bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A novel process of successive inter-strand template switches explains complex mutations and creates hair-pins

Ari Löytynoja[1] and Nick Goldman[2]

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

[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

Abstract

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 has been the abundance of complex mutations comprising multiple near-by base substitutions or insertion-deletions. We devised a generalized mutation model to study the role of template switch events in the origin of mutation clusters. Applied to the human genome, the novel model finds strong evidence for inter-strand template switch events and identifies new types of mutations that create short inversions, some flanked by a pair of inverted repeats. This new, local template switch process can create complex mutation patterns, including secondary structures, and explains apparent high frequencies of compensatory substitutions and multi-nucleotide mutations without invoking positive selection. Many of these complex mutations are polymorphic in humans but their detection with current sequencing methodologies is difficult.

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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)[1;2;3;4]. Explanations for the mutation clusters vary from an error-prone polymerase[3] to indels being mutagenic[5]. In bacteria, mutations creating perfect inverted repeats occur with high frequency[6] and the mechanism behind this is thought to involve intra- or inter-strand template switching during DNA replication[7] (Fig. 1a, b). 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[6], earlier studies have not considered the mechanism significant in the evolution of higher organisms[8]. These conclusions were based on limited data, however, and on an assumption that the mechanism creates perfect inverted repeats only. 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)[9], replication slippage[10] and microhomology-mediated break-induced replication (MMBIR)[11], but mutations attriuted to these typically involve major genomic rearrangements[12;13]. Even with the underlying biological mechanism uncertain, we realized that the existence and properties of a mutation process 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). When 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).

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This modeling of the template switch process 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. More specifically, 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 inverted and direct repeats flanked by dissimilar sequence and one case causing inversion of a sequence fragment only.

To test whether biological data support the proposed mechanism, we implemented a computational tool 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. 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[14;15] (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.

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Discovery of four-point mutation events. We found 4,901 candidate events, spread across all human chromosomes. 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 (Fig. 2c–d). While any complex mutation can be generated with a combination of 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. 2, 10) 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).

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. 1; 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"; 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 ob-

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

served in real sequence data is the order of switch points ① and ④: in all unobserved cases, 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 (cf. Supplementary Fig. 1). The discovery that this is not happening, as well as the numerous cases of inversion of spacer sequences, 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[7], the "intra-strand" mechanism has been the dominant hypothesis[6]. 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. 2) 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. 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[16]), 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. Although the total number of events, 88, was lower than between human and chimp (as expected, given the much smaller time since the last common ancestor), the proportions of different event types were similar and again only the six types not requiring opening of the new helix were found (Table 1; two humans comparison).

Focusing on the 88 candidate events, we manually studied the Caucasian male sequence data mapped onto the reference genome[17]. 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

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of them heterozygous, i.e. the sequence data containing reads consistent with both Venter and GRCh37 alleles (Fig. 3a, b; Supplementary Table 1; Supplementary Data 1); 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. 4). We then looked at the same loci in the 1000 Genomes (1kG) data[18] 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.

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 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). 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[18] 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). The

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

variation data confirm the two earlier cases as combinations of independent mutations (Supplementary Fig. 4) 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. 5). 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. 6)—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 the populations (Supplementary Fig. 7). This demonstrates that, 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.

Discussion

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 (Supplementary Fig. 1), the model significantly extends the one previously proposed for bacteria. First, unlike the bacterial model, pre-existing sequence similarity is not required and

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the process can thus create completely novel repeats (cf. 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[9;11;13]. 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 multinucleotide substitution (MNS). Some cases of MNSs have been explained with positive selection[19;20] while involvement of Pol ζ has been suggested to explain spatial differences in mutation frequency[3]. 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.*[21] as local template switch events. It has been shown that switch events can take place between distant loci[9;11;13] 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. 8) and the process we identified provides an alternative to the suggestion of indels being mutagenic and triggering near-by base substitutions[5].

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[22;23;24]. 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[25;26;20]. 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. The typical signature of the process, tight clusters of differences, make read mapping and subsequent variant calling challenging. This is demonstrated by phase 3 of the 1000 Genomes Project[18], which provides significant improvements in comparison to earlier releases but still contains errors and inconsistencies around the regions

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

that 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[21,22] 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. 9), our approach maximized the length of ②→③ fragment and reported this upper limit of the strand-switch event length. To generate controls (see below), we reconstructed the human query from the chimp reference with imperfect copying of the forward strand only. The computational tool used for the analyses is described 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[27] or dustmasked[28] 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:

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(*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[29] and following the UCSC analysis pipeline[30]. The four-point mutations were identified using the same approach as with human-chimp data. The 1000 Genomes variation data from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ were analysed using bcftools[31] and selected regions of resequencing data from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878 were visualized using samtools[32]. 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[33].

Other computational analyses. DNA secondary structures were predicted with the ViennaRNA package[34], using the command 'RNAfold –noconv –noGU -P dna_mathews2004.par'. The length distribution (Supplementary Fig. 2) and the allele frequencies (e.g. Fig. 3d) were visualized with R[35].

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Averof, M., Rokas, A., Wolfe, K. H. & Sharp, P. M. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**, 1283–1286 (2000).

2. Whelan, S. & Goldman, N. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**, 2027–2043 (2004).

3. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **24**, 1445–1454 (2014).

4. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

5. Tian, D. *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105–108 (2008).

6. Dutra, B. E. & Lovett, S. T. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. *J. Mol. Biol.* **356**, 300–311 (2006).

7. Ripley, L. S. Frameshift mutation: determinants of specificity. *Annu. Rev. Genet.* **24**, 189–213 (1990).

8. Ladoukakis, E. D. & Eyre-Walker, A. The excess of small inverted repeats in prokaryotes. *J. Mol. Evol.* **67**, 291–300 (2008).

9. Lee, J. A., Carvalho, C. M. B. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).

10. Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C. & Cooper, D. N. Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. *Hum. Mutat.* **26**, 362–373 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

11. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).

12. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).

13. Costantino, L. *et al.* Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88–91 (2013).

14. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).

15. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).

16. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).

17. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

18. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

19. Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S. & Kondrashov, A. S. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature* **429**, 558–562 (2004).

20. Meer, M. V., Kondrashov, A. S., Artzy-Randrup, Y. & Kondrashov, F. A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464**, 279–282 (2010).

21. Schrider, D. R., Hourmozdi, J. N. & Hahn, M. W. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* **21**, 1051–1054 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

22. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).

23. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).

24. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).

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

26. Dixon, M. T. & Hillis, D. M. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.* **10**, 256–267 (1993).

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

28. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).

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

30. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

31. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

33. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

34. Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

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

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Author Contributions: N.G. devised the extended model. A.L. implemented the method, performed the analyses and wrote the first draft of the paper. Both authors were involved in study design, discussed the results and contributed to the final manuscript.

Competing financial interests: The authors declare no competing financial interests.

Corresponding author: Correspondence to Ari Löytynoja

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

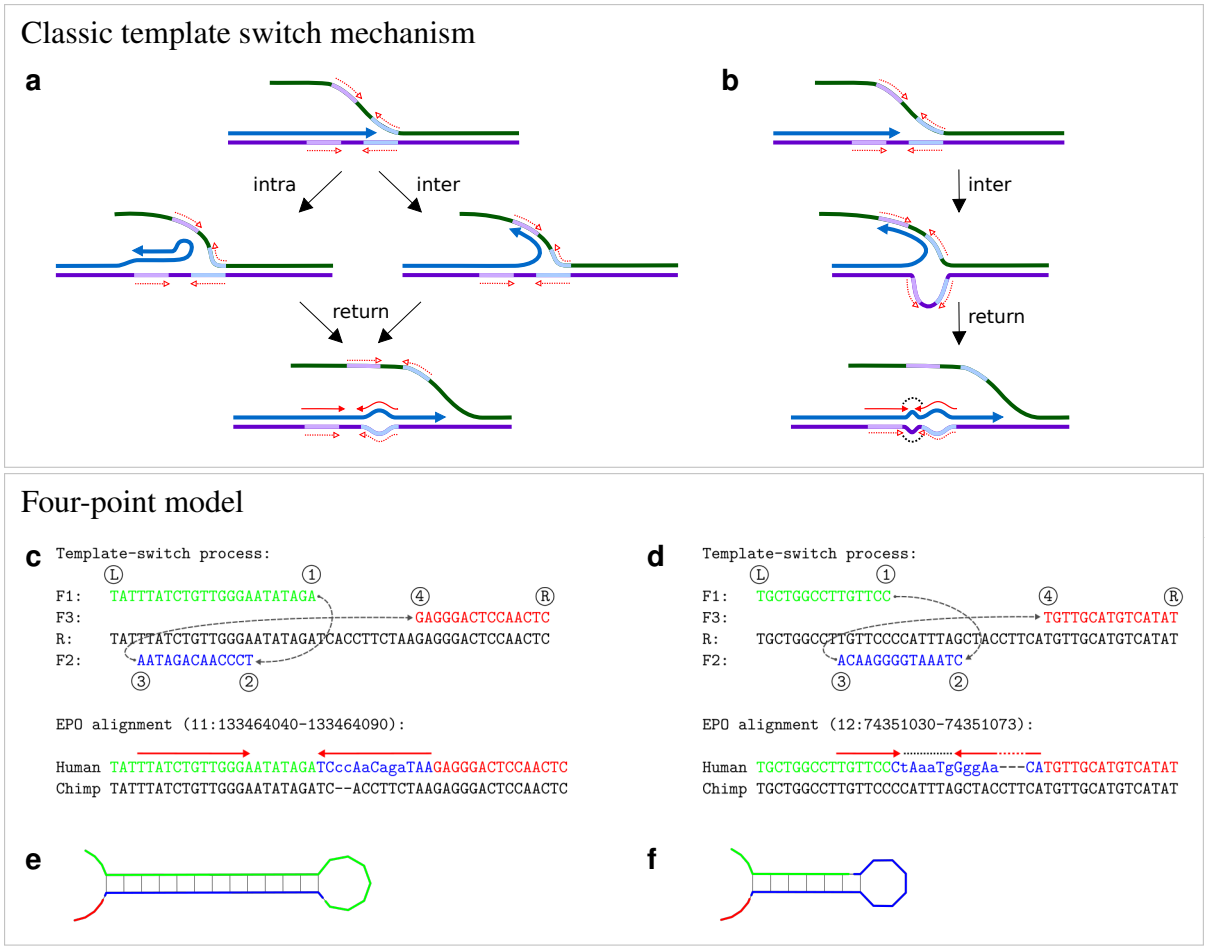

Fig. 1: Classic template switch mechanism and the new four-point model. **a**, **b**, The classic template switch mechanism creates perfect inverted repeats. **a**, DNA replication (blue arrow) exchanges template and converts a nearly perfect inverted repeat (dashed red arrows) into a perfect one (solid 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; **a**, left) or R (inter-strand switch; **a**, right, or **b**). The model perfectly explains 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 for 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

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

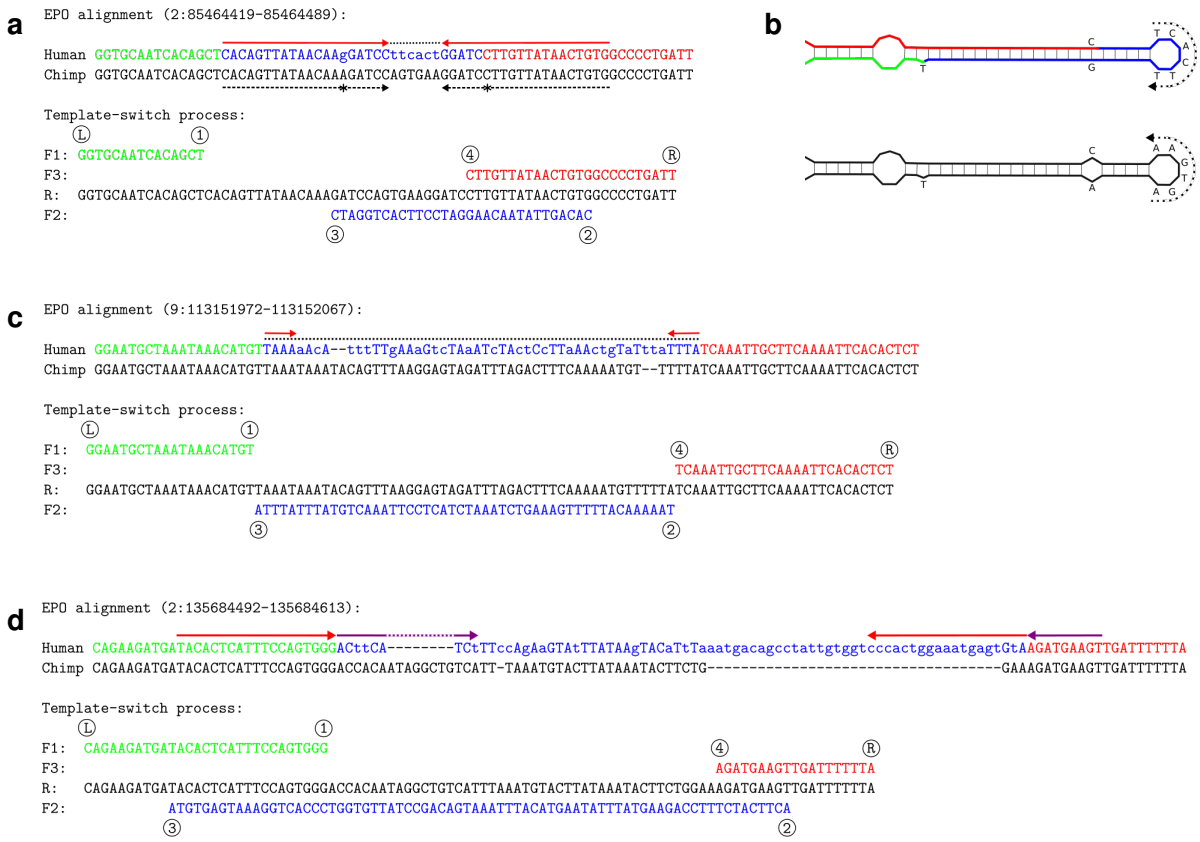

Fig. 2: Example events detected in human. **a**, A near-perfect inverted repeat in chimp (dashed black arrows, the one mismatch indicated with asterisks) has been converted into a perfect inverted repeat (red arrows) in human (top). The cluster of six additional dissimilarities (dotted line) in fact 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**, 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

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

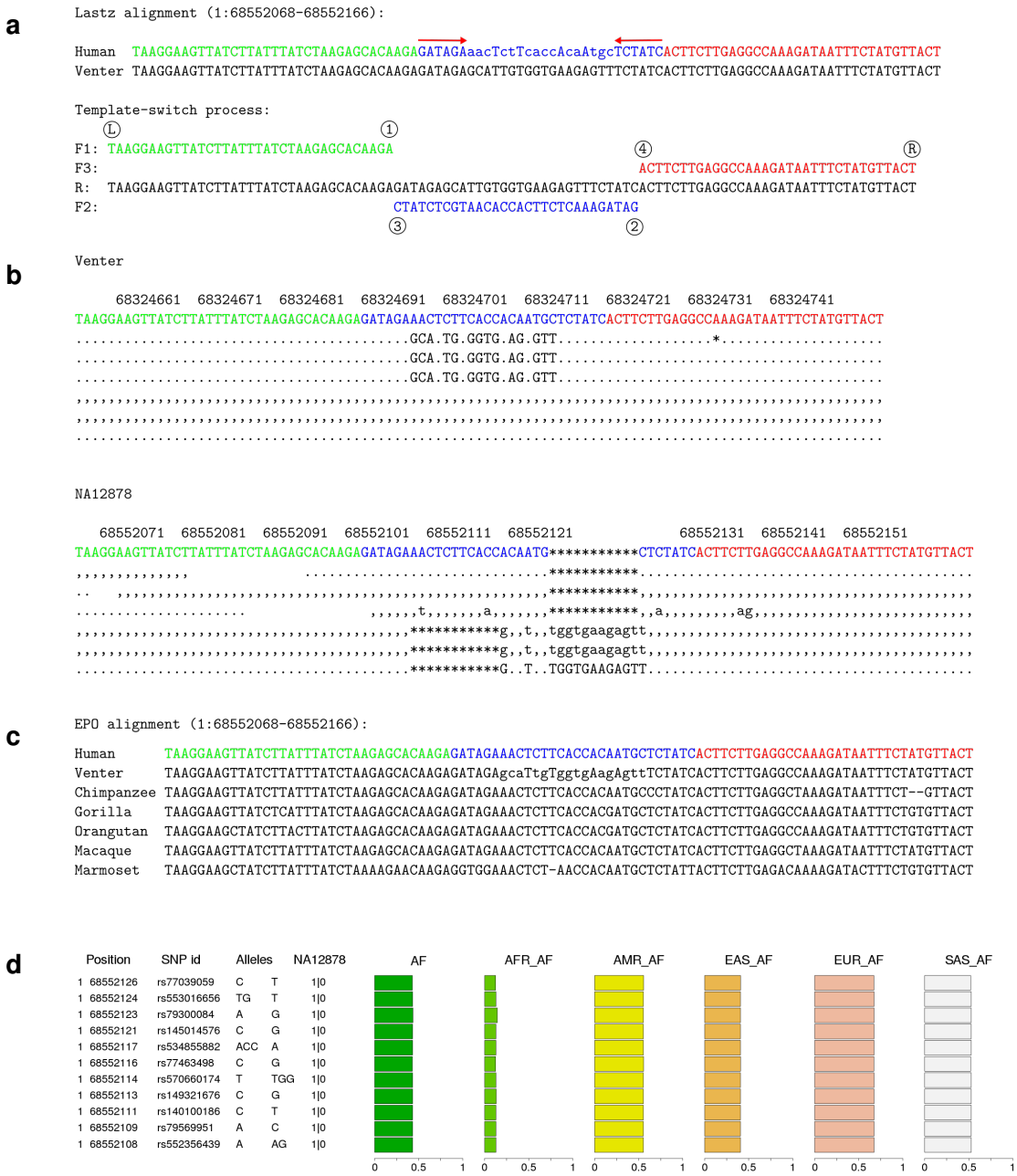

Fig. 3: A template switch mutation event with variable allele frequencies in human populations. **a**, Four-point model explanation of a complex mutation between the human reference GRCh37 (denoted Human) and a Caucasian male (Venter). See Fig. 1 for a description of the notation used. **b**, A subset 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**, Type 1-3-2-4 event can be reversible and 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).

bioRxiv preprint doi: https://doi.org/10.1101/038380; this version posted February 1, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

event type	output	human vs. chimp	two 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 & inv. spacer	0.22	0.15
1-3-2-4	inverted fragment	0.45	0.48
3-1-4-2	two inv. repeats & inv. spacer	0.01	0.01
events total		802	88