UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population

Permalink

https://escholarship.org/uc/item/4ms3s7pr

Journal

Proceedings of the National Academy of Sciences of the United States of America, 117(5)

ISSN

0027-8424

Authors

Kessler, Michael D
Loesch, Douglas P
Perry, James A
et al.

Publication Date

2020-02-04

DOI

10.1073/pnas.1902766117

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population

Michael D. Kessler[a,b,c,d], Douglas P. Loesch[a,b,c], James A. Perry[b,c], Nancy L. Heard-Costa[e,f], Daniel Taliun[g], Brian E. Cade[h,i], Heming Wang[h,i], Michelle Daya[j], John Ziniti[k], Soma Datta[k], Juan C. Celedón[l], Manuel E. Soto-Quiros[m], Lydiana Avila[m], Scott T. Weiss[k,n], Kathleen Barnes[j], Susan S. Redline[h,o,p], Ramachandran S. Vasan[f], Andrew D. Johnson[f,q], Rasika A. Mathias[r,s], Ryan Hernandez[t], James G. Wilson[u], Deborah A. Nickerson[v], Goncalo Abecasis[w], Sharon R. Browning[x], Sebastian Zöllner[y,z], Jeffrey R. O'Connell[b,c], Braxton D. Mitchell[b,c,aa], National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium[1], TOPMed Population Genetics Working Group[2], and Timothy D. O'Connor[a,b,c,d,3]

[a]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; [b]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [c]Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD 21201; [d]University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201; [e]Department of Neurology, Boston University School of Medicine, Boston, MA 02118; [f]Framingham Heart Study, Framingham, MA 01702; [g]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109; [h]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115; [i]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142; [j]Department of Medicine, University of Colorado Denver, Aurora, CO 80045; [k]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115; [l]Division of Pediatric Pulmonary Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213; [m]Department of Pediatrics, Hospital Nacional de Niños, 10103 San José, Costa Rica; [n]Department of Medicine, Harvard Medical School, Boston, MA 02115; [o]Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115; [p]Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215; [q]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, The Framingham Heart Study, Framingham, MA 01702; [r]Division of Allergy and Clinical Immunology, The Johns Hopkins School of Medicine, Baltimore, MD 21224; [s]Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD 21218; [t]Quantitative Life Sciences, McGill University, Montreal, QC H3A OG4, Canada; [u]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216; [v]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [w]School of Public Health, University of Michigan, Ann Arbor, MI 48109; [x]Department of Biostatistics, University of Washington, Seattle, WA 98195; [y]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109; [z]Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109; and [aa]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD 21201

Edited by Stephen T. Warren, Emory University School of Medicine, Atlanta, GA, and approved December 17, 2019 (received for review February 19, 2019)

De novo mutations (DNMs), or mutations that appear in an individual despite not being seen in their parents, are an important source of genetic variation whose impact is relevant to studies of human evolution, genetics, and disease. Utilizing high-coverage whole-genome sequencing data as part of the Trans-Omics for Precision Medicine (TOPMed) Program, we called 93,325 single-nucleotide DNMs across 1,465 trios from an array of diverse human populations, and used them to directly estimate and analyze DNM counts, rates, and spectra. We find a significant positive correlation between local recombination rate and local DNM rate, and that DNM rate explains a substantial portion (8.98 to 34.92%, depending on the model) of the genome-wide variation in population-level genetic variation from 41K unrelated TOPMed samples. Genome-wide heterozygosity does correlate with DNM rate, but only explains <1% of variation. While we are underpowered to see small differences, we do not find significant differences in DNM rate between individuals of European, African, and Latino ancestry, nor across ancestrally distinct segments within admixed individuals. However, we did find significantly fewer DNMs in Amish individuals, even when compared with other Europeans, and even after accounting for parental age and sequencing center. Specifically, we found significant reductions in the number of C→A and T→C mutations in the Amish, which seem to underpin their overall reduction in DNMs. Finally, we calculated near-zero estimates of narrow sense heritability (h^2), which suggest that variation in DNM rate is significantly shaped by nonadditive genetic effects and the environment.

de novo mutations | Amish | mutation rate | recombination | diversity

De novo mutations (DNMs) appear constitutively in an individual despite not being seen in their parents, and their identification and study are critically important to our understanding of human genomic evolution (1–11). For example, it is necessary to understand the rate at which DNMs accumulate in order to calibrate evolutionary models of species divergence. DNMs are also implicated in many diseases, including rare genetic disorders (8, 12, 13) and common complex diseases, such as autism and schizophrenia (13–15). Early studies indirectly inferred mutation rate estimates from patterns of rare Mendelian diseases

Author contributions: M.D.K., N.L.H.-C., B.E.C., M.D., J.C.C., M.E.S.-Q., L.A., S.T.W., K.B., S.S.R., R.S.V., A.D.J., R.A.M., R.H., J.G.W., D.A.N., G.A., S.R.B., S.Z., J.R.O., B.D.M., N.H.L.a.B.I.T.-O.f.P.M.T.C., T.P.G.W.G., and T.D.O. designed research; M.D.K., D.P.L., and J.R.O. performed research; M.D.K. and J.R.O. contributed new reagents/analytic tools; M.D.K., D.P.L., J.A.P., N.L.H.-C., D.T., B.E.C., H.W., M.D., J.Z., S.D., J.C.C., M.E.S.-Q., S.T.W., K.B., A.D.J., R.A.M., R.H., G.A., S.R.B., S.Z., J.R.O., B.D.M., and T.D.O. analyzed data; and M.D.K. and T.D.O. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: Raw sequence data are available through dbGaP as part of the TOPMed program, and is available via study-specific accessions. These are: Whole Genome Sequencing and Related Phenotypes in the "Framingham Heart Study" (phs000974) and "NHLBI TOPMed, Genetics of Cardiometabolic Health in the Amish" (phs000956), "NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados" (phs001143), "NHLBI TOPMed: The Cleveland Family Study (WGS)" (phs000954), and "NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica" (phs000988). Centralized alignments and variant calls are also available via these dbGaP accessions. After working with the TOPMed Consortium, we have created DNM call-set files with chromosome, position, reference allele, and alternative allele, as well as meta data files with paternal age, maternal age, sequencing center, European ancestry proportion, African ancestry proportion, Native American ancestry proportion, and self-reported ancestry. The cohort-specific files are available upon request.

[1]A complete list of the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium can be found in *SI Appendix*.

[2]A complete list of the TOPMed Population Genetics Working Group can be found in *SI Appendix*.

[3]To whom correspondence may be addressed. Email: timothydoconnor@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1902766117/-/DCSupplemental.

First published January 21, 2020.

Significance

Here we provide the most diverse human de novo mutation call set to date, and use it to quantify the genome-wide relationship between local mutation rate and population-level rare genetic variation. While we demonstrate that the human single-nucleotide mutation rate is similar across numerous human ancestries and populations, we also discovered a reduced mutation rate in the Amish founder population, which shows that mutation rates can shift rapidly. Finally, we find that variation in mutation rates is not heritable, which suggests that the environment may influence mutation rates more significantly than previously realized.

or from DNA substitution rates between species (4, 16, 17). More recently, modern sequencing technologies have enabled the use of pedigree data to directly estimate the number of new mutations found across the genome (17–21). These pedigree-based studies have identified both paternal and maternal age effects (8, 22), and most recently estimate contributions of 1.51 and 0.37 DNMs per year of paternal and maternal age, respectively (19). While these effects are often explained on the basis of DNA replication errors, recent studies have found that DNA repair processes are likely to be major contributors to the mutations that accrue in both paternal and maternal gametes (21, 23–25). Some of this repair-associated mutation accumulation has been found to be due to maternal age-dependent DNA damage in oocytes and to maternal age-dependent postzygotic mutation increases (25). These studies have also identified specific mutation patterns, such as C→G transversions, that strongly associate with DNA double-strand breaks and repair (24, 25). In addition, these repair-associated mutations have been found to cluster together in distinct patterns, to predominate in certain genomic regions, and to be associated with recombination (21, 23, 24), which has itself been shown to influence mutation rates (26–28).

Other features have also been reported to influence variation in mutation rates across the genome. GC content was recently shown to directly increase single-nucleotide and structural mutation rates, and was also shown to increase recombination rates (29). Chromatin structure has also been shown to associate with DNA mutations, with DNA replication times associating significantly with point mutations (30), and nucleotide positioning found to significantly modulate mutation rate (31). Recent work has tied a number of these features together by providing resolute and individualized recombination maps, and using them to demonstrate a positive relationship between maternal age and the rates and locations of meiotic crossovers (32). Specifically, older mothers have increased recombination that also shifts toward late replicating regions and low GC regions, and numerous loci seem to genetically influence meiotic recombination. Other recent work evaluates nucleotide content, histone and chromatin features, replication timing, and recombination to provide one of the clearest pictures to date of the factors that shape genome-wide variability in the human mutation rate (23). With regard to base content, Amos (33) has proposed the "Heterozygote Instability" hypothesis, which challenges the assumption that population size and mutation rate are independent, and suggests their interdependence on the basis of heterozygosity. According to this hypothesis, the occurrence of gene conversion events at heterozygous sites during meiosis could locally increase mutation rates, and Amos (34) uses substitution rates to provide support for this. Yang et al. (35) test this hypothesis using parent–offspring sequencing of *Arabidopsis*, rice, and honey bee, and show support for the relationship between heterozygosity and mutation rate by demonstrating an ∼3.5-fold higher mutation rate in heterozygotes compared with homozygotes, and mutation occurring closer to heterozygous sites and crossover events.

While identifying these mutational correlates have helped us better understand the biological processes that drive mutation, genetic estimates of mutation rate are one-half the magnitude of those originally inferred phylogenetically (8, 17–19, 36–38). This has raised questions about the accuracy of these genetics estimates, as well as about the accuracy of human evolutionary time points calculated using phylogenetic estimates. While it has been proposed that the failure of genetic methods to account for postzygotic mutations in the parent might bias estimates down and partly explain this discrepancy (19), recent work using sibling recurrence suggests only minor mutation rate estimate increases when accounting for a substantial portion of these mutations (39). This discrepancy has also raised the possibility that mutation rates have evolved more rapidly than previously assumed, and that molecular clock-type analyses are therefore flawed (17, 36). For example, analyses of base pair substitution patterns have identified mutational differences between human populations, and showed most notably that the rate of TCC→TTC transitions appears to have increased in Europeans thousands of years ago for some finite period of time (40, 41).

These findings provide a rationale for how mutation rates might differ between human populations. However, since most studies of DNMs have used data from small cohorts of individuals with predominantly European ancestry (8, 10, 42), little is known about the role of DNMs in the evolution and health of populations of predominantly non-European ancestry, and it is unclear whether DNM rates vary across different human populations. To address this and other questions about mutation, we used a high-coverage whole-genome sequencing (WGS) dataset generated by the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) program (43) to directly estimate and analyze DNM accumulation across multiple human ancestries and populations. After analyzing genome-wide patterns of mutation using a call set of 93,325 single-nucleotide variant (SNV) DNMs, we compared DNM counts and rates across five TOPMed cohorts that represent European, African, and Native American (Latino) ancestry individuals, and that include Amish individuals from a founder population with European ancestry. We also estimate the correlation between heterozygosity and SNV DNM count, and then test whether mutation rate is a heritable trait in anticipation of using genome-wide association studies (GWAS) to look for mutation rate-modifying loci.

Results

TOPMed Dataset and Positive Correlation between DNMs and Parental Age. Using WGS data for 1,465 individuals and their parents from the TOPMed initiative (43), we identified a DNM call set and compared DNM accumulation rates across ancestral background (Table 1). The analyzed individuals belong to five TOPMed cohorts with varied ancestral backgrounds: 1) The Amish, which are an isolated European founder population; 2) the Barbados Asthma Genetics Study (BAGS), which consists of individuals with predominantly African ancestry; 3) the Cleveland Family Study (CFS), which consists of both European and African American individuals; 4) the Genetic Epidemiology of Asthma in Costa Rica and the Childhood Asthma Management Program, which are collectively referred to as the CRA cohort and consist of admixed Latino individuals; and 5) the Framingham Heart Study (FHS), which consists of individuals with European American ancestry. For our analyses, we treated the CFS study as two separate cohorts (signified as CFS_AFR and CFS_EUR, respectively). These make up our six analysis cohorts (Amish, BAGS, CFS_AFR, CFS_EUR, CRA, FHS) (Table 1). After removing samples with DNM counts that were extreme outliers (often due to pedigree errors) (*SI Appendix*), we were left with a DNM call set of 93,325 SNVs across 1,449

Table 1. Cohort characteristics and mutation estimates

Study name	TOPMed project	No. of children (after outlier removal)	No. of Nuclear Families (after outlier removal)	Average paternal age at conception (years ± SD)	Average maternal age at conception (years ± SD)	Populations	Mutation rate	Mutation rate 95% CI	Parental age effect	Parental age effect 95% CI	Variance explained
AMISH	Genetics of Cardiometabolic Health in the Amish	115 (115)	59 (59)	29.24 ± 5.10	27.03 ± 5.24	Old Order Amish large extended pedigrees	1.13E-08	1.084E-08, 1.176E-08	1.313	0.901, 1.726	0.398
BAGS	Barbados Asthma Genetics Study	210 (208)	125 (124)	31.74 ± 7.12	27.27 ± 5.92	African Ancestry (from Barbados)	1.27E-08	1.233E-08, 1.311E-08	1.503	1.273, 1.733	0.727
CFS_AFR	The Cleveland Family Study	31 (31)	22 (22)	28.27 ± 6.93	24.96 ± 4.76	African American	1.20E-08	1.090E-08, 1.315E-08	1.932	1.283, 2.582	0.806
CFS_EUR	The Cleveland Family Study	100 (99)	52 (52)	29.97 ± 5.01	27.97 ± 4.92	European American	1.22E-08	1.177E-08, 1.270E-08	1.613	1.223, 2.003	0.716
CRA	Genetic Epidemiology of Asthma in Costa Rica (GACRS), Childhood Asthma Management Program (CAMP)	316 (310)	278 (276)	29.74 ± 6.56	26.77 ± 6.02	Costa Rican (Latino/ Hispanic)	1.22E-08	1.187E-08, 1.250E-08	1.591	1.410, 1.772	0.696
FHS	Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study	693 (686)	678 (672)	29.50 ± 5.28	27.39 ± 4.77	European American	1.21E-08	1.189E-08, 1.232E-08	1.739	1.564, 1.913	0.495

Study cohorts and metadata are described. The six cohorts used in this study derive their names ("Study name") from five TOPMed projects ("TOPMed project"), and represent a diversity of populations and ancestries ("Populations"). Sample sizes are shown ("No. of children") along with mean paternal and maternal age values per cohort (after the removal of DNM outliers). BAGS individuals have the highest average paternal age, which seems to explain their elevated DNM rate, and CFS_AFR individuals have the lowest maternal ages. The estimated average mutation rate and 95% CI per cohort is also shown (calculated after removal of outliers), as are parental age effects (estimated using paternal age alone, due to confounding between paternal and maternal ages), and the proportion of DNM variance explained by this parental age effect after accounting for Poisson variation.

individuals. This equates to an average of about 64.4 (95% CI: 63.6 to 65.2) mutations per individual, which is consistent with previous findings (8, 22, 24). These 1,449 individuals come from 1,201 independent nuclear families, and for our analysis we treated each child and their two parents as a trio. To control for any potential confounding effects from this sibling data, we repeated all of our analyses that focused on per individual DNM measures after randomly choosing one child per family. The results from these repeated analyses are qualitatively the same as those from our full analysis, and for simplicity, we only report results from analyses run with our full dataset (except when noted within our heritability models).

Using this call set, we calculated per sample mutation rate estimates as SNV DNM count divided by the number of autosomal bases with depth ≥10 and quality ≥30, which serves as a good representation of the number of bases evaluated for DNMs by our filter heuristic (see *SI Appendix* for additional details). This results in a mean SNV DNM rate estimate of 1.215×10^{-8} (95% CI: 1.201×10^{-8} to 1.230×10^{-8}) mutations per base pair per generation. While our DNM rate estimate is slightly lower than previous genetic estimates (see *SI Appendix* for additional details), our estimate is generally concordant with recent genetic estimates (19, 20, 37), and lends support to the accuracy of our filtering approach. As expected based on previous studies (8, 19), we found a highly significant association between DNM rate per individual and paternal age (linear regression, $R^2 = 0.410$, $P < 2.98 \times 10^{-162}$) (*SI Appendix*, Fig. S1*A*), which provides an additional degree of validation for our approach and our call set. While the high correlation in our dataset between paternal and maternal age makes it difficult to evaluate the separate effect of each on DNM count, linear modeling does succeed in identifying significant paternal and maternal age effects that are consistent in magnitude with those of recent studies (1.35 mutations per year of father's age, and 0.42 mutations per year of mother's age) (19, 21, 22). DNM totals per individual did not differ significantly on the basis of the sex of the individual for whom the DNMs were called (*SI Appendix*, Fig. S1*B*), their year of birth (*SI Appendix*, Fig. S1*C*), or the age of their DNA (i.e., the individual's age) at the time of collection (*SI Appendix*, Fig. S1*D*).

DNM Mutation Types and Patterns. Using variant effect predictor annotations (44) from the TOPMed Consortium (43) for loss-of-function (LOF) variants found within the genomes of the 1,201 single offspring trios, we found 8,499 LOF variants (*SI Appendix*, Fig. S2). These 1,201 offspring also have 77,015 DNMs. Looking at the intersection of these DNM and LOF call sets, we found 66 LOF DNM mutations. Therefore, 0.086% of DNMs are LOF mutations, while 0.778% of LOF variants are DNMs, which suggests that DNMs contribute significantly each generation to the total number of segregating LOF variants. Using the set of 93,099 DNMs that remain after counting only once those DNMs that are recurrent in siblings, and mapping our DNM set over to hg38 so as to be compatible with recent genetic maps, we found that 1,582 (1.7%) DNMs are in coding bases, which is consistent with the proportion of the genome comprised of coding sequence.

We then used this set of 93,099 DNMs to evaluate mutational patterns across the genome. Consistent with previous findings (19, 21), the most frequent mutation types in our DNM call set are C→T (43.31%) and T→C (25.42%) transitions (Fig. 1*A* and Table S1). All mutations showed robust associations with paternal age, which we used as a proxy for the total parental age effect due to the previously described confounding between paternal and maternal age (Fig. 1*A*, red stars). While DNM-type composition is similar across the genome (*SI Appendix*, Fig. S3), regions with significant deviations from the mean may serve as good candidates for the identification and investigation of atypical mutational processes (chromosomes [chrs] 2, 9, 16, and 19) (*SI Appendix*, Fig. S3*B*). The influence of base context on mutational frequency can be better appreciated when viewing each DNM as a 3-mer by considering the bases in the human reference genome immediately preceding and following each mutation (19, 21, 40, 41, 45) (Fig. 1*B*). For example, T→C mutations preceded by an A appear to be more common than might have been predicted based on their central base pair mutation-type alone. While CpG to TpG transitions already comprise four of the five most common 3-mer DNMs, their mutational potential is particularly highlighted by normalizing each 3-mer DNM count for background 3-mer frequency, which demonstrates an excess of CpG to TpG transitions compared to the expectation based on genome frequency (*SI Appendix*,

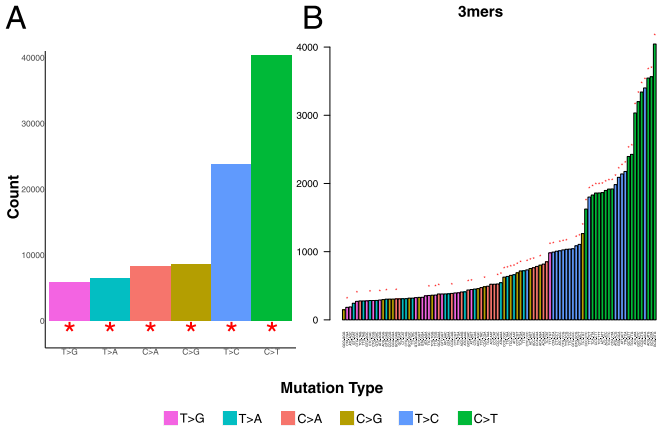

Fig. 1. Distribution of single-base and 3-mer mutation types across SNV DNM call set. (*A*) The distribution of single-base mutation type counts across our SNV DNM call set is shown. Colors represent mutation type, and stars represent associations with paternal age (red, $P < 0.05$ after Bonferroni correction). (*B*) The counts across our DNM call set for each of 96 3-mer mutation types is shown. Colors represent the center base mutation, and are the same as those in *A*. Stars represent associations with paternal age (red, $P < 0.05$ after Bonferroni correction).

Fig. S4). This postnormalization-inflated mutational pattern can also be seen for other CpG mutations (*SI Appendix*, Fig. S4, CpG to GpG in gold, and CpG to ApG in salmon).

To evaluate the relationship between mutation and recombination, we used our DNM call set to estimate local mutation rate across 1-Mb windows, and then tested the correlation between these rates and local recombination rate estimates derived from the recently published deCODE genetic map (32). We found a significant positive correlation in which regions of higher recombination have more DNMs (*SI Appendix*, Fig. S5*A*) ($R^2 = 0.057$, $P = 3.4 \times 10^{-36}$) (see *SI Appendix* for additional details). This relationship varies across the autosomes, with chrs 1, 7, 8, 9, and 16, 18, 19 showing strong positive correlations in which recombination explains as much as 39% of the variation in DNM rates (*SI Appendix*, Fig. S5*B*). Conversely, this correlation is either absent or limited across a number of the other autosomes.

Evolutionary theory predicts that mutation, natural selection, and genetic drift act in concert to shape genomic variation, and that recombination can shape variation by breaking up linked variants and enabling selection to act on them more efficiently (46, 47). Therefore, we tested the degree to which local recombination and mutation rate estimates explain the distribution of genomic variation seen at the human population level. To do this, we used genomic variation data from the TOPMed Consortium that was ascertained by performing WGS on ∼41K diverse unrelated individuals (43). We first calculated the total number of rare variants (AF < 0.005) within 1-Mb windows across the genome, and used standardized *z*-scores derived from these counts as a measure of localized levels of recent human genomic variation. Local recombination and mutation rates explain up to 37.85% (95% CI: 33.13 to 42.58) of the local genomic variation in segregating rare variants ($P < 7.25 \times 10^{-277}$) (Fig. 2). As expected, DNM rate explains the majority of this signal, accounting for 30.52% (95% CI: 26.08 to 34.92) of the variation in rare variants after regressing out the effects of local recombination rate ($P < 9.0 \times 10^{-214}$). Regions with the highest levels of rare variation, such as megabases 1 to 7 on chr8 and 1 to 9 and 78 to 90 on chr16, have high DNM rates, and are largely comprised of regions in the top 10% of recombination values (Fig. 2). When including both GC content and replication timing in the model as covariates, we can explain up to 41.28% (95% CI: 37.13 to 45.16) of the variation in rare variant totals. Even after adjusting for recombination rate, GC content, and replication timing, which is conservative given the interconnected relationships between these variables, DNMs still explain 27.95% (95% CI: 23.90 to 32.30) of the variation in rare variants. Similar models per chromosome can explain between 60% and 72% of the variation in rare variant totals across chromosomes that have segments with high local DNM rates (chrs 8, 9, 16, and 19) (Table S2), with DNM rate as the dominant explanatory variable. Interestingly, replication timing seems to best explain variation in rare variant levels across chrs 12 and 14.

In contrast to these rare variant models, when we ran similar models with standardized *z*-scores derived from the counts of common variants as our dependent variable, the most dominant explanatory variable is recombination rate. That is, when using recombination rate, GC content, replication timing, and DNM rate as covariates to predict the distribution of common variation across the genome, we explain a similar proportion of the variation in common variants as we did in rare variants ($R^2 = 0.3942$, 95% CI: 0.3486 to 0.4389), but the relative contributions of DNM rate and recombination rate seem to invert. In the former set of models of the genomic distribution of rare variation, recombination plus replication timing explain 11.8% (95% CI: 9.85 to 14.00) of variation, whereas DNM rate explains 31.56% (95% CI: 27.41 to 35.80) of the remaining variation. Conversely, in the models of the genomic distribution of common variation, recombination rate plus replication timing explain 30.03% (95% CI: 26.28 to 33.76) of variation, whereas DNM rate explains only the remaining 12.47% (95% CI: 8.98 to 16.01). This is consistent with evolutionary predictions, as common variants typically represent older variants, and their distribution has been shaped more by recombination, selection, or drift than by mutation. These patterns hold when using linear regression to implement an adjustment for coding proportion and mappability concerns, with each model (DNM rate, recombination rate, replication timing, and GC as covariates) explaining about 32% of the variation in common and rare variants, respectively. DNM rates still explain the majority of the variation in rare variant levels (17.77%, 95% CI: 13.28 to 22.57, after adjusting for recombination rate, replication timing, and GC content), and recombination rates still explain the largest portion of variation in common variant levels (14.32%, 95% CI: 11.51 to 17.45, after adjusting for DNM rate, replication timing, and GC content).

The Relationship between DNM Rates and Heterozygosity. In testing the relationship between heterozygosity and DNM rate across all individuals, we found a significant positive correlation ($P < 0.002$) (*SI Appendix*, Fig. S6*A*). However, while this persists after adjusting for parental age ($P < 0.025$), heterozygosity only explains ∼0.3% of the variation in DNM count. This relationship is also entirely driven by the Amish population, and no significant relationship persists when removing Amish individuals from the analysis that adjusts for parental age ($P > 0.13$). Furthermore, we found no relationship between heterozygosity and DNM count when looking intracohort (see *SI Appendix*, Fig. S6*B* for a representation of this across individuals from the FHS, our largest cohort). The number of kilobases in an individual's genome found within runs of homozygosity (ROH) also correlates significantly with DNM rate ($P < 2.0 \times 10^{-4}$) (*SI Appendix*, Fig. S7*A*) and persists after adjusting for parental age ($P < 3.1 \times 10^{-4}$). However, as was the case with heterozygosity, this is entirely driven by the Amish, who happen to be outliers for ROH. Similar to heterozygosity, additional analysis suggests that this association is confounded and not reflective of a causal relationship. ROH only explains ∼1% of the variation in DNMs, and its association with DNMs is no longer significant after filtering out the Amish ($P > 0.14$, $R^2 = 0.0017$) or when evaluating intracohort ($P \geq 0.112$).

DNM Rate Comparisons Across Ancestrally Diverse Cohorts. When comparing DNM rate across all six ancestrally diverse cohorts (Table 1), we found significant differences ($P < 1.6 \times 10^{-3}$, ANOVA) (Fig. 3

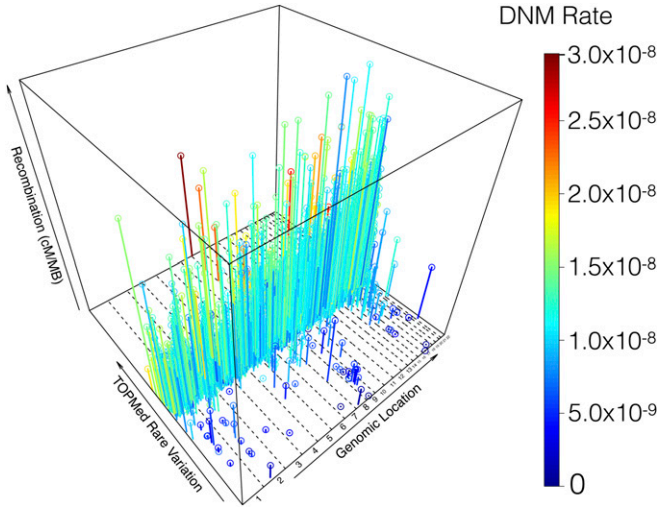

Fig. 2. City plot of rare variation, recombination rate, and DNM rate across the genome. The relationship between DNM rate (blue to red color range), rare variation (*y* axis, ranging from −5.83 to 9.06 *z*-scores), and recombination rate (*z* axis, ranging from 2.73×10^{-14} to 6.12 cM/Mb) across the genome (*x* axis, dotted vertical lines divide autosomes 1 to 22) is shown. In moving from low to high rare variation levels across the *y* axis, a blue to red gradient can be seen, which reflects the significant correlation between DNM rate and population-level rare variation. Furthermore, regions with high DNM rates and high variation levels generally have taller bars, which reflects the positive relationship between DNM rate, variation level, and recombination (a few exceptions to this can be seen as taller blueish bars). Regions with the highest variation levels in the genome, such as those on chromosomes 8 and 16, have the highest DNM rates.

and Table 1), even after accounting for parental age ($P < 0.01$, ANOVA). However, this appears to be driven by the Amish, who have significantly fewer DNMs per individual than the other cohorts ($P < 0.05$ between the Amish and the four largest populations [BAGS, CFS_EUR, CRA, FHS] after correcting for multiple testing). A reduced DNM rate persists in the Amish even when comparing DNM rates across population within regions of the genome estimated not to be in ROH in either parent's genome ($P < 0.033$, ANOVA) (*SI Appendix*, Fig. S7*B*). Therefore, while recent work suggests heterozygosity as a mutational driver, and increased ROH in the Amish is an appealing explanation for their reduced mutation rate, on the whole, differences in heterozygosity or ROH do not seem to be sufficient to explain the reduced mutation rate in the Amish.

To control for sequencing center differences that could potentially influence the number of DNMs identified, we performed a subanalysis comparing DNMs between cohorts that were sequenced at the same center. This subanalysis did not include individuals from the BAGS cohort, since they alone were sequenced directly by Illumina. This within-sequence center analysis also revealed significantly lower DNM rates in the Amish compared with individuals from FHS (European ancestry; $P < 0.005$, ANOVA), even after adjusting for parental age effects ($P < 0.002$, ANOVA), whereas we found no significant differences in DNM rates before and after adjusting for parental age effects between individuals from CFS_EUR (European ancestry), individuals from CFS_AFR (African ancestry), and individuals from CRA (Latino ancestry; $P > 0.56$, ANOVA). While the BAGS cohort appears to have an elevated mutation rate at first glance, this is a result of them having higher average paternal ages (Table 1). While we are underpowered to see smaller differences in mutation rate, we have reasonable power to see moderate to large differences between all populations other than CFS_AFR (*SI Appendix*, Fig. S8), so these results do suggest that differences between ancestry thought to influence the DNM rate (such as heterozygosity, demographic history, and so forth) are not driving large differences in the accumulation of DNMs.

The amount of variation in DNM rate attributable to parental age effects after adjustment for Poisson variation is lower in our dataset (57.7%) than previously reported (>80%) (8). This seems to be due to the fact that there is heterogeneity in the parental age effects across our sample set (Table 1), which itself becomes a contributor to the variation in DNM rate. In accordance with this, and consistent with the reduced DNM rate we found in the Amish, we estimate a lower parental age effect in the Amish than in the other cohorts (Table 1). However, we have limited power to detect this difference, which only reaches significance when comparing effect sizes determined by Poisson regression between the Amish and FHS ($P < 0.035$).

Evaluation of Batch Effects and Technical Artifacts. To evaluate the robustness of the observation of a DNM rate reduction in the Amish to technical artifacts, we took a multipronged approach. First, we validated the consistency of our DNM call set by using two offspring samples from the FHS cohort that had undergone repeat sequencing at another sequence center (University of Washington [UW]), as well as one offspring sample from the Amish cohort that has a monozygotic twin in our dataset. For these three samples, we evaluated the proportion of DNMs that are called as heterozygous sites in the validation sample, as well as the concordance between DNMs called in the initial and validation samples. In two of the three samples, 100% of DNMs (one Amish, one FHS) were called as heterozygous sites in the validation sample. In the third sample (FHS), 5 of 50 DNMs were called as homozygous reference in the validation sample. However, upon further inspection, all five of these DNMs had read counts and genotype likelihoods suggesting that they were heterozygous sites. With regard to DNM concordance, 95.3% of DNMs were called in both the initial and

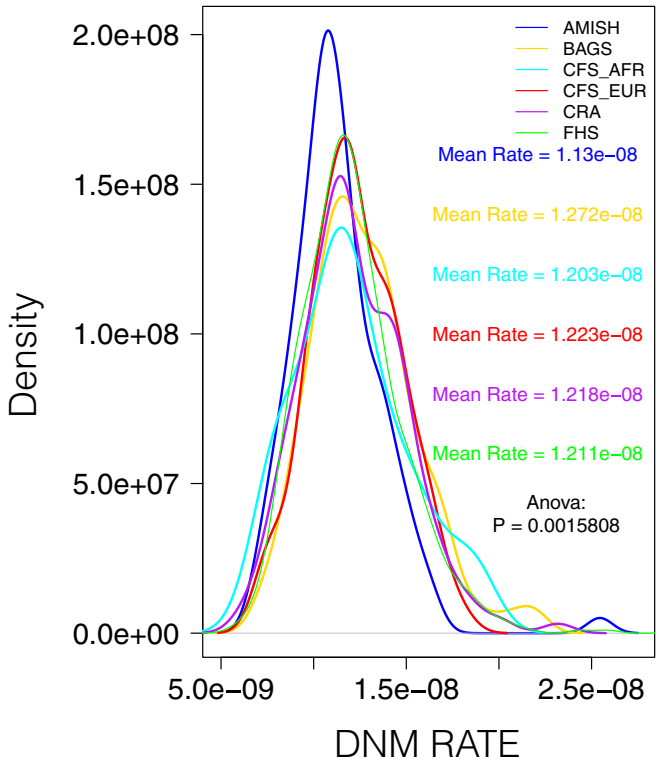

Fig. 3. DNM rates across diverse cohorts. DNM rates per individual show significant differences across cohort, which are driven by a reduction in the Amish.

validation samples. Notably, this number is significantly higher (98.2%) when excluding the five DNMs noted above as being improperly called as homozygous reference by the variant caller despite strong evidence of heterozygosity, which is unlikely to be a common issue.

We also assessed how a number of quality control measures vary across population. Specifically, we compared measures of average genome-wide sequence depth (*SI Appendix*, Fig. S9*A*), average depth of variants called autosome-wide and passing filters (*SI Appendix*, Fig. S9*B*), number of bases genome-wide with a minimum depth ≥ 10 (*SI Appendix*, Fig. S9*C*), genome-wide Ts/Tv ratio (*SI Appendix*, Fig. S9*D*), number of million reads genome-wide per sample (*SI Appendix*, Fig. S9*E*), percent of the genome covered (*SI Appendix*, Fig. S9*F*), percent of the genome with a depth of 10 (*SI Appendix*, Fig. S9*G*), average depth of genome-wide bases with Q ≥ 30 (*SI Appendix*, Fig. S9*H*), percent of the genome with Q ≥ 20 (*SI Appendix*, Fig. S9*I*), and the number of autosomal bases with depth ≥ 10 and Q ≥ 30 (*SI Appendix*, Fig. S9*J*). On the whole, our study cohorts feature similar quality metrics, and the Amish are not a notable outlier in any metric (*SI Appendix*, Fig. S9). While the BAGS cohort seems to be a moderate outlier in Ts/Tv and number of reads (*SI Appendix*, Fig. S9 *D* and *E*), and a more pronounced outlier with regard to percent genome covered and number of bases with Q ≥ 30 (*SI Appendix*, Fig. S9 *F–H*), adjusting for these (as well as all of the other) quality control measures did not significantly influence our signal. Similarly, when we used the number of bases per sample with depth ≥ 10 + Q ≥ 30 to normalize our DNM counts into DNM rate estimates, as described earlier, we actually saw an increased signal of a reduced DNM rate in the Amish.

DNM Rates by Ancestral Proportions of Individuals from Admixed Populations. To test more directly how interancestral differences might influence DNM accumulation, we used linear regression to assess the relationships between DNM rate and African ancestry proportion, and DNM rate and Native American ancestry proportion. Within the BAGS, CFS_AFR, and CRA individuals, which represent the three cohorts with significant African ancestry, we found no significant correlation between African ancestry proportion and DNM rate after adjusting for parental age ($P > 0.36$, ANOVA) (*SI Appendix*, Fig. S10*A*). This absence of a relationship persists when looking only within BAGS individuals ($P > 0.72$) or only within CRA individuals ($P > 0.19$), and there was also no relationship when comparing Native American ancestry proportion and DNM rate before ($P > 0.76$) or after ($P > 0.088$) adjusting for parental age within Latino ancestry individuals from the CRA cohort (*SI Appendix*, Fig. S10*B*).

Since we are also underpowered to see small-to-moderate differences in single-base or 3-mer mutation frequencies between populations, we instead tested the correlation between 3-mer mutation type and European ancestry proportion within all individuals of admixed African ancestry (BAGS, CRA, CFS_AFR cohorts). This should increase our power to detect mutational associations with ancestral background, and specifically allow us to test for a signal of ancestral association for the TCC→TTC 3-mer mutation that was previously shown to positively correlate with European ancestry (40, 41). Notably, we did not find a significant association between TCC→TTC 3-mer mutation and European ancestry (P ≥ 0.60). Interestingly, when testing this association in individuals from BAGS, which is our cohort with the highest average proportion of African ancestry, we found a negative correlation ($\beta = -4.08$, 95% CI: −6.70 to −1.70, $P < 0.0014$, uncorrected, Poisson regression) (*SI Appendix*, Fig. S11). Overall, this result is interesting when contrasted with recent evidence that an ancestry-specific pulse increased the occurrence of these 3-mer mutations in Europeans (41).

Shifts in the Mutational Spectrum of the Amish Founder Population. To further assess the reduced DNM rates seen in the Amish, we compared DNM counts for each single-base mutation type across cohort. ANOVA revealed differences in mutation counts across cohort, with differences in C→A ($P < 0.03$), T→C ($P < 8 \times 10^{-4}$), and T→G ($P < 3.2 \times 10^{-8}$) mutations being most significant (*SI Appendix*, Fig. S12). These differences are largely driven by decreases in these mutations in the Amish and increases in these mutations in individuals from the BAGS cohort. To further evaluate these differences while controlling for sequencing center, we compared mutations between the Amish and FHS cohorts, and found the reduced number of C→A ($P < 1 \times 10^{-2}$) and T→C ($P < 3 \times 10^{-4}$) mutations in the Amish to persist (*SI Appendix*, Fig. S13). Interestingly, the reduction in T→C mutations explains about 45% of the overall DNM reduction seen in the Amish.

Local Ancestry Analysis Does Not Identify Ancestrally Distinct Mutation Rates. To look more closely at whether mutation rates differ between ancestries, we compared the rates of DNM accumulation across ancestrally distinct genomic segments within admixed samples. Using local ancestry assignments, we counted DNMs across all possible diploid ancestral segment combinations (e.g., homozygous African, heterozygous African and European, and so forth) (*SI Appendix*, Fig. S14), and limited each comparison to individuals with at least 80 Mb of each diploid category. The resulting intraindividual comparisons allow for the control of unmeasured variables that may otherwise confound interindividual analyses, such as environmental exposures that might influence mutation rate. Therefore, the only variable that should differ between diploid segments is their ancestral origin, which should allow us to isolate and test for any effects of this ancestral background on local sequence context (i.e., this is a test for *cis* effects). After estimating DNM rates per individual by normalizing DNM counts by diploid category base total (i.e., rates per base pair), we did not find evidence of ancestry specific differences in DNM rate (*SI Appendix*, Fig. S15). While we did find a significant reduction in the mutation rate in African/European segments compared with African/African segments ($P = 7.33 \times 10^{-4}$), other comparisons between African and European segments showed no differences. Given the absence of consistent differences in DNM rate across local ancestral segments, these results do not provide compelling support for the existence of ancestry-based differences in DNM rate.

The Heritability of DNM Accumulation. To assess whether we could use GWAS to detect any genetic loci that might underpin the interindividual variation we see in DNM accumulation, we first wanted to test for what proportion of the variation in DNM accumulation was explained by genetics. One measure of this is narrow-sense heritability (h^2), which represents the proportion of phenotypic variation explained by additive genetic effects (48), and which can be used as a null model for running associations at every locus (i.e., each locus is tested for the proportion of h^2 that it explains). Treating the number of DNMs per individual as a phenotype reflecting DNM accumulation, we used the MMAP software (49) to estimate the h^2 of DNM accumulation via a restricted maximum likelihood-based (REML) method (*SI Appendix*). When using all samples across all cohorts, we used paternal age at offspring's conception, maternal age at offspring's conception, and cohort label as fixed effects, and a genetic relatedness matrix estimated from the sequence data. We estimate an $h^2 = 0.0$ (SE = 0.028), and we reaffirmed a paternal age effect ($P < 8.51 \times 10^{-55}$), a maternal age effect ($P < 6.42 \times 10^{-7}$), and an effect of Amish cohort status ($P < 1.13 \times 10^{-3}$) (Table 2 and *SI Appendix*, Table S3). When running this base model on each cohort separately, h^2 estimates are zero for each cohort other than CRA ($h^2 = 0.42$, $P = 0.122$) (Table S4). However, the h^2 result for CRA was 0.0 when further restricting to only unrelated subjects (Table S5).

To evaluate whether the portion of mutational variation explained by parental age might be shaped by additive genetic factors, we repeated these heritability models without including

Table 2. Heritability model across all cohorts

Variable	Value	Variable	Value
Mean DNMs	64.39	$\beta_{paternal\_age}$ *P* value	8.51E-55
SD	14.94	$\beta_{maternal\_age}$	0.45
Minimum	14.00	$\beta_{maternal\_age}$ SE	0.09
Maximum	164.00	$\beta_{maternal\_age}$ *P* value	6.42E-07
Kurtosis	3.64	β_{Amish}	−4.06
DNMs > 3 SD	12.00	β_{Amish} SE	1.24
DNMs < 3 SD	1.00	β_{Amish} *P* value	1.13E-03
Sample size	1,389	β_{BAGS}	0.78
h^2	0.00	β_{BAGS} SE	1.13
h^2 *P* value	0.00	β_{BAGS} *P* value	0.49
h^2 *P* value SE	0.03	β_{CFS_AFR}	1.63
Proportion variance explained by covariates	0.45	β_{CFS_AFR} SE	2.33
Adjusted proportion variance explained by vovariates	0.45	β_{CFS_AFR} *P* value	0.48
ln likelihood	−4,067.84	β_{CRA}	−0.39
Intercept	13.36	β_{CRA} SE	1.33
Intercept SE	1.79	β_{CRA} *P* value	0.77
Intercept *P* value	1.27E-13	β_{FHS}	−0.32
$\beta_{paternal\_age}$	1.32	β_{FHS} SE	0.78
$\beta_{paternal\_age}$ SE	0.08	β_{FHS} *P* value	0.68

Results are shown from heritability models run with MMAP across all samples with paternal and maternal ages available (n = 1,389). Heritability is estimated as zero (h^2 = 0.00), with an SE of 0.03). These models confirm that paternal age at offspring's conception ($P = 8.51 \times 10^{-55}$), maternal age at offspring's conception ($P = 6.42 \times 10^{-7}$), and Amish cohort status ($P = 1.13 \times 10^{-3}$) are significantly correlated with DNM total per individual.

parental ages as fixed effects. Interestingly, these models estimated nonzero heritability ($h^2 = 0.137$, $P < 3.93 \times 10^{-2}$), with the BAGS and CFS_EUR cohorts contributing most significantly to this nonzero heritability estimate (Tables S3 and S4). However, when repeating analyses with only one offspring randomly chosen per nuclear family (i.e., removing sibling data) (Table S6), or with only unrelated samples (Table S5), h^2 is zero in nearly all models. Exceptions include the Amish and CFS_EUR per cohort analyses, but these have notably small sample sizes as well as nonsignificant ($P > 0.129$) h^2 estimates with wide SEs. Overall, this suggests that the significant nonzero heritability we see when not adjusting for parental age effects is driven by the shared parental contribution of siblings and or confounding due to relatedness, and not by genotypic similarity. While statistical power is a concern as a result of our sample sizes, we do have reasonable power (≥0.6) to detect moderate h^2 across all samples as well as across larger cohorts and cohort combinations (*SI Appendix*, Fig. S16). Simulations across the known pedigree of nearly 6,000 Amish individuals also suggest that we would only expect to estimate zero heritability across our samples if the true heritability was ≤0.18 (empirical $P <$ 0.05) (*SI Appendix*, Fig. S17).

Discussion

Here we call DNMs across 1,465 individuals from diverse cohorts sequenced through the TOPMed program. Our call set is determined by a filtering heuristic that is similar to previous approaches (19), its specificity is supported by a Bayesian approach implemented in the TrioDeNovo software (50) that called 99.39% of our DNMs, and its overall accuracy is supported by validation sequencing across two repeated samples and one pair of monozygotic twins. In addition to this, quality assessments done as part of the TOPMed Consortium sequence analysis efforts used repeatedly sequenced samples to demonstrate similar genotypic concordance rates within and between sequence centers (43). While this was done for both variants passing calling filters and variants failing calling filters, variants passing calling filters have even larger levels of concordance than those failing calling filters, which reflects the efficacy of these calling filters in identifying errors. This suggests that even the differences in some quality measures mentioned above, such as the average depth of Q30 bases in BAGS, are already effectively handled by our variant calling quality control. This is directly in accordance with the facts that controlling for these covariates doesn't qualitatively change our signal, and that using DP10Q30 metrics to estimate normalized per sample mutation rates in fact increases our signal of a reduced mutation rate in the Amish. Other TOPMed Consortium results using principal components analysis to evaluate the influence of sequence center on genotypic variance further support a very limited influence of sequencing center (43).

Despite the predominance of C→T mutations among DNMs, we did not find any differences between populations or ancestral backgrounds in C→T mutations. This is notable given the high frequency of this mutational class, and the concomitant increased power to detect differences. This suggests that processes driving cytosine deamination are fairly conserved, and is consistent with the assertion by others that CpG mutation rates may serve as a better "molecular clock" than the base substitution rates that have previously been used (17, 40, 45, 51). Similarly, we did not find differences in other single-base and 3-mer mutation types across ancestral background. While we did have somewhat limited statistical power to see smaller effect sizes (*SI Appendix*, Fig. S8), we did have reasonable power to see moderate-to-large differences when comparing most populations, and potentially even had good power to see small-to-moderate differences when comparing our larger populations. For example, when comparing DNM rates between the BAGS and FHS cohorts (our largest African ancestry and European ancestry cohorts, respectively), we had 80% power to see an effect size difference reflected by a Cohen's d of 0.3. This represents an effect size small enough so that only 61.79% of the DNM rates in the population with the larger DNM rate would be larger than the mean in the other population (Cohen's U_3), there would only be a 58.4% chance that a random individual from the population with the larger DNM rate has a higher DNM rate than a random person from the other population (probability of superiority), and there would be an 88.08% overlap between the DNM rate distributions of the two populations (52). This is generally considered a small-to-moderate effect size, and helps to appreciate what kind of DNM rate differences we have power to see. However, we likely only have this level of power to detect differences in the overall SNV DNM rate (and possibly certain

single-base mutation rates), as we are notably more underpowered to detect differences in 3-mer rates due to their reduced count sizes. Nonetheless, we did not see significant mutation rate differences between populations with distinct ancestral backgrounds, which suggests that overall and single-base mutation rate differences are not likely to be large.

When leveraging quantitative estimates of genetic ancestry within a framework with increased power to see potential differences, we also did not find significant associations between European ancestry proportion and 3-mer mutation type across individuals from admixed cohorts (i.e., BAGS, CFS_AFR, CRA). This result is interesting when contrasted with recent findings of Europeans having had a significantly higher rate of TCC→TTC mutations in the past (40, 41). First, the recent identification of a significant batch effect in the 1000 Genomes data used to identify these mutational differences raises some questions about their legitimacy (53). Nevertheless, assuming TCC→TTC mutations did in fact increase in Europeans at some point in the past, our results suggest that there is currently no difference, which is consistent with findings that the increase in Europeans happened as a pulse that ended around 2,000 y ago (41). It is also possible that differences in the rate of this mutation across populations has been influenced by differential shifts in mean parental age at conception, which is consistent with recent findings by Jónsson et al. (19) and Agarwal and Przeworski (54) that this mutation is enriched for sex differences. Ultimately, additional research is needed to better understand the evolutionary history of TCC→TTC and other mutations across divergent human populations.

The correlations we found between local recombination and DNM rates are concordant with research showing that processes underpinning recombination are mutagenic (26, 27, 32), and our models comparing population-level variation with recombination rates, DNM rates, replication timing, and GC content add further resolution to the relationship between these intertwined covariates and genomic evolution. Areas of the genome with the highest local DNM rate are in regions with high recombination that have recently been identified as featuring specific recombination-based processes that drive mutation (24). Alternatively, regions with high recombination rate estimates that have low DNM rates (Fig. 2, tall blue bars) are good candidates for the identification of genomic features that may explain a reduced mutation rate, including contexts in which recombination itself potentially drives less mutagenesis. While recent research into the relationship between DNMs and fine-scale local recombination rates identified up to 50-fold increases in DNM rates around meiotic crossover events (32), the fact that these events are rare is likely why we only explain about 5% of the variation in DNMs on the basis of recombination. This recombination study also found 35 loci influencing recombination rate, which may lead one to ask why we did not find any signal of genetic influence over DNM accumulation (i.e., we estimate $h^2 = 0$), given that recombination is mutagenic and associates with DNMs in our dataset. First, this is likely due to the fact that recombination only explains a small portion of the overall variation in DNMs, and small effect-size modulators of recombination are unlikely to be detected as correlates of DNM rate without very large sample sizes. Second, we adjusted for parental age effects, which might eliminate from many of our heritability models the portion of variation in DNM rate influenced by meiotic recombination. Furthermore, multivariate models have been able to explain >50% of the variation in regional mutation rates on the basis of genomic features, such as GC content, exon density, sex chromosome status, recombination, and distance to telomeres (55). Given that we excluded most centromeric/telomeric segments from our analyses, didn't consider exon density or sex chromosome status, and used metadata from multiple studies across multiple populations, we expected our model to explain a lower proportion of the variation in mutation rate than these models. Nonetheless, these and other covariates potentially explain the existence of high recombination regions with low DNM rates.

Via these models, we also estimated the quantitative contributions of mutation and recombination to population-level variation, and were able to explain up to 38% of the variation across the genome in population-level rare variation using local recombination and DNM rates, and 41% when including replication timing as an additional covariate in the model. The remaining proportion of variation is likely the result of a combination of genetic drift, selection, or measurement error. Of this explained variation, ∼30.5% is attributable to variation in local DNM rates. Regions with high mutation rates have the most variation in the genome (e.g., chr8, chr16), and often feature high recombination rates that likely increase mutation accumulation. Numerous recent studies have found complementary results that describe the types of mutations predominating in high-variation regions, and identify the likely sources of these mutations (21, 23, 24). Some of this recent work demonstrated that clustered DNMs contribute very significantly to the clustering of genomic SNPs (specifically C→G SNPs), although clustered DNMs are estimated to make up less than 2.5% of DNMs (24) and may only have limited impact on the findings presented here. Here, we show that single-nucleotide DNMs contribute profoundly to the entire genomic distribution of common and rare single-nucleotide variation, but that recombination is a larger driver of the variation in common variant levels than is mutation. This is consistent with evolutionary theory about the ability of recombination to disconnect linked variants and enable selection to act more efficiently, as well as the expectation that common variants are older and have been subjected to nonmutational evolutionary forces for longer.

While the Heterozygosity Instability hypothesis (33, 34, 56, 57) and recent related findings (35) predict that increasingly heterozygous genomes will have higher mutation rates, we only found evidence of a modest relationship between heterozygosity and DNM rate. Furthermore, we did not find the differences in DNM rate between ancestral background that differences in heterozygosity across ancestry would have predicted, nor a significant correlation between heterozygosity and DNM rate between samples from the same population. While runs of homozygosity do seem to shape DNM rates more so than levels of heterozygosity, this relationship is entirely driven by the Amish, and additional analysis of mutation rate across non-ROH–bearing regions of the genome suggests that this is likely due to confounding between the high ROH in the Amish and whatever else is reducing their mutation rate. Therefore, while this finding of a potential relationship between ROH and DNM rate does raise the possibility that the absence of heterozygosity may drive significant reductions in DNM rate, additional research is needed to more directly address this question.

In our heritability analyses, we estimate the heritability of DNM rates to be zero across nearly all models. Due to sample size concerns, we conducted power analyses and used simulations to estimate the likelihood of estimating zero heritability across increasing values of true heritability. For the full dataset, most power calculation approaches (*SI Appendix*, Fig. S16 *A–D*) suggest we have good power to detect low to moderate levels of heritability [i.e., ($h^2 \geq 0.1$)], although we likely have very little power to detect heritability across the smaller cohorts. Similarly, given sample sizes similar to that of our full dataset, a trait with an $h^2 \geq 0.13$ will be estimated as zero ≤20% of the time. Even in circumstances with limited power, we would still expect h^2 estimates centered off of zero when working with traits that have true heritability greater than zero. Therefore, given our simulation results, and that our h^2 estimates are consistently centered around 0, we find it more likely that the true heritability of DNM rates across our dataset are low or approaching zero. While future initiatives might still consider large-scale GWAS efforts in order to search for local mutation events under genetic control, and larger studies with increased power are needed to confirm the heritability of DNM rates, we did not find evidence that genetic similarity explains the variation in mutation rates we see within or between populations.

Despite being a founder population that diverged from other Europeans only very recently (58), the Amish show a mutation rate reduction of about 7%. This reduction persists when controlling

for parental age effects and sequence quality metrics, and seems to be driven by reductions in C→A and T→C mutations. Together with our estimation that DNM rate has zero narrow-sense heritability, this suggests that the environment may play a bigger role in modulating the mutation rate than previously appreciated. The Amish lifestyle features preindustrial era aspects, and while modern Amish communities are diverse and have adapted to the usage of some modern items, they continue to limit the influence of technology in their daily lives (59, 60). Given this, it is possible that the Amish are exposed to fewer mutagens, and that this "clean living" may be partially responsible for the reduced mutation rate we report here. For example, studies have shown that rural areas, such as those similar to the areas occupied by the Amish, have fewer carcinogens and mutagens than industrialized areas (61–63). Recent analysis of mutation patterns has also called into question the canonical view that DNMs rise predominantly from replicative errors, and suggests that exogenous mutagens may play a larger role in mutation accumulation than previously appreciated (25). If the Amish do in fact experience less environmentally driven mutagenesis, then one would predict a significant reduction in the rate of cancer in the Amish. This is exactly what has been found in multiple Old Order Amish populations, with a particularly large reduction in cancer rates found in men (64, 65). A similar reduction in overall mortality has been found in Amish men compared with FHS men, which is hypothesized to be due to lifestyle factors, such as reduced tobacco use and increased physical activity (59). Given that DNM mutation in sperm is the single largest driver of DNM accumulation, an Amish environment that potentially limits DNA damage in Amish men is consistent with a lower DNM rate. In accordance with this, the Amish have the lowest estimated parental age effect (Table 1). This is also consistent with the recent finding of significant variability in parental age effects across ancestrally similar families, which also suggests the possibility that environmental factors influence DNM rates (66).

It is important to note that considerations of the environment as a potential explanation for the reduced DNM rate we detect in the Amish are entirely speculative, and that batch effects or other technical artifacts remain possible despite our significant efforts to control for them. Nonetheless, our findings as well as the aforementioned context do suggest environmental influence as a possible explanation, and one that is worthy of additional consideration and follow-up. The fact that the only signal of DNM heritability we detect is driven by siblings when not accounting for parental age effects further suggests that the environmental similarity shared by siblings (including parental age effects) is significantly influencing DNM rates. In sum, the mutational differences we found in the Amish stand in contrast to the relative homogeneity seen across the other diverse human populations we analyzed, and suggest that additional work is needed to better appreciate the forces shaping human mutational processes at fine scales.

Methods

WGS was performed using samples previously collected and consented across 90 NHLBI-funded research projects as part of the TOPMed program (43). Using variant data from a jointly called variant call set for samples from five of these TOPMed research projects, we implemented a filtering heuristic to call 93,325 DNMs across 1,449 samples after the removal of outliers and pedigree errors. We then used two resequenced samples and one monozygotic twin pair to validate our call set by evaluating the proportion of DNMs in one sample that were called as heterozygous sites in the second sample, and the percent of DNMs that were called by our filtering approach in both the initial and validation samples. After demonstrating high concordance between our call set and DNMs called with the TrioDeNovo Bayesian mutation caller, we compared single-nucleotide DNM counts and rates using ANOVA, *t* tests, and linear and Poisson regression. DNM rates were estimated as the number of single-nucleotide DNMs autosome-wide divided by the number of autosomal bases with depth $\geq$ 10 and quality $\geq$ 30. To test whether more heterozygous genomes have a higher DNM rate, we calculated genome-wide heterozygosity scores for each of the 1,449 samples included in our analysis, and compared these with estimated DNM rates per individual using linear regression. For comparison within only a particular ancestry, samples were subset down accordingly. For genome-wide analyses, we used the University of California, Santa Cruz's liftOver tool (67) to lift our DNM call set over to hg38 coordinates, and calculated local DNM rates, local recombination rates (32), local replication timing rates (68), GC content, and rare variation levels (43) for 1-Mb windows across the autosomes. To call local ancestry across our samples, we first phased the data using the Eagle2 algorithm as implemented in the Eagle software (69), combined WGS data from the 1000 Genomes Project (70) with high coverage WGS data from the Peruvian Genome Project (71), and used this genotype data as reference input to the RFMIX software (72). Heritability was estimated using MMAP (49) and GCTA (73) with DNM count per individual as the quantitative mutation phenotype. Additional methodological details are described in *SI Appendix*. Data can be accessed via dbGaP (https://www.ncbi.nlm.nih.gov/gap/) using the "phs" accession numbers listed for each study cohort in the Acknowledgments.

ACKNOWLEDGMENTS. Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (phs000974) and "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish" (phs000956) were performed at the Broad Institute of MIT and Harvard (3R01HL092577-06S1, 3R01HL121007-01S1). WGS for "NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados" (phs001143) was performed by Illumina, Inc. (3R01HL104608-04S1). WGS for "NHLBI TOPMed: The Cleveland Family Study (WGS)" (phs000954) and "NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica" (phs000988) were performed at the University of Washington Northwest Genomics Center (3R01HL098433-05S1, 3R37HL066289-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity quality control, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. A full list of TOPMed collaborators can be found at https://www.nhlbiwgs.org/topmed-banner-authorship. M.D.K. was supported by NIH grant T32CA154274. M.D.K. and T.D.O. were supported by funding from the Center for Health Related Informatics and Bioimaging at the University of Maryland School of Medicine, institutional support for the Institute for Genome Sciences and Program in Personalized Genomic Medicine at the University of Maryland School of Medicine, NIH Genomic Commons Award OT3 OD025459-01, NHLBI Trans-Omics for Precision Medicine Program High-performance Grant U01 HL137181-01, and National Human Genome Research Institute Genomic Innovator Grant 1 R35 HG010692-01 (to T.D.O.). D.P.L. was supported by NIH T32HL007698. This work was further supported by grants to the Amish research program (R01 HL121007, R01 AG18728, and U01 HL072515); a grant for the study of Asthma in Costa Rica (1P01HL132825-01 to S.T.W.); grants to study sleep apnea (R01-HL113338 to S.S.R., R35-HL135818 from Sleep Research Society Foundation to S.S.R. and with support for H.W., and K01-HL135405 from American Thoracic Society Foundation to B.E.C.); and Framingham Heart Study Grant HHSN268201500001 (to R.S.V.). Sequencing was funded by grants to the Genetics of Cardiometabolic Health in the Amish study (3R01HL121007-01S1), The Genetics and Epidemiology of Asthma in Barbados study (3R01HL104608-04S1), The Cleveland Family Study (3R01HL098433-05S1), The Genetic Epidemiology of Asthma in Costa Rica study (3R37HL066289-13S1), and Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study (3R01HL092577-06S1).

1. F. Tajima, The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**, 1457–1465 (1996).
2. D. E. Reich *et al.*, Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**, 135–142 (2002).
3. H. Ellegren, N. G. Smith, M. T. Webster, Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13**, 562–568 (2003).
4. A. S. Kondrashov, Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
5. I. Bray, D. Gunnell, G. Davey Smith, Advanced paternal age: How old is too old? *J. Epidemiol. Community Health* **60**, 851–853 (2006).
6. M. Lynch, Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
7. M. Lynch, Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 961–968 (2010).
8. A. Kong *et al.*, Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
9. L. B. Alexandrov *et al.*; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). Correction in: *Nature* **502**, 258 (2013).
10. P. F. Palamara *et al.*; Genome of the Netherlands Consortium, Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet.* **97**, 775–789 (2015).

11. Y. Field *et al.*, Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
12. J.-B. Rivière *et al.*; Finding of Rare Disease Genes (FORGE) Canada Consortium, De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat. Genet.* **44**, 934–940 (2012).
13. J. A. Veltman, H. G. Brunner, De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
14. J. Sebat *et al.*, Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
15. K. E. Samocha *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
16. J. B. Haldane, The rate of spontaneous mutation of a human gene. 1935. *J. Genet.* **83**, 235–244 (2004).
17. L. Ségurel, M. J. Wyman, M. Przeworski, Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
18. J. C. Roach *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
19. H. Jónsson *et al.*, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
20. D. F. Conrad *et al.*; 1000 Genomes Project, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
21. J. M. Goldmann *et al.*, Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).
22. W. S. Wong *et al.*, New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
23. J. Carlson *et al.*; BRIDGES Consortium, Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **9**, 3753 (2018).
24. J. M. Goldmann *et al.*, Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
25. Z. Gao *et al.*, Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9491–9500 (2019).
26. M. J. Lercher, L. D. Hurst, Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340 (2002).
27. B. Arbeithuber, A. J. Betancourt, T. Ebner, I. Tiemann-Boege, Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2109–2114 (2015).
28. I. Hellmann, I. Ebersberger, S. E. Ptak, S. Pääbo, M. Przeworski, A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
29. D. A. Kiktev, Z. Sheng, K. S. Lobachev, T. D. Petes, GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7109–E7118 (2018).
30. A. Koren *et al.*, Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
31. C. Li, N. M. Luscombe, Nucleosome positioning stability is a significant modulator of germline mutation rate variation across the human genome. bioRxiv:10.1101/494914 (15 May 2019).
32. B.V. Halldorsson *et al.*, Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
33. W. Amos, Heterozygosity and mutation rate: Evidence for an interaction and its implications: The potential for meiotic gene conversions to influence both mutation rate and distribution. *BioEssays* **32**, 82–90 (2010).
34. W. Amos, Variation in heterozygosity predicts variation in human substitution rates between populations, individuals and genomic regions. *PLoS One* **8**, e63048 (2013).
35. S. Yang *et al.*, Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463–467 (2015).
36. A. Scally, R. Durbin, Revising the human mutation rate: Implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
37. C. D. Campbell *et al.*, Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
38. V. M. Narasimhan *et al.*, Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).
39. H. Jónsson *et al.*, Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
40. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3439–3444 (2015).
41. K. Harris, J. K. Pritchard, Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).
42. S. Besenbacher *et al.*, Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).
43. D. Taliun *et al.*, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *bioRxiv*:10.1101/563866 (6 March 2019).
44. W. McLaren *et al.*, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
45. D. G. Hwang, P. Green, Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13994–14001 (2004).
46. E. J. B. Williams, L. D. Hurst, The proteins of linked genes evolve at similar rates. *Nature* **407**, 900–903 (2000).
47. W. G. Hill, A. Robertson, The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
48. N. R. Wray, P. M. Visscher, Estimating trait heritability. *Nat. Educ.* **1**, 29 (2008).
49. C. Sun, P. M. VanRaden, J. B. Cole, J. R. O'Connell, Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One* **9**, e103934 (2014).
50. Q. Wei *et al.*, A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* **31**, 1375–1381 (2015).
51. S.-H. Kim, N. Elango, C. Warden, E. Vigoda, S. V. Yi, Heterogeneous genomic molecular clocks in primates. *PLoS Genet.* **2**, e163 (2006).
52. K. Magnusson, Interpreting Cohen's *d* effect size: An interactive visualization. https://rpsychologist.com/d3/cohend/. Accessed 22 November 2019.
53. L. Anderson-Trocmé *et al.*, Legacy data confounds genomics studies. *Mol. Biol. Evol.*, msz201 (2019).
54. I. Agarwal, M. Przeworski, Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 17916–17924 (2019).
55. K. D. Makova, R. C. Hardison, The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
56. W. Amos, Heterozygosity increases microsatellite mutation rate. *Biol. Lett.* **12**, 20150929 (2016).
57. W. Amos, J. Flint, X. Xu, Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genet.* **9**, 72 (2008).
58. R. Agarwala, L. G. Biesecker, K. A. Hopkins, C. A. Francomano, A. A. Schaffer, Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. *Genome Res.* **8**, 211–221 (1998).
59. B. D. Mitchell *et al.*, Living the good life? Mortality and hospital utilization patterns in the Old Order Amish. *PLoS One* **7**, e51560 (2012).
60. S. M. Nolt, *The Amish: A Concise Introduction* (JHU Press, 2016).
61. C. A. Menzie, B. B. Potocki, J. Santodonato, Exposure to carcinogenic PAHs in the environment. *Environ. Sci. Technol.* **26**, 1278–1284 (1992).
62. P. S. Nielsen, H. Okkels, T. Sigsgaard, S. Kyrtopoulos, H. Autrup, Exposure to urban and rural air pollution: DNA and protein adducts and effect of glutathione-S-transferase genotype on adduct levels. *Int. Arch. Occup. Environ. Health* **68**, 170–176 (1996).
63. V. A. Tsihrintzis, R. Hamid, Modeling and management of urban stormwater runoff quality: A review. *Water Resour. Manage.* **11**, 136–164 (1997).
64. R. F. Hamman, J. I. Barancik, A. M. Lilienfeld, Patterns of mortality in the Old Order Amish. I. Background and major causes of death. *Am. J. Epidemiol.* **114**, 845–861 (1981).
65. J. A. Westman *et al.*, Low cancer incidence rates in Ohio Amish. *Cancer Causes Control* **21**, 69–75 (2010).
66. T. A. Sasani *et al.*, Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *bioRxiv*:10.1101/552117 (17 February 2019).
67. A. S. Hinrichs *et al.*, The UCSC genome browser database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
68. A. Koren *et al.*, Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
69. P.-R. Loh *et al.*, Reference-based phasing using the Haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
70. A. Auton *et al.*; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
71. D. N. Harris *et al.*, Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6526–E6535 (2018).
72. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
73. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).