Genome-wide detection of tandem DNA repeats expanded in autism

Supplementary Notes and Figures

Brett Trost[1,2*], Worrawat Engchuan[1,2*], Charlotte M. Nguyen[1,2,3*], Bhooma Thiruvahindrapuram[1,2*], Egor Dolzhenko[4], Ian Backstrom[1], Mila Mirceta[1,3], Bahareh A. Mojarad[1], Yue Yin[1], Alona Dov[1,3], Induja Chandrakumar[1], Tanya Prasolava[1], Natalie Shum[1,3], Omar Hamdan[1,2], Giovanna Pellecchia[1,2], Jennifer L. Howe[1,2], Joseph Whitney[1,2], Eric W. Klee[5,6], Saurabh Baheti[5], David G. Amaral[7], Evdokia Anagnostou[8], Mayada Elsabbagh[9], Bridget A. Fernandez[10], Ny Hoang[1,3], M. E. Suzanne Lewis[11,12], Xudong Liu[13], Calvin Sjaarda[13], Isabel M. Smith[14], Peter Szatmari[15,16,17], Lonnie Zwaigenbaum[18], David Glazer[19], Dean Hartley[20], A. Keith Stewart[6,21], Michael A. Eberle[4], Nozomu Sato[1], Christopher E. Pearson[1,3], Stephen W. Scherer[1,2,3,22], Ryan K. C. Yuen[1,2,3]

[1]Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada
[2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada
[3]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[4]Illumina Inc, San Diego, CA, USA
[5]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[6]Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA
[7]MIND Institute and Department of Psychiatry and Behavioral Sciences, University of California Davis School of Medicine, Sacramento, CA, USA
[8]Holland Bloorview Kids Rehabilitation Hospital, University of Toronto, Toronto, ON, Canada
[9]Montreal Neurological Institute, McGill University, Montreal, QC, Canada
[10]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada
[11]Medical Genetics, University of British Columbia (UBC), Vancouver, BC, Canada
[12]BC Children's Hospital Research Institute, Vancouver, BC, Canada
[13]Department of Psychiatry, Queen's University, Kingston, ON, Canada
[14]Dalhousie University / IWK Health Centre, Halifax, NS, Canada
[15]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[16]Centre for Addiction and Mental Health, Toronto, ON, Canada
[17]Department of Psychiatry, The Hospital for Sick Children, Toronto, ON, Canada
[18]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada
[19]Verily Life Sciences, South San Francisco, CA, USA
[20]Autism Speaks, New York, NY, USA
[21]Division of Hematology, Mayo Clinic, Rochester, MN, USA
[22]McLaughlin Centre, University of Toronto, Toronto, ON, Canada

*These authors contributed equally
Corresponding author: Ryan K. C. Yuen ryan.yuen@sickkids.ca

Supplementary Notes

Sample curation

We removed 39 samples with Mendelian error or sex mismatch. Samples sequenced from PCR-based libraries and/or using the Illumina HiSeq 2000/2500 platforms had more apparent tandem repeats detected per sample (Extended Data Fig. 2), so we removed them from subsequent analyses. For consistency of DNA source, we removed 243 samples from cell line-derived DNA (i.e., only DNA extracted from whole blood was used). Histograms and QQ plots showed that a normal distribution of EHdn-detected tandem repeat counts was best approximated by removing samples having a count >3 standard deviations from the mean (N=249; Extended Data Fig. 3). For multiplex families, we retained only one affected child per family (the individual with the earliest sample ID).

Characteristics of large tandem repeats transmitted from the parents

We defined large tandem repeat transmission events as when a tandem repeat that was above the 99th percentile according to length in a parent was transmitted to the child, with the child's repeat also being above the 99th length percentile. Since these events are rare (<0.1% in the population), the parent should be heterozygous for the large tandem repeat in each family. We found that large tandem repeats in genic regions were more likely to be transmitted than other large tandem repeats in ASD-affected individuals (SSC: OR=1.20; p=3.9x10^{-5}, SSC and MSSNG combined: OR=1.18, p=1.6x10^{-7}, Extended Data Fig. 6 and Supplementary Table 11). Similar to the rare tandem repeat expansion burden comparison between ASD-affected individuals and their unaffected siblings, the large, transmitted tandem repeats were enriched in exons (MSSNG: OR=1.98; p=5x10^{-4}, SSC and MSSNG combined: OR=1.63; p=7x10^{-7}, Extended Data Fig. 6 and Supplementary Table 11) and splicing with borderline statistical significance (MSSNG: OR=1.32; p=0.07, combined set: OR=1.12; p=0.10, Extended Data Fig. 6 and Supplementary Table 11). In terms of gene sets, both the nervous system development (SSC: OR=1.31; p=9x10^{-4}; FWER=0.03, SSC and MSSNG combined: OR=1.34; p=5.7x10^{-7}; FWER=1.8x10^{-5}, Extended Data Fig. 7 and Supplementary Table 12) and cardiovascular system or muscle (MSSNG: OR=1.54; p=1.4x10^{-4}; FWER=5x10^{-3}, combined set: OR=1.35; p=6x10^{-6}; FWER=2x10^{-4}; Extended Data Fig. 7 and Supplementary Table 12) were recapitulated in the transmission tests.

As noted in the main text, transmitted large tandem repeats were also enriched in SSC unaffected siblings, but were more likely to be further expanded in ASD-affected individuals. This effect was not statistically significant in genic expansions overall (OR=1.12; p=0.47), but they tend to be expanded more frequently in individuals with ASD than their unaffected siblings within exons (OR=2.38; p=0.08), splice sites (OR=5.49; p=0.02), and the nervous system development and cardiovascular system or muscle gene sets (OR=1.62; p=0.10).

Analysis of X-linked tandem repeat loci

When X-linked tandem repeat loci were considered (separately in males and females) in individuals of European ancestry from SSC, we found that affected individuals were enriched in genic tandem repeat expansions compared to their unaffected siblings, although this was not statistically significant (males: OR=1.37; p=0.63; females: OR=1.59; p=0.46). Power

calculations showed that N=96,084 males and N=16,387 females would be needed to observe a statistically significant effect. In males, the sole enriched gene set was nervous system development (p=0.03); no gene sets were significantly enriched in females (Supplementary Table 13).

Genotyping of disease loci

Given that EHdn cannot detect tandem repeats <150 bp, we used ExpansionHunter to genotype known tandem repeat disease loci, some of which have disease-causing size thresholds <150 bp. Of 49 loci for which a disease-causing size threshold is known, 18 (36.7%) of them have a disease-causing size threshold <150 bp. There were 22 loci with at least one ASD-affected individual or unaffected sibling (European individuals in SSC only) whose largest allele exceeded that threshold, and 15 (68.2%) of them have a disease-causing size threshold <150 bp. No loci were statistically significant in terms of more ASD-affected individuals than unaffected siblings having an allele exceeding the disease-causing threshold (Fisher's exact test) (Supplementary Table 14). The most promising locus with high odds ratio (OR=3.28) was the tandem repeat at *DMPK*, which was captured by analysis with EHdn.

Additional genetic findings and phenotypes

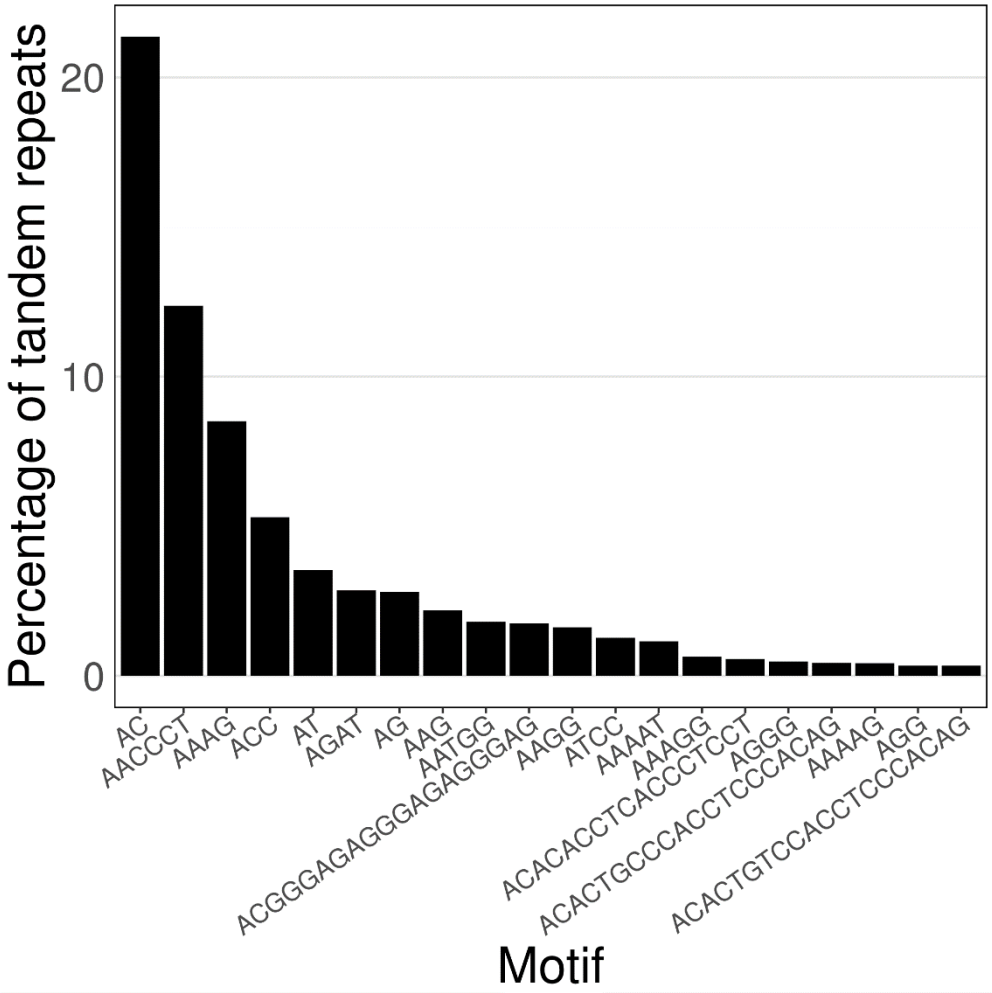
We estimated that large sample sizes would be required to provide sufficient statistical power to replicate individual loci identified in Table 1 and Supplementary Table 5; for instance, to achieve 80% power at α=0.05, N=7,206 and N=23,575 ASD-affected individuals (plus an equivalent number of unaffected individuals) would be required for *CACNB1* and *FGF14*, respectively.

Towards correlating the genetic findings herein with the phenotypes in the MSSNG cohort, we note that all 4 males with clinical information available in the database with CGG repeat expansions in *FMR1* were indicated as having fragile X syndrome (Supplementary Fig. 6). Similarly, one of the probands (family 1-1039) with a rare tandem repeat expansion of $(CTG)_{\sim 950}$ detected in *DMPK* was reported as having DM1 and other developmental problems (Supplementary Fig. 6). Her mother, who carries a repeat of $(CTG)_{\sim 180}$ in *DMPK* (Fig. 2g), also reported a history of difficulties in motor coordination (i.e., genetic anticipation). A pedigree with individuals having expanded tandem repeats at *FXN* is presented in Supplementary Fig. 6. In three other novel loci (*FGF14*, *CACNB1*, and *CDON*), detailed clinical information was available for a total of 12 affected individuals with rare tandem repeat expansions. Of 9 individuals for whom information on motor function was available, 6 (67%) have motor delay or motor issues, and 9 out of the 12 (75%) have psychiatric or behavioral problems (Supplementary Table 15).

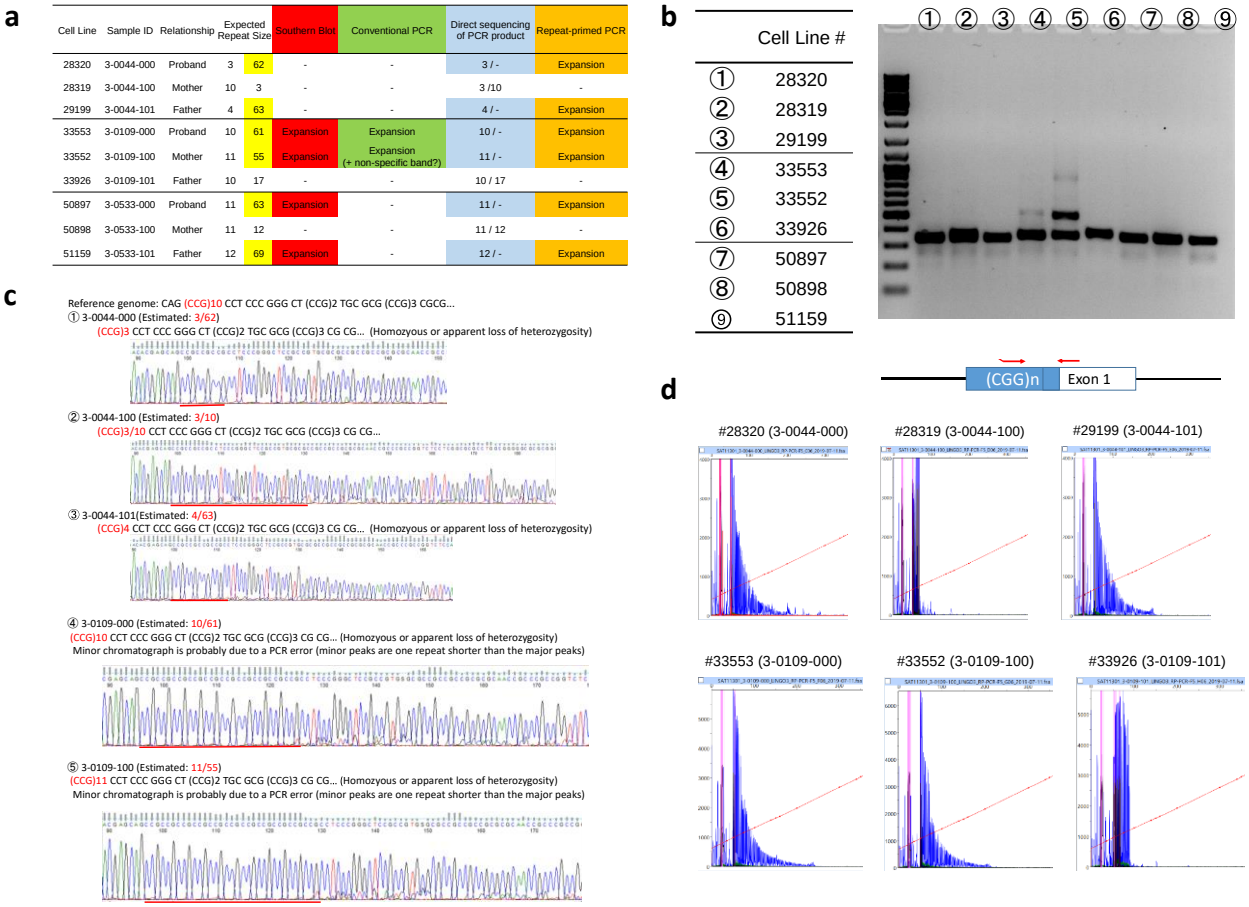
In essence, selected individuals in Table 1 should be clinically assessed for the respective OMIM conditions (e.g., myotonic dystrophy and ataxia) coupled to the genes with rare tandem repeat expansions. Looking beyond, all other individuals molecularly identified with the same variants should be assessed for ASD, including those parents carrying tandem repeat length at the 99^{th} percentile of the length distribution (Fig. 2a). The larger list of candidate disease loci in Supplementary Table 5 also suggests further genotype and phenotype studies necessary for proper medical management and counselling in ASD.

Although not statistically significant, there was a trend of more rare repeat expansions detected in children with older fathers (Supplementary Fig. 7).

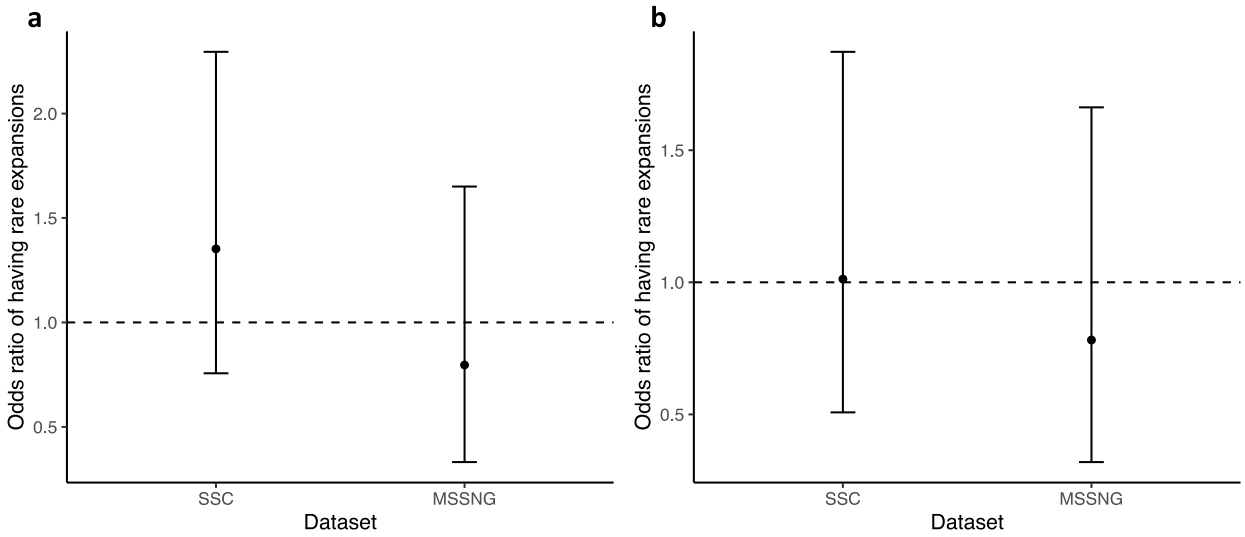
Supplementary Figures



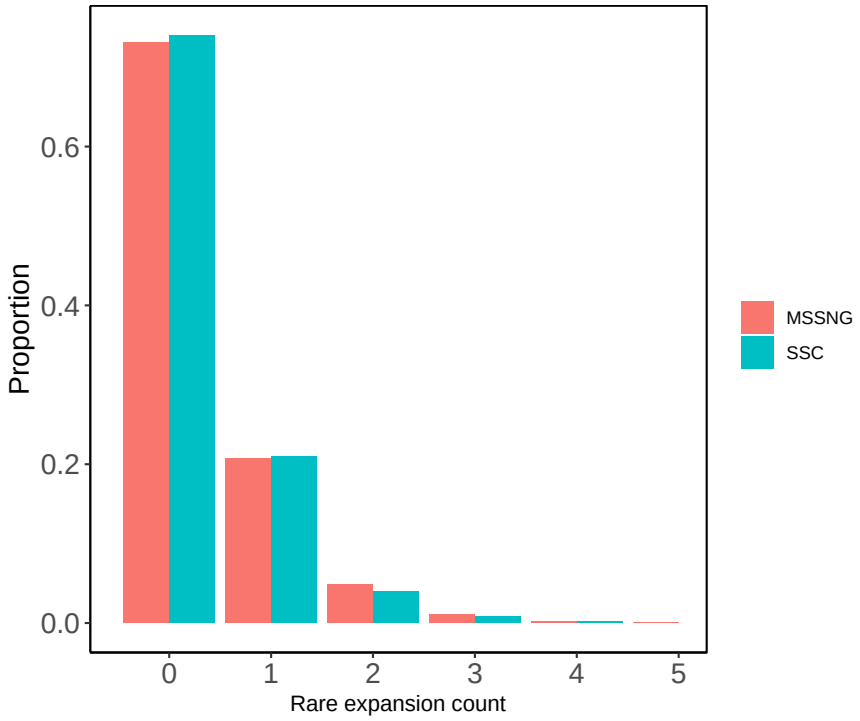
Supplementary Figure 1 | Distribution of repeat units (motifs) for the tandem repeats detected by ExpansionHunter Denovo. The 20 most common repeat units are shown.



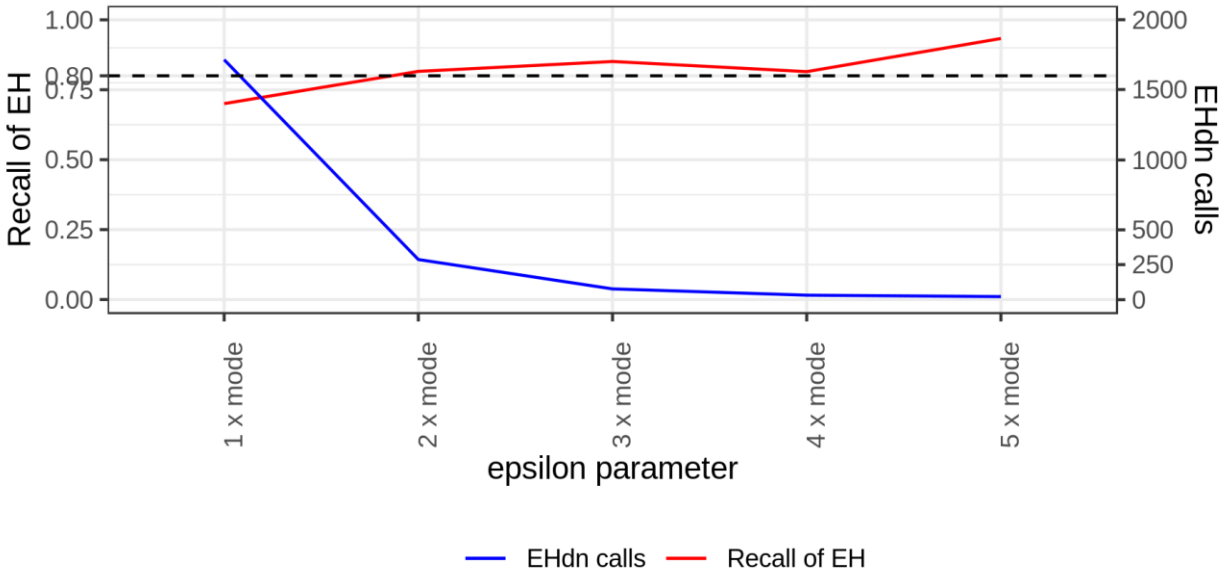
Supplementary Figure 2 | Methods for sizing the CGG repeat in *LINGO3*. **a**, Summary of results of repeat size analyses. PCR-free short-read sequence data with ExpansionHunter not only correctly determined the length of short CGG repeats, but also detected long CGG repeats. We detected that the presence of a small deletion adjacent to the repeat in two individuals hindered detection of the CGG repeat expansion by Southern blot. **b**, Results of PCR amplification of the CGG repeat in *LINGO3*. Due to the extremely high CG content of the region, long CGG repeats could not be amplified. **c**, Sanger sequencing of the PCR-amplified CGG repeat in *LINGO3*. Note that the bias of PCR towards preferential amplification of shorter amplicons made the chromatogram of longer alleles less prominent. **d**, Repeat-primed PCR design and results for the CGG repeat in *LINGO3*. The predictions for tandem repeat expansions made by ExpansionHunter were consistent with the repeat-primed PCR. Southern blot analysis of the large *LINGO3* expansions are shown in Fig. 2f. Repeat sizing of PCR-amplifiable samples by Sanger sequencing was performed once. Repeat-primed PCR experiments were consistently reproduced at least twice.



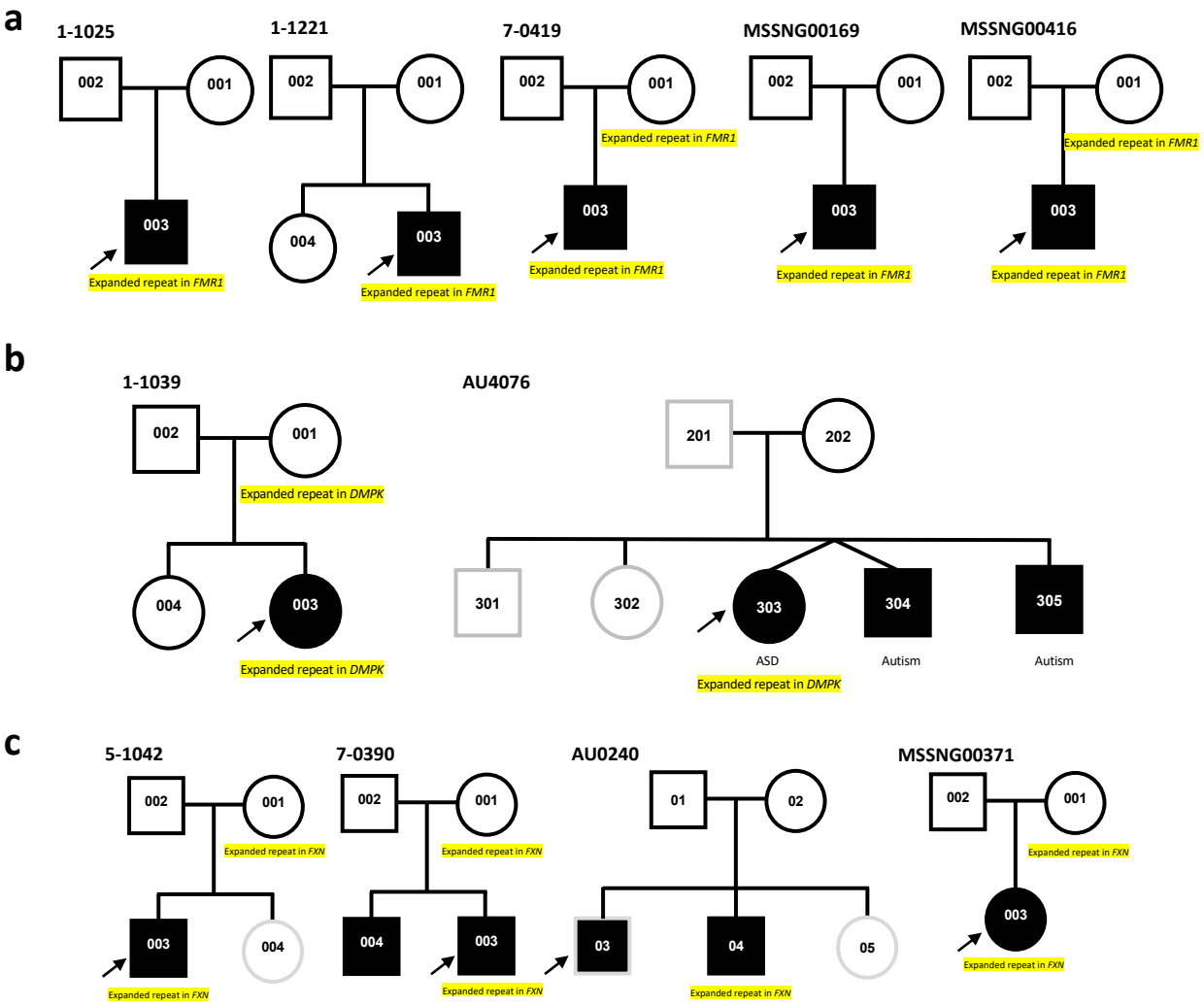
Supplementary Figure 3 | Odds ratios that an affected individual with a *de novo* genic deletion (a) or loss of function (LoF) variant (b) also had a rare tandem repeat expansion. There was no statistically significant difference observed in any test (SSC deletions: OR=1.35; p=0.24, MSSNG deletions: OR=0.8; p=0.61, SSC LoF: OR=1.0; p=1.0, and MSSNG LoF: OR=0.78; p=0.6) (Fisher's exact test). The number of individuals included for SSC and MSSNG were 1,845 and 1,566, respectively. Error bars indicate 95% confidence intervals.



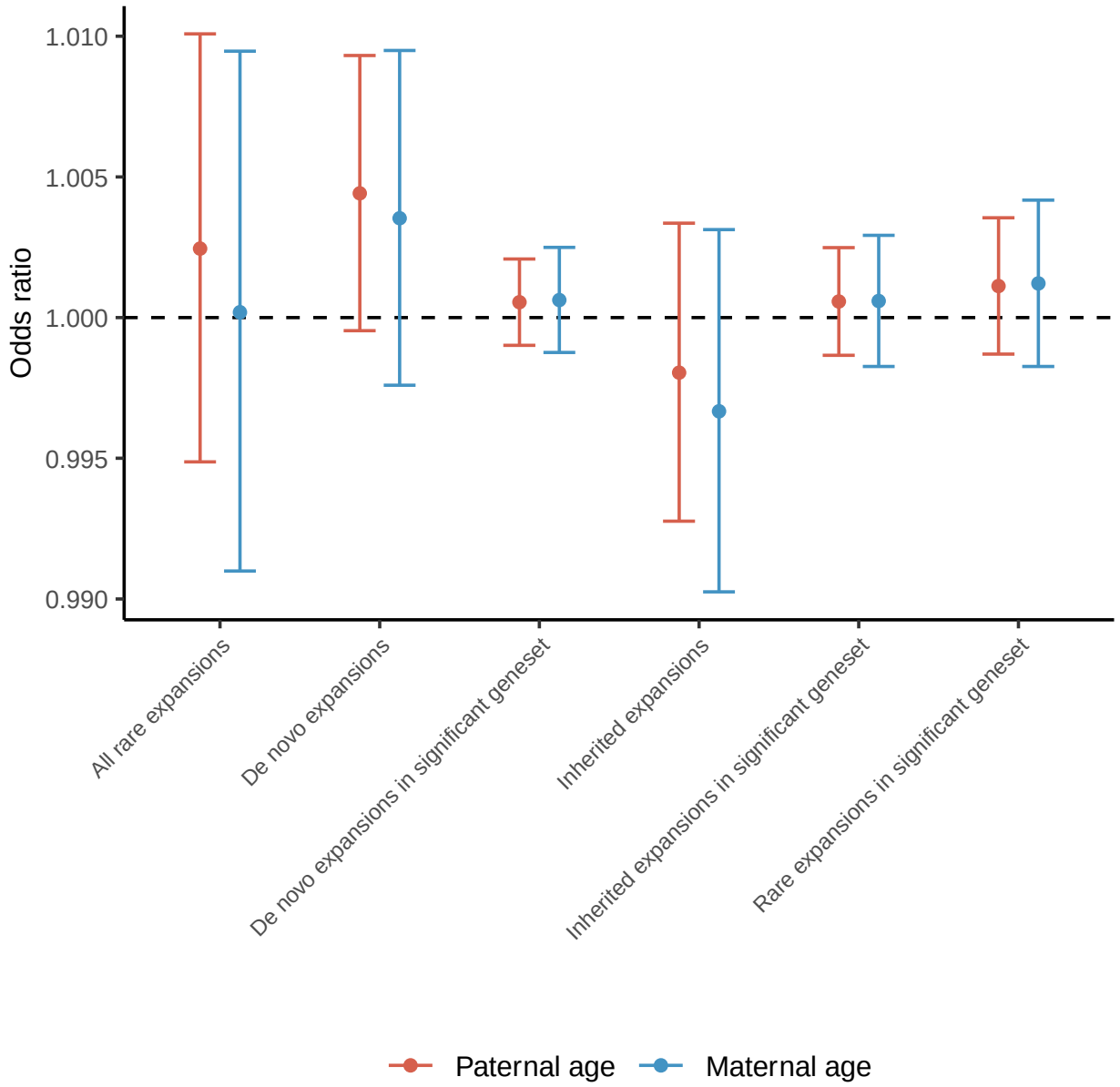
Supplementary Figure 4 | Number of rare tandem repeat expansions per subject in ASD-affected individuals. No difference was found in the detection rate of rare tandem repeat expansions between MSSNG (N=2,962) and SSC (N=4,279) (p=0.45, with population admixture as a covariate). An ANOVA test comparing two logistic regression models was used.



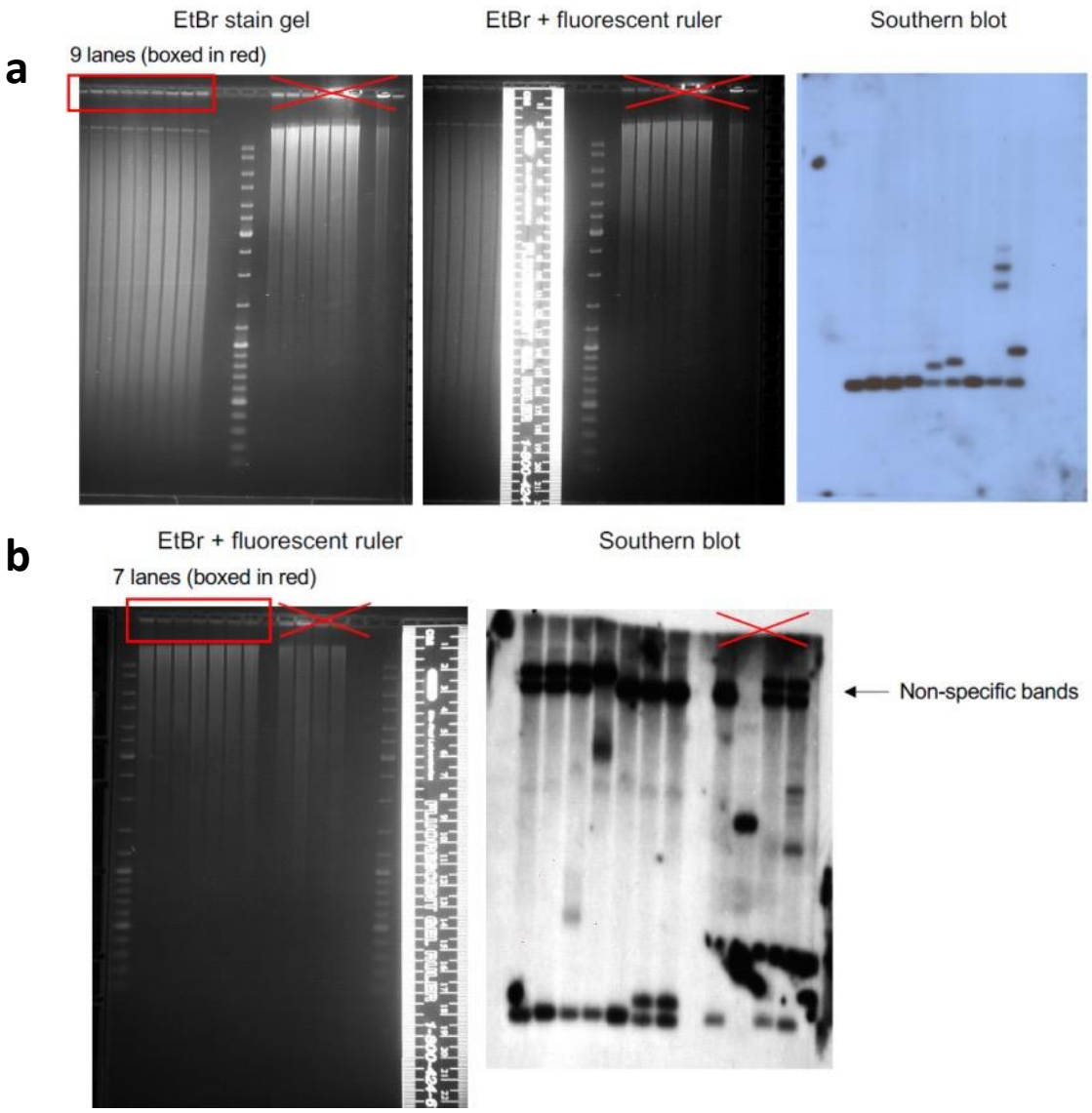
Supplementary Figure 5 | Effect of different DBSCAN epsilon values on recall of outliers detected in ExpansionHunter data (left y-axis) and number of tandem repeat loci detected by EHdn (right y-axis). Dashed line indicates 80% recall. Only tandem repeats with corresponding Tandem Repeats Finder coordinates and motif size >=3 were compared.



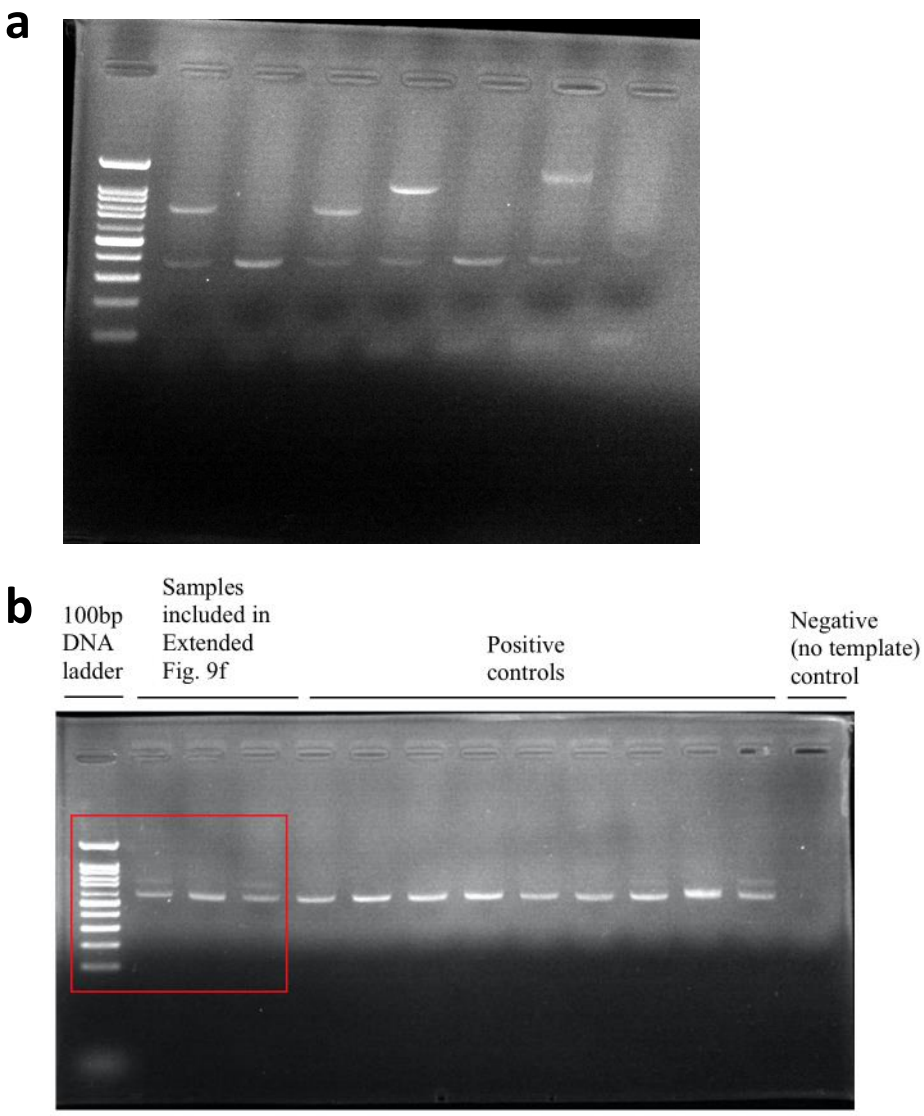
Supplementary Figure 6 | Pedigrees of families with rare tandem repeat expansions in top genes. a, Families with rare tandem repeat expansions in *FMR1*. There is clinical information available for 1-1025-003, 1-1221-003, MSSNG00169-003 and MSSNG00416-003. All of them are diagnosed with Fragile X syndrome. **b**, Families with rare tandem repeat expansions in *DMPK*. In the family with clinical information recorded (1-1039), 1-1039-003 was reported to have ASD, myotonic dystrophy, delayed development, and nocturnal hypoventilation. We experimentally validated no presence of repeat expansion in *DMPK* in 1-1039-004, AU4076304, and AU4076305. For the individual in grey, the corresponding sample was not available for testing. Individual 1-1039-003 was reported to have DM1; we could not confirm DM1 status in other individuals with *DMPK* expansions due to incomplete phenotype information or because testing was done at an early age. **c**, Families with rare tandem repeat expansions in *FXN*. Among families with clinical information recorded (7-0390 and AU0240), 7-0390-003 was reported to have ASD, anxiety, asthma, and fine motor delays. AU024004 was reported to have ASD and athetosis. For the individual in grey, the corresponding sample was not available for testing. Motifs of all expanded repeats detected are AAGGAG.



Supplementary Figure 7 | Correlation analysis between parental age and number of tandem repeat expansions detected. An expansion in a child with ASD was defined as *de novo* if the maximum repeat size of the corresponding parents was below the 75th percentile. Only samples with parental age information were included here (N=826). Error bars represent 95% confidence intervals. The correlation between age and number of tandem repeat expansions was estimated by linear regression.



Supplementary Figure 8 | Raw source images of Southern blots. Original gel images for **a**, Fig. 1f and **b**, Fig. 2g. Chopped images are indicated by red lines on the Southern blots.



Supplementary Figure 9 | Raw source images of gel electrophoresis. Original gel images for **a**, Extended Fig. 9b and **b**, Extended Fig. 9f. Chopped images are indicated by red lines.