De novo mutation rates at the single-mutation resolution in a human *HBB* gene-region associated with adaptation and genetic disease

Daniel Melamed, Yuval Nov, Assaf Malik, et al.

Genome Res. published online January 14, 2022
Access the most recent version at doi:10.1101/gr.276103.121

De novo mutation rates at the single-mutation resolution in a human *HBB* gene-region associated with adaptation and genetic disease

Daniel Melamed,[1,2] Yuval Nov,[3] Assaf Malik,[4] Michael B. Yakass,[5,6]
Evgeni Bolotin[1,2], Revital Shemer[7], Edem K. Hiadzi,[6]
Karl L. Skorecki,[8] Adi Livnat[1,2*]

[1]Department of Evolutionary and Environmental Biology, University of Haifa, 3498838, Israel,
[2]Institute of Evolution, University of Haifa, 3498838, Israel
[3]Department of Statistics, University of Haifa, 31905, Israel
[4]Bioinformatics Unit, Faculty of Natural Sciences, University of Haifa, 3498838 Israel
[5]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP),
Department of Biochemistry, Cell & Molecular Biology, University of Ghana,
Legon P.O. Box LG 54, Ghana
[6]Assisted Conception Unit, Lister Hospital & Fertility Centre,
Accra P.O. Box CT 966, Ghana
[7]The Ruth and Bruce Rappaport Faculty of Medicine & Research Institute,
Technion - Israel Institute of Technology, Haifa, 3525433, Israel
[8]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, 1311502 Israel
* To whom correspondence should be addressed; E-mail: alivnat@univ.haifa.ac.il

Running title: Rates of target *de novo* mutations

While it is known that the mutation rate varies across the genome, previous estimates were based on averaging across various numbers of positions. Here we describe a method to measure the origination rates of target mutations at target base positions and apply it to a 6-bp region in the human hemoglobin subunit beta (*HBB*) gene and to the identical, paralogous hemoglobin subunit delta (*HBD*) region in sperm cells from both African and European donors. The *HBB* region of interest (ROI) includes the site of the hemoglobin S (HbS) mutation, which protects against malaria, is common in Africa and has served as a classic example of adaptation by random mutation and natural selection. We found a significant correspondence between *de novo* mutation rates and past observations of alleles in carriers, showing that mutation rates vary substantially in a mutation-specific manner that contributes to the site frequency spectrum. We also found that the overall point mutation rate is significantly higher in Africans than in Europeans in the *HBB* region studied. Finally, the rate of the 20A→T mutation, called the "HbS mutation" when it appears in *HBB*, is significantly higher than expected from the genome-wide average for this mutation type. Nine instances were observed in the African *HBB* ROI, where it is of adaptive significance, representing at least three independent originations; no instances were observed elsewhere. Further studies will be needed to examine mutation rates at the single-mutation resolution across these and other loci and organisms and to uncover the molecular mechanisms responsible.

It is widely known that mutation rates vary across the genome at multiple scales (Hodgkinson and Eyre-Walker, 2011; Rahbari et al., 2016; Carlson et al., 2018) and are affected by multiple factors, from the mutation type (Gojobori et al., 1982; Bulmer, 1986), to the local genetic context (Gojobori et al., 1982; Bulmer, 1986; Blake et al., 1992; Hwang and Green, 2004; Rahbari et al., 2016; Carlson et al., 2018) to the general location in the genome (Wolfe et al., 1989; Matassi et al., 1999; Lercher et al., 2001; Ellegren et al., 2003). Although this

knowledge is highly advanced now compared with what was known a mere decade ago (Campbell et al., 2012; Michaelson et al., 2012; Francioli et al., 2015; Rahbari et al., 2016; Carlson et al., 2018), it could be enhanced further. In particular, rate measurements to date all have been based on averages of various kinds, such as an average across the genome (Nachman and Crowell, 2000; Rahbari et al., 2016), or across the instances of any particular motif (Hwang and Green, 2004; Carlson et al., 2018), or, in certain cases, across the entire stretch of a gene (Haldane, 1949; Vogel and Motulsky, 1997; Kondrashov, 2003). In contrast, technological limitations have precluded measuring mutation rates at particular base positions and of particular mutations at such positions. However, such high-resolution knowledge of the mutation-rate variation would bear on multiple open questions in genetics and evolution—from the relative importance of mutation-rate variation to the site frequency spectrum (SFS) (Lek et al., 2016; Harpak et al., 2016; Mathieson and Reich, 2017), to its importance for adaptive evolution and parallelism (Inoue et al., 2001; Crow et al., 2009; Dumas et al., 2012; Losos, 2017; Xie et al., 2019; Kratochwil et al., 2019; Kratochwil and Meyer, 2019; Lind, 2019), to its contribution to recurrent genetic disease and cancer (Lupski, 1998; McClellan and King, 2010; Veltman and Brunner, 2012; Shendure and Akey, 2015).

The most precise way of measuring mutation rates, free of biases due to past natural selection or random genetic drift events, is offered by *de novo* mutations—mutations that appeared for the first time in their carrier (Rahbari et al., 2016; Goldmann et al., 2016). These mutations are usually detected by studies comparing the genomes of children to those of their parents, a.k.a. "trio studies" (Roach et al., 2010; Conrad et al., 2011). However, because each individual carries only a small number (e.g., several dozen in humans) of *de novo* mutations scattered across the genome, the chance of encountering any particular target mutation of interest is miniscule, rendering it impractical to measure rates of target mutations using such studies.

To overcome this barrier, we have developed a method that enables identifying and counting, with high accuracy, ultra-rare genetic variants of choice in extremely narrow regions of interest (ROIs) within large populations of cells, such as a single target mutant in 100 million genomes. Since this method has both an error rate lower than the human mutation rate and sufficient yield

for the purpose, it enables measuring the frequencies of target mutations of choice in human sperm samples by counting their *de novo* instances at a single-digit resolution. For variants that are not expected to affect sperm fertility and viability (as in the case below), this frequency is the evolutionarily relevant mutation rate in males. Note that aside from this evolutionary application, ultra-accurate methods of mutation-detection are sought after for early detection of cancer, non-invasive prenatal testing, early identification of virus within host, and more (Salk et al., 2018).

As a first target for this method, we chose two sites: a 6-bp region spanning 3 codons within the human hemoglobin subunit beta (*HBB*) gene that is of great importance for adaptation and hematologic disease, and the identical, paralogous region within the hemoglobin subunit delta (*HBD*) gene. The former region includes, among others, the site of the HbS mutation. The most iconic balanced polymorphism mutation (Pauling et al., 1949; Ingram, 1957; Allison, 1954; Hartl and Clark, 2007; Cavalli-Sforza and Feldman, 2003; Feng et al., 2004), the HbS mutation is an A to T transversion (GAG→GTG, Glu→Val) in codon 6 of *HBB* causing sickle-cell anemia in homozygotes (Pauling et al., 1949) and providing substantial protection against severe malaria in heterozygotes (Allison, 1954; Piel et al., 2010; Flint et al., 1998; Kwiatkowski, 2005). Malaria, in turn, has been a leading cause of human morbidity and mortality, often causing more than a million deaths per year in the recent past, with Africa bearing the brunt of the disease burden (Carter and Mendis, 2002), and thus has been the strongest known agent of selection in humans in recent history (Kwiatkowski, 2005). Besides the HbS mutation, many other mutations, both point mutations and indels, are also known at this site, many of which are involved in hematologic illness (Hardison et al., 2002; Hardison and Miller, 2002). In contrast to *HBB*, mutations in *HBD* have a more limited effect and are not thought to confer resistance to malaria, as the *HBD*'s lower expression levels make it account for less than 3% of the circulating red blood cell hemoglobin in adults (Steinberg and Adams, 1991). While the population prevalence of the *HBB* mutations, whether beneficial or detrimental, is normally attributed to natural selection, so far it has not been possible to examine to what degree, if at all, mutational phenomena may also be relevant to their prevalence. To address this gap, we sought

to characterize the rates of mutations, including the HbS mutation, in the *HBB* and *HBD* ROIs in sperm samples of both African and European donors.

Results

To substantially reduce the false positive rate due to PCR amplification or high-throughput sequencing errors, following extraction of the DNA from the sperm of the donors, we first remove the majority of wild-type (WT) ROI molecules from each sample. Specifically for the target sites, we use the restriction enzyme (RE) Bsu36I, which cleaves the WT sequence CCTGAGG at positions 16–22 of *HBB* and the paralogous positions of *HBD* while leaving the HbS mutant and other mutants in these positions intact. Besides substantially reducing the false positive rate, this WT depletion has the additional benefit of reducing the sequencing costs by the same factor, because it removes the majority of fragments whose sequences are known to be WT (Fig. 1, Supplemental Text and Figs. S1–S4).

Importantly for the mutation rate calculation, we keep track of the number of WT molecules removed by accurately calculating the protected mutants' enrichment factor on a per sample basis. For this purpose, we generate two mixtures, each of which includes, in addition to the DNA studied, known amounts of mock DNA that is resistant to the RE digestion (Supplemental Text S2 and Fig. S2). Next, we apply the same protocol to the two mixtures, with the exception that the RE digestion step is applied to only one of them (Supplemental Text S2 and Fig. S2). The ratio of the ratios of sensitive to resistant molecules identified for the two mixtures after treatment at the sequence analysis step provides the enrichment factor of the protected mutants (Supplemental Text S2 and Fig. S2). This enrichment factor, multiplied by the number of WT molecules called, with the addition of the small number of mutants called, provides the number of cells analyzed (Supplemental Text S2 and Fig. S2). We set up the system in such manner that the calculation of the enrichment factor depends only on quantities that are precisely known, including volume measurements (Supplemental Text S2 and Fig. S2) and numbers of WT and mutant molecules called during the barcode-based sequence analysis stage as described below.

Following this mutation enrichment step, we attach unique barcodes to the DNA fragments

in order to reduce error by consensus sequencing of copies originating from the same original fragment. For this purpose, we build on and improve the Maximum Depth Sequencing method (MDS) (Jee et al., 2016), which allows one to focus on a narrow region of interest (ROI) and whose key idea is to attach the barcodes directly to a cleaved end of one of the two strands of each original target DNA fragment via a DNA polymerase–assisted extension reaction, as opposed to including the barcode only in the first copy of the DNA by extending the target-specific primer that carries it. In this manner, also errors that occur during the first, critical copying step are detected via consensus sequencing of reads sharing the same barcode (Jee et al., 2016) (Fig. 1, Supplemental Text and Fig. S1). To all of the above we add multiple innovations that increase sequencing accuracy, handle the large amounts of genomic DNA required and enable accurate measurement of the Bsu36I enrichment factor per sample as needed for the mutation rate calculation (Supplemental Figs. S1–S5). We refer to this whole method as Mutation Enrichment followed by upscaled Maximum Depth Sequencing—MEMDS (for a complete protocol, see Supplemental Text S1–S9 and Methods).

Finally, following sequence analysis (Supplemental Figs. S6–S10 and Table S2) the number of appearances of any mutation that confers resistance to the restriction enzyme is counted and divided by the calculated number of cells analyzed, providing the evolutionarily relevant *de novo* origination rate for each specific mutation in males per donor and per group of donors (Supplemental Figs. S11–S13 and Table S3). Following previous literature, we ignore G→T and C→T mutations in the barcoded strand (C→A and G→A in the sequenced strand) because they are thought to reflect not lasting mutations but the experimental disruption of an ongoing *in vivo* process of base damage and repair as well as *in vitro* mutations due to guanine oxidation and cytosine deamination (Arbeithuber et al., 2016; Jee et al., 2016) (Supplemental Text S8 and Figs. S12–S14). In addition, we exclude C→A, the complement of G→T, due to its association with the latter and its frequent appearance in the data (Supplemental Text S8 and Fig. S12). Following normal loss of material of ∼65%, true positives of non G→T, C→T and C→A mutations are identified with a false positive rate (error rate) $< 2.5 \times 10^{-9}$ per base (Fig. 2). Overall, MEMDS surpasses recent cutting-edge methods in both accuracy (Fig. 2A) and yield (Fig. 2B)

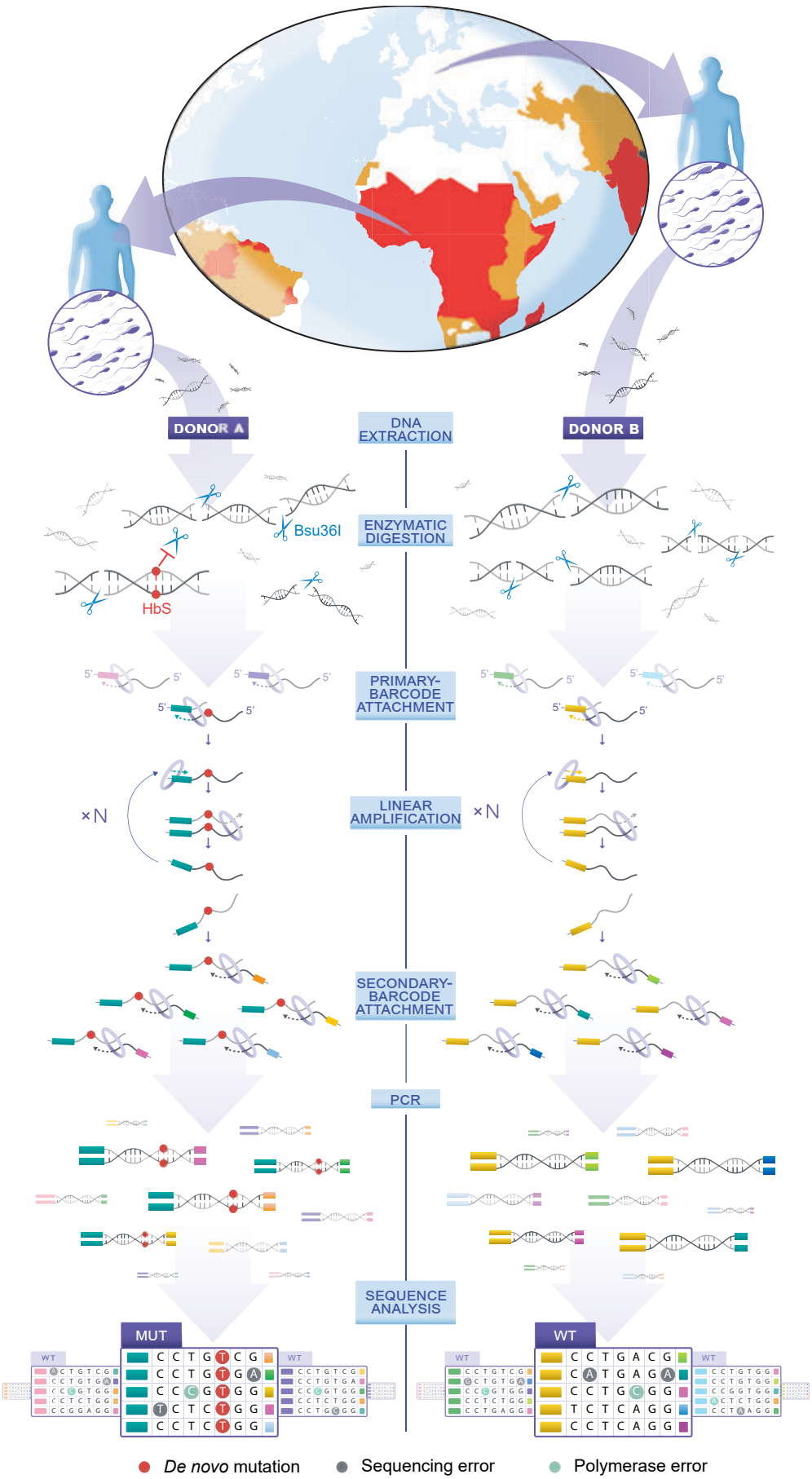

Figure 1: Experiment overview. Sperm samples are obtained from world regions with high or low malaria infection burden (malaria impact map adjusted from the CDC map; CDC Division of Parasitic Diseases & Malaria, 2019). Whole-genome DNA is extracted and an amount equivalent to 60–80 million sperm cells per donor is subjected to Bsu36I digestion. Bsu36I cleaves the DNA at multiple sites, including the *HBB* and *HBD* ROIs, which carry a specific recognition sequence. The HbS mutation blocks Bsu36I digestion and is thus enriched over the wild-type (WT). A primary barcode is added directly to each antisense DNA strand that carries the *HBB* or *HBD* ROI via a DNA polymerase–assisted fill-in reaction. Since each barcode consists of a random sequence of nucleotides, each of the numerous target fragments has its own, unique barcode, illustrated by a unique color on the left-end side of the representation of each barcoded fragment. Multiple single-strand copies are each generated directly from each uniquely barcoded target fragment by linear amplification. A secondary barcode composed of a random sequence of nucleotides is added to the other end of each of these copies by a single primer-extension reaction, illustrated by a unique color on the right-end side of each barcoded fragment. Thus, only full-length fragments (i.e., mutant or WT ROI sequences that evaded Bsu36I digestion) carry both the primary and the secondary barcodes and can be amplified by PCR for high-throughput sequencing. At the sequence analysis step, sequencing reads representing the PCR products of the linearly amplified copies are grouped together into families (see boxes), where in each family reads share the same primary barcode sequence. Sporadic sequencing errors or DNA-polymerase errors generated during linear or subsequent amplification steps are unlikely to be repeated in multiple copies and are removed. *De novo* mutations, such as the HbS mutation, are easily identified by their appearance in multiple reads from distinct linear-amplification events. For a complete description of the library preparation protocol, which includes additional steps, see Supplemental Figs. S1–S3.

(see also Supplemental Fig. S11).

With the help of this method, we examined a total of more than half a billion gene fragments individually taken from sperm of 12 donors. Since one of the samples was a mixture from two African donors with a total number of cells similar to the other African samples, we consider it here as a single sample of mixed African origins, bringing the total to 11 samples, 7 from African and 4 from European donors (Supplemental Table S1). The numbers of cells scanned and *de novo* mutations observed per person are shown in Table 1.

Average per ROI mutation rates: The average per base point mutation rates in the *HBB* and *HBD* ROIs are 3.3×10^{-8} and 2.79×10^{-8} respectively, significantly higher by $\sim$2.6-fold ($P < 2 \times 10^{-8}$, 95% CI 2.4×10^{-8}–4.4×10^{-8}) and $\sim$2.2-fold ($P < 6.7 \times 10^{-5}$, 95% CI 1.9×10^{-8}–4×10^{-8}, two-sided binomial exact test) than 1.25×10^{-8}, which we use as an estimate of the genome-wide per base per generation point-mutation rate (Supplemental Text S10). The average indel rates in these ROIs were 1.1×10^{-8} and 4.3×10^{-9} respectively,

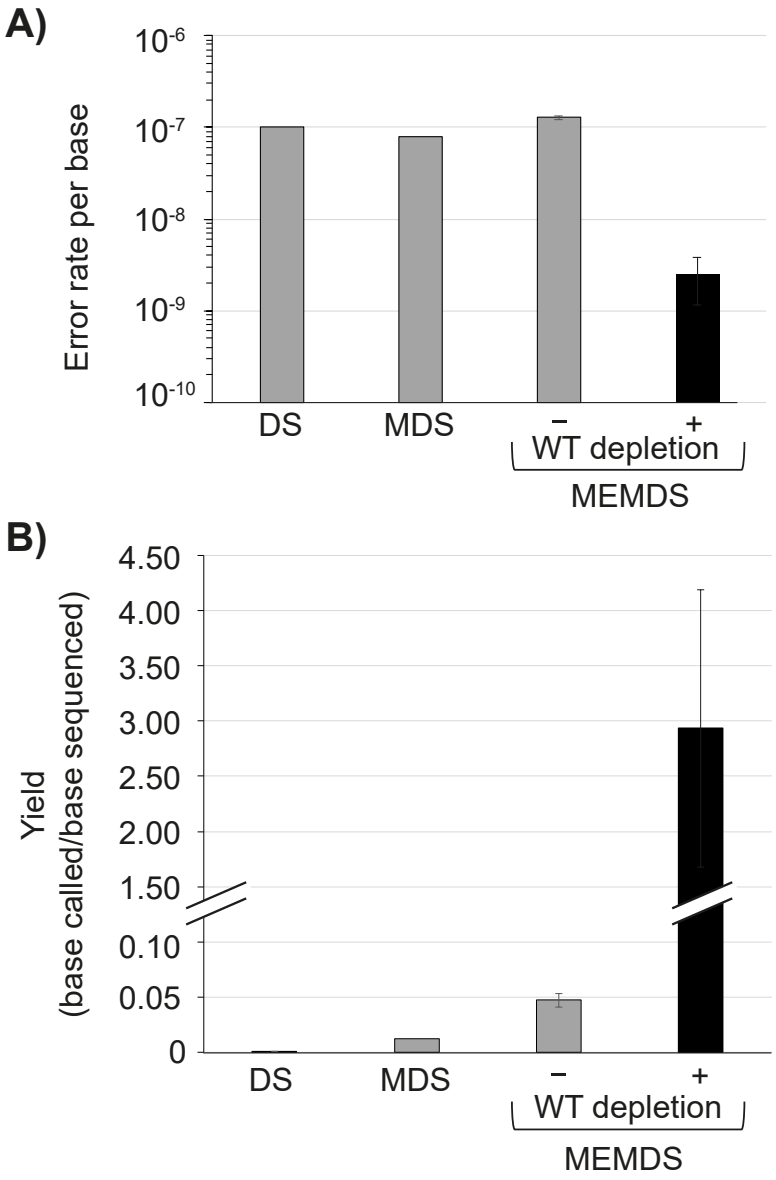

Figure 2: Accuracy and yield of MEMDS compared with current cutting-edge methods for studying target regions. A) Under a highly conservative estimate, MEMDS increases accuracy by at least 40-fold compared to Duplex Sequencing (DS) (Kennedy et al., 2014) and Maximum-Depth Sequencing (MDS) (Jee et al., 2016). B) MEMDS also increases yield per sequenced base (i.e., the number of MEMDS confirmed bases divided by the number of paired-end sequenced bases) by orders of magnitude over both DS and MDS (Kennedy et al., 2014; Jee et al., 2016). Notice that in MEMDS, the yield can be higher than 1 because the mutation-enrichment factor is accurately calculated (see Supplemental Text S2) and the base identity is known for the ROI sequences that were digested and removed from the final sequencing libraries (they have the restriction enzyme recognition sequence). Although the accuracy of DS has been improved in the context of sequencing large parts of the genome (Abascal et al., 2021), yield considerations and targeted DNA capture limitations preclude applying that method to narrow ROIs and target mutations (see Kennedy et al., 2014 and Supplemental Text S1).

			Position (or center of indel) ►	Point mutations												Indels					
				16		17		18			20			21	22	15	16	17-18	18	20 or 23	
			Cells scanned ▼	16C>G	16C>T	17C>G	17C>T	18T>G	18T>A	18T>C	20A>G	20A>T[1]	20A>C	21G>C	22G>C	14_16 del	16 del	17_18 del	18 del	20 del	19_21 or 22_24 del[2]
AFR	*HBB*	AFR1	22.3x10^6					1	1			5									11
		AFR2	30.9x10^6								2	2									
		AFR3	29.2x10^6				1					2	1								1
		AFR4	23.8x10^6		1	2				1											2
		AFR5	19.4x10^6			1	1				1								1		
		AFR6	25.0x10^6		1	1				1	1			1							
		AFR7	39.1x10^6	3	1	1	1				1									1	4
	HBD	AFR1	17.0x10^6	1		1				2					1						1
		AFR2	24.5x10^6	1														1			1
		AFR3	22.7x10^6	1	7																1
		AFR4	19.0x10^6	2			1			1											
		AFR5	14.5x10^6																	1	1
		AFR6	22.9x10^6			1	2			1	2										
		AFR7	32.8x10^6	1						1											1
EUR	*HBB*	EUR1	39.0x10^6	2												1					10
		EUR2	39.1x10^6								1		1				1				2
		EUR3	49.0x10^6			5				1											3
		EUR4	16.5x10^6																		4
	HBD	EUR1	33.8x10^6		1						1										
		EUR2	31.8x10^6																		2
		EUR3	44.3x10^6	1																	3
		EUR4	14.6x10^6							2											1

1. HbS
2. Hb-Leiden

Table 1: *HBB* and *HBD* ROI mutation counts. Counts of *de novo* mutations identified by MEMDS in DNA from 11 sperm samples, 7 from African (AFR) and 4 from European (EUR) donors. The numbers next to the donor labels refer to the calculated number of haploid individual genomes scanned by MEMDS. Some of the mutations have been observed before in carriers and have common names when they appear in *HBB*. These are 16C→G, Hb-Gorwihl; 16C→T, Hb-Tyne; 17C→G, Hb-Warwickshire; 17C→T, Hb-Aix-les-Bains; 20A→G, Hb-Lavagna; 20A→T, HbS; 20A→C, Hb-G-Makassar; 22G→C, Hb Bellevue III and 19_21del or 22_24del, Hb-Leiden. Note that Hb-Leiden can result from deletion of either positions 19–21 or positions 22–24, which include the same GAG sequence, both of which can be enriched and captured by MEMDS.

significantly higher by ∼9-fold ($P < 4.3 \times 10^{-25}$, 95% CI 8×10^{-9}–1.5×10^{-8}) and ∼3.4-fold ($P < 1.8 \times 10^{-4}$, 95% CI 2.3×10^{-9}–7.3×10^{-9}; two-sided binomial exact test) than the expected 1/10 of the point mutation rate (Supplemental Text S10). The average point mutation rate of the *HBB* ROI is not significantly higher than that of the *HBD* ROI ($P = 0.49$, two-sided Fisher's exact test), and the average indel rate of the former is significantly higher by ∼2.6-fold than that of the latter ($P = 0.0015$, OR 95% CI 1.42 – 5.01; two-sided Fisher's exact test).

Basic characteristics of mutation-rate variation: The variance in the rates of *de novo* point mutations is higher than expected from the genome-wide average (GWA) rates of these mutations (e.g., Harris, 2015; Harris and Pritchard, 2017) and their relative rates are different than expected from the GWA rates ($P < 10^{-6}$ in an omnibus multinomial test, adjusted for the excluded mutations, compared to the rates of Rahbari et al., 2016), even when adjusting the latter for the 3-mer, 5-mer and 7-mer nucleotide contexts ($P < 10^{-5}$ in all cases, compared to the rates of Carlson et al., 2018). The overall *de novo* rates of the 6 observed deletion types are highly non-uniform ($P < 10^{-6}$, multi-sample proportion test).

Correspondence between *de novo* rates and observations of alleles in carriers: The HbS and Hb-Leiden mutations both have been notably observed on multiple different genetic backgrounds in human populations, the former particularly in Africans (Flint et al., 1998; Hardison et al., 2002; Hardison and Miller, 2002). Here, they are the point mutation of highest *de novo* rate in the African *HBB* ROI and the deletion mutation of highest *de novo* rate in any gene and ethnicity. Furthermore, of the 23 potential deletions of up to size 3 that are observable by our method per ROI, only 5 deletions (16delC, 17_18delCT, 18_19delTG, 19_21delGAG or the equivalent 22_24delGAG—the Hb-Leiden mutation—and 20delA) have been reported to date on the HbVar database—a large collection of hemoglobin variants (Hardison et al., 2002; Hardison and Miller, 2002)—all in *HBB*; and of these deletion types, a significantly higher fraction is observed here *de novo* compared to deletion types not reported on HbVar (Supplemental Text S11). Pooling together both the *HBB* and *HBD* ROIs given the similarity of *de novo* indel types observed between them, this effect is significant both with ($P = 0.0078$, OR 95% CI 2.17–818.08, two-sided Fisher's exact test) and without ($P = 0.024$, OR 95% CI 1.44–

653.93, two-sided Fisher's exact test) the Hb-Leiden mutation, showing that the correspondence between *de novo* rates and alleles in populations extends beyond the HbS and Hb-Leiden mutations. Although the same analysis cannot be repeated for the point mutations because of the smaller number of observable mutation types and the synonymous vs. non-synonymous mutation confound, further observations are in the expected direction (Supplemental Text S11). The correspondence observed could not have been predicted from the mutations' GWA rates, even when adjusting for the genetic context (Supplemental Text S10–S11).

Between-population comparisons: In order to provide a conservative statistical test of a population-level difference that excludes individual- or sample-level variation alone as accounting for the result, we compared the per person overall point mutation rates in the *HBB* ROI between the African and European groups. Results showed that these rates were significantly higher in the African than in the European group both with ($P = 0.0061$) and without ($P = 0.043$, two-sided Wilcoxon rank sum test) counting the HbS mutation. Next, pooling together cells from all donors within each population to estimate the overall point-mutation rate in the *HBB* ROI shows it to be significantly higher by 2.57-fold in the African than in the European donors ($P < 0.006$, OR 95% CI 1.27–5.49, two-sided Fisher's exact test). Thus, there is a significant population-level difference between the continental groups in the overall point mutation rate in this narrow ROI that is not attributable to individual- or sample-level variation. In contrast, in the *HBD* ROI, the number of mutations was not high enough to establish such a difference above and beyond individual- or sample-level variation ($P = 0.18$, two-sided Wilcoxon rank sum test). In contrast to the *HBB* overall point mutation rate, the overall indel rate did not vary significantly between these groups in either ROI (P=0.35 and P=1, respectively, two-sided Fisher's exact test).

Position 20 mutation rates: Two particularly notable mutations are the HbS and Hb-Leiden mutations (details below). Considering codons 6 and 7 equivalent with respect to the latter mutation, both mutations can be said to affect position 20. Using the conservative test above-mentioned to exclude sample-level variation alone as accounting for the result, the overall per person point mutation rates at position 20 specifically are significantly higher in the *HBB* than in

the *HBD* ROI in Africans ($P = 0.017$, two-sided Wilcoxon rank sum test) but not in Europeans ($P = 1$). In the former, the overall point mutation rate at position 20 pooled across individuals is $\sim 6.1\times$ higher in *HBB* than in *HBD* ($P = 0.0061$, OR 95%CI: 1.50–37.14, two-sided Fisher's exact test). In the case of the overall indel rates at position 20, although the pooled rates are significantly higher in *HBB* than in *HBD* for both Africans and Europeans ($P = 0.044$, OR 95% CI: 1.03–6.54 and $P = 0.027$, OR 95% CI: 1.11–7.02 respectively; two-sided Fisher's exact tests), sample-level variation cannot be excluded as the source of the differences ($P = 1$ and $P = 0.69$ for Africans and Europeans, respectively; two-sided Wilcoxon rank sum tests).

Rates of the Hb-Leiden mutation: The 3 bp in-frame deletion variant of either codon 6 or codon 7 that is called "the Hb-Leiden mutation" when it occurs in *HBB* recurs noticeably more often than other mutations (comparing its per person rates to those of all other deletions combined to exclude sample-level variation, $P < 0.0005$, two-sided Wilcoxon rank sum test). Pooled across individuals, it appears at rates of 1.11×10^{-7} and 3.96×10^{-8} in the *HBB* and *HBD* ROIs respectively, $\sim 88.86\times$ and $\sim 31.66\times$ higher than the 1.25×10^{-9} estimate ($P = 4.04 \times 10^{-58}$, 95% CI 7.82×10^{-8}–1.53×10^{-7}; and $P = 1.62 \times 10^{-13}$, 95% CI 1.98×10^{-8}–7.08×10^{-8}), where the *HBB* rate is significantly ($\sim 2.81\times$) higher than the *HBD* rate ($P = 0.002$, OR 95% CI 1.40–5.63, two-sided Fisher's exact test).

Rates of the HbS mutation: The 20A→T mutation called "the HbS mutation" when it appears in the *HBB* ROI appears 9 times in the African *HBB* ROI and no times in the other cases combined (the European *HBB* ROI and the European and African *HBD* ROIs) ($P = 0.023$, 95% CI 1.5077–Inf; two-sided Fisher's exact test classifying each individual and gene case as having [> 0] or not having [$= 0$] *de novo* 20A→T in sperm and comparing the fractions of these classes between the groups). The rate of the HbS mutation in the overall group (Africans and Europeans combined)—2.7×10^{-8}—is $19.6\times$ higher ($P < 2 \times 10^{-9}$, rate 95% CI 1.24×10^{-8}–5.13×10^{-8}) than expected from the GWA for this mutation type (Supplemental Text S10), and its rate in the African group specifically—4.74×10^{-8}—is $\sim 35\times$ higher than expected from its GWA ($P = 1.2 \times 10^{-11}$, rate 95% CI 2.17×10^{-8}–9.0×10^{-8}; two-sided binomial exact test). In the African group, it is the mutation that deviates the most (Table S4) from its GWA among

the 12 observable point mutations, where its *de novo* rate varies significantly across samples ($P = 0.0025$, multi-sample proportion test), from 0 to 2.24×10^{-7} (the latter rate being $\sim$163$\times$ faster than expected; $P = 2.23 \times 10^{-10}$, 95% CI 7.27×10^{-8}–5.23×10^{-7}, two-sided binomial exact test). Note that the evolutionarily relevant mutation rate depends on the fraction of the mutation in sperm *per se*, not on whether it repeats because of independent originations or due to an early appearance followed by duplications during spermatogenesis. That being said, the minimal number of independent originations of the HbS mutation is 3, given that 3 individuals produced it *de novo*, and the corresponding minimal rate of independent occurrence of the HbS mutation in the sperm samples (a rate lower than the actual evolutionarily relevant mutation rate observed) across all individuals is 9.01×10^{-9}. This rate is still $\sim$6.5$\times$ higher than the genome-wide evolutionarily relevant mutation rate for this mutation type ($P = 0.011$, 95% rate CI 1.86×10^{-9}–2.63×10^{-8}, two-sided binomial exact test).

Discussion

The data exposes an ultra-high resolution correspondence between *de novo* mutation rates and past observations of alleles in carriers (Flint et al., 1998; Hardison et al., 2002; Hardison and Miller, 2002, Results and Supplemental Text S11), suggesting that these rates contribute to the prevalence of these mutations in populations. This correspondence could not have been predicted from the GWA rates of these mutation types even when adjusting for the local genetic context (Supplemental Text S10–S11). Consideration of the deletions observed clarifies this point. While past literature featured a single microdeletion rate decreasing with size (Gu and Li, 1995; Kondrashov, 2003; Lynch, 2010), sized-based rate-variation cannot explain the aforementioned correspondence obtained for same-sized deletions, the higher rate of the Hb-Leiden mutation compared to the smaller deletions, or the extent of rate-variation observed. Thus, the correspondence aforementioned, together with the fact that, in these ROIs, the rates of some mutations (e.g., those of the HbS and Hb-Leiden mutations) deviate much more than others from their corresponding GWA rates show that mutation-specific rates vary not only in the case of large rearrangement mutations (Gu et al., 2008; Zhang et al., 2009) but also in the cases of point

mutations and microindels. This rate variation could not have been seen using average-based measures (Kondrashov, 2003; Lynch, 2010) and establishes the relevance of mutation-specific point mutation and microindel rates to the site frequency spectrum (SFS) (Lek et al., 2016; Harpak et al., 2016; Mathieson and Reich, 2017).

The overall point mutation rate in the *HBB* ROI is significantly higher in the African than in the European group even under a nonparametric comparison which shows that the difference cannot be attributed to individual- or sample-level variation alone. Thus, it represents a significant population-level difference between the groups. This difference, occurring in an extremely narrow region spanning 3 codons of great importance for adaptation and genetic disease, is at least two orders of magnitude larger than previously reported differences in GWA mutation rates between continental groups (Harris, 2015; Harris and Pritchard, 2017). The correspondence between mutation-specific *de novo* rates and observations of alleles in carriers as well as this large difference in the overall point mutation rate between populations in a narrow region establish the importance of measuring mutation-rate variation at an ultra-high resolution.

Potential contributions to mutation rates from gross-level biological or environmental factors, such as age or pesticides, cannot sufficiently explain the results. First, the two populations are similar in ages (Supplemental Table S1). Second, any mutation-specific effect, like the correspondence between *de novo* rates and observations of alleles in carriers, cannot be explained by such macro-level factors, as the latter cannot be expected to affect the rates of equivalent mutations such as 20A→T in *HBB* vs. *HBD* differently. Third, the overall point mutation rate difference between the populations is also unlikely to be explained by them, because if on their own such macro-level factors had affected the ROIs, they should have affected the entire genome similarly, yet GWA differences in point mutation rates between continental groups are smaller than the ROI-specific differences observed here (Harris, 2015; Harris and Pritchard, 2017). Note that if macro-level factors affect mutation rates in interaction with mutation-, locus-, individual- and/or population-specific factors, then such specific factors must be assumed in any case. Thus, rather than suggesting involvement of macro-level factors, the data suggests a complex picture of mutation rates involving mutation-specific influences.

In addition, while the replication of mutations during spermatogenesis (clonal dependence) may make some contribution to the data, in practice it is insufficient in order to account for the significant results. First, the significance of the continental difference in the overall point mutation rates in *HBB* is impervious to any sample-level variation, including clonal dependence, as shown by the non-parametric between-population comparison described in the results section. Second, the correspondence between mutation rates and observations of alleles in carriers cannot be driven by it. On the contrary, in the absence of a cellular-level mechanism that induces specific mutations in a population-specific manner in accord with the cellular generation during spermatogenesis, differences in mutation timing during spermatogenesis could only add noise to the patterns observed, and thus any presence of clonal dependence would only make it more difficult to obtain significance for such patterns and in that sense is conservative to finding a pattern. Thus, more likely, the significance of these patterns is driven by independent originations of the mutations. These independent originations are consistent with mutation-specific rates being influenced by genetic and/or epigenetic factors (Livnat, 2013, 2017).

The prevalence of a mutation of heterozygote advantage in a population and of reading-frame conservation in a coding sequence have generally been considered to be outcomes of selection. However, here, both the HbS mutation, which provides strong malaria protection in heterozygotes, and the Hb-Leiden mutation, which is an in-frame deletion, are frequent not because of selection but because of frequent *de novo* origination. Indeed, that the rate of the in-frame Hb-Leiden mutation is much higher than that of all other observed deletions, which are frameshift deletions, demonstrates reading-frame conservation that is due not to selection (Lek et al., 2016) but rather to mutational phenomena. This observation provides a concrete example of "mutational conservation"—evolutionary conservation due to mutational reasons which, if it occurs more broadly, could offer an explanation for the puzzling observation of reading-frame conservation bias in pseudogenes (Zhang and Gerstein, 2003).

The fact that the genetic sequences at and adjacent to the ROIs are identical for the two populations and for the two genes yet the mutation rates vary significantly between the populations and between the genes suggests that what affects these mutation rates in the germline includes

more than this local DNA sequence and in that sense is complex (Livnat, 2013, 2017). These results are consistent with the observation that the variation of the mutation rates across loci is partly cryptic (not explained by the local DNA context) (Hodgkinson et al., 2009; Hodgkinson and Eyre-Walker, 2011), especially in the case of A↔T transversions (Hodgkinson et al., 2009), which include the HbS mutation-type (A→T). Combining the multiple insights discussed, the results suggest that mutation rates are both mutation-specific and influenced in a complex manner by the genetic and/or epigenetic background (Livnat, 2013, 2017).

The *HBB* region spanning 3 codons is of particular importance for adaptation and genetic disease: it is the site of mutations that provide strong protection against malaria (HbS and HbC, the latter not observable by our method) and/or increase the risk for hematologic disease (Flint et al., 1998; Hardison et al., 2002; Hardison and Miller, 2002). Thus, it is of interest that the overall point mutation rate in this region is significantly higher than expected, and that it is significantly higher in the African than in the European population. These results provide a clear case of a connection between mutation rates and adaptive evolution, thus moving beyond previous literature on the relevance of mutation rates to adaptive evolution and its repeatability (Crow et al., 2009; Dumas et al., 2012; Xie et al., 2019; Kratochwil et al., 2019; Kratochwil and Meyer, 2019; Lind, 2019).

The results underscore the importance of mapping the mutation-rate variation at an ultra-high resolution. It is beyond this fact that several observations on the HbS mutation specifically can be mentioned. First, if one assumes that the HbS rate is the same for both of the continental groups, the data shows that it is significantly higher by nearly 20-fold than expected from the GWA for this mutation type, in both Africans and Europeans. Any amount of hypothetical clonal dependence does not change this estimate of the observed evolutionarily relevant mutation rate, because the latter does not depend on the cause of the recurrence of the mutation in the sperm. Even the observed minimal rate of independent HbS originations in sperm is still significantly larger by $6.5\times$ than the evolutionarily relevant GWA rate for this mutation type. Consideration of the local genetic context does not change this conclusion (Supplemental Text S10). Thus, while the classical explanation of the HbS case relied only on selection, even un-

der the most conservative assumptions the overall HbS mutation rate observed here is notably higher than expected.

Second, given the significant continental difference in the overall point mutation rate between the groups, it would be surprising if the HbS mutation specifically does not show a continental effect. Consistent with this, in our samples, using the methodology described, we observe no instances of it in Europeans but 9 instances of it in total in Africans, amounting to a rate $\sim 35\times$ higher than expected from the GWA of this mutation type in the latter. Further consistent with a continental difference in the HbS mutation rate, it fits with the broader correspondence between *de novo* rates and observations of alleles in populations that HbS is most frequent in Africans and in some other populations in the Asian malaria belt (Flint et al., 1998) and appears *de novo* in our African but not in our European samples, while Hb-Leiden has been observed across the globe (Hardison et al., 2002; Hardison and Miller, 2002) and appears *de novo* in both our African and European samples.

Third, in the African *HBB* ROI, out of 12 observable point mutations, the HbS mutation has the rate that deviates the most from the corresponding GWA rate (Table S4).

Fourth, it is striking that despite at least three independent occurrences of the HbS mutation in the *HBB* ROI, not a single case of the equivalent 20A→T mutation in the *HBD* ROI was observed in any donor, African or European. Accordingly, we note that the binary test establishing the significantly higher concentration of the 20A→T mutation in the African *HBB* ROI as opposed to all other cases (the European *HBB* ROI or the *HBD* ROIs), which is impervious to any individual- or sample-level variance including clonal dependence, suggests that the 20A→T mutation arises more frequently where it is of adaptive significance than where it is not, though data does not suffice to tell whether this effect is due to a population-level difference or due to a locus-based difference or both.

Knowing that the HbS mutation is advantageous in heterozygotes under malarial pressure, how shall we interpret these results? One possibility is that, for a reason unrelated to adaptation, some individuals have a genomic fragility in *HBB* that generates the HbS mutation at a high rate. Accordingly, it is merely a coincidence that HbS provides protection against malaria, even more

so if that fragility applies more to Africans.

Another possibility is modifier theory (Feldman and Liberman, 1986; Altenberg et al., 2017), according to which alleles affecting the mutation rate may be favored by selection under certain conditions. (Leigh Jr, 1970; Moxon et al., 1994). However, since the benefit of a modifier allele that increases the mutation rate is tied to the excess beneficial mutations it helps generate, and since mutations are rare, it is normally expected that, for selection to be effective, it must act on a modifier allele that increases the mutation rate across a long enough stretch of the genome with which it remains linked for a long enough period of time, so that many different mutations potentially induced by this allele over space and time are factored into its selective benefit (Hodgkinson and Eyre-Walker, 2011; Martincorena and Luscombe, 2013; Walsh and Lynch, 2018). Thus, modifier theory does not predict an increase in the rate of particular DNA mutations at specific base positions, let alone in sexual, complex organisms, nor the complex genetic and/or epigenetic influences on such mutation rates suggested by the current data (cf. Leigh Jr, 1970; Moxon et al., 1994; Altenberg et al., 2017; Walsh and Lynch, 2018). On the contrary, the "reduction principle"—the first-order principle in modifier theory—underscores the general difficulty of accounting for increased mutation rates (Feldman and Liberman, 1986; Altenberg et al., 2017).

Finally, a recently proposed theory predicted that mutation-specific origination rates are influenced by the complex genetic and epigenetic background, that genetic relatedness in mutational tendencies exist, and that the HbS mutation arises more frequently in Africans than in Europeans (Livnat, 2013, 2017). It holds that novelty in evolution arises from emergent interactions which are then simplified through the generations by mutational mechanisms while being checked by natural selection (Livnat, 2017), one hypothetical example being that A→I RNA editing can mechanistically increase the A→G mutation rate in the corresponding positions (cf. Popitsch et al., 2020). Based on these and other previous work (Livnat and Papadimitriou, 2016), we hypothesize that recurring, evolved processes acting on DNA and/or RNA through epigenetic modifications (Klose and Bird, 2006), RNA editing (Nishikura, 2010) and other mechanisms may lead directly to their own replacement and simplification via DNA

mutations that arise in the course of evolution from these processes' molecular nature, mechanistically linking regulatory activity with structural mutational changes—though whether and by what specific mechanism this "replacement" hypothesis explains the HbS case specifically (alternative decoding of A→I editing, Licht et al. 2019, or other mechanisms) is yet to be investigated. This raises the possibility that a mutation of adaptive value such as the HbS one need not initiate the process of adaptation but can arise later in an evolutionary process where adaptations and mutation-specific rates jointly evolve (Livnat, 2013, 2017), and thus studies on the fundamental nature of mutation need to test for not only a short-term response to environmental pressures (Luria and Delbrück, 1943; Cairns et al., 1988) but also a long-term one.

Unlike previous methods that could explore only diffuse relationships between long-term selection pressures and the evolution of GWA mutation rates, the present method offers the refined ability needed to explore such relationships, if they exist, at the mutation-specific resolution. Because this method examines the mutation-specific resolution for the first time, it provides only initial estimates of mutation rates, which will require further investigation and refinement. Furthermore, it cannot be applied currently to all mutations, because it requires a special RE for each ROI. However, given the numerous REs available and their short recognition sequences, which imply large representation of these sequences across the genome, it likely applies across many loci and organisms. Therefore, some of the most important tasks now are to examine the high-resolution mutation rate variation across additional loci of interest and to explore the molecular mechanisms responsible.

Methods

For the experimental design and different stages of library preparation, see Supplemental Text S1–S3 and Figs. S1–S3. All of the oligos for the sperm DNA library preparation described in Supplemental Text S14 were ordered from IDT with standard desalting purity, unless otherwise mentioned. All enzymes were obtained from New England Biolabs (NEB). Plasmid mini-prep, PCR purification and agarose gel extraction were carried out with QIAGEN kits.

Spike-in plasmids preparation

Four puc19-based plasmids were generated. Two (ALP13 and ALP17) were designed to carry the *HBB* genomic segment from position -203 to +223 relative to the mRNA translation start site, with the Bsu36I restriction site CCTGAGG replaced with TTATGTT and ACGAGAC, respectively; and two others (ALP16 and ALP18) were designed to carry the *HBD* genomic segment from position -59 to +220 relative to the mRNA translation start site, with the Bsu36I-restriction site replaced with TTATGTT and ACGAGAC, respectively. To prepare the spike-in mixture, the four plasmids were linearized by BamHI, mixed in equal amounts and diluted to 10 femtograms/μl for the AFR1, AFR3, AFR5, AFR6, AFR7, EUR3 and EUR4 samples and 5 femtograms/μl for all other samples.

Collection of sperm samples

Semen samples from Africans were collected in the Assisted Conception Unit of the Lister Hospital & Fertility Centre in Accra, Ghana following clinical standards, and semen samples from Europeans were purchased from Fairfax, a large US cryobank, with the approvals of the Institutional Review Board of the Noguchi Memorial Institute for Medical Research (NMIMR-IRB 081/16-17) at the University of Ghana, Legon, the Rambam Health Care Center Helsinki Committee, Haifa (0312-16-RMB) and the Israel Ministry of Health (20188768). Donors with a history of cancer or infertility or with high fever in the 3 months prior to donation were excluded. Informed consent was obtained from all participants and personal identifying information was removed and replaced with codes at the source.

DNA extraction from sperm cells

The DNA isolation protocol was modified from Weyrich (2012). A semen sample from a single donor was divided into 500μl aliquots in multiple screw-capped tubes. The sperm aliquots were washed twice with 70% ethanol to remove seminal plasma. The remaining cells were rotated overnight at 50°C in a 700μl lysis buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl2, 50 mM EDTA, 1% SDS) containing 0.5% Triton X-100 (Fisher BioReagents BP151-100), 50 mM Tris(2-carboxyethyl) phosphine hydrochloride (TCEP; Sigma-Aldrich 646547) and 1.75

mg/mL Proteinase K (Fisher BioReagents BP1700-100). Lysates were centrifuged at $21,000\times$ g for 10 minutes at room temperature and supernatants were united in a single tube. DNA purification from the cleared lysate was carried out using QIAGEN Blood & Cell Culture DNA Maxi Kit (13362). Specifically, 5 mL lysate were supplemented by 15 mL of buffer G2 (800 mM guanidine hydrochloride, 30 mM Tris-HCl pH 8.0, 30 mM EDTA pH 8.0, 5% Tween 20, 0.5% Triton X-100), vortexed thoroughly and allowed to gravity-flow through a single Genomic-tip 500/G column pre-equilibrated by 10 mL of buffer QBT (750mM NaCl, 50 mM MOPS pH 7.0, 15% isopropanol [v/v]). Resin was washed twice by 15 mL of Buffer QC (1 M NaCl, 50 mM MOPS pH 7.0, 15% isopropanol [v/v]) and elution was carried out by 15 mL of Buffer QF pre-warmed to 50°C (1.25 M NaCl, 50 mM Tris-HCl pH 7.0, 15% isopropanol [v/v]). DNA was precipitated by adding 10.5 mL room temperature isopropanol to the elute, inverting the tube 10 times, and using a sterile tip to spool and transfer the DNA to a screw-capped tube containing 500µl of buffer EB (10 mM Tris-HCl pH 8.5). The DNA was allowed to dissolve overnight at room-temperature. For each donor, a small aliquot from the extracted DNA was PCR amplified and Sanger sequenced to verify the exact sequence of the *HBB* and *HBD* regions and to confirm that the donors were homozygous for the WT sequence for both ROIs.

Enzymatic digestion

For the Bsu36I-treated sample (Supplemental Text S1–S3), ~ 264 µg sperm DNA, equivalent to 80 million haploid cells (For AFR2 a DNA amount equivalent to 60 million cells was used), were mixed with a plasmid spike-in mixture (0.2 pg for AFR1 and 0.1 pg for other donors) and equally divided in a 96-well plate. Bsu36I digestion was carried out overnight at 37°C according to the manufacturer's instructions using 5 units per well. Then, each well was supplemented by 6 units of HpyCH4III to generate the primary barcode attachment site, and digestion continued for three more hours. For the Bsu36I–untreated reaction, 13.2 µg sperm DNA (and 9.9 µg for AFR2), representing 5% of the DNA amount used for the Bsu36I digest, were mixed with 6 times the volume of plasmid spike-in mixture, aliquoted to 5 tubes and incubated overnight with 2 units SalI-HF per tube instead of Bsu36I to allow for similar conditions of DNA digestion without affecting the Bsu36I and HpyCH4III sites. Then, each well was supplemented by 6

units of HpyCH4III and digestion continued for three more hours followed by DNA purification.

Primary barcode labeling and linear amplification

Direct barcode labeling and linear amplification of the digested *HBB* and *HBD* strands were carried out in a single reaction in 96-well plates. Each well contained about 1μg of digested DNA, 0.1 μM primary barcode oligo (oligo A; see Supplemental Text S14) and 1 μM of 5'-phosphorothioate-protected primer for linear amplification (oligo B). The reaction was carried out with Q5 high-fidelity polymerase according to the manufacturer's instructions, using the following thermocycler parameters: initial denaturation at 98°C for 20 seconds, followed by 16 cycles of 98°C for 5 seconds, 68°C for 15 seconds, and 72°C for 20 seconds. For each donor, each of the Bsu36I-treated and untreated samples was labeled by an oligo A with a different Donor Identifier-1 (ID-1) sequence, which was also not shared by samples from other donors, providing each donor and each condition with a unique identifier sequence.

5'-exonuclease treatment

To eliminate non 5'-phosphorothioate-protected strands, following purification, 15μg DNA aliquots from the post linearly amplified product of the Bsu36I-treated sample were incubated each at 37°C in the presence of 15 units of Lambda exonuclease, 30 units of T7 exonuclease and 90 units of RecJF exonuclease in 1x CutSmart buffer for 2.5 hours. The post linearly amplified product of the Bsu36I-untreated sample was incubated at the same conditions with 10 units of Lambda exonuclease, 20 units of T7 exonuclease and 60 units of RecJF exonuclease.

Secondary barcode labeling and 3'-exonuclease treatment

Following purification, the DNA was aliquoted into a 96-well plate (1 μg per well). A single primer extension reaction was carried out using 0.5 μM of the secondary barcode primer (oligo C) and Q5 high-fidelity polymerase according to manufacturer's instructions. The following thermocycler parameters were used: initial denaturation at 98°C for 20 seconds, followed by a single cycle of 98°C for 5 seconds, 68°C for 15 seconds, and 72°C for 40 seconds. To remove excess oligo C, immediately after the thermocycler temperature dropped to 16°C, 20 units of thermolabile Exo I were added directly to each well together with the relabeling control primer

(oligo D) in a known amount equivalent to 0.66% of the secondary barcode primer. After incubation of one hour at 37°C, the thermolabile Exo I was heat-inactivated for one minute at 80°C and the DNA was purified. For each donor, each of the Bsu36I-treated and untreated samples was labeled by an oligo C with a different Donor Identifier-2 sequence (ID-2), which was also not shared by samples from other donors, resulting in each donor and each condition having a unique Identifier-2 sequence.

PCR amplification and sequencing

The first PCR reaction of the dual-barcode labeled product was carried out using oligo E and oligo F1 as primers and Q5 high-fidelity polymerase, according to manufacturer's instructions. The following thermocycler parameters were used: initial denaturation at 98°C for 30 seconds, followed by 10 cycles of 98°C for 5 seconds, 72°C for 15 seconds, 72°C for 30 seconds, and a final extension at 72°C for 30 seconds. Amplification products were purified and the second PCR reaction was carried out using 25% of the first PCR product as template, the amplification primers E and F2, and Q5 high-fidelity polymerase according to the manufacturer's instructions (different F2 primers were used in order to add a unique Illumina index sequence to each Bsu36I-treated and untreated sample). The following thermocycler parameters were used: initial denaturation at 98°C for 30 seconds, followed by 24 cycles (with the exception of the EUR4 sample that was amplified by 17 cycles) of 98°C for 5 seconds, 70°C for 15 seconds, 72°C for 30 seconds, and a final extension at 72°C for 1 minute. PCR products were agarose-gel purified and further concentrated by a DNA clean & concentrator kit (Zymo Research). DNA libraries prepared from the Bsu36I-treated and untreated samples of the same donor were mixed in equal amounts and paired-end sequenced with 20% PhiX by Illumina MiSeq 300 cycles kit (V2) at the Technion Genome Center (TGC). For each donor, two or three MiSeq runs were performed to reach a minimum of 10 million reads per treatment (specifically, all but AFR5 and EUR3 were sequenced two times) and the resulting FASTQ sequences were joined prior to the sequence analysis step.

Sequence analysis

Illumina paired-end (PE) reads were merged via Pear (Zhang et al., 2014) using the default model for the detection of significantly aligned regions and Phred score corrections. Merged sequences were trimmed from Illumina adapters using cutadapt (Martin, 2011), and quality-filtered by Trimmomatic (Bolger et al., 2014) using a sliding window size of 3 and a Phred quality threshold of 30. Quality filtered sequences were trimmed to remove the 5' edge up to position 18, a sequence which includes the 14 bases of the primary barcode and the 4 bases of ID-1, while adding this information to the read's header. Only sequences with the correct ID-1 and first three bases of *HBB* or *HBD* sequences were maintained. Similarly, sequences were trimmed from 9 bp at their 3' edge, which include the 5 bases of the secondary barcode and the 4 bases of ID-2, while adding this information to the read's header. Only sequences with the correct ID-2 were maintained. Trimmed sequences were sorted to *HBB* or *HBD* sequence pools, based on the occupying bases at positions 33-38 of the coding sequence (CGTTAC for *HBB* and TGTCAA for *HBD*), allowing one mismatch and frameshifts of up to -3 or +3. Successfully sorted sequences were mapped to either the *HBB* or *HBD* reference sequence (obtained by Sanger sequencing aliquots from the matching donor samples) using BWA (Li, 2013) (parameters -M -t), and high-quality mutations (Phred score $\geq$ 28) were noted. Reads were grouped by their primary barcodes to 'families' and processed according to the workflow depicted in Figure S9.

Data access

All raw sequencing data generated in this study have been submitted to the NCBI database of Genotypes and Phenotypes (dbGAP; https://www.ncbi.nlm.nih.gov/gap/) under accession number phs002391.v1.p1. For final processed data see Supplemental Datasheets and Supplemental Text S15. Software is available at GitHub (https://github.com/livnat-lab/HBB_HBD) and as Supplemental Code.

Competing interests statement

The authors declare no competing interests.

Acknowledgments

We thank Marc Feldman for comments on a previous draft, Rami Reshef for infrastructural resources, Mary Otoo and Joshua Adoboe for help with sample collection, Sara Zelig, Alan Templeton and Nick Pippenger for technical comments, and Kim Weaver for extensive help. This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Author contributions: DM and AL invented the method and designed the studies; DM performed all experiments except for RS's; RS processed the EUR4 sample; YN created software tools for data analysis; AM created the computational pipeline for mutation calling; EB improved the pipeline; YN and AL provided statistical tools; MY, EH, KS and AL obtained IRB and Helsinki approvals; MY and EH collected samples; DM, YN, EB, AM and AL analyzed the results; DM and AL drafted the paper; DM, YN, KS and AL revised the draft; KS provided general advice; KS and AL acquired funding; AL conceived of the project and the replacement hypothesis and supervised the project.

References

Abascal F, Harvey LM, Mitchell E, Lawson AR, Lensing SV, Ellis P, Russell AJ, Alcantara RE, Baez-Ortega A, Wang Y, et al.. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**: 405–410.

Allison AC. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. *BMJ Brit Med J* **1**: 290–294.

Altenberg L, Liberman U, and Feldman MW. 2017. Unified reduction principle for the evolution of mutation, migration, and recombination. *Proc Natl Acad Sci USA* **114**: E2392–E2400.

Arbeithuber B, Makova KD, and Tiemann-Boege I. 2016. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res* **23**: 547–559.

Blake R, Hess ST, and Nicholson-Tuell J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol* **34**: 189–200.

Bolger AM, Lohse M, and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes. *Mol Biol Evol* **3**: 322–329.

Cairns J, Overbaugh J, and Miller S. 1988. The origin of mutants. *Nature* **335**: 142–145.

Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ, Sudmant PH, Shendure J, et al.. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**: 1277–1281.

Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ, et al.. 2018. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**: 3753.

Carter R and Mendis K. 2002. Evolutionary and historical aspects of the burden of malaria. *Clin Microbiol Rev* **15**: 564–94.

Cavalli-Sforza LL and Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* **33**: 266–275.

CDC Division of Parasitic Diseases & Malaria. 2019. "Where Malaria Occurs," http://www.cdc.gov/malaria/about/distribution.html Accessed 1/24/2019.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al.. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712.

Crow KD, Amemiya CT, Roth J, and Wagner GP. 2009. Hypermutability of *HoxA13A* and functional divergence from its paralog are associated with the origin of a novel developmental feature in zebrafish and related taxa (Cypriniformes). *Evolution* **63**: 1574–1592.

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney J, Jackson J, Sikela M, Raznahan A, Giedd J, et al.. 2012. Duf1220-domain copy number implicated in human brain-size pathology and evolution. *Am J Hum Genet* **91**: 444–454.

Ellegren H, Smith NG, and Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* **13**: 562–568.

Feldman MW and Liberman U. 1986. An evolutionary reduction principle for genetic modifiers. *Proc Natl Acad Sci USA* **83**: 4824–4827.

Feng Z, Smith D, McKenzie F, and Levin S. 2004. Coupling ecology and evolution: malaria and the S-gene across time scales. *Math Biosci* **189**: 1–19.

Flint J, Harding RM, Boyce AJ, and Clegg JB. 1998. The population genetics of the haemoglobinopathies. *Baillière's Clin Haem* **11**: 1–51.

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Van Duijn CM, Swertz M, Wijmenga C, Van Ommen G, et al.. 2015. Genome-wide patterns and properties of *de novo* mutations in humans. *Nat Genet* **47**: 822.

Gojobori T, Li WH, and Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol* **18**: 360–369.

Goldmann JM, Wong WS, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LE, Hoischen A, Roach JC, et al.. 2016. Parent-of-origin-specific signatures of *de novo* mutations. *Nat Genet* **48**: 935.

Gu W, Zhang F, and Lupski JR. 2008. Mechanisms for human genomic rearrangements. *PathoGenetics* **1**: 4.

Gu X and Li WH. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *Journal of Molecular Evolution* **40**: 464–473.

Haldane JBS. 1949. The rate of mutation of human genes. *Hereditas* **35**: 267–273.

Hardison R and Miller W. 2002. "Globin Gene Server," http://globin.cse.psu.edu/ Accessed 10/5/2019.

Hardison RC, Chui DH, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W, and Wajcman H. 2002. HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Hum Mutat* **19**: 225–233.

Harpak A, Bhaskar A, and Pritchard J. 2016. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genet* **12**: e1006489.

Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci USA* **112**: 3439–3444.

Harris K and Pritchard JK. 2017. Rapid evolution of the human mutation spectrum. *Elife* **6**.

Hartl DL and Clark AG. 2007. *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts, 4th edition.

Hodgkinson A and Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756–766.

Hodgkinson A, Ladoukakis E, and Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol* **7**.

Hwang DG and Green P. 2004. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* **101**: 13994–14001.

Ingram V. 1957. Gene mutations in human hemoglobin: The chemical difference between normal and sickle hemoglobin. *Nature* **180**: 326–328.

Inoue K, Dewar K, Katsanis N, Reiter LT, Lander ES, Devon KL, Wyman DW, Lupski JR, and Birren B. 2001. The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. *Genome Res* **11**: 1018–1033.

Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, and Nudler E. 2016. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534**: 693.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA, et al.. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**: 2586.

Klose RJ and Bird AP. 2006. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* **31**: 89–97.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**: 12–27.

Kratochwil CF, Liang Y, Urban S, Torres-Dowdall J, and Meyer A. 2019. Evolutionary dynamics of structural variation at a key locus for color pattern diversification in cichlid fishes. *Genome Biol Evol* **11**: 3452–3465.

Kratochwil CF and Meyer A. 2019. Fragile DNA contributes to repeated evolution. *Genome Biol* **20**: 39.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* **77**: 171–192.

Leigh Jr EG. 1970. Natural selection and mutability. *Am Nat* **104**: 301–305.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al.. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Lercher MJ, Williams EJ, and Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol* **18**: 2032–2039.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997 .

Licht K, Hartl M, Amman F, Anrather D, Janisiw MP, and Jantsch MF. 2019. Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res* **47**: 3–14.

Lind PA. 2019. Repeatability and predictability in experimental evolution. In *Evolution, Origin of Life, Concepts and Methods* (ed. P Pontarotti), pp. 57–83. Springer International Publishing.

Livnat A. 2013. Interaction-based evolution: how natural selection and nonrandom mutation work together. *Biol Direct* **8**: 24.

Livnat A. 2017. Simplification, innateness, and the absorption of meaning from context: how novelty arises from gradual network evolution. *Evol Biol* **44**: 145–189.

Livnat A and Papadimitriou C. 2016. Evolution and learning: used together, fused together. A response to Watson and Szathmáry. *Trends Ecol Evol* **31**: 894–896.

Losos JB. 2017. *Improbable Destinies: Fate, Chance, and the Future of Evolution*. Penguin, New York.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**: 417–422.

Luria SE and Delbrück M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**: 491.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107**: 961–968.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**: 10–12.

Martincorena I and Luscombe NM. 2013. Non-random mutation: the evolution of targeted hypermutation and hypomutation. *Bioessays* **35**: 123–130.

Matassi G, Sharp PM, and Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol* **9**: 786–791.

Mathieson I and Reich D. 2017. Differences in the rare variant spectrum among human populations. *PLoS Genet* **13**: e1006581.

McClellan J and King MC. 2010. Genetic heterogeneity in human disease. *Cell* **141**: 210–217.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al.. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**: 1431–1442.

Moxon ER, Rainey PB, Nowak MA, and Lenski RE. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4**: 24–33.

Nachman MW and Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.

Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**: 321–349.

Pauling L, Itano HA, Singer SJ, and Wells IC. 1949. Sickle-cell anemia, a molecular disease. *Science* **110**: 543–548.

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, Weatherall DJ, and Hay SI. 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun* **1**: 104.

Popitsch N, Huber CD, Buchumenski I, Eisenberg E, Jantsch M, Von Haeseler A, and Gallach M. 2020. A-to-I RNA editing uncovers hidden signals of adaptive genome evolution in animals. *Genome Biol Evol* **12**: 345–357.

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, et al.. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet* **48**: 126.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al.. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**: 636–639.

Salk JJ, Schmitt MW, and Loeb LA. 2018. Enhancing the accuracy of Next-Generation Sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**: 269.

Shendure J and Akey JM. 2015. The origins, determinants, and consequences of human mutations. *Science* **349**: 1478–1483.

Steinberg M and Adams JI. 1991. Hemoglobin A2: origin, evolution, and aftermath. *Blood* **78**: 2165–2177.

Veltman JA and Brunner HG. 2012. *de novo* mutations in human genetic disease. *Nat Rev Genet* **13**: 565–575.

Vogel F and Motulsky A. 1997. *Human genetics: problems and approaches*. Springer-Verlag, Berlin.

Walsh B and Lynch M. 2018. *Evolution and Selection of Quantitative Traits*. Oxford University Press, Oxford, UK.

Weyrich A. 2012. Preparation of genomic DNA from mammalian sperm. *Curr Protoc Mol Biol* **98**: 2–13.

Wolfe KH, Sharp PM, and Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285.

Xie KT, Wang G, Thompson AC, Wucherpfennig JI, Reimchen TE, MacColl AD, Schluter D, Bell MA, Vasquez KM, and Kingsley DM. 2019. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**: 81–84.

Zhang F, Carvalho CMB, and Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. *Trends Genet* **25**: 298–307.

Zhang J, Kobert K, Flouri T, and Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**: 614–620.

Zhang Z and Gerstein M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31**: 5338–5348.