11. I. Lucchitta, G. H. Curtis, M. E. Davis, S. W. Davis, B. Turrin, *Quat. Res.* **53**, 23 (2000).
12. V. Polyak, C. Hill, Y. Asmerom, *Science* **319**, 1377 (2008).
13. J. Pederson, R. Young, I. Lucchitta, L. S. Beard, G. Billingsley, *Science* **321**, 1634b (2008).
14. R. A. Young, *Tectonophysics* **61**, 25 (1979).
15. D. P. Elston, R. A. Young, *J. Geophys. Res.* **96**, 12389 (1991).
16. A. R. Potochnik, in *Colorado River Origin and Evolution*, R. A. Young, E. E. Spamer, Eds. (Grand Canyon Association, Grand Canyon, AZ, 2001), pp. 17–22.
17. R. A. Young, in *Colorado River Origin and Evolution*, R. A. Young, E. E. Spamer, Eds. (Grand Canyon Association, Grand Canyon, AZ, 2001), pp. 7–16.
18. R. A. Young, in *Geology of Grand Canyon, Northern Arizona*, D. P. Elston, G. H. Billingsley, R. A. Young, Eds. (28th International Geological Congress Fieldtrip Guidebook T115/315, American Geophysical Union, Washington, DC, 1989), pp. 166–173.
19. R. A. Young, in *Late Cenozoic Drainage History of the Southwestern Great Basin and Lower Colorado River Basin: Geologic and Biologic Perspectives*, M. C. Reheis, R. Hershler, D. M. Miller, Eds. (Geological Society of America Special Paper, Geological Society of America, Boulder, CO, 2008), vol. 439, pp. 319–333.
20. P. A. Pearthree, J. E. Spencer, J. E. Faulds, P. K. House, *Science* **321**, 1634 (2008).
21. B. Wernicke, *Geol. Soc. Am. Bull.* **123**, 1288 (2011).
22. M. A. House, B. P. Wernicke, K. A. Farley, *Nature* **396**, 66 (1998).
23. R. M. Flowers, B. P. Wernicke, K. A. Farley, *Geol. Soc. Am. Bull.* **120**, 571 (2008).
24. D. L. Shuster, K. A. Farley, *Earth Planet. Sci. Lett.* **217**, 1 (2004).
25. K. A. Farley, D. L. Shuster, E. B. Watson, K. H. Wanser, G. Balco, *Geochem. Geophys. Geosyst.* **11**, Q10001 (2010).
26. K. A. Farley, D. L. Shuster, R. A. Ketcham, *Geochim. Cosmochim. Acta* **75**, 4515 (2011).
27. D. L. Shuster, R. M. Flowers, K. A. Farley, *Earth Planet. Sci. Lett.* **249**, 148 (2006).
28. R. M. Flowers, R. A. Ketcham, D. L. Shuster, K. A. Farley, *Geochim. Cosmochim. Acta* **73**, 2347 (2009).
29. R. M. Flowers, D. L. Shuster, B. P. Wernicke, K. A. Farley, *Geology* **35**, 447 (2007).
30. Materials and methods are available as supplementary materials on *Science* Online.
31. T. A. Dumitru, I. R. Duddy, P. F. Green, *Geology* **22**, 499 (1994).
32. S. A. Kelley, C. E. Chapin, K. Karlstrom, in *Colorado River Origin and Evolution*, R. A. Young, E. E. Spamer, Eds. (Grand Canyon Association, Grand Canyon, AZ, 2001), pp. 37–42.

Acknowledgments: This work was supported by NSF grant EAR-1019896 to K.A.F. The data reported in this paper are tabulated in the supplementary materials. We thank B. Wernicke for discussion.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1229390/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S6
References (*33*–*40*)

27 August 2012; accepted 7 November 2012
Published online 29 November 2012;
10.1126/science.1229390

Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders

Brian J. O'Roak,[1] Laura Vives,[1] Wenqing Fu,[1] Jarrett D. Egertson,[1] Ian B. Stanaway,[1] Ian G. Phelps,[2,3] Gemma Carvill,[2,3] Akash Kumar,[1] Choli Lee,[1] Katy Ankenman,[4] Jeff Munson,[4] Joseph B. Hiatt,[1] Emily H. Turner,[1] Roie Levy,[1] Diana R. O'Day,[2] Niklas Krumm,[1] Bradley P. Coe,[1] Beth K. Martin,[1] Elhanan Borenstein,[1,5,6] Deborah A. Nickerson,[1] Heather C. Mefford,[2,3] Dan Doherty,[2,3] Joshua M. Akey,[1] Raphael Bernier,[4] Evan E. Eichler,[1,7]* Jay Shendure[1]*

Exome sequencing studies of autism spectrum disorders (ASDs) have identified many de novo mutations but few recurrently disrupted genes. We therefore developed a modified molecular inversion probe method enabling ultra-low-cost candidate gene resequencing in very large cohorts. To demonstrate the power of this approach, we captured and sequenced 44 candidate genes in 2446 ASD probands. We discovered 27 de novo events in 16 genes, 59% of which are predicted to truncate proteins or disrupt splicing. We estimate that recurrent disruptive mutations in six genes—*CHD8*, *DYRK1A*, *GRIN2B*, *TBR1*, *PTEN*, and *TBL1XR1*—may contribute to 1% of sporadic ASDs. Our data support associations between specific genes and reciprocal subphenotypes (*CHD8*-macrocephaly and *DYRK1A*-microcephaly) and replicate the importance of a β-catenin–chromatin-remodeling network to ASD etiology.

There is considerable interest in the contribution of rare variants and de novo mutations to the genetic basis of complex phenotypes such as autism spectrum disorders (ASDs). However, because of extreme genetic heterogeneity, the sample sizes required to implicate any single gene in a complex phenotype are extremely large (*1*). Exome sequencing has identified hundreds of ASD candidate genes on the basis of de novo mutations observed in the affected offspring of unaffected parents (*2*–*6*). Yet, only a single mutation was observed in nearly all such genes, and sequencing of over 900 trios was insufficient to establish mutations at any single gene as definitive genetic risk factors (*2*–*6*).

To address this, we sought to evaluate candidate genes identified by exome sequencing (*2*, *3*) for de novo mutations in a much larger ASD cohort. We developed a modified molecular inversion probe (MIP) strategy (Fig. 1A) (*7*–*9*) with novel algorithms for MIP design; an optimized, automatable work flow with robust performance and minimal DNA input; extensive multiplexing of samples while sequencing; and reagent costs of less than $1 per gene per sample. Extensive validation using several probe sets and sample collections demonstrated 99% sensitivity and 98% positive predictive value for single-nucleotide variants at well-covered positions, i.e., 92 to 98% of targeted bases (figs. S1 to S7 and tables S1 to S9) (*10*).

We applied this method to 2494 ASD probands from the Simons Simplex Collection (SSC) (*11*) using two probe sets [ASD1 (6 genes) and ASD2 (38 genes)] to target 44 ASD candidate genes (*12*). Preliminary results using ASD1 on a subset of the SSC implicated *GRIN2B* as a risk locus (*3*). The 44 genes were selected from 192 candidates (*2*, *3*) by focusing on genes with disruptive mutations, associations with syndromic autism (*13*), overlap with known or suspected neurodevelopmental copy number variation (CNV) risk loci (*13*, *14*), structural similarities, and/or neuronal expression (table S3). Although a few of the 44 genes have been reported to be disrupted in individuals with neurodevelopmental or neuropsychiatric disorders (often including concurrent dysmorphologies), their role in so-called idiopathic ASDs has not been rigorously established. Twenty-three of the 44 genes intersect a 49-member β-catenin–chromatin-remodeling protein-protein interaction (PPI) network (*2*) or an expanded 74-member network (figs. S8 and S9) (*3*, *4*).

We required samples to successfully capture with both probe sets, yielding 2446 ASD probands with MIP data, 2364 of which had only MIP data and for 82 of which we had also sequenced their exomes (*2*, *3*). The high GC content of several candidates required considerable rebalancing to improve capture uniformity (*12*) (figs. S3A and S10). Nevertheless, the reproducible behavior of most MIPs allowed us to identify copy number variation at targeted genes, including several inherited duplications (figs. S11 and S12 and table S10).

To discover de novo mutations, we first identified candidate sites by filtering against variants observed in other cohorts, including non-ASD exomes and MIP-based resequencing of 762 healthy, non-ASD individuals (*12*). The remaining

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. [2]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA 98195, USA. [3]Seattle Children's Hospital, Seattle, WA 98105, USA. [4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA 98195, USA. [5]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. [6]Santa Fe Institute, Santa Fe, NM 87501, USA. [7]Howard Hughes Medical Institute, Seattle, WA 98195, USA.

*To whom correspondence should be addressed. E-mail: shendure@uw.edu (J.S.); eee@gs.washington.edu (E.E.E.)

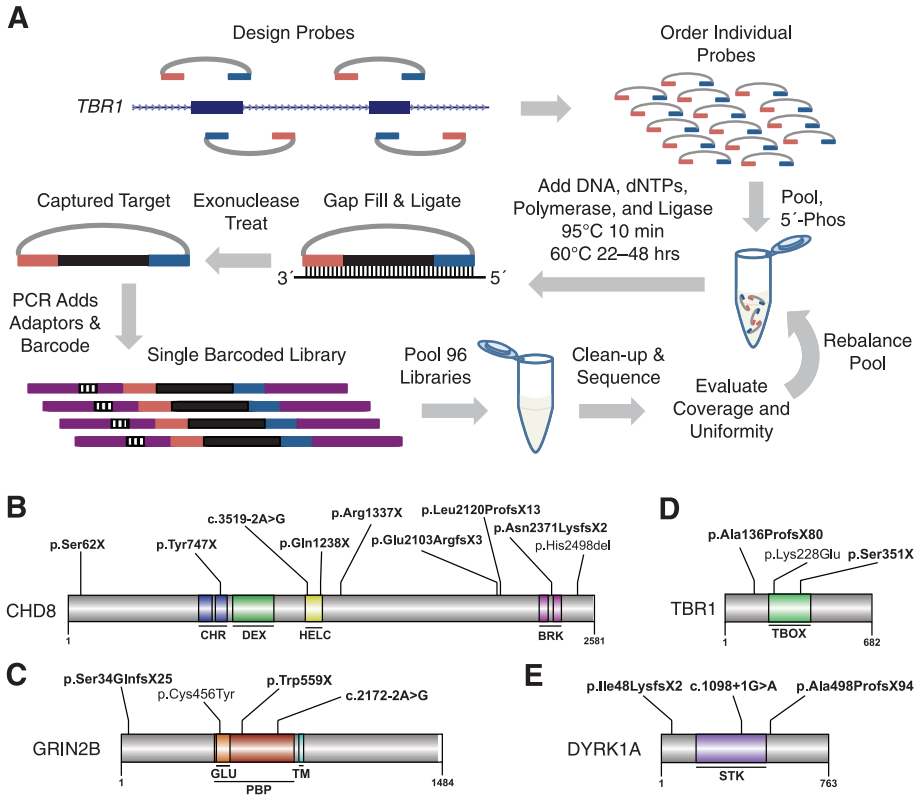

Fig. 1. Massively multiplex targeted sequencing identifies recurrently mutated genes in ASD probands. (**A**) Schematic showing design and general work flow of a modified MIP method enabling ultra-low-cost candidate gene resequencing in very large cohorts (figs. S1 to S7 and tables S1 to S9) (*10*). (**B** to **E**) Protein diagrams of four genes with multiple de novo mutation events. Significant protein domains for the largest protein isoform are shown (colored regions) as defined by SMART (*23*) with mutation locations indicated. (B) CHD8. (C) GRIN2B. (D) TBR1. (E) DYRK1A. Bold variants are nonsense, frameshifting indels or at splice sites (intron-exon junction is indicated). Domain abbreviations: CHR, chromatin organization modifier; DEX, DEAD-like helicases superfamily; HELC, helicase superfamily C-terminal; BRK, domain in transcription and CHROMO domain helicases; GLU, ligated ion channel L-glutamate– and glycine-binding site; PBP, eukaryotic homologs of bacterial periplasmic substrate binding proteins; TM, transmembrane; STK, serine-threonine kinase catalytic; TBOX, T-box DNA binding.

candidates were further tested by MIP-based resequencing of the proband's parents and, if potentially de novo, confirmed by Sanger sequencing of the parent-child trio (*10*, *12*). We discovered 27 de novo mutations that occurred in 16 of the 44 genes (Fig. 1, B to E; Table 1; and table S11). Consistent with an increased sensitivity for MIP-based resequencing, six of these were not reported in exome-sequenced individuals (Table 1, tables S5 and S11, and fig. S13) (*3*, *4*, *6*). Notably, the proportion of de novo events that are severely disruptive, i.e., coding indels, nonsense mutations, and splice-site disruptions (17/27 or 0.63), is four times the expected proportion for random de novo mutations (0.16, binomial $P = 4.9 \times 10^{-8}$) (table S12) (*15*).

Given their extremely low frequency, accurately establishing expectation for de novo mutations in a locus-specific manner through the sequencing of control trios is impractical. We therefore developed a probabilistic model that incorporates several factors: the overall rate of mutation in coding sequences, estimates of relative locus-specific rates based on human-chimpanzee fixed differences (fig. S14 and table S13), and other factors that may influence the distribution of mutation classes, e.g., codon structure (*12*). We applied this model to estimate (by simulation) the probability of observing additional de novo mutations during MIP-based resequencing of the SSC cohort. To compare expectation and observation, we treated missense mutations as one class and severe disruptions as a second class. Thus, we could evaluate the probability at a given

Table 1. Six genes with recurrent de novo mutations. Assay is the primary assay that identified the variant. Abbreviations: M, male; F, female; Mut, mutation type; Fs, frameshifting indel; Ns, nonsense; Sp, splice site; Aa, single–amino acid deletion; Ms, missense; EX, exome; HGVS, Human Genome Variation Society nomenclature; NVIQ, nonverbal intellectual quotient.

Proband	Sex	Gene	Mut	Assay	HGVS	NVIQ
12714.p1	M	*CHD8**	Ns	MIP	p.Ser62X	78
13986.p1	M	*CHD8**	Fs	MIP	p.Tyr747X	38
11654.p1	F	*CHD8**	Sp	MIP‡ (*4*)	c.3519-2A>G	41
13844.p1	M	*CHD8**	Ns	EX	p.Gln1238X	34
14016.p1	M	*CHD8**	Ns	MIP	p.Arg1337X	92
12991.p1	M	*CHD8**	Fs	MIP	p.Glu2103ArgfsX3	67
12752.p1	F	*CHD8**	Fs	EX	p.Leu2120ProfsX13	93
14233.p1	M	*CHD8**	Fs	MIP	p.Asn2371LysfsX2	19
14406.p1	M	*CHD8**	Aa	MIP	p.His2498del	98
12099.p1	M	*DYRK1A**	Fs	MIP‡ (*4*)	p.Ile48LysfsX2	55
13890.p1	F	*DYRK1A**	Sp	EX	c.1098+1G>A	42
13552.p1	M	*DYRK1A**	Fs	MIP§ (*6*)	p.Ala498ProfsX94	66
11691.p1	M	*GRIN2B*†	Fs	MIP‡‖ (*3*)	p.Ser34GlnfsX25	62
13932.p1	M	*GRIN2B*†	Ms	MIP	p.Cys456Tyr	55
12547.p1	M	*GRIN2B*†	Ns	MIP‖	p.Trp559X	65
12681.p1	F	*GRIN2B*†	Sp	EX	c.2172-2A>G	65
14433.p1	M	*PTEN*	Ms	MIP	p.Thr131Ile	50
14611.p1	M	*PTEN*	Fs	MIP	p.Cys136MetfsX44	33
11390.p1	F	*PTEN*	Ms	EX	p.Thr167Asn	77
12335.p1	F	*TBL1XR1**	Ms	EX	p.Leu282Pro	47
14612.p1	M	*TBL1XR1**	Fs	MIP	p.Ile397SerfsX19	41
11480.p1	M	*TBR1*†	Fs	EX	p.Ala136ProfsX80	41
13814.p1	M	*TBR1*†	Ms	MIP	p.Lys228Glu	78
13796.p1	F	*TBR1*†	Fs	MIP‡ (*4*)	p.Ser351X	63

*Part of 49-member connected component reported in (*3*). †Part of expanded 74-member connected component. ‡,§Proband was exome sequenced by cited study and variant was ‡not reported or §reported. ‖Variant reported in MIP screen from (*3*).

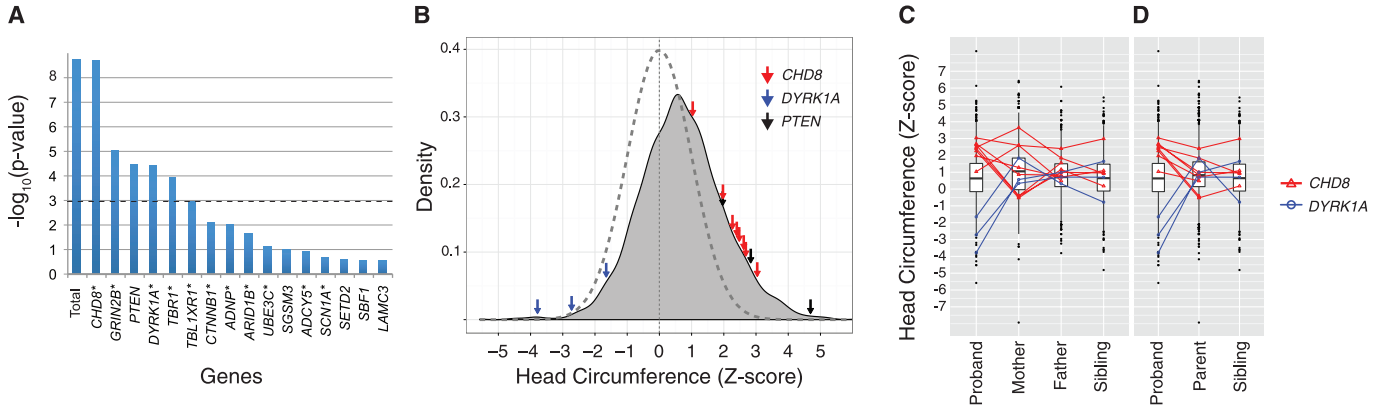

Fig. 2. Locus-specific mutation probabilities and associated phenotypes. (**A**) Estimated *P* values for the observed number of additional de novo mutations identified in the MIP screen of 44 ASD candidate genes. Probabilities shown are for observing *x* or more events, of which at least *y* belong to the severe class. The observed numbers of mutations in all 44 genes ("Total") and *CHD8* were not seen in any of 5×10^8 simulations. Based on the simulation mean (0.0153), the Poisson probability for seven or more severe class *CHD8* mutations is 3.8×10^{-17}. Dashed line Bonferroni corrected significance threshold for $\alpha = 0.05$. *Gene product in the 74-member PPI connected component. (**B** to **D**) Standardized head circumference (HC) *Z* scores for SSC. (B) All probands screened with superimposed normal distribution (dashed). HC *Z* scores for individuals with de novo truncating and/or splice mutations highlighted for *CHD8* (red arrows), *DYRK1A* (blue arrows), and *PTEN* (black arrows). (C and D) Box and whisker plots of the HC *Z* scores for the SSC. Mutations carriers are shown and linked to their respective family members. (C) All family members. (D) Only proband sex–matched family members.

locus of observing at least *x* de novo mutations, of which at least *y* belong to the severe class.

We found evidence of mutation burden—a higher rate of de novo mutation than expected—in the overall set of 44 genes (observed $n = 27$ versus mean expected $n = 5.6$, simulated $P < 2 \times 10^{-9}$) (Fig. 2A). The burden was driven by the severe class (observed $n = 17$ versus mean expected $n = 0.58$, simulated $P < 2 \times 10^{-9}$). Most severe class mutations intersected the 74-member PPI network (16 out of 17), although only 23 out of 44 genes are in this network (binomial $P = 0.0002$) (*12*). Furthermore, 21 out of 27 mutations occurred in network-associated genes (binomial $P = 0.004$). Of the six individual genes (*CHD8*, *GRIN2B*, *DYRK1A*, *PTEN*, *TBR1*, and *TBL1XR1*) with evidence of mutation burden [alpha of 0.05 after a Holm-Bonferroni correction for multiple testing (Fig. 2A); *TBL1XR1* is borderline significant with a more conservative Bonferroni correction], five fall within the β-catenin–chromatin-remodeling network. In our combined MIP and exome data, ~1% (24 out of 2573) of ASD probands harbor a de novo mutation in one of these six genes, with *CHD8* representing 0.35% (9 out of 2573) (Fig. 1B and Table 1).

For these analyses, we conservatively used the highest available empirical estimate of the overall mutation rate in coding sequences (*3*). With the exception of *TBL1XR*, these results were robust to doubling the overall mutation rate or to using the upper bound of the 95% confidence interval of the locus-specific rate estimate for each of these genes (*10*). Moreover, we obtained similar results regardless of whether parameters were estimated from rare, segregating variation or from de novo mutations in unaffected siblings (*10*), as well as with a sequence composition model based on genome-wide de novo mutation (*16*). Exome sequencing of non-ASD individuals (unaffected siblings or non-ASD cohorts) further supports these conclusions (table S14) (*10*).

We also validated 23 inherited, severely disruptive variants in the 44 genes (table S15). Two probands with such variants carry de novo 16p11.2 duplications (table S16). Combining de novo and inherited events, severe class variants were observed at twice the rate in MIP-sequenced probands as compared with MIP-sequenced healthy, non-ASD individuals (Fisher's exact test, $P = 0.083$). Severe class variants were not transmitted to 14 out of 20 unaffected siblings (binomial $P = 0.058$) (table S15). However, larger cohorts than currently exist will be needed to fully evaluate these modest trends.

We analyzed phenotypic data on probands with mutations in the six implicated genes. Each was diagnosed with autism on the basis of current, strict, gold-standard criteria. No obvious dysmorphologies or recurrent comorbidities were present. Probands tended to fall into the intellectual disability range for nonverbal IQ (NVIQ) (mean 58.3) (Table 1). However, for *CHD8*, probands were found to have NVIQ scores ranging from profoundly impaired to average (mean 62.2, range 19 to 98).

Given the previously observed microcephaly in our index *DYRK1A* mutation case, macrocephaly in both probands with *CHD8* mutations (*3*), and the association of these traits with other syndromic loci (*13*, *17*), we reexamined head circumference (HC) in the larger set of probands with protein-truncation or splice-site de novo events using age- and sex-normalized HC *Z* scores (*12*) (Fig. 2B). For *CHD8* ($n = 8$), we observed significantly larger head sizes relative to individuals screened without *CHD8* mutations (two-sample permutation test, two-sided $P = 0.0007$). De novo *CHD8* mutations are present in ~2% of macrocephalic (HC > 2.0) SSC probands ($n = 366$), which suggests a useful phenotype for patient subclassification. For *DYRK1A* ($n = 3$), we observed significantly smaller head sizes relative to individuals screened without *DYRK1A* mutations (two-sample permutation test, two-sided $P = 0.0005$). Comparison of head size in the context of the families (Fig. 2, C and D, and table S17) provides further support for this reciprocal trend (*10*). These findings are also consistent with case reports of patients with structural rearrangements and mouse transgenic models that implicate *DYRK1A* and *CHD8* as regulators of brain growth (*18*–*21*). Macrocephaly was also observed in individuals with de novo and inherited *PTEN* mutations (*22*).

Our data support an important role for de novo mutations in six genes in ~1% of sporadic ASDs. As the SSC was specifically established for simplex families and as its probands generally have higher cognitive functioning than has been reported in other ASD cohorts (*11*), it is unknown how our findings will translate into other cohorts. Furthermore, whereas our data implicate specific loci in ASDs, they are insufficient to evaluate whether the observed de novo mutations are sufficient to cause ASDs (tables S16 and S18).

Exome sequencing and CNV studies suggest that there are hundreds of relevant genetic loci for ASDs. Technologies and study designs directed at identifying de novo mutations, both for the discovery of ASD candidate genes, as well as for their validation, provide sufficient power to implicate individual genes from a relatively small number of events. The analytical framework described here can be applied to any other disorder—simple or complex—for which de novo coding mutations are suspected to contribute to risk. In addition, the experimental methods presented here are broadly useful for the rapid and economical resequencing of candidate

genes in extremely large cohorts, as may be required for the definitive implication of rare variants or de novo mutations in any genetically complex disorder.

References and Notes

1. G. V. Kryukov, A. Shpunt, J. A. Stamatoyannopoulos, S. R. Sunyaev, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3871 (2009).
2. B. J. O'Roak *et al.*, *Nat. Genet.* **43**, 585 (2011).
3. B. J. O'Roak *et al.*, *Nature* **485**, 246 (2012).
4. S. J. Sanders *et al.*, *Nature* **485**, 237 (2012).
5. B. M. Neale *et al.*, *Nature* **485**, 242 (2012).
6. I. Iossifov *et al.*, *Neuron* **74**, 285 (2012).
7. E. H. Turner, C. Lee, S. B. Ng, D. A. Nickerson, J. Shendure, *Nat. Methods* **6**, 315 (2009).
8. G. J. Porreca *et al.*, *Nat. Methods* **4**, 931 (2007).
9. S. Krishnakumar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9296 (2008).
10. See supplementary text on *Science* Online.
11. G. D. Fischbach, C. Lord, *Neuron* **68**, 192 (2010).
12. Materials and methods are available as supplementary materials on *Science* Online.
13. C. Betancur, *Brain Res.* **1380**, 42 (2011).
14. G. M. Cooper *et al.*, *Nat. Genet.* **43**, 838 (2011).
15. M. Lynch, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 961 (2010).
16. A. Kong *et al.*, *Nature* **488**, 471 (2012).
17. C. A. Williams, A. Dagli, A. Battaglia, *Am. J. Med. Genet. A.* **146A**, 2023 (2008).
18. R. S. Møller *et al.*, *Am. J. Hum. Genet.* **82**, 1165 (2008).
19. B. W. van Bon *et al.*, *Clin. Genet.* **79**, 296 (2011).
20. F. Guedj *et al.*, *Neurobiol. Dis.* **46**, 190 (2012).
21. M. E. Talkowski *et al.*, *Cell* **149**, 525 (2012).
22. J. Zhou, L. F. Parada, *Curr. Opin. Neurobiol.* **22**, 873 (2012).
23. I. Letunic, T. Doerks, P. Bork, *Nucleic Acids Res.* **40** (Database issue), D302 (2012).

Acknowledgments: We thank the National Heart, Lung, and Blood Institute, NIH Grand Opportunity (GO) Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the Women's Health Initiative Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010); we also thank B. Vernot, M. Dennis, T. Brown, and other members of the Eichler and Shendure labs for helpful discussions. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to phenotypic data on the Simons Foundation Autism Research Initiative (SFARI) Base. Approved researchers can obtain the SSC population dataset described in this study (https://ordering.base.sfari.org/~browse_collection/archive[ssc_v13]/ui:view) by applying at https://base.sfari.org. This work was supported by grants from the Simons Foundation (SFARI 137578, 191889 to E.E.E., J.S., and R.B.), NIH HD065285 (E.E.E. and J.S.), NIH NS069605 (H.C.M.), and R01 NS064077 (D.D.). E.B. is an Alfred P. Sloan Research Fellow. E.E.E. is an Investigator of the Howard Hughes Medical Institute. Scientific advisory boards or consulting affiliations: Ariosa Diagnostics (J.S.), Stratos Genomics (J.S.), Good Start Genetics (J.S.), Adaptive Biotechnologies (J.S.), Pacific Biosciences (E.E.E.), SynapDx (E.E.E.), DNAnexus (E.E.E.), and SFARI GENE (H.C.M.). B.J.O. is an inventor on patent PCT/US2009/30620: Mutations in contactin associated protein 2 are associated with increased risk for idiopathic autism. Raw sequencing data available at the National Database for Autism Research, NDARCOL1878.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1227764/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S14
Tables S1 to S18
References (*24*–*100*)

23 July 2012; accepted 1 November 2012
Published online 15 November 2012;
10.1126/science.1227764

Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell

Chenghang Zong,[1]* Sijia Lu,[1]*† Alec R. Chapman,[1,2]* X. Sunney Xie[1]‡

Kindred cells can have different genomes because of dynamic changes in DNA. Single-cell sequencing is needed to characterize these genomic differences but has been hindered by whole-genome amplification bias, resulting in low genome coverage. Here, we report on a new amplification method—multiple annealing and looping-based amplification cycles (MALBAC)—that offers high uniformity across the genome. Sequencing MALBAC-amplified DNA achieves 93% genome coverage ≥1x for a single human cell at 25x mean sequencing depth. We detected digitized copy-number variations (CNVs) of a single cancer cell. By sequencing three kindred cells, we were able to identify individual single-nucleotide variations (SNVs), with no false positives detected. We directly measured the genome-wide mutation rate of a cancer cell line and found that purine-pyrimidine exchanges occurred unusually frequently among the newly acquired SNVs.

Single-molecule and single-cell studies reveal behaviors that are hidden in bulk measurements (*1*, *2*). In a human cell, the genetic information is encoded in 46 chromosomes. The variations occurring in these chromosomes, such as single-nucleotide variations (SNVs) and copy-number variations (CNVs) (*3*), are the driving forces in biological processes such as evolution and cancer. Such dynamic variations are reflected in the genomic heterogeneity among a population of cells, which demands characterization of genomes at the single-cell level (*4*–*6*). Single-cell genomics analysis is also necessary when the number of cells available is limited to few or one, such as prenatal testing samples (*7*, *8*), circulating tumor cells (*9*), and forensic specimens (*10*).

Prompted by rapid progress in next-generation sequencing techniques (*11*), there have been several reports on whole-genome sequencing of single cells (*12*–*16*). These methods have relied on whole-genome amplification (WGA) of an individual cell to generate enough DNA for sequencing (*17*–*21*). However, WGA methods in general are prone to amplification bias, which results in low genome coverage. Polymerase chain reaction (PCR)–based WGA introduces sequence-dependent bias because of the exponential amplification with random primers (*17*, *18*, *22*). Multiple displacement amplification (MDA), which uses random priming and the strand-displacing ϕ29 polymerase under isothermal conditions (*19*), has provided improvements over PCR-based methods but still exhibits considerable bias, again due to nonlinear amplification.

Here we report a new WGA method, multiple annealing and looping-based amplification cycles (MALBAC), which introduces quasilinear preamplification to reduce the bias associated with nonlinear amplification. Picograms of DNA fragments (~10 to 100 kb) from a single human cell serve as templates for amplification with MALBAC (Fig. 1). The amplification is initiated with a pool of random primers, each having a common 27-nucleotide sequence and 8 variable nucleotides that can evenly hybridize to the templates at 0°C. At an elevated temperature of 65°C, DNA polymerases with strand-displacement activity are used to generate semiamplicons with variable lengths (0.5 to 1.5 kb), which are then melted off from the template at 94°C. Amplification of the semiamplicons gives full amplicons that have complementary ends. The temperature is cycled to 58°C to allow the looping of full amplicons, which prevents further amplification and cross-hybridizations. Five cycles of preamplification are followed by exponential amplification of the full amplicons by PCR to generate micrograms of DNA required for next-generation sequencing (Fig. 1). In the PCR, oligonucleotides with the common 27-nucleotide sequence are used as the primers.

We used MALBAC to amplify the DNA of single SW480 cancer cells. With ~25x mean

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA. [2]Program in Biophysics, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to the work.
†Present address: Yikon Genomics, 1 China Medical City Avenue, TQB Building, 5th floor, Taizhou, Jiangsu, China.
‡To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu