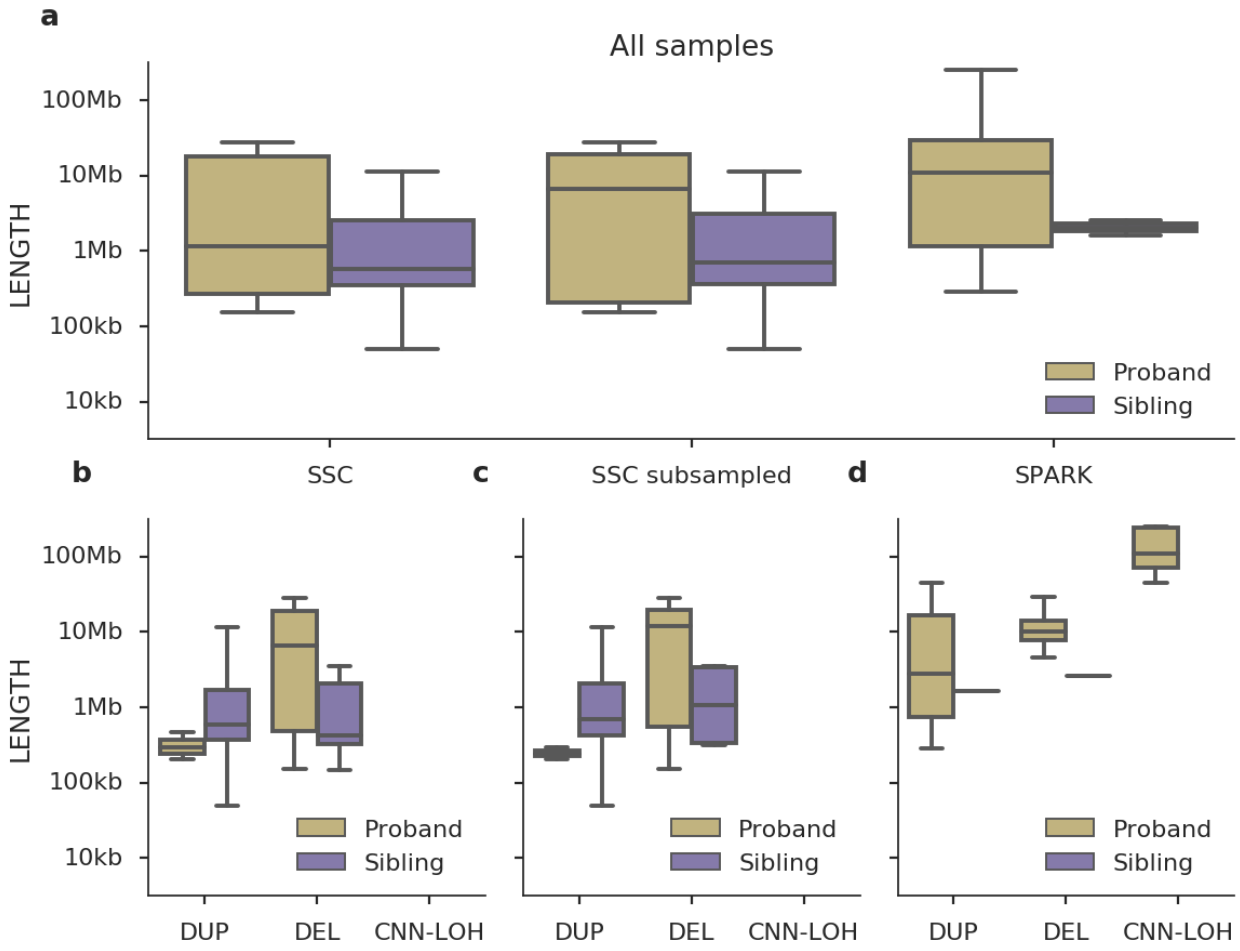
Supplementary Figure 5: Proportion of mCNVs that were located on the paternal haplotype and maternal haplotype, respectively; data are rate ± 95% CI. For this analysis, CNN-LOH events are considered to be located on the haplotype which is duplicated (i.e. the haplotype that becomes homozygous in the mosaic state). Of the 64 mCNVs detected, 28 were located on the paternal haplotype and 35 on the maternal haplotype. The haplotype of one event could not be established because parental genotypes were unavailable. The p-value is calculated using a two-sided binomial test that assumes mCNVs arise with equal probability on either parental haplotype.



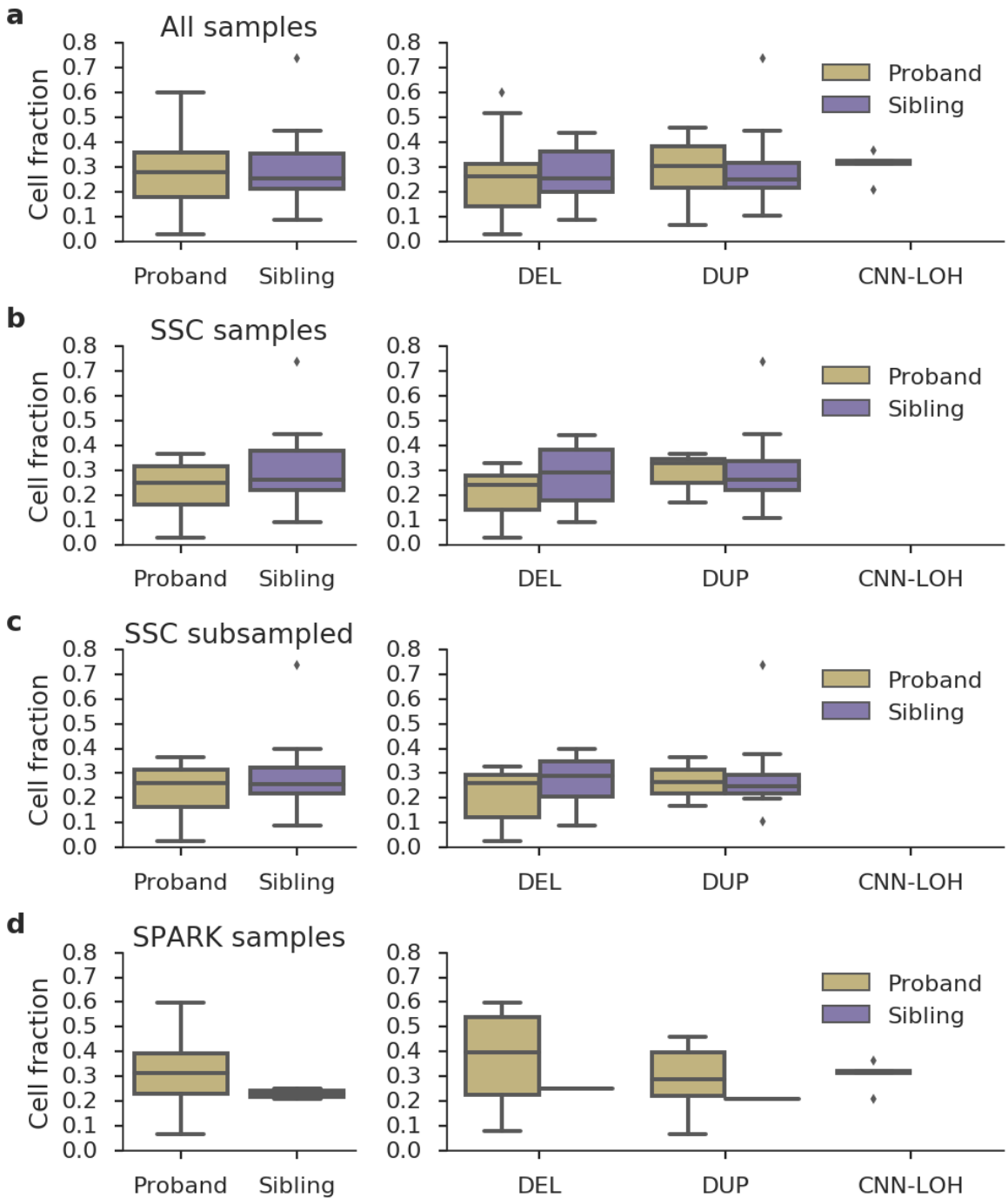
Supplementary Figure 6: Discovery of mosaic CNVs in SSC samples was robust to genotyping density. **a**, Events successfully recalled (blue) and not recalled (green) in SSC samples after randomly subsampling genotyped sites to the density of the SPARK array (~630K variants). **b**, recall rate after subsampling SSC arrays. **c**, Average change in boundary, (change in left boundary + change in right boundary) / 2, for each subsampled event as a function of event cell fraction. Average boundary change decreases as cell fraction increases (Spearman R = -0.63; P = 4.2×10^{-4}). **d,** concordance (defined as Jaccard similarity between original calls and recalled events after subsampling) as a function of cell fraction.



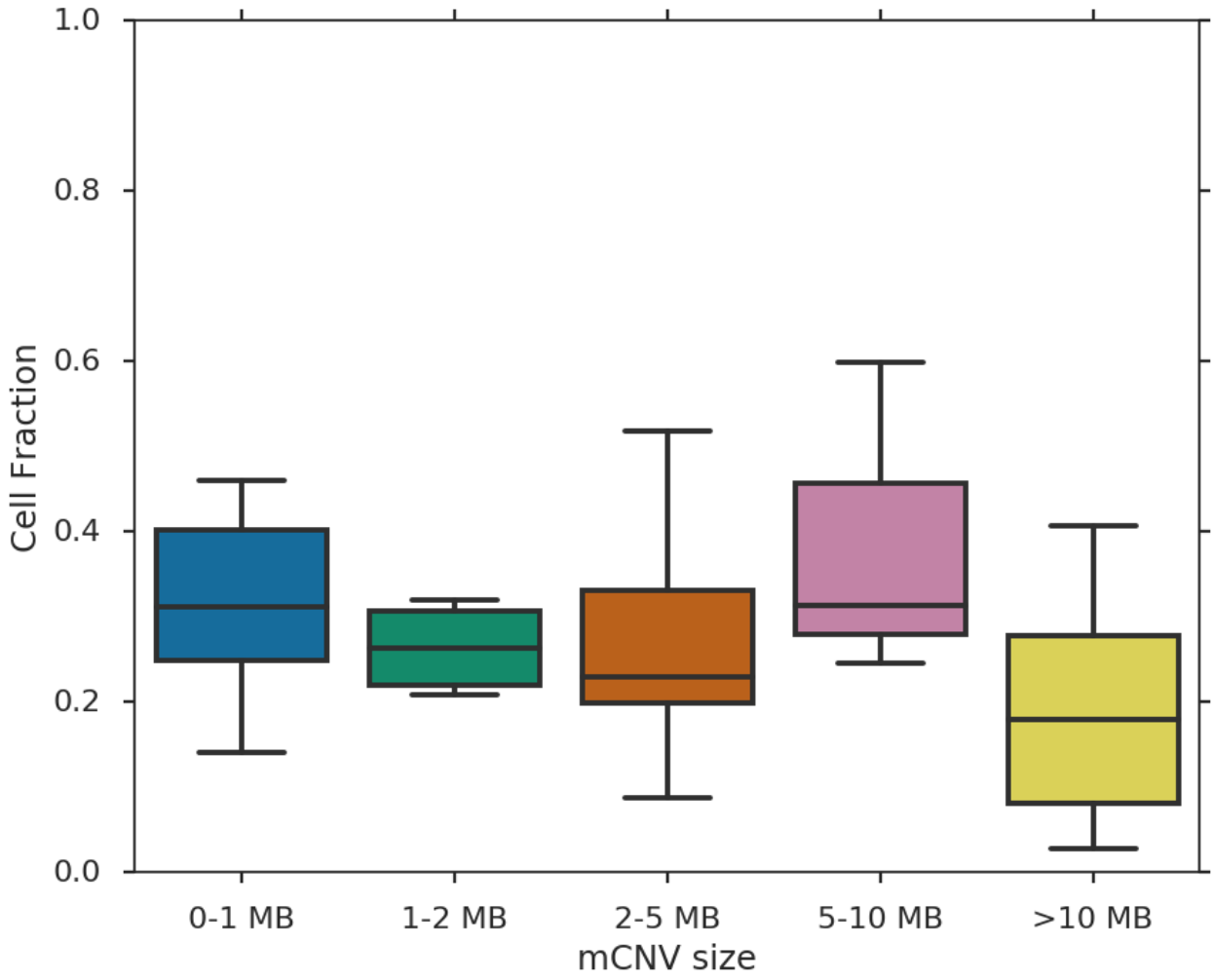
Supplementary Figure 7: Plots as in Fig. 1 but excluding SSC mCNVs that were not detected after subsampling SSC genotyped positions to the density of genotyped positions in SPARK samples (n=6 events excluded; n=58 events included of which 42 were in probands and 16 were in siblings); see Fig. 1 legend for definitions of statistical tests. Qualitative results do not change after excluding these events. P-values in **b** are slightly larger than in Fig. 1, and P-values in **c** are unchanged because the SSC events that were not recalled were smaller than the 4 Mb threshold. See Methods for box plot definitions.



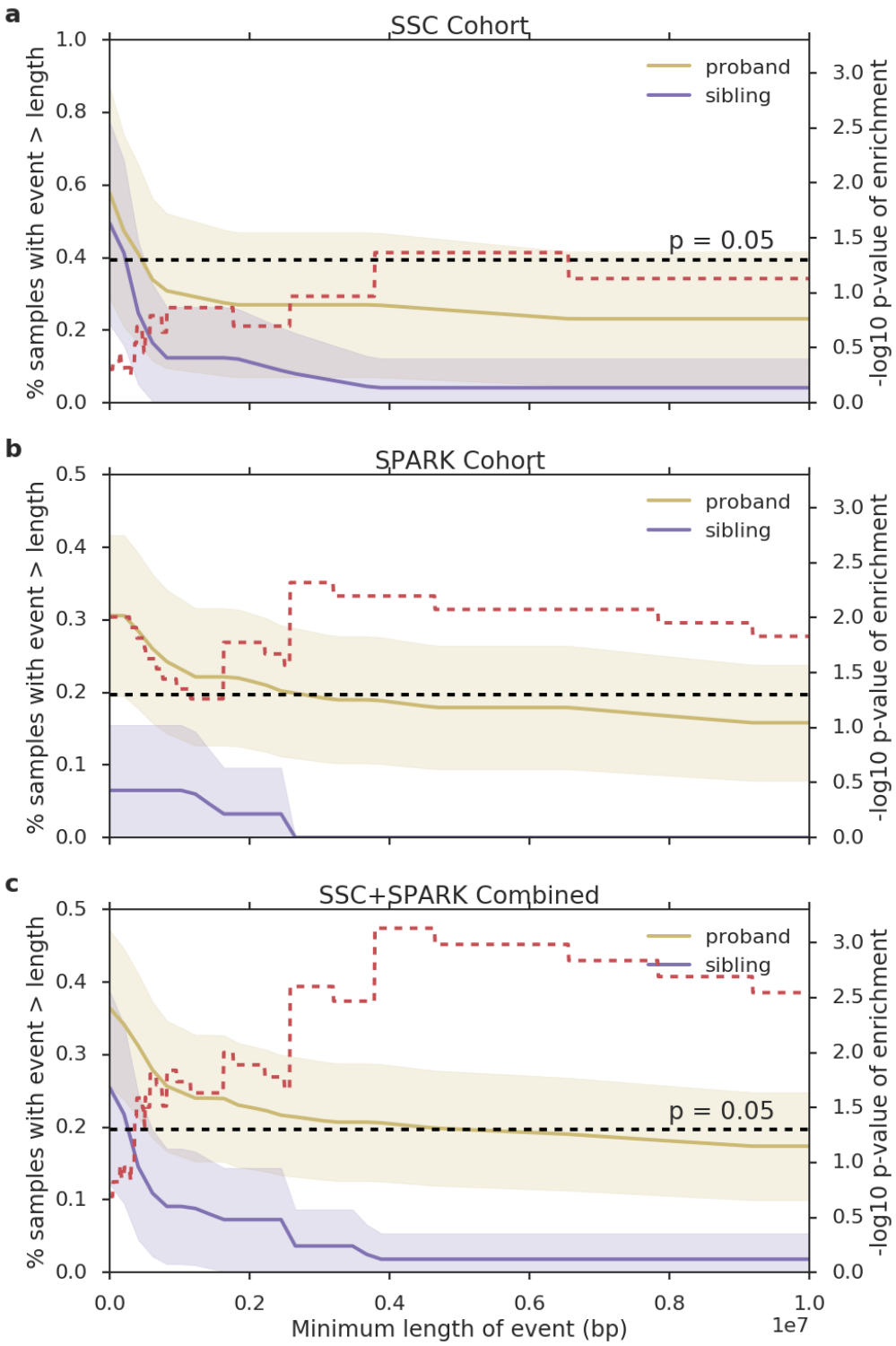
Supplementary Figure 8: **a,** Length distribution of mosaic CNVs in SSC samples, in SSC samples after subsampling to SPARK genotyping density, and in SPARK samples (n SSC=33, n SSC subsampled=27, n SPARK=31). After subsampling SSC genotypes to standardize detection sensitivity in SSC and SPARK, the median length difference between events in SSC vs. SPARK probands is not significant (median in SSC subsampled probands: 6.55 Mb; median in SPARK probands: 10.96 Mb; P=0.057, one-sided Mann-Whitney U-test). **b, c, d** Events in SSC, SSC after subsampling genotypes, and SPARK, respectively, stratified by copy-number state (see Supplementary Table 1 and 2 for sample sizes for each copy state). See Methods for box plot definitions.



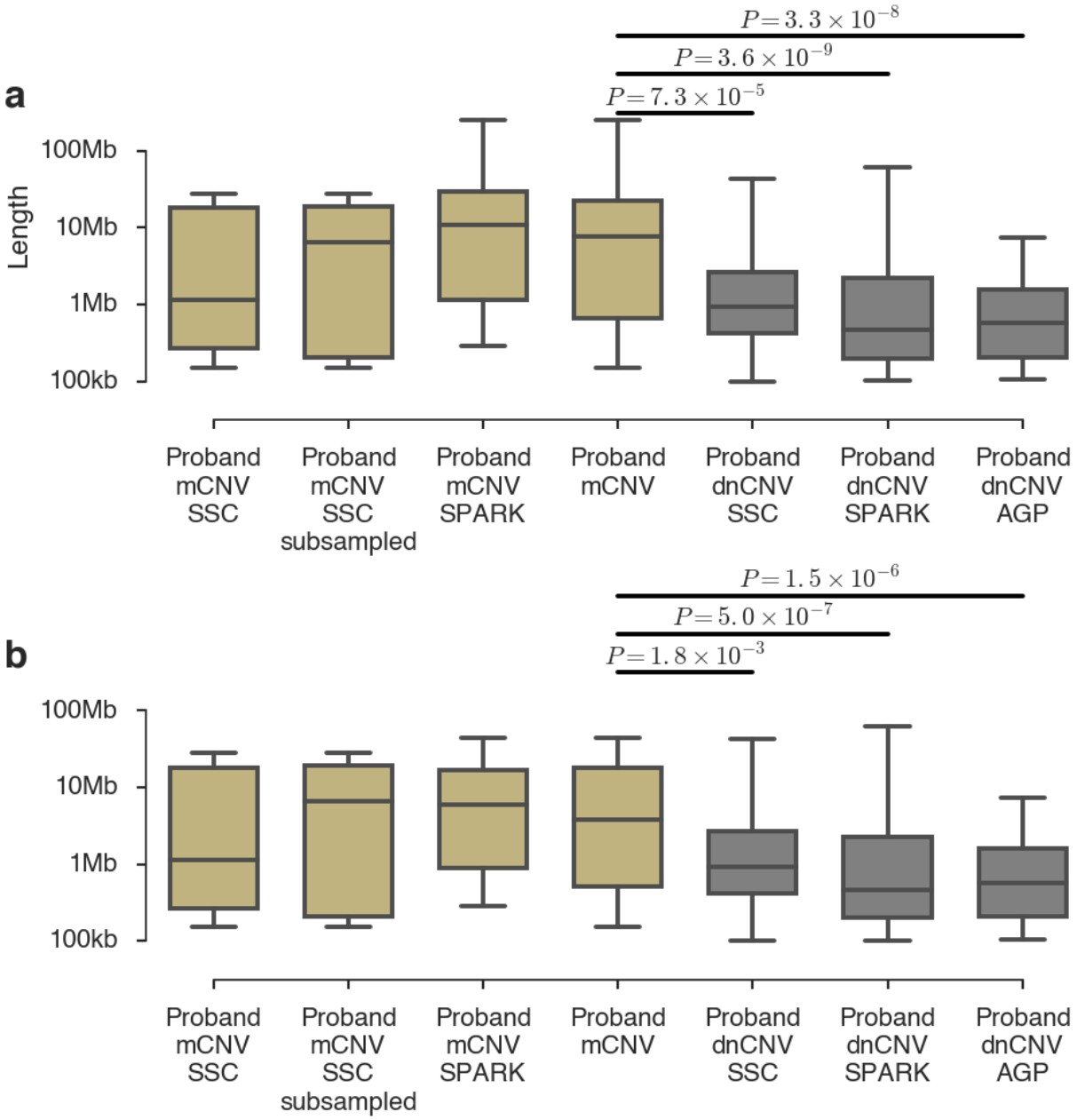
Supplementary Figure 9: The cell fraction of mosaic CNVs stratified by ASD status (left column) and by CNV type (right column) for **a,** all samples, **b,** SSC samples, **c,** SSC samples after subsampling genotypes, and **d,** SPARK samples (see Supplementary Table 1 and 2 for sample sizes). No differences are statistically significant between probands and siblings. See Methods for box plot definitions.



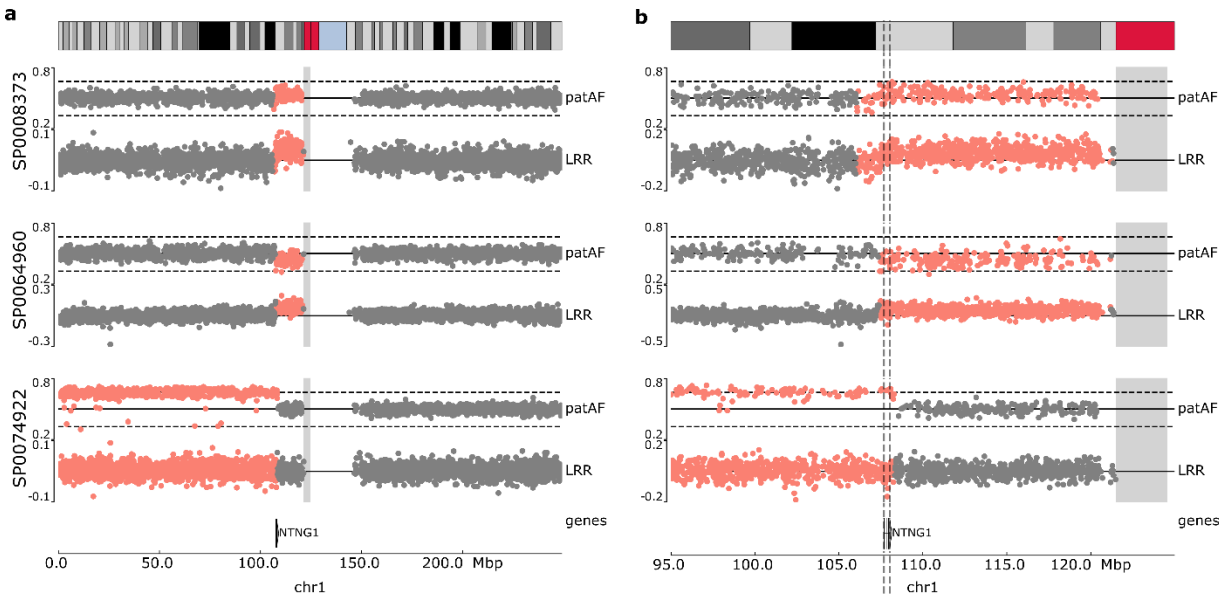
Supplementary Figure 10: Cell fraction distribution of mCNVs stratified by length of event (see Supplementary Table 1 for sample sizes). CNN-LOH events were excluded because we imposed a strict minimum cell fraction of 0.2 for CNN-LOH events in order to filter potential clonal hematopoiesis events. See Methods for box plot definitions.



Supplementary Figure 11: **a,** Burden by length in SSC, **b,** SPARK, and **c,** a combined analysis. Data as mean (solid line) ± 95% CI (shaded regions). The dashed red line provides the -log10 p-value (corresponding to the y-axis on the right) of the burden test at a given minimum mCNV length using one-sided Fisher's exact tests. The black line indicates the P = 0.05 significance level. P-values are not corrected for choice of size threshold.



Supplementary Figure 12: a, Length of mosaic CNVs in SSC probands (n=16), SSC probands after sensitivity correction (n=13), and SPARK probands (n=29) compared to length of *de novo* CNVs in either SSC probands (n=228), SPARK probands (n=330) or Autism Genome Project (AGP) probands (n=159). **b,** as in **a** but with mosaic CNN-LOH events removed. P-values calculated using one-sided Mann-Whitney U-tests. See Methods for box plot definitions.



Supplementary Figure 13: **a**, Paternal allele fraction (patAF) and LRR plots from three individuals with events which either start (SP0008373, SP0064960) or end (SP0074922) immediately adjacent to *NTNG1*. The former two are duplications while the latter is a CNN-LOH. **b,** Zoom in of neighborhood around *NTNG1* showed that all three events encompass *NTNG1* (boundaries marked by dashed grey lines) but do not terminate within *NTNG1*.

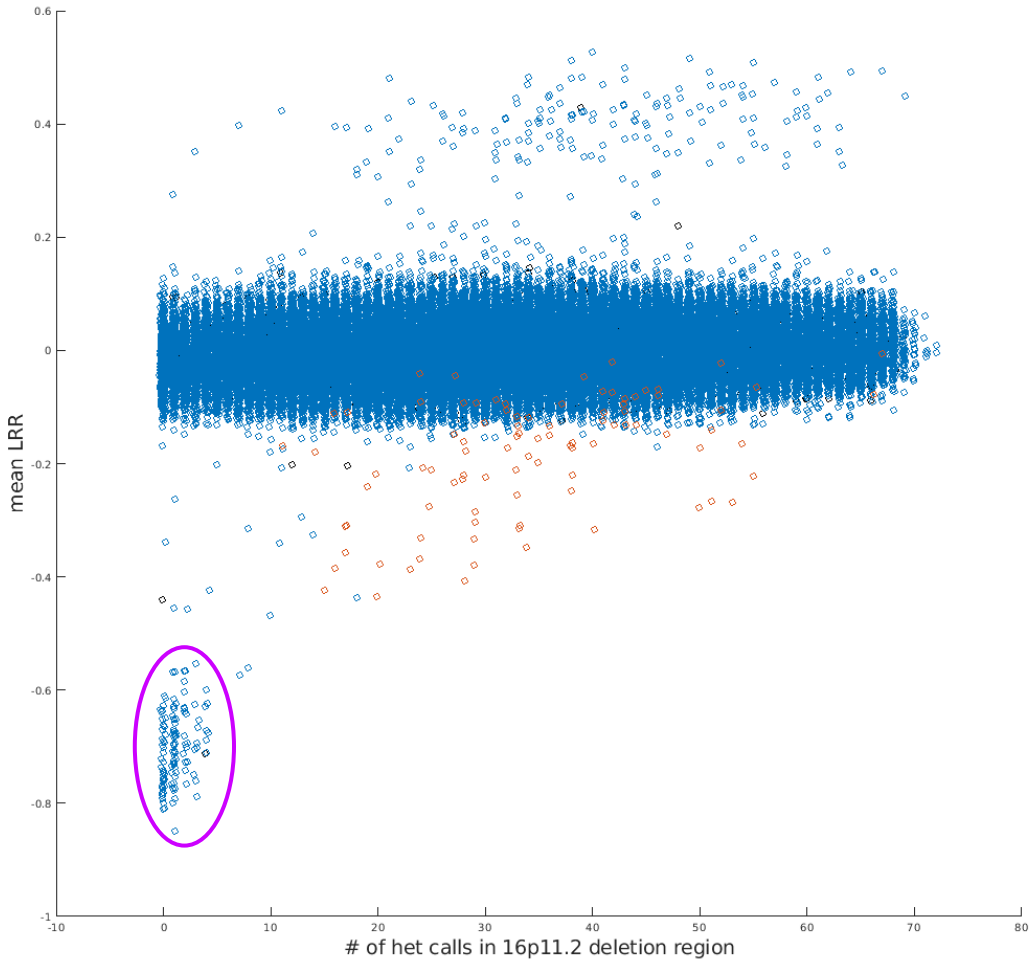
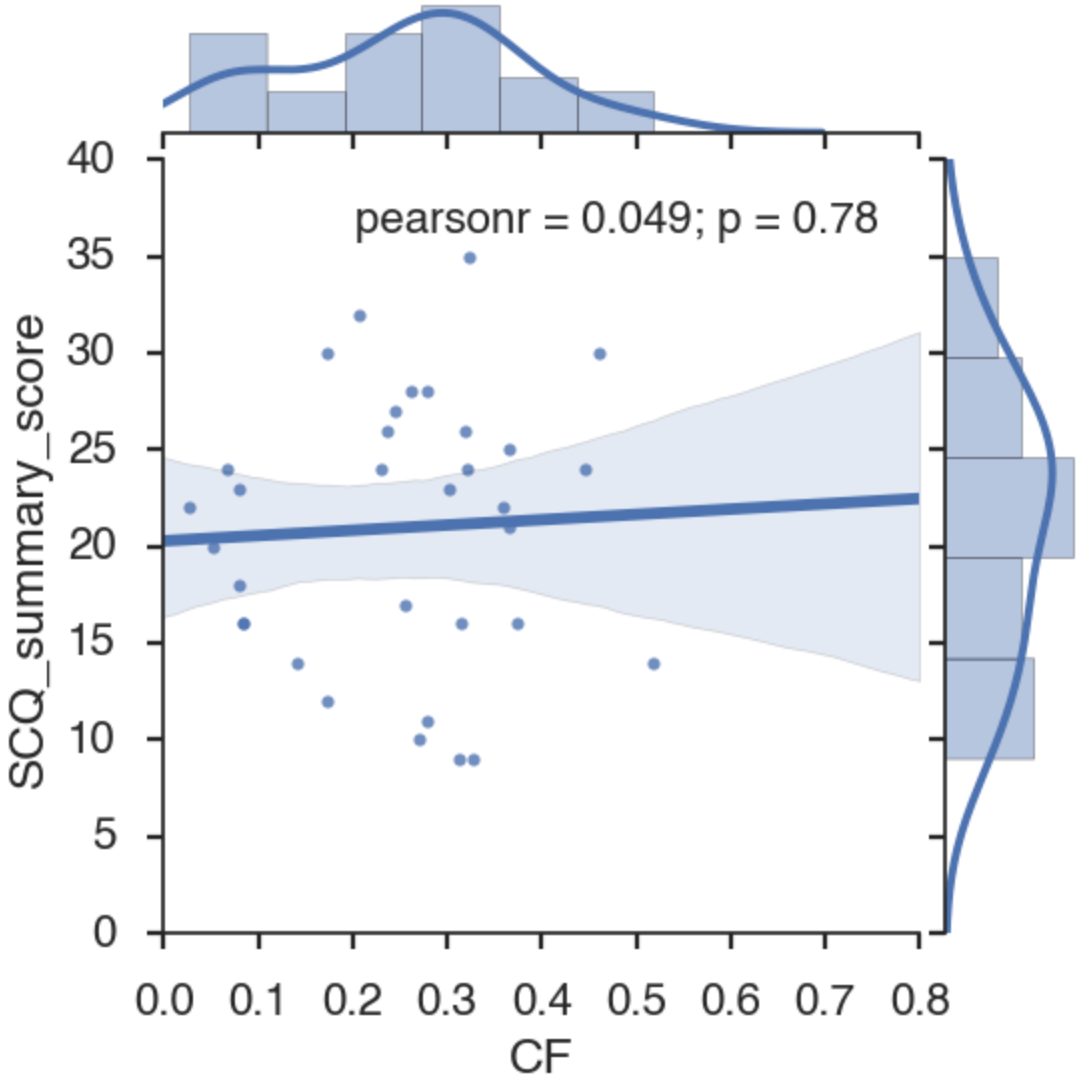
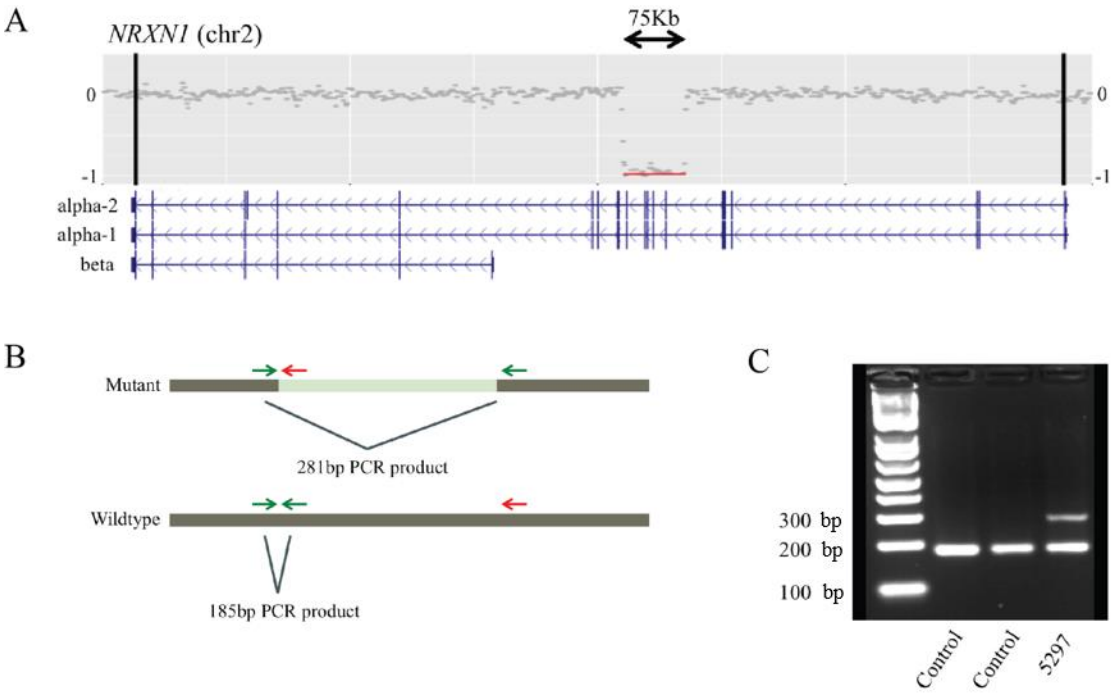


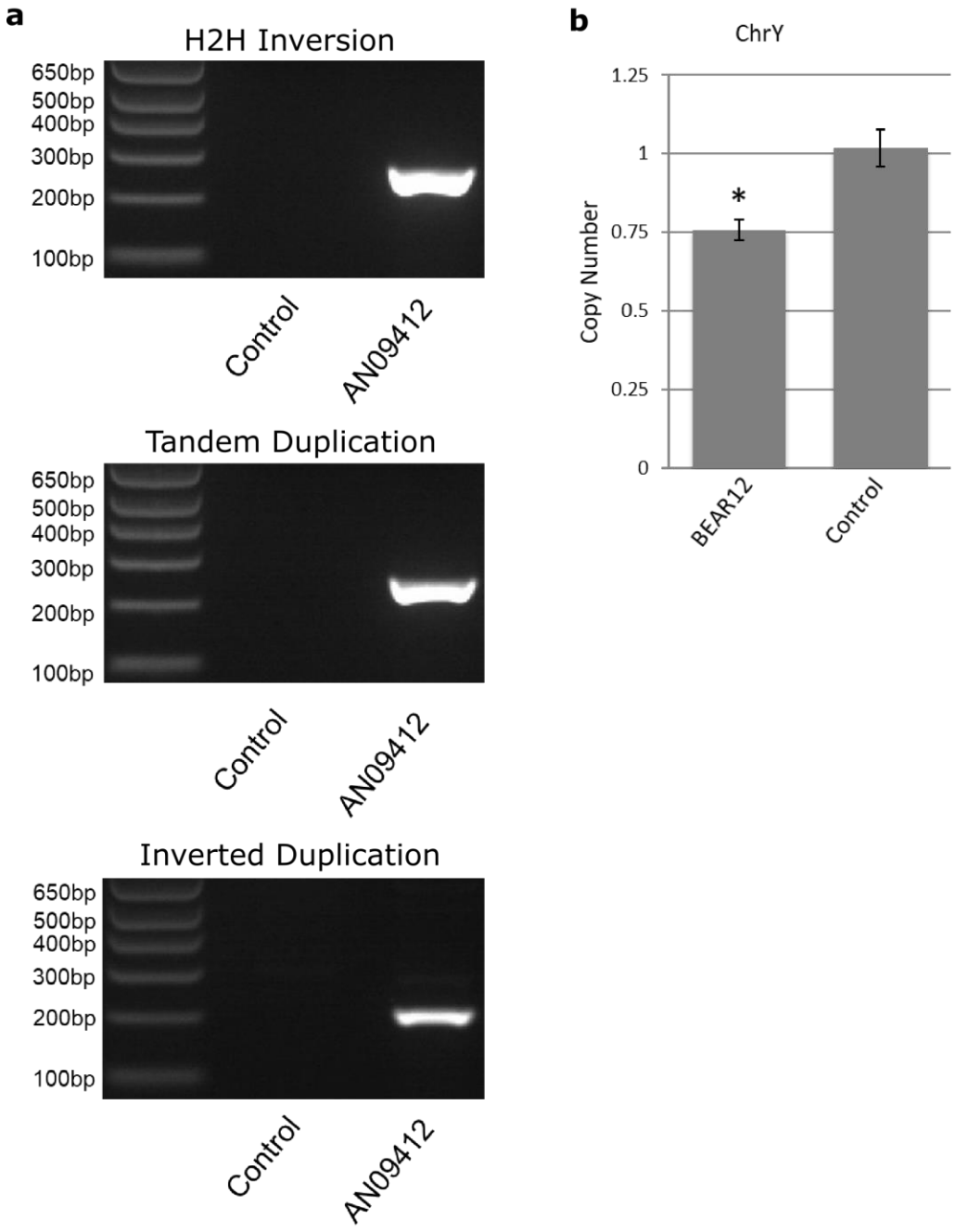
Figure 14: Mean genotyping intensity (LRR) vs. number of heterozygous SNP calls in the *16p11.2* deletion region (chr16:29,655,864-30,195,048, hg19) in UK Biobank samples. Samples with <5 hets and mean LRR <-0.5 were identified as germline *16p11.2* deletion carriers (purple circle). Red markers indicate carriers of mosaic *16p11.2* deletions previously identified[1].



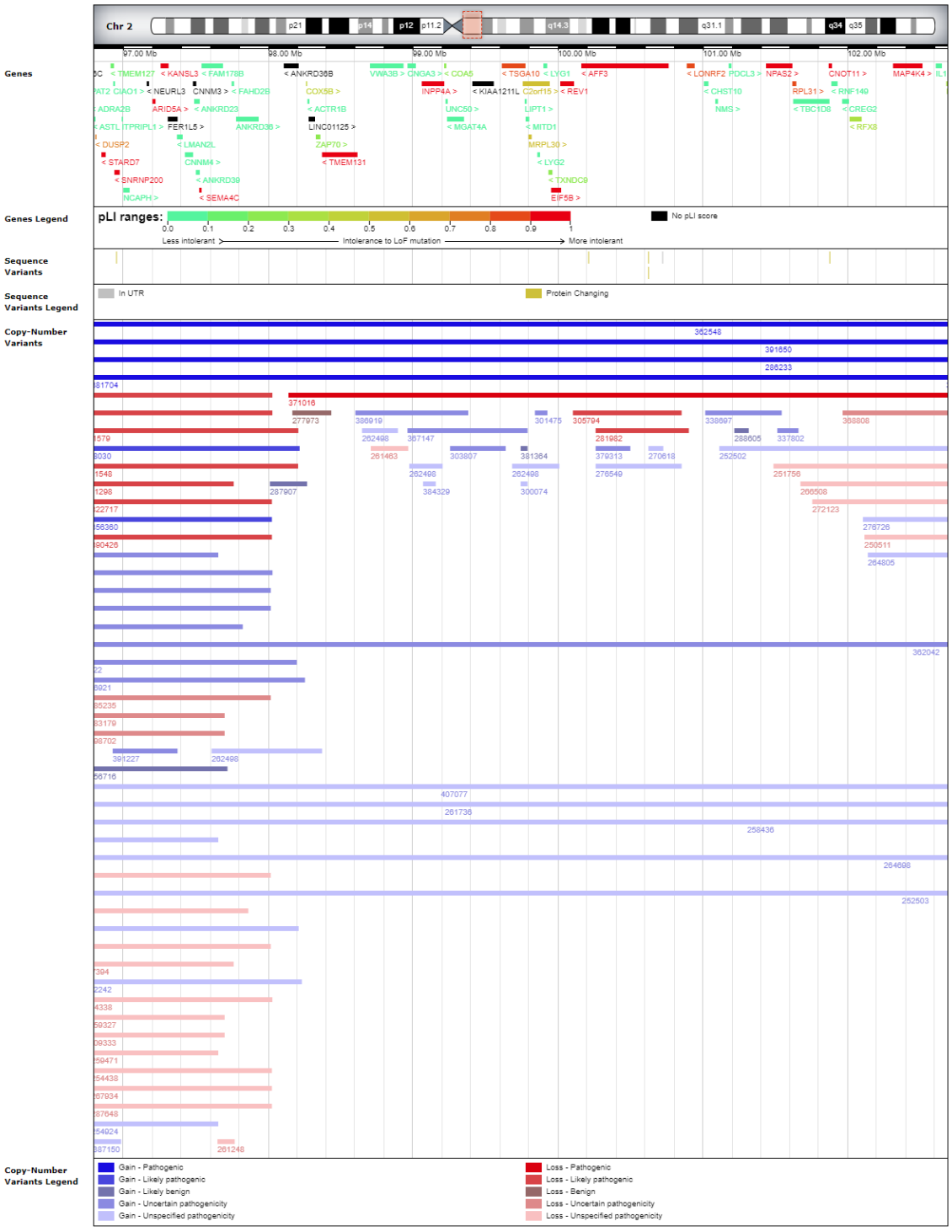
Supplementary Figure 15: Correlation between mosaic event cell fraction in probands and ASD phenotypic severity as quantified by SCQ summary score. Regression mean (solid line) ± 95% CI (shaded region).



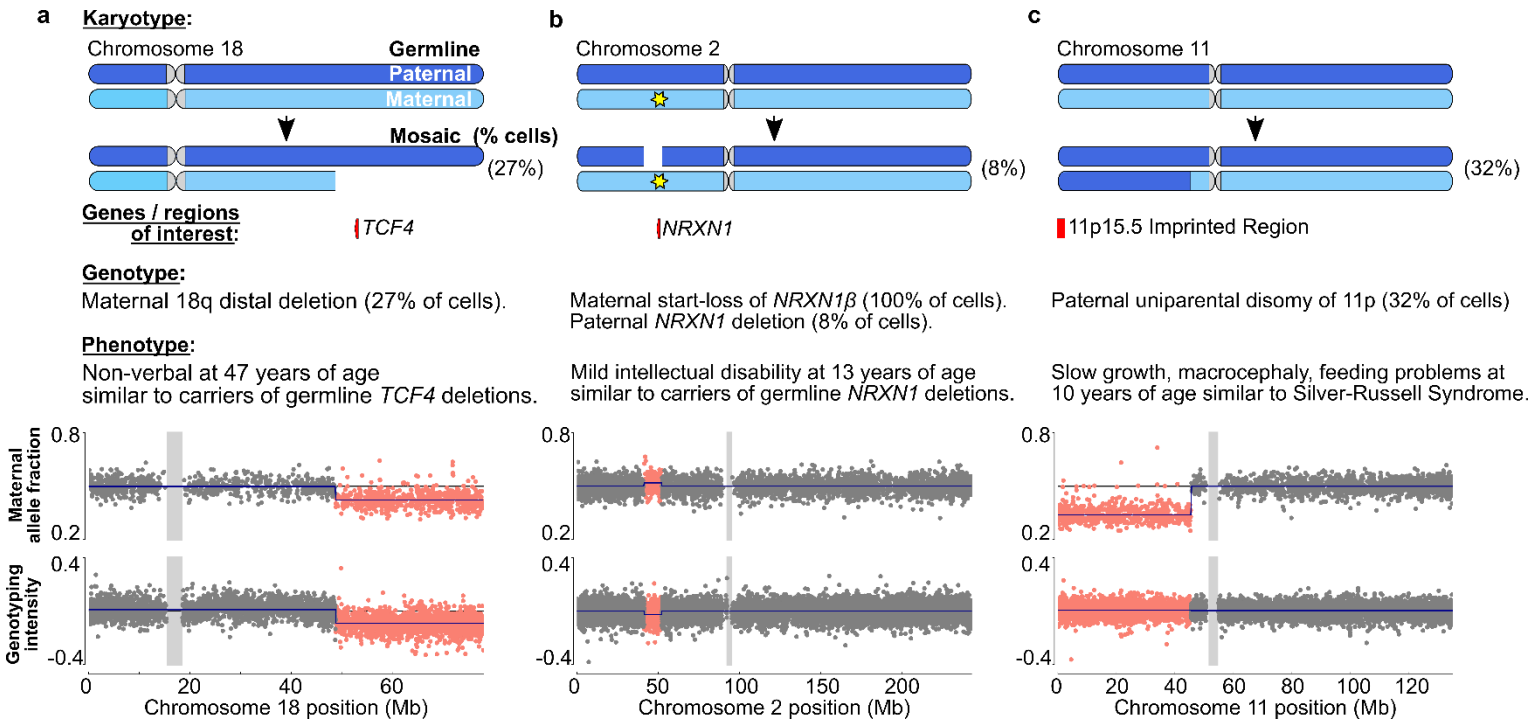
Supplementary Figure 16: Validation of a germline *NRXN1* deletion in brain WGS sample UMB5297. A, Difference from diploid copy number as calculated from WGS read depth. The red line indicates the ~75 kb germline deletion (2:50,731,007-50,805,387) within the alpha form of *NRXN1*. B, design of primers for mutant sequence resulting from the germline deletion and wildtype sequence present in the human reference genome. C, Gel electrophoresis of PCR products from mutant and wildtype primers. An extra band is visible in UMB5297, resulting from PCR amplification of the mutant sequence. Results were confirmed with three independent experimental replicates.



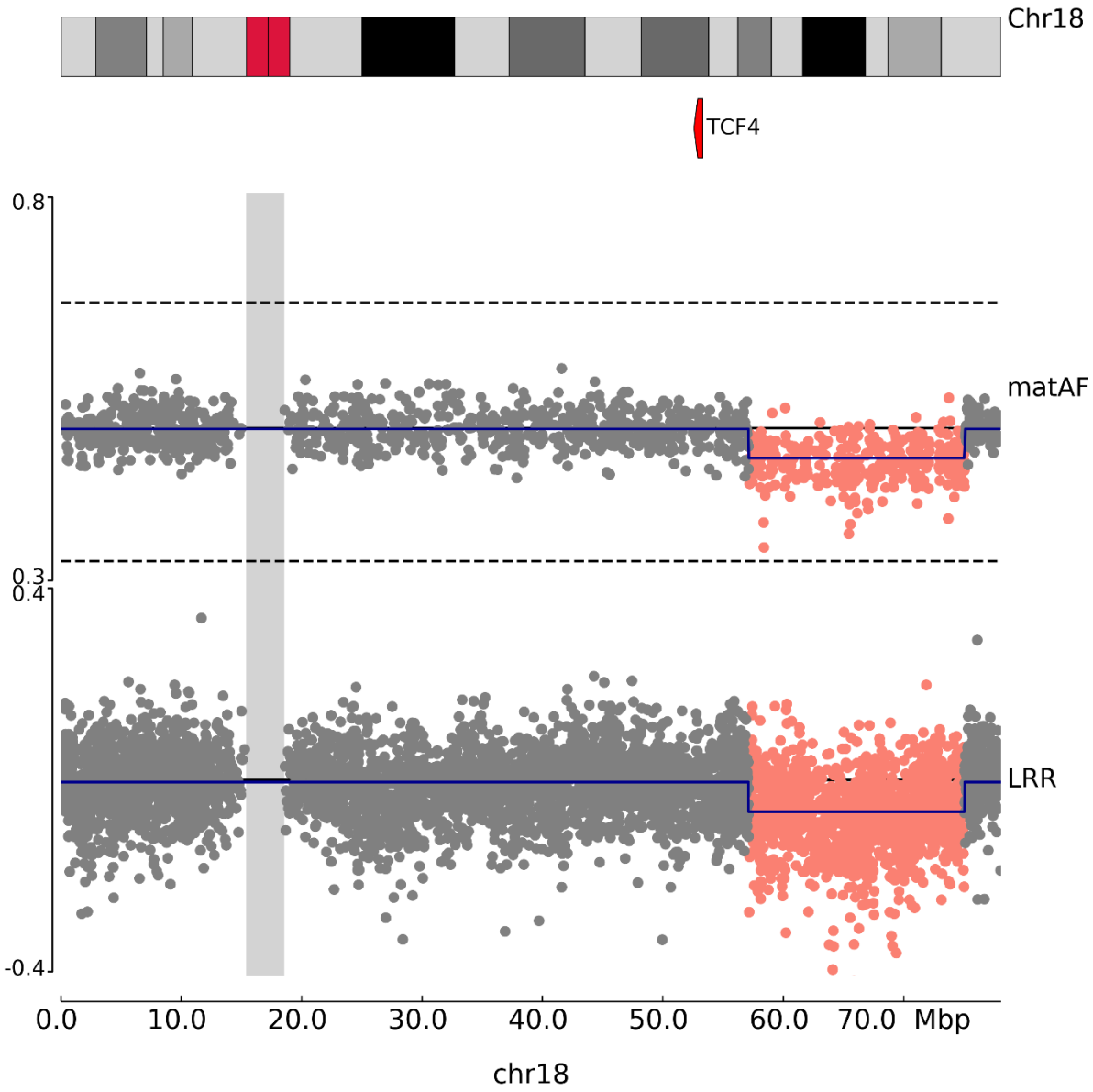
Supplementary Figure 17: Additional validation experiments for mCNVs discovered in WGS data. **a**, Gel electrophoresis of PCR products from primers for the breakpoints corresponding to the T2T inversion (top panel), tandem duplication (middle panel) and inverted duplication (bottom panel) in DNA from AN09412 and DNA from a control brain. White lines are visible when a successful PCR product was formed**.** Results were confirmed with three independent experiments per event. Original images of the gels are provided as Supplementary Data 1. **b,** copy number of chrY in brain tissue from ABN_XVTN and tissue from a control brain inferred from ddPCR. Data are mean ± approximate 95% CI from n=3 experimental replicates.



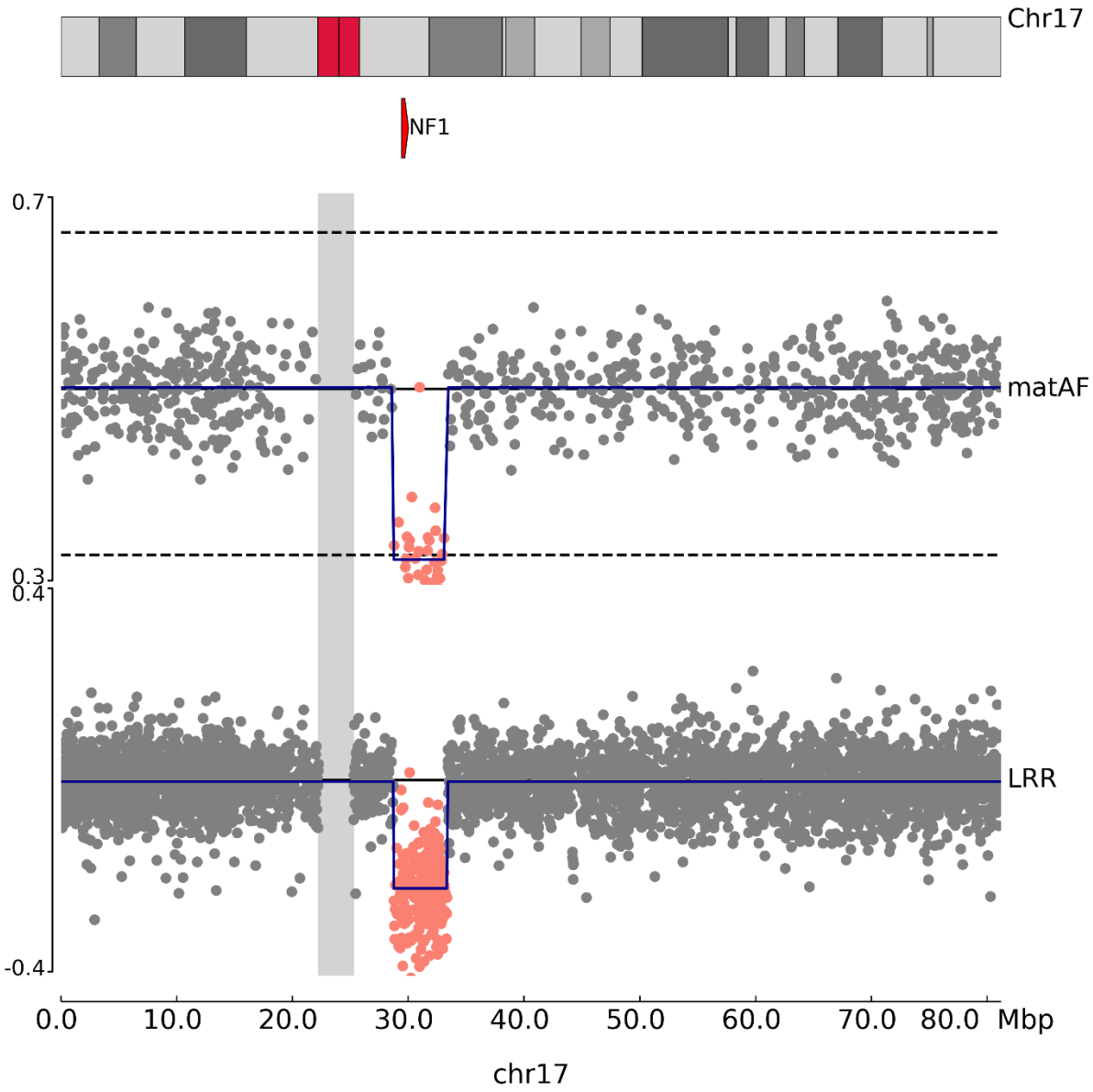
Supplementary Figure 18: Image from the DECIPHER browser[2] of reported pathogenic / likely pathogenic gains (bright blues) and pathogenic / likely pathogenic losses (bright reds) disrupting the same genomic region as the complex mosaic duplication identified in AN09142. Benign CNVs (grey blue / grey red) and CNVs of uncertain or unspecified pathogenicity are also included (light blues / light reds).



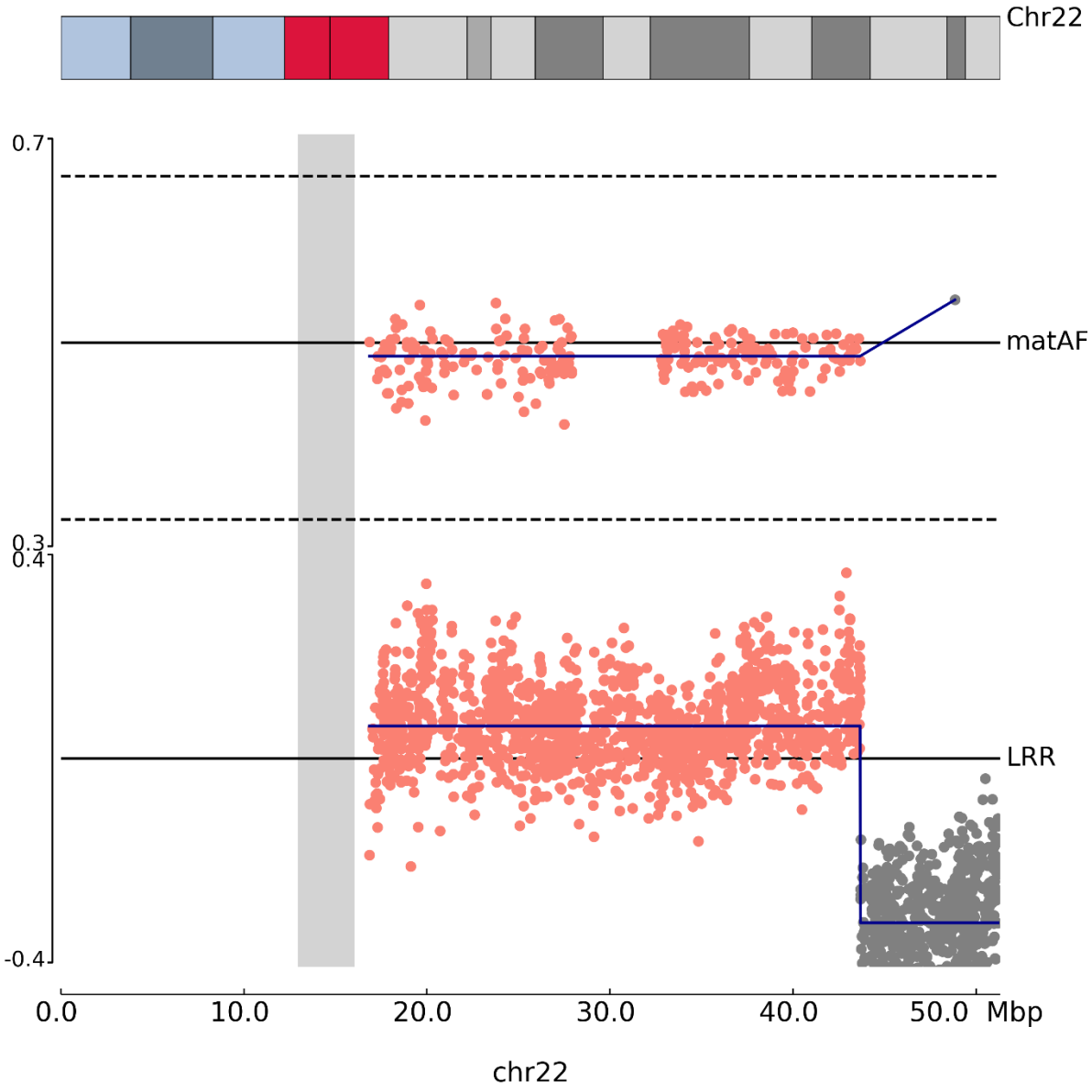
Supplementary Figure 19: a, 18q distal deletion. **b,** *NRXN1* deletion. **c,** 11p CNN-LOH. Top, diagrams of mosaic mutations altering inherited chromosomes in a fraction of cells. Paternal and maternal haplotypes are colored dark and light blue, respectively, with genes or regions of interest labeled below. Middle, description of mutations and observed clinical phenotypes. Bottom, maternal allele fraction at heterozygous SNPs (binned into groups of two adjacent SNPs) and total genotyping intensity (log R-ratio; LRR) at all SNPs genotyped on the chromosome (binned into groups of four adjacent SNPs); SNPs within the mCNV are highlighted.



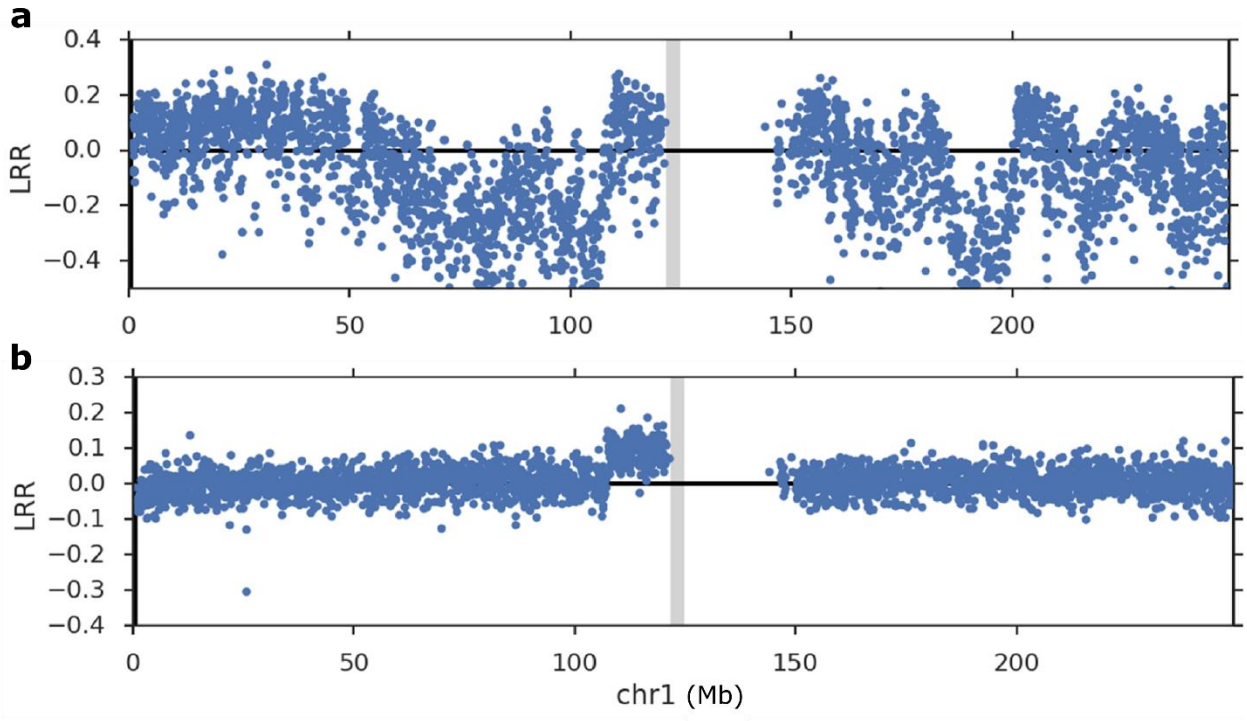
Supplementary Figure 20: Maternal allele fraction (matAF) and LRR plots of the chr18q distal deletion carried by SSC proband 12246.p1. The matAF plot includes only heterozygous SNPs. The boundaries of the event estimated by MoChA are chr18:57,102,326-75,041,151. The location of *TCF4* (chr18:52,942,850-53,068,756) is indicated with a red triangle.



Supplementary Figure 21: Maternal allele fraction (matAF) and LRR plots of mosaic deletion in SPARK proband SP0095456 overlapping *NF1* (indicated by red triangle). The matAF plot contains only heterozygous SNPs.



Supplementary Figure 22: Maternal allele fraction (matAF) and LRR plots of chr22 in SPARK proband SP0025588. The proband carries a mosaic duplication of paternal chr22 in 6.7% of cells (indicated by pink dots), and a germline de novo deletion on his maternal 22q haplotype (indicated by grey dots in LRR plot); a germline deletion converts all SNPs to a hemizygous state, and thus no heterozygous SNPs are present within the germline deletion except a few genotyping errors.



Supplementary Figure 23: Log-R Ratio (LRR) across chromosome 1 in SPARK proband SP0064960 **a,** before and **b,** after applying principal-components denoising to the signal (Methods). After denoising, a mosaic duplication is evident from 107 Mb to 121 Mb.

2. Supplementary Notes

1. 13391.s1 chr4 event and its relationship to *TET2*

We investigated whether the complex, chromothripsis-like event found in SSC sibling 13391.s1 (Supplementary Table 1, Supplementary Fig. 4a) might be a clonal hematopoietic event. One duplicated segment (chr4:106,390,734-107,509,199) occurred in the neighborhood of *TET2* (chr4:106,067,842-106,200,960), a common target of driver mutations in clonal hematopoiesis[3,4] (Supplementary Fig. 4b); we hypothesized that the event might affect *TET2* and thus have resulted in clonal expansion in blood. We thus obtained whole-genome sequencing for this individual from SFARI Base and manually inspected the region for the exact location of the left-hand breakpoint depicted in Supplementary Fig. 4c. We identified split reads and discordant reads indicating a rearrangement in which chr4:106,286,478 is connected to chr4:125,620,145, consistent with the breakpoints estimated from BAF and LRR genotyping-intensity measures by MoChA (chr4:106,390,734 and chr4:125,610,859, respectively). While *TET2* does not appear to be altered by the event, we cannot rule out that regulation of *TET2* may be disrupted, resulting in aberrant expression.

We also contacted the Simons Foundation Autism Research Initiative to obtain additional information on the individual. They noted the individual was ~6 years old at time of assessment. An assessing clinician noted mild delay in gross motor and personal care per the Vineland Adaptive Behavior Scales but no other clinically significant behavioral findings. The family did not note any significant medical or psychiatric history for the child. The child did not receive an IQ test.

We thus found no conclusive evidence either for the event arising due to clonal hematopoiesis or for it being an early embryonic event, which we suspect would manifest phenotypically. Given the age of the child at assessment, it is possible that a developmental or neuropsychiatric phenotype was not yet identifiable. We were unable to reach the family for further follow-up.

Given that we could not conclusively determine that the event was clonal hematopoietic, we opted to be conservative and include the event in all main figure analyses. Exclusion of the event results in burden p-values that are slightly more significant and does not change any findings reported in the main text.

2. Recalling mCNVs from subsampled SSC genotypes

We measured the robustness of our mosaic CNV detection algorithm to genotyping density by randomly subsampling genotyped variants from the SSC arrays to match the density of genotyped variants on the SPARK array (630K genotyped variants, Supplementary Table 12). Genotyped positions were randomly retained with probability P_{array}, such that the expected number of retained positions was equal to the number of genotyped positions in SPARK samples (626,789 positions). We then reran the MoChA algorithm on each SSC mCNV carrier and evaluated whether the original event was recalled in the subsampled data. An event was considered recalled if 1) it was still detected in the subsampled genotype data; 2) the subsampled event call reciprocally overlapped the original event by >75%; and 3) the two events were of the same type (e.g. gain or loss).

We recalled 27 of the 33 original mCNVs in SSC (Supplementary Figure 6a,b; Supplementary Table 2). Moreover, CNV boundaries were highly consistent between the original and subsampled events (median difference between event boundaries: 11.3 kb). As expected, differences between original and subsampled event boundaries decreased with increasing event cell fraction (Spearman R = -0.63; P = 4.2×10^{-4}, Supplementary Figure 6c). This translated into a high concordance (aka intersection of events divided by union of events) between original and subsampled events (median Jaccard similarity: 0.987; Supplementary Figure 6d).

The six events not recalled were all <1 Mb in size (median size: 474 kb; max size: 748 kb) consistent with the challenge of identifying short mosaic CNVs. Since the events were all below the 4 Mb threshold for which we calculated burden in our primary analyses, the exclusion of these events does not alter the statistical significance of the burden analyses presented in main text (Supplementary Figure 7).

3. Robustness of length difference between mCNVs in probands vs. siblings

Because relatively few mCNVs were detected in SPARK siblings relative to SSC siblings, the conclusion that mCNVs in probands are significantly longer than in siblings may be confounded by a combination of two factors: 1) our increased sensitivity to call small mCNV events in SSC samples compared to SPARK due to the higher genotyping density of the SSC arrays, and 2) the

discovery of chromosome-level CNN-LOH events in SPARK Probands but not in SSC samples. We thus accounted for each confounder separately and together. Events in probands remain significantly longer than in siblings after 1) including only SSC events that were identified after subsampling SSC genotypes to the density of the SPARK array ($P = 4.0\times10^{-3}$ Mann-Whitney U-test); including only gains and losses but not CNN-LOH events in the comparison ($P = 6.7\times10^{-3}$ Mann-Whitney U-test); and 3) including only subsampled SSC events and gains and losses in the comparison (P = 0.015, Mann-Whitney U-test).

4. Cell fraction distribution of mCNVs.

The mosaic fraction of mCNVs in probands and siblings had a median of ~0.25 (Supplementary Fig. 9). We suspected this feature may reflect statistical limitations of our methods and not underlying biology because 1) MoChA's sensitivity to low cell-fraction mCNVs increases as mCNV length increases; 2) the copy number state of low-cell fraction mCNVs is not easily inferred (Supplementary Fig. 2); and 3) we had imposed a strict minimum cell fraction of 0.2 on CNN-LOH events taken forward to analysis. We therefore stratified the cell-fraction distribution by mCNV size (Supplementary Fig. 10) after removing CNN-LOH events. Consistent with the hypothesis that the cell fraction distribution is a result of methodological limitations, the median cell fraction decreased with increasing length ($P = 1.3\times10^{-3}$, comparison of cell fractions between mCNVs 0-1 Mb in size vs. mCNVs >10 Mb in size, Mann-Whitney U-test).

5. Choosing a size threshold for burden analyses

In a pooled analysis of mCNVs in SSC and SPARK samples, we found that probands carry a burden of mCNVs >400 kb in size (38 proband carriers, 8 sibling carriers; OR = 2.17, 95% confidence interval = 1.01-4.65; P = 0.026 one-sided Fisher's exact test). This burden threshold is unchanged when accounting for increased sensitivity to small CNVs in SSC by removing events that were not detected after subsampling SSC genotyped positions to match the genotyping density of SPARK arrays (35 proband carriers, 7 sibling carriers; OR = 2.28, 95% CI = 1.01-5.14; P = 0.025). However, this burden is driven almost exclusively by events >4 Mb; after excluding the events >4 Mb in size, the burden for events >400 kb is no longer significant (13 proband carriers, 7 sibling carriers; P-value = 0.73). Indeed, after excluding events >4 Mb, there is no threshold for which the difference is significant. Given the rarity of mosaic CNVs,

analysis in a larger cohort will be necessary to confidently determine a lower bound on mCNV size for which probands carry a burden relative to siblings. We thus opted to use the conservative size threshold of 4 Mb throughout.

6. Identification of germline de novo CNVs in SPARK samples

To confirm the conclusion that mosaic CNVs in probands are significantly longer than *de novo* CNVs in probands, we identified putative dnCNVs from SPARK probands and siblings using MoChA. To do so, we first limited to SPARK probands and siblings for which both parents were also genotyped (N=6661 probands and N=3074 siblings), and we removed samples with evidence of contamination with DNA from another individual (Methods). To filter out inherited CNVs, we removed any event that reciprocally overlapped an event in a biological parent by at least 50% or reciprocally overlapped a CNV from the 1K Genomes Project Phase 3 CNV data set by at least 50%. We aggressively filtered out potentially mosaic events by 1) removing all events with an absolute LRR deviation from zero less than 0.11 and 2) removing duplications with a BAF deviation from zero less than 0.15 and a LRR deviation from zero less than 0.3. We further required that any duplication overlap at least 5 heterozygous SNP sites and that any deletion contain at most 10 heterozygous SNP calls.

Importantly, the above filters were designed to remove events that would confound a comparison of mosaic and germline *de novo* CNVs. However, such strict choices limit our sensitivity to detect true *de novo* CNVs. Nonetheless, our set of dnCNVs in probands and siblings demonstrates several features consistent with known patterns in ASD simplex quartets. 1) Probands carry a significant burden of dnCNVs relative to their siblings (295 Proband carriers vs. 96 sibling carriers; $P = 1.1\times10^{-3}$ one-sided Fisher's exact test); 2) Proband dnCNVs are significantly longer than those in siblings ($P = 4.5\times10^{-5}$ Mann-Whitney U-test); and 3) dnCNVs in probands include numerous examples of classic dnCNVs including three *16p11.2* CNVs, 18 duplications within *15q11-13*, one *7q11.23* CNV, and two *1q21.1* CNVs. We therefore believe that our SPARK dnCNV callset contains high-quality *de novo* CNVs and thus provided a reasonable validation set to confirm that mosaic CNVs are longer than *de novo* CNVs.

7. Mosaic CNV recurrence analysis

While most mCNVs we detected had unique, non-recurrent breakpoints (Supplementary Table 1), we observed two nearly identical mosaic duplications of 1pcen in SP0008373 and SP006490 (Supplementary Fig. 13a, top two panels). The event in SP0008373 had an estimated cell fraction of 17.1% while the event in SP006490 had an estimated cell fraction of 27.8%. We confirmed using `PLINK --genome`[5] that the two samples were unrelated (PI_HAT = 0.0398). Note, to calculate this statistic, we first calculated the minor allele frequency for each SNP across all genotyped individuals in SPARK and supplied these estimates to `PLINK –genome` using the `--read-freq` option. Additionally, we observed a CNN-LOH event in SP0074922 (Supplementary Fig. 13a, bottom panel) which ended in the same vicinity as the two duplications started. Thus, we observed three breakpoints within a 3 Mb region (chr1:106,052,011-108,310,224). While this region includes *NTNG1*, a gene that has been associated with ASD[6], the probability of observing three or more breakpoints within a 3 Mb region is P=0.107 under the assumption that breakpoints are distributed uniformly at random across the genome.

8. Lack of mosaic analogues of ASD-associated germline *de novo* CNVs

We sought to test whether the lack of mosaic analogues of ASD-associated germline *de novo* CNVs (ASD-dnCNVs) was significant. Compared to the reported recurrence rates of ASD-dnCNVs among SSC probands (55 / 132, Sanders et al. 2015 Table 1, ref [7]), the observed rate of mosaic analogues (0 / 40, Supplementary Table 4) is significantly less than expected ($P = 4.23\times10^{-6}$ one-sided Fisher's exact test), were mosaic CNVs to have the same recurrence patterns as germline CNVs in ASD. Considering only *16p11.2* CNVs – the dnCNV most observed in ASD probands – the observed rate of mosaic analogues is still significantly less than expected (0 / 40 mosaic analogues vs. 13 / 132 ; P = 0.037 one-sided Fisher's exact test). The rate of ASD-dnCNVs and mosaic analogues was not significantly different among siblings (0 / 19 mosaic analogues vs 3 / 34 ASD-dnCNVs; P = 0.28 one-sided Fisher's exact test).

9. Analysis of mosaic CNVs in *16p11.2* in the UK Biobank

16p11.2 de novo germline CNVs (both gains and losses) are strongly associated with ASD[7,8], yet we observed no mosaic analogues of such events in either SSC or SPARK probands or siblings. We looked for mosaic analogues of the germline *16p11.2* CNVs (duplications or deletions)

among mCNVs identified in the larger UK Biobank cohort[1]. We observed 73 events contained within the boundaries the extended *16p11.2* CNV locus (chromosome 16: 28,000,000-31,000,000, hg19), of which all were deletions. The carriers of these events were heavily biased to be female (Observed: 56 females, 17 males; Expected: 40 females, 33 males; P = 8.7e-5, Fisher's Exact Test) as has been previously reported[9]. We next checked whether these events could have arisen due to age-related clonal hematopoiesis; we observed a small, non-significant increase in prevalence of *16p11.2* losses with age (mean age = 57.3 (s.e.m. = 0.89) years in carriers; mean age = 56.5 years in the full cohort), in contrast with the strong increase in prevalence of other mosaic events with age (mean age = 59.5 (s.e.m. = 0.1) years in carriers of other events), suggesting an early-developmental origin of these events.

10. Putative damaging variants within mosaic CNVs

Beyond directly disrupting a gene by deletion, bifurcation, or dosage alterations, a mosaic CNV can convert a damaging variant that is heterozygous in the germline state into a hemizygous or homozygous variant in the mosaic state: suppose an individual inherited a damaging variant on paternal 1p; if the individual acquires a mosaic UPD of paternal 1p (CNN-LOH in which the paternal haplotype replaces the maternal haplotype), the damaging variant will be homozygous (i.e. present on both haplotypes) in the mosaic cells. If the individual acquires a mosaic deletion of 1p on the maternal haplotype, the damaging variant will be hemizygous in the mosaic cells. We thus searched for putative damaging variants (stop-gain, start-loss, frameshift, splice-site, or missense variant with CADD Phred score >20, ref. 8) converted to a hemizygous or homozygous state by a mCNV (Methods).

We found several examples of this phenomenon (Supplementary Table 7). SPARK proband SP0069140 carried a *NRXN1* start-loss variant on the maternal haplotype that was converted to a hemizygous state through a mosaic deletion of the paternal *NRXN1*. Additionally, every CNN-LOH event converted at least one putative damaging variant to a homozygous state. However, none of these variants occurred within known ASD genes, and their clinical relevance was of unknown significance.

11. Germline CNVs in brain tissue with plausible connection to ASD

We identified 9 disease-relevant germline SVs in ASD brains (Supplementary Table 10), revealing potential causes of disease in several previously unsolved cases. These include a case of germline 15q-duplication in case AN06365, a well-documented syndromic cause of autism usually occurring in the de novo state[10], as well as a 10Mb germline copy gain in case UMB1638 in chromosome *20q13.2-13.33*. Duplications of this region have been previously identified in individuals with ASD[11,12] . A small germline deletion in the ASD risk gene *NRXN1* was found in one ASD case, UMB5297 and validated via PCR, Supplementary Fig. 16). Importantly, this 75kb deletion affects five critical exons and is likely causative of disease in this case[13–15].

12. Mosaic CNVs correlate with individual-level clinical observations

We looked for evidence directly linking specific mCNVs to reported clinical symptoms. By cross-referencing the genes disrupted by individual mCNVs with known syndromes associated with those genes, we identified four probands in which the mosaic mutation appeared directly linked to the proband's symptoms (Fig. 19, Supplementary Fig. 20).

Two probands carried mosaic 18q distal deletions removing 29.2 and 17.9 Mb of sequence from chromosome 18 in 27% and 16% of cells, respectively (Fig. 19a, Supplementary Fig 20). Germline 18q distal deletions (OMIM: 601808) are well-characterized causes of intellectual disability and ASD[16]. While causal genes in this region have not been fully characterized, larger deletions which encompass the gene *TCF4* and all distal genes produce profound intellectual disability and little to no verbal communication (Pitt-Hopkins Syndrome, OMIM:610954), while smaller deletions encompassing fewer genes result in relatively mild cognitive impairment[17,18]. Of the two mosaic 18q distal deletions, the larger one extended beyond *TCF4* and the smaller one did not. Consistent with their germline analogues, the proband with the *TCF4* deletion was non-verbal, while the proband with the smaller deletion had an IQ in the normal range (full-scale IQ = 97, NVIQ = 98) and mild adaptive impairment by the Vineland Adaptive Composite Standard Score (VSS=66).

One proband carried a mosaic deletion of *NRXN1* (OMIM:614332), which has a well-documented (but incompletely penetrant) association with ASD, intellectual disability and speech delay[19,20]. The proband's mosaic deletion encompassed the entirety of *NRXN1* on his paternal haplotype in 8% of cells. Furthermore, on the maternal haplotype, the individual carried

an inherited, rare start-loss variant of the beta isoform of *NRXN1* (Fig. 19b, Supplementary Table 7, Supplementary Note 10). At age 13, the proband was reported to have an IQ in the range of 55-69 and slight language delay consistent with at least mild intellectual disability. Furthermore, the proband had ADHD, a condition also often associated with *NRXN1* deletions[21]. The mother exhibited no evidence of intellectual disability despite carrying the start-loss variant in *NRXN1*. This finding is consistent with previous reports that *NRXN1* LoF variants are incompletely penetrant[22,23] and suggests that the observed germline-mosaic compound heterozygosity contributes to the proband's clinical symptoms.

Another proband carried an acquired paternal uniparental disomy (UPD) of nearly the entirety of 11p in 32% of cells (Fig. 19c). The 11p15.5 region contains numerous paternally and maternally imprinted genes, and germline disruption of this region is known to produce syndromic growth disorders: Beckwith-Wiedemann syndrome (BWS, OMIM:130650; an overgrowth condition associated with hypermethylation) and Silver-Russell syndrome (SRS, OMIM:180860; an undergrowth condition associated with hypomethylation)[24]. The proband exhibited abnormally slow growth, macrocephaly, and feeding difficulties, all of which are common symptoms of SRS. SRS has also been associated with increased risk of ASD and intellectual disability[25]. While paternal 11p UPD is usually associated with BWS and maternal 11p UPD is usually associated with SRS, cases in which imprinting disruption led to the opposite phenotype have been reported[26]. Interestingly, we observed one other case of a mosaic 11p UPD impacting *11p15.5* in a sibling with a reported genetic condition (and therefore excluded from our main analyses) (Supplementary Table 14). This individual also had a reported (unspecified) growth disorder.

These case studies reinforce our observation of an overall burden of large mCNVs in ASD probands with concrete examples in which specific mCNVs potentially underlie the disorder via a variety of plausible mechanisms. We also explored six other cases in which mCNVs deleted ASD genes but a direct connection to reported phenotypes was less clear due to the phenotypic heterogeneity of ASD[27] and the limited phenotype data provided for each proband (Supplementary Notes 13 and 14, Supplementary Fig. 21 and 22).

13. Additional mosaic CNVs with plausible connections to proband phenotype

ASD, related neuropsychiatric disorders, and co-morbid medical conditions are phenotypically diverse, and the limited, standardized phenotypic information provided for each proband in SSC and SPARK is generally not detailed enough to allow clinical diagnosis of particular syndromes. We were therefore unable to confirm with high confidence that four mCNVs which clearly disrupted known ASD genes were likely responsible for the observed phenotype. However, in each case, we found that the observed phenotypes were fairly consistent with the mCNV being causative.

Rare mutations in *FOXP1* are associated with ASD[6] and intellectual disability[28,29]. SSC individual 11270.p1 carries a 19.0 Mb deletion encompassing *FOXP1* in 28% of cells. Consistent with germline disruption of *FOXP1*, the proband has evidence of significant intellectual disability (Non-verbal IQ, NVIQ=49; Vineland Adaptive Summary Score, VSS=63, reported phrased speech delay).

Loss of function mutations and microdeletions affecting *SETD5* have been implicated in ASD[30], ID[31], and developmental delay[32]. SSC proband 13362.p1 carries a 6.6 Mb deletion encompassing *SETD5* in 31% of cells. While the proband has reported speech delay and cognitive function below the population average (FSIQ=82, NVIQ=83, VSS=86), they do not meet the criteria for ID. We hypothesize that this may represent a mosaic phenotype which is milder than an equivalent germline analogue.

De novo and mosaic LOF variants in *BAZ2B* have been associated with ASD[33,34] and moderate intellectual disability[35]. SSC proband 11671.p1 carries an 11.9 Mb deletion encompassing *BAZ2B* in 33% of cells. Consistent with previous reports, the proband has a medical diagnosis of mild ID (DSM IV diagnosis code 317).

SPARK proband SP0095456 carried a 4.6 Mb deletion encompassing *NF1* in 52% of cells (Supplementary Fig. 21). Mosaic *NF1* microdeletions have previously been reported; unlike their germline counterparts, mosaic *NF1* losses are not associated with increased risk of intellectual disability or medical conditions including congenital heart abnormalities or neurofibroma tumors[36]. The mCNV we detected is significantly larger than most mosaic *NF1*-microdeletions (typically 1.2 Mb)[36,37]; nonetheless the proband's phenotype is mild, consistent with those previous reports: a learning disability but no cognitive impairment or medical conditions. While *NF1* deletions are often observed in clonal hematopoiesis, the mCNV we

detected seems unlikely to be present only in leukocytes due to the young age of the proband and the high cell fraction of the event. Indeed, in the UK Biobank, we observed 49 focal *NF1* deletions (occurring between 29.42 and 39.70 Mb on chr17). Of these, only eight (16%) had cell fraction estimated to exceeded that of the event observed in SP0095456 (>52%). Seven of the eight individuals were older than 60 years of age, and all eight were older than 55 years of age. The proband was 7 years of age at time of sample evaluation.

14. Other events with unverified disruption of ASD genes or connection to phenotype

Two probands carry events whose breakpoints appeared to bifurcate ASD genes. In this case, the CNV would act like a loss-of-function variant. However, we were unable to confirm the exact location of the breakpoints because genome sequencing (either whole-exome or whole-genome) was not available for these samples. A third proband carried an event which appeared to be a mosaic rescue of a germline deletion, but the phenotypic data available for the proband was not sufficiently detailed to confirm if the rescue resulted in a milder than expected phenotype.

SPARK proband SP0016887 carried a 667 kb duplication in 26% of cells whose left breakpoint appeared to bifurcate *TRIO*. *TRIO* has been previously reported to be enriched for variants (inherited and *de novo*) likely to affect ASD risk[30]. The proband exhibits speech delay (single words at 31 months). However, neither WES nor WGS sequencing was available to confirm the exact location of the left breakpoint of this event.

SSC proband 13674.p1 carried a 22.9 Mb deletion of 4p in 8.3% of cells. Simultaneously we also detected a 23.7 Mb mosaic event of 12p in 8.5% of cells (Supplementary Table 15). While we were unable to classify the chr12 event due to its low cell-fraction, we suspect that it is a duplication arising from an unbalanced translocation wherein the first 23 Mb of chr12 was duplicated and replaced the first 23 Mb of 4p. The chr12 event contains multiple genes previously implicated in ASD including *C12orf57, CACNA1C*, *GRIN2B*; additionally, the left breakpoint appears to bifurcate the ASD gene *SOX5*. Neither WES nor WGS was available to confirm this hypothesis.

SPARK proband SP0025588 carried a 7.5 Mb *de novo* germline deletion of maternal 22q (*22q13.2-q13.33*). Such events have been reported in multiple cases of individuals with developmental disorders[38] and typically cause Phelan-McDermid syndrome[39] (OMIM: 606232).

Interestingly, the proband also exhibited a mosaic rescue genotype: 6.7% of cells carried a duplicated paternal chr22 and thus returned 22q to a diploid state (Supplementary Fig. 22). We hypothesize that this mosaic rescue should decrease the severity of the individual's phenotype. While the proband has reported symptoms consistent with Phelan-McDermid Syndrome including being non-verbal at 14 years old, having intellectual disability and motor delay, the limited phenotype data provided no evidence either for or against decreased disease severity.

3. Members of the Brain Somatic Mosaicism Network Consortium

The Brain Somatic Mosaicism Network (BSMN) Consortium was supported by NIMH grants U01MH106874, U01MH106876, U01MH106882, U01MH106883, U01MH106883, U01MH106884, U01MH106891, U01MH106891, U01MH106891, U01MH106892, U01MH106893, U01MH108898 awarded to: Nenad Sestan (Yale University), Flora Vaccarino (Yale University), Fred Gage (Salk Institute for Biological Studies), Christopher Walsh (Boston Children's Hospital), Peter J. Park (Harvard University), Jonathan Pevsner (Kennedy Krieger Institute), Andrew Chess (Icahn School of Medicine at Mount Sinai), John V. Moran (University of Michigan), Daniel Weinberger (Lieber Institute for Brain Development), and Joseph Gleeson (University of California, San Diego).

Members of The Brain Somatic Mosaicism Network:

Boston Children's Hospital: Christopher Walsh, Javier Ganz, Mollie Woodworth, Pengpeng Li, Rachel Rodin, Robert Hill, Sara Bizzotto, Zinan Zhou

Harvard University: Alice Lee, Alison Barton, Alissa D'Gama, Alon Galor, Craig Bohrson, Daniel Kwon, Doga Gulhan, Elaine Lim, Isidro Ciriano Cortes, Lovelace J. Luquette, Maxwell Sherman, Michael Coulter, Michael Lodato, Peter Park, Rebeca Monroy, Sonia Kim, Yanmei Dou

Icahn School of Medicine at Mt. Sinai: Andrew Chess, Attila Gulyas-Kovacs, Chaggai Rosenbluh, Schahram Akbarian

Kennedy Krieger Institute: Ben Langmead, Jeremy Thorpe, Jonathan Pevsner, Sean Cho

Lieber Institute for Brain Development: Andrew Jaffe, Apua Paquola, Daniel Weinberger, Jennifer Erwin, Jooheon Shin, Richard Straub, Rujuta Narurkar

Mayo Clinic: Alexej Abyzov, Taejeong Bae

NIMH: Anjene Addington, David Panchision, Doug Meinecke, Geetha Senthil, Lora Bingaman, Tara Dutka, Thomas Lehner

Rockefeller University: Laura Saucedo-Cuevas, Tara Conniff

Sage Bionetworks: Kenneth Daily, Mette Peters

Salk Institute for Biological Studies: Fred Gage, Meiyan Wang, Patrick Reed, Sara Linker

Stanford University: Alex Urban, Bo Zhou, Xiaowei Zhu

Universitat Pompeu Fabra: Aitor Serres, David Juan, Inna Povolotskaya, Irene Lobon, Manuel Solis, Raquel Garcia, Tomas Marques-Bonet

University of California, Los Angeles: Gary Mathern

University of California, San Diego: Eric Courchesne, Jing Gu, Joseph Gleeson, Laurel Ball, Renee George, Tiziano Pramparo

University of Michigan: Diane A. Flasch, Trenton J. Frisbie, Jeffrey M. Kidd, John B. Moldovan, John V. Moran, Kenneth Y. Kwan, Ryan E. Mills, Sarah Emery, Weichen Zhou, Yifan Wang

University of Virginia: Aakrosh Ratan, Mike McConnell

Yale University: Flora Vaccarino, Gianfilippo Coppola, Jessica Lennington, Liana Fasching, Nenad Sestan, Sirisha Pochareddy

4. References

1. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *bioRxiv* 653691 (2019) doi:10.1101/653691.
2. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics* **84**, 524–533 (2009).
3. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine* **371**, 2477–2487 (2014).
4. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine* **371**, 2488–2498 (2014).
5. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
6. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
7. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
8. 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).
9. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
10. Finucane, B. M. *et al.* 15q Duplication Syndrome and Related Disorders. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

11. Guo, H. *et al.* Genome-wide copy number variation analysis in a Chinese autism spectrum disorder cohort. *Sci Rep* **7**, 44155 (2017).
12. Maini, I. *et al.* Prematurity, ventricular septal defect and dysmorphisms are independent predictors of pathogenic copy number variants: a retrospective study on array-CGH results and phenotypical features of 293 children with neurodevelopmental disorders and/or multiple congenital anomalies. *Ital J Pediatr* **44**, 34 (2018).
13. Ching, M. S. L. *et al.* Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 937–947 (2010).
14. Lowther, C. *et al.* Molecular characterization of NRXN1 deletions from 19,263 clinical microarray cases identifies exons important for neurodevelopmental disease expression. *Genet Med* **19**, 53–61 (2017).
15. Viñas-Jornet, M. *et al.* A common cognitive, psychiatric, and dysmorphic phenotype in carriers of NRXN1 deletion. *Mol Genet Genomic Med* **2**, 512–521 (2014).
16. Semrud-Clikeman, M. *et al.* Cognitive ability predicts degree of genetic abnormality in participants with 18q deletions. *Journal of the International Neuropsychological Society* **11**, 584–590 (2005).
17. Zweier, C. *et al.* Haploinsufficiency of TCF4 Causes Syndromal Mental Retardation with Intermittent Hyperventilation (Pitt-Hopkins Syndrome). *The American Journal of Human Genetics* **80**, 994–1001 (2007).
18. Feenstra, I. *et al.* Genotype–phenotype mapping of chromosome 18q deletions by high-resolution array CGH: An update of the phenotypic map. *American Journal of Medical Genetics Part A* **143A**, 1858–1867 (2007).

19. Dabell, M. P. *et al.* Investigation of NRXN1 deletions: Clinical and molecular characterization. *American Journal of Medical Genetics Part A* **161**, 717–731 (2013).
20. Chen, X. *et al.* Molecular Analysis of a Deletion Hotspot in the NRXN1 Region Reveals the Involvement of Short Inverted Repeats in Deletion CNVs. *Am J Hum Genet* **92**, 375–386 (2013).
21. Schaaf, C. P. *et al.* Phenotypic spectrum and genotype–phenotype correlations of NRXN1 exon deletions. *Eur J Hum Genet* **20**, 1240–1247 (2012).
22. Kim, H.-G. *et al.* Disruption of Neurexin 1 Associated with Autism Spectrum Disorder. *The American Journal of Human Genetics* **82**, 199–207 (2008).
23. Harrison, V. *et al.* Compound heterozygous deletion of NRXN1 causing severe developmental delay with early onset epilepsy in two sisters. *Am. J. Med. Genet. A* **155A**, 2826–2831 (2011).
24. Gicquel, C. *et al.* Epimutation of the telomeric imprinting center region on chromosome 11p15 in Silver-Russell syndrome. *Nat Genet* **37**, 1003–1007 (2005).
25. Price, S. M., Stanhope, R., Garrett, C., Preece, M. A. & Trembath, R. C. The spectrum of Silver-Russell syndrome: a clinical and molecular genetic study and new diagnostic criteria. *Journal of Medical Genetics* **36**, 837–842 (1999).
26. Azzi, S. *et al.* Multilocus methylation analysis in a large cohort of 11p15-related foetal growth disorders (Russell Silver and Beckwith Wiedemann syndromes) reveals simultaneous loss of methylation at paternal and maternal imprinted loci. *Hum Mol Genet* **18**, 4724–4733 (2009).

27. Nazeen, S., Palmer, N. P., Berger, B. & Kohane, I. S. Integrative analysis of genetic data sets reveals a shared innate immune component in autism spectrum disorder and its co-morbidities. *Genome Biology* **17**, 228 (2016).
28. Hamdan, F. F. *et al.* De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. *Am. J. Hum. Genet.* **87**, 671–678 (2010).
29. Horn, D. *et al.* Identification of FOXP1 deletions in three unrelated patients with mental retardation and significant speech and language deficits. *Human Mutation* **31**, E1851–E1860 (2010).
30. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
31. Grozeva, D. *et al.* De Novo Loss-of-Function Mutations in SETD5, Encoding a Methyltransferase in a 3p25 Microdeletion Syndrome Critical Region, Cause Intellectual Disability. *The American Journal of Human Genetics* **94**, 618–624 (2014).
32. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
33. Iossifov, I. *et al.* Low load for disruptive mutations in autism genes and their biased transmission. *PNAS* **112**, E5600–E5607 (2015).
34. Krupp, D. R. *et al.* Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *The American Journal of Human Genetics* **101**, 369–390 (2017).
35. Bowling, K. M. *et al.* Genomic diagnosis for children with intellectual disability and/or developmental delay. *Genome Med* **9**, 43 (2017).
36. Kehrer-Sawatzki, H., Mautner, V.-F. & Cooper, D. N. Emerging genotype-phenotype relationships in patients with large NF1 deletions. *Hum. Genet.* **136**, 349–376 (2017).

37. Vogt, J. *et al.* SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biology* **15**, R80 (2014).
38. Luciani, J. J. *et al.* Telomeric 22q13 deletions resulting from rings, simple deletions, and translocations: cytogenetic, molecular, and clinical analyses of 32 new observations. *Journal of Medical Genetics* **40**, 690–696 (2003).
39. Phelan, M. C. *et al.* 22q13 deletion syndrome. *American Journal of Medical Genetics* **101**, 91–99 (2001).
40. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *Journal of Medical Genetics* **56**, 131–138 (2019).