24. G. Zhang et al., Science **346**, 1311–1320 (2014).
25. C. R. Linnen et al., Science **339**, 1312–1316 (2013).

# HUMAN GENOMICS

# Health and population effects of rare gene knockouts in adult humans with related parents

Vagheesh M. Narasimhan,[1] Karen A. Hunt,[2]* Dan Mason,[3]* Christopher L. Baker,[4]* Konrad J. Karczewski,[5,6]* Michael R. Barnes,[7] Anthony H. Barnett,[8] Chris Bates,[9] Srikanth Bellary,[10] Nicholas A. Bockett,[2] Kristina Giorda,[11] Christopher J. Griffiths,[2] Harry Hemingway,[12,13] Zhilong Jia,[7] M. Ann Kelly,[14] Hajrah A. Khawaja,[7] Monkol Lek,[5,6] Shane McCarthy,[1] Rosie McEachan,[3] Anne O'Donnell-Luria,[5,6] Kenneth Paigen,[4] Constantinos A. Parisinos,[2] Eamonn Sheridan,[3] Laura Southgate,[2] Louise Tee,[14] Mark Thomas,[1] Yali Xue,[1] Michael Schnall-Levin,[11] Petko M. Petkov,[4] Chris Tyler-Smith,[1] Eamonn R. Maher,[15,16] Richard C. Trembath,[2,17] Daniel G. MacArthur,[5,6] John Wright,[3] Richard Durbin,[1]†‡ David A. van Heel[2]†‡

Examining complete gene knockouts within a viable organism can inform on gene function. We sequenced the exomes of 3222 British adults of Pakistani heritage with high parental relatedness, discovering 1111 rare-variant homozygous genotypes with predicted loss of function (knockouts) in 781 genes. We observed 13.7% fewer homozygous knockout genotypes than we expected, implying an average load of 1.6 recessive-lethal-equivalent loss-of-function (LOF) variants per adult. When genetic data were linked to the individuals' lifelong health records, we observed no significant relationship between gene knockouts and clinical consultation or prescription rate. In this data set, we identified a healthy *PRDM9*-knockout mother and performed phased genome sequencing on her, her child, and control individuals. Our results show that meiotic recombination sites are localized away from *PRDM9*-dependent hotspots. Thus, natural LOF variants inform on essential genetic loci and demonstrate *PRDM9* redundancy in humans.

Complete gene knockouts, typically caused by homozygous loss-of-function (LOF) genotypes, have helped researchers identify the function of many genes, predominantly through studies in model organisms and of severe Mendelian-inherited diseases in humans. However, information on the consequences of knocking out most human genes is still lacking. Naturally occurring complete gene knockouts offer the opportunity to study the effects of lifelong germline gene inactivation in living humans. A survey of LOF variants in adult humans revealed ~100 predicted LOF genotypes per individual, describing ~20 genes that carry homozygous predicted LOF alleles and hence are likely to be completely inactivated (*1*). Almost all of these homozygous genotypes were located at common variants with allele frequency >1%, in genes whose loss is likely to have weak or neutral effects on fitness and health (*1*). In contrast, rare predicted LOF genotypes were usually heterozygous and, thus, their overall effect on gene function is not known. A large exome sequencing aggregation study [conducted by the Exome Aggregation Consortium (ExAC)] of predominantly outbred individuals identified 1775 genes with homozygous predicted LOF genotypes in 60,706 individuals (*2*). Furthermore, 1171 genes with complete predicted LOF were identified in 104,220 Icelandic individuals (*3*), and modest enrichment for homozygous predicted LOF genotypes was shown in Finnish individuals (*4*). However, even in these large samples, homozygous predicted LOF genotypes tend to occur at variants of moderate allele frequency (~1%). Hence, these approaches will not readily assess knockouts in most genes, which are lacking such variants.

To identify knockouts created by rare homozygous predicted loss-of-function (rhLOF) variants, we sequenced the exomes of 3222 UK-dwelling adults of Pakistani heritage who were characterized as healthy, type 2 diabetic, or pregnant (*5*). These individuals have a high rate of parental relatedness (often through parents who are first cousins); thus, a substantial fraction of their autosomal genome occurs in long homozygous regions inferred to be identical by descent from a recent common ancestor (autozygous). We linked each person's genotype to their health care and epidemiological records, with the aims of (i) describing the properties and assessing the health effects of naturally occurring knockouts in an adult population; (ii) understanding the architecture of gene essentiality in the human genome, through the characterization of the population genetics of LOF variants; and (iii) conducting a detailed study of a *PRDM9* gene knockout, which plays a role in human meiotic recombination (*6*).

On average, 5.6% of the coding genome was autozygous, much higher than the percentage in outbred populations with European heritage (Fig. 1A and fig. S4). We identified, per individual, an average of 140.3 nonreference predicted

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [3]Bradford Institute for Health Research, Bradford Teaching Hospitals National Health Service (NHS) Foundation Trust, Bradford BD9 6RJ, UK. [4]Center for Genome Dynamics, The Jackson Laboratory, Bar Harbor, ME 04609, USA. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [7]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [8]Diabetes and Endocrine Centre, Heart of England NHS Foundation Trust and University of Birmingham, Birmingham B9 5SS, UK. [9]TPP, Mill House, Troy Road, Leeds LS18 5TN, UK. [10]Aston Research Centre for Healthy Ageing, Aston University, Birmingham B4 7ET, UK. [11]10X Genomics, 7068 Koll Center Parkway, Suite 415, Pleasanton, CA 94566, USA. [12]Farr Institute of Health Informatics Research, London NW1 2DA, UK. [13]Institute of Health Informatics, University College London, London NW1 2DA, UK. [14]School of Clinical and Experimental Medicine, University of Birmingham, Birmingham B15 2TT, UK. [15]Department of Medical Genetics, University of Cambridge and National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, Box 238, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [16]Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [17]Faculty of Life Sciences and Medicine, King's College London, London SE1 1UL, UK.

*These authors contributed equally to this work. †These authors contributed equally to this work. ‡Corresponding author. Email: rd@sanger.ac.uk (R.D.); d.vanheel@qmul.ac.uk (D.A.v.H.)

LOF genotypes comprising 16.1 rare heterozygotes (minor allele frequency <1%), 0.34 rare homozygotes, 83.2 common heterozygotes, and 40.6 common homozygotes. Nearly all rhLOF genotypes were found within autozygous segments (94.9%) (5), and the mean number of rhLOF variants per individual was proportional to the rate of autozygosity (Fig. 1B). Overall, we identified 1111 rhLOF genotypes at 847 variants (575 annotated as LOF variants in all GENCODE basic transcripts) in 781 different protein-coding genes (Fig. 1C) (5) in 821 individuals. Autozygous segments were observed across all coding regions, with a density distribution that was not significantly different from random (5) (Shapiro-Wilk test; $P = 0.112$). From these values, we estimate that 41.5% of individuals with 6.25% autozygosity (expected mean for an individual whose parents are first cousins but are otherwise outbred) will have one or more rhLOF genotypes (Fig. 1B).

The majority (422) of the identified genes with rhLOF genotypes had not been previously reported, although 167 had been reported as containing homozygous or compound heterozygous LOF genotypes in Iceland, and 299 appeared in the ExAC data. In total, 107 rhLOF genes were common to all three data sets (5), suggesting a subset of genes that are either tolerant of LOF variants and/or have higher rates of mutation. Eighty-nine rhLOF genotypes were homozygous without observed heterozygotes, and we identified three individuals who each had five rhLOF genotypes. On the basis of these observations, we predict that in 100,000 individuals with first-cousin–related parents of the same genetic ancestry, at least one knockout would occur in ~9000 of the ~20,000 human protein-coding genes (fig. S3) (5).

We observed a lower density of annotated rare LOF variants within autozygous tracts (where the genotypes are homozygous) compared with outside autozygous tracts (where they are typically heterozygous), indicating direct negative selection on a fraction of homozygotes (Fig. 2A). We matched each of the 16,708 rare annotated LOF (heterozygous and homozygous) variants to a randomly selected synonymous variant of the same allele frequency and observed 842 rare LOF variants with at least one homozygous genotype versus an average of 975.5 rare synonymous variants with at least one homozygote, which indicates a deficit of 13.7% [95% confidence interval (CI): 8 to 20%] of variants with rhLOF genotypes (Fig. 2B) (5). We attribute this deficit to some rhLOF genotypes resulting in early lethality or severe disease and thus being incompatible with our selection criteria for healthier adults, although our data do not inform whether these cases are due to fewer high-penetrance or more low-penetrance variants. This deficit is higher than in the Icelandic population (6.4%) (3), consistent with that analysis being biased toward more common variants already subject to selection.

We then combined the calculated deficit rate with the observed number of heterozygous annotated LOF variants, integrating across allele frequencies, to obtain a direct estimate of the recessive lethal load per person. This finding suggests that a typical individual from the population we sampled carries 1.6 recessive annotated LOF lethal-equivalent variants in the heterozygous state (5). Our estimate is similar to previous approximations of the lethal load—calculated by correlating the number of miscarriages, stillbirths, and infant mortalities with the level of autozygosity (Fig. 2C) (7, 8)—and is also similar to measurements in other species (9). Using epidemiological data from 13,586 mothers from the same Born in Bradford birth cohort studied in our genetic analysis, we esti-

mated 0.5 lethal equivalents resulting in miscarriage, stillbirth, or infant mortality per individual in our population (5). The difference between our two estimates can be accounted for by the fact that the first includes embryonic deaths, whereas the second involves fatalities occurring only after a registered pregnancy, which suggests that there are twice as many embryonically lethal recessive mutations as those that result in fetal or infant death. Controlling for other effects by comparing to synonymous mutations, we see a significant but moderate decrease ($R_{A/B}$ jackknife test; $P = 0.04$) in the rhLOF mutational load in our Pakistani-heritage population data set compared with outbred populations from the 1000 Genomes Project, although the observed decrease is less than that caused by the historic bottleneck in the Finnish population (FIN in Fig. 2D) (5).

We examined 215 rhLOF genes from our data set that have an exact 1:1 mouse:human gene ortholog. Analysis of mouse gene knockout data revealed 52 genes for which a lethal mouse phenotype had been reported on at least one genetic background (10). Whether or not the mouse ortholog knockout is lethal has no bearing on alteration of protein function, duplication, or changes in gene expression (5). Genes containing rhLOF showed 50% fewer molecular interactions compared with all genes in the STRING interactome data set (Kruskal-Wallis test; $P = 3.4 \times 10^{-9}$); this result was predominantly driven by genes in the binding interaction category (Kruskal-Wallis $P = 9.3 \times 10^{-11}$) (table S4). We saw a similar reduction in the Icelandic data, in contrast to both known pathogenic LOF variants and pathogenic gain-of-function variants reported in the Orphanet reference portal, which showed increased overall molecular interactions (Kruskal-Wallis $P = 1.1 \times 10^{-6}$ and $2 \times 10^{-12}$, respectively) (table S4) (5). Furthermore,
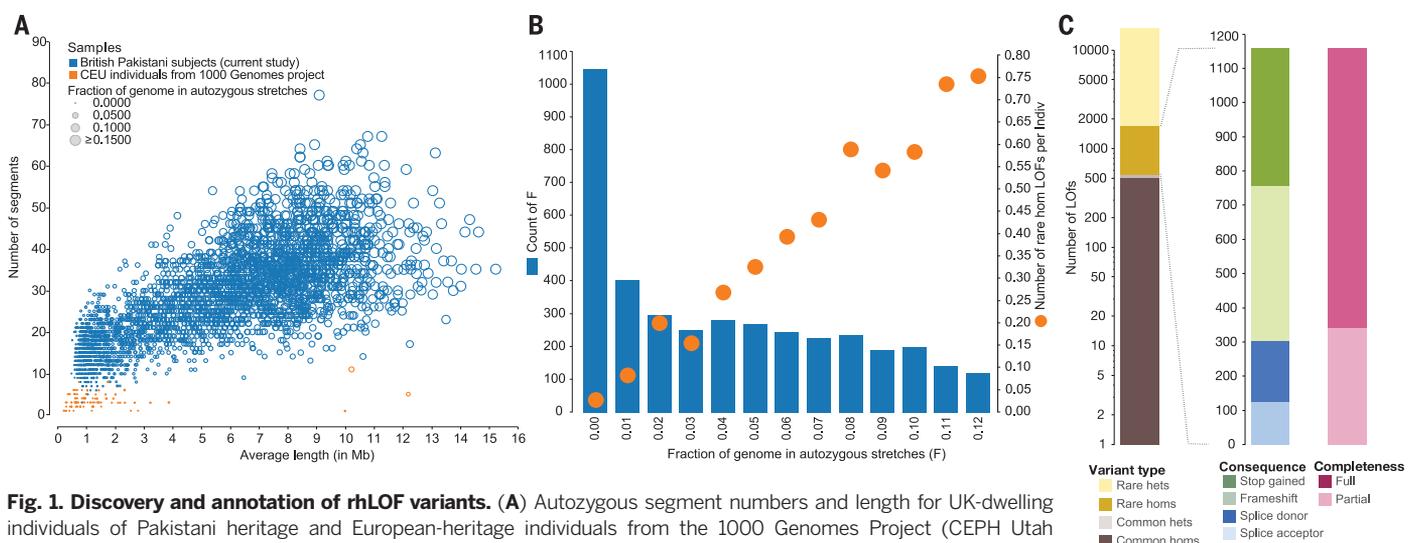
**Fig. 1. Discovery and annotation of rhLOF variants.** (**A**) Autozygous segment numbers and length for UK-dwelling individuals of Pakistani heritage and European-heritage individuals from the 1000 Genomes Project (CEPH Utah residents with ancestry from Northern and Western Europe; CEU). (**B**) Autozygosity and rhLOF in 3222 individuals. The graph shows the number (count) of individuals (left *y* axis, blue columns) binned by fraction of the autozygous genome (F; *x* axis) plotted against the mean number of rhLOF genotypes per individual (right *y* axis, orange circles). (**C**) Distribution of LOF variants by variant type (allele frequency and heterozygous-only versus those containing a homozygous genotype), predicted protein consequence, and transcript completeness (i.e., whether predicted for a full or partial set of GENCODE basic transcripts for the gene).

rhLOF genes that are drug targets have an 11.4% rate of going from phase 1 clinical trials to approval versus 6.7% for all target-indication pairs ($\chi^2$ test; $P = 0.046$), yet we observed no difference in the proportion of genes known or predicted to be druggable targets (11) when rhLOF genes (15%) were compared to all genes (13%, $P = 0.098$) (5).

In participants from the Born in Bradford study, where full health record data was available, we observed 52 individuals with 54 rhLOF genotypes in recessive disease genes, using the Online Mendelian Inheritance in Man (OMIM) catalog "confirmed" category. We expected that these variants would be enriched for false-positive observations (1). After a quality-control analysis of the sequence-based genotype calls (5), we inspected the annotation of these variants (1) (table S2). Of the 54 rhLOF genotypes, we considered 16 to be possible genome annotation errors (i.e., incorrectly described as LOF variants) (5) (table S2). Only six of the remaining 38 rhLOF variants were linked to a definite lifetime primary health record diagnosis consistent with the OMIM phenotype, although a further three genotypes were suggestively compatible (table S3). We suggest that the remaining 29 are due to a combination of incomplete penetrance (12–16), late onset of disease (i.e., not yet having occurred), individuals with mild symptoms not seeking medical attention, unrecognized technical issues with sequencing or annotation (e.g., tissue-specific alternative splicing), or dubious

evidence to support the gene-phenotype assignment (in table S3, we assess the available evidence for these possibilities).

We next assessed electronic health records in the Born In Bradford adults, focusing on the time since study recruitment (5). Drug prescription rate and clinical staff consultation rate have previously been shown to correlate strongly with health status (17). We compared individuals with one or more rhLOF genotype ($n = 638$) to individuals without rhLOF ($n = 1524$) and found no association with prescription rate [logistic regression, odds ratio (OR): 1.001, 95% CI: 0.988 to 1.0144] or consultation rate (OR: 1.017, 95% CI: 0.996 to 1.038), nor any associations in rhLOF subgroups (5).

One individual in our study was a healthy adult mother with a predicted rare homozygous LOF mutation in *PRDM9*, which we confirmed experimentally (5) (fig. S7, A and B). PRDM9 is the major known determinant of the genomic locations of meiotic recombination events in humans and mice through its DNA binding site zinc finger domain (6, 18, 19). We excluded the possibility that this rhLOF variant was from a somatic loss-of-heterozygosity event on the basis that this individual's genotype is heterozygous, not homozygous, on both sides of the 25-Mb autozygous region (fig. S7C). Her lifetime primary- and secondary-care health records were unremarkable. Her genotype predicts protein truncation in the SET methyltransferase domain (thus lacking the DNA-binding zinc finger domain), which

we confirmed by in vitro expression analysis (fig. S8A). We observed no increase in H3K4Me3 global methylation upon transfection (20) of the truncation allele (fig. S8A) and that R345Ter specifically disrupted *PRDM9*-dependent H3K4Me3 methylation at hotspots (fig. S8B).

We performed long-range molecularly phased whole-genome sequencing to determine the locations of meiotic recombination in the maternal gamete transmitted from mother to child and identified 39 candidate crossovers in the process (5). Using maps of double-strand breaks (DSBs) and a maximum likelihood model to account for variability in region size and hotspot density (18), we estimated that only 5.9% (2 log unit CI: 0 to 24%) of the observed *PRDM9*-knockout–duo maternal gamete crossovers matched DSB sites from homozygotes with the wild-type *PRDM9-A* allele (5). In comparison, for a control mother-child Centre d'Etude du Polymorphisme Humain (CEPH) pedigree duo homozygous for *PRDM9-A*, we estimated that 52.1% (CI: 36 to 69%) of the crossovers occurred in known DSB sites. Using similar methods, we saw that 18.5% of the crossovers observed in the *PRDM9*-knockout duo (CI: 1 to 42%) occurred in linkage-disequilibrium–based hotspots versus 75.7% in the control duo [CI: 57 to 89%, consistent with a previously published estimate of an average of 60% of crossovers occurring at hotspots (18)] (5).

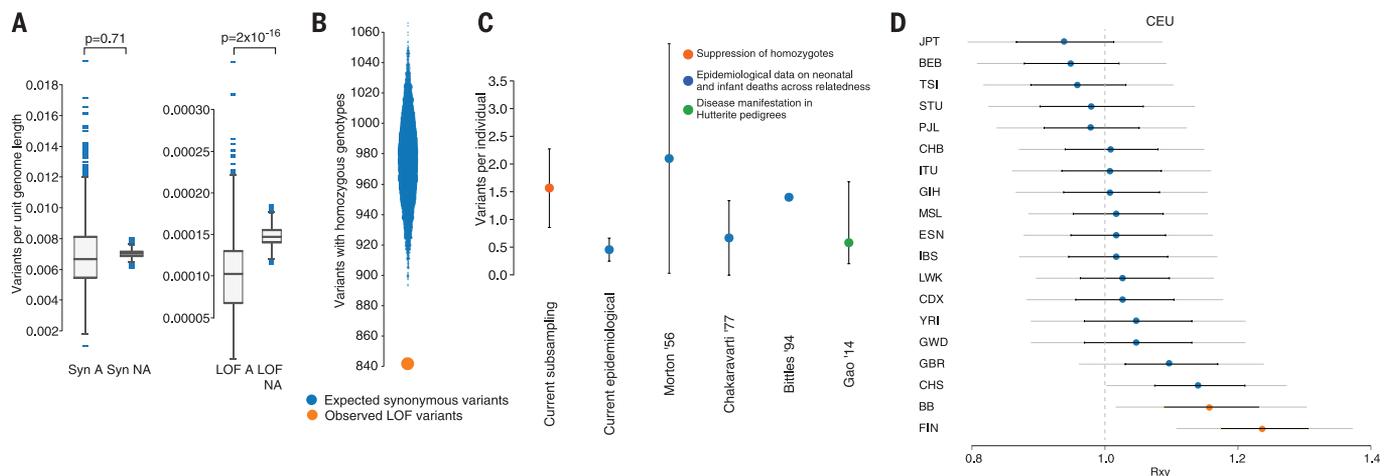*Prdm9*-knockout mice demonstrate abnormal location of recombination hotspots with enrichment



**Fig. 2. Population genetic analysis of rhLOF variants.** (**A**) Comparison of the number of LOF variants per unit length in autozygous regions (LOF A) with the expected rate from nonautozygous sections (LOF NA), showing suppression of rhLOFs (*t* test). A similar analysis of synonymous (Syn) variants revealed no significant differences. (**B**) Observed number of variants with homozygote genotypes in 16,708 rare LOF variants (orange circle) versus a frequency-matched subsampling of synonymous variants (blue violin plot). (**C**) Quantification of the recessive lethal load carried, on average, by a single individual. Orange circle, direct subsampling estimate for rhLOF variants from current study; blue circles, epidemiological estimates from correlating infant mortality rates to estimated autozygosity in current and published data; green circle, direct estimate from a large Hutterite pedigree. Black bars denote 95% CIs. (**D**) Relative number of derived LOF alleles that are frequent in one population and not another (under neutrality, the expectation is 1.0). Results were calculated for 1000 Genomes Project populations and the current Birmingham/Bradford Pakistani heritage population (BB), as compared with the CEU population. Error bars represent ±1 (black) or 2 (gray) SEs. Significant differences ($R_{A/B}$ jackknife test) from CEU population data are denoted by orange circles. JPT, Japanese (Tokyo, Japan); BEB, Bengali (Bangladesh); TSI, Toscani (Italy); STU, Sri Lankan Tamil (UK); PJL, Punjabi (Lahore, Pakistan); CHB, Han Chinese (Beijing, China); ITU, Indian Telugu (UK); GIH, Gujarati Indian (Houston, TX); MSL, Mende (Sierra Leone); ESN, Esan (Nigeria); IBS, Iberian (Spain); LWK, Luhya (Webuye, Kenya); CDX, Chinese Dai (Xishuangbanna, China); YRI, Yoruba (Ibadan, Nigeria); GWD, Gambian (Western Divisions in the Gambia); GBR, British (England and Scotland); CHS, Southern Han Chinese (China); FIN, Finnish (Finland).

at gene promoters and enhancers, fail to properly repair DSBs, and are infertile (both sexes sterile) (*21, 22*). Dogs, which lack *PRDM9*, retain recombination hotspots that occur in regions with high GC content (*23*), unlike in humans or knockout mice. It has been speculated that meiotic recombination in dogs is controlled by an ancestral mammalian mechanism and that PRDM9 competes and usurps these sites when active in noncanids (*23, 24*). However, we did not see increased overlap in our *PRDM9*-knockout–duo crossover intervals with promoters and their flanking regions or enrichment in GC content, as compared to the control duo (*5*). Thus, the existence of a healthy and fertile *PRDM9*-deficient adult human points to differences in humans versus both mice and dogs and supports the possibility of alternative mechanisms of localizing human meiotic crossovers (*25, 26*).

Together, these data suggest that apparent rhLOF genotypes identified by exome or genome sequencing of adult populations require cautious interpretation. Although this class of variants has the greatest predicted effect on protein function, the loss of most proteins is relatively harmless to the individual. Even in previously annotated disease genes, predicted rare LOF homozygotes may not always be as clinically relevant as often considered. This finding is becoming increasingly important now that exome- and genome-sequencing studies of healthier adults are rapidly expanding. We anticipate that further efforts to identify naturally occurring human knockouts—whether in bottlenecked populations or, more efficiently, in individuals with related parents, as described here—will yield both new data relevant to clinical interpretation and new biological insights,

as exemplified by our investigation of a *PRDM9*-deficient healthy and fertile woman.

## REFERENCES AND NOTES

1. D. G. MacArthur *et al.*, *Science* **335**, 823–828 (2012).
2. Exome Aggregation Consortium *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. 10.1101/030338 (2015); http://biorxiv.org/content/early/2015/10/30/030338.
3. P. Sulem *et al.*, *Nat. Genet.* **47**, 448–452 (2015).
4. E. T. Lim *et al.*, *PLOS Genet.* **10**, e1004494 (2014).
5. See supplementary materials on *Science* Online.
6. E. D. Parvanov, P. M. Petkov, K. Paigen, *Science* **327**, 835 (2010).
7. R. Chakraborty, A. Chakravarti, *Hum. Genet.* **36**, 47–54 (1977).
8. A. H. Bittles, J. V. Neel, *Nat. Genet.* **8**, 117–121 (1994).
9. A. R. McCune *et al.*, *Science* **296**, 2398–2401 (2002).
10. B.-Y. Liao, J. Zhang, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6987–6992 (2008).
11. M. R. Nelson *et al.*, *Nat. Genet.* **47**, 856–860 (2015).
12. J. Flannick *et al.*, *Nat. Genet.* **45**, 1380–1385 (2013).
13. H. L. Rehm *et al.*, *N. Engl. J. Med.* **372**, 2235–2242 (2015).
14. K. Walter *et al.*, *Nature* **526**, 82–90 (2015).
15. J. J. Johnston *et al.*, *Am. J. Hum. Genet.* **96**, 913–925 (2015).
16. S. L. Van Driest *et al.*, *JAMA* **315**, 47–57 (2016).
17. S. L. Brilleman *et al.*, *Br. J. Gen. Pract.* **63**, e274–e282 (2013).
18. F. Baudat *et al.*, *Science* **327**, 836–840 (2010).
19. S. Myers *et al.*, *Science* **327**, 876–879 (2010).
20. C. L. Baker *et al.*, *PLOS Genet.* **11**, e1005512 (2015).
21. K. Hayashi, K. Yoshida, Y. Matsui, *Nature* **438**, 374–378 (2005).
22. K. Brick, F. Smagulova, P. Khil, R. D. Camerini-Otero, G. V. Petukhova, *Nature* **485**, 642–645 (2012).
23. E. Axelsson, M. T. Webster, A. Ratnakumar, C. P. Ponting, K. Lindblad-Toh, *Genome Res.* **22**, 51–63 (2012).
24. I. L. Berg *et al.*, *Nat. Genet.* **42**, 859–863 (2010).
25. F. Pratto *et al.*, *Science* **346**, 1256442 (2014).
26. M. Lichten, *Science* **350**, 913 (2015).