Drastic mutations can be hidden in short-read mapping: thousands of mutation clusters in human genome are explained by short-range template switching

Ari Löytynoja[1,*]

1 Institute of Biotechnology, University of Helsinki, Finland

* ari.loytynoja@helsinki.fi

Abstract

Variation within human genomes is distributed unevenly 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. We reanalyzed haplotype-resolved genome assemblies representing 25 human populations and multinucleotide variants aggregated from 140,000 human sequencing experiments. We found local template switching to explain thousands of complex mutation clusters across the human genome, the loci segregating within and between populations with a small number appearing as de novo mutations. We developed computational tools for genotyping candidate template switch loci using short-read sequencing data and for identification of template switch events using both short-read data and genotype data. These tools will enable building a catalogue of affected loci and studying the cellular mechanisms behind template switching both in healthy organisms and in disease. Strikingly, we noticed that widely-used analysis pipelines for short-read sequencing data – capable of identifying single nucleotide changes – may miss TSM-origin inversions of tens of base pairs, potentially invalidating medical genetic studies searching for causative alleles behind genetic diseases.

Author summary

Mutations are not randomly distributed in genomes and they often appear as clusters of nearby changes. We earlier showed that a poorly known mechanism in DNA replication can create short inverted copies of nearby sequence and that these events then show as mutation clusters in sequence comparison. Using the latest DNA sequencing and variation data we show that the human genome contains thousands of mutation clusters consistent with this mechanism and that novel mutations are created at a significant rate. Strikingly we observed that widely used methods for processing DNA sequencing data may completely miss these mutations. This has significance e.g. in medical genetic studies aiming to identify mutations causing genetic diseases.

Introduction

Twenty years after the publication of the draft genomes [1, 2], the Telomere-to-Telomere Consortium reported the first truly complete sequence of a human genome [3, 4]. Every genome is unique, however, and the challenge now is to understand the global genomic diversity in the human population [5, 6] and to build a comprehensive pangenome to represent this variation [7]. While single nucleotide variation (SNV) and short insertion-deletions (indels) have been resolved relatively accurately with classical short-read sequencing [8], the characterization of structural variants (SVs) and low-complexity sequences has been lacking. The Human Genome Structural Variation Consortium developed a method [9] for phased diploid genome assembly with combination of long-read PacBio whole-genome sequencing (WGS) [10] and Strand-seq phasing [11] data. Ebert *et al.* [12] applied this method to 32 diverse human individuals and produced 64 assembled haplotypes, i.e. maternal and paternal copies of the genome. With the help of phased diploid genomes, they massively expanded the catalogue of known SVs and, having the information of co-inherited nearby sequence differences, could study in detail the different subclasses of complex variants. Such accurate characterization of multiple genomes will lay the foundations for understanding the role of complex mutations in human phenotypes and disease [13, 14, 15] and in evolution [13, 16, 17]. More generally, the unprecedented resolution of the new genome data will

open novel possibilities for understanding the mutational mechanisms of complex eukaryotic genomes [18, 19, 20].

While many subclasses of variants studied by Ebert *et al.* are clearly independent, mutations are known to be clustered [20] and their separation in the analysis stage may dismiss crucial information. We compared earlier two independent assemblies of the human genome [21, 22] and found mutation clusters consistent with DNA-replication related template switching [23]. The mechanisms, originally described and studied in bacteria [24, 25], resembles FoSTeS (Fork Stalling and Template Switching) [26] and MMBIR (Microhomology-Mediated Break-Induced Replication) [27] but is local and the reciprocal switches typically take place 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 [24, 25] (Fig. S1). Mutations compatible with the TSM mechanism appear as multiple nearby sequence changes and may be represented as combinations of SNPs, MNPs (single- and multi-nucleotide polymorphisms) and short indels in variant data. We showed [23] that the 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 [28] contain consistent de novo mutations (DNMs). The haplotype-resolved data of Ebert *et al.* and the variation information compiled since then by massive sequencing studies [5, 6] provide an opportunity to revisit the topic and study the TSM mechanism's role in generation of genomic variation. As haplotype-resolved genome data are still rare, it is of interest to know how reliably the TSM patterns found with de-novo-assembled genomes can be identified using the old-school short-read DNA sequencing.

Results

TSMs explain thousands of haplotypic mutation clusters

Utilising the variant data of Ebert *et al.* [12] (called here "HaplotypeSV data"), we combined the SNVs, indels and SVs, and for each maternal and paternal genome copy, identified the clusters of sequence differences (see Methods). For each mutation cluster locus within a haplotype, we placed the identified variants into their genomic context, 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') – we applied the four-point alignment (FPA) algorithm [23] to find the best solution involving a template switch, testing both alleles as the ancestral state (Fig. 1C,D; see Methods). Despite the attempts to evaluate TSMs under a probabilistic model [29], it is unclear when a potential TSM solution should be considered more plausible than a combination of SNVs and indels. We set an arbitrary rule that the TSM candidates need to (i) 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 (ii) have a ②→③ region, inferred to be copied from the other template, of at least 8 bases long. Despite seeing strong evidence for many VNTRs (variable number of tandem repeats) evolving through template switching (Fig. S3), we required the candidate ②→③ regions to be contain all four bases, thus removing hits within low-complexity sequence.

Per individual, we found 3,049–3,934 to mutation patterns consistent with the TSM mechanisms, the longest of them 318 bp in length (Table 1, 'ALL'). Many of these were in proximity of repeat elements and could potentially have been affected by gene conversion. After removing the loci masked as repetitive sequence in the reference genome, there were 872–1,121 inferred TSM events per individual (Fig. 1E), the maximum length varying from 97-180 and the mean length being 11.56 bp (Table 1, 'UNMASKED'; Fig. S4). Homozygosity of the inferred TSM loci varied from 26.7–42.3% (Fig. 1E), and the proportion of singletons from 2.1–9.0% (Table S1). Although comparisons on 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%, while those for Europeans

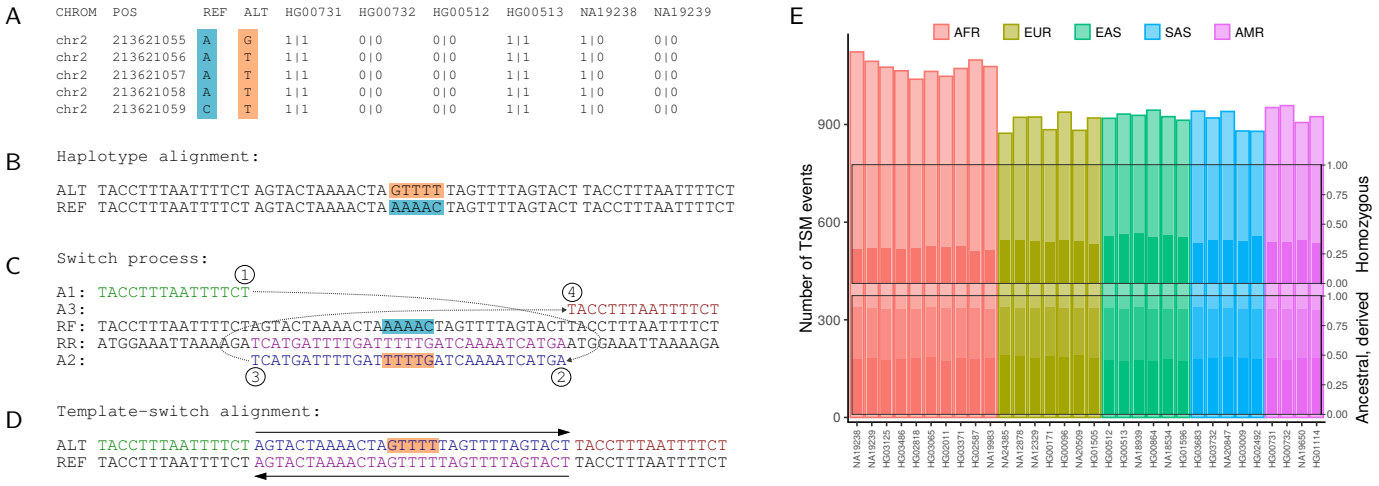


Fig 1. TSMs in HaplotypeSV data. (**A**) HaplotypeSV data contain a cluster of five SNVs that segregate as a unit among the individuals (first six shown). (**B**) Within their genomic context, REF (cyan) and ALT (orange) alleles create two haplotypes. (**C**) The FPA algorithm [23] attempts to explain 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 (full height, left); proportion of homozygous loci (middle, right); and proportion of ancestral (dark shading), derived (light shading) and unpolarized (no shading) alleles (bottom, right).

(EUR, including Azkenazi NA24385) and Asians (AAS, including South and East Asians) were higher and lower, respectively (Table 1). The greater variation of African populations and the old age of the TSM patterns was 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 1).

A unique feature of TSMs is that often the alternative alleles can be polarized, i.e. have the ancestral and derived state determined, without an outgroup (Fig. S2). We could define the ancestral allele in 87.7–90.3% of the cases and found the proportion of loci with the ALT variant being the ancestral allele ranging from 49.7–55.3 across the individuals (Fig. 1E; Table S1). On superpopulation level, Europeans had the highest average proportion of ancestral variants (Table 1), probably reflecting the European origin of the GRCh38 reference genome and thus it capturing a greater amount of derived European TSM variation.

Table 1. TSM statistics for HaplotypeSV data.

	Superpopulation	AFR	EUR	AAS
	Haplotypes	20	14	22
ALL	TSMs/individual	3783	3135	3173
	unique TSMs	12,501	7,370	8,956
	maximum length	318	188	188
	average length	13.14	13.14	13.15
UNMASKED	TSMs/individual	1,074	906	919
	unique TSMs	3,133	1,899	2,246
	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%	–

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 [30] as well as one of the reference samples of the Genome in a Bottle (GIAB) initiative [31]. We genotyped the TSM loci identified in NA12878 in different GIAB datasets and in the pedigree 1463 in order to assess their de novo vs. standing, in vitro vs. in vivo, and somatic vs. germline origin. Utilising the HaplotypeSV genotype data for NA12878, we reconstructed the genome context for the alternative alleles and, for each individual in the family, mapped the reads extracted from that particular loci against the two alternative alleles. Outside the mutation locus, the short reads should map evenly between the two alleles, but the reads overlapping the position that differ between the alleles are expected map to one allele only (Fig. 2A,B). By comparing the read coverage across the two alternative alleles, we could identify the test individual as homozygote for REF allele (R|R), heterozygote (R/A), or homozygote for ALT allele (A|A). We adopted a strategy similar to base calling and required 99% posterior probability for the genotype calls (see Methods).

We started with the Illumina Platinum [30] and 10X Genomics Chromium [32]

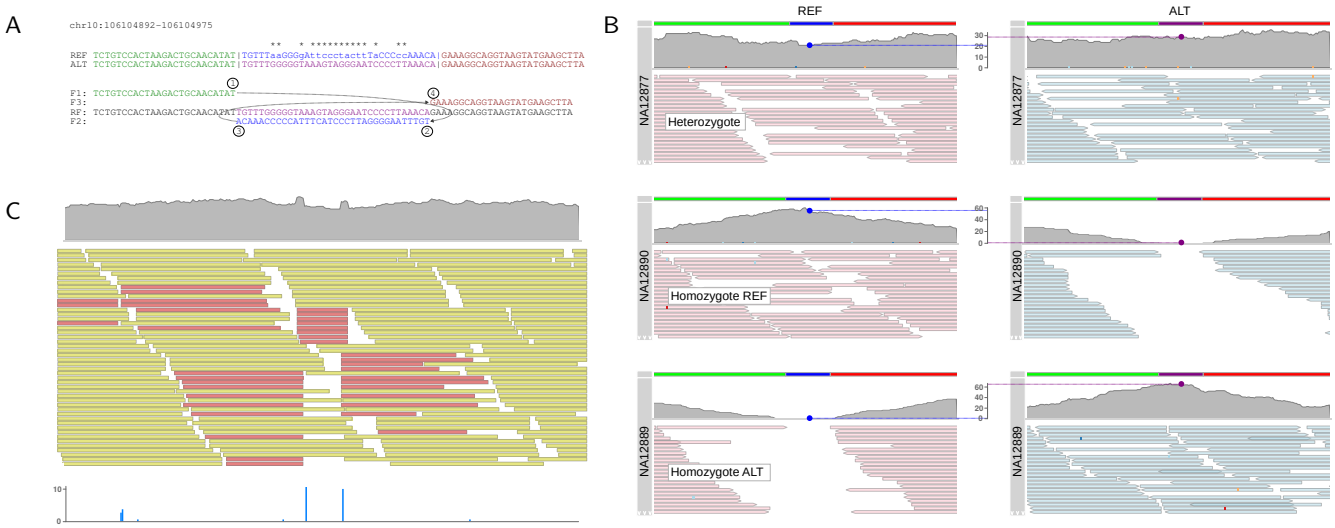


Fig 2. Genotyping of a TSM locus. (**A**) TSM solution for a complex mutation in chr10. (**B**) The two haplotypes with alternative central parts (blue and magenta; top) are used as the reference for re-mapping of reads extracted from the locus. NA12877 (top) is heterozygous and mapping coverage is uniform for the two alleles (gray). NA12890 (middle) and NA12889 (bottom) are homozygous for the REF (left, reads in pink) and ALT (right, reads in light blue) alleles, respectively. Genotype is inferred of the mapping coverage in the middle of the locus (blue and magenta dots and dotted lines). (**C**) NA12878 Platinum data with full and clipped reads in yellow and red, respectively, and number of clip ends as blue bars (bottom).

datasets for NA12878 and genotyped the TSM-like loci found in the HaplotypeSV data. Focusing on the 864 loci that were not masked as repeats or low-complexity sequence and had enough read coverage in both datasets, we found the two short-read datasets to provide the genotype and agree on it in 831 cases (Fig. 3A; Table S2). While 14 loci could not be genotyped with the short-read data using the 99% posterior probability cut-off, the variant pattern in the HaplotypeSV data was inconsistent in 84 cases, and the genotype of the loci could not determined. Of the 780 loci inferred to be either heterozygous or homozygous for the ALT allele in the HaplotypeSV data, the three datasets agreed in 735 cases (Table S2). Eight loci recorded variable in the HaplotypeSV data were inferred to be homozygous for the REF allele with both short-read datasets. Many of these were in repetitive sequence context and all loci were found segregating among other samples in the HaplotypeSV data ruling out the possibility of DNMs. A small number of the TSM candidate loci showed inconsistency among different datasets, possibly due to incorrectly mapped reads (Fig. S5).

To further verify the genotyping approach, we performed a similar analysis on the parents and the children of NA12878. On the parents, we genotyped the 783 loci found

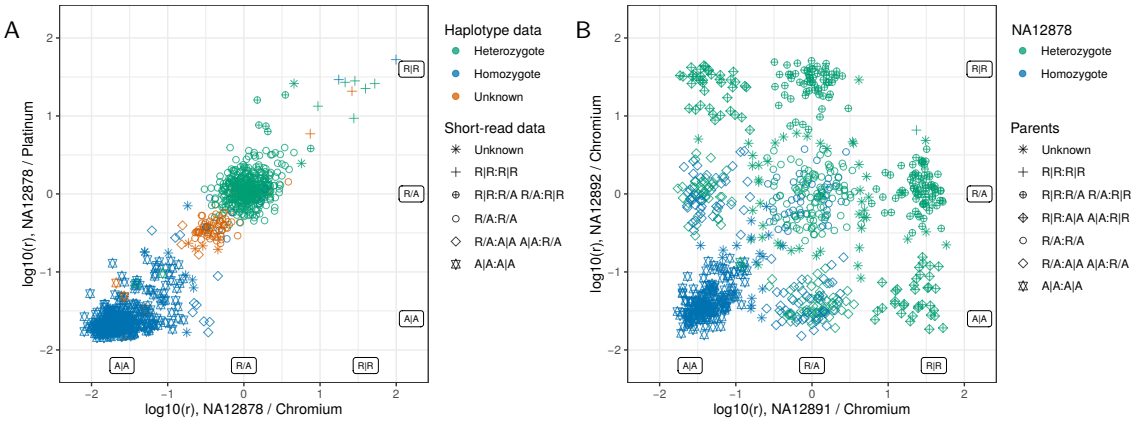


Fig 3. Confirmation of HaplotypeSV TSM loci with short-read data. (**A**) Ratio of REF and ALT allele mapping coverage ($r = depth_{\mathrm{REF}}/depth_{\mathrm{ALT}}$) reflects the genotype: $r \approx 1$ (and thus $log_{10}(r) \approx 0$) for a heterozygote; $r \ll 1$ and $r \gg 1$ for the two types of homozygotes. The Log10-ratios agree for the NA12878 Platinum and Chromium datasets, 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 datasets; 17 and 10 loci are called homozygous REF with at least one dataset 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 (blue), both parents contain at least one ALT allele. 'Unknown' indicates inconsistent variant data or low posterior probability. One pseudo-count was added to all values to avoid divisions by zero.

in the HaplotypeSV data and called to contain at least one ALT allele in the NA12878 Chromium data. We found strong agreement between the inferred genotypes and, of the 307 loci called homozygous for the ALT allele in NA12878, 290 could be genotyped with the both parental Chromium data and each of them contained at least one ALT allele (Fig. 3B; Table S3). The only locus appearing as a novel mutation according to parental data (R|R:R|R in Fig. 3B) was a false positive and in a manual verification one parent was found to contain two supporting reads. Genotyping of the loci in the eleven children (CEPH pedigree 1463) required their transfer from GRCh38 to GRCh37. The liftover worked and the parents' alignment data (Platinum/GRCh37) contained enough reads to genotype 843 of the TSM-like loci found variable in NA12878 in the HaplotypeSV data. With the exception of loci that were inferred to be heterozygous in both parents (R/A,R/A), the observed genotypes matched nearly perfectly the expectation under Hardy-Weinberg equilibrium (Table 2). An excess of heterozygotes in genotype data is a classical mark of an artefact caused by sequencing reads originating from different loci, e.g. 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.

Table 2. 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*
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

*Loci with posterior probability below 0.99 considered as unknown

Novel TSM loci can be discovered with short-read data

Many of the TSM loci detected in HaplotypeSV data had not been originally called with the short-read data despite the two alleles differing over several tens of bases. A closer look on 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. 2A) is long enough: the mapping algorithm then produces a descent mapping for the read core by clipping the mismatching overhangs (Fig. 2C). While this is clearly undesirable behaviour, 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, we took SvABA [33], a tool developed for detection of structural variants by local reassembly of short sequencing reads, and integrated our FPA algorithm [23] to search for TSM patterns in the resulting contigs. We independently identified clusters of clip sites in the Illumina Platinum [30] and 10X Genomics Chromium [32] short-read data for NA12878 and clusters of variants in the corresponding variant data, and applied SvABA-FPA on each candidate locus (see Methods). With this, we found 1,054 and 26,222 candidate TSM loci in the Platinum and Chromium data, respectively, the significantly different numbers reflecting the technological differences between the standard and the linked-reads sequencing as well as the downstream bioinformatic methods used [34, 35]. We applied rigorous filtering (see Methods) and reduced, respectively, the numbers to 214 and 757 high-confidence

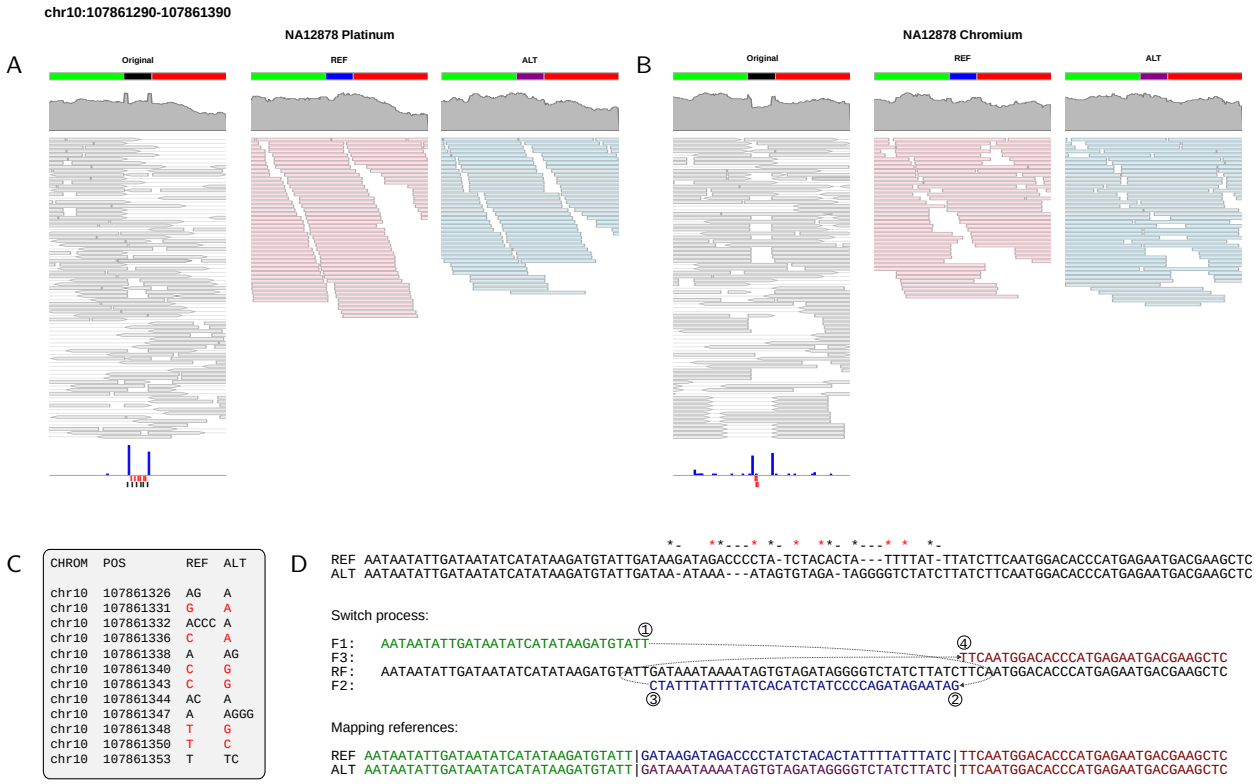


Fig 4. TSM candidate identified with short-read data. (**A**) In Platinum data (left), a region in chr10 has excess of soft clips (bottom, blue bars) and called variants (SNVs in red, indels in black); the mapping coverage (top, in gray) also shows atypical patterns. (**B**) Similar signals are seen in Chromium data. (**C**) 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 by Ebert *et al.* [12].

loci. Encouragingly, 181 of the loci were shared among the short-read datasets, and 121 and 148 of the loci overlapped with the 883 HaplotypeSV cases, the filtering thus removing 79.7 and 97.1 % of the candidate loci but only 37.0 and 34.8 % of the overlaps.

We studied in more detail the 59 cases that had the ②→③ fragment of at least 25 bases long (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. 4A,B) while in a few rare cases the variant calling had correctly captured the differences between the two TSM haplotypes (Fig. 4C,D). Some of the heterozygous loci were not called in any variant call set despite significant differences between the two haplotypes (Fig. S6, Data S1), and, even more strikingly, several loci homozygous for the ALT allele were uncalled in different variant call sets,

including the HaplotypeSV data (Fig. S7). Although nominally of the same origin, the studied sequencing datasets derive from lymphoblastoid cell lines and the differences may indicate activity of the TSM mechanism in vitro cell cultures [36]. Among the 59 cases with the longest ②→③ fragment, 17 loci appeared to be DNMs in the Chromium data whereas no unique TSM loci were inferred in the Platinum data (Data S1). While this is consistent with the lower total number of hits in Platinum data, the result may also reflect the differences in the sequencing technologies and bioinformatic pipelines [32]. Although the lower mapping coverage for ALT alleles (Fig. S8, Data S1) is expected from DNMs, we cannot exclude the possibility of artefacts in the underlying data.

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. 4B), but none of the manually confirmed cases supported by both short-read datasets was truly novel and absent from the dbSNP database [37] (Data S1). While dbSNP correctly lists most studied loci as multinucleotide variants (MNVs), all the loci were also present as multiple SNVs and indels and some only as multiple independent MNV or indels (Fig. S9, 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 [38].

The Genome Aggregation Database (gnomAD, v2.1.1) [5] 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.* [39] identified gnomAD SNVs appearing within proximity of 1-10 bp, and provide them as phased variant pairs. Using these data, we selected variant pairs with similar allele counts (ACs; ΔAC $\leq$10%), merged the data of different variant-pair distances, and for all clusters of three variants, created the alternative alleles and searched for a TSM solution to explain the differences (see Methods). Among the 91,300 clusters of three or more SNVs, we could find a TSM solution for 4,425 loci. Among these, we had a closer look at the 192 cases where the ②→③ region was at least 25 bp long and the REF and ALT allele differed at least by four SNVs (Data S2).

Fig 5. Complex gnomAD MNV explained by a single TSM. (**A**) Wang *et al.* [39] identified 15 SNVs within a 29 bp interval in chr12. (**B**) All the differences (top, marked with asterisks) can be explained with a TSM inverting 39 bp-long fragment in place (middle, bottom).

Among the studied cases, the TSM mechanism perfectly explained MNVs consisting of of up to 15 base changes (Fig. 5) while the longest inferred TSM events had the ②→③ region of over a hundred bases (Fig. S10). Consistent with our earlier findings, 36% of the studied TSM loci contained low complexity sequence or VNTRs, but the base changes, including those in the two longest events of 119 and 125 bp (Fig. S10B,C), were not necessarily explainable by a simple slippage mechanism; within complex sequence, the longest TSM event was 90 bp long (Fig. S10A). Interestingly, we found a locus, present in a single individual, where 16 of the 19 gnomAD MNV base changes could be explained with two adjacent TSM events of 62 and 32 bp separated by a forward fragment of 16 bp (Fig. S11), indicating that also local TSMs can be chained into complex mutations [13]. As the gnomAD MNV data lack indels, no TSM patterns involving sequence length changes were included and the real number of mutation clusters explainable by the TSM mechanism is likely much higher.

We replicated the analysis with the DNM data from 33 large, three-generation CEPH families by Sasani *et al.* [40]. We did not find TSM-like patterns among the germline DNMs of the 70 second-generation individuals but saw four consistent patterns (Fig. S12) 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.* estimated that, in average, there are 70.1 de novo SNVs and 5.9 de novo indels per genome; the germline status of third-generation variants cannot be verified, but assuming the 20:1 ratio of non-germline to germline DNMs [18], the observed TSM-like patterns are likely to be somatic.

Discussion

Our analyses extend the previous works [23, 29] and show that human genomes have thousands of mutations patterns consistent with the DNA-replication related template switching [23] and that these loci segregate within and between populations. We studied the haplotype-resolved genotype data by Ebert *et al.* [12], two independent WGS experiments on the reference individual NA12878 and her extended family [30, 32], and variant information in the gnomAD database [5, 39] and from a DNM study [40]. While we found Ebert *et al.*'s HaplotypeSV data to have captured large numbers of TSM patterns and mostly reflect the haplotypic differences between the chromosomes, the call set was far from perfect and we could identify many additional TSM-like loci using the independent short-read datasets. We identified de novo TSMs both in short-read data and in data of Sasani *et al.*, though not able to confirm their germline status, and saw consistent singleton TSM patterns in gnomAD variant data [5, 39]. Overall, our 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 [20, 41].

While 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. There are anecdotal cases of human genetic diseases caused by TSM-like mutations [42], however, and a striking finding of our study was that the DNA sequencing pipelines capable of identifying SNVs may miss inversions of a few tens of bases. The difficulty of detecting certain TSM patterns using the standard analysis tools potentially invalidates e.g. medical genetic studies searching for causative alleles behind genetic diseases. On the other, Seplyarskiy *et al.* [43] 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 times higher frequency of TSMs in leading strand seen bacteria [44, 45] applies to other organisms, or if the mechanism is related e.g. to coordination or clashes of transcription and replication systems [46, 47]. Local TSMs have many similarities with the FoSTeS [26] and MBIR [27] mechanisms – and

possibly with chromoanasynthesis in cancer [48, 49] – but due to their short length are expected to be more benign [13]. With our novel tools 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 [43]. Such analyses should bring us closer to understanding the mechanisms underlying the template switching.

Methods

Data The Ebert *et al.* 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. The Chromium data (bam;vcf) were downloaded from `http://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016` and `http://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/10Xgenomics_ChromiumGenome_LongRanger2.0_06202016`. The Platinum data (bam;vcf) were downloaded from `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/illumina_platinum_pedigree/data/CEU`, via `http://github.com/Illumina/PlatinumGenomes`, and from dbGaP (`http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001224.v1.p1`; authorized access for A.L.). The gnomAD data were downloaded through `https://gnomad.broadinstitute.org/downloads#v2-multi-nucleotide-variants` and the Sasani *et al.* data through `https://github.com/elifesciences-publications/ceph-dnm-manuscript/tree/master/data`. The 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).

Analysis of haplotype data The genotype data of Ebert *et al.* (i.e., HaplotypeSV data) were processed with a custom python script (tsm_scan_SV2.py) internally utilising bwa [34], samtools and bcftools [50]. 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 (i) creates the two alleles for each locus using the reference sequence and the HaplotypeSV variant information, (ii) extracts the reads mapped to the locus in a bam alignment file and maps these reads to the two alternative alleles, and (iii) 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 on identical context with 500 bp of flanking sequence from the reference genome.

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

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

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 $Y = \{R|R, R/A, A|A\}$:

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

.

Analysis of short-read data The FPA algorithm [23], written in `C++` and available through `https://github.com/ariloytynoja/fpa`, was integrated in the SvABA tool [33]. Clusters of more than one base mismatch or more than ten clipped reads within a 20-bp interval were identified in vcf and bam data using custom bash and awk scripts. The 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; the bash/awk scripts for filtering are provided in the supplementary data.

For the loci passing the filtering, the REF and ALT alleles were placed on 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.

Analysis of CEPH 1463 data The coordinates of candidate loci were transferred from GRCh38 to GRCh37 with CrossMap v0.2.9 [51] using the Ensembl chain file. The same methods were used to genotype the loci.

Analysis of MNV data The MNV data from gnomAD and by Sasani *et al.* 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) to the reference allele and searches for a TSM solution, reciprocally, to create one of the alleles from the other. Tandem repeats within the ②→③ region of selected hits were searched with trf [52].

Data and code availability All material will be deposited in Github. The python/awk scripts used, the source code for the SvABA-FPA tool, and the instructions for their use are provided in the supplementary material. With the exception of the CEPH 1643 children, all data utilised in this study are publicly available and the instructions for their download and analysis are provided in the supplementary data.

Ethics statement A.L. had authorized access to the dbGaP genome data for the CEPH 1643 offspring (`http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001224.v1.p1`) under project name "Properties of de novo template switch mutations". The offspring data were analysed in a secure computer environment only accessible by the author, and only summary statistics of genotype inheritance are reported.

Acknowledgments

We 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 #322681.

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304–1351.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

3. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome; 2021.

4. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation; 2021.

5. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–443.

6. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290–299.

7. Miga KH, Wang T. The Need for a Human Pangenome Reference Sequence. Annu Rev Genomics Hum Genet. 2021;.

8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–59.

9. Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat Biotechnol. 2021;39(3):302–308.

10. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–138.

11. Falconer E, Lansdorp PM. Strand-seq: a unifying tool for studies of chromosome segregation. Semin Cell Dev Biol. 2013;24(8-9):643–652.

12. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372(6537).

13. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–481.

14. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125–138.

15. Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. J Hum Genet. 2020;65(1):3–10.

16. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39(10):1256–1260.

17. Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, et al. Local adaptation and archaic introgression shape global diversity at human structural variant loci; 2021.

18. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011;43(7):712–714.

19. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13(8):565–575.

20. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15:47–70.

21. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–945.

22. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5(10):e254.

23. Löytynoja A, Goldman N. Short template switch events explain mutation clusters in the human genome. Genome Res. 2017;27(6):1039–1049.

24. Ripley LS. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc Natl Acad Sci U S A. 1982;79:4128–4132.

25. Dutra BE, Lovett ST. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. J Mol Biol. 2006;356:300–311.

26. Lee JA, Carvalho CMB, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 2007;131(7):1235–1247.

27. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009;5(1):e1000327.

28. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, et al. Multi-nucleotide de novo Mutations in Humans. PLoS Genet. 2016;12(11):e1006315.

29. Walker CR, Scally A, De Maio N, Goldman N. Short-range template switching in great ape genomes explored using pair hidden Markov models. PLoS Genet. 2021;17(3):e1009221.

30. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. 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. 2017;27(1):157–164.

31. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025.

32. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27(5):757–767.

33. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28(4):581–591.

34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;.

35. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. 2019;29(4):635–645.

36. Shirley MD, Baugher JD, Stevens EL, Tang Z, Gerry N, Beiswanger CM, et al. Chromosomal variation in lymphoblastoid cell lines. Hum Mutat. 2012;33(7):1075–1086.

37. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–311.

38. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. Comparison of phasing strategies for whole human genomes. PLoS Genet. 2018;14(4):e1007308.

39. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun. 2020;11(1):2539.

40. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife. 2019;8.

41. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014;24(9):1445–1454.

42. Menardi C, Schneider R, Neuschmid-Kaspar F, Klocker H, Hirsch-Kauffmann M, Auer B, et al. 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. 1997;10:251–255.

43. Seplyarskiy VB, Soldatov RA, Koch E, McGinty RJ, Goldmann JM, Hernandez RD, et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. Science. 2021;.

44. Rosche WA, Trinh TQ, Sinden RR. Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. J Mol Biol. 1997;269(2):176–187.

45. Seier T, Padgett DR, Zilberberg G, Sutera VA Jr, Toha N, Lovett ST. Insights into mutagenesis using Escherichia coli chromosomal lacZ strains that enable detection of a wide spectrum of mutational events. Genetics. 2011;188(2):247–262.

46. Hamperl S, Bocek MJ, Saldivar JC, Swigut T, Cimprich KA. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. Cell. 2017;170(4):774–786.e19.

47. Chen YH, Keegan S, Kahli M, Tonzi P, Fenyö D, Huang TT, et al. Transcription shapes DNA replication initiation and termination in human cells. Nat Struct Mol Biol. 2019;26(1):67–77.

48. Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell. 2011;146(6):889–903.

49. Holland AJ, Cleveland DW. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat Med. 2012;18(11):1630–1638.

50. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2). doi:10.1093/gigascience/giab008.

51. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006–1007.

52. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–580.

53. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.

54. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–192.

55. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. In: Mathé E, Davis S, editors. Statistical Genomics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 335–351.

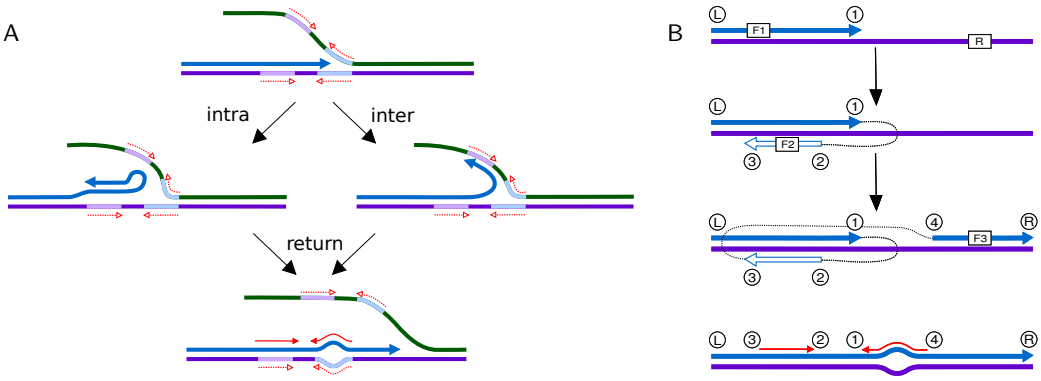


Fig S1. Classical vs. four-point model model of template switching. (**A**) In the classical model of Ripley [24], the replication (blue arrow) switches the template and starts going backwards either along the newly synthesized strand (intra) or along the opposite strand (inter). After a short while, the replication switches back to the original strand (return) and continues normally. As the result, the near-perfect inverted repeat (red dashed arrows) have been converted into a perfect inverted repeat (red arrows), causing mismatches between the sequences (bulge). (**B**) The four-point model of Löytynoja and Goldman [23] is a computationally tractable representation of the switch process and relaxes some of the assumptions underlying the Ripley model. The model assumes four points, ① - ④, where the replication process changes from one template to another, and copying of the fragment ②→③ in reverse-complement manner. The outcome of the switch process depends on the relative positions of points ①-④, and the type 3-2-1-4 (read from left to right) shown here is consistent with the process shown in a and can take place either intra- or inter-strand. See Löytynoja and Goldman [23] for enumeration and outcomes of all possible switch point combinations.

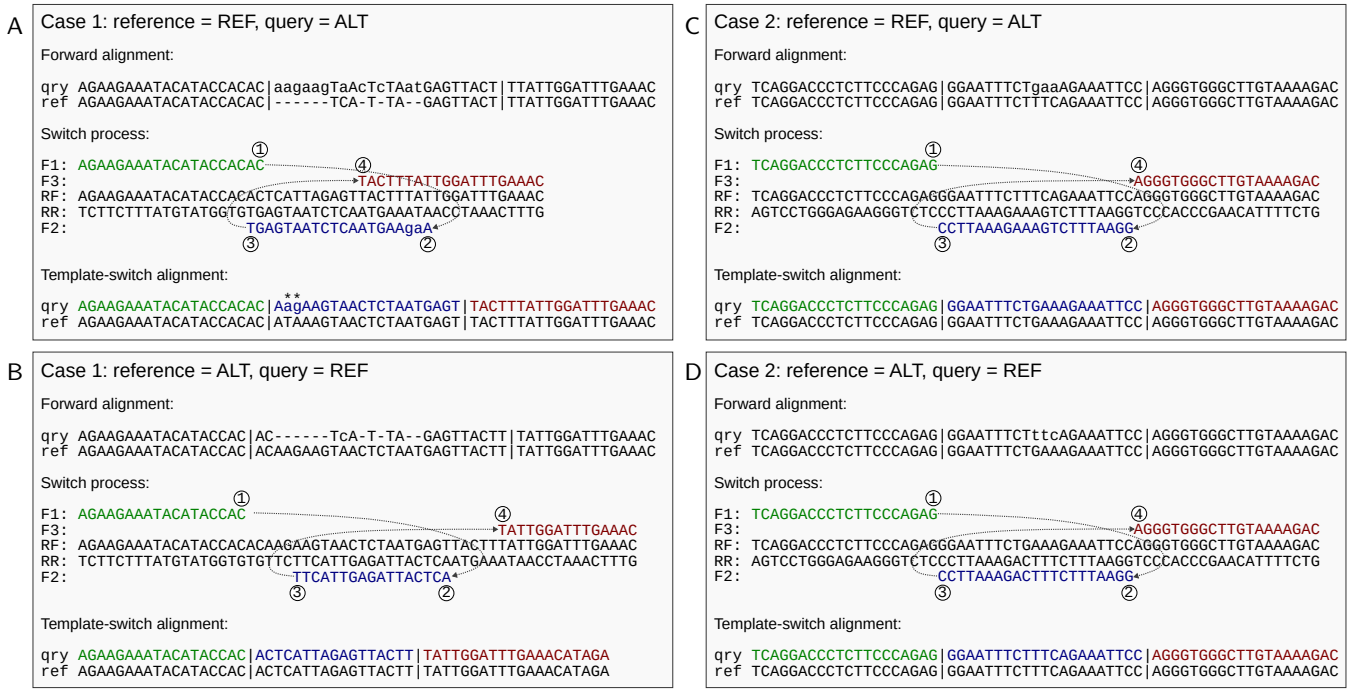
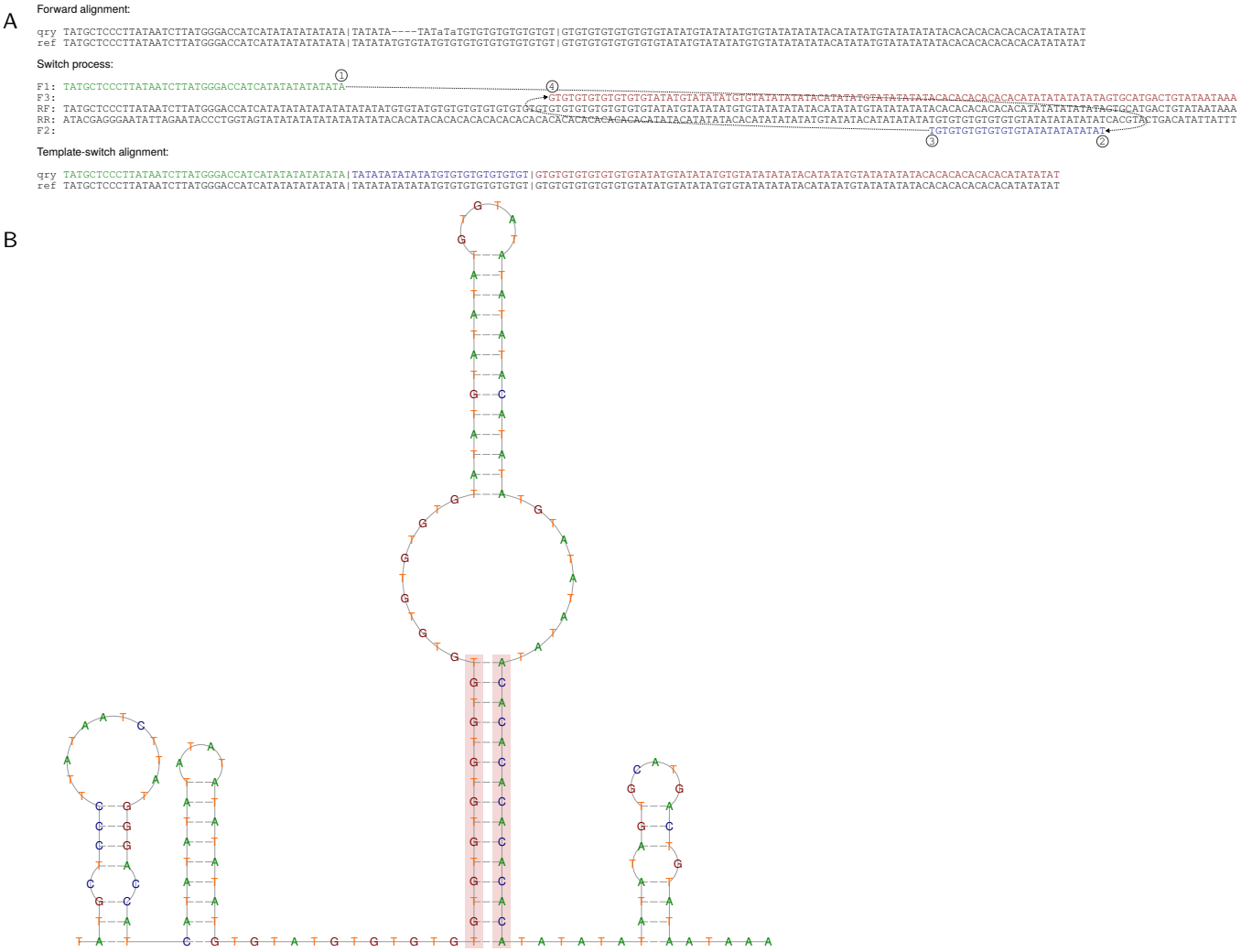


Fig S2. Direction of TSMs. (**A**) The complex mutation (top) is partially explained by the TSM mechanism when the REF allele is used as the reference (middle); the solution (bottom) contains two mismatches (marked with asterisks). (**B**) With the ALT allele as the reference, the complex mutation is fully explained, indicating that ALT is the ancestral state and REF is a derived state. (**C,D**) Inversions in place cannot be polarised as both arrangements fully explain the complex mutation pattern with a single TSM event.


A
Forward alignment:
qry TATGCTCCCTTATAATCTTATGGGACCATCATATATATATATA|TATATA----TATaTaTGTGTGTGTGTGTGT|GTGTGTGTGTGTGTGTATATGTATATATGTGTATATATATACATATATGTATATATATACACACACACACATATATAT
ref TATGCTCCCTTATAATCTTATGGGACCATCATATATATATATA|TATATATGTGTATGTGTGTGTGTGTGTGTGT|GTGTGTGTGTGTGTGTATATGTATATATGTGTATATATATACATATATGTATATATATACACACACACACATATATAT
Switch process:
F1: TATGCTCCCTTATAATCTTATGGGACCATCATATATATATATA
F3:
RF: TATGCTCCCTTATAATCTTATGGGACCATCATATATATATATATATATATATGTGTATGTGTGTGTGTGTGTGTGTGTGTGTATATGTATATATGTGTATATATATACATATATGTATATATATACACACACACACATATATATATATATAGTGCATGACTGTATAATAAA
RR: ATACGAGGGAATATTAGAATACCCTGGTAGTATATATATATATATATATATACACATACACACACACAC

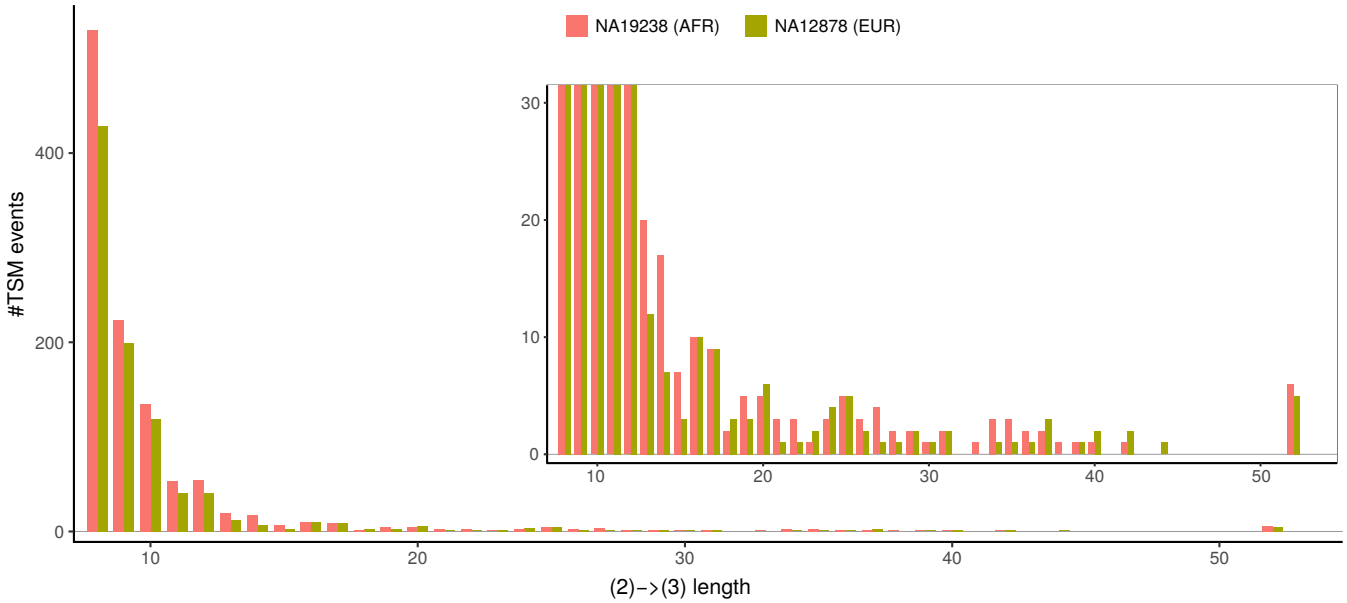


Fig S4. Lengths of inferred TMS events. A great majority of inferred TSM events are 10 bp or shorter in length. Despite the higher numbers of TSM loci identified in African samples, the length distribution of TSM events in individuals of African (NA19238) and European (NA12878) origin are similar. The inset shows lower values. Cases longer than 50 bp are grouped.

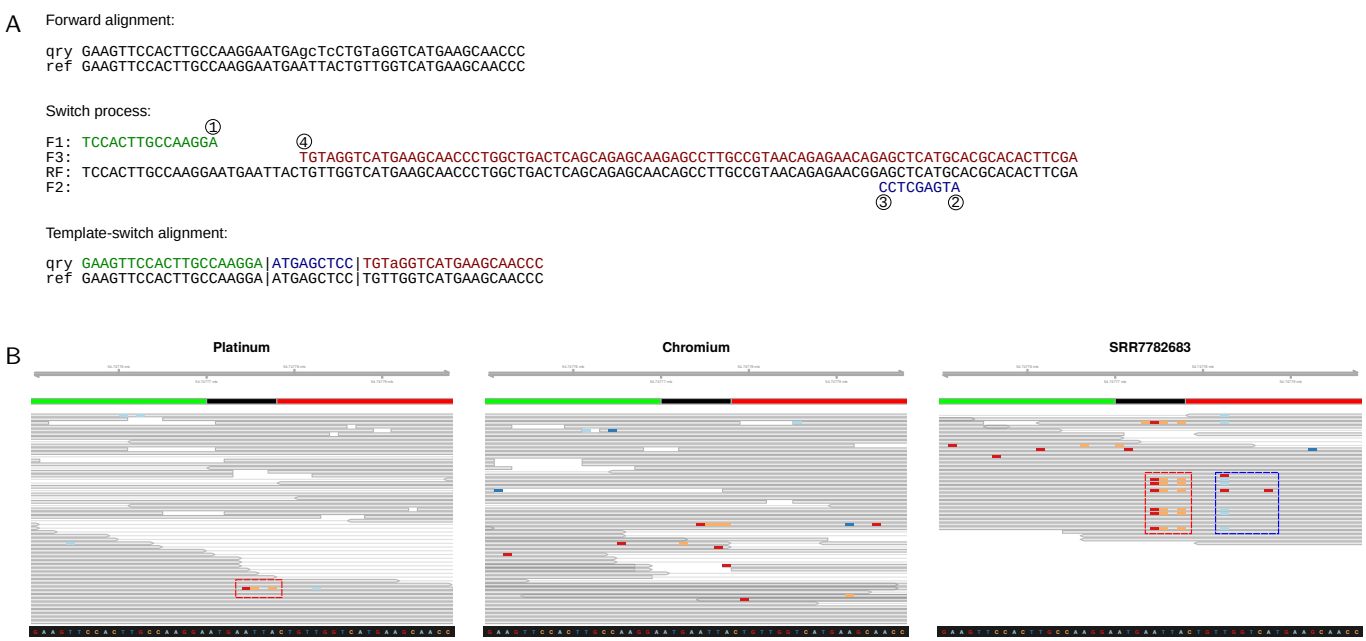


Fig S5. Inconsistency among different short-read datasets. (**A**) The TSM solution explaining a cluster of three base differences (and not explaining one nearby difference). (**B**) The Platinum data have one supporting read (left; red box) whereas the Chromium data have none; an independent dataset has multiple supporting reads (red box) but these are associated with flanking changes (blue box) creating more than two haplotypes. A plausible explanation is mismapping of reads from another locus. Data were visualised with IGV [54].

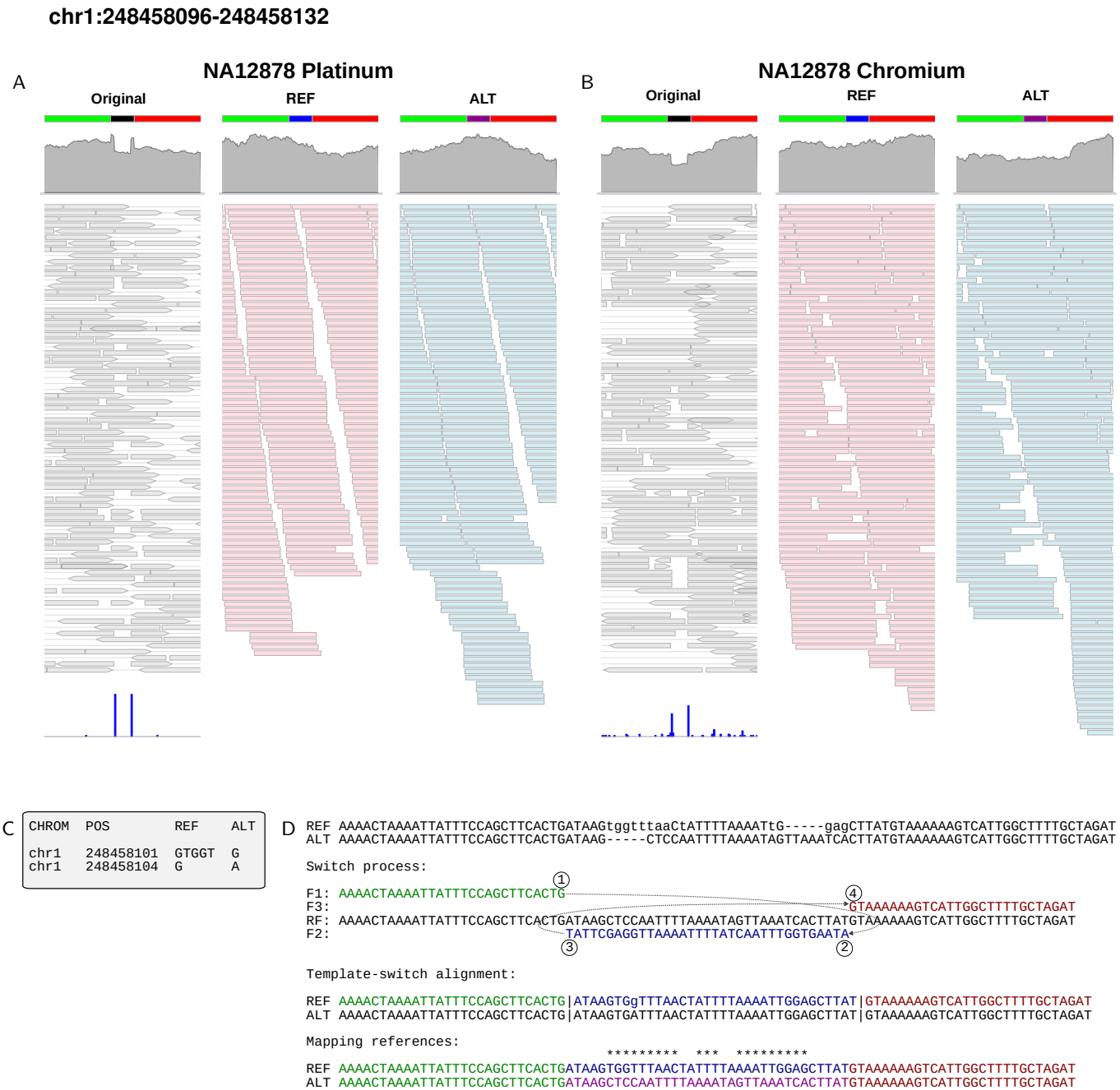


Fig S6. A heterozygous TSM event identified in NA12878 with short-read data. (**A**) Platinum data (left) and (**B**) Chromium data (left) have excess of soft clips (bottom, blue bars) and atypical patterns in the mapping coverage (top, in gray). (**C**) Two variants were called in this region by Ebert *et al.* [12] but these fail to explain the haplotype differences. (**D**) De novo assembly of the reads creates two locally highly dissimilar haplotypes (top). All but one difference 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. Data were visualised with Gviz [55].

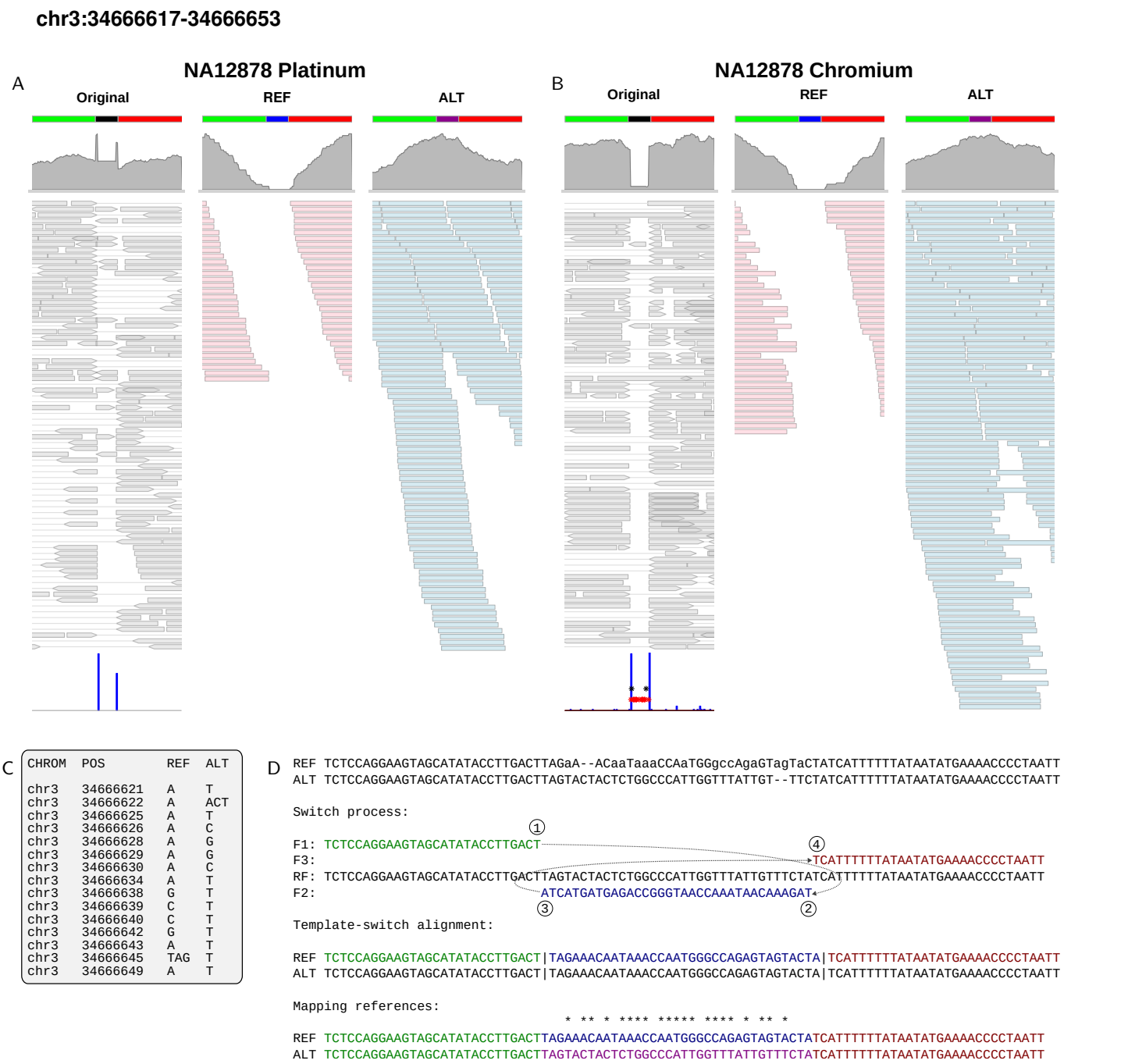


Fig S7. A homozygous TSM event identified in NA12878 with short-read data. (**A**) Platinum data (left) and (**B**) Chromium data (left) have excess of soft clips (bottom, blue bars) and atypical patterns in the mapping coverage (top, in gray). (**C**) Fifteen variants fully explaining the differences were called in the Chromium data while no variants were indicated in this region in the Platinum data or by Ebert *et al.* [12] (**D**) De novo assembly of the reads creates two locally highly dissimilar haplotypes (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 on the ALT haplotype only (**A,B**; middle, right). Data were visualised with Gviz [55].

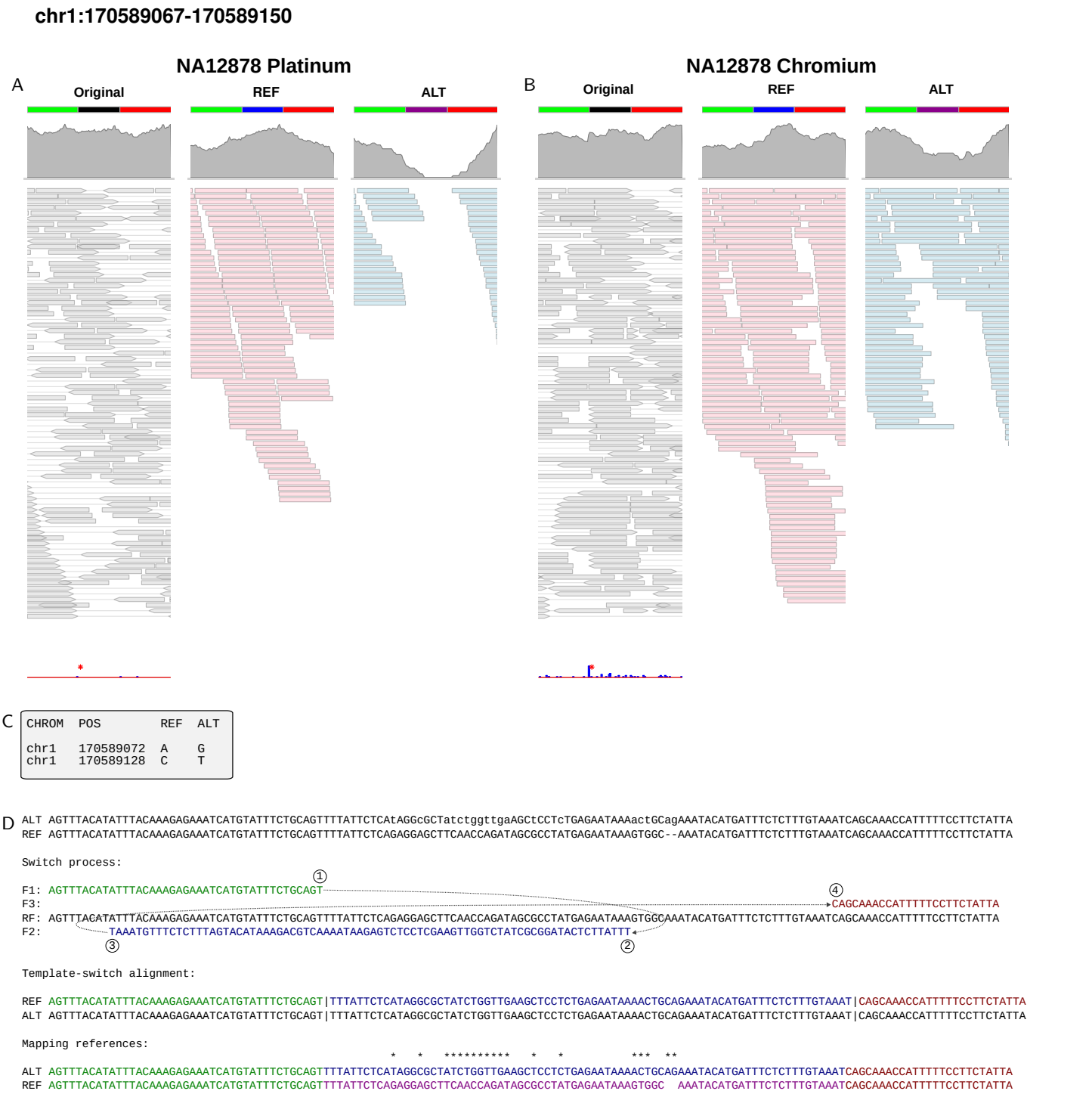


Fig S8. A potential de novo TSM event identified in NA12878 Chromium short-read data. (**A**) Platinum data (left) show no signal whereas (**B**) Chromium data (left) have some soft clips (bottom, blue bars) in this region. (**C**) One or two SNVs were called in this region by different datasets. (**D**) De novo assembly of Chromium reads creates two locally highly dissimilar haplotypes (top). All differences can be explained with a TSM event (middle). Using two haplotypes with alternative central parts (blue and magenta; bottom) as the reference, a fraction of Chromium reads map in full length on the ALT haplotype (**B**; middle, right). Data were visualised with Gviz [55].

A

CHR	START	END	ID	REF, ALT(s)
chr1	248458100	248458100	rs1294550520	A, C
chr1	248458100	248458126	rs1553320400	AGTGGTTTAACTATTTTAAAATTGGAG, AG
chr1	248458101	248458101	rs945532251	G, C
chr1	248458102	248458101	rs370077035	-, C, CTCCAATTTTAAAA
chr1	248458102	248458104	rs1558344009	TGG, CTCCAA
chr1	248458102	248458109	rs774785163	TGGTTTAA, -
chr1	248458102	248458126	rs766858754	TGGTTTAACTATTTTAAAATTGGAG, CTCCAATTTTAAAATAGTTAAATCA
chr1	248458103	248458103	rs55778336	G, C, T
chr1	248458103	248458108	rs1312056438	GGTTTA, -
chr1	248458104	248458104	rs11488140	G, A, C
chr1	248458105	248458104	rs75431795	-, AA
chr1	248458108	248458108	rs61834525	A, T
chr1	248458108	248458111	rs1558344025	AACT, TAAA
chr1	248458109	248458109	rs1325281402	A, C
chr1	248458110	248458110	rs78430540	C, A
chr1	248458110	248458110	rs1283571306	C, -
chr1	248458111	248458111	rs79278902	T, A
chr1	248458112	248458111	rs759219409	-, CCA, TCA
chr1	248458112	248458112	rs1217534201	A, C, T
chr1	248458112	248458112	rs1408812771	A, -
chr1	248458114	248458114	rs77716920	T, A
chr1	248458114	248458117	rs1558344038	TTTA, AGTT
chr1	248458115	248458115	rs78627666	T, G
chr1	248458117	248458117	rs61834526	A, T
chr1	248458119	248458119	rs1259249878	A, C
chr1	248458120	248458120	rs1442840868	A, G
chr1	248458122	248458122	rs56040868	T, C
chr1	248458122	248458124	rs752225071	TGG, -
chr1	248458122	248458126	rs386642020	TGGAG, AGTTAAATCA, CA
chr1	248458123	248458124	rs58657081	GG, -
chr1	248458124	248458124	rs1042592330	G, A, C
chr1	248458126	248458126	rs866296781	G, A
chr1	248458126	248458126	rs58743262	G, -
chr1	248458127	248458126	rs373041964	-, TTAAACCA, TTAAATCA
chr1	248458127	248458128	rs1553320408	CT, CTCCAATTTTAAAATAGTTAAATCACT
chr1	248458129	248458129	rs1003776664	T, C
chr1	248458130	248458130	rs1208388863	A, G

```
Alignment:

REF GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTGATAAGTGGTTTAACTATTTTAAAATTGGAGCTTATGTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC
ALT GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTGATAAGCTCCAATTTTAAAATAGTTAAATCACTTATGTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC

Switch process:

                                                        ①                                                       ④
F1: GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTG
F3:                                                                                                GTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC
RF: GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTGATAAGCTCCAATTTTAAAATAGTTAAATCACTTATGTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC
F2:                                                        TATTCGAGGTTAAAATTTTATCAATTTGGTGAATA
                                                        ③                                                       ②

Template-switch alignment:

REF GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTG|ATAAGTGgTTTAACTATTTTAAAATTGGAGCTTAT|GTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC
ALT GAAGCAAAAAAACACCAAAAAACTAAAATTATTTCCAGCTTCACTG|ATAAGTGATTTAACTATTTTAAAATTGGAGCTTAT|GTAAAAAAGTCATTGGCTTTTGCTAGATTCTTCATATATTCCTTC
```

B

CHR	START	END	ID	REF, ALT(s)
chr1	248437366	248437371	rs1282560120	TTACTT, TT
chr1	248437373	248437383	rs1362375318	AGTTGTGGAAG, AG
chr1	248437375	248437375	rs1403073841	T, A, C
chr1	248437376	248437376	rs1282373520	T, G
chr1	248437378	248437378	rs1451470978	T, C
chr1	248437381	248437382	rs1379895247	AA, A
chr1	248437388	248437387	rs879403957	-, TG
chr1	248437389	248437390	rs1234910043	CC, -
chr1	248437390	248437391	rs879504096	CA, -
chr1	248437396	248437397	rs369794123	CT, CTTCCACAACT
chr1	248437398	248437397	rs1205190376	-, TCC
chr1	248437399	248437398	rs1287619838	-, CAACTG

```
Alignment:

REF AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTTTAGTTGTGGAAGAAAA--GCCATTTTCT---------AGAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA
ALT AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTTTAG---------AAAATGGC--TTTTCTTCCACAACTAGAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA

Switch process:

                                                        ①                                                       ④
F1: AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTT
F3:                                                                                                GAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA
RF: AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTTTAGAAAATGGCTTTTCTTCCACAACTAGAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA
F2:                                                        ATCTTTTACCGAAAAGAAGGTGTTGAT
                                                        ③                                                       ②

Template-switch alignment:
REF AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTT|TAGTTGTGGAAGAAAAGCCATTTTCTA|GAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA
ALT AATCAAAACTTCTCTATCGTTCTAAAGATTCCCCAAGTGATTTTACTT|TAGTTGTGGAAGAAAAGCCATTTTCTA|GAAGTGACACTTCATGTGAAAATACTGCTGGGGAGGAAACATAATGA
```

Fig S9. Description of TSM variants in the dbSNP database. (**A**) TSM locus is described in full with a MNV of 25 bp (cyan) but the same changes are also given as multiple SNVs and indels. (**B**) TSM haplotypes can be reconstructed from four different variants (cyan, magenta) but the region has also additional variants annotated.

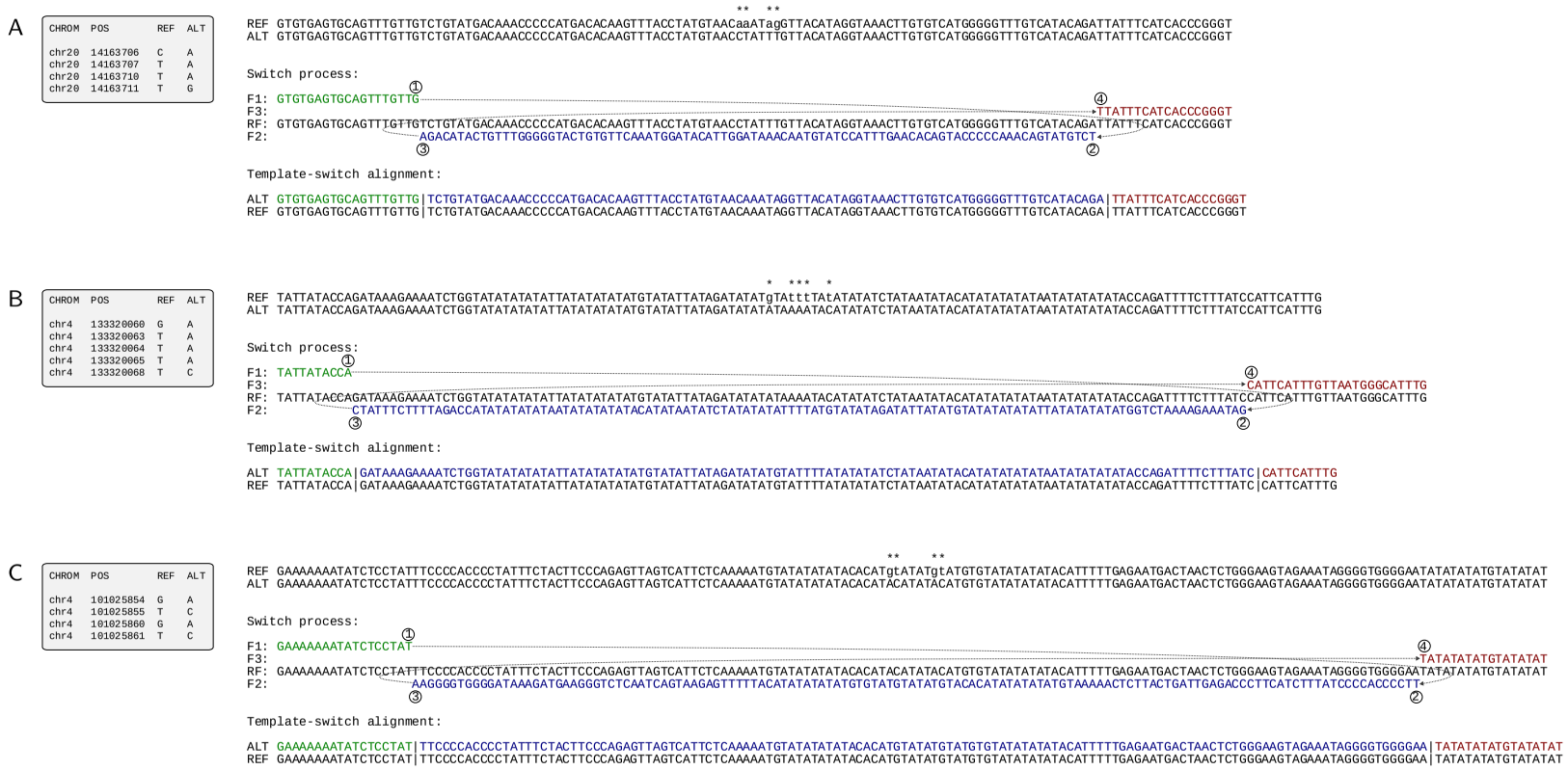


Fig S10. The three longest TSM events explaining gnomAD MNVs. (**A**) The longest TSM event within complex sequence is 90 bp long and explains a MNV of four base differences. (**B,C**) The longest TSM events involving VNTRs, 119 and 125 bp in length, explain MNVs of five and four base differences.

A

CHROM	POS	REF	ALT
chr10	53250728	A	G
chr10	53250729	A	C
chr10	53250732	G	A
chr10	53250734	A	G
chr10	53250738	C	G
chr10	53250743	T	C
chr10	53250757	C	A
chr10	53250766	A	T
chr10	53250767	G	A
chr10	53250770	G	A
chr10	53250771	C	A
chr10	53250796	T	G
chr10	53250797	T	C
chr10	53250800	T	C
chr10	53250801	A	T
chr10	53250804	A	T
chr10	53250805	C	G
chr10	53250810	T	G
chr10	53250811	T	C

B

```
                                                  **  * *   *    *            *        **  **                        **  **  **    **
ALT ACTTGAAATGGAAGAAAGGGTTTGGATAGgcAGaAgCACgCCAAcATGAAAAGTGAGGaCCTACATCtaTGaaATCATGGGACTCCTTCTAGGTGATgcCActGAtgTAGGgcCTCACTTTTCATGTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG
REF ACTTGAAATGGAAGAAAGGGTTTGGATAGAAAGGAACACCCCAATATGAAAAGTGAGGCCCTACATCAGTGGCATCATGGGACTCCTTCTAGGTGATTTCATAGAACTAGGTTCTCACTTTTCATGTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG

Switch process:

                  ①                                                                  ④                ⑤                                  ⑧
FW: ACTTGAAATGGAAGA                                                                  TGGGACTCCTTCTAGG                                   GTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG
RF: ACTTGAAATGGAAGAAAGGGTTTGGATAGAAAGGAACACCCCAATATGAAAAGTGAGGCCCTACATCAGTGGCATCATGGGACTCCTTCTAGGTGATTTCATAGAACTAGGTTCTCACTTTTCATGTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG
BW:                                                  TACTTTTCACTCCGGGATGTAGTCACCGTAGT                ACTAAAGTATCTACATCCAGGAGTGAAAAGTACAACCGCACGAAGACGGATAGGTTTGGGAA
                                                     ⑦                                ⑥                ③                  **    *                                                  ②

Template-switch alignment:

                                                                                  *     **
ALT ACTTGAAATGGAAGA|AAGGGTTTGGATAGGCAGAAGCACGCCAACATGAAAAGTGAGgACCTAcaTCTATGAAATCA|TGGGACTCCTTCTAGG|TGATGCCACTGATGTAGGGCCTCACTTTTCAT|GTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG
REF ACTTGAAATGGAAGA|AAGGGTTTGGATAGGCAGAAGCACGCCAACATGAAAAGTGAGAACCTAGTTCTATGAAATCA|TGGGACTCCTTCTAGG|TGATGCCACTGATGTAGGGCCTCACTTTTCAT|GTTGGCGTGCTTCTGCCTATCCAAACCCTTGAGATCCAGG
```

Fig S11. Large gnomAD MNV explained by two TSMs. (**A**) A cluster of 19 base changes in gnomAD MNV data. (**B**) Of the 19 base differences (top), 16 can be explained with two TSM events (middle, bottom). The DNA replication proceeds from left to right (green); at ①, the replication jumps to ② in the opposite strand and proceeds in reverse-complement manner until ③ (blue); it then jumps to ④ in the original template and proceeds until ⑤ (magenta); the replication again changes the template (either to the opposite strand or to the newly synthesized strand), jumps to ⑥ and continues until in reverse-complement manner until ⑦ (orange); the replication makes a final jump to ⑧ in the original template and continues normally (red).

```
A 1:246254743-246254746

  Alignment:
  qry CTTTTCCACTCGGAAGACCCCTCTCcATtGAGAGAGAGCTGGTTTTCTTTATTC
  ref CTTTTCCACTCGGAAGACCCCTCTCAATGGAGAGAGAGCTGGTTTTCTTTATTC

  Switch process:

  F1: TCCTCCTTTCCTGTCTGCACCCACCATACCTCCTGTCTTTTCCACTCGGAAGACCC
  F3:                                                          AGAGCTGGTTTTCTTTATTCTTCTGCCTATTAAACCTCAGCACAAAAAAAAAAAA
  RF: TCCTCCTTTCCTGTCTGCACCCACCATACCTCCTGTCTTTTCCACTCGGAAGACCCCTCTCAATGGAGAGAGAGCTGGTTTTCTTTATTCTTCTGCCTATTAAACCTCAGCACAAAAAAAAAAAA
  F2:                                                            GAGAGTTACCTCTC

  Template-switch alignment:

  qry CTTTTCCACTCGGAAGACCC|CTCTCCATTGAGAG|AGAGCTGGTTTTCTTTATTC
  ref CTTTTCCACTCGGAAGACCC|CTCTCCATTGAGAG|AGAGCTGGTTTTCTTTATTC


  2:111493177-111493179

  Alignment:
  qry CATCGGCTCCCAGGGGCAAGgctcccGgAGGCACAGCCCAGGTTCCTCTT
  ref CATCGGCTCCCAGGGGCAAG------GAAGGCACAGCCCAGGTTCCTCTT

  Switch process:

  F1: CATCGGCTCCCAGGGGCAAG
  F3:                      GCACAGCCCAGGTTCCTCTTCTAAGGAGTGACTTTCTCTCTCCCAAGGATGAGCCTGCTGGGCTAAGTCCTCAGACTCCGGGAGCATCCTGGCTGCTG
  RF: CATCGGCTCCCAGGGGCAAGGAAGGCACAGCCCAGGTTCCTCTTCTAAGGAGTGACTTTCTCTCTCCCAAGGATGAGCCTGCTGGGCTAAGTCCTCAGACTCCGGGAGCATCCTGGCTGCTG
  F2:                                                                                                                GAGGCCCTCG

  Template-switch alignment:

  qry CATCGGCTCCCAGGGGCAAG|GCTCCCGGAG|GCACAGCCCAGGTTCCTCTT
  ref CATCGGCTCCCAGGGGCAAG|GCTCCCGGAG|GCACAGCCCAGGTTCCTCTT


  8:15047723-15047727

  Alignment:
  qry ATTTAATTTTATATCCCTTT-GGtAAAGCTTGTTATACCTTTAAGTAAC
  ref ATTTAATTTTATATCCCTTTGGGAAAAGCTTGTTATACCTTTAAGTAAC

  Switch process:

  F1: GTCTGCTTTACCTGAAATTCTCCTCTGTACACTCAAACCCTACTCTCACCACCATAACACCAACGGCAAAATCCTTGGCTCATTTAATTTTATATCCCTTT
  F3:                                                                                                         TTGTTATACCTTTAAGTAAC
  RF: GTCTGCTTTACCTGAAATTCTCCTCTGTACACTCAAACCCTACTCTCACCACCATAACACCAACGGCAAAATCCTTGGCTCATTTAATTTTATATCCCTTTGGGAAAAGCTTGTTATACCTTTAAGTAAC
  F2:     CGAAATGG

  Template-switch alignment:

  qry ATTTAATTTTATATCCCTTT|GGTAAAGC|TTGTTATACCTTTAAGTAAC
  ref ATTTAATTTTATATCCCTTT|GGTAAAGC|TTGTTATACCTTTAAGTAAC


  X:1316241-1316242

  Alignment:
  qry CTGAGCTTGCTCAGACACACCCTCcaGGGAAGGCAGGACACCCTCCCTC
  ref CTGAGCTTGCTCAGACACACCCTCTGGGGAAGGCAGGACACCCTCCCTC

  Switch process:

  F1: ATTACTTTCAGATGGGGCCCTGGAGGATGACAGAAGGCTGAGCTTGCTCAGACACAC
  F3:                                                           AAGGCAGGACACCCTCCCTCCCACCTCGAAGGCAGGACACCCTCCCTCCCACCTCCC
  RF: ATTACTTTCAGATGGGGCCCTGGAGGATGACAGAAGGCTGAGCTTGCTCAGACACACCCTCTGGGGAAGGCAGGACACCCTCCCTCCCACCTCGAAGGCAGGACACCCTCCCTCCCACCTCCC
  F2:                   GGGACCTCC

  Template-switch alignment:

  qry CTGAGCTTGCTCAGACACAC|CCTCCAGGG|AAGGCAGGACACCCTCCCTC
  ref CTGAGCTTGCTCAGACACAC|CCTCCAGGG|AAGGCAGGACACCCTCCCTC

B 19:5181074-5181092

  Alignment:
  qry GGCTCACGCCTGTAAcTCCAGCACTTTGGGaGGcTGAGGCGGGCAGATCACAAGGTCAGGAGA
  ref GGCTCACGCCTGTAATTCCAGCACTTTGGGGGGTTGAGGCGGGCAGATCACAAGGTCAGGAGA

  Switch process:

  F1: GATCCACCCGCCTCAGCCTCCCAAAGTGCAATGTATCTATTTCTATGTAATTGTATCACATGTATATATATATTTTTAAGAGAGACAGGGTCTTGGGGCCGGGCGCGGTGGCTCACGCCTGTAACTCCA
  F3:                                                                                                                                     CAGATCACAAGGTCAGGAGA
  RF: GATCCACCCGCCTCAGCCTCCCAAAGTGCAATGTATCTATTTCTATGTAATTGTATCACATGTATATATATATATTTTTAAGAGAGACAGGGTCTTGGGGCCGGGCGCGGTGGCTCACGCCTGTAATTCCAGCACTTTGGGGGGTTGAGGCGGGCAGATCACAAGGTCAGGAGA
  F2:       GGGCGGAGTCGGAGGGTTTCACG

  Template-switch alignment:

  qry GGCTCACGCCTGTAAcTCCA|GCACTTTGGGAGGCTGAGGCGGG|CAGATCACAAGGTCAGGAGA
  ref GGCTCACGCCTGTAATTCCA|GCACTTTGGGAGGCTGAGGCGGG|CAGATCACAAGGTCAGGAGA
```

Fig S12. De novo MNVs explainable with the TSM mechanism. (**A**) Four MNVs in the third-generation individuals of Sasani *et al.* [40] that can be explained with a single TSM. (**B**) A TSM solution that can explain two of the three clustered de novo SNVs in a third-generation individual of Sasani *et al.* [40]

Table S1. Number and length of inferred TSM events as well as proportion of homozygous loci and resolved and ancestral (of former) allelic state in different samples. Only unmasked loci are included.

				Length (bp)		Allele (%)		Polarization (%)	
ID	Population	Region	n	average	maxim.	homozyg.	singleton	resolved	ancestral
HG02011	ACB	AFR	1046	10.67	171	30.50	7.58	89.50	49.68
NA19983	ASW	AFR	1077	10.77	171	27.48	8.41	88.03	53.74
HG03125	ESN	AFR	1073	10.43	171	29.45	7.12	88.01	51.32
HG03371	ESN	AFR	1070	10.44	171	31.40	8.43	89.09	52.57
HG02587	GWD	AFR	1097	10.81	180	26.71	9.59	89.44	51.73
HG02818	GWD	AFR	1038	10.64	100	29.38	8.07	88.07	53.01
HG03065	MSL	AFR	1060	10.48	171	30.94	8.81	89.84	53.09
HG03486	MSL	AFR	1064	10.90	180	28.85	8.39	88.17	52.08
NA19238	YRI	AFR	1121	10.44	97	28.37	8.60	90.29	51.28
NA19239	YRI	AFR	1092	10.47	171	29.30	8.91	89.03	52.46
NA12329	CEU	EUR	922	10.70	180	35.03	3.07	88.84	52.32
NA12878	CEU	EUR	921	10.49	171	36.48	2.61	89.48	54.42
HG00171	FIN	EUR	883	10.28	171	34.65	3.63	90.05	54.27
HG00096	GBR	EUR	937	10.67	171	36.07	3.11	89.23	53.17
HG01505	IBS	EUR	920	10.79	180	32.83	4.03	89.46	52.61
NA20509	TSI	EUR	882	10.47	180	35.49	4.52	89.12	55.34
NA24385	AZK	EUR	872	10.48	171	36.24	2.71	89.58	54.86
HG00864	CDX	EAS	942	10.61	171	39.17	4.33	88.67	50.42
NA18534	CHB	EAS	923	10.59	171	39.98	4.34	87.99	52.15
HG00512	CHS	EAS	917	10.52	171	39.91	2.82	89.12	50.67
HG00513	CHS	EAS	929	10.69	180	41.87	3.57	88.41	50.24
NA18939	JPT	EAS	927	10.50	171	42.29	3.16	89.76	50.30
HG01596	KHV	EAS	910	10.78	171	39.01	4.12	88.50	50.00
HG03009	BEB	SAS	878	10.45	171	35.76	3.85	88.52	52.76
NA20847	GIH	SAS	938	10.55	171	35.82	4.59	89.15	53.46
HG03732	ITU	SAS	920	10.37	171	35.87	3.57	88.15	52.77
HG02492	PJL	SAS	879	10.80	180	39.59	3.14	87.94	52.39
HG03683	STU	SAS	939	10.22	97	33.97	4.06	89.80	50.89
HG01114	CLM	AMR	923	10.64	171	33.69	3.86	87.66	53.83
NA19650	MXL	AMR	906	10.58	171	35.76	3.43	88.41	52.93
HG00731	PUR	AMR	949	10.63	171	34.56	3.09	88.55	52.91
HG00732	PUR	AMR	955	10.60	171	34.87	3.23	88.20	51.36

Table S2. Genotypes obtained with different datasets for variable TSM loci in NA12878 HaplotypeSV data.

		Platinum			
HaplotypeSV	Chromium	R|R	R/A	A|A	NA
NA	R|R	2	0	0	0
NA	R/A	0	71	1	3
NA	A|A	0	2	4	0
NA	NA	0	0	0	1
R/A	R|R	6	1	0	0
R/A	R/A	5	425	0	0
R/A	A|A	0	0	2	0
R/A	NA	1	1	0	0
A|A	R|R	2	0	0	0
A|A	R/A	0	9	7	0
A|A	A|A	0	3	310	3
A|A	NA	0	1	4	0

Table S3. Parental genotypes for HaplotypeSV TSM loci confirmed variable within NA12878 Chromium data.

		NA12892			
NA12878	NA12891	R|R	R/A	A|A	NA
R/A	R|R	1	80	35	4
R/A	R/A	65	113	53	14
R/A	A|A	39	42	2	2
R/A	NA	1	20	3	2
A|A	R|R	0	0	0	0
A|A	R/A	0	42	46	1
A|A	A|A	0	43	159	4
A|A	NA	0	3	6	3