Thousands of human mutation clusters are explained by short-range template switching

Ari Löytynoja

Institute of Biotechnology, University of Helsinki, FI-00014 Helsinki, Finland

Variation within human genomes is unevenly distributed, and variants show spatial clustering. DNA replication–related template switching is a poorly known mutational mechanism capable of causing major chromosomal rearrangements as well as creating short inverted sequence copies that appear as local mutation clusters in sequence comparisons. In this study, haplotype-resolved genome assemblies representing 25 human populations and multinucleotide variants aggregated from 140,000 human sequencing experiments were reanalyzed. Local template switching could explain thousands of complex mutation clusters across the human genome, the loci segregating within and between populations. During the study, computational tools were developed for identification of template switch events using both short-read sequencing data and genotype data, and for genotyping candidate loci using short-read data. The characteristics of template-switch mutations complicate their detection, and widely used analysis pipelines for short-read sequencing data, normally capable of identifying single nucleotide changes, were found to miss template-switch mutations of tens of base pairs, potentially invalidating medical genetic studies searching for a causative allele behind genetic diseases. Combined with the massive sequencing data now available for humans, the novel tools described here enable building catalogs of affected loci and studying the cellular mechanisms behind template switching in both healthy organisms and disease.

[Supplemental material is available for this article.]

Twenty years after the publication of the draft genomes (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), the Telomere-to-Telomere Consortium reported the first truly complete sequence of a human genome (Aganezov et al. 2022; Nurk et al. 2022). However, every genome is unique, and the challenge now is to understand the global genomic diversity in the human population (Karczewski et al. 2020; Taliun et al. 2021) and to build a comprehensive pangenome to represent this variation (Miga and Wang 2021). Whereas single nucleotide variation (SNV) and short insertion-deletions (indels) have been resolved relatively accurately with classical short-read sequencing (Bentley et al. 2008), the characterization of structural variants (SVs) and low-complexity sequences has been lacking. The Human Genome Structural Variation Consortium developed a method (Porubsky et al. 2021) for phased diploid genome assembly with a combination of long-read Pacific Biosciences (PacBio) whole-genome sequencing (WGS) (Eid et al. 2009) and Strand-seq phasing (Falconer and Lansdorp 2013) data. Ebert et al. (2021) applied this method to 32 diverse human individuals and produced 64 assembled haplotypes, that is, maternal and paternal copies of the genome. With the help of phased diploid genomes, they massively expanded the catalog of known SVs and, having the information of coinherited nearby sequence differences, could study in detail the different subclasses of complex variants. Such accurate characterization of multiple genomes will lay the foundation for understanding the role of complex mutations in human phenotypes and disease (Zhang et al. 2009; Weischenfeldt et al. 2013; Sakamoto et al. 2020) and in evolution (Perry et al. 2007; Zhang et al. 2009; Yan et al. 2021). More generally, the unprecedented resolution of the new genome data will open novel possibilities for understanding the mutational mechanisms of the complex eukaryotic genomes (Conrad et al. 2011; Veltman and Brunner 2012; Ségurel et al. 2014).

Although many subclasses of variants studied by Ebert et al. (2021) are clearly independent, mutations are known to be clustered (Ségurel et al. 2014), and their separation at the analysis stage may dismiss crucial information. We compared earlier two independent assemblies of the human genome (International Human Genome Sequencing Consortium 2004; Levy et al. 2007) and found mutation clusters consistent with DNA-replication related template switching (Löytynoja and Goldman 2017). The mechanism, originally described and studied in bacteria (Ripley 1982; Dutra and Lovett 2006), resembles FoSTeS (Fork Stalling and Template Switching) (Lee et al. 2007) and MMBIR (Microhomology-Mediated Break-Induced Replication) (Hastings et al. 2009) but is local, and reciprocal switches typically occur within a region of a few tens or hundreds of base pairs (bp). Unlike the chromosomal rearrangements created by FoSTeS and MMBIR, the footprint of local template switch mutations (TSMs) are short inversions, either happening in place or creating a reverse complement copy of a nearby sequence region; these are believed to be produced by the replication briefly switching either to the complementary DNA strand or going backwards along the nascent DNA strand (Supplemental Fig. S1; Ripley 1982; Dutra and Lovett 2006). In sequence comparisons, mutations compatible with the TSM mechanism appear as multiple nearby sequence changes and may be represented as combinations of SNPs, MNPs (single- and multinucleotide polymorphisms) and short indels in variant data. We showed earlier (Löytynoja and Goldman 2017) that the

Corresponding author: ari.loytynoja@helsinki.fi
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276478.121.

TSM candidates identified between the two assemblies segregate in the 1000 Genomes population data; that TSM-like mutation patterns can be identified in genotype data; and that parent-offspring trios (Besenbacher et al. 2016) contain consistent de novo mutations (DNMs). Here, I revisit the topic using the haplotype-resolved data of Ebert et al. (2021) and the variation information compiled by massive sequencing studies (e.g., Karczewski et al. 2020; Taliun et al. 2021) and study the TSM mechanism's role in generation of genomic variation. As haplotype-resolved genome data are still rare, I assess how reliably the TSM patterns found with de novo–assembled genomes can be identified using traditional short-read DNA sequencing.

Results

TSMs explain thousands of haplotypic mutation clusters

The SNVs, indels, and SVs of the variant data of Ebert et al. (2021) (called here "HaplotypeSV data") (Table 1) were combined, and for each maternal and paternal genome copy, the clusters of sequence differences were identified (see Methods). For each mutation cluster locus within a haplotype, the identified variants were placed in their genomic background, creating two alleles with 150 bp of sequence context (Fig. 1A,B). Finally, having the two alleles—the region from the GRCh38 reference (called "REF") and the alternative allele formed by the mutation cluster (called "ALT")—the four-point alignment (FPA) algorithm (Löytynoja and Goldman 2017) was applied to find the best solution involving a template switch, testing both alleles as the ancestral state (Fig. 1C,D; see Methods). Despite attempts to evaluate TSMs under a probabilistic model (Walker et al. 2021), it is unclear when a potential TSM solution should be considered more plausible than a combination of SNVs and indels. Here, the TSM candidates were required to (1) be supported by at least two sequence changes, of which at least one is a base change, and as a result show a higher sequence identity across the whole region than the original forward-aligned solution; and (2) have a ②→③ region, inferred to be copied from the other template, of at least 8 bases long (Supplemental Fig. S1). Despite strong evidence of many VNTRs (variable number of tandem repeats) evolving through template switching (Supplemental Fig. S2), the candidate ②→③ regions were required to contain all four bases, thus removing hits within low-complexity sequences.

Per individual, 3049–3934 mutation patterns consistent with the TSM mechanism were identified, the longest of them 318 bp in length (Table 2, "ALL"). Many of these were in proximity to repeat elements and, due to mismapping of sequencing reads, could potentially be nonindependent events counted multiple times. After removing the loci masked as repetitive sequence in the reference genome, there were 872–1121 high-quality TSM events per individual (Fig. 1E), the maximum length varying from 97–180 bp and the mean length being 11.56 bp (Table 2, "UNMASKED"; Supplemental Fig. S3). Homozygosity of inferred TSM loci varied from 26.7%–42.3% (Fig. 1E), and the proportion of singletons from 2.1%–9.0% (Supplemental Table S1). Although comparisons at the superpopulation level are affected by the different sample sizes, African (AFR) individuals expectedly had the lowest level of homozygosity and fixed loci, 29.2% and 0.7%, and the highest proportion of singleton loci, 7.8% (The 1000 Genomes Project Consortium 2015), whereas those for Europeans (EUR, including Azkenazi NA24385) and Asians (AAS, including South and East Asians) were higher and lower, respectively (Table 2). The greater variation of African populations and the old age of the TSM patterns were reflected in the sharing of loci across the superpopulations: AFR shared 45.6% and 49.9% of its TSM loci with EUR and AAS, respectively, whereas the latter two shared 75.3%–80.7% of their TSM loci with other superpopulations (Table 2). A unique feature of TSMs is that often alternative alleles can be polarized, that is, have the ancestral and derived state determined, without an outgroup (Supplemental Fig. S4). The ancestral allele could be defined in 87.7%–90.3% of the cases and the proportion of loci with the ALT variant being the ancestral allele was found ranging from 49.7–55.3 across the individuals (Fig. 1E; Supplemental Table S1). At the superpopulation level, Europeans had the highest average proportion of ancestral variants (Table 2), probably reflecting the European origin of the GRCh38 reference genome and thus capturing a greater amount of derived European TSM variation.

Considering the candidate TSMs in the unmasked part of the reference genome, there were in total 2200 unique loci. Whereas the annotation revealed 96.2% of the mutation loci to be intergenic or intronic (Supplemental Table S2), 24 loci were inferred as coding. On closer look at three cases that had the ②→③ region of at least 12 bp long, segregation of the ALT variants revealed one of these as a false positive, but the two others seemed real. The inferred TSM events created a cluster of linked base substitutions and indels in the variant data and appeared at the lowest possible frequency, heterozygous in a single individual; the mutations resulted in an early stop codon in an alternatively spliced exon of gene *ANP32E* (Fig. 2) and changes in a nonsense-mediated decay destined transcript of gene *PHF21A*.

Table 1. Description of the data sets analyzed

Data set	Source	Type	Platform	Read length	#Samples	#Variants
HaplotypeSV	Ebert et al. 2021	VCF	PacBio + Strand-seq	NA	32	18,241,436
Platinum NA12878	Eberle et al. 2017	VCF	Illumina	100 PE	1	4,167,900
Chromium NA12878	Weisenfeld et al. 2017	VCF	Illumina Linked-Reads	150 PE	1	4,898,662
1KGP phase 3 GRCh38	1000G Consortium 2015	VCF	NA	NA	2504	80,000,000+
gnomAD_MNV full	Wang et al. 2020	TSV	NA	NA	125,000+	1,792,248
gnomAD_MNV 3+	Wang et al. 2020	TSV	NA	NA	125,000+	91,300
second_gen.dnms	Sasani et al. 2019	TSV	Illumina	150 PE	70	4671
third_gen.dnms	Sasani et al. 2019	TSV	Illumina	150 PE	350	24,975
Platinum NA12878	Eberle et al. 2017	BAM	Illumina	100 PE	1	50x
Platinum CEPH 1463	Eberle et al. 2017	BAM	Illumina	100 PE	17	50x
Chromium NA12878	Weisenfeld et al. 2017	BAM	Illumina Linked-Reads	151 PE	1	56x
Chromium NA1289[12]	Marks et al. 2019	BAM	Illumina Linked-Reads	151 PE	2	~30x
PacBio GIAB NA12878	PacBio (PRJNA540705)	BAM	PacBio CCS Sequel II	NA	1	~30x

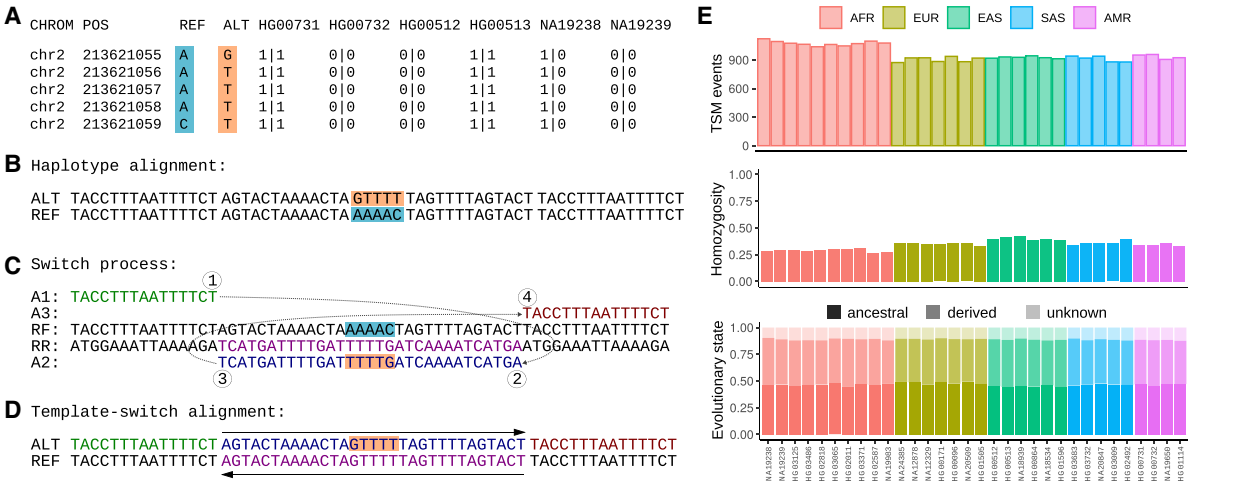

Figure 1. TSMs in HaplotypeSV data. (*A*) HaplotypeSV data contain a cluster of five SNVs that segregate as a unit among individuals (the first six are shown). (*B*) Within their genomic context, REF (cyan) and ALT (orange) alleles create two haplotypes. (*C*) The FPA algorithm attempts to explain the observed differences with two reciprocal template-switch events. In this case, the ALT sequence (A1, A2, A3; shown in green, blue, and red) can be created from the REF sequence by copying the A2 fragment in reverse-complement (RR; shown in magenta). (*D*) The template-switch solution fully explains the cluster of five base differences. (*E*) Total numbers of inferred TSM events in different HaplotypeSV samples after removal of low-complexity sequences (*top*); proportion of homozygous loci (*middle*); and proportion of ancestral (dark), derived (intermediate), and unpolarized (light) alleles (*bottom*).

Known TSM loci can be genotyped with short-read data

One individual in the HaplotypeSV data, NA12878, is the mother of a three-generation 17-member pedigree (CEPH 1463) sequenced to high coverage (Eberle et al. 2017) as well as one of the reference samples of the Genome in a Bottle (GIAB) initiative (Zook et al. 2016). The TSM loci identified in NA12878 were genotyped in different GIAB data sets and in the pedigree 1463 to assess their de novo versus standing, in vitro versus in vivo, and somatic versus germline origin. Using HaplotypeSV genotype data for NA12878, the genome contexts for the alternative alleles were reconstructed and, for each individual in the family, the reads extracted from that particular locus were mapped against the two alternative alleles. Outside the mutation locus, short reads should map evenly between the two alleles, but reads that overlap the positions differing between the alleles are expected to map to one allele only (Fig. 3A,B). By comparing the read coverage across the two alternative alleles, the test individual could be identified as homozygote for REF allele (R|R), heterozygote (R/A), or homozygote for ALT allele (A|A). A strategy similar to base calling was adopted and a 99% posterior probability was required for credible genotype calls (see Methods; Supplemental Fig. S5).

Starting with the Illumina Platinum (Eberle et al. 2017) and 10x Genomics Chromium (Weisenfeld et al. 2017; Marks et al. 2019) data sets for NA12878 (Table 1), and the TSM-like loci found in the HaplotypeSV data were genotyped. Focusing on the 864 loci that were not masked as repeats or low-complexity sequences and that had enough read coverage in both data sets, the two short-read data sets were found to provide the genotype and agree on it in 831 cases (Fig. 4A; Supplemental Table S3). Whereas 14 loci could not be genotyped with the short-read data using the 99% posterior probability cutoff, the variant pattern in the HaplotypeSV data was inconsistent in 84 cases, and the genotype of the loci could not be determined. Of the 780 loci inferred to be either heterozygous or homozygous for the ALT allele in the

Table 2. TSM statistics for HaplotypeSV data

Superpopulation		AFR	EUR	AAS
Haplotypes		**20**	**14**	**22**
ALL	TSMs/individual	3783	3135	3173
	Unique TSMs	12,501	7370	8956
	Maximum length	318	188	188
	Average length	13.14	13.14	13.15
UNMASKED	TSMs/individual	1074	906	919
	Unique TSMs	3133	1899	2246
	Maximum length	180	180	180
	Average length	10.61	10.56	10.55
	Singletons	7.8%	2.7%	3.2%
	Fixed	0.7%	1.8%	2.8%
	Homozygous	29.2%	35.3%	38.5%
	Ancestral	52.0%	53.9%	51.5%
	Shared with AFR	–	75.3%	69.6%
	Shared with EUR	45.6%	–	68.2%
	Shared with AAS	49.9%	80.7%	–

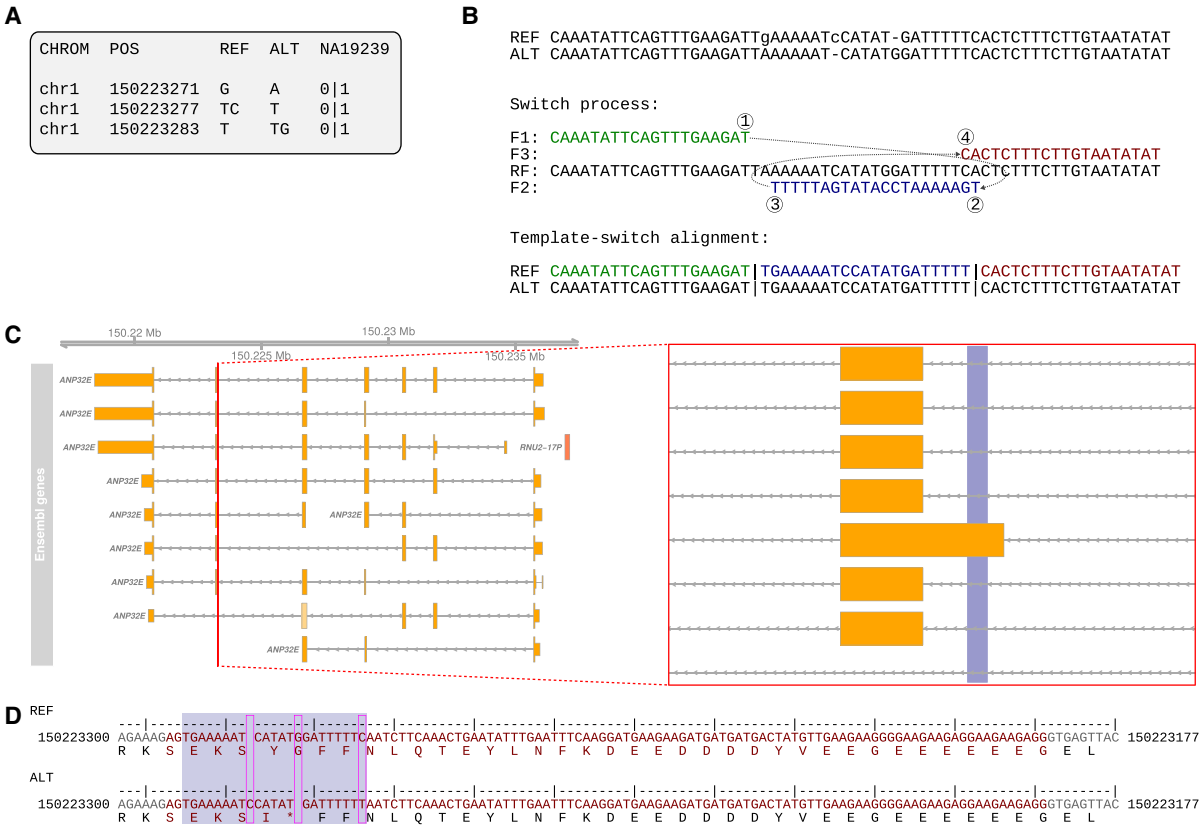

Figure 2. TSM causing an early stop codon. (*A*) The three variants included in the HaplotypeSV data appear together in one individual, NA19239. (*B*) All changes are explained by a template-switch event. (*C*) The affected sites are within an alternatively spliced exon of gene *ANP32E*, highlighted in blue. (*D*) The changes, shown in magenta, alter the reading frame and cause an early stop codon (*bottom*). Coding sites are shown in red.

HaplotypeSV data, the three data sets agreed in 735 cases (Supplemental Table S3). Eight loci recorded variable in the HaplotypeSV data were inferred to be homozygous for the REF allele in both short-read data sets. Many of these were in a repetitive sequence context and all loci were found to segregate among other samples in the HaplotypeSV data, ruling out the possibility of DNMs. Although the reasons for inconsistent HaplotypeSV genotypes could not be resolved, 74 of 84 cases showed intermediate depth ratios in short-read data, that is, r was not $\tilde{1}$, $\ll 1$, or $\gg 1$ (Fig. 4A; orange symbols around −0.5, −0.5). If these were de novo mutations, one would expect the ratios to be reverse, and a more likely explanation is cross-mapping from duplicated sequence blocks. A small number of candidate TSM loci showed inconsistency between different data sets, possibly due to similarly incorrectly mapped reads (Supplemental Fig. S6).

To further verify the genotyping approach, a similar analysis was performed on the parents and children of NA12878. On the parents, the 783 loci found in the HaplotypeSV data were genotyped and called to contain at least one ALT allele in the NA12878 Chromium data. The inferred genotypes showed strong agreement, and, of the 307 loci called homozygous for the ALT allele in NA12878, 290 could be genotyped with Chromium data of both parents and each of them contained at least one ALT allele (Fig. 4B; Supplemental Table S4). The only locus that appeared as a novel mutation according to parental data (R|R:R|R in Fig. 4B) was a false positive, and in manual verification, one parent was found to contain two supporting reads. Genotyping of the loci in the 11 children (CEPH pedigree 1463) required their transfer from GRCh38 to GRCh37. The lift-over worked and the parents' (NA12877 and NA12878) alignment data (Platinum/GRCh37) contained enough reads to genotype 843 of the TSM-like loci found variable in NA12878 in the HaplotypeSV data. Except for loci that were inferred to be heterozygous in both parents (R/A, R/A), the observed genotypes matched nearly perfectly the expectation under the Mendelian segregation ratio (Table 3). An excess of heterozygotes in genotype data is a classic mark of artifact caused by sequencing reads originating from different loci, for example, due to duplication of genome regions. Seeing those among the CEPH 1463 children suggests that a proportion of inferred variants for NA12878 in the HaplotypeSV data are erroneous: The TSM events causing those variants may be real, but they have happened in duplicated copies, not within the loci where the calls were made.

Novel TSM loci can be discovered with short-read data

Many of the TSM loci detected in the HaplotypeSV data had not originally been called with the short-read data, despite the two alleles differing over several tens of bases. A closer look at these revealed that alternative TSM haplotypes can show nearly uniform sequencing coverage across the region if the ②→③ region of the inferred TSM process (Fig. 3A) is long enough: The mapping algorithm then produces a decent mapping for the read core by clipping the mismatching overhangs (Fig. 3C). Although this is

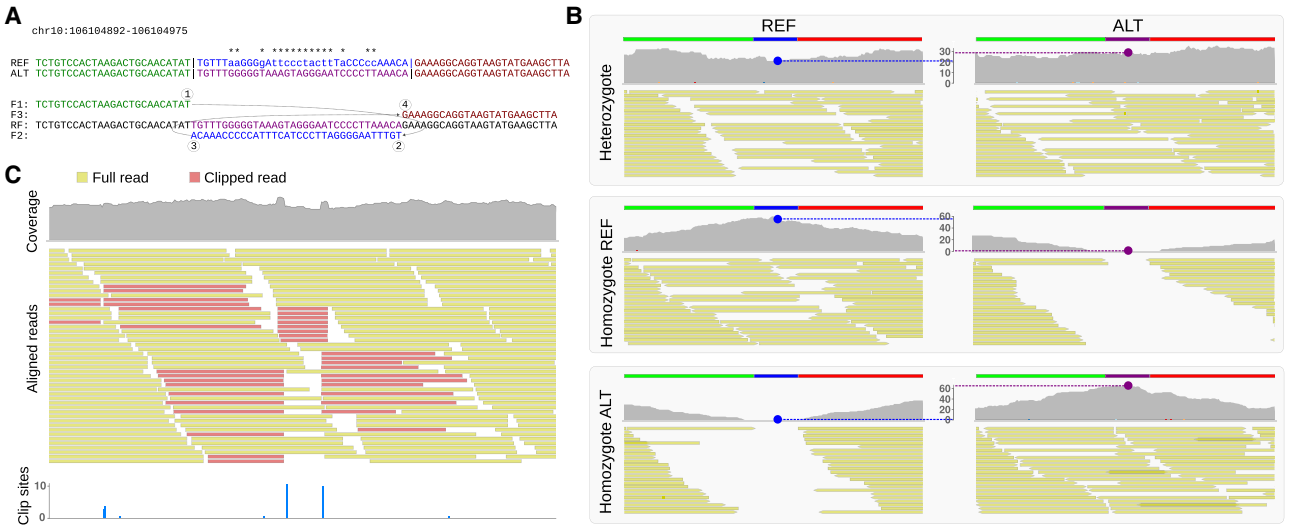

Figure 3. Genotyping of a TSM locus. (*A*) TSM solution for a complex mutation on Chromosome 10. The REF and ALT alleles share the *left* (green) and *right* (red) flanking region but have a different central part (blue and magenta). The differences in the central part are explained by a TSM event (*below*). (*B*) The two haplotypes with alternative central parts (blue and magenta bars) are used as the reference for remapping of reads extracted from the locus. In a heterozygous individual (*top*), reads are mapped evenly to the two alleles, giving uniform coverages (gray). In individuals homozygous for the REF allele (*middle*) and ALT allele (*bottom*), reads covering the central part are mapped predominantly to one allele only. Genotype is inferred of the mapping coverage in the middle of the locus (blue and magenta dots and dotted lines) using a function similar to that for genotype calling on nucleotides. (*C*) NA12878 Platinum data show fairly normal mapping coverage (gray, *top*), but a closer look at the reads (full and clipped reads in yellow and red, respectively) reveals an anomaly. The cluster of clip sites (blue; *bottom*) allows computational identification of the locus.

clearly undesirable behavior, the observation suggested a strategy to identify potential TSM loci in short-read data by searching for clusters of clip sites, extracting the reads mapped to the region, and then producing a local reassembly of the reads; in positive cases, the dissimilarities between the assembled sequences could be explained by a TSM event.

To test that, using SvABA, a tool developed for the detection of structural variants by local reassembly of short sequencing reads (Wala et al. 2018), the FPA algorithm was integrated (Löytynoja and Goldman 2017) to search for TSM patterns in the resulting contigs. Clusters of clip sites were identified in short-read data and clusters of variants in the corresponding variant data, performing the analysis independently on the Illumina Platinum (Eberle et al. 2017) and 10x Genomics Chromium (Weisenfeld et al. 2017) resources for NA12878. SvABA-FPA was then applied on each candidate locus (see Methods), and 1054 and 26,222 candidate TSM loci were found in the Platinum and Chromium data, respectively. Of these, 211 and 755 loci, respectively, passed the sequence-based filtering and 203 and 299 loci the subsequent mapping depth-based genotyping (being either heterozygous or

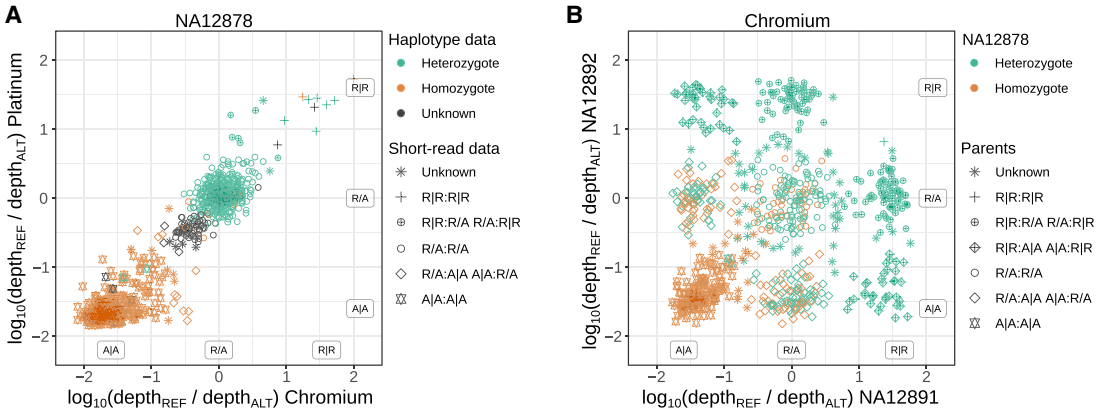

Figure 4. Confirmation of HaplotypeSV TSM loci with short-read data. (*A*) Ratio of REF and ALT allele mapping coverage ($r = depth_{REF}/depth_{ALT}$) reflects the genotype: $r \approx 1$ (and thus $log_{10}(r) \approx 0$) for a heterozygote; $r \ll 1$ and $r \gg 1$ for the two types of homozygotes. The $\log_{10}$ ratios agree for the NA12878 Platinum and Chromium data sets, and 95.5% of the 883 loci identified in HaplotypeSV data are called to contain at least one ALT allele with 99% posterior probability in both data sets; 17 and 10 loci are called homozygous REF with at least one data set and by both (*top-right* corner). (*B*) Of the 783 variable loci in NA12878 Chromium data, all but one locus are called to contain at least one ALT allele in the parents. In the 290 loci homozygous for ALT in NA12878 (orange), both parents contain at least one ALT allele. "Unknown" indicates inconsistent variant data or posterior probability below 0.99. One pseudocount was added to all values to avoid divisions by zero. In the legend, the inferred genotypes for the two data sets are separated by a colon, and X_1/X_2 represents the alternative arrangements of the two alleles, $X_1|X_2$ and $X_2|X_1$.

Table 3. Genotypes (%) of TSM loci in CEPH 1463 children

Parental genotypes	n	Offspring expected			Offspring observed			
		R|R	R/A	A|A	R|R	R/A	A|A	NA[a]
R|R,R|R	7	100	0	0	72.7	24.7	0	2.6
R|R,R/A	169	50	50	0	48.6	50.5	0.5	0.4
R|R,A|A	37	0	100	0	0	98.5	1.5	0
R/A,R/A	214	25	50	25	16.8	63.5	17.5	2.2
R/A,A|A	206	0	50	50	0.4	53.0	45.1	1.6
A|A,A|A	177	0	0	100	0	0.5	98.5	1.1

[a]Loci with posterior probability below 0.99 considered as unknown.

homozygous for the ALT allele) (see Supplemental Fig. S7; Methods). The fact that 456 of the 755 Chromium hits passing the first stage of filtering could not be confirmed with read mapping indicates that the alternative haplotypes were created by a small number of reads. Whether these reads had originally been misaligned or represent low-frequency TSM mutations, possibly originating in vitro, could not be confirmed with the current data; similar hits not found in the Platinum data may be explained by different read lengths (Table 1) or technological differences between the standard and linked-reads Illumina sequencing and downstream bioinformatic methods used (Li 2013; Marks et al. 2019). A majority of the loci passing all filtering, 89.7% and 70.6% for Platinum and Chromium, respectively, were shared either with the other short-read approach or with the HaplotypeSV set (Fig. 5). On the other hand, the large number of candidate loci, 690 of the total 843, identified uniquely with the HaplotypeSV data suggests that a great majority of the loci do not contain misaligned and soft-clipped reads: if no variants were called at those loci in the Platinum and Chromium sets, no signal was present to include the loci for the short-read-based TSM search.

Fifty-nine cases that had the ②→③ fragment of at least 25 bases long were studied in more detail (Supplemental Data S1). Most of the cases showed the expected signal in the original alignment data that was removed when the same reads were mapped against the two alternative haplotypes (Fig. 6A,B), whereas in a few rare cases the variant calling had correctly captured the differences between the two TSM haplotypes (Fig. 6C,D). Some of the heterozygous loci were not called in any variant call set despite significant differences between the two haplotypes (Supplemental Fig. S8; Supplemental Data S1), and several loci homozygous for the ALT allele were uncalled in different variant call sets, including the HaplotypeSV data (Supplemental Fig. S9). As an example of the latter, 67 TSM candidates (Supplemental Fig. S7D) found in and confirmed by both Platinum and Chromium data were not present in the HaplotypeSV set. A closer look was taken at the 16 loci with the ②→③ fragment of at least 25 bases long, examining the loci in PacBio long-read sequencing data and 1000 Genomes variation data (Table 1; Supplemental Data S2; The 1000 Genomes Project Consortium 2015). All the studied cases were confirmed to be genuine mutation clusters by the PacBio data, and two types of errors in HaplotypeSV calls were observed: (1) The representation of the mutation cluster was incomplete, thus not allowing correct reconstruction of the alternative allele and discovery of the TSM event (Supplemental Fig. S10A); and (2) complete lack of variant calls (Supplemental Fig. S10B). In an extreme, the variants were missing from HaplotypeSV data despite being called in full for NA12878 in the 1000 Genomes data and clearly supported by all sequencing data (Supplemental Fig. S10C).

TSMs explain thousands of clusters in variation databases

A significant fraction of the TSM candidate loci were missing from the original variant call sets (Fig. 6C)—16 of the 59 closely studied loci were missing from both sets (Supplemental Data S1)—but none of the manually confirmed cases supported by both short-read data sets was completely absent from the dbSNP database (Sherry et al. 2001). Although dbSNP correctly lists many studied loci as multinucleotide variants (MNVs), all loci were also present as multiple SNVs and indels and some only as multiple independent MNVs or indels (Supplemental Fig. S11; Supplemental Data S1). The latter is slightly surprising as variants originating from a complex mutation are expected to be present in their entirety or not at all, and similar allele frequencies should allow phasing and grouping independent variants into MNVs (Choi et al. 2018).

The Genome Aggregation Database (gnomAD, v2.1.1) (Karczewski et al. 2020) contains variation found among 125,000 exomes and 15,700 genomes of unrelated individuals sequenced as part of various disease-specific and population genetic studies. Wang et al. (2020) identified gnomAD SNVs appearing within a proximity of 1–10 bp and provides them as phased variant pairs (Table 1). Using these data, variant pairs with similar allele counts (ACs; $\Delta AC \leq 10\%$) were selected, the data of different variant-pair distances were merged, and for all clusters of at least three variants, the alternative alleles were created, and a search was made for a TSM solution to explain the differences (see Methods). Among the 91,300 clusters of three or more SNVs, a TSM solution was found for 4425 loci. Among these, a closer look was taken at the 192 cases where the ②→③ region was at least 25 bp long and the REF and ALT alleles differed at least by four SNVs (Supplemental Data S3).

Among the studied cases, the TSM mechanism perfectly explained MNVs consisting of up to 15 base changes (Fig. 7A,B) whereas the longest inferred TSM events had the ②→③ region of over a hundred bases (Supplemental Fig. S12). Consistent with previous findings, 36% of the TSM loci studied contained low

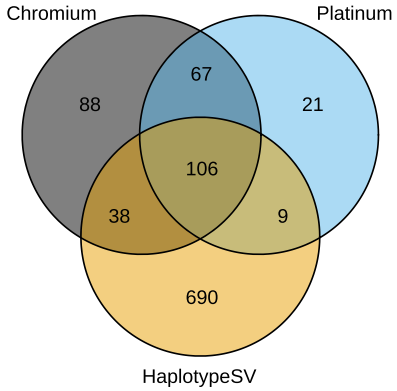

Figure 5. The number of TSM candidate loci identified in different data. The diagram shows the overlap of TSM candidate loci identified using the Platinum and Chromium short-read data and the HaplotypeSV genotype data for the reference individual NA12878. The Platinum and Chromium loci have passed sequence-based filtering and all sets have passed the mapping depth–based genotyping.

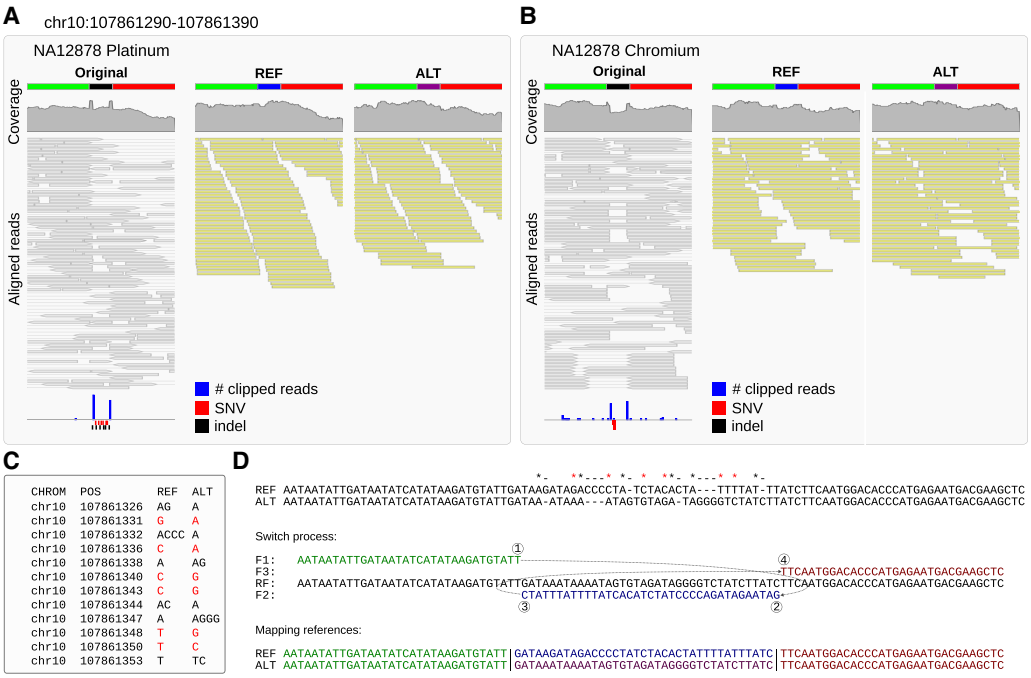

Figure 6. TSM candidates identified with short-read data. (*A*) In the Platinum data (*left*), a region in Chromosome 10 has an excess of soft clips (*bottom*, blue bars) and called variants (SNVs in red, indels in black); the mapping coverage (*top*, in gray) also shows atypical patterns. (*B*) Similar signals are seen in Chromium data. (*C*) The Platinum variant data contain six SNVs and six indels within 28 bp. (*D*) De novo assembly of the reads creates two locally highly dissimilar haplotypes compatible with the called variants (*top*). All differences can be explained with a TSM event, an inversion in place (*middle*). Using two haplotypes with alternative central parts (blue and magenta; *bottom*) as the reference, extracted reads map in full length (*A*,*B*; *middle*, *right*) with roughly even coverages. No variants for NA12878 were called in this region in HaplotypeSV data.

complexity sequence or VNTRs, but base changes, including those in the two longest events of 119 and 125 bp (Supplemental Fig. S12B,C), were not necessarily explainable by a simple slippage mechanism; within a complex sequence, the longest TSM event was 90 bp long (Supplemental Fig. S12A). A locus in Chromosome 19, present in a single individual, was of particular interest as 16 of the 19 base changes of gnomAD MNVs could be explained with two adjacent 62- and 32-bp TSM events separated by a 16-bp forward fragment (Supplemental Fig. S13). This indicates that, similarly to larger arrangements (Lee et al. 2007; Zhang et al. 2009), local TSMs can also be chained into complex combinations. One should note that the gnomAD MNV data lack indels and no TSM patterns involving sequence length changes were included, meaning that the real number of mutation clusters explainable by the TSM mechanism is likely much higher.

In replicating the analysis with the DNM data from 33 large, three-generation CEPH families by Sasani et al. (2019) (Table 1), although no TSM-like patterns were found among the germline DNMs of the 70 second-generation individuals, four consistent patterns (Supplemental Fig. S14) were observed among the 24,975 de novo SNVs and small indels of the 350 third-generation individuals. As most variants were isolated and there were only 218 DNM clusters, this gives a frequency of 1.8% for TSMs among the clustered DNMs. Sasani et al. (2019) estimated that, on average, there are 70.1 de novo SNVs and 5.9 de novo indels per genome, but the germline status of third-generation variants cannot be verified. In general, the vast majority of mutations are somatic, with Conrad et al. (2011) estimating a ratio of 20:1 for non-germline to germline DNMs, and the observed TSM-like patterns are also likely to be somatic.

Discussion

My analyses extend the previous works (Löytynoja and Goldman 2017; Walker et al. 2021) and show that human genomes have thousands of mutation patterns consistent with DNA replication-related template switching and that these loci segregate within and between populations. I studied the haplotype-resolved genotype data by Ebert et al. (2021), two independent WGS experiments on the reference individual NA12878 and her extended family (Eberle et al. 2017; Weisenfeld et al. 2017; Marks et al. 2019), and variant information in the gnomAD database (Karczewski et al. 2020; Wang et al. 2020) and from a DNM study (Sasani et al. 2019). Whereas I found the HaplotypeSV data of Ebert et al. (2021) to have captured large numbers of TSM patterns and mostly to reflect the haplotypic differences between the chromosomes, the call set was far from perfect and I could identify many additional TSM-like loci using the independent short-read data sets. The characteristics of the loci missing from the HaplotypeSV data raise serious questions about the completeness of the Ebert et al. (2021) variant set, and some of the omitted TSM-like mutation patterns were perfectly visible in the sequencing data and called in full for exactly the same individual in the 1000G variant data. In addition to short-read data, de novo TSMs were identified in the data of Sasani et al. (2019), although their germline status could not be confirmed, and consistent singleton TSM patterns were seen in gnomAD variant data (Karczewski et al. 2020; Wang et al. 2020). Overall, the results demonstrate that, despite their relatively low rate, the TSM mechanism explains a significant fraction of MNVs seen in human variation data and thus contributes to correlation in local mutation frequencies (Harris and Nielsen 2014; Ségurel et al. 2014).

Figure 7. Complex gnomAD MNV explained by a single TSM. (*A*) Wang et al. (2020) identified 15 SNVs within a 29-bp interval in Chromosome 12. (*B*) All the differences (*top*, marked with asterisks) can be explained with a TSM inverting a 39-bp-long fragment in place (*middle*, *bottom*).

Whereas the ability of TSMs to create reverse-complement copies of short sequence fragments is highly interesting, the mechanistic origin of sequence differences is irrelevant to many applications of genome analysis. My analyses revealed that TSM-like mutations are found within genes and other potentially functional elements, and there are anecdotal reports of TSM-like mutations causing genetic diseases in humans (Menardi et al. 1997). Given this, a worrying finding of the study was that the DNA sequencing pipelines capable of identifying SNVs may miss inverted fragments of a few tens of bases. The difficulty of detecting certain TSM patterns comes from the ability of mapping tools to align the affected reads in their expected context after excessive clipping of the read ends. If this mapping error goes unnoticed in variant calling, it can invalidate, for example, medical genetic studies searching for causative alleles behind genetic diseases. On the other hand, Seplyarskiy et al. (2021) recently proposed an approach to separate the signals of different mutation processes, and among others, described signals from "asymmetric resolution of bulky DNA damage" and from "asymmetric replication errors." Such asymmetric signal could be created by template switching if the 10× higher frequency of TSMs in leading strand seen in bacteria (Rosche et al. 1997; Seier et al. 2011) applies to other organisms, or if the mechanism is related, for example, to coordination or clashes of transcription and replication systems (Hamperl et al. 2017; Chen et al. 2019). Local TSMs have many similarities with the FoSTeS (Lee et al. 2007) and MMBIR (Hastings et al. 2009) mechanisms —and possibly with chromoanasynthesis in cancer (Liu et al. 2011; Holland and Cleveland 2012)—but due to their short length, they are expected to be more benign (Zhang et al. 2009). With novel tools shown here, TSM patterns can be identified in different types of genome data, enabling analyses of their genome-wide distribution and possible correlation with different cellular processes (Seplyarskiy et al. 2021). Such analyses should bring us closer to understanding the mechanisms underlying template switching.

Methods

Data

Ebert et al. (2021) genotype data (VCF; freeze3) were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/integrated_callset/, combining SNV, indel, and SV calls. Chromium data (BAM;VCF) were downloaded from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016/ and https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/technical/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016. Platinum data (BAM;VCF) were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/illumina_platinum_pedigree/data/CEU, via GitHub (https://github.com/Illumina/PlatinumGenomes) and from the NCBI dbGaP database (https://www.ncbi.nlm.nih.gov/gap/) under study ID phs001224.v1.p1. Pacific Biosciences' contribution to the NIST GIAB initiative was downloaded from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/PacBio_SequelII_CCS_11kb. 1000 Genomes variant data were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/phase3_liftover _nygc_dir/. gnomAD data were downloaded via https://gnomad.broadinstitute.org/downloads#v2-multi-nucleotide-variants and Sasani et al. (2019) data through GitHub (https://github.com/elifesciences-publications/ceph-dnm-manuscript/tree/master/data). Reference genomes were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/reference (GRCh38) and from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference (GRCh37 and 38). Description of the data sets is given in Table 1. With the exception of the CEPH 1463 children, all data used in this study are publicly available and the instructions for their download and analysis are provided in the Supplemental Code. The authorized access to the CEPH 1463 offspring data was under the project name "Properties of de novo template switch mutations." The data were analyzed in a secure computer environment accessible only by the author, and only summary statistics of genotype inheritance are reported.

Analysis of haplotype data

The genotype data of Ebert et al. (2021) (i.e., HaplotypeSV data) were processed with a custom Python script (tsm_scan_SV2.py) internally utilizing BWA (Li 2013), SAMtools, and BCFtools (Danecek et al. 2021). The script detects clusters of variants and creates alternative haplotypes by replacing parts of the reference genome sequence with the variant bases. The reconstructed haplotypes (with 150 bp of flanking sequence) are then compared to the reference allele, and a TSM solution is searched reciprocally to create one of the alleles from the other. Each individual and chromosome was processed independently.

The candidate loci in NA12878 were processed with another Python script (tsm_alleles2.py) that (1) creates the two alleles for each locus using the reference sequence and the HaplotypeSV variant information, (2) extracts the reads mapped to the locus in a BAM alignment file and maps these reads to the two alternative

alleles, and (3) computes the mapping coverage statistics for the two alleles, recording the mapping depth at the region differing between the alleles as well as immediately upstream and downstream of the differing region. The REF allele was taken from the reference genome, and the ALT allele was created by placing the NA12878 variants (each haplotype separately) into a copy of that; the differing regions were then placed in identical context with 500 bp of flanking sequence from the reference genome.

The genotype of the locus was inferred from the mapping coverage of the two alternative alleles using a strategy similar to base calling. A minimum coverage of 10 reads for upstream, downstream, and within the differing region for either of the alleles was required. Then, the probabilities for the three possible genotypes were obtained as

$$Pr\{G = R|R\} = (1-\epsilon)^{d_r}\epsilon^{d_a}$$
$$Pr\{G = R/A\} = 0.5^{d_r} \times 0.5^{d_a}$$
$$Pr\{G = A|A\} = \epsilon^{d_r}(1-\epsilon)^{d_a}$$

where ε is the error rate of 0.01 for the read being mapped to a wrong allele, and d_r and d_a are the mapping coverages for the REF and the ALT alleles. The inferred genotype X, where $X \in \{R|R, R/A, A|A\}$ (standing for homozygote REF, heterozygote, and homozygote ALT, respectively), was then the genotype with the highest probability if that was at least 99% of the total probability

$$Pr\{G = X\} = \max\begin{cases} Pr\{G = R|R\} \\ Pr\{G = R/A\} \\ Pr\{G = A|A\} \end{cases}$$

if for $Y = \{R|R, R/A, A|A\}$

$$\frac{Pr\{G = X\}}{\sum Pr\{G = Y\}} > 0.99.$$

The variant annotation was performed using R (R Core Team 2020) and the package VariantAnnotation (Obenchain et al. 2014) and the software BEDTools (Quinlan and Hall 2010) based on Ensembl v. 106 gene models (Homo_sapiens.GRCh38.106.chr.gff3). Only one count of each annotation class was considered for each locus.

Analysis of short-read data

The FPA algorithm (Löytynoja and Goldman 2017), written in C++ and available through GitHub (https://github.com/ariloytynoja/fpa), was integrated into the SvABA tool (Wala et al. 2018). Clusters of more than one base mismatch were identified in VCF data using BCFtools and an awk script counting variants within 20-bp intervals. Similarly, clusters of more than 10 clipped reads were identified in BAM data using SAMtools and an awk script calculating the positions of soft clips based on the CIGAR string and another script identifying clusters of positions within 20-bp intervals. Candidate loci were targeted with SvABA-FPA and, for contigs showing two or more base differences in comparison to the reference sequence, TSM solutions were computed using both the REF and the ALT allele (i.e., SvABA-created contig) as the ancestral type. The resulting TSM candidates were filtered and those overlapping with repeat elements or low-complexity sequence, or having short length or low identity at flanking sequences were removed. More precisely, loci intersecting with masked sequence or assembled contigs having low sequence complexity (Trifonov's complexity with order 5 >0.25; computed with program SeqComplex [Caballero et al. 2014]) were removed; TSMs were compared to forward alignments and required to be better, containing at least two edit events (of which at least one base mismatch) less than the nontemplate-switching alignment; of those, ones with ①–④ distance longer than 250 bp or shorter than 5 bp and upstream/downtream flanking regions or the ②→③ fragment showing sequence identity below 95% or being <10 bp long were discarded.

For the loci passing the sequence-based filtering, the REF and ALT alleles were placed in identical context with 500 bp of flanking sequence. Alignment data (in BAM format) were genotyped by extracting the reads and mapping them to the two alleles using the custom Python script. The genotype was inferred from the read coverage as explained above, and loci inferred as heterozygous or homozygous for the ALT allele were retained. Overlapping loci were counted using the R package GenomicRanges (Lawrence et al. 2013). The custom scripts for all steps and instructions for their use are provided in the Supplemental Code.

Analysis of CEPH 1463 data

The coordinates of candidate loci were transferred from GRCh38 to GRCh37 with CrossMap v0.2.9 (Zhao et al. 2014) using the Ensembl chain file. The same methods were used to genotype the loci.

Analysis of MNV data

The MNV data from gnomAD (Wang et al. 2020) and by Sasani et al. (2019) were processed similarly to the HaplotypeSV data with a Python script (tsm_scan_dnm.py). The script detects clusters of variants and creates alternative haplotypes by replacing parts of the reference genome sequence with the variant bases. It then compares the reconstructed haplotypes (with 150 bp of flanking sequence) with the reference allele and, reciprocally, searches for a TSM solution to create one of the alleles from the other. Tandem repeats were searched with trf (Benson 1999) and hits containing them in the ②→③ region were discarded.

Software availability

The Python/awk scripts used, the source code for the SvABA-FPA tool, and the instructions for their use are provided as Supplemental Code and deposited on GitHub (https://github.com/ariloytynoja/short-range-template-switching).

Competing interest statement

The author declares no competing interests.

Acknowledgments

I thank CSC – IT Center for Science and the University of Helsinki IT Center for the computing resources and the secure analysis environment, and the UH Biodata Analytics Unit for helpful discussions. This work was enabled by the Academy of Finland grant number 322681.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* **376:** eabl3533. doi:10.1126/science.abl3533

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580. doi:10.1093/nar/27.2.573

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole

human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59. doi:10.1038/nature07517
Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, et al. 2016. Multi-nucleotide *de novo* mutations in humans. *PLoS Genet* **12:** e1006315. doi:10.1371/journal.pgen.1006315
Caballero J, Smit AFA, Hood L, Glusman G. 2014. Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic Acids Res* **42:** e99. doi:10.1093/nar/gku356
Chen YH, Keegan S, Kahli M, Tonzi P, Fenyö D, Huang TT, Smith DJ. 2019. Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol* **26:** 67–77. doi:10.1038/s41594-018-0171-0
Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. 2018. Comparison of phasing strategies for whole human genomes. *PLoS Genet* **14:** e1007308. doi:10.1371/journal.pgen.1007308
Conrad DF, Keebler JEM, 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:** 712–714. doi:10.1038/ng.862
Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10:** giab008. doi:10.1093/gigascience/giab008
Dutra BE, Lovett ST. 2006. *Cis* and *trans*-acting effects on a mutational hotspot involving a replication template switch. *J Mol Biol* **356:** 300–311. doi:10.1016/j.jmb.2005.11.071
Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang H-Y, Humphray SJ, Halpern AL, et al. 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27:** 157–164. doi:10.1101/gr.210500.116
Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117
Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138. doi:10.1126/science.1162986
Falconer E, Lansdorp PM. 2013. Strand-seq: a unifying tool for studies of chromosome segregation. *Semin Cell Dev Biol* **24:** 643–652. doi:10.1016/j.semcdb.2013.04.005
Hamperl S, Bocek MJ, Saldivar JC, Swigut T, Cimprich KA. 2017. Transcription-replication conflict orientation modulates R-loop levels and activates distinct DNA damage responses. *Cell* **170:** 774–786.e19. doi:10.1016/j.cell.2017.07.043
Harris K, Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res* **24:** 1445–1454. doi:10.1101/gr.170696.113
Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5:** e1000327. doi:10.1371/journal.pgen.1000327
Holland AJ, Cleveland DW. 2012. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat Med* **18:** 1630–1638. doi:10.1038/nm.2988
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945. doi:10.1038/nature03001
Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581:** 434–443. doi:10.1038/s41586-020-2308-7
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118
Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131:** 1235–1247. doi:10.1016/j.cell.2007.11.037
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254. doi:10.1371/journal.pbio.0050254
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].
Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, et al. 2011. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146:** 889–903. doi:10.1016/j.cell.2011.07.042
Löytynoja A, Goldman N. 2017. Short template switch events explain mutation clusters in the human genome. *Genome Res* **27:** 1039–1049. doi:10.1101/gr.214973.116
Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29:** 635–645. doi:10.1101/gr.234443.118
Menardi C, Schneider R, Neuschmid-Kaspar F, Klocker H, Hirsch-Kauffmann M, Auer B, Schweiger M. 1997. Human APRT deficiency: indication for multiple origins of the most common Caucasian mutation and detection of a novel type of mutation involving intrastrand-templated repair. *Hum Mutat* **10:** 251–255. doi:10.1002/(SICI)1098-1004(1997)10:3<251::AID-HUMU15>3.0.CO;2-Z
Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annu Rev Genomics Hum Genet* **22:** 81–102. doi:10.1146/annurev-genom-120120-081921
Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987
Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. 2014. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30:** 2076–2078. doi:10.1093/bioinformatics/btu168
Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39:** 1256–1260. doi:10.1038/ng2123
Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, et al. 2021. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* **39:** 302–308. doi:10.1038/s41587-020-0719-5
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033
R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Ripley LS. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc Natl Acad Sci* **79:** 4128–4132. doi:10.1073/pnas.79.13.4128
Rosche WA, Trinh TQ, Sinden RR. 1997. Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. *J Mol Biol* **269:** 176–187. doi:10.1006/jmbi.1997.1034
Sakamoto Y, Sereewattanawoot S, Suzuki A. 2020. A new era of long-read sequencing for cancer genomics. *J Hum Genet* **65:** 3–10. doi:10.1038/s10038-019-0658-5
Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. 2019. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife* **8:** 46922. doi:10.7554/eLife.46922
Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15:** 47–70. doi:10.1146/annurev-genom-031714-125740
Seier T, Padgett DR, Zilberberg G, Sutera VA Jr, Toha N, Lovett ST. 2011. Insights into mutagenesis using *Escherichia coli* chromosomal *lacZ* strains that enable detection of a wide spectrum of mutational events. *Genetics* **188:** 247–262. doi:10.1534/genetics.111.127746
Seplyarskiy VB, Soldatov RA, Koch E, McGinty RJ, Goldmann JM, Hernandez RD, Barnes K, Correa A, Burchard EG, Ellinor PT, et al. 2021. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373:** 1030–1035. doi:10.1126/science.aba7408
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308–311. doi:10.1093/nar/29.1.308
Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM., Kang HM, et al. 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590:** 290–299. doi:10.1038/s41586-021-03205-y
Veltman JA, Brunner HG. 2012. *De novo* mutations in human genetic disease. *Nat Rev Genet* **13:** 565–575. doi:10.1038/nrg3241
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351. doi:10.1126/science.1058040
Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, et al. 2018. SvABA: genome-

wide detection of structural variants and indels by local assembly. *Genome Res* **28:** 581–591. doi:10.1101/gr.221028.117

Walker CR, Scally A, De Maio N, Goldman N. 2021. Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLoS Genet* **17:** e1009221. doi:10.1371/journal.pgen.1009221

Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, et al. 2020. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat Commun* **11:** 2539. doi:10.1038/s41467-019-12438-5

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14:** 125–138. doi:10.1038/nrg3373

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27:** 757–767. doi:10.1101/gr.214874.116

Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, McCoy RC. 2021. Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife* **10:** e67615. doi:10.7554/eLife.67615

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10:** 451–481. doi:10.1146/annurev.genom.9.081307.164217

Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30:** 1006–1007. doi:10.1093/bioinformatics/btt730

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Received December 9, 2021; accepted in revised form June 21, 2022.