scientific reports

OPEN

# Characteristics and possible mechanisms of formation of microinversions distinguishing human and chimpanzee genomes

Nadezhda A. Potapova[1✉], Alexey S. Kondrashov[2] & Sergei M. Mirkin[3✉]

**Genomic inversions come in various sizes. While long inversions are relatively easy to identify by aligning high-quality genome sequences, unambiguous identification of microinversions is more problematic. Here, using a set of extra stringent criteria to distinguish microinversions from other mutational events, we describe microinversions that occurred after the divergence of humans and chimpanzees. In total, we found 59 definite microinversions that range from 17 to 33 nucleotides in length. In majority of them, human genome sequences matched exactly the reverse-complemented chimpanzee genome sequences, implying that the inverted DNA segment was copied precisely. All these microinversions were flanked by perfect or nearly perfect inverted repeats pointing to their key role in their formation. Template switching at inverted repeats during DNA replication was previously discussed as a possible mechanism for the microinversion formation. However, many of definite microinversions found by us cannot be easily explained via template switching owing to the combination of the short length and imperfect nature of their flanking inverted repeats. We propose a novel, alternative mechanism that involves repair of a double-stranded break within the inverting segment via microhomology-mediated break-induced replication, which can consistently explain all definite microinversion events.**

A pure inversion is a mutation that replaces a segment of the genome with its reverse complement sequence. Inversions are routinely found in studies of genetic variation both within species[1] and between species[2]. Because the rate of inversions is much lower than that of single nucleotide substitutions and of short deletions and insertions[3,4], they comprise only a small proportion of all genetic changes. Thus, to detect large number of inversions, it may be necessary to compare genomes of species that are separated from each other by a relatively long evolutionary path.

Inversions can vary in length dramatically[5–7], which leads to different potential pitfalls that complicate ascertaining long and short inversions. Detection of long inversions can be hindered by incorrect genome assembly because most non-human genomes were sequenced via short-read platforms. Only comparison of high-quality genomes and usage of stringent parameters for genome assembly and alignment can produce reliable results. For example, 252 long inversions caused by retrotransposons were detected with high confidence on the human-chimpanzee path[8].

Detection of short inversions (with lengths below 100 nucleotides), sometimes referred to as "ultra microinversions"[5], "submicroscopic inversions"[7] or "pico inversions"[6], requires distinguishing them from products of other mechanisms that can also produce segments of low similarity ("bubbles") in genome alignments, such as clustered small-scale mutations, mutational hotspots or simply a large number of independent small-scale mutations that occurred by chance within a short segment of the genome[2,9,10]. Moreover, multiple nucleotide substitutions, if they accumulated within a microinversion after it has occurred, can render it unrecognizable. Thus, identifying microinversions that occurred on a long evolutionary path may be practically impossible[2,11–13].

Conversely, the more similar the genomes of two compared species are, the shorter the microinversions are that can be reliably detected. For example, human and chimpanzee genomes, which are > 98% identical if long deletions and insertions are ignored, and ~ 95% identical, if all genetic differences are considered[14], provide an

[1]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia 127051. [2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA. [3]Department of Biology, Tufts University, Medford, MA 02155, USA. ✉email: nadezhdalpotapova@gmail.com; sergei.mirkin@tufts.edu

excellent system for ascertaining microinversions. Several comparative analyses were performed for these species, some of which reported thousands of microinversions[5,9,13,15]. Note, however, that the outcomes of these analyses differ significantly, likely due to different methods of filtering used in those studies.

Resolving the differences in the detection and analysis of microinversion is important for several reasons. First, it was hypothesized that microinversions may play a role in human genetic diseases and are overrepresented in cancer genomes[13,16]. Second, being such rare events, microinversions are very useful for phylogenetic inferences[2] because the probability of homoplasy, due to repeated origins of the same microinversion, is extremely low. Third, they may help elucidate genetic relationships between human populations as sensitive markers for separating populations[17].

Here we describe definite microinversion that occurred after the divergence of humans and chimpanzees, which were identified using a set of extra stringent criteria. Majority of them were pure inversions, i.e. human genome sequences matched exactly to the reverse-complemented chimpanzee sequences. All these microinversions were flanked by perfect or nearly perfect inverted repeats, strongly suggesting that they play a key role in the origin of microinversions. While template switching at inverted repeats was previously discussed as a possible mechanism for the microinversion formation[18–20], many of definite microinversions detected by us cannot be easily explained by this mechanism. We, thus, propose a novel, alternative mechanism stipulating that microinversions emerge in course of double-strand break repair.

## Results

### Identifying microinversions between human-chimpanzee and chimpanzee-bonobo genome pairs.

Ascertainment of microinversions by comparison of similar genomes appears to be a straightforward task because a microinversion produces a segment of the alignment where the sequence of one species is (nearly) identical to reverse-complemented sequence of the other species. Still, one needs to distinguish microinversions from other phenomena that can produce similar outcomes[9].

As long as a tool that is used for comparing genomes is not specifically searching for microinversions, they present in the alignment as “bubbles”, or segments of low similarity. Long alignments routinely contain multiple bubbles, which are particularly conspicuous when the two compared genomes are generally very similar to each other. However, most of these bubbles are not products of microinversions. Thus, the key problem is to specifically recognize bubbles which resulted from microinversions.

How can a bubble emerge? A microinversion, an individual event of some other kind that involves many nucleotide sites, and a clump of many separate single-nucleotide substitutions[11], and/or other small-scale mutations are the three options. Difficulty in discriminating between these options is the main reason why previous studies reported vastly different numbers of microinversions on the evolutionary path connecting humans and chimpanzees[5,6,21].

How to distinguish between these three options? We believe that it is impossible to reliably recognize a microinversion that is too short. Indeed, a one nucleotide long microinversion cannot be distinguished from the corresponding transversion. It is a moot question, therefore, whether such microinversions exist. In general, a microinversion of a genome segment of length N looks like a succession of N single-nucleotide substitutions. Obviously, when N is small, even a bubble consisting of two genome segments of equal lengths such that the sequence of one segment coincides with the reverse-complemented sequences of the other segment (say, AG–CT, with N = 2) does not necessarily imply that a microinversion occurred. Several independent substitutions, or a complex mutation that produced such a pattern accidentally, cannot be ruled out. By contrast, a long enough microinversion (say, of N = 27, resulting in a bubble GCAGATCATCCTTTATTCCTACCTTGT—ACAAGG TAGGAATAAAGGATGATCTGC, see Table S1) is clearly recognizable because the probability of an accidental origin of such a bubble by any other means is vanishingly low.

What is the minimal length of a microinversion that can be reliably recognized as such by comparing the human and chimpanzee genomes? Because these two genomes differ from each other at ~ 1 nucleotide site out of 100, the probability that a bubble of length N appears as a result of independent single-nucleotide substitutions is a succession of N mismatches, is $(1/100)^N$. Thus, a priori, in the alignment of length of ~ 1 billion nucleotides, this mechanism can generate bubbles of lengths only up to 5 nucleotides, and one might optimistically conclude that a microinversion of N > 5 can be reliably recognized.

However, this does not seem to be the case. Human-chimpanzee genome alignment contains 85,410 bubbles with lengths above 5 nucleotides, including 28,113 with lengths above 10 (Table S2). Clearly, a vast majority of these bubbles are not microinversions, since the sequence of one genome segment bears no similarity to the reverse-complemented sequence of the other segment. This should not be surprising because heterogeneity in the rate of evolution along the genome leads to non-random clumping of mismatches and, thus, causes overrepresentation of long bubbles[22]. Also, large-scale mutations different from microinversions likely played a significant role in generating bubbles of N > 5. Unfortunately, we do not know a priori how many bubbles are expected to appear in these ways.

While the comparison of the two segments that constitute a bubble can reveal its origin, false-positive microinversions remain a concern. This is because a segment of one genome can be identical or very similar to the reverse-complemented segment of the other genome by sheer accident. If we are dealing with K bubbles, and the probability that a bubble originated by other means but looks like a microinversion is P, we can be confident that every bubble that looks like a microinversion indeed evolved in this way only if $P << 1/K$, because the expected number of false positives is $<< 1$ in this case.

Under the simplest assumptions, in a bubble of length N, the segment of one genome coincides accidentally with a reverse-complemented segment of the other genome with probability $(1/3)^N$, if the bubble contains no matches, or with probability $(1/4)^N$, if random matches between the segments are also allowed. Thus, the

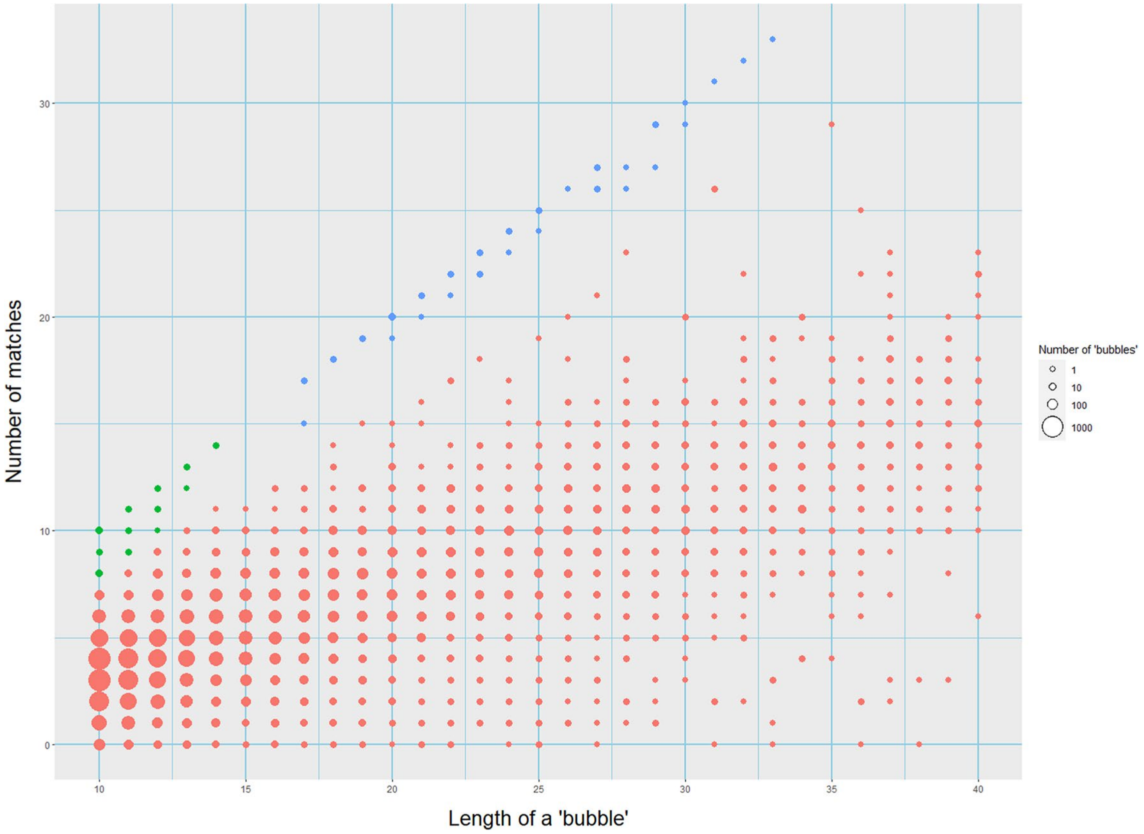


**Figure 1.** Microinversions in the human-chimpanzee genome alignment. The x-axis corresponds to the length of a "bubble" from 10 to 40 nucleotides; the y-axis corresponds to the number of matches between the reverse complement of chimpanzee DNA sequence and the corresponding human DNA sequence. Blue circles correspond to microinversions that we consider to be definite, green circles—likely microinversions. Sizes of circles corresponds to the number of "bubbles".

probability for a bubble of length 10 to look like a product of a microinversion by accident is ~ $10^{-5}$ for bubbles with no matches or ~ $10^{-7}$ for bubbles with random matches. Because the total number of bubbles of length 10 and above in the human-chimpanzee genome alignment is only ~ $10^{4}$, i.e. exceeds those that could have originated by random chance by at least and order of magnitude (Table S2), we assume that practically all bubbles that look like microinversions, are likely definite microinversions. In other words, it seems likely that the minimal length of a bubble in the human-chimpanzee genome alignment that can be confidently interpreted as a microinversion if one of the genome segments (nearly) coincides with the reverse-complemented sequence of the other segment is close to 10.

Our data are in agreement with this simple probabilistic reasoning. Table S2 shows bubbles of 5-to 40-nucleotides with different numbers of reverse-complemented matches, while Fig. 1 is a plot representation of the bubbles ranging from 10 to 40 nucleotides. First, vast majority of them (designated by red circles in Fig. 1 or unhighlighted in Table S2) do not possess any excess of such matches and, instead, approximately conform to the binomial distribution with the average number of such matches equal to ~ 1/3 of the length of the bubble. Second, there is a small fraction of bubbles made entirely or almost entirely of reverse-complemented matches (blue and green circles in Fig. 1, and the highlighted diagonal in Table S2). The latter bubbles were likely produced by microinversions while the bubbles of the first kind—by other mechanisms.

As expected, the kind of a bubble becomes clearly recognizable when N is large enough. The two kinds of bubbles form separate modes starting from N ~ 10, and starting from N ~ 15, the chances that a bubble accidentally looks like a product of a microinversion become very low. In contrast, it seems to be impossible to claim with confidence that a bubble with N < 10 was produced by a microinversion, although some of them likely were. To be conservative, we concentrate on microinversions of lengths 15 and above (blue circles in Fig. 1), while the range of N from 10 to 14 (green circles in Fig. 1) is considered by us as a gray area.

We conclude that among a very large number of potential microinversions listed in the Table S2, only 59 events of lengths from 17 to 33 seem to be definite microinversions. In the majority (45) of them, there was a complete match between a human sequence and the corresponding reverse complement chimpanzee's sequence, while in the remaining 14 cases, there were only one-to-two small differences (mismatches or gaps).

Orangutan genome was used as an outgroup to determine in which lineage, human or chimpanzee, a particular microinversion occurred. Among them, 21 occurred in the human lineages and 28 in the chimpanzee lineage, and 10 could not be assigned due to lack of the outgroup sequences (Table S1).

We did not observe any biases in terms of genome structure or functioning. 31 out of 59 microinversions were located within introns, which is not very different from random expectation because introns occupied ~ 43% of analyzed sequences. Only one gene, *SC5D*, carried a microinversion located in an exon, but it does not overlap with the coding segment and thus does not affect the protein.

The same approach was used to discover microinversions from the chimpanzee—bonobo genome alignment (Table S3). Using the same stringent criteria as for human-chimpanzee analysis, we found 9 microinversions ranging in length from 17 to 25 nucleotides.

### Definite microinversions are flanked by perfect or nearly perfect inverted repeats.

Notably, each definite microinversion that we found is flanked by inverted repeat sequences (Fig. 2, Table S1). The length of these inverted repeats varies from 3 to 75 nucleotides per flank, 34 of which contain mismatches or small gaps.

With only two exceptions, the inner edge of an inverted repeat is located within – 5 to + 4 nucleotides relative to the outer boundaries of the microinversion. The distances between the outer boundaries of microinversions and the inner edges of inverted repeats are correlated (Pearson's correlation coefficient, calculated in Fig. 3, is 0.64).

These inverted repeats can be subdivided into three loose classes, based on their length. First, there are relatively long repeats, in which the inverted repeat segment is at least 8 nucleotides or more: 23 of 59 microinversions are surrounded by such repeats. Obviously, these repeats are extremely unlikely to occur by chance, even if we take into account some flexibility in their locations relative to a microinversion.

Secondly, there are 28 repeats of the intermediate length, in which an inverted repeat sequence ranges from 4 to 7 nucleotides. They are also highly unlikely to occur by chance given their proximity to the outer edges of the microinversion.

Finally, there are 8 microinversions that are flanked by only 3 nucleotide-long inverted repeat sequences. Notably, however, 6 of them are directly adjacent to the microinversion edges, the random chance of which is 1/64. In the remaining cases, inverted repeats overlap with the microinversion edges on 2 or 4 nucleotides, respectively. Thus, we conclude that even the shortest inverted repeats flanking microinversions are not spurious as well.

Similarly, all 9 microinversions revealed by comparison between chimpanzee and bonobo genomes are also flanked by inverted repeats ranging from 3 to 10 nucleotides (Table S3).

## Discussion

We found 59 microinversions of lengths from 17 to 33 on the evolutionary path connecting humans and chimpanzees (Fig. 1). These data suggest that during human-chimpanzee divergence definite microinversions appeared at the rate of ~ $2 \times 10^{-13}$ per nucleotide per generation (taking into account the genome size of ~ $3 \times 10^{9}$ nucleotides and assuming that the two lineages together went through ~ $10^{6}$ generations[23]). Thus, microinversions occur five orders of magnitude less frequently than single-nucleotide substitutions, the rate of which is $10^{-8}$ per nucleotide per generation rate. We realize that the actual rate of microinversion occurrence must be somewhat higher than we deduced, given that we ignored shorter events and could not analyze hard-to-sequence genomic regions, but it is still orders of magnitude less than the rate of point substitutions, which is not surprising, as a microinversion is obviously a much more complex and improbable event.

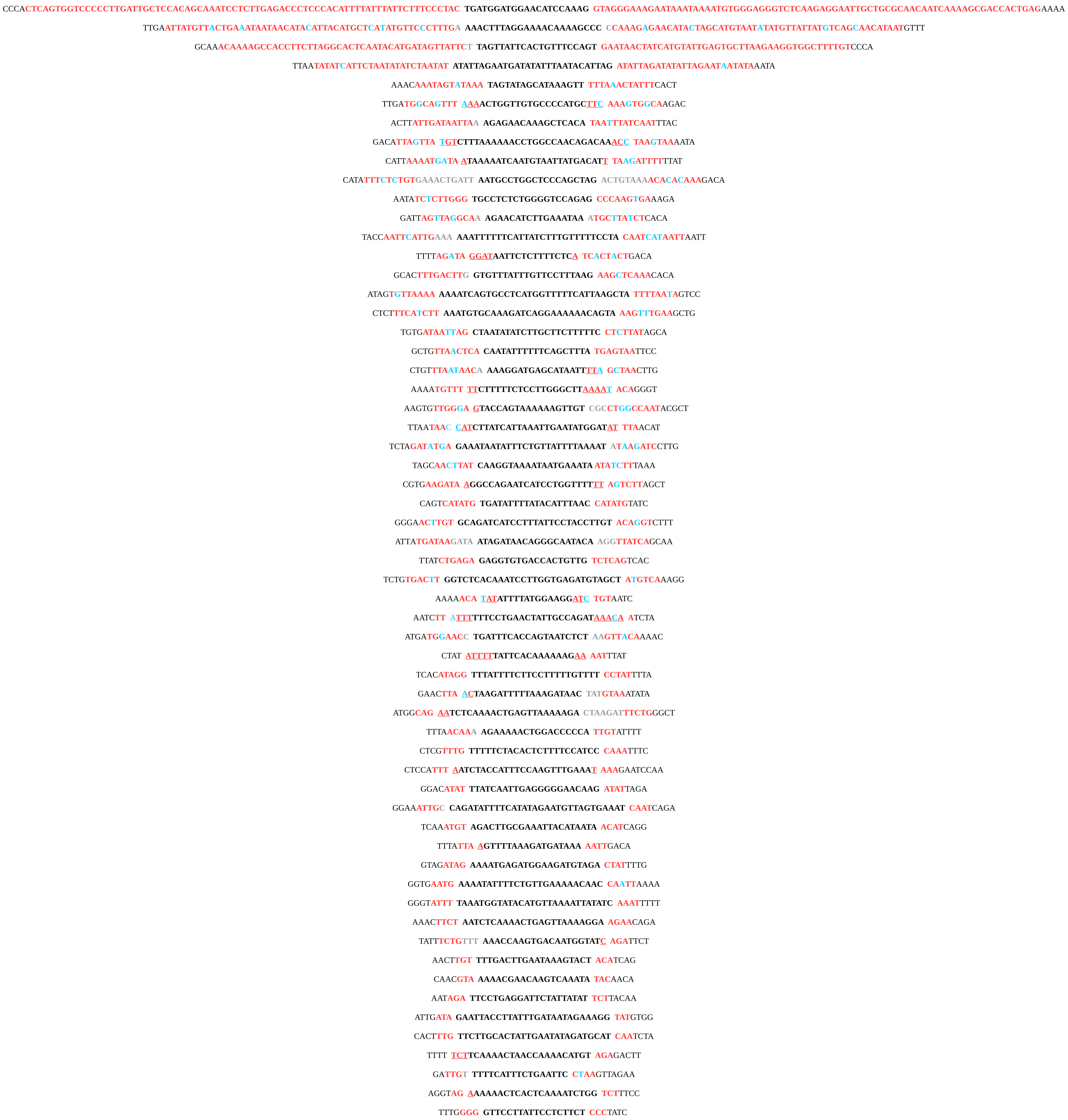
Our data are consistent with the constant rate of the origin of microinversions. The human and the chimpanzee lineages accumulated similar numbers of them, which is to be expected because the overall rates of genome evolution of these two lineages are similar[24]. Observing 9/59 times less microinversions on the path connecting chimpanzees and bonobos agrees well with this path being ~ 4 times shorter than that between humans and chimpanzees[25].

In 45 of the 59 microinversion, the human genome segment precisely coincided with the reverse-complemented chimpanzee genome segment, while in the remaining 14 there is one or two differences, either a mismatch or a gap. At the moment of its occurrence, a microinversion can be pure, if the replacement of the genome segment with the reverse-complemented sequence is not accompanied by any other mutations. Alternatively, a microinversion can be impure to start with, if its occurrence was accompanied by other, small-scale mutations, as it is usually the case for long inversions[8,15,26]. Our data indicate that, in contrast to long inversions, definite microinversions are usually born pure.

Indeed, the divergence between human and chimpanzee genomes is ~ 1%. Because the characteristic length of our microinversions is ~ 20 nucleotides, we expect, if small-scale differences within microinversions emerge independently, that 1/5 of microinversions carry 1 mismatch, 1/25 carry 2 mismatches, etc. Our data conform to these expectations. Moreover, among the 3 small-scale differences within the segments of the human-chimpanzee alignment that correspond to a microinversion that can be attributed to a particular lineage, 6 occurred not in the lineage where the microinversion happened, but in the other one (Table S1).

Note that the number of microinversions that we found is much smaller than that in a recent study of Walker et al.[27]. This is because our approaches are radically different. Walker et al. strived to find all plausible microinversions by employing hidden Markov models for this purpose. Their set contains over 4000 microinversions, some of which are as short as 6 nucleotides. Obviously, even if such a short segment of the human-chimpanzee alignment contains sequences that are exactly reverse-complementary to each other, one cannot be certain that it, indeed, is the result of a microinversion, although it is the most likely explanation. In contrast, we tried to avoid any possibility of false positive events and identified only definite microinversions relying on simple probabilistic arguments described in "Results" section. Therefore, we only analyzed microinversion longer than 15 nucleotides, for which the probability of false discovery is vanishingly small (Fig. 1).

All definite microinversions that we detected appeared to be flanked by perfect or nearly perfect inverted repeats. Roughly 39% of these inverted repeats are relatively long (> 8 nucleotides per flank), 50% are of an



```
CCCACTCAGTGGTCCCCCTTGATTGCTCCACAGCAAATCCTCTTGAGACCCTCCCACATTTTATTTATTCTTTCCCTAC TGATGGATGGAACATCCAAAG GTAGGGAAAGAATAAATAAAATGTGGGAGGGTCTCAAGAGGAATTGCTGCGCAACAATCAAAAGCGACCACTGAGAAAA
TTGAATTATGTTACTGAAATAATAACATACATTACATGCTCATATGTTCCCTTTGA AAACTTTAGGAAAACAAAAGCCC CCAAAGAGAACATACTAGCATGTAATATATGTTATTATGTCAGCAACATAATGTTT
GCAAACAAAAGCCACCTTCTTAGGCACTCAATACATGATAGTTATTCT TAGTTATTCACTGTTTCCAGT GAATAACTATCATGTATTGAGTGCTTAAGAAGGTGGCTTTTGTCCCA
TTAATATATCATTCTAATATATCTAATAT ATATTAGAATGATATATTTAATACATTAG ATATTAGATATATTAGAATAATATAAATA
AAACAAATAGTATAAA TAGTATAGCATAAAGTT TTTAAACTATTTCACT
TTGATGGCAGTTT AAAACTGGTTGTGCCCCATGCTTC AAAGTGGCAAGAC
ACTTATTGATAATTAA AGAGAACAAAGCTCACA TAATTTATCAATTTAC
GACATTAGTTA TGTCTTTAAAAAACCTGGCCAACAGACAAACC TAAGTAAAATA
CATTAAAATGATA ATAAAAATCAATGTAATTATGACATT TAAGATTTTTTAT
CATATTTCTCTGTGAAACTGATT AATGCCTGGCTCCCAGCTAG ACTGTAAAACACACAAAGACA
AATATCTCTTGGG TGCCTCTCTGGGGTCCAGAG CCCAAGTGAAAGA
GATTAGTTAGGCAA AGAACATCTTGAAATAA ATGCTTATCTCACA
TACCAATTCATTGAAA AAATTTTTTCATTATCTTTGTTTTTCCTA CAATCATAATTAATT
TTTTAGATA GGATAATTCTCTTTTCTCA TCACTACTGACA
GCACTTTGACTTG GTGTTTATTTGTTCCTTTAAG AAGCTCAAACACA
ATAGTGTTAAAA AAAATCAGTGCCTCATGGTTTTTCATTAAGCTA TTTTAATAGTCC
CTCTTTCATCTT AAATGTGCAAAGATCAGGAAAAAACAGTA AAGTTTGAAGCTG
TGTGATAATTAG CTAATATATCTTGCTTCTTTTTC CTCTTATAGCA
GCTGTTAACTCA CAATATTTTTTCAGCTTTA TGAGTAATTCC
CTGTTTAATAACA AAAGGATGAGCATAATTTTA GCTAACTTG
AAAATGTTT TTCTTTTTCTCCTTGGGCTTAAAAT ACAGGGT
AAGTGTTGGGA GTACCAGTAAAAAAGTTGT CGCCTGGCCAATACGCT
TTAATAAC CATCTTATCATTAAATTGAATATGGATAT TTAACAT
TCTAGATATGA GAAATAATATTTCTGTTATTTTAAAAT ATAAGATCCTTG
TAGCAACTTAT CAAGGTAAAATAATGAAATA ATATCTTTAAA
CGTGAAGATA AGGCCAGAATCATCCTGGTTTTTT AGTCTTAGCT
CAGTCATATG TGATATTTTATACATTTAAC CATATGTATC
GGGAACTTGT GCAGATCATCCTTTATTCCTACCTTGT ACAGGTCTTT
ATTATGATAAGATA ATAGATAACAGGGCAATACA AGGTTATCAGCAA
TTATCTGAGA GAGGTGTGACCACTGTTG TCTCAGTCAC
TCTGTGACTT GGTCTCACAAATCCTTGGTGAGATGTAGCT ATGTCAAAGG
AAAAACA TATATTTTATGGAAGGATC TGTAATC
AATCTT ATTTTTTCCTGAACTATTGCCAGATAAACA ATCTA
ATGATGGAACC TGATTTCACCAGTAATCTCT AAGTTACAAAAC
CTAT ATTTTTATTCACAAAAAAGAA AATTTAT
TCACATAGG TTTATTTTCTTCCTTTTTGTTTT CCTATTTTA
GAACTTA ACTAAGATTTTTAAAGATAAC TATGTAAATATA
ATGGCAG AATCTCAAAACTGAGTTAAAAAGA CTAAGATTTCTGGGCT
TTTAACAAA AGAAAAACTGGACCCCCA TTGTATTTT
CTCGTTTG TTTTTCTACACTCTTTTCCATCC CAAATTTC
CTCCATTT AATCTACCATTTCCAAGTTTGAAAT AAAGAATCCAA
GGACATAT TTATCAATTGAGGGGGAACAAG ATATTAGA
GGAAATTGC CAGATATTTTCATATAGAATGTTAGTGAAAT CAATCAGA
TCAAATGT AGACTTGCGAAATTACATAATA ACATCAGG
TTTATTA AGTTTTAAAGATGATAAA AATTGACA
GTAGATAG AAAATGAGATGGAAGATGTAGA CTATTTTG
GGTGAATG AAAATATTTTCTGTTGAAAAACAAC CAATTAAAA
GGGTATTT TAAATGGTATACATGTTAAAATTATATC AAATTTTT
AAACTTCT AATCTCAAAACTGAGTTAAAAGGA AGAACAGA
TATTTCTGTTT AAACCAAGTGACAATGGTATC AGATTCT
AACTTGT TTTGACTTGAATAAAGTACT ACATCAG
CAACGTA AAAACGAACAAGTCAAATA TACAACA
AATAGA TTCCTGAGGATTCTATTATAT TCTTACAA
ATTGATA GAATTACCTTATTTGATAATAGAAAGG TATGTGG
CACTTTG TTCTTGCACTATTGAATATAGATGCAT CAATCTA
TTTT TCTTCAAAACTAACCAAAACATGT AGAGACTT
GATTGT TTTTCATTTCTGAATTC CTAAGTTAGAA
AGGTAG AAAAAACTCACTCAAAATCTGG TCTTTCC
TTTGGGG GTTCCTTATTCCTCTTCT CCCTATC
```

**Figure 2.** Definite microinversions between human and chimpanzee genomes are always flanked by inverted repeats. Microinversion sequences are shown in bold black font; inverted repeats are shown in red bold font, and their overlaps with microinversions are underlined; mismatches within inverted repeats are shown in blue bold font; sequences separating microinversions from inverted repeats are shown in grey bold font; sequences external to inverted repeats are shown in regular black font.

intermediate length (4 to 8 nucleotides per flank) and the remaining 11% are short (3 nucleotides per flank). Most importantly, the inner edges of these inverted repeats juxtapose with the boundaries of the microinversions in practically every case (Figs. 2, 3). We conclude that the combination of the inverted repeat lengths and precise location makes their chance occurrence at microinversions highly unlikely.

Thus, the following two observations are principal when considering a mechanism(s) responsible for the microinversion formation. First, clean origin of most our microinversions implies that whatever mechanism caused this rearrangement, the inverted segment was copied precisely. Second, since an inverted segment is surrounded by inverted repeats in all the cases, they are most likely involved in the process.

The conventional wisdom about the inversion formation is that a DNA segment flanked by two homologous DNA sequences in an inverted orientation would flip upon recombination between these flanking sequences[28].

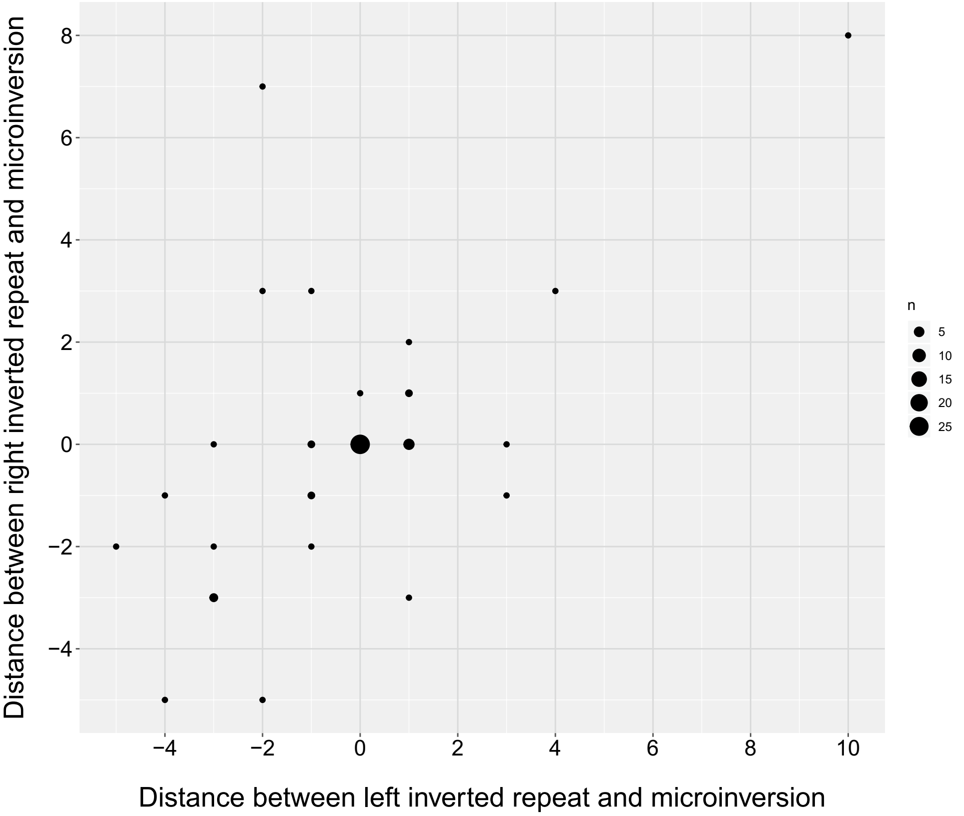


**Figure 3.** Positions of flanking inverted repeats are correlated with the outer boundaries of microinversions. The x-axis shows the distances in nucleotides between microinversions and inverted repeats on their left; the y-axis shows the distances between microinversions and inverted repeats on their right. Zero value for the left or right repeat means that a repeat is located immediately adjacent to the microinversion; positive value shows that there is a distance between a repeat and the microinversion; negative value shows that a repeat partly overlaps with the microinversion.

Two branches of recombination were shown to be involved. Homologous recombination (HR) between long inverted DNA elements, for example two non-LTR retrotransposons, are known to cause inversions[29]. Note, however, that the minimal homology length for HR in eukaryotes is somewhere between 50 and 250 base pairs[30,31]. Since only two of our 59 microinversions are flanked by the repeats exceeding this 50 bp, we can effectively rule out this mechanism. On the opposite end of the spectrum is site-specific recombination, where specialized DNA recombinases carry out DNA strand exchange between short inverted repeat (IR) sequences resulting in inversions[32]. While this mechanism is mostly used by prokaryotes, it is not without precedence in humans. For example, Rag1/Rag2 recombinase makes inversions during V(D)J recombination of immunoglobulin light chain genes[33]. We don't think, however, that site-specific recombination is the cause of microinversions. One counterargument is that given the lack of sequence similarity between IRs flanking different microinversions, one would have to assume the involvement of multiple site-specific recombinases. Even more importantly, two IRs need to be aligned in a parallel way for the recombination to proceed, thus, the central DNA segment must form a loop[34]. This is physically impossible for the microinversions of our size (< 33 bp), given that the persistence length of DNA corresponds to ~ 145 base pairs[35].

Perfect or imperfect inverted repeats were shown to induce template switching events during DNA replication. In a nutshell, the unique sequence composition of the inverted repeat allows DNA polymerase to jump between its halves situated on either template or nascent DNA strands leading to mutations and genome rearrangements[18]. It was suggested that template switching may account for the inversions of DNA segments flanked by inverted repeats[7,19,20,36]. Most recently, Walker et al.[27] described a method for modelling and detecting mutations that appeared upon template switching events, which identified thousands of such events (4017 unique cases) in the great apes genomes. Various genome rearrangements of different lengths, including thousands of potential microinversions, were detected among those template-switching events. Note, however, that the majority of these microinversions were shorter than 15 nucleotides, many as short as 1 nt-long, which, as discussed above, are hard to ascertain as definite inversions.

For the microinversion to occur via template switching, two template switches at inverted repeats during DNA replication must occur as discussed in Refs.[7,27,36]. Figure 4 shows how two template-switching events during DNA replication can create a microinversion. First, the leading DNA polymerase jumps from the first half of the flanking IR in the leading strand template to the identical sequence in the fold-back lagging strand template. Note that this jump is facilitated by the trombone-like configuration of the DNA strands at the replication fork[37]. Second,

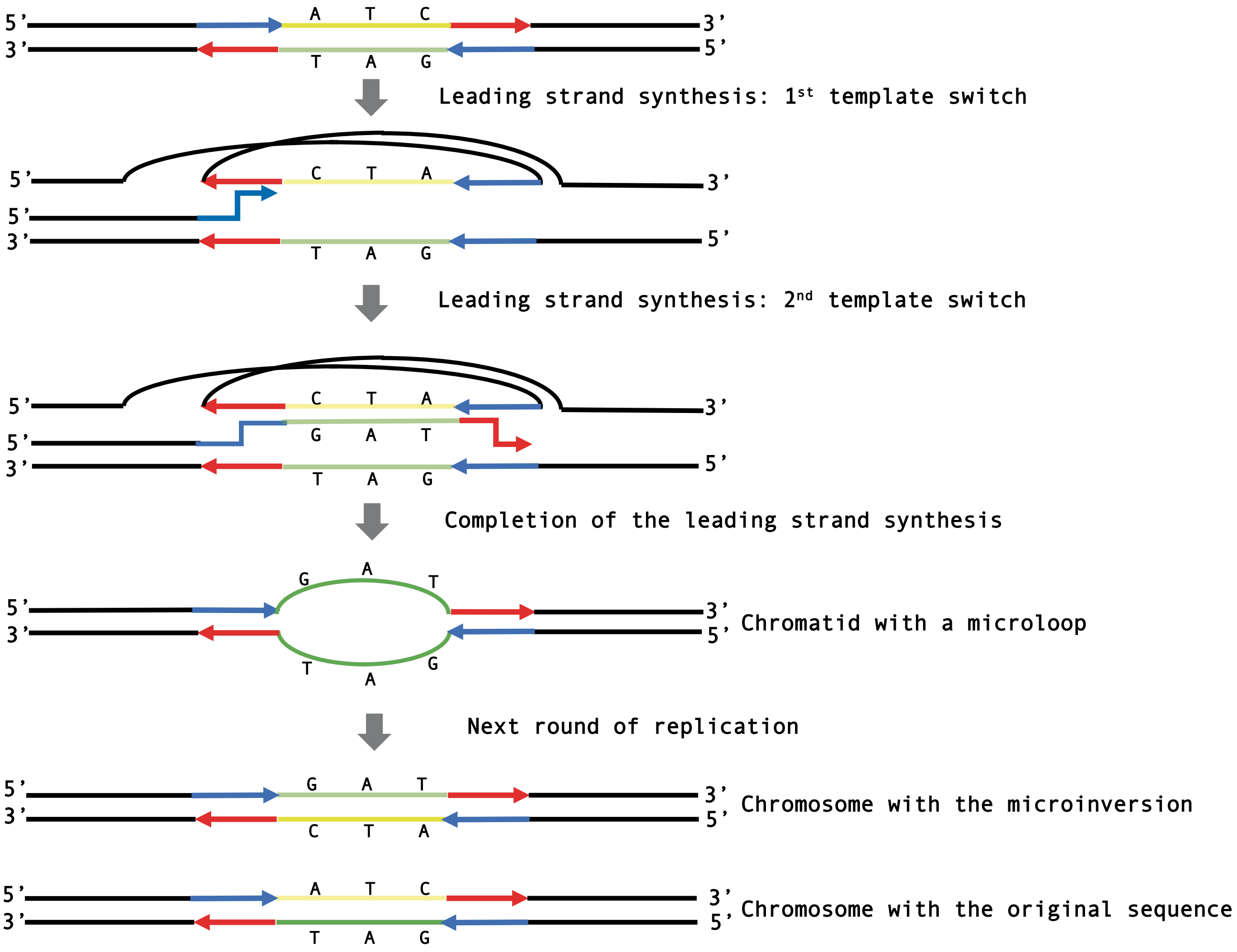

5'
3'
A T C
T A G
3'
5'
Leading strand synthesis: 1st template switch
5'
5'
3'
C T A
T A G
3'
5'
Leading strand synthesis: 2nd template switch
5'
5'
3'
C T A
G A T
T A G
3'
5'
Completion of the leading strand synthesis
5'
3'
G A T
T A G
3'
5'
Chromatid with a microloop
Next round of replication
5'
3'
G A T
C T A
3'
5'
Chromosome with the microinversion
5'
3'
A T C
T A G
3'
5'
Chromosome with the original sequence

**Figure 4.** Template-switching model for the formation of microinversions. Complementary strands of a DNA segment to undergo inversion are shown as yellow and green lines. Complementary strands of flanking inverted repeats are shown as red and blue arrows; the tips of these arrows designate 3′ ends of inverted repeats (see text for details).

the leading DNA polymerase synthesizes the reverse complement of its anticipated template until it reaches the second half of the IR. Third, the DNA polymerase jumps to the identical sequence in the leading strand template resuming normal replication. This sequence of events gives rise to a newly synthesized DNA molecule with a microloop consisting of two inverted strands at the site of future microinversion. After the second round of replication, two DNA molecules arise, one of which contains the microinversion.

While similar models were discussed in earlier studies[7,19,27,36], they have two potential caveats. First, since two jumps of DNA polymerase must occur within a half of an inverted repeat, one would expect it to be sufficiently long. It is hard to imagine how the two jumps could realistically happen within a sequence that is only 3 or 4 nt-long. The second problem is that template switching of imperfect inverted repeats is known to convert them into perfect ones[38]. We believe, therefore, that the template-switch model could explain the formation of those definite microinversions that are flanked by longer, perfect inverted repeats (such as microinversions in the upper part of Fig. 2), while it is hard to apply to explain the formation of microinversions that are flanked by very short IRs, or longer, but imperfect IRs.

As an alternative, it was previously discussed that inversions can result from an incorrect repair of two double-strand breaks (DSBs) formed within flanking inverted repeats via non-homologous end-joining (NHEJ)[9]. We don't think, however, that this scenario applies to our microinversions, since formation of two DSBs separated by ~ 20 nucleotides is highly unlikely even in a strong DNA damaging environment such as ionizing radiation. A more likely scenario—repair of a single DSB positioned proximal or distal to a hairpin-forming sequence—was shown to form foldback inversions (a.k.a. inverted duplications)[39], that are principally different from the microinversions described here.

We propose a new mechanism for the microinversion formation during DSB repair that is applicable to all definite microinversions found by us (Fig. 5). The trigger is a DSB in the region to undergo inversion, which is present in one sister chromatid in the G2 phase of the cell cycle. Subsequent end-resection would expose the flanking inverted repeat in a single-stranded state. This is followed by an asymmetric invasion of the IR from the left flank of the inverting segment into the sister chromatid, such that it pairs with the right flank IR. The invaded

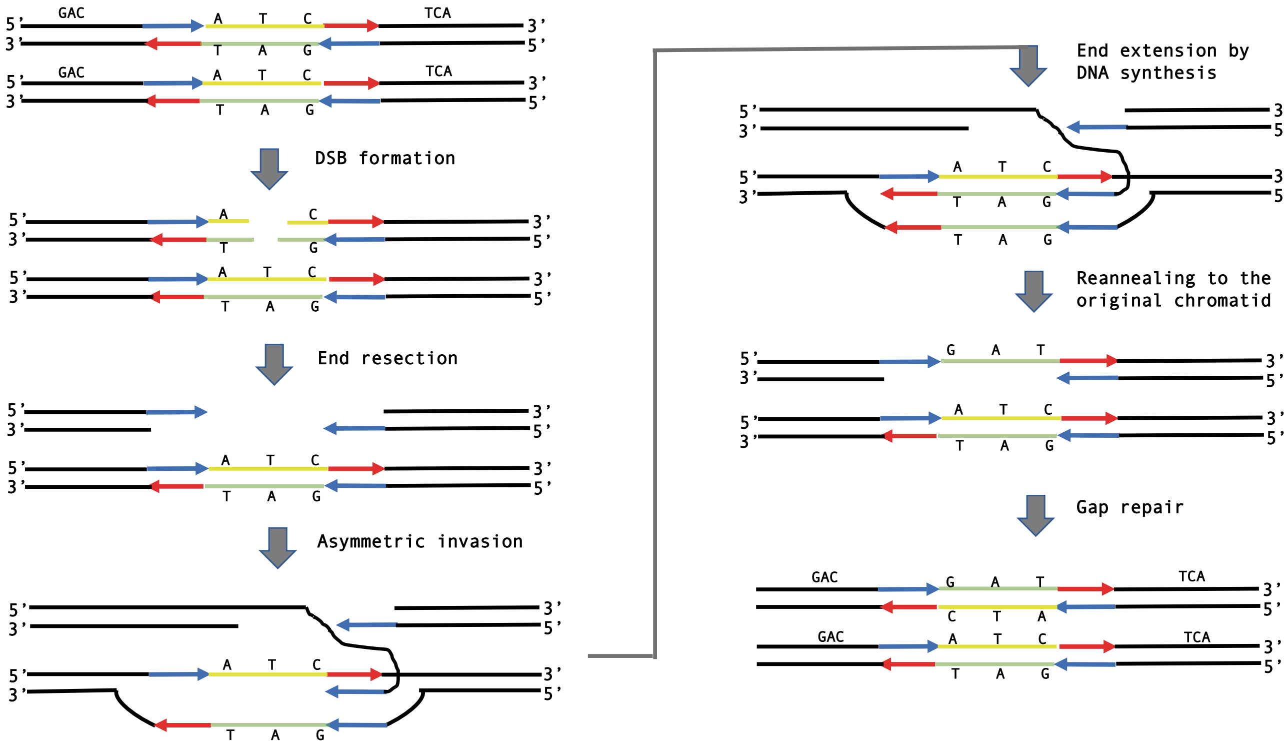


**Figure 5.** Microhomology-mediated BIR model for the microinversion formation. Complementary strands of a DNA segment to undergo inversion are shown as yellow and green lines. Complementary strands of flanking inverted repeats are shown as red and blue arrows; the tips of these arrows designate 3′ ends of inverted repeats. See text for details.

DNA strand is then extended to the left flank IR followed by its reannealing with the original chromatid. The following gap repair synthesis completes the inversion process. Note that in contrast to the template-switching model, our DSB repair model does not involve the presence of single-stranded loop-outs, that are known to be hypermutagenic[40]. Consequently, the DSB repair model is consistent with the lack of mutations in the inverting segments characterized in this study.

At a first glance, this model is reminiscent of the synthesis-dependent strand annealing (SDSA), which leads to the formation of non-crossover products during homologous recombination[41]. There are two significant differences, however. First, the invasion and pairing occur between IRs on the opposite sides of the break, in contrast with regular SDSA, where DNA segments on the same side of the break are being paired. We believe that this asymmetric invasion should not be a problem for microinversions, the length of which is well within the limits of single-stranded DNA flexibility[42]. Second, in canonical SDSA the region of homology at the invading end should be significantly longer (> 50 nucleotides)[41] than IRs flanking our microinversions. Altogether, our model is much more like an alternative pathway of DSB repair called break-induced replication (BIR)[43], more specifically microhomology-mediated BIR[44,45]. While initially proposed to explain genome rearrangements observed in patients with hereditary diseases[44], microhomology-mediated BIR was recently characterized experimentally, and the microhomologies required for invasion and strand extension were comparable with our flanking IRs[46,47]. We believe therefore, the model presented in Fig. 5 could satisfactorily explain the majority of the definite microinversion events observed in this study.

Of note, while DNA repair events leading to the microinverstion formation shown in Fig. 5 take place between sister chromatids in the G2 phase prior to mitosis, we speculate that similar microhomology-mediated BIR events could also happen between homologous chromosomes in meiosis, resulting in non-crossover products with microinversions. Finally, this model can be particularly applicable to multiple microinversions observed in cancer genomes[16], given that DSBs are among the main drivers of the genome instability causing cancer development[48].

## Methods

Human (hg38) and common chimpanzee (panTro5) genomes (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz; http://hgdownload.soe.ucsc.edu/goldenPath/panTro5/bigZips/panTro5.fa.gz) and their pairwise alignment (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/vsPanTro5/), as well as data for chimpanzee (panTro4) and bonobo (panPan2) (genomes https://hgdownload.soe.ucsc.edu/goldenPath/panTro4/bigZips/panTro4.fa.gz; https://hgdownload.soe.ucsc.edu/goldenPath/panPan2/bigZips/panPan2.fa.gz and alignment; http://hgdownload.soe.ucsc.edu/goldenPath/panPan2/vsPanTro4/), were obtained from the UCSC Genome Browser. Annotation for human genome (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/

knownCanonical.txt.gz) and data on simple repeats (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/simpleRepeat.txt.gz) were also obtained from the same source.

The alignment of human and chimpanzee genomes available from the UCSC Genome Browser consists of alternating regions aligned by LASTZ[49] and regions that were left unaligned. This alignment, which represented the point of departure for our analysis, did not attempt to discover microinversions, because reverse-complemented sequence segments were not considered. Thus, a microinversion usually was represented by a "bubble", a pair of unaligned genome segments.

However, in some cases the quality of the alignment of a genome segment with the corresponding reverse-complemented sequence was high enough to include these "false" alignments into the UCSC Genome Browser, as exemplified by this case:

```
----TGTCTTTAAAAAACCTGGCCA--------ACAGACAAACC
GGTTTGTCTGT--------TGGCCAGGTTTTTTAAAGACA----
```

Here, the reverse-complemented chimpanzee genome segment matches the human genome segment precisely, suggesting that a microinversion took place. However, the two genome segments can also be aligned, and this alignment was presented by the UCSC Genome Browser. That is why it is necessary to treat not only genome segments that were left unaligned but also loose segments of the alignment as possible microinversions.

In the human-chimpanzee genome alignment available from the UCSC Genome Browser, only ~ 80% of the genomes are aligned to each other. We first refined this alignment by aligning the corresponding human and chimpanzee genome segments of lengths above 15 nucleotides that were left unaligned, as long as the difference between their lengths did not exceed 2%. Biopython package pairwise2 (with a matrix BLOSUM62 and penalties for gap opening and extending – 10 and – 0.5, respectively) was used for this purpose.

After this, we ascertained microinversions from the refined alignment as follows. We first searched for suspicious "bubbles" in it, flanked by segments where the alignment is reliable and unambiguous. Each mismatch within an alignment was treated as the possible beginning of a bubble. We compared regions of alignments that started from a mismatch M and assigned a penalty of – 0.5 for a match and + 1 for a mismatch or a gap (these parameters resulted in the maximal number of detected microinversions), until the total score became negative. After this, we defined a potential bubble as the segment of the alignment from the initial mismatch M to the position where the total score was maximal. A potential bubble was considered to be real if the following conditions were met: its length was at least 5 positions on the alignment, and it was flanked by alignment segments of lengths above 100 nucleotides that were of a good quality (without tracks of gaps or simple repeats). Search for the next bubble commenced from the next mismatch after the end of the previous bubble, if one was detected, or, otherwise, after the mismatch M.

We ignored bubbles in which the genome segment of at least one of the two species was at least partially masked by the RepeatMasker, as well as those located in poorly assembled regions of genomes ("Unknown" chromosomal regions in at least one species). Every remaining bubble was investigated for the possible presence of a microinversion. We made the reverse complement of the chimpanzee genome segment which resides within a bubble and aligned it to the corresponding human genome segment by the Smith-Waterman algorithm[50] using Biopython package pairwise2 as described above. Those bubbles for which the resulting alignment contained less than 50% of gaps were treated as potential microinversion, although only a small proportion of them are likely to be real microinversions.

## Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files.

Received: 20 August 2021; Accepted: 28 December 2021
Published online: 12 January 2022

## References

1. Dobzhansky, T. & Sturtevant, A. H. Inversions in the chromosomes of drosophila Pseudoobscura. *Genetics* **23**, 28–64 (1938).
2. Chaisson, M. J., Raphael, B. J. & Pevzner, P. A. Microinversions in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19824–19829. https://doi.org/10.1073/pnas.0603984103 (2006).
3. Brandler, W. M. *et al.* Frequency and complexity of de novo structural mutation in autism. *Am. J. Hum. Genet.* **98**, 667–679. https://doi.org/10.1016/j.ajhg.2016.02.018 (2016).
4. Stenson, P. D. *et al.* The human gene mutation database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677. https://doi.org/10.1007/s00439-017-1779-6 (2017).
5. Hara, Y. & Imanishi, T. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *BMC Evol. Biol.* **11**, 308. https://doi.org/10.1186/1471-2148-11-308 (2011).
6. Hou, M., Yao, P., Antonou, A. & Johns, M. A. Pico-inplace-inversions between human and chimpanzee. *Bioinformatics* **27**, 3266–3275. https://doi.org/10.1093/bioinformatics/btr566 (2011).
7. Kolb, J. *et al.* Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome Res.* **17**, 469–483. https://doi.org/10.1007/s10577-009-9039-9 (2009).
8. Lee, J., Han, K., Meyer, T. J., Kim, H. S. & Batzer, M. A. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE* **3**, e4047. https://doi.org/10.1371/journal.pone.0004047 (2008).
9. Braun, E. L. *et al.* Homoplastic microinversions and the avian tree of life. *BMC Evol. Biol.* **11**, 141. https://doi.org/10.1186/1471-2148-11-141 (2011).
10. Macdonald, S. J. & Long, A. D. Fine scale structural variants distinguish the genomes of Drosophila melanogaster and *D. pseudoobscura*. *Genome Biol.* **7**, R67. https://doi.org/10.1186/gb-2006-7-7-R67 (2006).

11. Chen, J. M., Ferec, C. & Cooper, D. N. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. *Hum. Mutat.* **30**, 1435–1448. https://doi.org/10.1002/humu.21088 (2009).
12. Giner-Delgado, C. *et al.* Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**, 4222. https://doi.org/10.1038/s41467-019-12173-x (2019).
13. He, F., Li, Y., Tang, Y. H., Ma, J. & Zhu, H. Identifying micro-inversions using high-throughput sequencing reads. *BMC Genomics* **17**, 4. https://doi.org/10.1186/s12864-015-2305-7 (2016).
14. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13633–13635. https://doi.org/10.1073/pnas.172510699 (2002).
15. Feuk, L. Inversion variants in the human genome: Role in disease and genome architecture. *Genome Med.* **2**, 11. https://doi.org/10.1186/gm132 (2010).
16. Qu, L., Zhu, H. & Wang, M. Micro-inversions in human cancer genomes. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018**, 1323–1326. https://doi.org/10.1109/EMBC.2018.8512514 (2018).
17. Qu, L. *et al.* The landscape of micro-inversions provide clues for population genetic analysis of humans. *Interdiscip. Sci.* **12**, 499–514. https://doi.org/10.1007/s12539-020-00392-6 (2020).
18. Lovett, S. T. Template-switching during replication fork repair in bacteria. *DNA Repair (Amst.)* **56**, 118–128. https://doi.org/10.1016/j.dnarep.2017.06.014 (2017).
19. Loytynoja, A. & Goldman, N. Short template switch events explain mutation clusters in the human genome. *Genome Res.* **27**, 1039–1049. https://doi.org/10.1101/gr.214973.116 (2017).
20. Tremblay-Belzile, S., Lepage, E., Zampini, E. & Brisson, N. Short-range inversions: rethinking organelle genome stability: Template switching events during DNA replication destabilize organelle genomes. *BioEssays* **37**, 1086–1094. https://doi.org/10.1002/bies.201500064 (2015).
21. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1**, e56. https://doi.org/10.1371/journal.pgen.0010056 (2005).
22. Silva, J. C. & Kondrashov, A. S. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.* **18**, 544–547. https://doi.org/10.1016/s0168-9525(02)02757-9 (2002).
23. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304. https://doi.org/10.1093/genetics/156.1.297 (2000).
24. Moorjani, P., Amorim, C. E., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10607–10612. https://doi.org/10.1073/pnas.1600374113 (2016).
25. Prufer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531. https://doi.org/10.1038/nature11128 (2012).
26. Calvete, O., Gonzalez, J., Betran, E. & Ruiz, A. Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in Drosophila. *Mol. Biol. Evol.* **29**, 1875–1889. https://doi.org/10.1093/molbev/mss067 (2012).
27. Walker, C. R., Scally, A., De Maio, N. & Goldman, N. Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLoS Genet.* **17**, e1009221. https://doi.org/10.1371/journal.pgen.1009221 (2021).
28. Griffiths, A. J. F., Gelbart, W. M., Miller, J. H. & Lewontin, R. C. *Modern Genetic Analysis* (W.H. Freeman, 1999).
29. Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin. Cancer Biol.* **20**, 211–221. https://doi.org/10.1016/j.semcancer.2010.03.001 (2010).
30. Kouprina, N. & Larionov, V. Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma* **125**, 621–632. https://doi.org/10.1007/s00412-016-0588-3 (2016).
31. Prado, F., Cortes-Ledesma, F., Huertas, P. & Aguilera, A. Mitotic recombination in *Saccharomyces cerevisiae*. *Curr. Genet.* **42**, 185–198. https://doi.org/10.1007/s00294-002-0346-3 (2003).
32. Craig, N. L. Site-specific inversion: Enhancers, recombination proteins, and mechanism. *Cell* **41**, 649–650. https://doi.org/10.1016/s0092-8674(85)80040-4 (1985).
33. Helmink, B. A. & Sleckman, B. P. The response to and repair of RAG-mediated DNA double-strand breaks. *Annu. Rev. Immunol.* **30**, 175–202. https://doi.org/10.1146/annurev-immunol-030409-101320 (2012).
34. Wasserman, S. A. & Cozzarelli, N. R. Biochemical topology: Applications to DNA recombination and replication. *Science* **232**, 951–960. https://doi.org/10.1126/science.3010458 (1986).
35. Vologodskaia, M. & Vologodskii, A. Contribution of the intrinsic curvature to measured DNA persistence length. *J. Mol. Biol.* **317**, 205–213. https://doi.org/10.1006/jmbi.2001.5366 (2002).
36. Sui, Y. *et al.* Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28191–28200. https://doi.org/10.1073/pnas.2018633117 (2020).
37. Yao, N. Y. & O'Donnell, M. Replisome dynamics and use of DNA trombone loops to bypass replication blocks. *Mol. Biosyst.* **4**, 1075–1084. https://doi.org/10.1039/b811097b (2008).
38. 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. https://doi.org/10.1016/j.jmb.2005.11.071 (2006).
39. Li, B. Z., Putnam, C. D. & Kolodner, R. D. Mechanisms underlying genome instability mediated by formation of foldback inversions in *Saccharomyces cerevisiae*. *Elife*. https://doi.org/10.7554/eLife.58223 (2020).
40. Saini, N. & Gordenin, D. A. Hypermutation in single-stranded DNA. *DNA Repair (Amst.)* **91–92**, 102868. https://doi.org/10.1016/j.dnarep.2020.102868 (2020).
41. McVey, M., Khodaverdian, V. Y., Meyer, D., Cerqueira, P. G. & Heyer, W. D. Eukaryotic DNA polymerases in homologous recombination. *Annu. Rev. Genet.* **50**, 393–421. https://doi.org/10.1146/annurev-genet-120215-035243 (2016).
42. Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271**, 795–799. https://doi.org/10.1126/science.271.5250.795 (1996).
43. Anand, R. P., Lovett, S. T. & Haber, J. E. Break-induced DNA replication. *Cold Spring Harb. Perspect Biol.* **5**, a010397. https://doi.org/10.1101/cshperspect.a010397 (2013).
44. 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. https://doi.org/10.1371/journal.pgen.1000327 (2009).
45. Sakofsky, C. J. & Malkova, A. Break induced replication in eukaryotes: Mechanisms, functions, and consequences. *Crit. Rev. Biochem. Mol. Biol.* **52**, 395–413. https://doi.org/10.1080/10409238.2017.1314444 (2017).
46. Sakofsky, C. J. *et al.* Translesion polymerases drive microhomology-mediated break-induced replication leading to complex chromosomal rearrangements. *Mol. Cell* **60**, 860–872. https://doi.org/10.1016/j.molcel.2015.10.041 (2015).
47. Segar, M. W., Sakofsky, C. J., Malkova, A. & Liu, Y. MMBIRFinder: A tool to detect microhomology-mediated break-induced replication. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**, 799–806. https://doi.org/10.1109/TCBB.2014.2359450 (2015).
48. Aparicio, T., Baer, R. & Gautier, J. DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst.)* **19**, 169–175. https://doi.org/10.1016/j.dnarep.2014.03.014 (2014).
49. Harris, R. S. Improved pairwise alignment of genomic DNA (2007).
50. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5 (1981).

## Acknowledgements

This work was supported by the Grant 5R35GM130322 from NIGMS to SMM. Computational resources of the "Makarich" HPC cluster were provided by the Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia. We are grateful to Andrey Mironov, Mitch McVey and Kate Mirkin for fruitful discussions, and to Julia Hisey for her excellent editorial help.

## Author contributions

N.A.P. designed the study, collected and analyzed the data and wrote the paper; A.S.K designed the study, analyzed the results and wrote the paper; S.M.M. analyzed the results, proposed the models and wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-04621-w.

**Correspondence** and requests for materials should be addressed to N.A.P. or S.M.M.