Short template switch events in human evolution cause complex mutation patterns

Ari Löytynoja[1,*] and Nick Goldman[2]

[1]Institute of Biotechnology, University of Helsinki, Helsinki, Finland

[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK

[*]Corresponding author: ari.loytynoja@helsinki.fi

Abstract

Background. Resequencing efforts are uncovering the extent of genetic variation in humans and provide data to study the evolutionary processes shaping our genome. One recurring puzzle in both intra- and inter-species studies is the high frequency of complex mutations comprising multiple nearby base substitutions or insertion-deletions. We devised a generalized mutation model to study the role of template switch events in the origin of such mutation clusters.

Results. Applied to the human genome, our model detects thousands of template switch events during the evolution of human and chimp from their common ancestor, and hundreds of events between two independently sequenced human genomes. While many of these are consistent with the inter-strand template switch mechanism proposed for bacteria, our model also identifies new types of mutations that create short inversions, some flanked by paired inverted repeats. This local template switch process creates numerous complex mutation patterns, including secondary structures, and explains multi-nucleotide mutations and compensatory substitutions without invoking positive selection. Detection of these complex mutations with current resequencing methodologies is difficult and we find many erroneous variant annotations in human reference data.

Conclusions. Previously unexplained short template switch events account for a large number of complex mutation patterns in human evolution, without invoking complicated and speculative mechanisms or implausible coincidence. We show that clustered sequence differences are challenging for mapping and variant calling methods. Template switch events such as those we have uncovered may have been neglected as an explanation for complex mutations because of biases in commonly used analyses. Incorporation of our model into analysis pipelines will lead to improved understanding of genome variation and evolution.

Keywords: template switch events, complex mutations, multi-nucleotide mutations, human evolution, human resequencing

Background

Mutations are not evenly distributed in genome sequences and tend to form clusters of base substitutions or combinations of multiple substitutions and insertion-deletions (indels) (Averof et al., 2000; Harris and Nielsen, 2014; Sudmant et al., 2015; Whelan and Goldman, 2004). Explanations for the mutation clusters vary from an error-prone polymerase (Harris and Nielsen, 2014) to indels being mutagenic (Tian et al., 2008). In bacteria, mutations creating perfect inverted repeats occur with high frequency (Dutra and Lovett, 2006) and the mechanism behind this is thought to involve intra- or inter-strand template switching during DNA replication (Fig. 1a, b) (Ripley, 1990). Both template switch types can cause sequence changes within the repeat (Fig. 1a), while the latter can additionally invert the ‘spacer’ sequence (the region between the repeat fragments; Fig. 1b). While changes creating novel repeats can appear as clusters of differences (Dutra and Lovett, 2006), earlier studies have not considered the mechanism significant in the evolution of higher organisms (Ladoukakis and Eyre-Walker, 2008). These conclusions were based on limited data, however, and on an assumption that the mechanism necessarily creates perfect inverted repeats. We compared human and chimp genomes and observed mutation clusters that create novel inverted repeats consistent with the mechanism proposed for bacteria. Many clusters could only partially be explained by the creation of an inverted repeat, however, and novel repeats were often flanked by indels or dissimilar sequence, inconsistent with the classical model.

Mechanisms triggering template switching have been proposed, e.g. replication fork stalling (FoSTeS) (Lee et al., 2007), replication slippage (Chen et al., 2005) and microhomology-mediated break-induced replication (MMBIR) (Hastings, Ira and Lupski, 2009), but mutations attributed to these typically involve major genomic rearrangements (Costantino et al., 2013; Hastings, Lupski, Rosenberg and Ira, 2009). Even with the underlying biological mechanism uncertain, we realized that the existence and properties of a template switch mutation process, capable of creating inverted repeats, could be studied using pairs of closely-related genome sequences. More specifically, we devised the ‘four-point model’, a generalized template switch model that projects the four sequence positions associated with a ‘switch-and-return’ event onto a reference sequence and then constructs a replication copy from the three fragments defined by these points. Assuming that replication proceeds from left (Ⓛ) to right (Ⓡ), and with points ① and ② indicating the location of the first switch event and ③ and ④ indicating the second (return) switch event, the replication copy then consists of fragments Ⓛ→①, ②→③, ④→Ⓡ (Fig. 1c–d). If fragment ②→③ overlaps with fragment Ⓛ→① or ④→Ⓡ, the mutation creates a novel inverted repeat that then may form a RNA secondary structure (Fig. 1e–f).

Modeling the template switch process like this has two major advantages. First, a model allows for a formal analysis of mutation events and their evaluation in comparison to alternative explanations. Second, our description of the process is general and has few *a priori* constraints for the template exchanges. Our projection of switch points onto a reference is impartial regarding the type of the switch event—either intra- or inter-strand—and the model only requires that the ②→③ fragment is copied in reverse-complement orientation. The possible outputs under the four-point model are defined by the relative order and distance of the switch points, and the classical mechanism proposed to explain inverted repeats in bacteria is a special case of our generalized model (cf. Fig. 1a,c). Supplementary Fig. 1 illustrates all the possible cases under the model, covering the scenarios described before (Fig. 1a–b) as well as several others, including creation of inverted and direct repeats flanked by dissimilar sequence and one case causing inversion of a sequence fragment only. An important characteristic of the model is that replacement of the ①→④ fragment with the reverse-complement of the ②→③ fragment by a single switch event can generate changes that, when viewed in a linear alignment, will appear as clusters of nearby substitutions and indels.

To test whether biological data support the proposed mechanism, we implemented a computational tool based on a custom dynamic programming algorithm that identifies clusters of differences between two aligned genomic sequences and then searches for an explanation of the region of dissimilarity in one sequence (replicate output) by copying a fragment from the other sequence (reference) in reverse-complement orientation, as achieved in the four-point model (see Methods, Supplementary Fig. 2 and Supplementary Algorithm 1). With two closely-related sequences, parallel mutation will be rare and we arbitrarily designate one sequence as the reference and assume that it represents the ancestral form around each mutation event in the replicate lineage. We applied this method to genome-wide Ensembl EPO alignments (Flicek et al., 2013; Paten et al., 2008) (v.71, 6 primates) of human and chimp, considering the chimp sequence the reference and the human sequence the mutated copy. We focused on the complex and unique regions of the genome and compared the solutions involving a template switch to the original linear sequence alignments. From the potential cases of template switch events detected, we filtered a set of high-confidence events (see Methods). To create a control to assess false positives, we used a proxy for observing the mutation patterns by chance: we computed the best solutions explaining the dissimilar sequence regions with the fragment ②→③ copied in reverse (i.e., not reverse-complement) orientation and evaluated these solutions using the same criteria. We performed similar analyses on resequenced human genome data, detecting numerous polymorphic loci. We then investigated the evidence for these mutations in variant calls from the 1000 Genomes (1kG) project (1000 Genomes Project Consortium et al., 2015).

Results and discussion

Discovery of four-point mutation events from human-chimp data. We found 4,901 candidate events, spread across all human chromosomes in the human-chimp comparisons. Some candidate events were consistent with the original mechanism proposed for bacteria and convert a near-perfect inverted repeat into a perfect one (see example in Fig. 2a–b) but the majority were associated with large sequence changes, causing multiple base differences and indels in linear alignments (e.g. Fig. 2c–d). While any complex mutation could be generated by a combination of simple, ‘traditional’, mutations, Occam’s razor suggests that a four-point model template switch mutation is a better explanation than multiple substitutions and indels occurring in such a cluster. However, we also noticed that matches shorter than 12–13 bases are often found by chance (Supplementary Figs. 3, 4) and, despite strict filtering (see Methods), our list of candidate events might contain false positives. To get an unbiased picture of the process, we removed events with ②→③ fragment shorter than 14 bases. This was done to improve the signal to noise ratio and does not mean that short template switch events could not happen: in contrast, many cases with a short ②→③ fragment appear highly convincing (e.g. Fig. 1c).

After this filtering, we assigned the 802 remaining candidate events to specific event types based on the relative positions of the switch points and computed their frequencies. We found that, of the 12 possible conformations of switch points, only six are present (Table 1, human vs. chimp comparison). Of these, two event pairs are mirror cases indistinguishable from one-another if both DNA strands are considered (see also Supplementary Fig. 1), and the six conformations observed therefore define four distinct switch event types. Type “1-4-3-2” (with its mirror case “3-2-1-4”; Supplementary Fig. 1a; e.g. Fig. 1c) creates an inverted repeat and accounts for 32% of the high-confidence events detected in the chimp-human comparison. Type “1-3-4-2” (with its mirror case “3-1-2-4”; Supplementary Fig. 1a; e.g. Fig. 1d) creates an inverted repeat separated by an inverted spacer sequence, accounting for 22% of events. The remaining two types are novel and only achievable under our four-point model: type “1-3-2-4”, accounting for 45% of events, only inverts a sequence fragment and creates no repeat (e.g. Fig. 2c), and type “3-1-4-2” creates two inverted repeats separated by an inverted spacer (e.g. Fig. 2d) and accounts for 1% of events.

The unifying feature of the event types theoretically possible under the model but not observed in real sequence data is that in the ordering of the switch points, ④ precedes ①. This is the hallmark of an event in which the second (return) template switch requires the opening of the newly synthesized DNA double helix (see Supplementary Fig. 1). In addition we observe numerous cases of inversion of spacer sequences; this cannot occur when ② precedes ①, a prerequisite of intra-strand switches. These discoveries suggest that template switches occur inter-strand: that is, the fragment ②→③ is copied from the opposite strand (Fig. 1). Although inversions of spacer sequences have been observed in bacteria (Ripley, 1990), the intra-strand mechanism has been the dominant hypothesis (Dutra and Lovett, 2006). It appears that this is not correct, at least for evolution since the human-chimp divergence. We also find that the relative frequencies of different event types are very different. In part this may be determined by factors such as the length distribution of the copied fragment (Supplementary Fig. 3) and type “3-1-4-2” requiring that the fragment ②→③ overlaps with both ① and ④. However, the frequencies of different event types may also reflect the properties of the mutation process, e.g. template switching benefiting from the proximity of the DNA strands, or the chances of the new mutation to escape error correction.

Identification of polymorphic mutations in human data. To understand whether template switch events are actively shaping human genomes, we analyzed human resequencing data and searched for polymorphic loci. We first aligned the human reference genome (GRCh37) to that of a Caucasian male (Venter, also denoted HuRef (Levy et al., 2007)), both based on classical capillary sequencing and assembled independently. We then considered Venter as the reference and identified clusters of mutations in GRCh37 that were consistent with different types of four-point model template switch events. After stringent filtering, we focused on a small number of high-confidence events to be studied more closely (see Methods). For these 88 events, the proportions of different event types were similar to those found in human-chimp comparisons. Again, only the six types not requiring opening of the new helix were found and the majority of events require inter-strand switches (Table 1; two humans comparison).

Still focusing on these 88 candidate events, we manually studied the Caucasian male sequence data mapped onto the reference genome (Li and Durbin, 2011). Despite some inconsistencies between the original Venter genome assembly and re-mapping of the sequence reads against GRCh37, we could resolve the genotype of the Caucasian male for 75 (85%) of the candidate events and found 40 of them heterozygous, i.e. the sequence data contain reads consistent with both Venter and GRCh37 alleles (Fig. 3a, b; Supplementary Table 1; see also Supplementary Data 1 available at http://loytynojalab.biocenter.helsinki.fi/software/fpa); in two cases the read data revealed that the mutations forming the cluster are not linked and are the result of two independent mutation events (Supplementary Fig. 5).

We then looked at the same loci in the 1kG data (1000 Genomes Project Consortium et al., 2015) and studied the alignment data for individual NA12878. We found that NA12878 has a non-reference allele at 46 loci (61%) and, with the exception of the two cases mentioned, all the changes are found within the same sequence reads. This finding has two implications. First, with two different sequencing technologies (capillary and Illumina) and analysis pipelines showing the same mutation patterns, we can reject the possibility that the observed events could be technical artefacts. Second, the agreement of short-read data with assembly based on long capillary reads suggests that the template switch mutation mechanism can be studied using modern resequencing data.

Elimination of mutation accumulation hypothesis. In principle, the perfect linkage of adjacent sequence changes in two unrelated individuals could also be explained by mutations being accumulated over a long period of time in complete absence of recombination. To rule that out, we assessed the maximum age of the mutation clusters using phylogenetic information (Fig. 3c). The EPO alignments contain data from at least two additional primate species for all but one of the 75 loci. The two alleles detected between the two humans GRCh37 and Venter segregate among the primate species in only one of these loci; in all 73 other cases, all primate sequences resemble one of the two human alleles while the second human allele is unique (Fig. 3c; Supplementary Data 2, http://loytynojalab.biocenter.helsinki.fi/software/fpa). Although some loci could be polymorphic in non-human primates, the result suggests that a great majority of the mutation events are young and the adjacent changes result from a single mutation event.

Mutation clusters in 1000 Genomes variation data. NA12878 is only one individual and a greater proportion of the 75 candidate loci may be truly polymorphic in larger samples. We investigated whether the mutations caused by template switch events are visible in variation data. Using the 1kG variant calls (1000 Genomes Project Consortium et al., 2015) we found that this is indeed the case: of the 75 confirmed events between the reference and the Caucasian male, the mutation pattern created by the event is completely explained by combinations of the 1kG variants (separate calls of indels and SNPs) at 35 loci, and partially explained at a further 15 loci. In most cases, the mutations at a locus have uniform allele frequencies within human populations, further demonstrating the perfect linkage and the single origin for the full mutation cluster (Fig. 3d; Supplementary Data 1, http://loytynojalab.biocenter.helsinki.fi/software/fpa). The variation data confirm the two earlier cases as combinations of independent mutations (Supplementary Fig. 5) but, for all other inconsistencies, alignment data show the incomplete mutation patterns and the non-uniform allele frequencies to be artefacts from erroneous mapping and variant calling (Supplementary Fig. 6). Such inconsistencies are expected when the variant calls are based on mapping of short reads containing multiple differences to a reference sequence, and demonstrate the difficulty of correctly detecting complex mutations using current analysis methods.

Despite highly uniform allele frequencies, the 1kG variant calls consider the template switch events that we identified to be clusters of independent mutations events—the largest clusters consisting of more than ten apparently independent mutation events (Supplementary Fig. 7)—and thus seriously exaggerate the estimates of local mutation rate. On the other hand, if alleles were correctly called, uniform frequencies at adjacent positions would indicate a shared history for a mutation cluster and potentially allow computational detection of events. To test this, we turned back to the events found between human and chimp and studied if any of these are still polymorphic in humans and show uniform allele frequencies (see Methods). We found several such events, the frequencies of the two haplotypes varying from close to 0 to nearly 1, and the frequencies differing significantly between populations (Supplementary Fig. 8). This finding demonstrates two things: first, a greater number of loci than were detected by a comparison of two human individuals are polymorphic and segregate amongst human populations; and second, if the read mapping and variant calling were perfect, variation data combined with variant sequence reconstruction could be used for *de novo* computational detection of template switch mutations. Under the same constraints, the approach could also be applied to resequencing data from trios.

Conclusions

Our generalized template switch model can explain a large number of complex mutation patterns—clusters of apparent base substitutions and indels—with a single mutation event. Although we do not find evidence of those events that require opening of the newly synthesized DNA strand, the model significantly extends the one previously proposed for bacteria. First, unlike the bacterial model, pre-existing sequence similarity is not required and the process can thus create completely novel repeats (see Fig. 2). This is consistent with the reported cases of major genomic rearrangements where microhomology of only two or three bases is observed at the switch points (Costantino et al., 2013; Hastings, Ira and Lupski, 2009; Lee et al., 2007). Second, the most common event type we detected only inverts a sequence fragment, with no or very short inverted repeats. Such mutations are not considered by the original bacterial model, which focuses on long inverted repeats.

When the template switch event does not involve loss or gain of sequence, the mutation pattern appears as a multi-nucleotide substitution (MNS). Some cases of MNSs have been explained with positive selection (Bazykin et al., 2004; Meer et al., 2010) while involvement of Pol ζ has been suggested to explain spatial differences in mutation frequency (Harris and Nielsen, 2014). Our results demonstrate that template switch mutations are also playing a role in the creation of clusters of adjacent substitutions. Interestingly, we cannot explain the cases of MNS shown in Schrider et al. (2011) as local template switch events. It has been shown that switch events can take place between distant loci (Costantino et al., 2013; Hastings, Ira and Lupski, 2009; Lee et al., 2007) and it is plausible that the same mechanism is still involved; a copy event from a distant locus would create a MNS but no local secondary structure. Many template switch events are associated with indels in the alignment (Supplementary Fig. 9) and the process we identified provides an alternative to the proposition of indels being mutagenic and triggering nearby base substitutions (Tian et al., 2008).

The proposed four-point model has consequences for our understanding of genome evolution and the methods used for studying it. It provides a one-step mechanism for the generation of hair-pin loops and, in combination with other mutations, provides a pathway to more complex secondary structures (Ding et al., 2014; Rouskin et al., 2014; Wan et al., 2014). The model also provides a mechanism for the evolution of existing DNA secondary structures and provides an explanation for the long-standing dilemma of exceptionally high rates for compensatory substitutions (Dixon and Hillis, 1993; Meer et al., 2010; Tillier and Collins, 1998). Interestingly, the mechanism may also maintain apparent DNA secondary structures without selective force.

A probable reason why template switch mutations have not received greater attention may be bias in commonly used analysis methods. Tight clusters of differences, the typical signature of the process, make read mapping and subsequent variant calling challenging. This is demonstrated by phase 3 of the 1kG Project (1000 Genomes Project Consortium et al., 2015), which provides significant improvements in comparison to earlier releases but, as we have shown, still contains errors and inconsistencies around the regions we have studied. The new mutation mechanism we propose could be modeled and considered in future analyses. With improvements in relevant algorithms, the full extent of local template switch events could be uncovered.

Methods

Discovery of four-point mutations. We downloaded the Ensembl (v.71) EPO alignments (Flicek et al., 2013; Paten et al., 2008) of six primates and included all blocks containing only one human and chimp sequence, covering in total 2.648 Gb of the human sequence and 94.8% of the EPO alignment regions. Keeping only human and chimp sequences, we identified alignment regions where two or more non-identical bases (mismatches or indels) occur within a 10-base window. For each such mutation cluster, we considered the surrounding sequence (for human and chimp, respectively, 100 and 200 bases up- and downstream from the cluster boundaries), and in accordance with our four-point model attempted to reconstruct the human query from the chimp reference with imperfect copying (allowing for mismatches and indels) of the forward strand and two freely placed template switch events. Candidate switch events were required to have high sequence similarity without alignment gaps and within the ②→③ fragment only mismatches were allowed. If exact positions of switch events could not be determined (Supplementary Fig. 10), our approach maximized the length of ②→③ fragment and reported this upper limit of the strand-switch event length. For comparison, we reconstructed the human query from the chimp reference with imperfect copying of the forward strand only (i.e., linear alignment) using the same scoring. A custom dynamic programming algorithm to determine the optimal four-point model explanation for each mutation cluster is described in Supplementary Fig. 2 and Supplementary Algorithm 1. The computational tool used for the analyses is available at http://loytynojalab.biocenter.helsinki.fi/software/fpa.

Filtering of events. For each mutation cluster, we recorded the coordinates of the inferred template switch events and computed similarity measures for the different parts of the template switch and forward alignments as well as the differences in the inferred numbers of mutations between the two solutions; we also recorded whether the regions include repeatmasked (Smit et al., 2013–2015) or dustmasked (Morgulis et al., 2006) sites, as well as the number of different bases included in the ②→③ fragments. We then selected a set of events as high-confidence candidates using the following criteria: (*i*) the switch points ① and ④ are at most 30 bases up- and downstream, respectively, from the cluster boundaries; (*ii*) the ②→③ fragment is at least 10 bases long; (*iii*) the ②→③ fragment as well as 40-base flanking regions up- and downstream show at least 95% identity between the sequences; (*iv*) the forward alignment indicates at least two differences (of which at least one a mismatch) more than the template switch alignment (which may also contain up to 5% mismatches); (*v*) the ②→③ fragment is not repeatmasked or dustmasked and contains all four bases. As a control to assist in assessing the occurrence of false positives, we repeated the analysis without complementing the ②→③ fragment: no biological function is known for reverse repeats and we consider them a proxy for the probability of observing a repeat of particular length by chance.

Identification of polymorphic mutations. The GRCh37 human reference and Venter Caucasian male genome sequences were aligned using Lastz (Harris, 2007) and following the UCSC analysis pipeline (Kent et al., 2002). The four-point mutations were identified using the same approach as with human-chimp data. The 1kG variation data from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ were analysed using bcftools (Li, 2011) and selected regions of resequencing data from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878 were visualized using samtools (Li et al., 2009). Mutation clusters with uniform allele frequencies were identified as follows: (*i*) 1kG variant calls were extracted for the mutation cluster plus 10 bases of flanking region; (*ii*) for each locus, runs of adjacent positions with less 10% difference in global allele frequency (AF) were recorded; and (*iii*) the runs of selected length (e.g. 3) with AF between 0.01 and 0.99 were outputted. The 1kG variant alleles were reconstructed using GATK (McKenna et al., 2010). Short-read alignment data, 1kG variant calls and primate sequence alignments for the candidate template-switch event loci are available at http://loytynojalab.biocenter.helsinki.fi/software/fpa.

Other computational analyses. DNA secondary structures were predicted with the ViennaRNA package (Lorenz et al., 2011), using the command 'RNAfold –noconv –noGU -P dna_mathews-2004.par'. The length distribution (Supplementary Fig. 3) and the allele frequencies (e.g. Fig. 3d) were visualized with R (R Core Team, 2014).

Acknowledgements

We thank Martin Taylor for help and comments in early stages of the study, and CSC - IT Center for Science, Finland, for computational resources.

Competing interests: The authors declare that they have no competing financial interests.

Authors' contributions: NG devised the extended model. AL implemented the method and performed the analyses. NG and AL designed the study, discussed the results and wrote the manuscript. Both authors read and approved the final manuscript.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. and Abecasis, G. R. (2015), 'A global reference for human genetic variation', *Nature* **526**(7571), 68–74. doi: 10.1038/nature15393.

Averof, M., Rokas, A., Wolfe, K. H. and Sharp, P. M. (2000), 'Evidence for a high frequency of simultaneous double-nucleotide substitutions', *Science* **287**, 1283–1286. doi: 10.1126/science.287.5456.1283.

Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S. and Kondrashov, A. S. (2004), 'Positive selection at sites of multiple amino acid replacements since rat-mouse divergence', *Nature* **429**, 558–562. doi: 10.1038/nature02601.

Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C. and Cooper, D. N. (2005), 'Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions', *Hum. Mutat.* **26**, 362–373. doi: 10.1002/humu.20230.

Costantino, L., Sotiriou, S. K., Rantala, J. K., Magin, S., Mladenov, E., Helleday, T., Haber, J. E., Iliakis, G., Kallioniemi, O. P. and Halazonetis, T. D. (2013), 'Break-induced replication repair of damaged forks induces genomic duplications in human cells', *Science* **343**, 88–91. doi: 10.1126/science.1243211.

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C. and Assmann, S. M. (2014), 'In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features', *Nature* **505**, 696–700. doi: 10.1038/nature12756.

Dixon, M. T. and Hillis, D. M. (1993), 'Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis', *Mol. Biol. Evol.* **10**, 256–267.

Dutra, B. E. and Lovett, S. T. (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.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. and Searle, S. M. J. (2013), 'Ensembl 2013', *Nucleic Acids Res.* **41**, D48–D55. doi: 10.1093/nar/gks1236.

Harris, K. and Nielsen, R. (2014), 'Error-prone polymerase activity causes multinucleotide mutations in humans', *Genome Res.* **24**, 1445–1454. doi: 10.1101/gr.170696.113.

Harris, R. (2007), '*Improved pairwise alignment of genomic DNA*. PhD thesis, Pennsylvania State University'.

Hastings, P. J., Ira, G. and Lupski, J. R. (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.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. and Ira, G. (2009), 'Mechanisms of change in gene copy number', *Nat. Rev. Genet.* **10**, 551–564. doi: 10.1038/nrg2593.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002), 'The human genome browser at UCSC', *Genome Res.* **12**, 996–1006. doi: 10.1101/gr.229102.

Ladoukakis, E. D. and Eyre-Walker, A. (2008), 'The excess of small inverted repeats in prokaryotes', *J. Mol. Evol.* **67**, 291–300. doi: 10.1007/s00239-008-9151-z.

Lee, J. A., Carvalho, C. M. B. and Lupski, J. R. (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, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H., Frazier, M. E., Scherer, S. W., Strausberg, R. L. and Venter, J. C. (2007), 'The diploid genome sequence of an individual human', *PLoS Biol.* **5**, e254. doi: 10.1371/journal.pbio.0050254.

Li, H. (2011), 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics* **27**, 2987–2993. doi: 10.1093/bioinformatics/btr509.

Li, H. and Durbin, R. (2011), 'Inference of human population history from individual whole-genome sequences', *Nature* **475**, 493–496. doi: 10.1038/nature10231.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* **25**, 2078–2079. doi: 10.1093/bioinformatics/btp352.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. and Hofacker, I. L. (2011), 'ViennaRNA package 2.0', *Algorithms Mol. Biol.* **6**, 26. doi: 10.1186/1748-7188-6-26.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010), 'The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res.* **20**, 1297–1303. doi: 10.1101/gr.107524.110.

Meer, M. V., Kondrashov, A. S., Artzy-Randrup, Y. and Kondrashov, F. A. (2010), 'Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness', *Nature* **464**, 279–282. doi: 10.1038/nature08691.

Morgulis, A., Gertz, E. M., Schäffer, A. A. and Agarwala, R. (2006), 'A fast and symmetric

DUST implementation to mask low-complexity DNA sequences', *J. Comput. Biol.* **13**, 1028–1040. doi: 10.1089/cmb.2006.13.1028.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I. and Birney, E. (2008), 'Genome-wide nucleotide-level mammalian ancestor reconstruction', *Genome Res.* **18**, 1829–1843. doi: 10.1101/gr.076521.108.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org*

Ripley, L. S. (1990), 'Frameshift mutation: determinants of specificity', *Annu. Rev. Genet.* **24**, 189–213. doi: 10.1146/annurev.ge.24.120190.001201.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J. S. (2014), 'Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo', *Nature* **505**, 701–705. doi: 10.1038/nature12894.

Schrider, D. R., Hourmozdi, J. N. and Hahn, M. W. (2011), 'Pervasive multinucleotide mutational events in eukaryotes', *Curr. Biol.* **21**, 1051–1054. doi: 10.1016/j.cub.2011.05.013.

Smit, A. F. A., Hubley, R. and Green, P. (2013–2015), *RepeatMasker Open-4.0.*
URL: *http://www.repeatmasker.org*

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E. and Korbel, J. O. (2015), 'An integrated map of structural variation in 2,504 human genomes', *Nature* **526**, 75–81. doi: 10.1038/nature15394.

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J. and Chen, J.-Q. (2008), 'Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes', *Nature* **455**, 105–108. doi: 10.1038/nature07175.

Tillier, E. R. and Collins, R. A. (1998), 'High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA', *Genetics* **148**, 1993–2002.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E. and Chang, H. Y. (2014), 'Landscape and variation of RNA secondary structure across the human transcriptome', *Nature* **505**, 706–709. doi: 10.1038/nature12946.

Whelan, S. and Goldman, N. (2004), 'Estimating the frequency of events that cause multiple-nucleotide changes', *Genetics* **167**, 2027–2043. doi: 10.1534/genetics.103.023226.

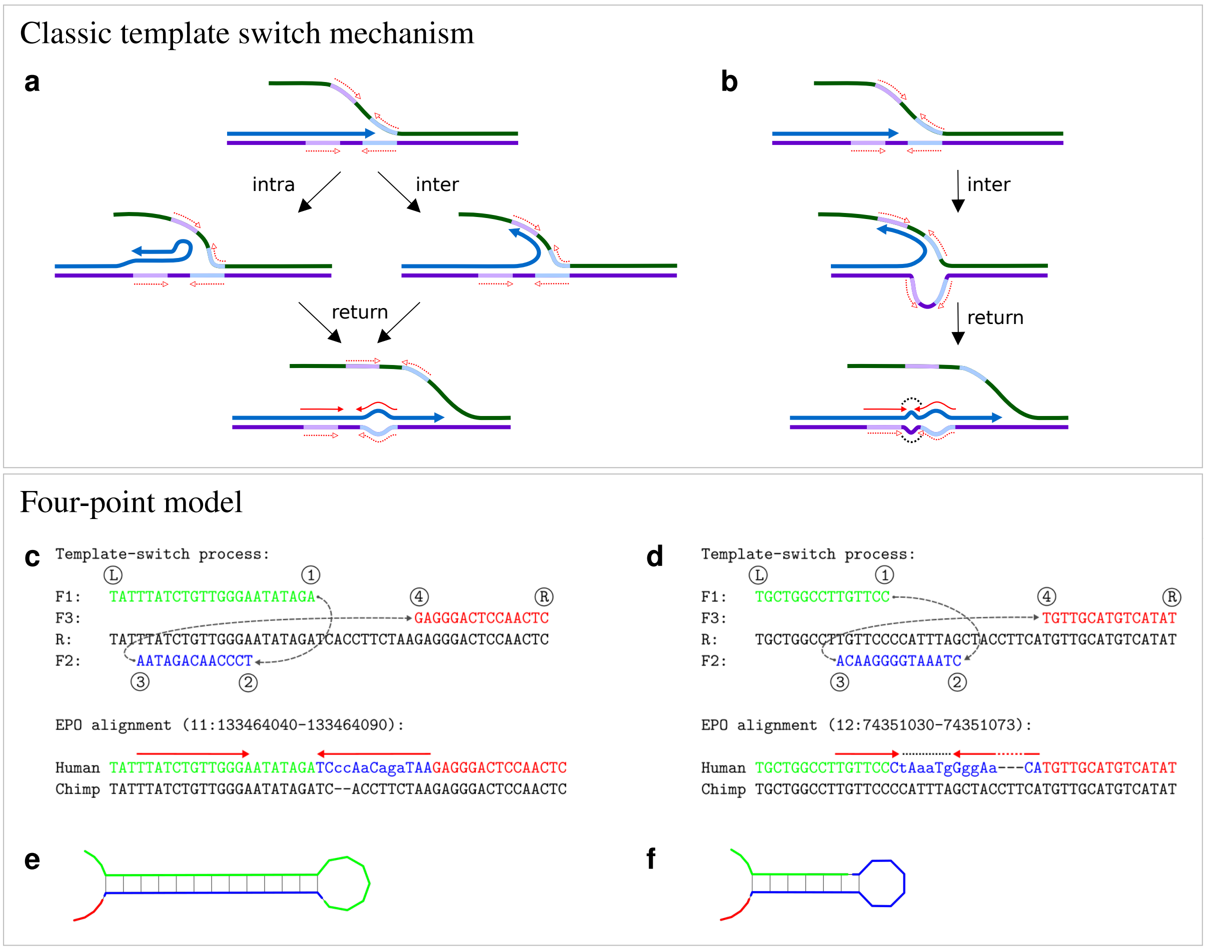

Classic template switch mechanism
a
intra
inter
return
b
inter
return
Four-point model
c
Template-switch process:
Ⓛ
①
F1: TATTTATCTGTTGGGAATATAGA
④
Ⓡ
F3: GAGGGACTCCAACTC
R: TATTTATCTGTTGGGAATATAGATCACCTTCTAAGAGGGACTCCAACTC
F2: AATAGACAACCCT
③
②
EPO alignment (11:133464040-133464090):
Human TATTTATCTGTTGGGAATATAGATCccAaCagaTAAGAGGGACTCCAACTC
Chimp TATTTATCTGTTGGGAATATAGATC--ACCTTCTAAGAGGGACTCCAACTC
d
Template-switch process:
Ⓛ
①
F1: TGCTGGCCTTGTTCC
④
Ⓡ
F3: TGTTGCATGTCATAT
R: TGCTGGCCTTGTTCCCCATTTAGCTACCTTCATGTTGCATGTCATAT
F2: ACAAGGGGTAAATC
③
②
EPO alignment (12:74351030-74351073):
Human TGCTGGCCTTGTTCCCtAaaTgGggAa---CATGTTGCATGTCATAT
Chimp TGCTGGCCTTGTTCCCCATTTAGCTACCTTCATGTTGCATGTCATAT
e
f

Fig. 1: Classic template switch mechanism and the new four-point model. **a**, **b**, The classic template switch mechanism creates perfect inverted repeats. **a**, DNA replication (blue arrow) exchanges template and converts a nearly perfect inverted repeat (dashed red arrows) into a perfect one (solid red arrows), causing a cluster of differences (bulge, bottom); this can happen by an intra-strand (left) or an inter-strand (right) switch. **b**, An inter-strand switch may invert the spacer of the repeat (black dots). **c**, **d**, The new four-point model generalizes the template switch mutation process. Template exchanges are described with four switch points (labelled ①–④) projected onto a reference sequence (R). The points define three sequence fragments (F1–F3) which, when concatenated, create a mutated output (mismatches shown in lower case in the human sequence). F1 and F3 are copied from R; F2 is copied complementary to either F1 (intra-strand switch) or R (inter-strand switch). The model can perfectly explain complex mutations observed in real data (bottom). **c**, Event "3-2-1-4", named for the order of the switch points along R, creates an inverted repeat (bottom; red arrows). **d**, Event "3-1-2-4" creates an inverted repeat (red arrows) separated by an inverted spacer (dotted line). **e**, **f**, Predicted secondary structures generated by the inverted repeats created in the Human sequences in **c**, **d**, respectively.

Links to original data:

c: http://grch37.ensembl.org/Homo%5Fsapiens/Location/Compara%5FAlignments?align=548&r=11:133333935-133333985

d: http://grch37.ensembl.org/Homo%5Fsapiens/Location/Compara%5FAlignments?align=548&r=12:74744810-74744853

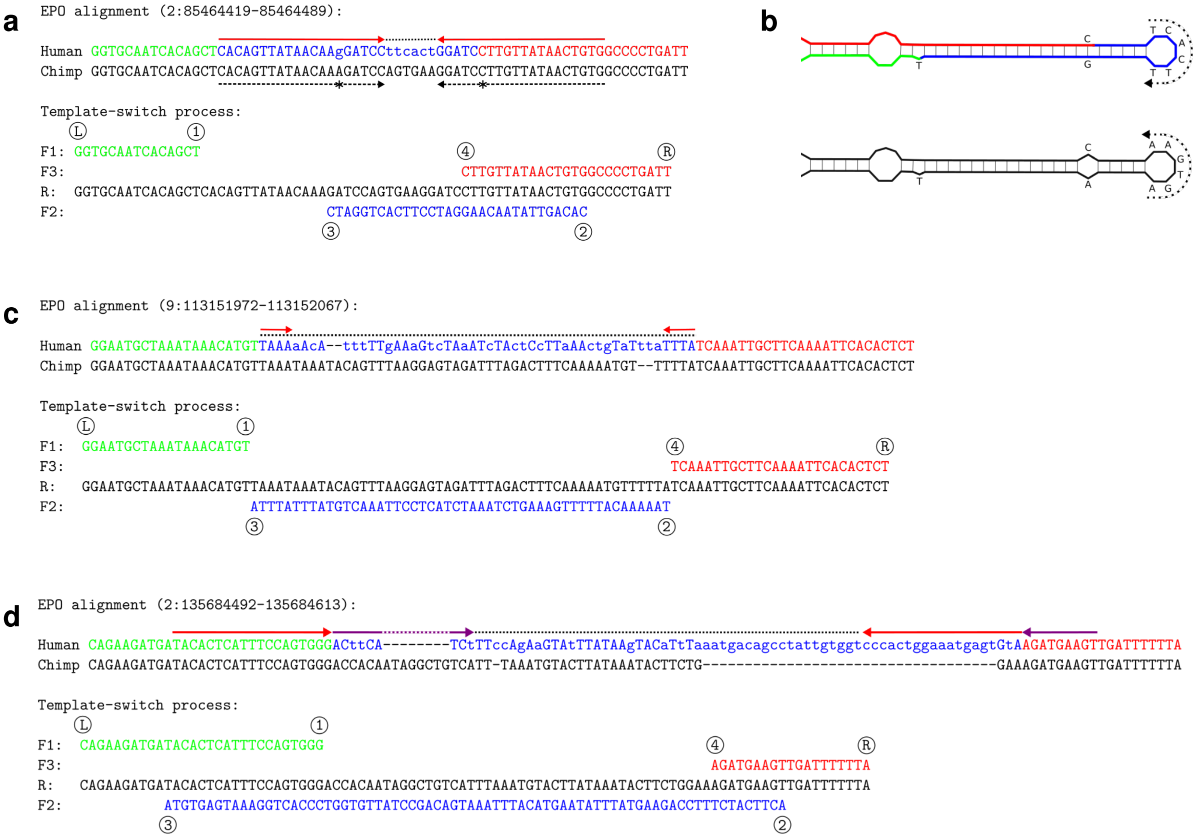
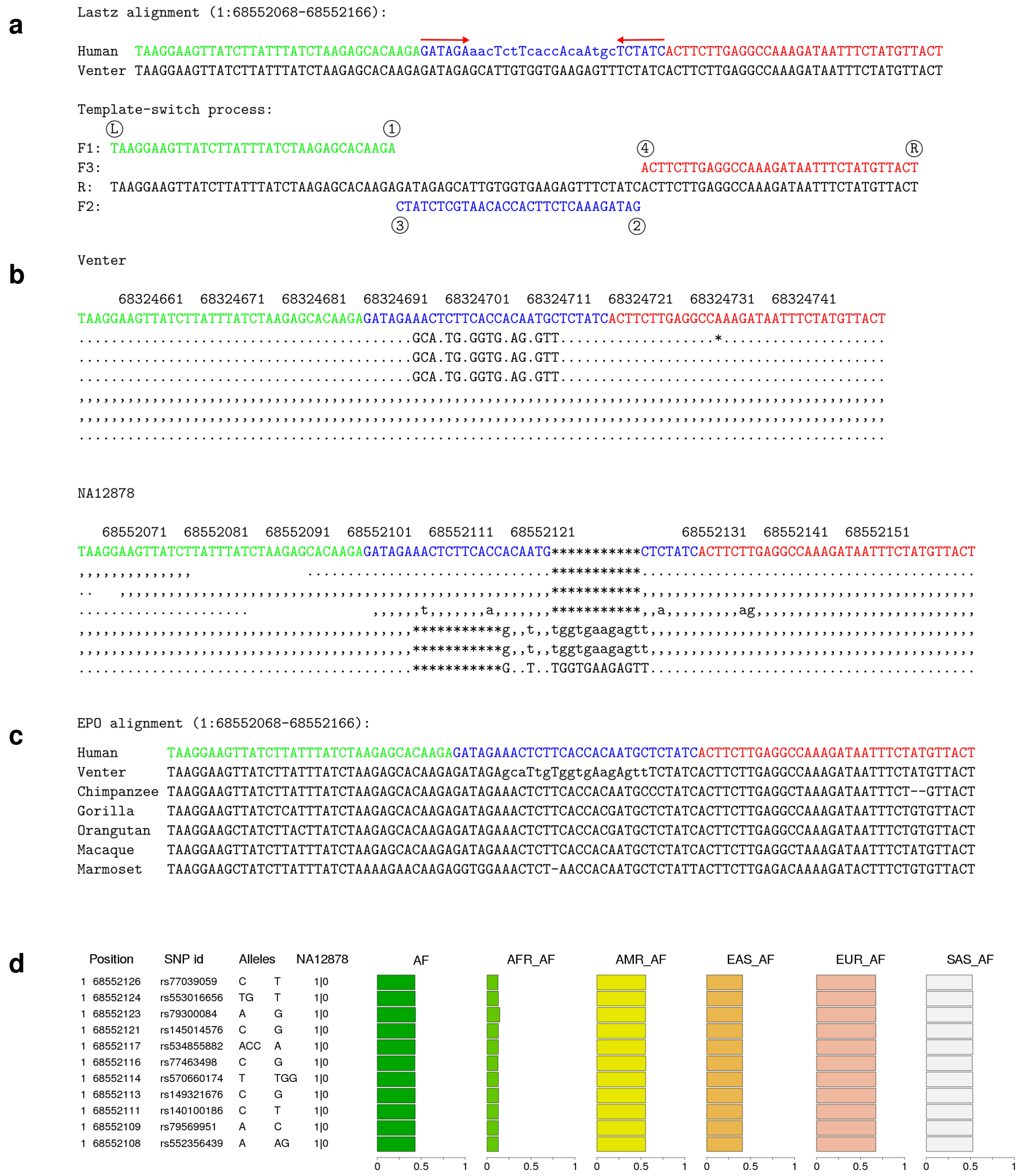

a
EPO alignment (2:85464419-85464489):
Human GGTGCAATCACAGCTCACAGTTATAACAAgGATCCttcactGGATCCTTGTTATAACTGTGGCCCCTGATT
Chimp GGTGCAATCACAGCTCACAGTTATAACAAAGATCCAGTGAAGGATCCTTGTTATAACTGTGGCCCCTGATT
Template-switch process:
Ⓛ
①
F1: GGTGCAATCACAGCT
④
Ⓡ
F3: CTTGTTATAACTGTGGCCCCTGATT
R: GGTGCAATCACAGCTCACAGTTATAACAAAGATCCAGTGAAGGATCCTTGTTATAACTGTGGCCCCTGATT
F2: CTAGGTCACTTCCTAGGAACAATATTGACAC
③
②
b
C
G
T
T
C
A
C
T
T
C
A
T
A
A
G
T
A
G
c
EPO alignment (9:113151972-113152067):
Human GGAATGCTAAATAAACATGTTAAAaAcA--tttTTgAAaGtcTAaATcTActCcTTaAActgTaTttaTTTATCAAATTGCTTCAAAATTCACACTCT
Chimp GGAATGCTAAATAAACATGTTAAATAAATACAGTTTAAGGAGTAGATTTAGACTTTCAAAAATGT--TTTTATCAAATTGCTTCAAAATTCACACTCT
Template-switch process:
Ⓛ
①
F1: GGAATGCTAAATAAACATGT
④
Ⓡ
F3: TCAAATTGCTTCAAAATTCACACTCT
R: GGAATGCTAAATAAACATGTTAAATAAATACAGTTTAAGGAGTAGATTTAGACTTTCAAAAATGTTTTTATCAAATTGCTTCAAAATTCACACTCT
F2: ATTTATTTATGTCAAATTCCTCATCTAAATCTGAAAGTTTTTACAAAAAT
③
②
d
EPO alignment (2:135684492-135684613):
Human CAGAAGATGATACACTCATTTCCAGTGGGACttCA--------TCtTTccAgAaGTAtTTATAAgTACaTtTaaatgacagcctattgtggtcccactggaaatgagtGtAAGATGAAGTTGATTTTTTA
Chimp CAGAAGATGATACACTCATTTCCAGTGGGACCACAATAGGCTGTCATT-TAAATGTACTTATAAATACTTCTG-----------------------------------GAAAGATGAAGTTGATTTTTTA
Template-switch process:
Ⓛ
①
F1: CAGAAGATGATACACTCATTTCCAGTGGG
④
Ⓡ
F3: AGATGAAGTTGATTTTTTA
R: CAGAAGATGATACACTCATTTCCAGTGGGACCACAATAGGCTGTCATTTAAATGTACTTATAAATACTTCTGGAAAGATGAAGTTGATTTTTTA
F2: ATGTGAGTAAAGGTCACCCTGGTGTTATCCGACAGTAAATTTACATGAATATTTATGAAGACCTTTCTACTTCA
③
②

Fig. 2: Example events detected in human. **a**, A near-perfect inverted repeat in chimp (dashed black arrows, the one mismatch indicated with asterisks) has been converted into a perfect inverted repeat (red arrows) in human (top). The cluster of six additional dissimilarities (dotted line) in fact represents perfect inversion of the 6-bp spacer sequence and makes the template switch (bottom) a likely explanation. **b**, Predicted DNA secondary structure before (chimp; bottom) and after (human; top) the template switch event. The dotted arrows indicate the reverse-complemented spacer region, which the four-point model explains with a single event. **c**, **d**, Additional complex mutation patterns (mismatches in lower case) that can be explained by a single template switch event. **c**, Event “1-3-2-4" only converts the spacer sequence. **d**, Event “3-1-4-2" converts the spacer sequence and creates two inverted repeats (red and magenta arrows).

Links to original data:
a: http://grch37.ensembl.org/Homo%5Fsapiens/Location/Compara%5FAlignments?align=548&r=2:85464419-85464489
c: http://grch37.ensembl.org/Homo%5Fsapiens/Location/Compara%5FAlignments?align=548&r=9:113151972-113152067
d: http://grch37.ensembl.org/Homo%5Fsapiens/Location/Compara%5FAlignments?align=548&r=2:135684492-135684613



```
Lastz alignment (1:68552068-68552166):

Human  TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAaacTctTcaccAcaAtgcTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
Venter TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAGCATTGTGGTGAAGAGTTTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT

Template-switch process:
    Ⓛ                                 ①
F1: TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGA                              ④                                Ⓡ
F3:                                                                  ACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
R:  TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAGCATTGTGGTGAAGAGTTTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
F2:                                        CTATCTCGTAACACCACTTCTCAAAGATAG
                                       ③                            ②
```

```
Venter

     68324661  68324671  68324681  68324691  68324701  68324711  68324721  68324731  68324741
TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACAATGCTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
.........................................GCA.TG.GGTG.AG.GTT...................*....................
.........................................GCA.TG.GGTG.AG.GTT........................................
.........................................GCA.TG.GGTG.AG.GTT........................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
...................................................................................................
```

```
NA12878

   68552071  68552081  68552091  68552101  68552111  68552121             68552131  68552141  68552151
TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACAATG***********CTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
,,,,,,,,,,,,,,              ..............................***********.........................................
..   ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,***********,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.....................               ,,,,,,t,,,,,,,a,,,,,,,***********,,a,,,,,,,,,ag,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,***********g,,t,,tggtgaagagtt,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,***********g,,t,,tggtgaagagtt,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................***********G..T..TGGTGAAGAGTT........................................
```

```
EPO alignment (1:68552068-68552166):

Human      TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACAATGCTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
Venter     TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAgcaTtgTggtgAagAgttTCTATCACTTCTTGAGGCCAAAGATAATTTCTATGTTACT
Chimpanzee TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACAATGCCCTATCACTTCTTGAGGCTAAAGATAATTTCT--GTTACT
Gorilla    TAAGGAAGTTATCTCATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACGATGCTCTATCACTTCTTGAGGCCAAAGATAATTTCTGTGTTACT
Orangutan  TAAGGAAGCTATCTTACTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACGATGCTCTATCACTTCTTGAGGCCAAAGATAATTTCTGTGTTACT
Macaque    TAAGGAAGTTATCTTATTTATCTAAGAGCACAAGAGATAGAAACTCTTCACCACAATGCTCTATCACTTCTTGAGGCTAAAGATAATTTCTATGTTACT
Marmoset   TAAGGAAGCTATCTTATTTATCTAAAAGAACAAGAGGTGGAAACTCT-AACCACAATGCTCTATTACTTCTTGAGACAAAAGATACTTTCTGTGTTACT
```

Position	SNP id	Alleles		NA12878
1 68552126	rs77039059	C	T	1|0
1 68552124	rs553016656	TG	T	1|0
1 68552123	rs79300084	A	G	1|0
1 68552121	rs145014576	C	G	1|0
1 68552117	rs534855882	ACC	A	1|0
1 68552116	rs77463498	C	G	1|0
1 68552114	rs570660174	T	TGG	1|0
1 68552113	rs149321676	C	G	1|0
1 68552111	rs140100186	C	T	1|0
1 68552109	rs79569951	A	C	1|0
1 68552108	rs552356439	A	AG	1|0

Fig. 3: A template switch mutation event with variable allele frequencies in human populations. **a**, Four-point model explanation of a complex mutation between the human reference GRCh37 (denoted Human) and a Caucasian male (Venter). Notation is as in Fig. 1. **b**, A subset of the original sequencing reads from the Caucasian male (top) and the 1kG individual NA12878 (bottom). Dots and commas indicate the read matching to the reference on the forward and reverse strand, upper- and lower-case characters denote the corresponding mismatches, and asterisks mark the alignment gaps. These reads reveal heterozygosity at the locus. **c**, The EPO alignment for primates reveals that the human reference (Human) is the ancestral form. As all other primates resemble the reference allele, the most parsimonious explanation is that the mutation (Venter) has happened in the human lineage since its divergence from the human-chimp ancestor. **d**, 1kG variation data explain this event as a cluster of 7 SNPs and 4 indels. The phased genotypes for NA12878 (1|0) indicate that the variant alleles are linked and all in the same haplotype. The single origin of the whole cluster is further supported by the uniform derived allele frequencies across the sites within all 1kG-data (AF) and within each superpopulation (AFR, AMR, EAS, EUR, SAS).

Table 1: Proportion of event types. Proportion of different event types among the high-confidence cases, for the comparisons of human vs. chimp and of two humans. Only one observed event type could happen via intra-strand switching (red star, its mirror case indicated with a black star). All other events can only happen inter-strand (see also Supplementary Fig. 1).

		human vs.	two
event type	output	chimp	humans
★ 1-4-3-2, ★ 3-2-1-4	inverted repeat	0.32	0.36
1-3-4-2, 3-1-2-4	inverted repeat and inverted spacer	0.22	0.15
1-3-2-4	inverted fragment	0.45	0.48
3-1-4-2	two inverted repeats and inverted spacer	0.01	0.01
events total		802	88